Mo-Lin Ge
Yang-Hui He *Eds.*



Dialogues Between Physics and Mathematics

C. N. Yang at 100

Springer

Dialogues Between Physics and Mathematics

Mo-Lin Ge • Yang-Hui He
Editors

Dialogues Between Physics and Mathematics

C. N. Yang at 100

Springer

Editors
Mo-Lin Ge
Chern Institute of Mathematics
Nankai University
Tianjin, China

Yang-Hui He
London Institute for Mathematical Sciences; City, University of London; School of Physics, Nankai University; Merton College, University of Oxford
Oxford, UK

ISBN 978-3-031-17522-0 ISBN 978-3-031-17523-7 (eBook)
https://doi.org/10.1007/978-3-031-17523-7

Mathematics Subject Classification: 00-02, 81-02, 11Gxx, 11Mxx, 57K10

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Cover illustration: Private collection

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Opto, domine, et hunc natalem et plurimos alios quam felicissimos agas.
- Pliny, Sec. Ep. Triiano Imp.

敬祝楊先生福如東海，壽比南山。

Preface

The breadth and depth of Professor Chen-Ning Frank Yang's contributions to theoretical physics have been legendary. His Nobel Prize-winning work on parity violation in 1957 is but a glimpse into his profound insights and influences. Though his initial contributions were to theoretical particle and nuclear physics, C. N. Yang's work had bestowed lasting effects on high energy physics, condensed matter physics, mathematical physics, as well as to pure mathematics, especially to topology, algebraic geometry, differential geometry, and group theory.

This volume, comprised of essays from world-renowned mathematicians and physicists, many of whom are friends and colleagues of C. N. Yang over the decades, is a celebration of his 100th birthday. After a brief summary of Prof. Yang's life and achievements, the book begins with recollections of his years at Stony Brook by Prof. Peter van Nieuwenhuizen, who succeeded Yang in 1999 as the director of the Institute for Theoretical Physics, now named the *C. N. Yang Institute for Theoretical Physics* at Stony Brook. Prof. van Nieuwenhuizen is one of the founders of the super-gravity theory, which is a brainchild between supersymmetry, Yang-Mills theory, and general relativity.

Chapter 2 is by the Fields Medalist Prof. Vladimir Drinfeld, entitled "A stacky approach to crystals", which is in the field of arithmetic geometry. It initiates the stacky approach to so-called crystalline cohomology (algebraic stacks are generalizations of algebraic varieties in that points on them form a groupoid rather than a set, and may thus have nontrivial automorphism groups). This is a p-adic version of de Rham cohomology theory, well-known to geometers and physicists, that governs differential forms entering Yang-Mills gauge theories. It is the hope of many leading mathematicians such as Minhyong Kim, Don Zagier, et al., that arithmetic geometry and number theory will play an increasing role in fundamental physics.

Chapter 3 is from the topologist and Norbert Wiener awardee Prof. Louis Kauffman, entitled "The Potts Model, the Jones Polynomial and Link Homology", and explores the relation between the Potts model in statistical mechanics, the Temperley-Lieb algebra, as well as the Jones polynomial and Khovanov homology in knot theory. This involves two subjects close to Prof. Yang's heart: (1) models

from statistical mechanics such as Ising and Potts and (2) groups from braids and knots.

Chapter 4 is from the physics Nobel Laureate Prof. Sir Anthony Leggett, entitled "The Penrose-Onsager-Yang Approach to Superconductivity and Superfluidity". As he states in the abstract, "While the work of Professor C. N. Yang in condensed matter physics may not be quite as universally known as that in the area of particle physics, it includes a number of seminal papers, of which the one I like best is the 1962 Reviews of Modern Physics paper entitled *Concept of off diagonal long range order and the quantum phases of liquid He and of superconductors*.... This paper in effect asks the general question "what exactly is it that superfluid helium and superconductors have in common, and what is at the root of their anomalous properties?" Although this question has also been answered by others in rather different terms, I believe that the answer given by Yang is by far the clearest and most intuitive."

Chapter 5 is by the mathematician Prof. Yuri Manin, recipient of numerous prizes, such as Nemmers, Schock, Bolyai, etc. Joint with N. Combe and M. Marcolli, the chapter is entitled "Quantum Operads" and introduces the study of quantum symmetries of operads. Quantum groups and braid groups are intimately related to the Yang-Baxter equations. From a categorical point of view, the study of operads that generalize the structure of groups should offer new insights into the construction of quantum codes and thus into quantum information theory.

Chapter 6 is from the physicist Prof. Jian-Wei Pan, a pioneer in experimental quantum sciences and winner of many awards such as the Zeiss Prize, the International Quantum Communication Award, R. W. Wood Prize, the Future Science Prize. It is a review on quantum computational advantage, bringing together mathematics, computational complexity, and quantum optics. Prof. Yang has had a long interest in the development of the quantum computer, both the theoretical and experimental aspects, and has appreciated the works of Pan.

Chapter 7 is from the physics Nobel Laureate Prof. Sir Roger Penrose, which begins with a review of how twistor theory is important in describing classical and quantum particles with spin and helicity. This twistor formalism has been over the years, a novel and complexified perspective on the structure of space-time and quantum gravity. There has been much recent activity on relating this approach to scattering amplitudes in the Yang-Mills and string theory. In this chapter, the quantum twistor algebra is shown to have a hitherto-unnoticed split-octonion algebra and G_2^* structure.

Chapter 8 is from the physicist Prof. Alexander Polyakov, winner of numerous awards such as the Breakthrough Prize, the Planck Medal, the Lorenz Medal, etc. Based on joint work with F. Popov, it is on "Kronecker anomalies and gravitational striction", which studies quantum field theories in which the number of degrees of freedom changes discontinuously across the momentum space. This leads to a concept which they call "Kronecker anomaly" and which leads to non-local effective actions that can be represented as a theory with random, self-tuning coupling constants.

Chapter 9 is by the physics Nobel Laureate Prof. Gerard 't Hooft, on "Projecting Local and Global symmetries to the Planck Scale". It is based on a recent series of papers by the authors on a novel approach to reconciling quantum mechanics with general relativity, a topic that had been in Prof. Yang's mind for decades. Very appropriately, the chapter begins with the words "In 1954, Chen-Ning Yang and Robert L. Mills published their idea that Maxwell's equations for the electromagnetic force could be generalised...."

Chapter 10 is by the physics Nobel Laureate Prof. Frank Wilczek, entitled "Gauge Symmetry in Shape Dynamics", and is on the subject of gauge symmetry—one of Prof. Yang's key contributions to physics—on deformable bodies. Indeed, as a tribute, the abstract begins with "C. N. Yang's ideas about local gauge symmetry and non-integrable phases have been enormously fertile sources of inspiration in fundamental physics and in the quantum theory of matter...."

Chapter 11 is by the Fields Medalist Prof. Edward Witten, provocatively entitled "Why Does Quantum Field Theory in Curved Spacetime Make Sense? And What Happens to the Algebra of Observables in the Thermodynamic Limit?" It provides a gentle introduction to some classical results about quantum field theory in curved spacetime and about the thermodynamic limit of quantum statistical mechanics. This fundamental question, unifying gauge theory and general relativity, has long been on the mind of C. N. Yang, as exemplified by his 1974 paper on the pure space formalism.

Chapter 12 is by the experimental physicist Prof. Qikun Xue, inaugural winner of the Chinese Future Science Award and of the Fritz London Memorial Prize. It is on the "Quantum Anomalous Hall Effect". Prof. Yang has long appreciated the work of Prof. Xue and has been a colleague of his at Tsinghua University. It was at Tsinghua that Prof. Xue made the experimental discovery of the quantum anomalous Hall effect (QAHE).

Finally, Chap. 13 is from the physicist Prof. Zhongxian Zhao, winner of the State Supreme Science and Technology Award and the Bernd T. Matthias Prize. It is based on joint work with J. Hu and X. Dong, entitled "Magic superconducting states in Cuprates", where superconductivity is stronger and more robust. Indeed, as discussed in Chap. 4, Prof. Yang had a long interest in superconductivity.

We have included a plenitude of photographs featuring Prof. Yang over the years, his times in the USA and China, his long service to Nankai and Tsinghua, as well as his life-long friendship with the late Professor Shiing-Shen Chern (1911–2004), who along with Prof. Yang stands as one of the giants of modern mathematics and physics. Indeed, the friendship between Prof. Chern and Prof. Yang, personally and intellectually, deserves a biography of its own. Yang's father was a teacher to Chern, and in return, Chern became a teacher to Yang. One of the crowning achievements of Prof. Yang, and indeed, of modern science, is the Yang-Mills theory, which is intimately related to Chern's theory of connections on bundles over manifolds.

This friendship is immortalized at the entrance of the *Chern Institute of Mathematics* at Nankai University, where a mural captures a moment of the countless dialogues between Professors Chern and Yang. This image, in part, is the inspiration of the title of this volume, celebrating Prof. Yang's birthday, and embracing the "dia-

logue" between physics and mathematics. Coincidentally, last year was the 110th anniversary of Prof. Chern's birth, and the *Nankai Symposium on Mathematical Dialogues* was held, of which we had been two of the organizers.[1]

As we pay our tributes to Prof. Chern, we also fondly remember three other dear friends of Prof. Yang, who had been waiting to celebrate this birthday and who had wished to contribute chapters. Sadly, they have of late been recalled to eternity. They are Professors Sir Michael Atiyah[2] (1929–2019), Freeman Dyson (1923–2020), and Ludvig Faddeev (1934–2017); we have included their photos in this volume. *Requiescant in pace, et lux perpetua luceat eis*.

We wish, above all, a heartfelt Happy 100th Birthday to Professor Chen-Ning Frank Yang! May you, and your beloved dialogues between mathematics and physics, be as in *Shih-Ching*, the Book of Odes:

如月之恆，如日之升。如南山之壽，不騫不崩。如松柏之茂，無不爾或承。
《詩經·小雅·天保》

Tianjin, China — Mo-Lin Ge
Oxford, UK — Yang-Hui He
August, 2022

[1] See eponymous Proceedings, Springer, 2022, ISBN 9789811923272.

[2] We were fortunate that Sir Michael was able to contribute, just before his passing, to a volume which we had the honour to co-edit with Prof. Yang: C.N.Yang, M.-L. Ge and Y.-H. He, *Topology and Physics*, World Scientific, 2019,ISBN 978-9813278509.

Chen Ning Yang 楊振寧

A Brief Curriculum Vitae

Sept 22, 1922	Born in Hefei, Anhui, China
1938–1942	Undergraduate, Southwest Associated University[3]
1942–1944	Graduate Student, SAU, China
1946–1948	PhD, University of Chicago, USA (Advisor: Edward Teller)
1949–1966	Institute for Advanced Study, Princeton, USA
1966–1999	Albert Einstein Professor, Institute of Theoretical Physics,[4] Stony Brook, SUNY, USA (Emeritus 1999–)
1986	Established the Theoretical Physics Section at Nankai Institute of Mathematics,[5] China (Invited by S.-S. Chern)
1986–	Distinguished Professor-at-large, Chinese University of Hong Kong
1995	Founded the Center for Advanced Studies of Tsinghua University, China
1998–	Professor of Tsinghua University, China

Main Awards

1957	Nobel Prize
1980	Rumford Premium
1986	National Medal of Science, USA
1993	Benjamin Franklin Medal
1994	Bower Award
1995	Einstein Medal
1996	N. Bogoliubov Prize
1997	Small Planet #3421 named after Chen Ning Yang
1999	Lars Onsager Prize
2001	King Faisal International Prize
2006	*You Bring Charm to the World* Award—Prize of Success for Life, China

Key Contributions

1948	Angular distribution in nuclear reaction
1949	Fermi-Yang model
1949	Selection rules of a particle decaying into two photons
1950	Reflection properties of Spin-1/2 fields (with J. Tiomno)
1951	Exact calculation of spontaneous magnetization of 2-dimensional Ising model

[3] SAU was a consortium, due to WWII, of Peking, Tsinghua and Nankai Universities, (1938–1946).

[4] Now named the C. N. Yang Institute of Theoretical Physics.

[5] Now dubbed the Chern Institute of Mathematics.

1952 Statistical theory of equations of state and phase transition I and II (circle theorem) (with T. D. Lee)
1954 Yang-Mills theory (with R. L. Mills)
1956 Parity non-conservation in weak interaction (with T. D. Lee)
1957 Pseudo potential calculation of hard-sphere Bosonic particles (with K. Huang and T. D. Lee)
1960 Imperfect Bose system
1969 Quantization of magnetic flux in superconducting cylinders (with N. Byers)
1962 Off-diagonal long-range order (ODLRO)
1963 Critical point in liquid-gas transition (with C. P. Yang)
1964 Phenomenology in CP violation (with T. T. Wu)
1966 Bethe hypothesis for anisotropic spin-spin chain (with C. P. Yang)
1967 Foundation of "Yang-Baxter equation" (Partly with C. P. Yang)
1975 Gauge fields and fiber bundles (with T. T. Wu)
1977 Self-dual gauge fields
1989 η-pair (with S. C. Zhang)
1990 $SO(4)$ symmetry in Hubbard model (with S. C. Zhang)
2003 Selected papers II
2008 曙光集 ("The Twilight Collection") collected works (in Chinese)
2018 晨曦集 ("The Dawn Collection") collected works (with Wong Fan, in Chinese) Second edition, 2021.

Membership/Foreign Membership of Academies

U. S. Academy of Sciences
Chinese Academy of Sciences, Beijing
Russian Academy of Sciences
Royal Society (London)
American Philosophical Society
Academia Sinica, Taiwan
Brazilian Academy of Sciences
Venezuelan Academy of Sciences
Polish Academy of Sciences
Royal Spanish Academy of Science
Korean Academy of Science and Technology
Pontifical Academy of Sciences
Japan Academy

Chairmanships

1965 Panel of Theoretical Physics, Physics Survey Committee, National Academy of Sciences
1972–1976 Division of Particles and Fields of the International Union of Pure and Applied Physics
1970–1971 Division of Particles and Fields of the American Physical Society
1980–1983 Fachbeirat of the Max Planck Institute of Physics, Munich

Presidencies

1977–1980	National Association of Chinese Americans (NACA)
1989–1994	Association of Asia Pacific Physical Societies
1996–	Asia Pacific Center for Theoretical Physics (APCTP)

Board Memberships

Governing Council of the Courant Inst. of Mathematical Science 1963–
Science Advisory Committee of IBM 1966–1971
High Energy Physics Advisory Panel (HEPAP) 1968–1970
Board of Trustees of Rockefeller University 1970–1976
Woods Hole Oceanographic Institution 1962–1978
Board of Trustees of Ben Gurion University 1981–
Board of Directors of the AAAS 1975–1979
Board of Trustees of the Salk Institute 1978–1989
Board of Directors of the Neuroscience Institute 1983–1988
Board of Directors of the Scientific American Inc. 1983–1990
Council of Scholars, The Library of Congress 1990–
Board of Trustees, Stony Brook Foundation 1981–1999
Board of Trustees, Brookhaven Science Associates 1998–1999
Board of Shaw Prize Foundation, Hong Kong 香港邵逸夫獎基金會 2002–2014

Lectureships

Gibbs Lecturer, American Mathematical Society
Loeb Lecturer, Harvard University (1957, 1972 and 1997)
Vanuxem Lecturer, Princeton University
Lincoln Lecturer, Bd. of Foreign Scholars (Fulbright Bd.) of the State Department
Pauli Lecturer, ETH, Zurich
Courant Lecturer, New York University
Fermi Lecturer, Scuola Normale Superiore, Pisa
Oppenheimer Memorial Lecturer, Los Alamos
Lorentz Professor, Leiden
Klein Memorial Lectures, Stockholm
Ta-You Wu Lecturer, University of Michigan
Goudsmit Lecturer, University of Michigan
Wu-Zhi Yang Lecturer, Fudan University, Shanghai
Primakoff Lecturer, APS
Oppenheimer Lecture, Berkeley

Honourary Doctorates

Princeton University (1958), Brooklyn Polytechnic Institute (1965), University of Wroclaw, Poland (1974), Gustavus Adolphus College, Minnesota (1975), University of Maryland (1979), University of Durham, England (1979), Fudan University, China (1984), Eidg. Technische Hochschule (ETH, 1987), Moscow State University

(1992), Drexel University (1995), Tsing Hua University, Taiwan (1996), Chiao Tung University, Taiwan (1996), Chinese University, Hong Kong (1997), University of Michigan (1998), State University of New York, Stony Brook (1999), Washington College (1999), Baptist University, Hong Kong (1999), Chung-Cheng University, Taiwan (2000), Hong Kong University of Science and Technology (2002), Ecole Normale Superieure, Paris (2003), Nanyang Technological University, Singapore (2008), Taiwan University (2015), University of Macau (2015)

Published Books and Collections

- Selected Papers 1945–1980 With Commentary, Freeman and Company, 1983
- 《讀書教學四十年》, 香港三聯書店 (*Forty Years of Reading and Teaching*, SDX Joint Pub. Hong Kong), 1985
- 《讀書教學再十年》, 台北時報出版公司 (*Ten More Years of Reading and Teaching*, Taipei Times Pub) , 1995
- 《楊振寧文集》, 上、下冊, 上海華東師範大學出版社 (*Collected Works, C.N.Yang, Vol. I & II*, East China Normal U. Press), 1998
- 《曙光集》, (簡體字版), 北京三聯書店; (繁體字版), 八方文化創作室 (*The Twilight Collection*, SDX Joint Pub. Beijing), 2008
- 《曙光集十年增版》, 北京三聯書店 (*The Twilight Collection, Extended*, SDX Joint Pub. Beijing), 2018
- *Selected Papers II With Commentaries*, World Scientific Publishing Co, 2013
- 《六十八年心路》, 北京三聯書店 (*Six-eight Years of my heart's journey*, SDX Joint Pub. Beijing), 2014
- 《楊振寧選集》, 華中科技大學出版社 (*Selected Works of C.N.Yang*, Central China Normal U. Press), 2016
- 《晨曦集》, 商務印書館; 八方文化創作室 (*The Dawn Collection*, Commercial Press, Hong Kong), 2018
- C. N. Yang, M. L. Ge and Y. H. He, *Topology and Physics,* (楊振寧、葛墨林、何楊輝，《拓撲和物理》) World Scientific, 2019
- *Selected Papers, 1945–1980, with Commentary*, 商印 Reprint 2020
- 《晨曦集》增版, 商務印書館 (The Dawn Collection, Extended), 2021



Private Collection



Private Collection



C. N. Yang speaking at the 2019 Xplorer Prize Ceremony (Private Collection)

The Friendship Between Professors S. S. Chern and C. N. Yang over the Years



C. N. Yang and S. S. Chern at dinner ca. 1990 (Private Collection)



C. N. Yang and S. S. Chern in Nankai University, ca. 1990 (Private Collection)



C. N. Yang and S. S. Chern in Chern's house, Nankai University, 1999 (Private Collection)



C. N. Yang and S. S. Chern in Nankai University, 2002 (Private Collection)



C. N. Yang's 70th birthday celebration at Nankai, 1992 (front row from left: Mrs. Chern, S. S. Chern, Ta-You Wu, T.-R. Wu, Mrs. S. Chen-Wu; back row from left: M.-G. Guang, C. N. Yang); (Private Collection)



C. N. Yang et al. at S. S. Chern's grave, Nankai University, 2005 (Private Collection)

Professor C. N. Yang and Various Friends



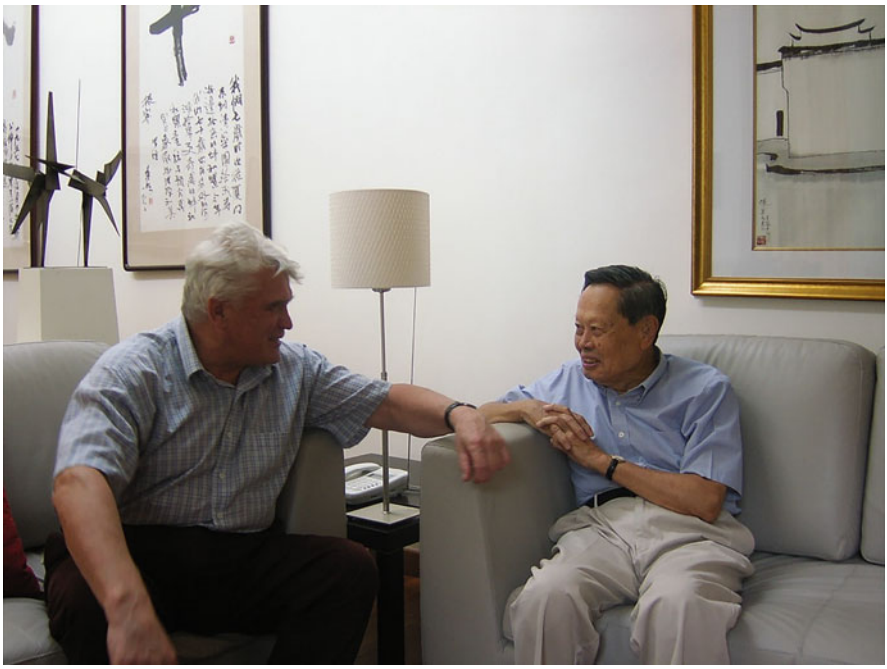
Sir & Lady Michael Atiyah at the Great Wall, circa 2000 (Photo by Mo-Lin Ge)



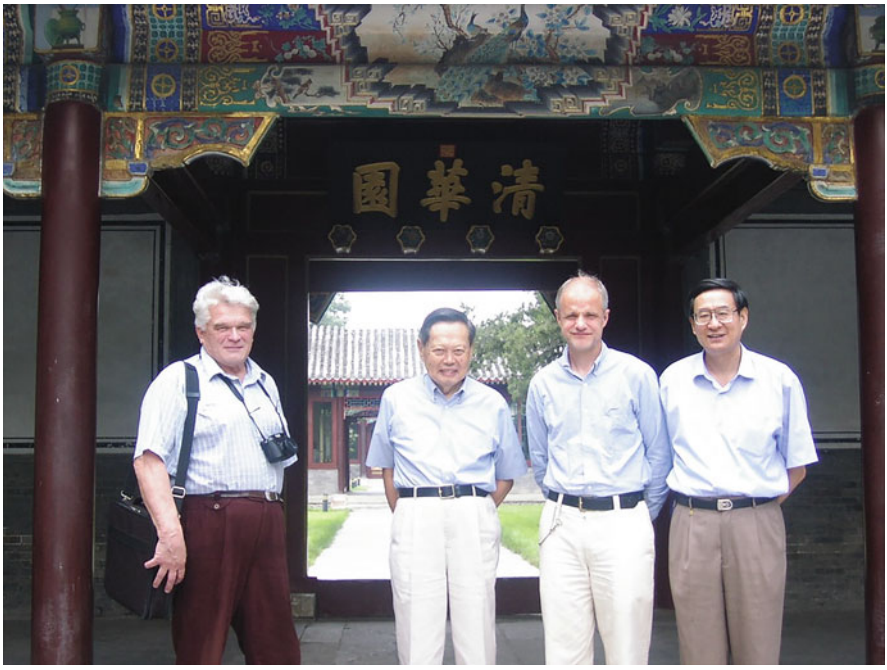
Sir M. Atiyah with M.-L. Ge at Nankai, circa 2000 (Private Collection)



Sir M. Atiyah with Y.-H. He in China, 2009 (Private Collection)



C. N. Yang and Ludwig Faddeev, 2008 (Photo by Mo-Lin Ge)



L. Faddeev, C. N. Yang, A. Niemi, M.-L. Ge, 2008, Tsinghua University Campus (Private Collection)



Freeman Dyson and M.-L. Ge before the portrait of S.S. Chern in the Chern Museum, Nankai, circa 2000 (Photo by Mo-Lin Ge)



C.N. Yang's 95th birthday celebration at the Chern Institute of Mathematics, Nankai, circa 2017 (Private Collection)

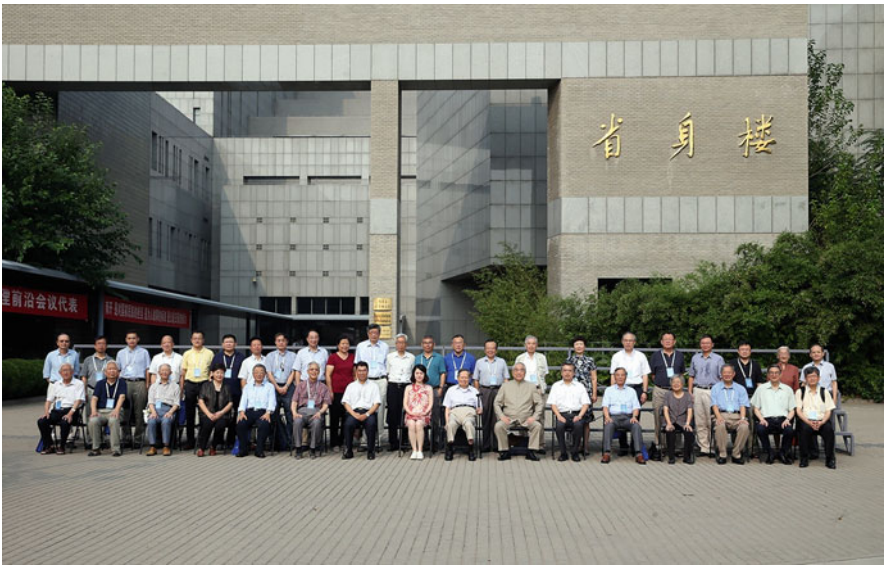




At a conference in Nankai University, 1987, with signature (Private Collection)



Ceremony of C. N. Yang receiving Honorary Professorship from Nankai University, 1986 (Private Collection)



Frontier of Physics Conference at the Chern Institute, Nankai University, with various Academicians of the Chinese Academy of Science, 2012, celebrating C.N. Yang's 90th birthday; to the left (not pictured) is the grave of S. S. Chern (Private Collection)



C. N. Yang and M.-L. Ge climbing the Great Wall, ca. 1990 (Private Collection)



M.-L. Ge and C. N. Yang, at the Chern-Yang Mural, ca. 2015, Chern Institute foyer, in front of the mural of Chern and Yang in Dialogue (Private Collection)

Acknowledgements

Immense gratitude is due to Jiakang Bao, Elli Heyes, Ed Hirst, and Steven Yan Xiao for their invaluable dedication in the type-setting, careful reading, and meticulous editorial assistance. We are also indebted to Dr. Rémi Lodh of Springer Nature for his help and advice, as well as all the anonymous referees for their useful comments and suggestions, throughout this unique project.

Contents

1 Frank Yang at Stony Brook and the Beginning of Supergravity 1
Peter van Nieuwenhuizen
1.1 Prologue 1
1.2 Frank Before Coming to Stony Brook 2
1.3 Early Years of Frank at Stony Brook 4
1.4 Supersymmetry and Quantum Gravity Before 1976 8
1.5 Some Recollections of the Path to Supergravity at Stony Brook 9
1.6 Supergravity Lives! 12
1.7 Epilogue 12
1.8 Later Years of Frank at Stony Brook 13
References 17

2 A Stacky Approach to Crystals 19
Vladimir Drinfeld
2.1 Introduction 19
2.1.1 A Theorem of Bhatt–Morrow–Scholze 19
2.1.2 A Generalization 20
2.1.3 Isocrystals 21
2.2 Crystals and Crystalline Cohomology 22
2.2.1 A Class of Schemes 22
2.2.2 Some Simplicial Formal Schemes 23
2.2.3 Notation and Terminology Related to Quasi-Coherent Sheaves 24
2.2.4 Formulation of the Results 26
2.2.5 Proof of Theorem 2.1(i) 27
2.2.6 Proof of Theorem 2.1(ii) 29
2.2.7 Proof of Theorem 2.1(iii) 30
2.2.8 $H^0_{\mathrm{cris}}(X, \mathrm{O})$ and the Ring of Constants 32
2.3 Isocrystals 34
2.3.1 A Class of Schemes 34

2.3.2 Coherent Crystals and Isocrystals 34
2.3.3 Local Projectivity 35
2.3.4 Proof of Proposition 2.4 36
2.3.5 Isocrystals as Vector Bundles 37
2.3.6 Banachian Games 39
2.3.7 Proof of Propositions 2.5 and 2.6 43
References 46

3 The Potts Model, the Jones Polynomial and Link Homology 49
Louis H. Kauffman
3.1 Introduction 49
3.2 Bracket Polynomial and Jones Polynomial 55
3.3 Khovanov Homology and the Cube Category 61
3.4 The Dichromatic Polynomial and the Potts Model 64
3.5 Khovanov Homology 73
3.6 Homology and the Potts Model 77
3.7 The Potts Model and Stosic's Categorification of the Dichromatic Polynomial 84
3.8 Imaginary Temperature, Real Time and Quantum Statistics 88
References 91

4 The Penrose–Onsager–Yang Approach to Superconductivity and Superfluidity 93
Anthony Leggett
4.1 Quantum Condensation: The Onsager-Penrose-Yang Approach 94
4.2 Some Considerations and Questions Raised by the Content of Sect. 4.1 97
4.3 What Is Special About Quantum Condensates? 100
4.4 Why Is Nature So Fond of "Simple" Quantum Condensation? Why Is "Fragmentation" So Rare? 103
4.5 When Does Fragmentation Occur? 106
4.6 Alternative Approaches to Quantum Condensation: Some Problems 107
4.6.1 ODLRO 107
4.6.2 Anomalous Averages 107
4.6.3 "Spontaneously Broken U(1) Symmetry" 108
References 112

5 Quantum Operads 113
Noémie Combe, Yuri I. Manin, and Matilde Marcolli
5.1 Introduction and Brief Survey 114
5.2 Quantum Structures in Symmetric Monoidal Categories 114
5.2.1 Monoidal (=Tensor) Categories $\boldsymbol{V}$ ([13], Sec. 2.2, 2.3) 114

5.2.2 Symmetric Monoidal Categories 115
5.2.3 Magmas, Comagmas, Bimagmas, Associativity and Commutativity for (co, bi)magmas in Symmetric Monoidal Categories ([13], Sec. 2.4) 115
5.2.4 Monoids, Comonoids, Bimonoids, and Hopf Algebras in Symmetric MonoIdal Categories ([13], Def. 2.7) 115
5.2.5 Quantum Quasigroups ([13], Sec. 3.1) 115
5.2.6 Quantum Loops 116
5.2.7 Functoriality ([13], Prop. 3.4) 116
5.2.8 Magmas etc. in the Categories of Sets with Direct Product 116
5.3 Monoidal Categories of Operads 116
5.3.1 Graphs and Their Categories 116
5.3.2 Operads and Categories of Operads (See [2], Sec. 1.6, p. 262) 117
5.3.3 Operads and Collections as Symmetric Monoidal Categories 117
5.3.4 Operads as Monoids 118
5.3.5 Comonoids in Operadic Setup 119
5.3.6 The Magmatic Operad (See [3]) 119
5.3.7 Quasigroup Monomials and Planar Trees 120
5.4 Moufang Loops and Operads 120
5.4.1 Moufang Monomials and Their Encoding by Labeled Graphs 120
5.4.2 Passage to Moufang Operad: Basic Identity 121
5.4.3 Moufang Collections (See [2], Sec. 1.5, pp. 259–261) 121
5.4.4 Latin Square Designs and Their Encoding by Graphs 121
5.4.5 From Loops to Latin Square Designs 122
5.5 Operadic Structures on Quantum States 122
5.5.1 Operads of Classical and Quantum Probabilities 122
5.5.2 Classical Probabilities from Quantum States 124
5.5.3 Non-unital Operads 125
5.5.4 The Q_P-Operad of Quantum States 126
5.5.5 The Q_Λ-Operad of Quantum States 129
5.5.6 Trees of Projective Quantum Measurements 130
5.5.7 Entropy Functionals 132
5.5.8 A_∞-Operad of Quantum Channels 133
5.6 Operads and Almost-Symplectic Quantum Codes 136
5.6.1 Rational and Binary Little Square Operads 136
5.6.2 Binary Little Square Operads and Almost Symplectic Spaces 139
5.6.3 Colored p-ary Little Squares 140

5.6.4 Operad Partial-Action on Quantum Codes 143
References 145

6 Quantum Computational Complexity with Photons and Linear Optics 147
Jian-Wei Pan
6.1 Introduction 147
6.2 The Mathematics: Permanent and Hafnian 149
6.2.1 Permanent 149
6.2.2 Hafnian 149
6.3 The Model: Boson Sampling 150
6.4 Single-Photon Boson Sampling Experiments 154
6.5 Quantum Computational Advantage with *Jiuzhang* 155
6.6 Applications 161
References 162

7 Quantized Twistors, G_2^*, and the Split Octonions 165
Roger Penrose
7.1 A Key Motivation for the Formulation of Twistor Theory 165
7.2 The 2-Spinor Formalism 167
7.3 Projective Twistor Space 170
7.4 Twistor Kinematics 174
7.5 Quantized Twistor Theory and Masless Fields 178
7.6 Split Octonions and G_2^* 183
References 188

8 Kronecker Anomalies and Gravitational Striction 191
Alexander M. Polyakov and Fedor K. Popov
8.1 Introduction and General Discussion 191
8.2 Kronecker Anomaly in Thermal Harmonic Oscillator 196
References 214

9 Projecting Local and Global Symmetries to the Planck Scale 215
Gerard 't Hooft
9.1 Introduction 215
9.2 Quantum Mechanics 217
9.3 Classical Models Underlying Quantum Mechanics 220
9.4 The Standard Model 223
References 226

10 Gauge Symmetry in Shape Dynamics 227
Frank Wilczek
10.1 Gauge Structure: Fundamental, Emergent, Productive 227
10.2 Dynamical Equation for Deformable Bodies 228
10.2.1 Referenced Angular Momentum 228
10.2.2 Inertia Tensor and Angular Motion 229
10.2.3 Gauge Symmetry and Gauge Field 231

10.2.4 Dynamical Equation 232
10.2.5 Three Dimensional Notation 233
10.2.6 Specializations 234
10.2.7 Angular Momentum and Energy 235
10.3 Extensions 236
10.3.1 Blobs, Media, and Swarms 236
10.3.2 Molecules and Nuclei 236
References 239

11 Why Does Quantum Field Theory in Curved Spacetime Make Sense? And What Happens to the Algebra of Observables in the Thermodynamic Limit? 241
Edward Witten
11.1 Introduction 241
11.2 Quantum Field Theory in Curved Spacetime 244
11.2.1 The Problem 244
11.2.2 Practicing with a Spin System 247
11.2.3 A System of Harmonic Oscillators 250
11.2.4 Back to Field Theory 253
11.2.5 What Is Quantum Field Theory in an Open Universe? 255
11.2.6 Non-Free Theories 259
11.3 Quantum Statistical Mechanics and the Thermodynamic Limit 260
11.3.1 The Thermofield Double 260
11.3.2 Surprises in the Thermodynamic Limit 264
11.3.3 Examples from Spin Systems 266
11.3.4 Relation to Quantum Field Theory 270
11.3.5 The Hagedorn Temperature 273
11.3.6 Density Matrices and Entropy 276
11.4 The Large N Limit and the Thermofield Double 280
References 282

12 Quantum Anomalous Hall Effect 285
Qikun Xue

13 Magic Superconducting States in Cuprates 291
Jiangping Hu, Xiaoli Dong, and Zhongxian Zhao
References 298

Chapter 1
Frank Yang at Stony Brook and the Beginning of Supergravity



Peter van Nieuwenhuizen

Abstract Frank Yang came to Stony Brook in 1966 as director of a newly created Institute for Theoretical Physics. He stayed for 32 years. (It was aptly renamed the C. N. Yang Institute upon his retirement from Stony Brook in 1998). The author was his deputy director for a decade, and succeeded him as director, to be succeeded in turn by George Sterman who is still at present the director. In this contribution in honor of "the sage of Stony Brook" (Dyson's words), I will tell some reminiscences and anecdotes of how life was under Frank's directorship. When the Yang Institute celebrated its 50th anniversary, Frank recalled in a brief video from China that for him the two most memorable events during his time at Stony Brook were the discovery of Supergravity and the dictionary between physics and mathematics which he established in discussions with Jim Simons. The latter can be found on the net [1]. I shall recall the beginning of Supergravity at Stony Brook, as a tribute to the discoverer of nonabelian gauge theories.

1.1 Prologue

At a meeting in North Carolina in 1976, a panel discussed the directions into which physics might evolve, and the chairman asked Frank Yang about his opinion about supersymmetry. Frank stood up and said approximately the following: *In my opinion any breakthrough in physics should be accompanied by a breakthrough in mathematics. I asked Chern [the mathematician from Berkeley who had studied with Frank's father] what he thought of anticommuting numbers. He said that mathematically there was not much to it.* And then Frank sat down...

Dan Freedman was in the audience. He was working with me and Sergio Ferrara on Supergravity, a new nonabelian gauge theory which combines General Relativity with the Fermi-Bose symmetry called supersymmetry. The superspace version of that theory uses anticommuting coordinates. There we were! What to do? Frank

P. van Nieuwenhuizen (✉)
C. N. Yang Institute for Theoretical Physics, Stony Brook University, Stony Brook, NY, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_1

never told anyone in his Institute what he or she should do, and yet, here he stated in his usual terse and pertinent comments what he thought of the work on supersymmetry (and string theory which became a hot subject a few years later). Since we worked on a version of the theory in ordinary space, we continued our research.

Some years later I was invited to give a talk on Supergravity at a celebration of the Institute for Advanced Studies (IAS) at Princeton. It was a daunting task with Nobel laureates in the audience. I explained that Supergravity can be quantized with the same methods that were developed by 't Hooft and Veltman for Yang–Mills theory. And I showed that the one loop corrections in the Maxwell–Einstein theory were divergent, but in the supersymmetric Maxwell–Einstein theory ($N = 2$ supergravity) the divergences miraculously canceled. Penrose was in the audience, and at the end of my talk he criticised my approach. He observed that quantum corrections change the gravitational field, and thus also the structure of the light-cones, whereas in my one-loop perturbation approach I had not taken this change into account. (This criticism stands even today as up to 7-loop corrections in $N = 8$ supergravity have been shown to be finite.) I stumbled with my response to this criticism, but Frank Yang came to the rescue. I do not remember his exact words but I believe he said that the mere fact that divergences canceled due to a new symmetry principle was a great result. That was Frank also: he supported the members of his Institute unconditionally, and encouraged and helped us whenever he could.

1.2 Frank Before Coming to Stony Brook

Frank came in 1945 to the US to study with Fermi.[1] He considered Fermi the ideal physicist: with both feet on the ground, he was both a theorist and an experimentalist. Frank proposed to write a thesis in experimental physics with Fermi, but that was not possible because Fermi was working at the Argonne National Laboratory which did not allow participation by foreign nationals. Instead, Frank joined a group building an accelerator, but that did not go well (I have heard that some physicists at Chicago quipped: where is Yang there is a Bang), and at the advice of Edward Teller he moved on to theoretical physics. He wrote in 1948 a dissertation on a topic of group theory[2] in physics *On the Angular Distribution in Nuclear Reactions and Coincidence Measurements*, and also pursued his interests in statistical mechanics. Of that time dates his 1949 article with Fermi *Are Mesons Elementary Particles?* [2].

[1] Frank first went to Columbia, where however he found out that Fermi had moved to Chicago, so Frank moved on to Chicago as well.

[2] Decades later I wrote an article on molecular vibrations of Platonic polytopes in 4 space instead of 3 space dimensions. I asked Frank a question about it, and got back a short complete answer using characters instead of the Young tableaux I was more familiar with.

In 1949 Frank became a postdoc at the Institute for Advanced Studies (IAS) in Princeton. Fermi advised him not to stay there too long because physics at the IAS was too abstract, and Chicago offered him a position, to start in 1950. However, Frank's girlfriend Chih-Li Tu was studying in New York so Frank decided to stay in Princeton (they were married in 1950 and had three children; she was a daughter of General Yu-Ming Tu).

He stayed in Princeton for 17 years. He wrote a series of physics-shattering papers of which I only quote the following: on the spin of the π^0 [3], on Majorana fields in weak interactions [4], on nonabelian gauge theory ("Yang–Mills theory") [5], on parity violation [6]. He wrote well over 100 other articles while at Princeton, most of which dealt with particle phenomenology and topics in statistical mechanics.

In the spring of 1965, Oppenheimer told Frank that he was preparing to retire from his post as Director of the IAS and wanted to recommend Frank to the board of Trustees to be his successor. Frank replied that he would not like to be the director of the Institute. Oppenheimer asked Yang to think it over and then to decide. Yang did, and in a letter to Oppenheimer he wrote: *It is quite uncertain that I shall make a good Director, while it is quite certain that I shall not enjoy the life of a Director.*

In 1964 and 1965, the State of New York created five Einstein Professorships for the universities of the State of New York. Stony Brook designate-president John Toll (a Princeton alumnus), with help from physics chairman Alex Pond (another Princeton graduate) and theoretical physicist Max Dresden (later the Executive Director at the ITP) and others vigorously pursued the possibility of attracting Frank as Einstein professor to a newly created Institute for Theoretical Physics at Stony Brook. Of course there was vigorous competition. For example, A. S. Wightman from Princeton University ended a recommendation letter for Frank as follows: *I cannot end this letter without an additional remark. The possibility of the departure of Yang from Princeton is viewed by all physicists here as a disaster to the Princeton community, and we will fight like tigers to try to keep him here.* There came the painful necessity of asking for letters of reference, painful for the letter writers because as Oppenheimer wrote in his letter of recommendation: ... *there are at most one or two men in the world who might rival him, and none who is better* ..., so writing a letter of recommendation was like carrying coals to Newcastle. We quote from some of the letters.

... Yang's contributions to the post-war development of physics theory are so well-known and so highly appreciated that it is pointless for one to review them. You are capturing one of the giants of 20th century physics. Every university in the world is envious ... [Geoffrey F. Chew from Berkeley]

... His most celebrated work is that with Lee on the failure of conservation of parity in weak interactions. His reasoning was superb. From a number of discrepancies in the behaviour of K-mesons which baffled everybody in the physics fraternity, Yang and Lee deduced that violation of parity was the only possible way out, although it completely contradicted previously held sacred tenets of physics. They then proceeded to develop a sound theory of the neutrino, and predicted the violation of parity also in β decay, a phenomenon which everybody had thought to

be completely explored and understood. Their prediction was immediately proved experimentally. A few years later it led to a much more profound theory of β *decay* ... [Hans A. Bethe from Cornell]

... It seems ridiculous for me to be writing a letter of recommendation for Frank Yang; anybody who has heard of me will certainly have heard of Yang. ... He is not only a productive worker himself, but he also takes interest in the work of younger Institute members and has helped many of them with suggestions and criticisms. He is not a man who opens his heart to strangers, but he is easily approachable and always ready to listen. ... [Freeman Dyson from the IAS]

... It was quite depressing to be a student with Frank. He did everything with such consummate ease that we all felt rather inadequate ... [M. L. Goldberger from Princeton]

... [He is] efficient in the highest sense. The questions he chooses to answer, turn out to be always key questions. The techniques he chooses or creates to deal with them are rarely more elaborate than the significance of the questions merits. This characteristic of his work is reminiscent of that Enrico Fermi ... [A. S. Wightman from Princeton]

1.3 Early Years of Frank at Stony Brook

In the spring of 1965 Frank visited Stony Brook a few times, everything went well: he liked the place and bought a house on the Sound, and started as Director of the ITP in the fall of 1966. Starting an Institute from scratch, facing the competition of celebrated Ivy League Universities and other famous Institutions, would have daunted anybody, but Frank, as I witnessed later at many occasions, was an eternal optimist, who did not lose time on speculation that he could fail. Great institutions need a central person around whom the institution grows, and Frank was that in every respect. The first bunch of four postdocs arrived (see Fig. 1.1). They worked in quantum field theory and phenomenology. Life at the Institute was mercifully free from long faculty meetings: all time was devoted to physics, and faculty meetings, if any, were brown-bag luncheon meetings which lasted less than half-an-hour. Brookhaven was nearby, and seminars were a combined Brookhaven-Stony Brook affair. Frank would arrive at seminars with a pad and a pencil, but he never seemed to write anything on his pad. His mere presence lifted the intellectual level of debate (Fig. 1.2).

Since the Institute was a new institution, there were few rules and much opportunity for initiatives. Frank, if needed, helped in all possible ways. Soon after I arrived, as an assistant professor, I introduced the "Friday Seminars", seminars for and by graduate students with some faculty present,[3] and after these seminars the participants would play a game of soccer (see Fig. 1.3). There was much social

[3] They continue up to today, and are now as active as then.

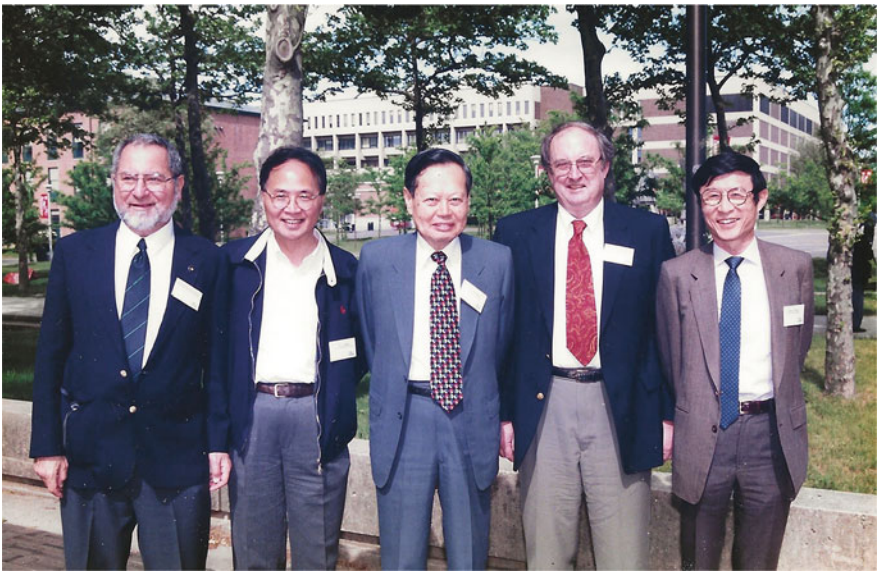


Fig. 1.1 The first four postdocs with Frank Yang. They started at Stony Brook in 1966. From left to right: Michael Nieto, Hwa-Tung Nieh, Frank Yang, Bill Bardeen, and Wu-Ki Tung. Photo taken at Frank's retirement from Stony Brook

contact between students and professors because most of the faculty were young. We had every second weekend a party at the home of a student or a professor. We also had "Saint Nicolas parties" (a Dutch custom) where participants wrote anonymous poems with jokes about each other, that were read in public. Some of these poems adorn today a wall of our common room. An example is the following poem about Frank and Rubik's cube.

He sat before the challenge,
His gaze upon it fixed
At stake, his fame and honor
For Maths-magic tricksHungary"
The Century's Greatest Dare

Without the slightest twist or turn,
Without a yawn or Flinch
For six and twenty hours
He didn't budge an inch

Then with a loud "Eureka",
He jumped up from his place...
A simple transformation,
Some non abelian space

He sat back to complete the task
With feverish turn & twist...
And twist and turn
And slide and slip,
He rubbed his aching wrist

And after near eternity,
The last piece found its groove
C.C.Yang had conquered
Rubik's cube
In 80,000 moves!!!

December 5, 1981



Fig. 1.2 Part of the 1967 Stony Brook Physics Department photo. From left to right: John Toll, Frank Yang, Paul Dirac (who was visiting Stony Brook that year), and Max Dresden

Physics in those years was, as so often, evolving in many directions whose future was unclear. Particle physicists had observed that

- strong interactions are short ranged (implying the absence of massless hadrons)
- they have several (exact or approximate) symmetries and corresponding conservation laws
- many hadrons had been found, most of them unstable (resonances), and with large mass and angular momentum (spin): the "hadron zoo" seemed to be increasing every day.



Fig. 1.3 To promote contact between professors and students, soccer games were held after the "Friday seminars". Left upper corner Max Dresden, kneeling on the right Dan Freedman, in the middle the author, on his left Steven Libby

It seemed clear at that time (but erroneous with hindsight) to rule out a description of these phenomena based on quantum field theory (QFT) because

- there were too many particles, unlike QED where there were only the photon and the electron
- it seemed that particles with spin $J > 2$ cannot be described by a renormalizable QFT (correct)
- the coupling constant of strong interactions was large, ruling out perturbation theory.

Then in 1968 arrived a bombshell by Veneziano, who wrote down a formula for the process $\pi + \pi \rightarrow \pi + \omega$ which satisfied all the theoretical requirements that had been extracted from rudimentary field theory and experimental evidence. This triggered a frenzy of activity of "dual models". Theoretical particle physics in the 1960s on dual models consisted of studies about free field theory, analyticity, Schwinger terms in current algebras, sum rules, etc. Regge theory was in; it explained successfully the dip in $\pi^- p \rightarrow \pi^0 n$, but later cuts were found necessary, and even later the Pomeron was added to account for background effects in scattering experiments. Conferences in those days were devoted to endless presentations of newly discovered resonances with their Argand diagrams, and I hated every minute of it.

Although most physicists had given up hope that quantum field theory could describe the strong interactions, a few physicists continued studying quantum Yang–Mills theory along the lines which has been so successful in 1947 for QED. (One of them was my advisor, Tini Veltman.) Frank had written a paper on the ζ limiting procedure, a field theory which violated causality at the microscopic level (but whose propagators turned out to coincide with the much later constructed propagators of spontaneously broken gauge theories).

Current algebra was used to describe the huge amount of experimental data. John Bell suggested that the currents of current algebra should couple to the fields of the 1954 Yang–Mills theory, and a few physicists started studying nonabelian gauge theories in a theoretical well-defined way, thus bringing the work of Yang and Mills to the foreground. In 1972 another bombshell occurred: 't Hooft showed that nonabelian gauge theories were renormalizable in the same way as QED, and with Veltman he developed a method of regulating the divergences in loop calculations (dimensional regularization). This opened the doors towards the Standard Model.

1.4 Supersymmetry and Quantum Gravity Before 1976

Several physicists tried in the fifty years between 1926 and 1976 to construct a theory of quantum gravity, but the method of renormalization, so successful for the non-gravitational interactions, does not work for the gravitational interactions because the gravitational coupling constant, the Newton constant G, is dimensional, and this leads in general to infinities. Thus in quantum gravity one can only hope that infinities cancel "miraculously", and if that does not happen, the theory is doomed. The word "miraculously" has a meaning for insiders, it means that further hidden symmetries are present in addition to those of Einstein's theory of gravity (General Relativity) which cancel the otherwise fatal infinities.

In the early 1970s, a series of discoveries were made which put the problem of quantum gravity in a new perspective. First of all, as already mentioned, 't Hooft and Veltman (1971) were able to construct an extension of the old method of renormalization which was consistent (meaning: it got rid of infinities) when applied to Yang–Mills theory (the theory of gauge fields that appears in the Standard Model).

The other discovery in the early 1970s was Supersymmetry. Two Soviet groups stumbled on a very interesting new concept, Bose-Fermi symmetry, a new symmetry between the two kinds of particles in our universe: bosons (such as photons, and the other gauge particles of the Standard Model) and fermions (electrons, neutrinos, quarks). Golfand and Likhtman (1970) wrote down what would be called today an interacting theory of an $N = 1$ massive multiplet containing a scalar, two spinors and a vector particle, all of the same mass. They made an incorrect statement about parity, but in the process they discovered an extension of the mathematics used to describe symmetries. (This branch of mathematics is called Group Theory and was developed in the nineteenth century. They were the first to write down what one

now calls a Superalgebra, with commutators (standard) and anticommutators (very non-standard). They were in a certain sense the successors of Jordan, who in 1928 had introduced anticommutators for the quantization of the electron.) The other group of Soviet physicists, Volkov and Akulov, proposed in 1973 an explanation of why neutrinos are massless (as they were supposed to be at that time): just as the theory for the Higgs boson needs and contains extra massless fields (would-be Goldstone bosons) which are eaten by gauge fields, which then become massive as a result of this banquet, they proposed that the neutrinos were the fermionic equivalents of these massless fields (would-be Goldstone fermions). Closer study revealed that the neutrinos of our world do not have the properties of would-be Goldstone fermions, but the idea of a Bose-Fermi symmetry (Supersymmetry) had been launched. Then Volkov and Soroka (1973) constructed a "nonlinear realization of a Supersymmetric theory with a super Higgs effect": spin 3/2 fields eat spin 1/2 fields and become massive. These Soviet papers were unknown in Europe and the US, and were only rediscovered after the very influential work of Wess and Zumino. Julius Wess and Bruno Zumino wrote in 1973 a series of papers (after Gervais and Sakita had constructed in 1971 a two-dimensional relativistic quantum field theory for strings with Supersymmetry) in which they were able to incorporate Supersymmetry into four-dimensional relativistic quantum field theory. Thus by the mid-1970s Quantum Field Theory had returned as the most successful description of the non-gravitational forces, and Supersymmetry had made its appearance. The time was ripe for the next step: combining gravity with Supersymmetry in the framework of relativistic quantum field theory.

1.5 Some Recollections of the Path to Supergravity at Stony Brook

In 1973 I came to the US to study as a postdoctoral fellow with Stanley Deser at Brandeis University (near Boston). 't Hooft and Veltman had just published an article in which they applied their quantization methods (which were developed for the Quantum Standard Model) to Einstein's theory of gravity, and to the coupling of this theory to a scalar field (such as the Higgs boson). They wrote: *A certain exhaustion however prevents us from further investigation, for the time being.*, a clear hint for a young physicist to throw his hat in the ring. Deser and I studied the quantization and renormalization (or, rather, finiteness) of many gravitational theories: GR coupled to Dirac's electron, to the gauge fields of Yang and Mills, and with Grisaru and others I extended these calculations to other cases. But try as we would, the infinities never canceled, and thus all these beautiful models were doomed by quantum mechanics. I still hoped to find the Holy Grail: a model ("the" model?) in which the infinities cancel. I tried an old theory of Hermann Weyl of the 1910s ("local scale invariance", now a key ingredient of string theory) but I did not complete this work (and would probably have failed; for insiders: because one needs

ghosts for ghosts). At that time (1974) I went to a summer workshop in Paris at the Ecole Normale (where I had been a Joliot-Curie fellow for a couple of years and to which I have kept my ties). There I met Daniel Z. Freedman (Dan), a theoretical physicist from Stony Brook, and we hit it off straight away. (Dan later told me that he had heard me the year before give a seminar on my work on quantum gravity at Princeton and had found it very interesting). After the workshop, Dan went back to Stony Brook, and I received an offer to come there as an Assistant Professor.

So in the fall of 1975 I came to Stony Brook to work on Quantum Gravity. The past year Dan had studied the articles on Supersymmetry for non-gravitational models, so what was more natural than to collaborate on Supersymmetry in Gravity? Wess and Zumino were working on it at CERN, and Zumino gave a lecture on it at a conference in Boston in the early fall of 1975. Arnowitt and Nath at Northeastern had published a theory which was a very direct extension of GR but which turned out (later) to be inconsistent. And Volkov and Soroka had "gauged" the Superalgebra mentioned above, to obtain the super-Higgs effect. So it was clear that there was competition. Dan went to Paris for a few months that fall on a sabbatical (again at the Ecole Normale) where he met Sergio Ferrara from CERN. Sergio had come to Paris to work on another project, and in a discussion with Dan he made an important suggestion about coupling spin 3/2 spin fields directly to gravity. Such was the situation when Dan returned at the end of the fall. We started working very hard, Dan contributing his insights about supersymmetry (about which I did not know much) and I contributing my experience in Quantum Gravity, in particular my expertise in how to describe fermions in gravitational fields. We exchanged frequent airmail letters with Sergio about our work (there was of course no internet and transatlantic phone calls were expensive and thus discouraged). The next summer I finally met him in Paris, and he became for many years my collaborator and friend. As mentioned above, for a supersymmetric theory of gravity one needs a new kind of fermion, with spin 3/2 instead of spin 1/2. Massive spin 3/2 fields had been studied by Rarita and Schwinger in 1941. (In a companion article, Kusaka used their theory to find out whether neutrinos have spin 3/2 or spin 1/2. The results clearly showed that neutrinos do not have spin 3/2. We now know that they have spin 1/2.) We worked very hard, everyday till late, all weekends, no free time. Sometimes things moved forward and at other times we hit roadblocks. Dan had a sign on his office door with "Supergravity lives" if things went OK, and "Supergravity is dead" when we got stuck. There was much support of the members of our Institute, in particular Frank helped whenever he could. But support was not enough, we needed a way to handle the enormously complicated equations we discovered. In the next paragraph, I shall now describe what is meant by "handle the equations". This part of the story is of necessity more technical than the rest, but it plays an essential role.

The issue we faced was to show that all supersymmetry variations of the "action" (a well-defined mathematical object) vanished. The linear variations (with one spin 3/2 field) vanished by a "heart-warming" calculation. Next came the cubic variations (with three spin 3/2 fields). They did not vanish, but we modified our starting point by adding quadratic terms to the spin 3/2 transformation law (later identified as "torsion", see below), and quartic terms to the action. After much work we were

able to show that the cubic variations did cancel, but new terms of fifth order in spin 3/2 fields turned up in the variation of the modified action. We showed (essentially by dimensional analysis) that further modifications of the starting point (quartic terms in the spin 3/2 transformation law, and sixth-order terms in the action) could not help in canceling the remaining fifth order variation, thus, we wrote, ... *it is critical to verify that it vanishes.* Dan, Sergio and I decided to publish our results, and submitted our article *Progress towards a Theory of Supergravity* to the Physical Review (the premier American physics journal) on 29 March 1976.

The issue was now of course whether the quintic variations did vanish. At that point we decided to use the computer. In Utrecht (where I got my PhD), I had learned to handle programs for formulas ("symbolic manipulation programs"), and had used them in Paris and Brandeis to investigate quantum gravity. Using a simple program (for insiders: a Fortran program, not even a symbolic manipulation program), I started applying it to the equations. In order that the theory of Supergravity (the name was obvious: Supersymmetric Gravity = SuperGravity) exist, there were about 2000 terms whose coefficients had to cancel. According to the theory of Supersymmetry, these coefficients were all integer numbers, so the computer program did not have to be accurate in decimals. There were computer bugs, and we spent days on end to simplify the computer program to bring the cost down. (The total bill was in the tens of dollars, a ridiculously small amount of money compared with the millions of dollars now spent on computers in physics). After a lot of work we came to a point where there were no further ambiguities: either the 2000 coefficients all vanished, or "Supergravity was definitely dead". One night (I worked every night) I was sitting in the small computer room in our Institute. The results came in on a printer that was connected by phone to the big (big for those days) CDC computer at Brookhaven. I realized it was now up or down. In the room was also a colleague who had recently been told he would not get tenure, and who was for obvious reasons depressed. In those gloomy surroundings I waited for the results. If any of the 2000 coefficients were nonzero, it would mean that half a year of frenetic work was in vain. And then the zeros started coming, in batches of 20. The first dozens of them were indeed all zero, but that we knew already, the crucial point was the last series of batches. More zeros, rising tension. Zeros near the end. I held my breath. And then it was all over: all 2000 numbers were zero: "Supergravity lives". I phoned Dan who was in a hotel for a conference in Chicago (it was by now the middle of the night) and told him the news. I heard a sleepy voice say, *That is great*, followed by a yawn. I went home, and I felt depressed. That may sound strange to outsiders, but I have noted that other physicists also sometimes feel depressed when after long and arduous work they finally succeed. It is in my opinion the scientific equivalent of post-partum depression. However, in the days after, the clouds lifted, and we knew that we had constructed a fundamental new theory of physics, which combined Einstein's General Relativity with a symmetry between the two classes of particles in our universe. We sent a "Note added" to our article with the message that all variations now canceled.

1.6 Supergravity Lives!

Up to this point I have described my day-to-day experiences in Stony Brook, but the world had not stood still. Our paper had been submitted on 29 March. A few weeks later, Dan received a letter from Bruno Zumino dated April 17 which he has kept in which Bruno congratulated us on our success and compared our work with work he had been doing with Wess on the same problem. And, again two weeks later, on 28 April, Deser and Zumino submitted a paper entitled *Consistent Supergravity* to the leading European journal Physics Letters B. In this article they referred to our work, but proposed a reformulation which stressed the role of "torsion". The notion of torsion has a long history in GR. Originally it was proposed by the mathematician Elie Cartan (of the Ecole Normale in Paris) in 1922, as a generalization of Einstein's theory of GR. Since that time many relativists have studied torsion further, but it never played a significant role. In Supergravity, on the other hand, it is an important option because it simplifies the formulation, eliminating the need to consider the quintic variations explicitly. In fact, setting torsion to zero supergravity becomes inconsistent. A few years later, I found with Paul Townsend who was then a postdoc at Stony Brook a third formulation of Supergravity which combines the virtues of the formulation of FNF (Freedman, van Nieuwenhuizen, and Ferrara) and that of DZ (Deser and Zumino). Their formulation with an independent torsion field is called the first-order formulation, and our formulation where torsion is due to pairs of fermionic spin 3/2 fields is called second order formulation. Dan called our third formulation the 1.5 order formalism. It is theoretically rather obvious, but for practical calculations it has turned out to be indispensable.

1.7 Epilogue

Since these stressful but exciting weeks in 1976, Supergravity has grown tremendously. It has become part of string theory, more precisely string theory becomes Supergravity at energies low enough that experiments can be done to test whether it exists and if so, its properties. On the one hand, it is not the Holy Grail because it contains (probably, the issue is not settled) infinities of quantum field theory. On the other hand, it is full of new hidden symmetries (dualities) which have led to major changes in our thinking. In 1998 it was discovered by Maldacena that Supergravity in 5 dimensions corresponds to a certain non-gravitational but supersymmetric theory in 4 dimensions (our world), and this "AdS/CFT correspondence" has been extended to many other models.

1.8 Later Years of Frank at Stony Brook

Also the ITP, "Frank's institute", grew and was accepted by the physics community on a par with other great institutions. The gamble of starting a new Institute from scratch with the world's most famous theoretical physicist had paid off. Frank worked tirelessly to attract bright new stars, but he also enthusiastically supported efforts by even young professors to recruit "hot postdocs". An anecdote illustrates his active help. In 1979 Dan and I decided to organize the (I believe) first conference on Supergravity. Of course it was to be held at Stony Brook, but there were some obstacles: the print shop could not print the poster for the workshop on time, there was no auditorium available, etc. We went to Frank, and he said, *One moment*, and phoned the president of Stony Brook, John Toll. Frank had hardly finished explaining that this workshop was important for the image of physics at Stony Brook, as John Toll told us to go ahead and mention that he urged the utmost cooperation. From then on the organization went smoothly, and the conference was a great success.

Frank stayed active in research. As George Sterman writes:

C. N. Yang's research while at Stony Brook produced at least two breakthroughs that continue to guide contemporary research. The first grew out of new methods he established for the solution of two-dimensional field theories, and the second from new insights he developed into the relationship of gauge field theories to geometry. The former was the origin of the celebrated Yang-Baxter equation, with applications far beyond two dimensions, the latter to a renewed dialog between mathematics and physics that continues to inform both disciplines in the twenty-first century. Over his Stony Brook years and beyond, his numerous influential papers, many involving the concept of magnetic monopoles, deepened and applied these insights.

In addition, over several decades, he and T. T. Chao studied aspects of hadronic scattering at high energy. Based on a geometrical model, and using a few elegant assumptions, such as "limiting fragmentation" they provided insights on fundamental aspects of hadronic scattering, such as total cross sections and their diffractive behavior in near-forward scattering that remain relevant today in the laboratory and for theory.

Many other papers reflect an abiding and growing interest and perspective on the history of physics, and the development of its seminal ideas, including the principle of gauge invariance and concepts flowing out of quantum mechanics. At the same time, his excitement for and encouragement of emerging laboratory breakthroughs in the 1990's, in high temperature superconductivity and Bose-Einstein condensation, were illuminating for all his colleagues.

Around 1990 Frank asked me to become his Deputy Director, and in this capacity we had discussions about which direction the Institute should go. He had become convinced that the future of theoretical physics lay in applied physics, rather than fundamental physics. For example, he was in favor of promoting accelerator physics, but the work at the Institute in traditional areas such as QCD, Ising models, monopoles and supergravity was so fertile that he shelved his idea. This was typical

of Frank: he would cheerfully accept ideas of others, but privately he kept his convictions. Consider the following anecdote written by Shou-Cheng Zhang who came to study with Frank at Stony Brook.

Even though he was extremely busy, Prof. Yang always made himself available to talk to the incoming graduate students. In my first meeting with him, he asked me what my interest was. I answered that the highest goal of theoretical physics is to pursue Einstein's dream and unify gravity with all other forces, and this is what I am interested in doing. To my great shock, not only did he discourage me from doing that, he even discouraged me from doing particle physics in general. He argued eloquently that physics is a broad discipline, and interesting problems can be found everywhere.

I left Prof. Yang's office in a deep crisis and confusion. Fortunately, he offered a course called "Selected Problems in Theoretical Physics" to the incoming graduate students. In this course, he did not talk at all about what I consider to be the "frontier problems" in physics, but instead he talked about problems like Bohm-Aharonov effect, duality of the Ising model, flux quantization of a superconductor, phases and holography, off-diagonal-long-range-order and the concept of gauge field and magnetic monopoles. What I learned the most in this course was the selection of the topics. These topics reflected his personal taste in physics, and that's something one cannot learn from books. In these topics I saw that the complexity of nature can be captured by simplicity and beauty of the theory. This is what theoretical physics is all about.

Frank had only a few PhD students, but his style of lecturing set an example for future faculty as this recollection of Derek Teaney shows who was first a student and later a professor at Stony Brook.

I arrived in Stony Brook in the fall of 1995 and registered for the canonical first year graduate course in Quantum Mechanics. But, it wasn't the canonical course: it was taught by Professor C. N. Yang. His course was in some ways very old school, and there was ***a lot*** *of homework every week. Indeed, Professor Yang attributed part of his success to his strong formal education in China, and he was committed to creating a similar academic culture at Stony Brook. By contrast, his lectures were not old school, but rather gave his own insights into almost every aspect of quantum mechanics. The standard topics, such as the properties of spherical harmonics, the Wigner-Weiskopff approximation for decaying states, the zeros of wave functions, and the properties of* $SU(2), \cdots$, *were presented in such a personal way, which just does not come from reading the standard books.*

He searched for research directions even in the familiar material of first year graduate quantum mechanics. During the course he rediscovered some interesting properties of the zeros of wave functions in one dimension, while lecturing on perturbation theory. He took a group of us aside and suggested it as a research direction, before he learned that the proposed methodology was already in use. In 1995 when Bose-Einstein condensates were created in atomic traps, he encouraged us all to "catch the moment".

What really came through in his lectures was his humility, and his respect for the intellectual traditions of the subject. He always lectured slowly, and with the

purpose of explaining the physics simply and correctly. He was not there to show off – he had too much respect for his "teacher", Enrico Fermi, and mentor, Abraham Pais, for that. (Stories of Fermi and Pais appeared throughout the course, bringing additional insight.) At a purely mechanical level, Professor Yang was not a perfect lecturer. But, I learned the most when he would get lost or confused on the board. In these instances he would stop, go even slower and back to basics, to finally fix the confusion. I realize now that this style is characteristic of a truly deep thinker, and it is a style we should all try to emulate.

The members of the Institute had great respect for Frank. Here is a recollection of Robert Shrock.

One of my main recollections of C. N. (Frank) Yang was his emphasis on the importance of symmetry in physics. I recall many eloquent discussions that he gave on gauge symmetry in particle physics. This was particularly fitting, in view of his seminal development of non-Abelian (Yang–Mills) theories. Yang also understood the importance of broken symmetries, as was evident from his inference with T. D. Lee in 1956, based on their analysis of experimental data, that parity was broken by weak interactions.

Another memory that I have is of Frank's great breadth of interests and research. In addition to his far-reaching contributions to particle physics, he made many very important contributions to statistical mechanics. I vividly remember how excited Frank was when Bose-Einstein condensation of dilute atomic vapors was observed in 1995. He was similarly enthusiastic about various advances in condensed matter physics, such as the discovery of carbon-60 fullerenes in 1985 and the discovery of high-temperature superconductivity in 1986.

Finally, I have a strong memory of Frank Yang's awareness that he bridged two different countries, China and the United States.

In 1998 Frank decided to retire from Stony Brook, and move to China where he continues right up to today to do research in physics. Our Institute was renamed "the C. N. Yang Institute for Theoretical Physics" in gratitude for 32 years of driving it to prominence. A symposium was organized where a huge crowd of 190 participants came, among which were many of his old friends, and many prominent physicists (see Fig. 1.4). The proceedings were published in [7]; they contain the talks at the symposium but also a complete publication list of Frank up to 1999 (Fig. 1.5).

Frank, you have been all your life in the words of Freeman Dyson a conservative revolutionary; preserving and cherishing the past, and removing old structures when needed to construct the future. But you also have made the life of countless physicists who passed through Stony Brook, a very special experience. And, although I never told you, you made the 24 years I was with you the best years of my life as a physicist.



Fig. 1.4 Participants of Yang's retirement symposium from Stony Brook in May 1999



Fig. 1.5 At the occasion of the 100th birthday of Frank Yang, the YITP (the Yang Institute for Theoretical Physics), the Physics Department, and the Simons Center for Geometry and Physics at Stony Brook organized a special colloquium with one hundred red and white balloons

Acknowledgments I thank George Sterman for a careful reading of this article, and Jacques Verbaarschot for finding the photo in Fig. 1.2, and Martin Roček for retrieving the poem about Frank and Rubik's cube.

References

1. J. Simons, C.N. Yang, Stony Brook Masters Series. https://youtu.be/zVWlapujbfo. Stony Brook University (2009)
2. E. Fermi, C.N. Yang, Are Mesons elementary particles? Phys. Rev. **76**, 1739 (1949)
3. C.N. Yang, Possible experimental determination of whether the neutral meson is scalar or pseudoscalar. Phys. Rev. **77**, 722 (1950)
4. C.N. Yang, J. Tiomno, Reflection properties of spin $\frac{1}{2}$ fields and a universal Fermi-type interaction. Phys. Rev. **79**, 495 (1950)
5. C.N. Yang, R. Mills, Conservation of isotopic spin and isotopic gauge invariance. Phys. Rev. **96**, 191 (1954)
6. T.D. Lee, C.N. Yang, Question of parity conservation in weak interactions. Phys. Rev. **104**, 254 (1956)
7. Symmetries & Modern Physics: Yang Retirement Symposium, ed. by A. Goldhaber, R. Shrock, J. Smith, G. Sterman, P. van Nieuwenhuizen, W. Weisberger (World Scientific, Singapore, 2003)

Chapter 2
A Stacky Approach to Crystals



Vladimir Drinfeld

To Professor C. N. Yang with deepest admiration

Abstract Inspired by a theorem of Bhatt–Morrow–Scholze, we develop a stacky approach to crystals and isocrystals on "Frobenius-smooth" schemes over $\mathbb{F}_p$. This class of schemes goes back to Berthelot-Messing and contains all smooth schemes over perfect fields of characteristic p. To treat isocrystals, we prove some descent theorems for sheaves of Banachian modules, which could be interesting in their own right.

Keywords F-isocrystal · Slope filtration · Tannakian category · Formal group · Semiperfect ring

2.1 Introduction

Fix a prime p.

2.1.1 A Theorem of Bhatt–Morrow–Scholze

Let X be a smooth scheme over a perfect field k of characteristic p. Let X_{perf} be the perfection of X, i.e.,

$$X_{\mathrm{perf}} := \varprojlim(\ldots \xrightarrow{\mathrm{Fr}} X \xrightarrow{\mathrm{Fr}} X \xrightarrow{\mathrm{Fr}} X).$$

V. Drinfeld (✉)
Department of Mathematics, University of Chicago, Chicago, IL, USA
e-mail: drinfeld@uchicago.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_2

Let $W(X_{\mathrm{perf}})$ be the p-adic formal scheme whose underlying topological space is that of X_{perf} (or of X) and whose structure sheaf is obtained by applying to $\mathcal{O}_{X_{\mathrm{perf}}}$ the functor of p-typical Witt vectors. Theorem 1.10 of [8] can be reformulated as a canonical isomorphism

$$R\Gamma_{\mathrm{cris}}(X, \mathcal{O}) = R\Gamma(W(X_{\mathrm{perf}})/\mathcal{G}, \mathcal{O}),$$

where $\mathcal{G}$ is a certain flat affine groupoid (in the category of p-adic formal schemes) acting on $W(X_{\mathrm{perf}})$, and $W(X_{\mathrm{perf}})/\mathcal{G}$ is the quotient stack.[1] If X is affine this gives an explicit complex computing $R\Gamma_{\mathrm{cris}}(X, \mathcal{O})$ (the complex is constructed using the nerve of $\mathcal{G}$).

One can define $\mathcal{G}$ to be the unique groupoid acting on $W(X_{\mathrm{perf}})$ such that the morphism $\mathcal{G} \to W(X_{\mathrm{perf}}) \times W(X_{\mathrm{perf}})$ is obtained by taking the divided power envelope of the ideal of the closed subscheme

$$X_{\mathrm{perf}} \times_X X_{\mathrm{perf}} \subset W(X_{\mathrm{perf}}) \times W(X_{\mathrm{perf}}).$$

We prefer to define the groupoid $\mathcal{G}$ by describing its nerve $\mathcal{A}_\bullet$ using Fontaine's functor A_{cris} (see Sect. 2.2.2.3, in which we follow [8]).

2.1.2 A Generalization

Our first main result is Theorem 2.1. It says that if X is as above then a crystal of quasi-coherent $\mathcal{O}$-modules on the absolute crystalline site of X is the same as an $\mathcal{O}$-module M on the stack $W(X_{\mathrm{perf}})/\mathcal{G}$, and the complex $R\Gamma(W(X_{\mathrm{perf}})/\mathcal{G}, M)$ identifies with the cohomology of the corresponding crystal.

We think of crystalline cohomology not in terms of the de Rham complex but in terms of the bigger and more tautological Čech-Alexander complex. This approach makes Theorem 2.1 almost obvious.

Following Berthelot-Messing [4, §1], we allow X to be any $\mathbb{F}_p$-scheme X which is *Frobenius-smooth* in the sense of Sect. 2.2.1. E.g., if k is a perfect $\mathbb{F}_p$-algebra then $\operatorname{Spec} k[[x_1, \ldots, x_n]]$ or any smooth scheme over $\operatorname{Spec} k$ is allowed.

Remark 2.1 As explained in Appendix section "The Isomorphism Between $W(X_{\mathrm{perf}})/\mathcal{G}$ and the Prismatization of X", for any Frobenius-smooth $\mathbb{F}_p$-scheme X, the stack $W(X_{\mathrm{perf}})/\mathcal{G}$ canonically identifies with the *prismatization* of X. The notion of prismatization is not used in the main body of this article (in fact, it did not exist when the original version of the article was written).

[1] The stacks $(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes_{\mathbb{Z}_p} (\mathbb{Z}/p^r\mathbb{Z})$ are usually *not* Artin stacks because the two canonical morphisms $\mathcal{G} \to W(X_{\mathrm{perf}})$ usually have infinite type.

2.1.3 Isocrystals

Now let X be a Noetherian $\mathbb{F}_p$-scheme. For such schemes Frobenius-smoothness is equivalent to Ω^1_X being locally free and coherent (see Sect. 2.3.1); let us assume this. E.g., if k is a perfect field of characteristic p then $\operatorname{Spec} k[[x_1, \ldots, x_n]]$ or any smooth scheme over $\operatorname{Spec} k$ is allowed.

2.1.3.1 What We Mean by an Isocrystal

Consider the category of crystals of finitely generated quasi-coherent $\mathcal{O}$-modules on the absolute crystalline site of X. Tensoring it by $\mathbb{Q}$, one gets a category denoted by $\operatorname{Isoc}(X)$, whose objects are called *isocrystals*. (Thus our isocrystals are not necessarily convergent in the sense of [19].)

2.1.3.2 The Result on Isocrystals

Our second main result (Theorem 2.2) provides a canonical equivalence

$$\operatorname{Isoc}(X) \xrightarrow{\sim} \operatorname{Bun}_{\mathbb{Q}}(W(X_{\mathrm{perf}})/\mathcal{G}), \tag{2.1}$$

where the right-hand side is, so to say, the category of vector bundles on $(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ (more details are explained in Sect. 2.1.3.3 below).

2.1.3.3 "Banachian Games" and $\operatorname{Bun}_{\mathbb{Q}}$

For any $\mathbb{Z}_p$-flat p-adic formal scheme[2] $\mathcal{Y}$, let $\operatorname{Bun}_{\mathbb{Q}}(\mathcal{Y})$ denote the category of finitely generated locally projective $(\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})$-modules. We prove that the assignment $\mathcal{Y} \mapsto \operatorname{Bun}_{\mathbb{Q}}(\mathcal{Y})$ satisfies fpqc descent (see Proposition 2.6). Because of this, the meaning of $\operatorname{Bun}_{\mathbb{Q}}(W(X_{\mathrm{perf}})/\mathcal{G})$ is clear.

More generally, in Sect. 2.3.6 we prove fpqc descent for the category of sheaves of Banachian[3] $(\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})$-modules, denoted by $\operatorname{QCoh}^{\flat}(\mathcal{Y}) \otimes \mathbb{Q}$ (see Corollary 2.5).

[2] See Sect. 2.2.3.1 for the precise meaning of these words. Note that $\mathcal{Y}$ does not have to be Noetherian.

[3] By a Banachian space over $\mathbb{Q}_p$ we mean a complete topological vector space such that the topology comes from a non-Archimedean norm.

2.2 Crystals and Crystalline Cohomology

2.2.1 A Class of Schemes

Lemma 2.1 *The following properties of an $\mathbb{F}_p$-scheme X are equivalent:*

(i) the Frobenius morphism $\mathrm{Fr} : X \to X$ *is syntomic (i.e., flat, of finite presentation, and with each fiber being a locally complete intersection);*

(ii) every point of X has a neighborhood isomorphic to $\operatorname{Spec} B$ *for some $\mathbb{F}_p$-algebra B with a finite p-basis in the sense of [4, Def. 1.1.1] (i.e., there exist $x_1, \dots, x_d \in B$ such that every element of B can be written uniquely as $\sum_{\alpha \in J} b_\alpha^p x^\alpha$, $b_\alpha \in B$, where $J := \{0, 1, \dots, p-1\}^d$ and for $\alpha = (\alpha_1, \dots, \alpha_d) \in J$ one sets $x^\alpha := x_1^{\alpha_1} \cdot \ldots \cdot x_d^{\alpha_d}$).*

$\mathbb{F}_p$-schemes satisfying the equivalent conditions of Lemma 2.1 will be called *Frobenius-smooth*. We are mostly interested in the following two classes of such schemes.

Example 2.1 X is a smooth scheme over a perfect field k of characteristic p.

Example 2.2 $X = \operatorname{Spec} k[[x_1, \dots, x_n]]$, where k is a perfect field of characteristic p.

Remark 2.2 The equivalent conditions of Lemma 2.1 still hold if in Examples 2.1–2.2 one replaces "perfect field" with "perfect $\mathbb{F}_p$-algebra".

Remark 2.3 In [4, §1] Berthelot and Messing study $\mathbb{F}_p$-algebras with a *not necessarily finite* p-basis; they prove that in some respects they are as good as smooth $\mathbb{F}_p$-algebras. More results of this type can be found in [9, §1]. In particular, any $\mathbb{F}_p$-algebra with a p-basis is formally smooth (see [4, §1.1.1]), and moreover, its cotangent complex is a free module concentrated in degree 0 (see [9, Lemma 1.1.2]).

Lemma 2.2 *Let X be an $\mathbb{F}_p$-scheme whose Frobenius endomorphism is finite. Let $n \in \mathbb{N}$. Let $\Delta \subset X^n$ be the diagonal X and $\Delta' := (\mathrm{Fr}_{X^n})^{-1}(\Delta) \subset X^n$. Let $I \subset \mathcal{O}_{\Delta'}$ be the ideal of $\Delta \subset \Delta'$. Then I is finitely generated.*

Let us recall that according to the definition from EGA, "finitely generated" really means "locally finitely generated".

Proof Let $\pi : \Delta' \hookrightarrow X^n \to X$ be one of the n projections. The morphism π is finite because Fr_X is. So the middle term of the exacts sequence

$$0 \to \pi_* I \to \pi_* \mathcal{O}_{\Delta'} \to \mathcal{O}_X \to 0$$

is a finitely generated $\mathcal{O}_X$-module. Since our exact sequence is locally split, this implies that $\pi_* I$ is a finitely generated $\mathcal{O}_X$-module. So I is a finitely generated ideal in $\mathcal{O}_{\Delta'}$. □

Corollary 2.1 *In the situation of Lemma 2.2 the ideal I is nilpotent on every quasi-compact open subset of X.*

2.2.2 Some Simplicial Formal Schemes

Let X be an $\mathbb{F}_p$-scheme.

2.2.2.1 The Simplicial Scheme $\mathcal{P}_\bullet$

Let X_{perf} be the perfection of X, i.e.,

$$X_{\text{perf}} := \varprojlim(\ldots \xrightarrow{\text{Fr}} X \xrightarrow{\text{Fr}} X \xrightarrow{\text{Fr}} X).$$

For an integer $n \geq 0$, let $[n] := \{0, 1, \ldots, n\}$ and let $\mathcal{P}_n \subset X_{\text{perf}}^{[n]}$ be the preimage of the diagonal $X \hookrightarrow X^{[n]}$; equivalently, $\mathcal{P}_n$ is the fiber product (over X) of $n+1$ copies of X_{perf}. The schemes $\mathcal{P}_n$ form a simplicial scheme $\mathcal{P}_\bullet = \mathcal{P}_\bullet(X)$ over X.

The scheme $\mathcal{P}_0 = X_{\text{perf}}$ is perfect. For any n, the scheme $\mathcal{P}_n$ is *semiperfect* in the sense of [20], i.e., the Frobenius endomorphism of $\mathcal{P}_n$ is a closed embedding. Note that the underlying topological space of $\mathcal{P}_n$ is that of X. As usual, the structure sheaf of $\mathcal{P}_n$ is denoted by $\mathcal{O}_{\mathcal{P}_n}$.

2.2.2.2 The Simplicial Formal Scheme $\mathcal{F}_\bullet$

Let $\mathcal{F}_n$ be the "Fontainization" of $\mathcal{P}_n$, i.e.,

$$\mathcal{F}_n = \varinjlim(\mathcal{P}_n \xrightarrow{\text{Fr}} \mathcal{P}_n \xrightarrow{\text{Fr}} \mathcal{P}_n \xrightarrow{\text{Fr}} \ldots).$$

Each $\mathcal{F}_n$ is a perfect formal scheme over $\mathbb{F}_p$, whose underlying topological space is that of X. These formal schemes form a simplicial formal scheme $\mathcal{F}_\bullet = \mathcal{F}_\bullet(X)$.

Corollary 2.1 implies that if Fr_X is finite then $\mathcal{F}_n$ is equal to the formal completion of $X_{\text{perf}}^{[n]} = X_{\text{perf}}^{n+1}$ along the closed subscheme $\mathcal{P}_n \subset X_{\text{perf}}^{[n]}$.

2.2.2.3 The Simplicial Formal Scheme $\mathcal{A}_\bullet$

Let $\mathcal{A}_n$ be the p-adic formal scheme obtained from $\mathcal{P}_n$ by applying Fontaine's functor A_{cris} (see [14, §2.2] or [11, §2.4]). So the underlying topological space of $\mathcal{A}_n$ is that of X, and the structure sheaf $\mathcal{O}_{\mathcal{A}_n}$ is the p-adic completion of the PD hull[4] of the surjection $W(\mathcal{O}_{\mathcal{F}_n}) \twoheadrightarrow \mathcal{O}_{\mathcal{P}_n}$ (as usual, PD stands for "divided powers" and W for the p-typical Witt vectors). The formal schemes $\mathcal{A}_n$ form a simplicial formal scheme $\mathcal{A}_\bullet = \mathcal{A}_\bullet(X)$.

By functoriality, the morphism $\mathrm{Fr}_X : X \to X$ induces a canonical simplicial morphism $F : \mathcal{A}_\bullet(X) \to \mathcal{A}_\bullet(X)$. The morphism $F : \mathcal{A}_n \to \mathcal{A}_n$ is usually *not an isomorphism* if $n > 0$ (if $n = 0$ it is an isomorphism because $\mathcal{A}_0 = W(X_{\mathrm{perf}})$).

2.2.3 Notation and Terminology Related to Quasi-Coherent Sheaves

2.2.3.1 p-adic Formal Schemes and Stacks

By a p-adic formal scheme (resp. p-adic formal stack) $\mathcal{Y}$ we mean a sequence of schemes (resp. stacks) $\mathcal{Y}_n$ equipped with isomorphisms $\mathcal{Y}_{n+1} \otimes_{\mathbb{Z}} \mathbb{Z}/p^n\mathbb{Z} \xrightarrow{\sim} \mathcal{Y}_n$. (Then $\mathcal{Y}_n$ is over $\mathbb{Z}/p^n\mathbb{Z}$ and $\mathcal{Y}_{n+1} \otimes_{\mathbb{Z}} \mathbb{Z}/p^n\mathbb{Z} = \mathcal{Y}_{n+1} \otimes_{\mathbb{Z}/p^{n+1}\mathbb{Z}} \mathbb{Z}/p^n\mathbb{Z}$). In this situation we often write $\mathcal{Y} \otimes \mathbb{Z}/p^n\mathbb{Z}$ instead of $\mathcal{Y}_n$.

$\mathcal{Y}$ is said to be $\mathbb{Z}_p$-flat if each $\mathcal{Y}_n$ is flat over $\mathbb{Z}/p^n\mathbb{Z}$.

2.2.3.2 The Notation QCoh($\mathcal{Y}$)

If $\mathcal{Y}$ is a scheme or stack we write $\mathrm{QCoh}(\mathcal{Y})$ for the category of quasi-coherent sheaves on $\mathcal{Y}$.

Now suppose that $\mathcal{Y}$ and $\mathcal{Y}_n$ are as in Sect. 2.2.3.1. Then we write $\mathrm{QCoh}(\mathcal{Y})$ for the projective limit of the categories $\mathrm{QCoh}(\mathcal{Y}_n)$ with respect to the pullback functors $\mathrm{QCoh}(\mathcal{Y}_{n+1}) \to \mathrm{QCoh}(\mathcal{Y}_n)$ (so an object of $\mathrm{QCoh}(\mathcal{Y})$ is a sequence of objects $\mathcal{F}_n \in \mathrm{QCoh}(\mathcal{Y}_n)$ with isomorphisms $\mathcal{F}_{n+1}/p^n\mathcal{F}_{n+1} \xrightarrow{\sim} \mathcal{F}_n$). Note that $\mathrm{QCoh}(\mathcal{Y})$ is a Karoubian additive category (but usually not an abelian one). We often write $\mathcal{F}/p^n\mathcal{F}$ instead of $\mathcal{F}_n$.

[4] The definition of PD hull is given in [2, §I.2.3]. Throughout this article, PD hulls are taken in the category of PD algebras over $(\mathbb{Z}_p, p\mathbb{Z}_p)$ (rather than over $(\mathbb{Z}_p, 0)$). In other words, we assume that $\gamma_n(p)$ is equal to $p^n/n! \in p\mathbb{Z}_p$.

2.2.3.3 $\mathbb{Z}_p$-Flatness

Let $\mathcal{Y}$ and $\mathcal{Y}_n$ be as in Sect. 2.2.3.1 and let $\mathcal{F} \in \mathrm{QCoh}(\mathcal{Y})$. We say that $\mathcal{F}$ is $\mathbb{Z}_p$-*flat* if each $\mathcal{F}_n$ is flat over $\mathbb{Z}/p^n\mathbb{Z}$. Let $\mathrm{QCoh}^\flat(\mathcal{Y}) \subset \mathrm{QCoh}(\mathcal{Y})$ be the full subcategory of $\mathbb{Z}_p$-flat objects.

2.2.3.4 Finite Generation

We say that $\mathcal{F} \in \mathrm{QCoh}(\mathcal{Y})$ is *finitely generated* if $\mathcal{F}_n$ is finitely generated[5] for each n (or equivalently, for $n = 1$). Let $\mathrm{QCoh}_{\mathrm{fin}}(\mathcal{Y}) \subset \mathrm{QCoh}(\mathcal{Y})$ be the full subcategory of finitely generated objects. Let $\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}) := \mathrm{QCoh}_{\mathrm{fin}}(\mathcal{Y}) \cap \mathrm{QCoh}^\flat(\mathcal{Y})$.

2.2.3.5 Cohomology

Let $\mathcal{Y}$ and $\mathcal{Y}_n$ be as in Sect. 2.2.3.1. Let $\mathcal{F}$ and $\mathcal{F}_n$ be as in Sect. 2.2.3.2. Then we define $R\Gamma(\mathcal{Y}, \mathcal{F})$ to be the homotopy projective limit of $R\Gamma(\mathcal{Y}_n, \mathcal{F}_n)$.

2.2.3.6 Equivariant Objects

Let $\Gamma \rightrightarrows \mathcal{Y}$ be a groupoid in the category of p-adic formal schemes. Assume that the canonical morphisms $\Gamma \to \mathcal{Y}$ are flat, quasi-compact, and quasi-separated. Then for each n we have the quotent stack $\mathcal{Y}_n/\Gamma_n$ (the words "quotient" and "stack" are understood in the sense of the fpqc topology). The stacks $\mathcal{Y}_n/\Gamma_n$ form a p-adic formal stack $\mathcal{Y}/\Gamma$. In this situation objects of $\mathrm{QCoh}(\mathcal{Y}/\Gamma)$ are also called Γ-*equivariant objects* of $\mathrm{QCoh}(\mathcal{Y})$.

2.2.3.7 Objects of QCoh($\mathcal{Y}$) as Sheaves

Let $\mathcal{Y}$ be a p-adic formal scheme and $\mathcal{F} \in \mathrm{QCoh}(\mathcal{Y})$. Let $\mathcal{F}_\infty$ be the projective limit of the sheaves $\mathcal{F}_n = \mathcal{F}/p^n\mathcal{F}$ in the category of presheaves on the topological space $\mathcal{Y}$; it is clear that $\mathcal{F}_\infty$ is a sheaf. As explained to me by M. Temkin,

$$\mathcal{F}_\infty/p^n\mathcal{F}_\infty = \mathcal{F}_n \tag{2.2}$$

(this will be proved in Sect. 2.2.3.8). Moreover,

$$R\Gamma(\mathcal{Y}, \mathcal{F}_\infty) = R\Gamma(\mathcal{Y}, \mathcal{F}), \tag{2.3}$$

[5] Let us recall that according to the definition from EGA, "finitely generated" really means "locally finitely generated".

where $R\Gamma(\mathcal{Y}, \mathcal{F})$ was defined in Sect. 2.2.3.5: indeed, for any open affine $U \subset \mathcal{Y}$ the derived projective limit of $H^0(U, \mathsf{F}_n)$ equals the usual one because the maps $H^0(U, \mathsf{F}_{n+1}) \to H^0(U, \mathsf{F}_n)$ are surjective.

Because of (2.2) and (2.3), we will usually not distinguish $\mathcal{F}_\infty$ from $\mathcal{F}$.

Let $\mathcal{O}_\mathcal{Y}$ denote the projective limit of the sheaves $\mathcal{O}_{\mathcal{Y}_n}$. Then the ringed space $(\mathcal{Y}, \mathcal{O}_\mathcal{Y})$ is a formal scheme in the sense of EGA, and $\mathcal{F}_\infty$ is a sheaf of $\mathcal{O}_\mathcal{Y}$-modules.

2.2.3.8 Proof of (2.2)

It suffices to show that for every open affine $U \subset \mathcal{Y}$ the sequence

$$H^0(U, \mathcal{F}_\infty) \xrightarrow{p^n} H^0(U, \mathcal{F}_\infty) \to H^0(U, \mathcal{F}_n) \to 0$$

is exact. Let $M := H^0(U, \mathcal{F}_\infty)$ and $M_n := H^0(U, \mathcal{F}_n)$. Then M is the projective limit of M_n. The transition maps $M_{n+1} \to M_n$ are surjective, so the map $M \to M_n$ is surjective for each n.

Let $M^n := \mathrm{Ker}(M \twoheadrightarrow M_n)$, so $M_n = M/M^n$. We have to show that $M^n = p^n M$.

Note that M is separated and complete with respect to the filtration formed by M^i, $i \in \mathbb{N}$. Since $M_i = M_{i+1}/p^i M_{i+1}$, we have

$$M^i = p^i M + M^{i+1}. \tag{2.4}$$

So $M^n = p^n M + M^{n+1} \supset p^n M$. On the other hand, if $x \in M^n$ then applying (2.4) for $i = n, n+1, n+2, \ldots$, we get

$$x = p^n m_0 + x_1, \quad x_1 = p^{n+1} m_1 + x_2, \quad x_2 = p^{n+2} m_2 + x_3, \ \ldots$$

for some elements $m_j \in M$ and $x_j \in M^{n+j}$. Then $x = p^n m$, where $m = \sum_{j=0}^{\infty} p^j m_j$ (this infinite series converges because $p^j m_j \in M^j$ and M is complete).

2.2.4 Formulation of the Results

2.2.4.1 Convention

By a *crystal* on an $\mathbb{F}_p$-scheme X we mean a crystal of quasi-coherent $\mathcal{O}$-modules on the absolute crystalline site $\mathrm{Cris}(X)$. For the definition of this site, see the end of §III.1.1.3 of [2].

The proof of the following theorem will be given in Sects. 2.2.5–2.2.7.

Theorem 2.1 *Let X be a Frobenius-smooth scheme (i.e., an $\mathbb{F}_p$-scheme satisfying the equivalent conditions of Lemma 2.1).*

(i) The simplicial formal scheme $\mathcal{A}_\bullet$ from Sect. 2.2.2.3 is the nerve[6] of a flat affine[7] groupoid $\mathcal{G}$ acting on $\mathcal{A}_0 = W(X_{\mathrm{perf}})$.

(ii) A crystal on X is the same as an object $M \in \mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathcal{G})$ (or equivalently, a $\mathcal{G}$-equivariant object of $\mathrm{QCoh}(W(X_{\mathrm{perf}}))$.

(iii) For M as above, $R\Gamma_{\mathrm{cris}}(X, M) = R\Gamma(W(X_{\mathrm{perf}})/\mathcal{G}, M)$.

In the case $M = \mathcal{O}$ statement (iii) is equivalent to [8, Thm. 1.10], but our proof is different (we think of crystalline cohomology in terms of the Čech-Alexander complex[8] rather than the de Rham complex).

Remark 2.4 An object $M \in \mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathcal{G})$ defines a collection of objects $M^n \in \mathrm{QCoh}(\mathcal{A}_n)$. For each $r \in \mathbb{N}$, the sheaves $M^n/p^r M^n$ form a cosimplicial sheaf[9] on the topological space X. Statement (iii) of the theorem can be reformulated as a canonical isomorphism

$$R\Gamma_{\mathrm{cris}}(X, M) \xrightarrow{\sim} \varprojlim_r R\Gamma(X, \mathrm{Tot}(M^\bullet/p^r M^\bullet)). \tag{2.5}$$

Remark 2.5 The stacks $(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes_{\mathbb{Z}_p} (\mathbb{Z}/p^r\mathbb{Z})$ are usually *not* Artin stacks because the two canonical morphisms $\mathcal{G} \to W(X_{\mathrm{perf}})$ usually have infinite type.

2.2.5 Proof of Theorem 2.1(i)

We can assume that $X = \mathrm{Spec}\, B$, where B is an $\mathbb{F}_p$-algebra with a finite p-basis.

2.2.5.1 The Simplicial Formal Scheme $\mathcal{X}_\bullet$

By [4, Prop. 1.1.7] or by [9, Remark 1.2.3(a)], X admits a lift $\mathcal{X} = \mathrm{Spf}\,\mathsf{B}$, where B is a flat p-adically complete $\mathbb{Z}_p$-algebra with $\mathsf{B}/p\mathsf{B} = B$. Fix $\mathcal{X}$. Let $\mathcal{X}_n$ be the p-adically completed PD hull of the (ideal of the) diagonal

$$X \hookrightarrow X^{n+1} \subset \mathcal{X}^{n+1} = \mathcal{X}^{[n]}.$$

[6] Recall that a groupoid (and more generally, a category) is uniquely determined by its nerve.

[7] This means that the two canonical morphisms $\mathcal{G} \to W(X_{\mathrm{perf}})$ are flat and affine.

[8] The Čech-Alexander complex is discussed in [15, §5.1, 5.5], [5, §2], and [1, §1.7-1.8].

[9] More precisely, a cosimplicial sheaf of modules over the cosimplicial sheaf of rings formed by the structure sheaves of the formal schemes $\mathcal{A}_n$.

Then $\mathcal{X}_n$ is a p-adic formal scheme whose underlying topological space is X. The formal schemes $\mathcal{X}_n$ form a simplicial formal scheme $\mathcal{X}_\bullet$.

Fix a p-basis $x_1, \dots, x_d \in B$. Let $\widetilde{x}_j \in \mathcal{B}$, $\widetilde{x}_j \mapsto x_j$. We have canonical embeddings $i_m : \mathcal{B} = H^0(\mathcal{X}, \mathcal{O}_{\mathcal{X}}) \hookrightarrow H^0(\mathcal{X}^{[n]}, \mathcal{O}_{\mathcal{X}^{[n]}})$.

Proposition 2.1 *As an $i_m(\mathcal{B})$-algebra, $H^0(\mathcal{X}_n, \mathcal{O}_{\mathcal{X}_n})$ is the p-adically completed algebra of PD polynomials over $i_m(\mathcal{B})$ with respect to the elements $i_r(\widetilde{x}_j) - i_m(\widetilde{x}_j)$, where $r \in \{0, \dots, n\}$, $r \neq m$, $1 \leq j \leq d$.*

Proof For $n = 2$ this is [4, Cor. 1.3.2(i)]. The general case follows. □

Corollary 2.2 *$\mathcal{X}_\bullet$ is the nerve of a flat affine groupoid $\mathcal{R}$ acting on $\mathcal{X}$.*

Lemma 2.3 *The formation of PD hulls commutes with flat base change.*

Proof This was proved by Berthelot [2, Prop. I.2.7.1]. □

2.2.5.2 End of the Proof

By [4, Prop. 1.2.6], the morphism $\pi : X_{\mathrm{perf}} \to X$ lifts to a morphism $\widetilde{\pi} : W(X_{\mathrm{perf}}) \to \mathcal{X}$. Since $\mathrm{Fr}_X : X \to X$ is flat, so are π and $\widetilde{\pi}$.

Let $\mathcal{P}_n$ be as in Sect. 2.2.2.1. The formal scheme $\mathcal{A}_n$ from Sect. 2.2.2.3 is the p-adically completed PD hull of $\mathcal{P}_n$ in $W(X_{\mathrm{perf}})^{[n]}$. So by Lemma 2.3 and flatness of $\widetilde{\pi}$, the diagram

$$\begin{array}{ccc} \mathcal{A}_n & \xrightarrow{\widetilde{\pi}_n} & \mathcal{X}_n \\ \downarrow & & \downarrow \\ W(X_{\mathrm{perf}})^{[n]} & \longrightarrow & \mathcal{X}^{[n]} \end{array} \tag{2.6}$$

is Cartesian. So Corollary 2.2 implies that $\mathcal{A}_\bullet$ is the nerve of a flat groupoid acting on $W(X_{\mathrm{perf}})$.

Remark 2.6 In the situation of Sect. 2.2.5.2, $\widetilde{\pi} : W(X_{\mathrm{perf}}) \to \mathcal{X}$ induces an isomorphism

$$W(X_{\mathrm{perf}})/\mathcal{G} \xrightarrow{\sim} \mathcal{X}/\mathcal{R}, \tag{2.7}$$

where $\mathcal{G}$ and $\mathcal{R}$ are as in Theorem 2.1(i) and Corollary 2.2, respectively. This follows from faithful flatness of $\widetilde{\pi}$ and the fact that the diagram (2.6) is Cartesian.

Remark 2.7 The isomorphism (2.7) does not depend on the choice of $\widetilde{\pi}$. Indeed, if $\widetilde{\pi}' : W(X_{\mathrm{perf}}) \to \mathcal{X}$ is another lift of π then the morphism $W(X_{\mathrm{perf}}) \xrightarrow{(\widetilde{\pi}, \widetilde{\pi}')} \mathcal{X} \times \mathcal{X}$

factors through $\mathcal{X}_1$; to see this, use the PD structure on the ideal of the subscheme $X_{\mathrm{perf}} \subset W(X_{\mathrm{perf}})$.

2.2.6 *Proof of Theorem 2.1(ii)*

2.2.6.1 General Remark

Recall that if Y is a semiperfect $\mathbb{F}_p$-scheme then $A_{\mathrm{cris}}(Y) \otimes (\mathbb{Z}/p^r\mathbb{Z})$ is the final object of the crystalline site of Y over $\mathbb{Z}/p^r\mathbb{Z}$ (see [14, §2.2] or [11, Prop. 2.2.1]). So a crystal on Y is the same as an object $N \in \mathrm{QCoh}(A_{\mathrm{cris}}(Y))$, and the crystalline cohomology of the crystal is just $R\Gamma(A_{\mathrm{cris}}(Y), N)$.

2.2.6.2 The Functor in One Direction

Let us apply Sect. 2.2.6.1 to the semiperfect schemes $\mathcal{P}_n$ from Sect. 2.2.2.1. Recall that $\mathcal{A}_n := A_{\mathrm{cris}}(\mathcal{P}_n)$. So given a crystal on X, its pullback to $\mathcal{P}_n$ can be viewed as an object $M^n \in \mathrm{QCoh}(\mathcal{A}_n)$. Moreover, the collection of objects M^n is compatible via $*$-pullbacks.[10] By the definition of the groupoid $\mathcal{G}$ (see Theorem 2.1(i)), such a compatible collection is the same as a $\mathcal{G}$-equivariant object of $\mathrm{QCoh}(\mathcal{A}_0) = \mathrm{QCoh}(W(X_{\mathrm{perf}}))$ or equivalently, an object of $\mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathcal{G})$. Thus we have constructed a functor

$$\{\text{Crystals on } X\} \to \mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathcal{G}). \tag{2.8}$$

It remains to prove that the functor (2.8) is an equivalence. The question is local, so we can assume that X is the spectrum of an $\mathbb{F}_p$-algebra with a finite p-basis.

2.2.6.3 Factorizing the Functor (2.8)

Let $\mathcal{X}$ and $\mathcal{X}_\bullet$ be as in Sect. 2.2.5.1. Each scheme of the form $\mathcal{X}_n \otimes (\mathbb{Z}/p^r\mathbb{Z})$ is a PD thickening of X, so a crystal M on X defines for each n an object $M_{\mathcal{X}_n} \in \mathrm{QCoh}(\mathcal{X}_n)$. This collection of objects is compatible via $*$-pullbacks.[11] By the definition of the groupoid $\mathcal{R}$ (see Corollary 2.2), such a compatible collection is the same as an $\mathcal{R}$-equivariant object of $\mathrm{QCoh}(\mathcal{X})$ or equivalently, an object of

[10] By this we mean that for every map $f : [m] \to [n]$ one has a canonical isomorphism $(f^+)^* M^m \xrightarrow{\sim} M^n$, where $f^+ : X^{[n]} \to X^{[m]}$ is induced by f, and these isomorphisms are compatible with composition of f's.

[11] Here we use that the map $\mathcal{X}_n \otimes (\mathbb{Z}/p^r\mathbb{Z}) \to \mathcal{X}_m \otimes (\mathbb{Z}/p^r\mathbb{Z})$ corresponding to any map $[m] \to [n]$ is a PD morphism.

$\mathrm{QCoh}(\mathscr{X}/\mathscr{R})$. Thus we have constructed a functor

$$\{\text{Crystals on } X\} \to \mathrm{QCoh}(\mathscr{X}/\mathscr{R}). \tag{2.9}$$

The functor (2.8) is the composition of (2.9) and the equivalence

$$\mathrm{QCoh}(\mathscr{X}/\mathscr{R}) \xrightarrow{\sim} \mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathscr{G})$$

corresponding to (2.7). It remains to show that (2.9) is an equivalence. This is a consequence of the next lemma.

Lemma 2.4 *Suppose that X is the spectrum of a $\mathbb{F}_p$-algebra with a finite p-basis. Let $\mathscr{X}$ be as in Sect. 2.2.5.1. Then every object of the absolute crystalline site of X admits a morphism to $\mathscr{X} \otimes (\mathbb{Z}/p^r\mathbb{Z})$ for some r.*

Proof Follows from [4, Prop. 1.2.6]. □

2.2.7 Proof of Theorem 2.1(iii)

Let M be a crystal on X. We can also think of M as an object of $\mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathscr{G})$.

2.2.7.1 General Remark

Let X_{cris} be the absolute crystalline topos of X (i.e., the category of sheaves on $\mathrm{Cris}(X)$). Let $Y_\bullet$ be a simplicial ringed topos over X_{cris}. Then one has a canonical morphism

$$R\Gamma_{\mathrm{cris}}(X, M) \to \mathrm{Tot}(R\Gamma(Y_\bullet, M^\bullet)), \tag{2.10}$$

where M^n is the pullback of M to Y_n. Moreover, this construction is functorial in $Y_\bullet$.

2.2.7.2 The Map in One Direction

Apply (2.10) for $Y_\bullet = (\mathcal{P}_\bullet)_{\mathrm{cris}}$, where $\mathcal{P}_\bullet$ is as in Sect. 2.2.2.1. As explained in Sect. 2.2.6.1, we can think of M^n as an object of $\mathrm{QCoh}(\mathcal{A}_n)$, where $\mathcal{A}_n := A_{\mathrm{cris}}(\mathcal{P}_n)$; it is this object that was denoted by M^n in Remark 2.4. Moreover, $R\Gamma_{\mathrm{cris}}(\mathcal{P}_n, M^n) = R\Gamma(\mathcal{A}_n, M^n)$ by Sect. 2.2.6.1. Thus we can rewrite (2.10) as a morphism

$$R\Gamma_{\mathrm{cris}}(X, M) \to \mathrm{Tot}(R\Gamma(\mathcal{A}_\bullet, M^\bullet)) = R\Gamma(W(X_{\mathrm{perf}})/\mathscr{G}, M). \tag{2.11}$$

It remains to prove that the morphism (2.11) is an isomorphism. The question is local, so we can assume that X is the spectrum of an $\mathbb{F}_p$-algebra with a finite p-basis.

2.2.7.3 End of the Proof

Let $\mathcal{X}$ and $\mathcal{X}_\bullet$ be as in Sect. 2.2.5.1. By Lemma 2.4, we get an isomorphism

$$R\Gamma_{\text{cris}}(X, M) \xrightarrow{\sim} \text{Tot}(R\Gamma(\mathcal{X}_\bullet, M_{\mathcal{X}_\bullet})) = \text{Tot}(\Gamma(\mathcal{X}_\bullet, M_{\mathcal{X}_\bullet})) \tag{2.12}$$

where $M_{\mathcal{X}_\bullet}$ is as in Sect. 2.2.6.3; $\text{Tot}(\Gamma(\mathcal{X}_\bullet, M_{\mathcal{X}_\bullet}))$ is called the *Čech-Alexander complex*.

We have

$$\text{Tot}(\Gamma(\mathcal{X}_\bullet, M_{\mathcal{X}_\bullet})) = R\Gamma(\mathcal{X}/\mathcal{R}, M),$$

where $\mathcal{R}$ is the groupoid from Corollary 2.2 and $\mathcal{X}/\mathcal{R}$ is the quotient stack (just as in Sect. 2.2.6.3, we can think of M as an object of $\text{QCoh}(\mathcal{X}/\mathcal{R})$). It remains to show that (2.11) is the usual isomorphism

$$R\Gamma(\mathcal{X}/\mathcal{R}, M) \to R\Gamma(W(X_{\text{perf}})/\mathcal{G}, M)$$

corresponding to the isomorphism of stacks (2.7). This follows from the next lemma.

Lemma 2.5 *Choose $\widetilde{\pi} : W(X_{\text{perf}}) \to \mathcal{X}$ as in Sect. 2.2.5.2 and let $\widetilde{\pi}_n : \mathcal{A}_n \to \mathcal{X}_n$ be as in diagram* (2.6)*. Then the map* (2.11) *equals the composition of* (2.12) *and the morphism*

$$\text{Tot}(R\Gamma(\mathcal{X}_\bullet, M_{\mathcal{X}_\bullet})) \to \text{Tot}(R\Gamma(\mathcal{A}_\bullet, M^\bullet))$$

that comes from the maps $\widetilde{\pi}_n^* : R\Gamma(\mathcal{X}_n, M_{\mathcal{X}_n}) \to R\Gamma(\mathcal{A}_n, M^n)$.

Proof For each n, $\mathcal{X}_n$ is an object of the crystalline topos X_{cris}. Let $X_{\text{cris}}/\mathcal{X}_n$ be the category of objects of X_{cris} over $\mathcal{X}_n$; this category is a ringed topos over X_{cris}. As n varies, we get a simplicial ringed topos $X_{\text{cris}}/\mathcal{X}_\bullet$ over X_{cris}. Moreover, we have a morphism

$$(\mathcal{P}_\bullet)_{\text{cris}} \to X_{\text{cris}}/\mathcal{X}_\bullet$$

of simplicial ringed topoi over X_{cris}.

The morphism (2.12) is the morphism (2.10) for $Y_\bullet = X_{\text{cris}}/\mathcal{X}_\bullet$. So the lemma follows from functoriality of the map (2.10) with respect to $Y_\bullet$. □

2.2.8 $H^0_{\mathrm{cris}}(X, O)$ and the Ring of Constants

As before, we assume that X is Frobenius-smooth in the sense of Sect. 2.2.1. Let us compute $H^0_{\mathrm{cris}}(X, O)$, where O is the structure sheaf on the absolute crystalline site of X.

2.2.8.1 The Ring of Constants

Let A be the ring of regular functions on X and $k := \bigcap_{n=1}^{\infty} A^{p^n}$. We call k the *ring of constants* of X.

Frobenius-smoothness implies that X is reduced. So A is reduced, and k is a perfect $\mathbb{F}_p$-algebra.

Lemma 2.6 *Let $X = \operatorname{Spec} B$, where B is an $\mathbb{F}_p$-algebra with a p-basis*[12] *$x_1, \ldots, x_m$. Let $\pi : X \to \mathbb{G}_a^m$ be the morphism defined by $x_1, \ldots, x_m$. Let $\widehat{\mathbb{G}}_a$ be the formal additive group over $\mathbb{F}_p$.*

(i) There is a unique action of $\widehat{\mathbb{G}}_a^m$ on X such that $\pi : X \to \mathbb{G}_a^m$ is $\widehat{\mathbb{G}}_a^m$-equivariant.
(ii) The ring of constants on X is equal to the ring of all $\widehat{\mathbb{G}}_a^m$-invariant regular functions on X.
(iii) A closed subscheme $Y \subset X$ is preserved by the $\widehat{\mathbb{G}}_a^m$-action if and only if for each $n \in \mathbb{N}$ there exists a closed subscheme $Z \subset X$ such that $Y = (\mathrm{Fr}_X^n)^{-1}(Z)$.

Proof

(i) $\widehat{\mathbb{G}}_a$ is the inductive limit of the finite group schemes $\mathrm{Ker}(\mathbb{G}_a^m \xrightarrow{\mathrm{Fr}^n} \mathbb{G}_a^m)$, $n \in \mathbb{N}$. For every $n \in \mathbb{N}$ the diagram

$$\begin{array}{ccc} X & \xrightarrow{\mathrm{Fr}^n} & X \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ \mathbb{G}_a^m & \xrightarrow{\mathrm{Fr}^n} & \mathbb{G}_a^m \end{array}$$

is Cartesian (for $n = 1$ by the definition of p-basis, the general case follows). Any action of $\mathrm{Ker}(\mathbb{G}_a^m \xrightarrow{\mathrm{Fr}^n} \mathbb{G}_a^m)$ on X is fiberwise with respect to the morphism $\mathrm{Fr}^n : X \to X$. So there is one and only one such action with the property that $\pi : X \to \mathbb{G}_a^m$ is $\widehat{\mathbb{G}}_a^m$-equivariant.

(ii–iii) The morphism $\mathrm{Fr}^n : X \to X$ is a torsor with respect to $\mathrm{Ker}(\mathbb{G}_a^m \xrightarrow{\mathrm{Fr}^n} \mathbb{G}_a^m)$.

□

[12] The definition of p-basis was given in Lemma 2.2.1.

Proposition 2.2 *Let X be a Frobenius-smooth $\mathbb{F}_p$-scheme. Let k be the ring of constants of X. Then the map $W(k) \to H^0_{\mathrm{cris}}(X, \mathcal{O})$ induced by the morphism $X \to \operatorname{Spec} k$ is an isomorphism.*

Proof We can assume that $X = \operatorname{Spec} B$, where B is an $\mathbb{F}_p$-algebra with a p-basis $x_1, \ldots, x_m$. By [4, Prop. 1.1.7], there exists a flat p-adically complete $\mathbb{Z}_p$-algebra $\widetilde{B}$ with $\widetilde{B}/p\widetilde{B} = B$. For each i choose a lift of x_i to $\widetilde{B}$; this lift will still be noted by x_i. By [4, Prop. 1.3.1], for each $n \in \mathbb{N}$ the module of differentials of $\widetilde{B}/p^n\widetilde{B}$ is a free $(\widetilde{B}/p^n\widetilde{B})$-module with basis dx_i, $1 \le i \le m$. So for each $i \le m$ there is a unique derivation $D_i : \widetilde{B} \to \widetilde{B}$ such that $D_i(x_j) = \delta_{ij}$. One has

$$H^0_{\mathrm{cris}}(X, \mathcal{O}) = \bigcap_{i=1}^{m} \operatorname{Ker}(\widetilde{B} \xrightarrow{D_i} \widetilde{B});$$

this follows, e.g., from [4, Prop. 1.3.3] (because $H^0_{\mathrm{cris}}(X, \mathcal{O})$ is the $\mathbb{Z}_p$-module of endomorphisms of the crystal $\mathcal{O}$).

Since $x_1, \ldots, x_m$ form a p-basis in B, for each $r \in \mathbb{N}$ the ring $\widetilde{B}$ is topologically generated by $x_1, \ldots, x_m$ and elements of the form $\widetilde{f}^{p^r}$, $\widetilde{f} \in \widetilde{B}$. This implies that for each $i \le m$ and $l \in \mathbb{N}$ one has

$$D_i^l(\widetilde{B}) \subset l! \cdot \widetilde{B},$$

so one has the commuting operators $D_i^{(l)} := (l!)^{-1} D_i^l$ acting on $\widetilde{B}$ and satisfying the Leibniz formula

$$D_i^{(l)}(\widetilde{f}\widetilde{g}) = \sum_{a+b=l} D_i^{(a)}(\widetilde{f}) D_i^{(b)}(\widetilde{g}), \quad \widetilde{f}, \widetilde{g} \in \widetilde{B}$$

and the relation $D_i^{(r)} D_i^{(s)} = \binom{r+s}{r} D_i^{(r+s)}$. These operators define an action of $\widehat{\mathbb{G}}_a^m$ on $\widetilde{B}$, where $\widehat{\mathbb{G}}_a$ is the additive formal group over $\mathbb{Z}_p$. The corresponding action of $\widehat{\mathbb{G}}_a^m$ on B is the one from Lemma 2.6(i).

To prove that the map $W(k) \to H^0_{\mathrm{cris}}(X, \mathcal{O})$ is an isomorphism, it suffices to check that if $\widetilde{f} \in \widetilde{B}$ is killed by $D_1, \ldots, D_m$ and $f \in B$ is the image of $\widetilde{f}$ then $f \in k$. It is clear that $\widetilde{f}$ is $\widehat{\mathbb{G}}_a^m$-invariant. So f is $\widehat{\mathbb{G}}_a^m$-invariant. By Lemma 2.6(ii), this means that $f \in k$. □

2.3 Isocrystals

2.3.1 A Class of Schemes

Proposition 2.3 *The following properties of an $\mathbb{F}_p$-scheme X are equivalent:*

(i) X is Noetherian and Frobenius-smooth (see Lemma 2.1 and the sentence after it);

(ii) X is Noetherian, and Ω^1_X is locally free and finitely generated.

These properties imply that the scheme X is regular and excellent.

Proof It is clear that (i)⇒(ii). The implication (ii)⇒(i) follows from Theorem 1 of [21], whose proof is based on [13, Prop. 1]. E. Kunz proved that (i) implies regularity and excellence, see Theorems 107-108 of [18, §42]. □

Throughout Sect. 2.3, we assume that X has the equivalent properties of Proposition 2.3. For instance, if k is a perfect field of characteristic p one can take X to be either $\operatorname{Spec} k[[x_1, \dots, x_n]]$ or a quasi-compact smooth k-scheme.

2.3.1.1 The Ring of Constants

In Sect. 2.2.8.1 we defined the ring of constants k by $k := \bigcap_{n=1}^{\infty} A^{p^n}$, where A is the ring of regular functions on X. If X has the properties of Proposition 2.3 then $k = \bigcap_{n=1}^{\infty} E^{p^n}$, where E is the ring of rational functions on X: indeed, if $f \in \bigcap_{n=1}^{\infty} E^{p^n}$ and D is the divisor of poles of f then D is divisible by p^n for all $n \in \mathbb{N}$, so $D = 0$. Thus if X is irreducible then the ring of constants is a perfect field; in general, it is a product of perfect fields.

2.3.2 Coherent Crystals and Isocrystals

Let X be an $\mathbb{F}_p$-scheme that has the equivalent properties of Proposition 2.3. By a *coherent crystal* on an $\mathbb{F}_p$-scheme X we mean a crystal of finitely generated[13] quasi-coherent $\mathcal{O}$-modules on the absolute crystalline site of X.

Let $\mathcal{G}$ be the groupoid on $W(X_{\mathrm{perf}})$ constructed in Theorem 2.1(i). We have the categories

$$\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \subset \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \subset \mathrm{QCoh}(W(X_{\mathrm{perf}})/\mathcal{G}),$$

(we are using the notation of Sects. 2.2.3.2–2.2.3.4).

[13] Let us recall that according to the definition from EGA, "finitely generated" really means "locally finitely generated".

Lemma 2.7

(i) The equivalence Sect. 2.2.6.2 identifies the category of coherent crystals on X with $\mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G})$.
(ii) The category $\mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G})$ *is abelian.*
(iii) $\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes \mathbb{Q} = \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes \mathbb{Q}$

Proof By Remark 2.6, $W(X_{\mathrm{perf}})/\mathcal{G}$ is locally isomorphic to a quotient of a Noetherian formal scheme by a flat groupoid. This implies (ii–iii). Statement (i) follows from Lemma 2.4. □

Definition 2.1 Objects of the category

$$\mathrm{Isoc}(X) := \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes \mathbb{Q} = \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G}) \otimes \mathbb{Q}$$

will be called *isocrystals* on X.

(Thus our isocrystals are not necessarily convergent in the sense of [19].)

2.3.3 Local Projectivity

Proposition 2.4 *Let $\mathcal{X}$ be a $\mathbb{Z}_p$-flat p-adic formal scheme with $\mathcal{X} \otimes_{\mathbb{Z}_p} \mathbb{F}_p = X$. Let M be a coherent crystal on X and $M_{\mathcal{X}}$ the corresponding coherent $\mathcal{O}_{\mathcal{X}}$-module. Then the $(\mathcal{O}_{\mathcal{X}} \otimes \mathbb{Q})$-module $M_{\mathcal{X}} \otimes \mathbb{Q}$ is locally projective.*

A proof will be given in Sect. 2.3.4.

Corollary 2.3 *Let $M \in \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G})$. Let $M_{W(X_{\mathrm{perf}})}$ be the corresponding module over $\mathcal{O}_{W(X_{\mathrm{perf}})}$. Then $M_{W(X_{\mathrm{perf}})} \otimes \mathbb{Q}$ is a locally projective*[14] *module over $\mathcal{O}_{W(X_{\mathrm{perf}})} \otimes \mathbb{Q}$.*

Proof The statement is local on X. So by [4, Prop. 1.2.6], we can assume that the canonical morphism $X_{\mathrm{perf}} \to X$ lifts to a morphism $W(X_{\mathrm{perf}}) \to \mathcal{X}$, where $\mathcal{X} \in \mathrm{FrmSch}^{\flat}$, $\mathcal{X} \otimes_{\mathbb{Z}_p} \mathbb{F}_p = X$. It remains to apply Proposition 2.4. □

Corollary 2.4 *If X is irreducible then* $\mathrm{Isoc}(X)$ *is a Tannakian category over* $\mathrm{Frac}\, W(k)$*, where k is the field of constants of X (in the sense of Sect. 2.3.1.1).*

Proof The fact that $\mathrm{Isoc}(X)$ is abelian follows from Lemma 2.7(ii).

$W(X_{\mathrm{perf}})/\mathcal{G}$ is locally isomorphic to a quotient of a Noetherian formal scheme by a flat groupoid, so the tensor category $\mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G})$ has internal $\underline{\mathrm{Hom}}$'s. Proposition 2.4 implies that for any $M, N \in \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathcal{G})$, the canonical

[14] A module over a sheaf of rings is said to be locally projective if it can be locally represented as a direct summand of a free module.

morphism

$$\underline{\mathrm{Hom}}(M, O) \otimes N \to \underline{\mathrm{Hom}}(M, N)$$

becomes an isomorphism in the category $\mathrm{Isoc}(X) := \mathrm{QCoh}_{\mathrm{fin}}(W(X_{\mathrm{perf}})/\mathscr{G}) \otimes \mathbb{Q}$. So $\mathrm{Isoc}(X)$ is rigid by [10, Prop. 2.3].

By Proposition 2.2, the endomorphism ring of the unit object of $\mathrm{Isoc}(X)$ equals $\mathrm{Frac}\, W(k)$.

Let $\alpha : \mathrm{Spec}\, E \to X$ be a point with E being a perfect field. Let $K := \mathrm{Frac}\, W(E)$. Let

$$\Phi_\alpha : \mathrm{Isoc}(X) \to \{K\text{-vector spaces}\}$$

be the functor obtained by composing $\alpha^* : \mathrm{Isoc}(X) \to \mathrm{Isoc}(\mathrm{Spec}\, E)$ with the obvious fiber functor $\mathrm{Isoc}(\mathrm{Spec}\, E) \to \{K\text{-vector spaces}\}$. Then Φ_α is an exact tensor functor. Since X is connected, Proposition 2.4 implies that if $M \in \mathrm{Isoc}(X)$ and $\Phi_\alpha(M) = 0$ then $M = 0$. So Φ_α is a fiber functor. □

2.3.4 Proof of Proposition 2.4

The proposition is well known if X is a scheme of finite type over a perfect field, see [19, Cor. 2.9] or [3, §2.3.4]. The proof given in *loc. cit.* uses the following fact: if E is a field of characteristic 0 and N is a finitely generated module over $E[[t_1, \ldots, t_n]]$ which admits a connection, then N is free. The proof given below is essentially the same (but organized using Fitting ideals).

2.3.4.1 Strategy

We can assume that $X = \mathrm{Spec}\, B$, $\mathscr{X} = \mathrm{Spf}\, \widetilde{B}$, where B and $\widetilde{B}$ are as in the proof of Proposition 2.2. Choose an exact sequence of $\widetilde{B}$-modules

$$0 \to N \to \widetilde{B}^l \to M_{\mathscr{X}} \to 0.$$

For $r \in \{0, \ldots, l\}$, let $I_r \subset \widetilde{B}$ be the image of the canonical map

$$\bigwedge^{l-r} N \otimes \mathrm{Hom}(\bigwedge^{l-r} \widetilde{B}^l, \widetilde{B}) \to \widetilde{B} \tag{2.13}$$

(so the ideals $I_0 \subset \ldots I_{l-1} \subset I_l = \widetilde{B}$ are the Fitting ideals of $M_{\mathscr{X}}$). Let

$$J_r := \{u \in \widetilde{B} \mid p^j u \in I_r \text{ for some } j\}.$$

Define the ideals $J'_0 \subset \ldots J'_{l-1} \subset J'_l = B$ by $J'_r := J_r/pJ_r$. The $(\mathcal{O}_{\mathscr{X}} \otimes \mathbb{Q})$-module $M_{\mathscr{X}} \otimes \mathbb{Q}$ is locally projective of constant rank r if and only if $J'_r = B$ and $J'_i = 0$ for $i < r$. So to prove Proposition 2.4, it suffices to show that the closed subscheme $\mathrm{Spec}(B/J'_r) \subset \mathrm{Spec}\, B$ is open for each r.

We will use the notation from the proof of Proposition 2.2; in particular, we have the derivations $D_i : \widetilde{B} \to \widetilde{B}$ and the differential operators $D_i^{(l)} : \widetilde{B} \to \widetilde{B}$ and $D_i^{(l)} : B \to B$.

Lemma 2.8 *$D_i(I_r) \subset I_r$ for all i and r.*

Proof Our $M_{\mathscr{X}}$ is a module over the ring of differential operators $B[D_1, \ldots, D_m]$. For each i there exists a $\widetilde{B}$-linear operator $L_i : \widetilde{B}^l \to \widetilde{B}^l$ such that the map $f : \widetilde{B}^l \to M_{\mathscr{X}}$ satisfies the identity

$$D_i \circ f = f \circ \nabla_i \,, \text{ where } \nabla_i := D_i + L_i \,.$$

Then $\nabla_i(N) \subset N$. Think of ∇ as a (not necessarily integrable) connection on $\widetilde{B}^l$ and N. Equip $\widetilde{B}$ with the trivial connection. Then (2.13) becomes a horizontal morphism, so its image I_r is preserved by the operators D_i . □

Our goal is to show that the closed subscheme $\mathrm{Spec}(B/J'_r) \subset \mathrm{Spec}\, B$ is open for each r. Since B is Noetherian, it suffices to prove the following

Lemma 2.9 *For each $r, n \in \mathbb{N}$ there exists a closed subscheme $Z \subset \mathrm{Spec}\, B = X$ such that* $\mathrm{Spec}(B/J'_r) = (\mathrm{Fr}_X^n)^{-1}(Z)$.

Proof Lemma 2.8 implies that $D_i^{(l)}(J_r) \subset J_r$ for all i and l (recall that $D_i^{(l)} = D_i^l/l!$). So $D_i^{(l)}(J'_r) \subset J'_r$. This means that the closed subscheme $\mathrm{Spec}\, B/J'_r \subset \mathrm{Spec}\, B$ is preserved by the $\widehat{\mathbb{G}}_a^m$-action from Lemma 2.6(i). It remains to use Lemma 2.6(iii). □

2.3.5 Isocrystals as Vector Bundles

2.3.5.1 The Category $\mathbf{Bun}_{\mathbb{Q}}(\mathscr{Y})$

Let $\mathrm{FrmSch}^{\flat}$ denote the category of $\mathbb{Z}_p$-flat p-adic formal schemes. For $\mathscr{Y} \in \mathrm{FrmSch}^{\flat}$, let $\mathrm{Bun}_{\mathbb{Q}}(\mathscr{Y})$ denote the category of finitely generated locally projective $(\mathcal{O}_{\mathscr{Y}} \otimes \mathbb{Q})$-modules.

We secretly think of $\mathrm{Bun}_{\mathbb{Q}}(\mathscr{Y})$ as the category of vector bundles on the "generic fiber" $\mathscr{Y} \otimes \mathbb{Q}$, which can hopefully be defined as some kind of analytic space (maybe in the sense of R. Huber, see [16] and [20, §2.1]). This is true if $\mathscr{Y}$ is a formal scheme of finite type over $W(k)$, where k is a perfect field of characteristic p (moreover, in this case $\mathscr{Y} \otimes \mathbb{Q}$ can be understood as an analytic space in the sense of Tate or

Berkovich). But we need non-Noetherian formal schemes (e.g., the formal scheme $W(X_{\mathrm{perf}})$ from Sect. 2.3.2).

Proposition 2.5 *Suppose that $\mathcal{Y} \in \mathrm{FrmSch}^\flat$ is affine. Then any object of $\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y})$ is a direct summand of $\mathcal{O}_{\mathcal{Y}}^n \otimes \mathbb{Q}$ for some $n \in \mathbb{N}$.*

The proof is given in Sect. 2.3.7.

2.3.5.2 Flat Descent for $\mathbf{Bun}_{\mathbb{Q}}(\mathcal{Y})$

Let $C^\bullet$ be a cosimplicial category, i.e., a functor from the simplex category Δ to the 2-category of categories. Then the projective limit of this functor is denoted by $\mathrm{Tot}(C^\bullet)$. So an object of $\mathrm{Tot}(C^\bullet)$ is an object of C^0 with an isomorphism between its two images in C^1 satisfying the cocycle condition (whose formulation involves C^2).

In particular, for a simplicial object $\mathcal{Y}_\bullet$ in $\mathrm{FrmSch}^\flat$ we have the category $\mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet))$.

Proposition 2.6 *Let $f : \mathcal{Y} \to \mathcal{Z}$ be a faithfully flat*[15] *quasi-compact morphism in $\mathrm{FrmSch}^\flat$. Let $\mathcal{Y}_n$ be the fiber product (over $\mathcal{Z}$) of $n+1$ copies of $\mathcal{Y}$ (in particular, $\mathcal{Y}_0 = \mathcal{Y}$). Then the functor*

$$\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Z}) \to \mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet)) \tag{2.14}$$

is an equivalence.

Proof of Full Faithfulness If $\mathcal{Y}$ and $\mathcal{Z}$ are affine then the sequence

$$0 \to H^0(\mathcal{Z}, \mathcal{O}_{\mathcal{Z}}) \to H^0(\mathcal{Y}_0, \mathcal{O}_{\mathcal{Y}_0}) \rightrightarrows H^0(\mathcal{Y}_1, \mathcal{O}_{\mathcal{Y}_1})$$

is exact by usual flat descent. This implies that the functor (2.14) is fully faithful. Essential surjectivity will be proved in Sect. 2.3.7. □

Let us note that Proposition 2.6 was substantially generalized by A. Mathew [17].

2.3.5.3 Equivariant Objects of $\mathbf{Bun}_{\mathbb{Q}}(\mathcal{Y})$

Suppose we have a groupoid $\Gamma \rightrightarrows \mathcal{Y}$ in $\mathrm{FrmSch}^\flat$ such that the two morphisms $\Gamma \to \mathcal{Y}$ are flat and quasi-compact. Let $\mathcal{Y}_\bullet$ be its nerve (this is a simplicial formal

[15] A morphism of p-adic formal schemes $f : \mathcal{Y} \to \mathcal{Z}$ is said to be flat if for every $r \in \mathbb{N}$ it induces a flat morphism $\mathcal{Y} \otimes (\mathbb{Z}/p^r\mathbb{Z}) \to \mathcal{Z} \otimes (\mathbb{Z}/p^r\mathbb{Z})$ (if $\mathcal{Y}$ and $\mathcal{Z}$ are $\mathbb{Z}_p$-flat it suffices to check this for $n = 1$). We do *not* require f to induce flat morphisms $H^0(\mathcal{Z}', \mathcal{O}_{\mathcal{Z}'}) \to H^0(\mathcal{Y}', \mathcal{O}_{\mathcal{Y}'})$, where $\mathcal{Y}' \subset \mathcal{Y}$ and $\mathcal{Z}' \subset \mathcal{Z}$ are open affines such that $\mathcal{Y}' \subset f^{-1}(\mathcal{Z}')$.

scheme with $\mathcal{Y}_0 = \mathcal{Y}$ and $\mathcal{Y}_1 = \Gamma$). Set

$$\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}/\Gamma) := \mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet)).$$

This notation is legitimate because by Proposition 2.6, the category $\mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet))$ depends only on the quotient stack $\mathcal{Y}/\Gamma$: indeed,

$$\mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet)) = \varprojlim_{\mathcal{X}} \mathrm{Bun}_{\mathbb{Q}}(\mathcal{X}),$$

where $\mathcal{X}$ runs through the category of objects of $\mathrm{FrmSch}^\flat$ equipped with a morphism to $\mathcal{Y}/\Gamma$. Objects of $\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}/\Gamma)$ are called Γ-*equivariant objects of* $\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y})$.

Now let $\mathcal{Y} = W(X_{\mathrm{perf}})$, $\Gamma = \mathcal{G}$. Let $\mathcal{Y}_\bullet$ be the nerve of $\mathcal{G}$. By Corollary 2.3, the essential image of the functor

$$\mathrm{Isoc}(X) = \mathrm{Tot}(\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}_\bullet)) \otimes \mathbb{Q} \to \mathrm{Tot}(\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}_\bullet) \otimes \mathbb{Q}) \tag{2.15}$$

is contained in $\mathrm{Tot}(\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}_\bullet)) = \mathrm{Bun}_{\mathbb{Q}}(W(X_{\mathrm{perf}})/\mathcal{G})$. So we get a tensor functor

$$\mathrm{Isoc}(X) \to \mathrm{Bun}_{\mathbb{Q}}(W(X_{\mathrm{perf}})/\mathcal{G}). \tag{2.16}$$

Theorem 2.2 *The functor* (2.16) *is an equivalence.*

The proof will be given in Sect. 2.3.6.2.

2.3.6 Banachian Games

We will use the notation and terminology of Sect. 2.2.3. As explained in Sect. 2.2.3.7, for a p-adic formal scheme $\mathcal{Y}$ it is harmless to identify an object $\mathcal{F} \in \mathrm{QCoh}(\mathcal{Y})$ with the corresponding sheaf F_∞ on $\mathcal{Y}$, where $\mathcal{F}_\infty$ is the projective limit of the sheaves $\mathcal{F}_n = \mathcal{F}/p^n\mathcal{F}$. We will do it sometimes.

2.3.6.1 One of the Goals

Let $\mathrm{FrmSch}^\flat_{\mathrm{qcqs}} \subset \mathrm{FrmSch}^\flat$ be the full subcategory of quasi-compact quasi-separated formal schemes. We will study the presheaves of categories

$$\mathcal{Y} \mapsto \mathrm{QCoh}^\flat(\mathcal{Y}) \otimes \mathbb{Q} \quad \text{and} \quad \mathcal{Y} \mapsto \mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}) \otimes \mathbb{Q}, \quad \mathcal{Y} \in \mathrm{FrmSch}^\flat_{\mathrm{qcqs}}.$$

One of our goals is to prove that these presheaves are fpqc sheaves (see Corollary 2.5 below).

Remark 2.8 $\mathrm{QCoh}^\flat(\mathrm{Spf}\,\mathbb{Z}_p)\otimes\mathbb{Q}$ is the category of *Banachian spaces* over $\mathbb{Q}_p$. By a Banachian space over a non-Archimedean field we mean a complete topological vector space such that the topology comes from a non-Archimedean norm. (Note that the topology determines an equivalence class of norms but not a particular norm.)

Remark 2.9 If $\mathcal{Y}\in\mathrm{FrmSch}^\flat$ is affine then $\mathrm{QCoh}^\flat(\mathcal{Y})\otimes\mathbb{Q}$ is the category of Banachian modules over the Banach $\mathbb{Q}_p$-algebra $H^0(\mathcal{Y},\mathcal{O}_{\mathcal{Y}})\otimes\mathbb{Q}$. For any $\mathcal{Y}\in\mathrm{FrmSch}^\flat_{\mathrm{qcqs}}$, Corollary 2.5 proved below allows one to identify $\mathrm{QCoh}^\flat(\mathcal{Y})\otimes\mathbb{Q}$ with the category of "quasi-coherent sheaves of Banachian $(\mathcal{O}_{\mathcal{Y}}\otimes\mathbb{Q})$-modules".

Proposition 2.7 *Let $f:\mathcal{Y}'\to\mathcal{Y}$ be a faithfully flat morphism in* $\mathrm{FrmSch}^\flat_{\mathrm{qcqs}}$. *Let $M\in\mathrm{QCoh}^\flat(\mathcal{Y})\otimes\mathbb{Q}$. Then $M\in\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y})\otimes\mathbb{Q}$ if and only if $f^*M\in\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}')\otimes\mathbb{Q}$.*

Proof The "only if" statement is obvious.

Say that $M_1,M_2\in\mathrm{QCoh}^\flat(\mathcal{Y})$ are *isogenous* if they are isomorphic in $\mathrm{QCoh}^\flat(\mathcal{Y})\otimes\mathbb{Q}$. To prove the "if" statement, we have to show that if $M\in\mathrm{QCoh}^\flat(\mathcal{Y})$ and f^*M is isogenous to some $N'\in\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y}')$ then M is isogenous to some $N\in\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathcal{Y})$. We have $f^*M\supset N'\supset p^nf^*M$ for some n. The submodule $N'/p^{n+1}f^*M\subset f^*(M/p^{n+1}M)$ is finitely generated. Since $\mathcal{Y}\otimes(\mathbb{Z}/p^{n+1}\mathbb{Z})$ is a quasi-compact quasi-separated scheme, $M/p^{n+1}M$ equals the sum of its finitely generated submodules. So there exists a finitely generated submodule $L\subset M/p^{n+1}M$ such that $f^*L\supset N'/p^{n+1}f^*M$. Since $N'\supset p^nf^*M$, we have $f^*L\supset f^*(p^nM/p^{n+1}M)$, so $L\supset p^nM/p^{n+1}M$ by fully faithfulness. Now let $N\subset M$ be the preimage of L. Then $N\supset p^nM$, and $N/p^{n+1}M=L$ is finitely generated. So N is finitely generated. It is clear that N is isogeneous to M. □

Proposition 2.8 *Let $\Gamma\rightrightarrows\mathcal{Y}$ be a flat groupoid in* $\mathrm{FrmSch}^\flat_{\mathrm{qcqs}}$. *Let $\mathcal{Y}_\bullet$ be its nerve. Then the functor*

$$\mathrm{Tot}(\mathrm{QCoh}^\flat(\mathcal{Y}_\bullet))\otimes\mathbb{Q}\to\mathrm{Tot}(\mathrm{QCoh}^\flat(\mathcal{Y}_\bullet)\otimes\mathbb{Q}) \tag{2.17}$$

is an equivalence.

The proof given below is similar to that of [19, Thm. 1.9].

Proof Full faithfulness is clear. Let us prove essential surjectivity.

Let $\pi_0,\pi_1:\mathcal{Y}_1\to\mathcal{Y}_0$ be the face maps and $e:\mathcal{Y}_0\to\mathcal{Y}_1$ the degeneration map (i.e., the unit of the groupoid). An object of $\mathrm{Tot}(\mathrm{QCoh}^\flat(\mathcal{Y}_\bullet)\otimes\mathbb{Q})$ is given by a pair (M,α), where $M\in\mathrm{QCoh}^\flat(\mathcal{Y}_0)$ and $\alpha\in\mathrm{Hom}(\pi_0^*M,\pi_1^*M)\otimes\mathbb{Q}$ satisfies the cocycle condition and the condition $e^*(\alpha)=\mathrm{id}_M$. Let $n\geq 0$ be such that $p^n\alpha\in\mathrm{Hom}(\pi_0^*M,\pi_1^*M)$. We will construct an $\mathcal{O}_{\mathcal{Y}}$-submodule $M'\subset M$ containing p^nM

such that $M' \in \mathrm{QCoh}(\mathscr{Y})$ and the morphism $\alpha : \pi_0^* M \to p^{-n}\pi_1^* M$ maps $\pi_0^* M'$ to $\pi_1^* M'$. This is clearly enough.

Let us introduce some notation. For $i \le j$ we have the morphism

$$\alpha : \pi_0^*(p^i M/p^j M) \to \pi_1^*(p^{i-n} M/p^{j-n} M).$$

It induces a morphism

$$\beta : p^i M/p^j M \to \Phi(p^{i-n} M/p^{j-n} M), \quad \Phi := (\pi_0)_* \pi_1^*. \tag{2.18}$$

Now let $M' \subset M$ be the submodule containing $p^n M$ such that

$$M'/p^n M = \mathrm{Ker}(M/p^n M \xrightarrow{\beta} \Phi(p^{-n} M/M)). \tag{2.19}$$

Then $\alpha(\pi_0^* M') \subset \pi_1^* M$, and M' is the biggest submodule of M with this property. Our goal is to prove that

$$\alpha(\pi_0^* M') \subset \pi_1^* M'. \tag{2.20}$$

Let us make some general remarks. Since $\mathscr{Y}_\bullet$ is the nerve of a flat groupoid, Φ is a left exact comonad acting on the category $\{N \in \mathrm{QCoh}(\mathscr{Y}) \mid p^r N = 0 \text{ for some } r\}$. In particular, we have the coproduct $\mu : \Phi \to \Phi^2$. The morphisms (2.18) satisfy the following identity,[16] which follows from the cocycle property of α: the composite maps

$$p^i M/p^j M \xrightarrow{\beta} \Phi(p^{i-n} M/p^{j-n} M) \xrightarrow{\Phi(\beta)} \Phi^2(p^{i-2n} M/p^{j-2n} M),$$

$$p^i M/p^j M \xrightarrow{\beta} \Phi(p^{i-n} M/p^{j-n} M) \to \Phi(p^{i-2n} M/p^{j-2n} M) \xrightarrow{\mu} \Phi^2(p^{i-2n} M/p^{j-2n} M)$$

are equal to each other. In particular, the composite maps

$$M/p^{2n} M \xrightarrow{\beta} \Phi(p^{-n} M/p^n M) \xrightarrow{\Phi(\beta)} \Phi^2(p^{-2n} M/M), \tag{2.21}$$

$$M/p^{2n} M \xrightarrow{\beta} \Phi(p^{-n} M/p^n M) \to \Phi(p^{-2n} M/M) \xrightarrow{\mu} \Phi^2(p^{-2n} M/M) \tag{2.22}$$

are equal to each other.

[16] If $n = 0$ this identity means that $p^i M/p^j M$ is a comodule over Φ.

Let us now prove (2.20). By (2.19), $M'/p^{2n}M$ is equal to the preimage of $\Phi(M/p^nM)$ under the map

$$M/p^{2n}M \xrightarrow{\beta} \Phi(p^{-n}M/p^nM). \tag{2.23}$$

So the first two arrows of (2.22) kill $M'/p^{2n}M$. Therefore the composite map (2.21) kills $M'/p^{2n}M$. So the morphism (2.23) maps $M'/p^{2n}M$ to

$$\mathrm{Ker}(\Phi(M/p^nM) \xrightarrow{\Phi(\beta)} \Phi^2(p^{-n}M/M)) = \Phi(M'/p^nM).$$

This means that the composite map

$$\pi_0^*(M'/p^{2n}M) \xrightarrow{\alpha} \pi_1^*(p^{-n}M/p^nM) \to \pi_1^*(p^{-n}M/M')$$

is zero, which is equivalent to (2.20). □

Corollary 2.5 *Let $f : \mathcal{Y} \to \mathcal{Z}$ be a faithfully flat morphism in* $\mathrm{FrmSch}^{\flat}_{\mathrm{qcqs}}$. *Let $\mathcal{Y}_n$ be the fiber product (over $\mathcal{Z}$) of $n+1$ copies of $\mathcal{Y}$. Then the functors*

$$\mathrm{QCoh}^{\flat}(\mathcal{Z}) \otimes \mathbb{Q} \to \mathrm{Tot}(\mathrm{QCoh}^{\flat}(\mathcal{Y}_\bullet) \otimes \mathbb{Q}), \tag{2.24}$$

$$\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Z}) \otimes \mathbb{Q} \to \mathrm{Tot}(\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Y}_\bullet) \otimes \mathbb{Q}), \tag{2.25}$$

are equivalences.

Proof Full faithfulness is clear. To prove essential surjectivity of (2.24), combine usual flat descent with Proposition 2.8 applied to the groupoid with nerve $\mathcal{Y}_\bullet$. Essential surjectivity of (2.25) follows from Proposition 2.7 and essential surjectivity of (2.24). □

2.3.6.2 Proof of Theorem 2.2

By Corollary 2.5, the functor (2.15) is an equivalence. Theorem 2.2 follows.

Proposition 2.9 *The categories* $\mathrm{QCoh}^{\flat}(\mathcal{Y})\otimes\mathbb{Q}$ *and* $\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Y})\otimes\mathbb{Q}$ *are Karoubian for every* $\mathcal{Y} \in \mathrm{FrmSch}^{\flat}_{\mathrm{qcqs}}$.

Proof Let $M \in \mathrm{QCoh}^{\flat}(\mathcal{Y})$, $r \geq 0$ an integer, $\pi \in p^{-r} \cdot \mathrm{End}\, M$, $\pi^2 = \pi$. Set

$$M' := p^r M + (p^r \cdot \pi)(M) \subset M.$$

Then $M' \in \mathrm{QCoh}^\flat(\mathscr{Y})$, $p^r M \subset M' \subset M$. Moreover, $\pi(M') \subset M'$, so we have a direct sum decomposition

$$M' = \pi(M') \oplus (1-\pi)(M')$$

in $\mathrm{QCoh}^\flat(\mathscr{Y})$. It gives the desired direct sum composition of M in $\mathrm{QCoh}^\flat(\mathscr{Y}) \otimes \mathbb{Q}$.

If M is in $\mathrm{QCoh}^\flat_{\mathrm{fin}}(\mathscr{Y})$ then so are M' and $\pi(M')$. □

2.3.7 *Proof of Propositions 2.5 and 2.6*

Lemma 2.10 *Let $\varphi : A \to B$ be a homomorphism of $\mathbb{Z}_p$-flat p-adically complete rings. Assume that the homomorphism $A/p^r A \to B/p^r B$ induced by φ is faithfully flat for all r (or equivalently, for $r = 1$). Let $I_A \subset A \otimes \mathbb{Q}$ be an ideal, and let $I_B \subset B \otimes \mathbb{Q}$ be the ideal generated by I_A. If I_B is generated by an idempotent in $B \otimes \mathbb{Q}$ then I_A is generated by an idempotent in $A \otimes \mathbb{Q}$.*

Proof Let $e \in B \otimes \mathbb{Q}$ be the idempotent that generates I_B. Then

$$e \in \mathrm{Ker}(B \rightrightarrows B \widehat{\otimes}_A B) \otimes \mathbb{Q} = A \otimes \mathbb{Q}.$$

Let us show that $I_A = e \cdot (A \otimes \mathbb{Q})$.

Since $A \subset B$ and $(1-e) \cdot I_B = 0$, we have $(1-e) \cdot I_A = 0$, so $I_A \subset e \cdot (A \otimes \mathbb{Q})$. It remains to show that

$$I_A + (1-e) \cdot (A \otimes \mathbb{Q}) = A \otimes \mathbb{Q}. \tag{2.26}$$

Let $J := A \cap (I_A + (1-e) \cdot (A \otimes \mathbb{Q}))$. Then J generates the unit ideal in $B \otimes \mathbb{Q}$, so $p^r \in JB$ for some r. But the map

$$A/(p^{r+1}A + J) \to B/(p^{r+1}B + J \cdot B)$$

is injective by the faithful flatness assumption. So $p^r \in J + p^{r+1}A$. Therefore $p^r \in J$, which is equivalent to (2.26). □

Corollary 2.6 *Let $\varphi : A \to B$ be as in Lemma 2.10. Let $f : A^l \otimes \mathbb{Q} \to A^m \otimes \mathbb{Q}$ be an A-module homomorphism. Let $f_B : B^l \otimes \mathbb{Q} \to B^m \otimes \mathbb{Q}$ be the base change of f. If* Coker f_B *is a projective $(B \otimes \mathbb{Q})$-module then* Coker f *is a projective $(A \otimes \mathbb{Q})$-module.*

Proof For each d apply Lemma 2.10 to the ideal of $A \otimes \mathbb{Q}$ generated by all minors of f of order d. □

Lemma 2.11 *For every* $\mathcal{Y} \in \mathrm{FrmSch}^{\flat}_{\mathrm{qcqs}}$, *the canonical functor*

$$\mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Y}) \otimes \mathbb{Q} \to \{\textit{all } (\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})\textit{-modules}\} \tag{2.27}$$

is fully faithful, and its essential image contains $\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y})$.

Proof Full faithfulness is clear. Let us show that every object $M \in \mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y})$ belongs to the essential image of (2.27). If M is a direct summand of $(\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})^n$ this is true by Proposition 2.9. The general case follows by Corollary 2.5. □

Lemma 2.11 provides a fully faithful embedding $\mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}) \hookrightarrow \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Y}) \otimes \mathbb{Q}$. It commutes with pullbacks.

The following lemma implies Propositions 2.5 and 2.6.

Lemma 2.12 *Let* $\mathcal{Y}, \mathcal{Z} \in \mathrm{FrmSch}^{\flat}$ *be affine and* $f : \mathcal{Y} \to \mathcal{Z}$ *a faithfully flat morphism. Let* $M \in \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Z}) \otimes \mathbb{Q}$. *Suppose that* $f^*M \in \mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y})$. *Then* M *is a direct summand of* $\mathcal{O}^n_{\mathcal{Z}} \otimes \mathbb{Q}$ *for some* n.

Proof Let $M_0 \in \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Z})$ be a representative for M. Since $\mathcal{Z}$ is affine, M_0 is generated by finitely many global sections, so we get an exact sequence

$$0 \to N_0 \to \mathcal{O}^m_{\mathcal{Z}} \to M_0 \to 0$$

in $\mathrm{QCoh}^{\flat}(\mathcal{Z})$. Let $N \in \mathrm{QCoh}^{\flat}(\mathcal{Z}) \otimes \mathbb{Q}$ correspond to N_0. Then

$$f^*N \in \mathrm{Bun}_{\mathbb{Q}}(\mathcal{Y}) \subset \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Y}) \otimes \mathbb{Q},$$

so $N \in \mathrm{QCoh}^{\flat}_{\mathrm{fin}}(\mathcal{Z})$ by Proposition 2.7. Thus we have an exact sequence

$$(\mathcal{O}_{\mathcal{Z}} \otimes \mathbb{Q})^l \xrightarrow{\varphi} (\mathcal{O}_{\mathcal{Z}} \otimes \mathbb{Q})^m \to M \to 0$$

inducing an exact sequence $(\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})^l \to (\mathcal{O}_{\mathcal{Y}} \otimes \mathbb{Q})^m \to f^*M \to 0$. It remains to apply Corollary 2.6. □

Acknowledgments I thank A. Beilinson, P. Berthelot, B. Bhatt, A. Mathew, A. Ogus, N. Rozenblyum, and M. Temkin for valuable advice and references. The author's research was partially supported by NSF grants DMS-1303100 and DMS-2001425.

Appendix

The Isomorphism Between $W(X_{\mathrm{perf}})/\mathscr{G}$ and the Prismatization of X

The Goal

Prismatization is a certain functor $X \mapsto \mathrm{WCart}_X$ from the category of p-adic formal schemes to the category of stacks, see [6] or [12, §1]. Let us note that in [12] this functor is denoted by $X \mapsto X^{\mathbb{\Delta}}$.

The stack WCart_X can always be presented (in many different ways) as a quotient of a formal scheme by a flat groupoid. The goal of this Appendix is to deduce from [6] that if X is a Frobenius-smooth $\mathbb{F}_p$-scheme then WCart_X has a *canonical* presentation of this type; namely, one has a canonical isomorphism

$$\mathrm{WCart}_X \xrightarrow{\sim} W(X_{\mathrm{perf}})/\mathscr{G}, \tag{2.28}$$

as promised in Remark 2.1.

Prismatization of Semiperfect $\mathbb{F}_p$-Schemes

Perfect Case

According to [6, Example 3.12], for any perfect $\mathbb{F}_p$-scheme X one has a canonical isomorphism $W(X) \xrightarrow{\sim} \mathrm{WCart}_X$, where $W(X)$ is the p-adic formal scheme whose underlying topological space is that of X and whose structure sheaf is obtained by applying to $\mathcal{O}_X$ the functor of p-typical Witt vectors.

General Case

More generally, for any semiperfect $\mathbb{F}_p$-scheme X one has a canonical isomorphism $A_{\mathrm{cris}}(X) \xrightarrow{\sim} \mathrm{WCart}_X$, where $A_{\mathrm{cris}}(X)$ is the p-adic formal scheme whose underlying topological space is that of X and whose structure sheaf is obtained by applying Fontaine's functor A_{cris} to $\mathcal{O}_X$. This is [6, Lemma 6.1] in the case that X is regular semiperfect (the only one that we need) and [6, Cor. 7.18] for arbitrary semiperfect schemes.

In these statements from [6] A_{cris} is not mentioned explicitly. Instead, it is proved that if X is affine then WCart_X is the Spf of the prismatic cohomology of X. But it is known from [7, §5] that the prismatic cohomology of an $\mathbb{F}_p$-scheme identifies with its crystalline cohomology; on the other hand, if R is a semiperfect $\mathbb{F}_p$-algebra

then the crystalline cohomology of Spec R equals $A_{\mathrm{cris}}(R)$ (in particular, it lives in degree zero[17]).

The Morphism $W(X_{\mathbf{perf}}) \to \mathbf{WCart}_X$

By section "Perfect Case", For any $\mathbb{F}_p$-scheme X, the morphism

$$X_{\mathrm{perf}} \to X \tag{2.29}$$

induces a morphism

$$W(X_{\mathrm{perf}}) = \mathrm{WCart}_{X_{\mathrm{perf}}} \to \mathrm{WCart}_X \tag{2.30}$$

(we have used section "Perfect Case").

From now on, suppose that X is Frobenius-smooth in the sense of Sect. 2.2.1. Then (2.29) is a quasi-syntomic cover, so by [6, Lemma 6.3], the map (2.30) is an fpqc cover (by this we mean that for every scheme S, every morphism $S \to \mathrm{WCart}_X$ lifts to a morphism $S \to W(X_{\mathrm{perf}})$ fpqc-locally on S).

The Čech Nerve of (2.30)

By definition, $\mathscr{G}$ is the groupoid acting on $W(X_{\mathrm{perf}})$ whose Čech nerve is the simplical formal scheme $\mathcal{A}_\bullet$ from Sect. 2.2.2.3. So to finish constructing the isomorphism (2.28), it remains to identify the Čech nerve of (2.30) with $\mathcal{A}_\bullet$.

Since X is Frobenius-smooth, the morphism (2.29) is flat. So by [6, Rem. 3.5], the Čech nerve of (2.30) is $\mathrm{WCart}_{\mathcal{P}_\bullet}$, where $\mathcal{P}_\bullet$ is the Čech nerve of (2.29). By section "General Case", we have $\mathrm{WCart}_{\mathcal{P}_\bullet} = A_{\mathrm{cris}}(\mathcal{P}_\bullet)$. Finally, $A_{\mathrm{cris}}(\mathcal{P}_\bullet) =: \mathcal{A}_\bullet$.

References

1. A. Beilinson, On the crystalline period map. Camb. J. Math. **1**(1), 1–51 (2013)
2. P. Berthelot, Cohomologie cristalline des schémas de caractéristique $p > 0$. Lecture Notes in Mathematics, vol. 407 (Springer-Verlag, Berlin-New York, 1974)
3. P. Berthelot, Cohomologie rigide et cohomologie rigide à support propre, part 1, Prépublication IRMAR 96-03. https://perso.univ-rennes1.fr/pierre.berthelot/

[17] In general, the "true" objects are the *derived* crystalline cohomology of Spec R and the *derived* version of $A_{\mathrm{cris}}(R)$. They are canonically isomorphic and live in nonpositive degrees; if R is *regular* semiperfect then they live in degree zero and are equal to their classical prototypes.

4. P. Berthelot, W. Messing, Théorie de Dieudonné cristalline. III. Théorèmes d'équivalence et de pleine fidélité, in *The Grothendieck Festschrift*, vol. I. Progress in Math., vol. 86 (Birkhäuser, Boston, 1990), pp. 173–247
5. B. Bhatt, A.J. de Jong, Crystalline cohomology and de Rham cohomology. arXiv:1110.5001
6. B. Bhatt, J. Lurie, The prismatization of p-adic formal schemes. arXiv:2201.06124. Version 1
7. B. Bhatt, P. Scholze, Prisms and prismatic cohomology. arXiv:1905.08229
8. B. Bhatt, M. Morrow, P. Scholze, Topological Hochschild homology and integral p-adic Hodge theory. Publ. Math. IHES **129**, 199–310 (2019)
9. A.J. de Jong, Crystalline Dieudonné module theory via formal and rigid geometry. Publ. Math. IHES **82**, 5–96 (1995)
10. P. Deligne, Catégories tannakiennes, in *Grothendieck Festschrift*, vol. II, Progr. Math., vol. 87 (Birkhäuser, Boston, 1990), pp. 111–195
11. V. Drinfeld, On a theorem of Scholze-Weinstein. arXiv:1810.04292
12. V. Drinfeld, Prismatization. arXiv:2005.04746
13. J. Fogarty, Kähler differentials and Hilbert's fourteenth problem for finite groups. Am. J. Math. **102**(6), 1159–1175 (1980)
14. J.-M. Fontaine, Le corps des périodes p-adiques, in Périodes p-adiques, 59–111, Astérisque **223** (Soc. Math. France, Paris, 1994)
15. A. Grothendieck, Crystals and the de Rham cohomology of schemes, in Dix exposés sur la cohomologie des schémas, 306–358, Adv. Stud. Pure Math., vol. 3 (North-Holland, Amsterdam, 1968)
16. R. Huber, A generalization of formal schemes and rigid analytic varieties. Math. Z. **217**(4), 513–551 (1994)
17. A. Mathew, Faithfully flat descent of almost perfect complexes in rigid geometry. J. Pure Appl. Algebra **226**(5), 106938 (2022)
18. H. Matsumura, *Commutative Algebra*. Mathematics Lecture Note Series, vol. 56 (The Benjamin/Cummings Publishing Company, Reading, MA, 1980)
19. A. Ogus, F-isocrystals and de Rham cohomology II—convergent isocrystals. Duke Math. J. **51**, 765–850 (1984)
20. P. Scholze, J. Weinstein, Moduli of p-divisible groups. Camb. J. Math. **1**(2), 145–237 (2013)
21. A. Tyc, Differential basis, p-basis, and smoothness in characteristic $p > 0$. Proc. AMS **103**(2), 389–394 (1988)

Chapter 3
The Potts Model, the Jones Polynomial and Link Homology



Louis H. Kauffman

This paper is dedicated to C. N. Yang with special birthday greetings

Abstract In the paper we explore how the Potts model in statistical mechanics is related to the Temperley-Lieb algebra, the Jones polynomial and Khovanov homology. This exploration is made possible because the underlying combinatorics for the bracket state sum for the Jones polynomial is shared by the Potts model for planar graphs.

3.1 Introduction

Partition functions in statistical mechanics take the form

$$Z_G = \sum_{\sigma} e^{(-1/kT)E(\sigma)}$$

where σ runs over the different physical states of a system G and $E(\sigma)$ is the energy of the state σ. The probability for the system to be in the state σ is taken to be

$$prob(\sigma) = e^{(-1/kT)E(\sigma)}/Z_G.$$

Since Onsager's work, showing that the partition function of the Ising model has a phase transition, it has been a significant subject in mathematical physics to study the properties of partition functions for simply defined models based on graphs G. The underlying physical system is modeled by a graph G and the states σ

L. H. Kauffman (✉)
Department of Mathematics, Statistics and Computer Science (m/c 249), University of Illinois at Chicago, Chicago, IL, USA
e-mail: kauffman@uic.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_3

are certain discrete labelings of G. The reader can consult Baxter's book [3] for many beautiful examples. The Potts model, discussed below, is a generalization of the Ising model and it is an example of a statistical mechnanics model that is intimately related to knot theory and to the Jones polynomial [11]. While there are many connections between statistical mechanics and the Jones polynomial, there have not been so many connections of statistical mechanics with categorifications of the Jones polynomial. In [18] the author began an exploration of this relationship. The present paper extends that work. For descriptions of relationships of knot theory with physical theory see [7, 8, 26].

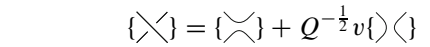
The partition function for the Potts model is given by the formula

$$P_G(Q,T) = \sum_{\sigma} e^{(J/kT)E(\sigma)} = \sum_{\sigma} e^{KE(\sigma)}$$

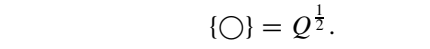
where σ is an assignment of one element of the set $\{1, 2, \cdots, Q\}$ to each node of the graph G, and $E(\sigma)$ denotes the number of edges of a graph G whose end-nodes receive the same assignment from σ. In this model, σ is regarded as a *physical state* of the Potts system and $E(\sigma)$ is the *energy* of this state. Here $K = J\frac{1}{kT}$ where J is plus or minus one (ferromagnetic and anti-ferromagnetic cases), k is Boltzmann's constant and T is the temperature. The Potts partition function can be expressed in terms of the dichromatic polynomial of the graph G. Letting

$$v = e^K - 1,$$

it is shown in [15, 16] that the dichromatic polynomial $Z[G](v, Q)$ for a plane graph can be expressed in terms of a bracket state summation of the form



with



Then the Potts partition function is given by the formula

$$P_G = Q^{N/2}\{K(G)\}$$

where $K(G)$ is an alternating link diagram associated with the plane graph G so that the projection of $K(G)$ to the plane is a medial diagram for the graph. This translation of the Potts model in terms of a bracket expansion makes it possible to examine how the Khovanov homology of the states of the bracket is related to the evaluation of the Potts partition function.

There are seven sections in this paper beyond the introduction. In Sect. 3.2 we review the bracket model of the Jones polynomial, its state summation and its relationship with the Temperley-Lieb algebra. In Sect. 3.3 we describe the

categorical framework for Khovanov homology. We will use this same framework for discussing the Potts model. In Sect. 3.4 we discuss the dichromatic (Tutte) polynomial for graphs and how it can be used to express the partition function of the Potts model. We express the dichromatic polynomial using the deletion contraction formula in the form shown below

$$Z[\bullet\!\!-\!\!-\!\!\bullet] = Z[\bullet\!\!\times\!\!\bullet] = Z[\bullet\!\supset\!\subset\!\bullet] + vZ[\bullet\!\frown\!\frown\!\bullet].$$

Here and edge of the graph is decorated with a knot theoretic crossing and the deletion and contraction edges are denoted as shown where the contraction edge has a decoration consisting in two arcs that do not touch the edge, while the deletion edge is decorated with arcs that intersect this edge. Making all possible such decorations of a plane graph gives a set of states for expressing the partition function. These decorations allow us to intermediate fluently between the medial graph/knot theory point of view and the graph theoretic, sum over subgraphs point of view. This section proves the relationship alluded to above between the dichromatic polynomial and the bracket expansion and the consequent formulation of the Potts model in terms of Temperley-Lieb algebra from this point of view. In the direction of the statistical mechnanics we show how the dichromatic polynomial yields the partition function for the Potts model as well.

In Sect. 3.5 we review the definition of Khovanov homology and observe, in parallel with [25], that it is very natural to begin by defining the Khovanov chain complex via enhanced states of the bracket polynomial model for the Jones polynomial. In fact, we begin with Khovanov's rewrite of the bracket state sum in the form

$$\langle \times \rangle = \langle \asymp \rangle - q \langle \rangle \langle \rangle$$

with $\langle \bigcirc \rangle = (q + q^{-1})$. We rewrite the state sum formula for this version of the bracket in terms of enhanced states (each loop is labeled $+1$ or -1 corresponding to q^{+1} and q^{-1} respectively) and show how, collecting terms, the formula for state sum has the form of a graded sum of Euler characteristics

$$\langle K \rangle = \sum_j q^j \sum_i (-1)^i \dim(C^{ij}) = \sum_j q^j \chi(C^{\bullet\, j}) = \sum_j q^j \chi(\mathcal{H}^{\bullet\, j})$$

where C^{ij} is a module with basis the set of enhanced states s with i smoothings of type B (see Sect. 3.2 for definitions), $\mathcal{H}^{\bullet\, j}$ is its homology and $j = j(s)$ where

$$j(s) = n_B(s) + \lambda(s)$$

where $n_B(s)$ denotes the number of B-smoothings in s and $\lambda(s)$ denotes the number of loops with positive label minus the number of loops with negative label in s. This formula suggests that there should be differentials $\partial : C^{i,j} \longrightarrow C^{i+1,j}$ so that j

is preserved under the differential. We show that the restriction $j(s) = j(\partial(s))$ uniquely determines the differential in the complex, and how this leads to the Frobenius algebra structure that Khovanov used to define the differential. The method by which we arrive at the graded Euler characteristic is particularly useful for our subsequent discussion. One should look for subcomplexes on which Euler characteristics can be defined, and one should attempt to shape the state summation so that these characteristics appear in the state sum.

In Sect. 3.6 we recall the definition of the dichromatic polynomial and the fact [15] that the dichromatic polynomial for a planar graph G can be expressed as a special bracket state summation on an associated alternating knot $K(G)$. We recall that, for special values of its two parameters, the Potts model can be expressed as a dichromatic polynomial. This sets the stage for seeing what is the role of Khovanov homology in the evaluations of the Potts model and of the dichromatic polynomial. In both cases (dichromatic polynomial in general and the Potts model in particular) one finds that the state summation does not so easily rearrange itself as a sum over Euler characteristics, as we have explained in Sect. 3.2. The states in the bracket summation for the dichromatic polynomial of $K(G)$ are the same as the states for the bracket polynomial of $K(G)$, so the Khovanov chain complex is present at the level of the states for the partition function. We use this relationship to derive a formula for the Euler characteristics of the homologies in the Khovanov complex in terms of the coefficients of the dichromatic polynomial. We show that

$$\chi(KH^{\bullet,\lambda}(K(G)) = \sum_{2i+2j-k-N/2=\lambda, k\leq 2i+j} (-1)^j C_k^{2i+j} \pi_{ij}$$

where $KH^{\bullet,\lambda}(K(G)$ denotes the Khovanov homology of the medial knot $K(G)$ associated to the planar graph G, π_{ij} are the coefficients of the Potts partition function for the graph G with N nodes, and C_r^n is the binomial coefficient $n!/(r!((n-r)!)$ This shows just how the dichromatic polynomial is related to the Khovanov homology. Since the dichormatic polynomial has the form

$$Z[G](Q, v) = \sum_{i,j} Q^i v^j \pi_{ij}$$

and one obtains the Potts model by taking Q to be the number of spins in the model and $v = e^K - 1$ as explained above, this shows also the relationship with the Potts model.

We clarify this relationship by using a two-variable bracket expansion, the *ρ-bracket*, that reduces to the Khovanov version of the bracket as a function of q when ρ is equal to one.S

$$[\diagup\!\diagdown] = [\diagdown\!\diagup] - q\rho[\rangle\langle]$$

with

$$[\bigcirc] = q + q^{-1}.$$

We can regard this expansion as an intermediary between the Potts model (dichromatic polynomial) and the topological bracket. The ρ-bracket can be rewritten in the following form:

$$[K] = \sum_{i,j} (-\rho)^i q^j \dim(C^{ij})$$

$$= \sum_j q^j \sum_i (-\rho)^i \dim(C^{ij}) = \sum_j q^j \chi_\rho(C^{\bullet j}),$$

where we define C^{ij} to be the linear span of the set of enhanced states with $n_B(s) = i$ and $j(s) = j$. Then the number of such states is the dimension $\dim(C^{ij})$. We have expressed this ρ-bracket expansion in terms of generalized Euler characteristics of the complexes $C^{\bullet j}$:

$$\chi_\rho(C^{\bullet j}) = \sum_i (-\rho)^i \dim(C^{ij}).$$

These generalized Euler characteristics become classical Euler characteristics when $\rho = 1$, and, in that case, are the same as the Euler characteristic of the homology. With ρ not equal to 1, we do not have direct access to the homology. In that case the polynomial is expressed in terms of ranks of chain modules, but not in terms of ranks of corresponding homology groups.

We analyze those cases of the Potts model where $\rho = 1$ (so that Euler characteristics of Khovanov homology appear as coefficients in the Potts partition function) and we find that at criticality this requires $Q = 4$ and $e^K = -1$ ($K = J/kT$ as in the second paragraph of this introduction), hence an imaginary value of the temperature. When we simply require that $\rho = 1$, not necessarily at criticality, then we find that for $Q = 2$ we have $e^K = \pm i$. For Q = 3, we have $e^K = \frac{-1 \pm \sqrt{3}i}{2}$. For $Q = 4$ we have $e^K = -1$. For $Q > 4$ it is easy to verify that e^K is real and negative. Thus in all cases of $\rho = 1$ we find that the Potts model has complex temperature values. Further work is called for to see how the evaluations of the Potts model at complex values influence its behaviour for real temperature values in relation to the Khovanov homology. Such relationships between complex and real temeperature evaluations already known for the accumlation of the zeroes of the partition function (the Lee-Yang zeroes [20]). In Sect. 3.8 we return to the matter of imaginary temperature.

In Sect. 3.7 we discuss Stosic's categorification of the dichromatic polynomial. Stosic's categorification involves using a differential motivated directly by the graphical structure, and gives a homology theory distinct from Khovanov homology.

We examine Stosic's categorification in relation to the Potts model and again show that it will work (in the sense that the coefficients of partition function are Euler characteristics of the homology) when the temperature is imaginary. Specifically, we find this behaviour for $K = i\pi + ln(q + q^2 + \cdots + q^n)$, and again the challenge is to find out what is the influence of this graph homology on the Potts model at real temperatures. The results of this section do not require planarity of the graph G. More generally, we see that in the case of this model, the differential in the homology is related to the combinatorial structure of the physical model. In that model a given state has graphical regions of constant spin (calling the discrete assignments $\{1, 2, \cdots, Q\}$ the *spins*). We interpret the partial differentials in this model as taking two such regions and making them into a single region, by adding an edge that joins them, and re-assigning a spin to the new region that is a combination of the spins of the two formerly separate regions. This form of partial differential is a way to think about relating the different states of the physical system. It is however, not directly related to classical physical process since it invokes a global change in the spin configuration of the model. (One might consider such global transitions in quantum processes.) The Stosic homology measures such global changes, and it is probably the non-classical physicality of this patterning that makes it manifest at imaginary temperatures. The direct physical transitions from state to state that are relevant to the classical physics of the model may indeed involve changes of the form of the regions of constant spin, but this will happen by a local change so that two disjoint regions with the same spin join into a single region, or a single region of constant spin bifurcates into two regions with this spin. These are the sort of local-classical-time transitions that one works with in a statistical mechanics model. Such transitions are part of the larger and more global transitions that are described by the differentials in our interpretation of the Stosic homology for the Potts model. As in the previous section, we obtain a formula that expresses the Euler characteristics of the Stosic homology in terms of the dichromatic coefficients: Let $[j, k]$ be the multinomial coefficient for $(1 + q + \cdots + q^n)^j = \sum_{k=0}^{nj} [j, k]q^k$. Then

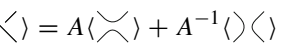
$$\chi(H^{\bullet\lambda}(G)) = \sum_{\lambda = ni+k : i,j,k} (-1)^j [j, k]\pi_{ij}$$

where $H^{\bullet\lambda}(G)$ denotes Stosic homology and π_{ij} are the dichromatic coeffients as above.

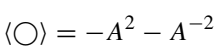
In Sect. 3.8 we formulate a version of Wick rotation for the Potts model so that it is seen (for imaginary temperature) as a quantum amplitude. In this way, for those cases where the temperature is pure imaginary, we obtain a quantum physical interpretation of the Potts model and hence a relationship of Khovanov homology with quantum amplitudes at the special values discussed in Sect. 3.6.

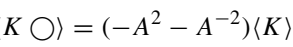
3.2 Bracket Polynomial and Jones Polynomial

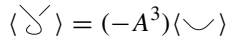
The bracket polynomial [14] model for the Jones polynomial [11–13, 26] is usually described by the expansion

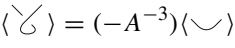


Here the small diagrams indicate parts of otherwise identical larger knot or link diagrams. The two types of smoothing (local diagram with no crossing) in this formula are said to be of type A (A above) and type B (A^{-1} above).









One uses these equations to normalize the invariant and make a model of the Jones polynomial. In an oriented link diagram we use the conventions shown in Fig. 3.2 to designate a sign of $+1$ or -1 to an oriented crossing. In the normalized version we define

$$f_K(A) = (-A^3)^{-wr(K)}\langle K\rangle / \langle \bigcirc \rangle$$

where the writhe $wr(K)$ is the sum of the oriented crossing signs for a choice of orientation of the link K. The writne is invariant under the second and third Reidemeister moves. The normalized version is invariant under all three Reidemeister moves as shown in Fig. 3.1. In the 1920s Alexander and Briggs showed that two links in three dimensional space are equivalent by continuous deformation through embeddings if and only if corresponding projection diagrams are equivalent by these graphical moves. Reidemeister in the first book on the theory of knots [21] made this theorem about the moves the foundation for a combinatorial approach to the theory of knots and links.

Since we shall not use oriented links in this paper, we refer the reader to [14] for the details about the writhe. One then has that $f_K(A)$ is invariant under the Reidemeister moves (again see [14]) and the original Jones polynomial $V_K(t)$ is given by the formula

$$V_K(t) = f_K(t^{-1/4}).$$

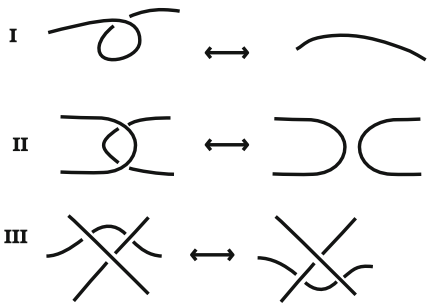


Fig. 3.1 Reidemeister moves

The Jones polynomial has been of great interest since its discovery in 1983 due to its relationships with statistical mechanics, due to its ability to often detect the difference between a knot and its mirror image and due to the many open problems and relationships of this invariant with other aspects of low dimensional topology.

The State Summation In order to obtain a closed formula for the bracket, we now describe it as a state summation. Let K be any unoriented link diagram. Define a *state*, S, of K to be the collection of planar loops resulting from a choice of smoothing for each crossing of K. There are two choices (A and B) for smoothing a given crossing, and thus there are $2^{c(K)}$ states of a diagram with $c(K)$ crossings. In a state we label each smoothing with A or A^{-1} according to the convention indicated by the expansion formula for the bracket. These labels are the *vertex weights* of the state. There are two evaluations related to a state. The first is the product of the vertex weights, denoted $\langle K|S\rangle$. The second is the number of loops in the state S, denoted $||S||$.

Define the *state summation*, $\langle K\rangle$, by the formula

$$\langle K\rangle = \sum_S < K|S > \delta^{||S||}$$

where $\delta = -A^2 - A^{-2}$. This is the state expansion of the bracket. It is possible to rewrite this expansion in other ways. For our purposes in this paper it is more convenient to think of the loop evaluation as a sum of *two* loop evaluations, one giving $-A^2$ and one giving $-A^{-2}$. This can be accomplished by letting each state curve carry an extra label of $+1$ or -1. We describe these *enhanced states* below. But before we do this, it will be useful for the reader to examine Fig. 3.7. In Fig. 3.7 we show all the states for the right-handed trefoil knot, labelling the sites with A or B where B denotes a smoothing that would receive A^{-1} in the state expansion.

Note that in the state enumeration in Fig. 3.7 we have organized the states in tiers so that the state that has only A-smoothings is at the left and the state that has only B-smoothings is at the right (Fig. 3.2).

Bracket, Braids and Temperley-Lieb Algebra With the help of Fig. 3.3 we recall the Artin braid group B_n on n strands, generated by the elementary braids

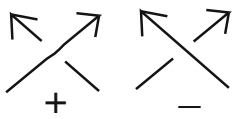


Fig. 3.2 Crossing signs

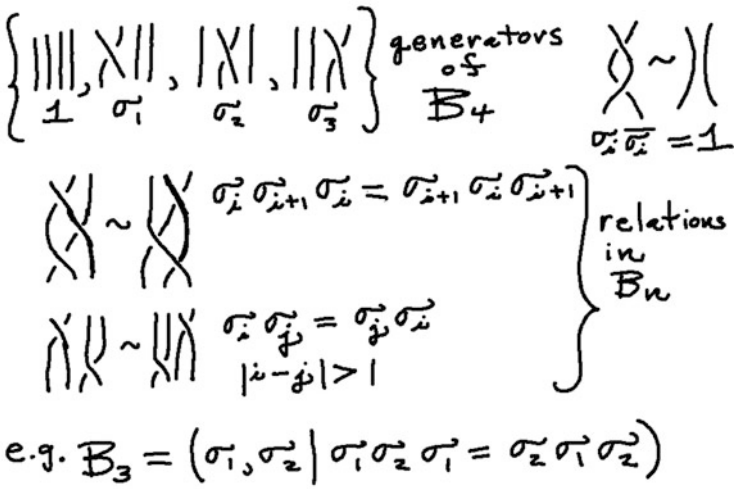


Fig. 3.3 Artin braid group

$\{\sigma_1, \cdots, \sigma_{n-1}\}$ and having presentation

$$B_n = (\sigma_1, \cdots, \sigma_{n-1} | \sigma_i \sigma_{i+1} \sigma_i = \sigma_{i+1} \sigma_i \sigma_{i+1}, \sigma_i \sigma_j = \sigma_j \sigma_i, |i-j| > 1).$$

The pattern of the bracket gives a way to produce a representation of the braid group to the Temperley-Lieb algebra TL_n. This is illustrated in Figs. 3.4, 3.5 and 3.6. In Fig. 3.4 we show the relations in the Temperley-Lieb algebra and diagrammatic interpretations for the generators and the relations. This algebra, TL_n, was originally defined in terms of the generators $U_i, i = 1, \cdots, n-1$ and the relations

1. $U_i U_j = U_j U_i$ for $|i-j| > 1$,
2. $U_i U_{i+1} U_i = U_i$ and $U_{i+1} U_i U_{i+1} = U_{i+1}$,
3. $U_i^2 = \delta U_i$.

For the sake of the bracket polynomial we take $\delta = A^2 - A^{-2}$ and write the algebra additively as a module over the ring $Z[A, A^{-1}]$.

In Fig. 3.5 we indicate how the algebra with these relations can be regarded as a faithful description of the composition of *planar connections* between two rows of n points. Call the multiplicative monoid of connection arcs $C(n)$. It consists

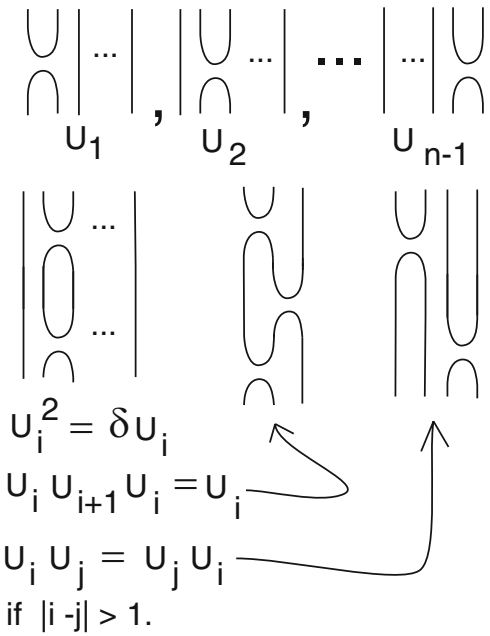


Fig. 3.4 Diagrammatic Temperley-Lieb algebra

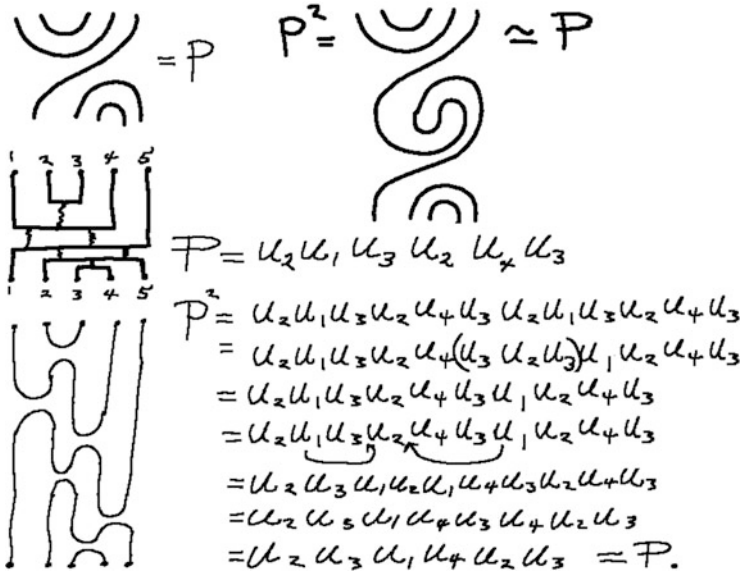


Fig. 3.5 Temperley-Lieb connection algebra

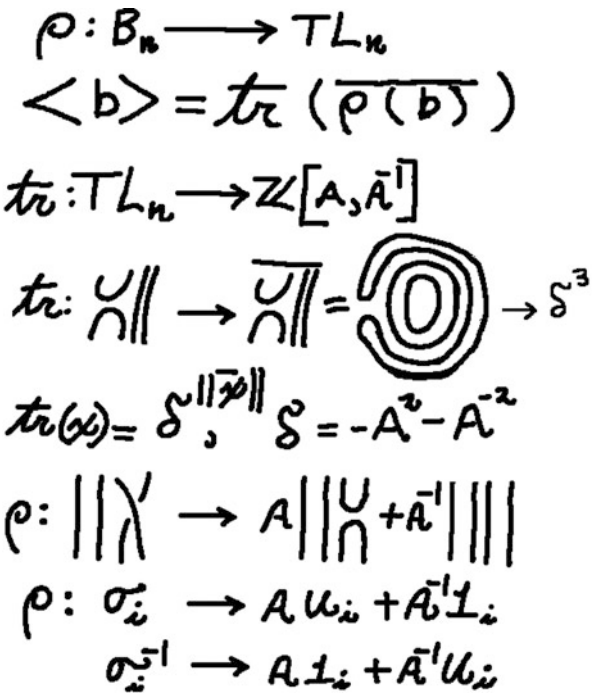


Fig. 3.6 Representation of the braid group to the Temperley-Lieb algebra

in non-crossing collections of planar arcs connecting the $2n$ points in a $1-1$ correspondence. Two elements in the connection monoid are *equal* if the point to point connections are the same. The positions of the arcs that represent the connections are irrelevant. Since we use non-crossing arcs, the arcs can be moved by planar topological isotopy. This explains why many identities related to this planar algebra appear topological. In the figure we illustrate an elememt P and its square P^2 obtained by attaching the bottom row of one factor to the top row of the other factor (this is the general form of multiplication in braids and in the Temperley-Lieb algebra). It is clear at the topological and at the connection combinatorics level that $P^2 = P$. We also illustrate a rectilinear procedure for converting any element in the connection monoid to a product of diagrammatic generators of the Temperley-Lieb algebra. This method of conversion can be used [17] as a canonical form for products of the generators U_i and can be used to prove that the free additive algebra generated by the connection monoid is exactly the Temperley-Lieb algebra given by generators and relations. The figure also shows a derivation of $P^2 = P$ done entirely inside the Temperley-Lieb algebra. As this example shows, the connection algebra viewpoint provides an extra freedom in handling combinatorial problems about the algebra.

Figure 3.6 illustrates representation of the braid group to the Temperley-Lieb algebra, $\rho : B_n \longrightarrow TL_n$, via these combinatorial methods. The representation given here is equivalent to the Jones representation [16]. The figure sketches the

definition of the representation and the trace $Tr : TL_n \longrightarrow Z[A, A^{-1}]$ given by $Tr(x) = \delta^{||\overline{x}||}$ where $\overline{x}$ denotes the closure of the Temperley Lieb element obtained by attaching the top row to the bottom row by an identity collection of strands. This has the same form as standard braid closure and we shall write $\overline{b}$ for the corresponding closure of a braid b. With that, we can state the fundamental formula

$$< \overline{b} >= Tr(\rho(b))$$

that expresses the bracket polynomial (and hence the Jones polynomial) in terms of the trace on the Temperley-Lieb algebra. Again, this formulation is equivalent to that of Jones, but the trace originally used by Jones came from normal forms for products of elements U_i. Remarkably, these normal forms are the ones that arise in our combinatorics from the rectilinear version of the connection structures as shown in Fig. 3.5.

Changing Variables Letting $c(K)$ denote the number of crossings in the diagram K, if we replace $\langle K \rangle$ by $A^{-c(K)}\langle K \rangle$, and then replace A^2 by $-q^{-1}$, the bracket is then rewritten in the following form:

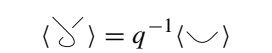


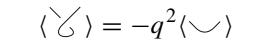
with $\langle \bigcirc \rangle = (q + q^{-1})$. It is useful to use this form of the bracket state sum for the sake of the grading in the Khovanov homology (to be described below). We shall continue to refer to the smoothings labeled q (or A^{-1} in the original bracket formulation) as *B-smoothings*.

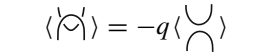
We catalog here the resulting behaviour of this modified bracket under the Reidemeister moves of Fig. 3.1.

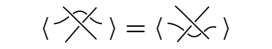
$$\langle \bigcirc \rangle = q + q^{-1}$$

$$\langle K \bigcirc \rangle = (q + q^{-1})\langle K \rangle$$









It follows that if we define

$$J_K = (-1)^{n_-} q^{n_+ - 2n_-} \langle K \rangle,$$

where n_- denotes the number of negative crossings in K and n_+ denotes the number of positive crossings in K, then J_K is invariant under all three Reidemeister moves. Thus J_K is a version of the Jones polynomial taking the value $q + q^{-1}$ on an unknotted circle.

Using Enhanced States We now use the convention of *enhanced states* where an enhanced state has a label of 1 or -1 on each of its component loops. We then regard the value of the loop $q + q^{-1}$ as the sum of the value of a circle labeled with a 1 (the value is q) added to the value of a circle labeled with an -1 (the value is q^{-1}). We can choose the less neutral labels of $+1$ and x so that

$$q^{+1} \Longleftrightarrow +1 \Longleftrightarrow 1$$

and

$$q^{-1} \Longleftrightarrow -1 \Longleftrightarrow x,$$

since an algebra involving 1 and x naturally appears later in relation to Khovanov homology. It does no harm to take this form of labeling from the beginning. The use of enhanced states for formulating Khovanov homology was pointed out by Oleg Viro in [25].

Consider the form of the expansion of this version of the bracket polynonmial in enhanced states. We have the formula as a sum over enhanced states s :

$$\langle K \rangle = \sum_s (-1)^{i(s)} q^{j(s)}$$

where $i(s)$ is the number of B-type smoothings in s and $j(s) = i(s) + \lambda(s)$, with $\lambda(s)$ the number of loops labeled 1 minus the number of loops labeled -1 in the enhanced state s.

One advantage of the expression of the bracket polynomial via enhanced states is that it is now a sum of monomials. We shall make use of this property throughout the rest of the paper.

3.3 Khovanov Homology and the Cube Category

We are going to make a chain complex from the states of the bracket polynomial so that the homology of this chain complex is a knot invariant. One way to see how such a homology theory arises is to step back and note that the collection of states

for a diagram K forms a category in the shape of a cube. A functor from such a category to a category of modules gives rise to a homology theory in a natural way, as we explain below.

Examine Figs. 3.7 and 3.8. In Fig. 3.7 we show all the standard bracket states for the trefoil knot with arrows between them whenever the state at the output of the arrow is obtained from the state at the input of the arrow by a single smoothing of a site of type A to a site of type B. The abstract structure of this collection of states is a category with objects of the form $\langle ABA \rangle$ where this symbol denotes one of the states in the state diagram of Fig. 3.7. In Fig. 3.8 we illustrate this cube category (the states are arranged in the form of a cube) by replacing the states in Fig. 3.7 by symbols $\langle XYZ \rangle$ where each literal is either an A or a B. A typical generating morphism in the 3-cube category is

$$\langle ABA \rangle \longrightarrow \langle BBA \rangle.$$

We formalize this way of looking at the bracket states as follows. Let $\mathcal{S}(K)$ denote a category associated with the states of the bracket for a diagram K whose objects are the states, with sites labeled A and B as in Fig. 3.7. A morphism in this category is an arrow from a state with a given number of A's to a state with fewer A's.

Let $\mathcal{D}^n = \{A, B\}^n$ be the n-cube category whose objects are the n-sequences from the set $\{A, B\}$ and whose morphisms are arrows from sequences with greater numbers of A's to sequences with fewer numbers of A's. Thus $\mathcal{D}^n$ is equivalent to

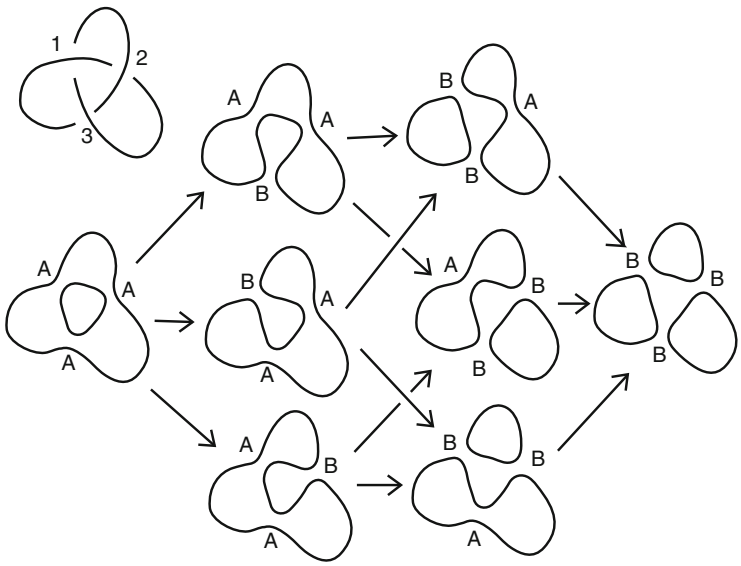


Fig. 3.7 Bracket states and Khovanov category

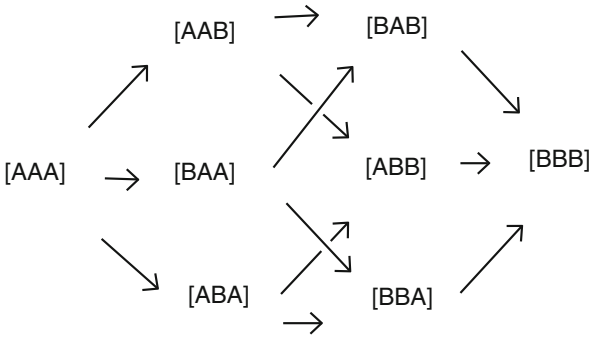


Fig. 3.8 Cube category

the poset category of subsets of $\{1, 2, \cdots n\}$. We make a functor $\mathcal{R} : \mathcal{D}^n \longrightarrow \mathcal{S}(K)$ for a diagram K with n crossings as follows. We map sequences in the cube category to bracket states by choosing to label the crossings of the diagram K from the set $\{1, 2, \cdots n\}$, and letting this functor take abstract A's and B's in the cube category to smoothings at those crossings of type A or type B. Thus each sequence in the cube category is associated with a unique state of K when K has n crossings. By the same token, we define a functor $\mathcal{S} : \mathcal{S}(K) \longrightarrow \mathcal{D}^n$ by associating a sequence to each state and morphisms between sequences corresponding to the state smoothings. With these conventions, the two compositions of these morphisms are the identity maps on their respective categories.

Let $\mathcal{M}$ be a category of modules, and let $\mathcal{F} : \mathcal{D}^n \longrightarrow \mathcal{M}$ be a functor. We point out that there is a natural structure of chain complex associated with the functor $\mathcal{F}$. First note that each object in $\mathcal{D}^n$ has the form

$$X = \langle X_0 \cdots X_{n-1} \rangle$$

where each X_i equals either A or B and we have morphisms

$$d_i : \langle X_0 \cdots X_i = A \cdots X_{n-1} \rangle \longrightarrow \langle X_0 \cdots \overline{X}_i = B \cdots X_{n-1} \rangle$$

whenever $X_i = A$ and (by definition) $\overline{X}_i = B$. We then define

$$\partial_i = \mathcal{F}(d_i) : \mathcal{F}\langle X_0 \cdots X_i \cdots X_{n-1} \rangle \longrightarrow \mathcal{F}\langle X_0 \cdots \overline{X}_i \cdots X_{n-1} \rangle$$

whenever d_i is defined. We then define the chain complex$C$ by

$$C^k = \bigoplus_X \mathcal{F}\langle X_0 \cdots X_{n-1} \rangle$$

where each sequence $X = \langle X_0 \cdots X_{n-1} \rangle$ has k B's. With this we define

$$\partial : C^k \longrightarrow C^{k+1}$$

by the formula

$$\partial x = \Sigma_{i=0}^{n-1} (-1)^{c(X,i)} \partial_i(x)$$

for $x \in CX = C\langle X_0 \cdots X_{n-1} \rangle$ and $c(X, i)$ denotes the number of A's in the sequence X that precede X_i.

We want $\partial^2 = 0$ and it is easy to see that this is equivalent to the condition that $\partial_i \partial_j = \partial_j \partial_i$ for $i \neq j$ whenever these maps and compositions are defined. We can assume that the functor $\mathcal{F}$ has this property, or we can build it in axiomatically by adding the corresponding relations to the cube category in the form $d_i d_j = d_j d_i$ for $i \neq j$ whenever these maps are defined. In the section on Khovnov homology we shall see that there is a natural way to define the maps in the state category so that this condition holds. Once we axiomatize this commutation relation at the level of the state category or the cube category, then the functor $\mathcal{F}$ will induce a chain complex and homology as above.

In this way, we see that a suitable functor from the cube category to a module category allows us to define homology that is modeled on the "shape" of the cube. The set of bracket states forms a natural functorial image of the cube category, and that makes it possible to define the Khovanov chain complex. In terms of the bracket states, we will map each state loop to a specific module V, and each state to a tensor power of V to the number of loops in the state. The details of this construction are in the section on Khovanov homology.

We use a specific construction for the Khovanov complex that is directly related to the enhanced states for the bracket polynomial. In this construction we will use the enhanced states, regarding each loop as labeled with either 1 or x for a module $V = k[x]/(x^2)$ associated with the loop (where $k = Z/2Z$ or $k = Z$.) Thus the two labelings of the loop will correspond to the two generators of the module V. A state that is a collection of loops will be associated with $V^{\otimes r}$ where r is the number of loops in the state. In this way we will obtain a functor from the state category to a module category, and at the same time it will happen that any single enhanced state will correspond to a generator of the chain complex.

Before describing Khovanov homology, we describe the Potts model and a graphical category for handling the Potts model.

3.4 The Dichromatic Polynomial and the Potts Model

Let G be a finite graph. The *dichromatic polynomial* of G is defined by the following recursion:

$$Z[G](v, Q) = Z[G'](v, Q) + vZ[G''](v, Q)$$

$$Z[\bullet \sqcup G] = QZ[G].$$

where G' is the result of deleting an edge from G, while G'' is the result of contracting that same edge so that its end-nodes have been collapsed to a single node. In the second equation, the dot $\bullet$ represents a graph with one node and no edges, and $\bullet \sqcup G$ represents the disjoint union of the single-node graph with the graph G.

The partition function $P_G(Q, T)$ for the Q-state Potts model in statistical mechanics on a graph G is equal to the dichromatic polynomial when

$$v = e^{J\frac{1}{kT}} - 1$$

where T is the temperature for the model and k is Boltzmann's constant. Here $J = \pm 1$ according as we work with the ferromagnetic or anti-ferromagnetic models (see [3] Chapter 12). This result was discovered by Kastelyn and Fortuin [6] and utilized by Temperley and Lieb [3, 23, 24]. For completeness, we recall the proof.

For simplicity we denote

$$K = J\frac{1}{kT}$$

so that

$$v = e^K - 1.$$

We will show that, when $Q = N$, a positive integer, then

$$P_G(Q, T) = Z[G](e^K - 1, Q).$$

The partition function is given by the formula

$$P_G(Q, T) = \sum_{\sigma} e^{K E(\sigma)}$$

where σ is an assignment of one element of the set $\{1, 2, \cdots, Q\}$ to each node of the graph G, and $E(\sigma)$ denotes the number of edges of the graph G whose end-nodes receive the same assignment from σ. In this model, σ is regarded as a *physical state* of the Potts system and $E(\sigma)$ is the *energy* of this state. Thus we can write $E(\sigma) = \sum_{<i,j>} \delta(\sigma_i, \sigma_j)$ where $< i, j >$ runs over all the edges of the graph G

with nodes denoted by i and j. Then, with $v = e^K - 1$,

$$P_G(Q,T) = \sum_{\sigma} e^{K \sum_{<i,j>} \delta(\sigma_i,\sigma_j)}$$

$$= \sum_{\sigma} \prod_{<i,j>} e^{K\delta(\sigma_i,\sigma_j)}$$

$$= \sum_{\sigma} \prod_{<i,j>} (1 + v\delta(\sigma_i,\sigma_j))$$

The contraction/deletion formula of the dichromatic polynomial now follows immediately by analysis of the terms $1 + v\delta(\sigma_i, \sigma_j)$ for each edge: If an edge had distinct spins at its nodes, it will contribute 1 to the product. If an edge has equal spins at its node, it will contribute $e^K = 1 + v$ to the product, and this will be, in the recursion, a joint contribution of the deleted graph and the contracted graph.

Thus we have the graph theoretic formulation of the Potts model in terms of the dichromatic polynomial:

$$P_G(Q,T) = Z[G](v, Q).$$

The dichromatic polynomial is equivalent to the *Tutte polynomial* of the graph [4] where there are only unit coefficients in the deletion/contraction recursion, and the variables appear in terms of loop and isthmus edges in the resulting graphs. Here, we will use the dichromatic version of the Tutte polynomial.

If the graph G is planar, it is convenient to rewrite the recursion as shown below. In this rewrite we have labeled edges of the graph prior to the recursion with a crossing and then the two possibilities for the recursion are labeled with *smoothings* of this crossing, where the endpoints are connected so that the two arcs do not cross one another. This is notational terminology from knot theory and we shall explain it in more detail below. The smoothing that intersects the edge corresponds to deletion of the edge. The smoothing the does not intersect the edge corresponds to the contraction of the edge.

$$Z[\bullet\!\!-\!\!\bullet] = Z[\bullet\!\!\times\!\!\bullet] = Z[\bullet\!\supset\!\subset\!\bullet] + vZ[\bullet\!\frown\!\bullet].$$

With this notation for the deletion and the contraction we can illustrate a *state summation* for the dichromatic polynomial as illustrated in Fig. 3.9. In this figure we show the calculation of the dichromatic polynomial for the graph $\bullet\!\!-\!\!\bullet$. We create a graph whose nodes are copies of the original graph G decorated with either a contraction or a deletion smoothing at each edge. Two nodes are connected by an arrow if one differs from the other by a change of one deletion to one contraction.

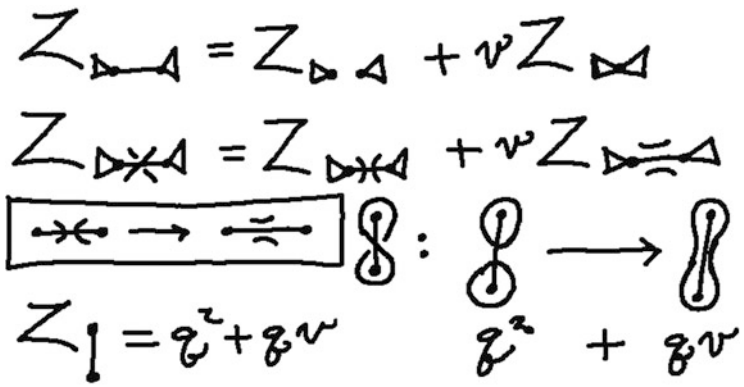
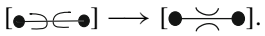


Fig. 3.9 Deletion contraction reformulation

Thus arrows of the sort shown below generate these hypergamous.



We call the hypergraph of a given graph the *category* of that graph and denote it by $Cat(G)$. In fact $Cat(G)$ does generate a category in the sense of category theory. Recall that any directed graph generates a category by adding an identity arrow to each node and defining the full set of arrows in the category to be in correspondence with the directed paths in the graph. In this sense we regard $Cat(G)$ as a category. As we illustrate in Figs. 3.10, 3.11, and 3.12, the *form* of this category is relevant to organizing the structure of the dichromatic polynomial. One would like to see if there are ways to measure the category that give new information about the graph and its dichromatic polynomial. How does this shed light on the partition function of the Potts model?

In Fig. 3.10 we show how this edge notation extends and connects with knot and link diagrams when the graphs are planar. We shall say *knot diagram* to refer to any four valent plane graph whose nodes are taken in the form of crossings as drawn in this figure. A crossing is a four valent node with a small arc deleted so that the local form of the crossing consists of one continuous arc and two cut arcs with cyclic order in the plane of the form (continuous, cut, continuous, cut). The drawing has the appearance of one arc weaving over another arc where the under crossing arc consists in the two arcs with the cut between them. Placing a crossing at each edge of the graph, we connect them consecutively around each face of the graph and obtain a link diagram $K(G)$, the *medial link diagram* for G. If the crossings are replaced by ordinary four valent nodes, then $K(G)$ is replaced by the *medial graph* of G. The figure also illustrates the *checkerboard shading* of $K(G)$ where each region of the knot diagram that contains a node of G receives a shading. One can start with a knot diagram K, shade it so that adjacent regions receive different colors. Then assign a graphical node to each shaded region and an edge between

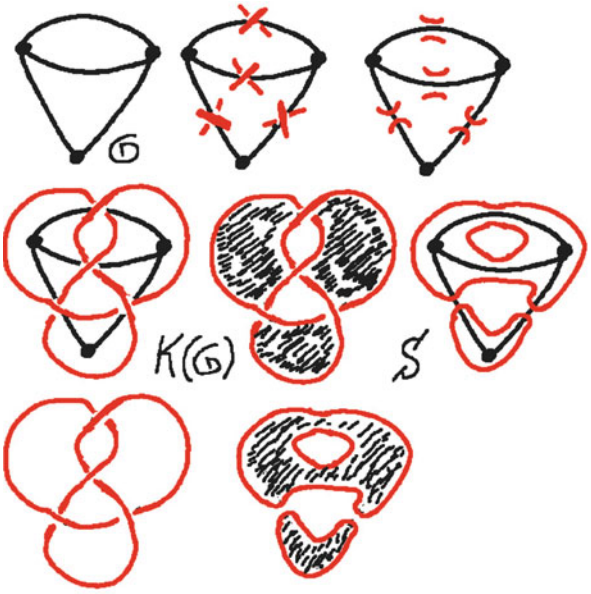


Fig. 3.10 Medial Tait checkerboard graph and K(G)

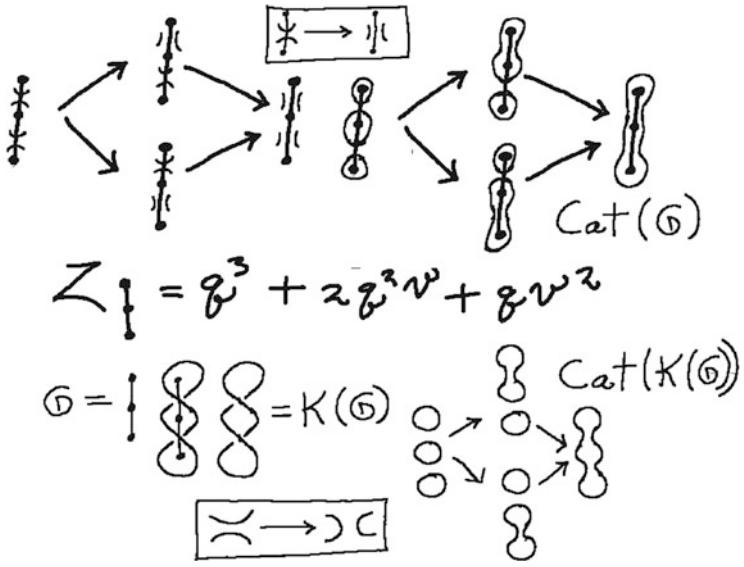


Fig. 3.11 Two edge graph

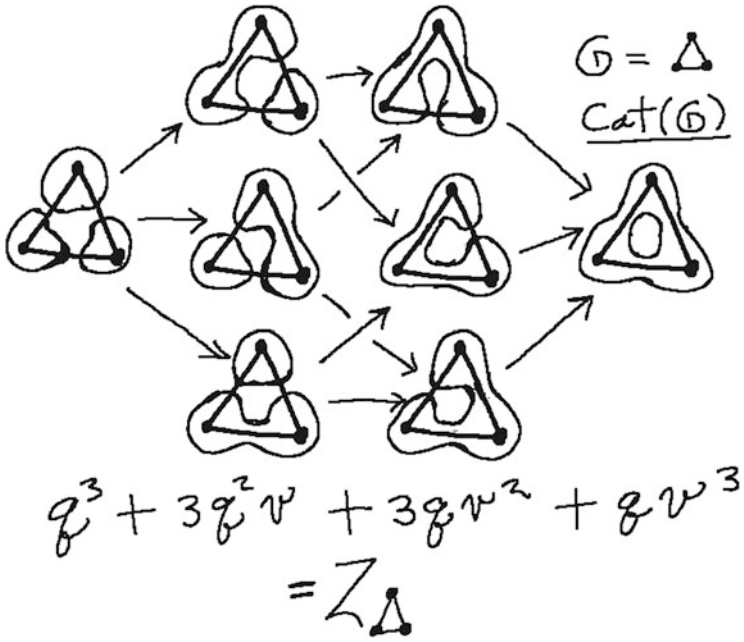


Fig. 3.12 Triangle graph

nodes whenever two shaded regions share a crossing. The resulting graph $G(K)$ has the property that $K = K(G(K))$ when K is an alternating knot diagram with crossings at the shaded regions of (negative) type as shown in the figure. Finally this figure illustrates how one of the states of G results in a configuration of loops obtained from $K(G)$ by choices of smoothings of the crossings of $K(G)$ according to the pattern of the state. Note that in smoothing a crossing we reconnect it so that no arcs locally cross one another.

Figures 3.11 and 3.12 illustrate the state calculations of the dichromatic polynomials of a two edge line graph and a triangle graph. Note that the category of the graph proceeds, in these drawings from left to right where at the left, all edges are deleted and the evaluation is equal to Q raised to the number of nodes in the graph. As one proceeds to the right the factors of v increase one at a time as we replace [●⊃⊂●] by [●⌒⌒●] one at a time. In each column the Q contribution can vary depending upon the number of connected subgraphs in the corresponding deletion. In a shaded diagram, a connected subgraph appears as a connected shaded component.

By the same token the bracket polynomial can be regarded as defined in terms a category of states of the knot diagram and calculated as illustrated in Fig. 3.13. The reader can take from this figure our conventions for referring to smoothings of type A and smoothings of type B. *In this figure we have illustrated the bracket as a calculation of a three variable polynomial in A, B and δ.* In this way the parallel

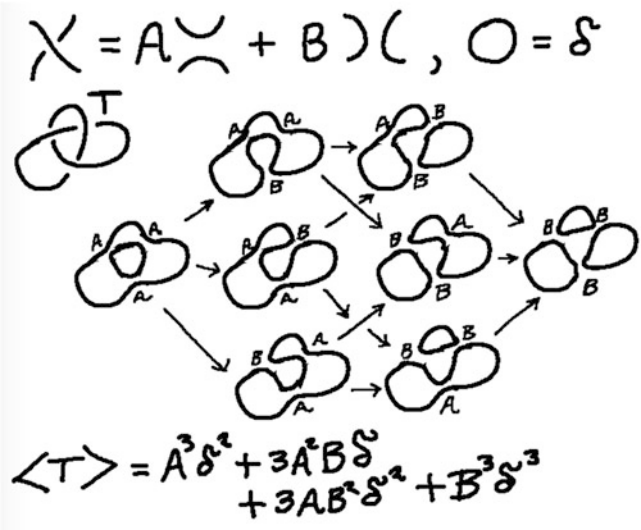


Fig. 3.13 Bracket states and Khovanov complex

with the expansion of the Potts model should be clear. The specialization $B = A^{-1}, \delta = -A^2 - A^{-2}$ gives a Laurent polynomial that is invariant under *regular isotopy* of knot and link diagrams. This is the topological equivalence relation on the diagrams generated by Reidemeister moves 2 and 3 as illustrated in Fig. 3.1.

In [15, 16] it is shown that the dichromatic polynomial $Z[G](v, Q)$ for a plane graph can be expressed in terms of a bracket state summation for $K(G)$ of the form

$$\{\times\} = \{\asymp\} + Q^{-\frac{1}{2}} v \{)(\}$$

with

$$\{\bigcirc\} = Q^{\frac{1}{2}}.$$

Here

$$Z[G](v, Q) = Q^{N/2}\{K(G)\}$$

where $K(G)$ is an alternating link diagram associated with the plane graph G so that the projection of $K(G)$ to the plane is a medial diagram for the graph. Here we use the opposite convention from [16] in associating crossings to edges in the graph.

We will now explain how this result is obtained. It is a key to seeing how the diagrammatic Temperley Lieb algebra is related to the Potts model and it is a key to our discussion in this paper of the relationships of the Potts model with Khovanov homology of knots and links.

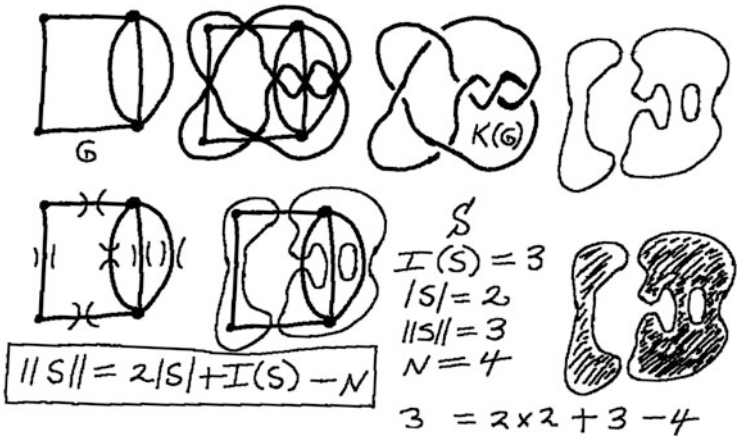


Fig. 3.14 Loop count and region count

The key to translating the dichromatic polynomial to the bracket state summation is the formula

$$||S|| = 2|S| + I(S) - N$$

where $||S||$ denotes the number of loops in a bracket state S. You can think of the bracket state S as obtained from a state of the graph G using the deletion and contraction markings that we have explained above. The $|S|$ denotes the number of shaded regions in the state S and correspondinging, it denotes the number of connected components in the deletion/contraction state of the graph, or equivalently the number of connected components in the subgraph H of G obtained by performing only the deletions. The quantity $I(S)$ is the number of edges of the subgraph H and is called the number of *internal smoothings* of S because these are the smoothings of S that occur with shading between them in the medial shading. Finally N is the number of nodes of the graph G. See Fig. 3.14 for an illustration of this formula. The formula is a direct consequence of Euler's Theorem for plane graphs. As we shall now see, this formula effects the translation we desire between counting shaded regions and counting state loops.

Thus we have the formula

$$Z[G] = \sum_S q^{|S|} v^{I(S)}$$

where a state S can interchangeable be seen as delineating a subgraph H of G or a loops state of $K(G)$. We have described the correspondence using the category $Cat(G)$ and the local smoothings so that on the bracket side we have that



This means that *the B smoothing in the knot diagram corresponds to the internal edges of the graph H and to the internal smoothings of the state S*.

We calculate:

$$Z[G] = \sum_S q^{|S|} v^{I(S)}$$

$$= \sum_S q^{(1/2)(||S||-I(S)+N)} v^{I(S)}$$

$$= q^{N/2} \sum_S (q^{1/2})^{||S||} (q^{-1/2} v)^{I(S)}.$$

$$= q^{N/2} \{K(G)\}$$

where

$$\{\times\} = \{)(\} + Q^{-\frac{1}{2}} v \{)(\}$$

with

$$\{\bigcirc\} = Q^{\frac{1}{2}}.$$

We have finished the proof of the formula for the Potts partition function in terms of the bracket state sum.

The proof of the formula $Z[G] = Q^{N/2}\{K(G)\}$ involves recounting boundaries of regions in correspondence with the loops in the link diagram. The advantage of the bracket expansion of the dichromatic polynomial is that it shows that this graph invariant is part of a family of polynomials that includes the Jones polynomial and it shows how the dichromatic polynomial for a graph whose medial is a braid closure can be expressed in terms of the Temperley-Lieb algebra. This in turn reflects on the structure of the Potts model for planar graphs, as we remark below.

By our above discussion we have a link diagrammatic formulation for the Potts partition function for planar graphs G

$$P_G(Q, T) = Q^{N/2}\{K(G)\}(Q, v = e^K - 1)$$

where N is the number of nodes in the graph G. This bracket expansion for the Potts model is very useful in thinking about the physical structure of the model. For example, since the bracket expansion can be expressed in terms of the Temperley-Lieb algebra one can use this formalism to express the expansion of the Potts model in terms of the Temperley-Lieb algebra. Figure 3.15 acts as a reminder of this formulation. The alternating medial knot $K(G)$ for a rectangular lattice can be seen as a plat closure (using maixma at the top and minima at the bottom)

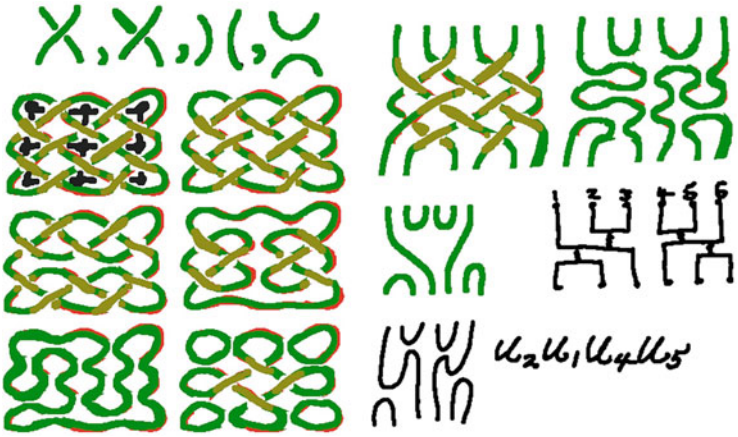


Fig. 3.15 Lattice graph

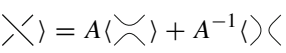
of a braid. Expanding the braid in the Temperley Lieb algebra using our bracket reformulaiton of the Potts model then gives an expansion of the Potts model in terms of the Temperley-Lieb algebra. This expansion is structurally the same as the original version via transfer matrices of Temperley and Lieb. This method clarifies the fundamental relationship of the Potts model and the algebra of Temperley and Lieb. Furthermore the conjectured critical temperature for the Potts model occurs for T when $Q^{-\frac{1}{2}}v = 1$. We see clearly in the bracket expansion that this value of T corresponds to a point of symmetry of the model where the value of the partition function does not depend upon the designation of over and undercrossings in the associated knot or link. This corresponds to a symmetry between the plane graph G and its dual.

3.5 Khovanov Homology

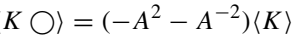
In this section, we describe Khovanov homology along the lines of [1, 19], and we tell the story so that the gradings and the structure of the differential emerge in a natural way. This approach to motivating the Khovanov homology uses elements of Khovanov's original approach, Viro's use of enhanced states for the bracket polynomial [25], and Bar-Natan's emphasis on tangle cobordisms [2]. We use similar considerations in our paper [5].

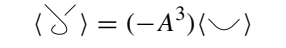
Two key motivating ideas are involved in finding the Khovanov invariant. First of all, one would like to *categorify* a link polynomial such as $\langle K \rangle$. There are many meanings to the term categorify, but here the quest is to find a way to express the link polynomial as a *graded Euler characteristic* $\langle K \rangle = \chi_q \langle \mathcal{H}(K) \rangle$ for some homology theory associated with $\langle K \rangle$.

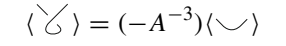
The bracket polynomial [14] model for the Jones polynomial [11–13, 26] is usually described by the expansion



and we have







Letting $c(K)$ denote the number of crossings in the diagram K, if we replace $\langle K \rangle$ by $A^{-c(K)}\langle K \rangle$, and then replace A^2 by $-q^{-1}$, the bracket will be rewritten in the following form:



with $\langle \bigcirc \rangle = (q + q^{-1})$. It is useful to use this form of the bracket state sum for the sake of the grading in the Khovanov homology (to be described below). We shall continue to refer to the smoothings labeled q (or A^{-1} in the original bracket formulation) as *B-smoothings*. We should further note that we use the well-known convention of *enhanced states* where an enhanced sbgtate has a label of 1 or X on each of its component loops. We then regard the value of the loop $q + q^{-1}$ as the sum of the value of a circle labeled with a 1 (the value is q) added to the value of a circle labeled with an X (the value is q^{-1}). We could have chosen the more neutral labels of $+1$ and -1 so that

$$q^{+1} \Longleftrightarrow +1 \Longleftrightarrow 1$$

and

$$q^{-1} \Longleftrightarrow -1 \Longleftrightarrow X,$$

but, since an algebra involving 1 and X naturally appears later, we take this form of labeling from the beginning.

To see how the Khovanov grading arises, consider the form of the expansion of this version of the bracket polynonmial in enhanced states. We have the formula as a sum over enhanced states s :

$$\langle K \rangle = \sum_s (-1)^{n_B(s)} q^{j(s)}$$

where $n_B(s)$ is the number of B-type smoothings in s, $\lambda(s)$ is the number of loops in s labeled 1 minus the number of loops labeled X, and $j(s) = n_B(s) + \lambda(s)$. This can be rewritten in the following form:

$$\langle K \rangle = \sum_{i,j} (-1)^i q^j \dim(C^{ij})$$

where we define C^{ij} to be the linear span (over $k = Z/2Z$ as we will work with mod 2 coefficients) of the set of enhanced states with $n_B(s) = i$ and $j(s) = j$. Then the number of such states is the dimension $\dim(C^{ij})$.

We would like to have a bigraded complex composed of the C^{ij} with a differential

$$\partial : C^{ij} \longrightarrow C^{i+1\,j}.$$

The differential should increase the *homological grading* i by 1 and preserve the *quantum grading* j. Then we could write

$$\langle K \rangle = \sum_j q^j \sum_i (-1)^i \dim(C^{ij}) = \sum_j q^j \chi(C^{\bullet\, j}),$$

where $\chi(C^{\bullet\, j})$ is the Euler characteristic of the subcomplex $C^{\bullet\, j}$ for a fixed value of j.

This formula would constitute a categorification of the bracket polynomial. Below, we shall see how *the original Khovanov differential* ∂ *is uniquely determined by the restriction that* $j(\partial s) = j(s)$ *for each enhanced state* s. Since j is preserved by the differential, these subcomplexes $C^{\bullet\, j}$ have their own Euler characteristics and homology. We have

$$\chi(H(C^{\bullet\, j})) = \chi(C^{\bullet\, j})$$

where $H(C^{\bullet\, j})$ denotes the homology of the complex $C^{\bullet\, j}$. We can write

$$\langle K \rangle = \sum_j q^j \chi(H(C^{\bullet\, j})).$$

The last formula expresses the bracket polynomial as a *graded Euler characteristic* of a homology theory associated with the enhanced states of the bracket state summation. This is the categorification of the bracket polynomial. Khovanov proves that this homology theory is an invariant of knots and links (via the Reidemeister moves of Fig. 3.1), creating a new and stronger invariant than the original Jones polynomial.

We will construct the differential in this complex first for mod-2 coefficients. The differential is based on regarding two states as *adjacent* if one differs from the other by a single smoothing at some site. Thus if (s, τ) denotes a pair consisting in an enhanced state s and site τ of that state with τ of type A, then we consider

all enhanced states s' obtained from s by smoothing at τ and relabeling only those loops that are affected by the resmoothing. Call this set of enhanced states $S'[s, \tau]$. Then we shall define the *partial differential* $\partial_\tau(s)$ as a sum over certain elements in $S'[s, \tau]$, and the differential by the formula

$$\partial(s) = \sum_\tau \partial_\tau(s)$$

with the sum over all type A sites τ in s. It then remains to see what are the possibilities for $\partial_\tau(s)$ so that $j(s)$ is preserved.

Note that if $s' \in S'[s, \tau]$, then $n_B(s') = n_B(s) + 1$. Thus

$$j(s') = n_B(s') + \lambda(s') = 1 + n_B(s) + \lambda(s').$$

From this we conclude that $j(s) = j(s')$ if and only if $\lambda(s') = \lambda(s) - 1$. Recall that

$$\lambda(s) = [s : +] - [s : -]$$

where $[s : +]$ is the number of loops in s labeled $+1$, $[s : -]$ is the number of loops labeled -1 (same as labeled with X) and $j(s) = n_B(s) + \lambda(s)$.

Proposition *The partial differentials $\partial_\tau(s)$ are uniquely determined by the condition that $j(s') = j(s)$ for all s' involved in the action of the partial differential on the enhanced state s. This unique form of the partial differential can be described by the following structures of multiplication and comultiplication on the algebra $\mathcal{A} = k[X]/(X^2)$ where $k = Z/2Z$ for mod-2 coefficients, or $k = Z$ for integral coefficients.*

1. *The element* 1 *is a multiplicative unit and* $X^2 = 0$.
2. $\Delta(1) = 1 \otimes X + X \otimes 1$ *and* $\Delta(X) = X \otimes X$.

These rules describe the local relabeling process for loops in a state. Multiplication corresponds to the case where two loops merge to a single loop, while comultiplication corresponds to the case where one loop bifurcates into two loops.

Proof Using the above description of the differential, suppose that there are two loops at τ that merge in the smoothing. If both loops are labeled 1 in s then the local contribution to $\lambda(s)$ is 2. Let s' denote a smoothing in $S[s, \tau]$. In order for the local λ contribution to become 1, we see that the merged loop must be labeled 1. Similarly if the two loops are labeled 1 and X, then the merged loop must be labeled X so that the local contribution for λ goes from 0 to -1. Finally, if the two loops are labeled X and X, then there is no label available for a single loop that will give -3, so we define ∂ to be zero in this case. We can summarize the result by saying that there is a multiplicative structure m such that $m(1, 1) = 1, m(1, X) = m(X, 1) = x, m(X, X) = 0$, and this multiplication describes the structure of the partial differential when two loops merge. Since this is the multiplicative structure of the algebra $\mathcal{A} = k[X]/(X^2)$, we take this algebra as summarizing the differential.

Now consider the case where s has a single loop at the site τ. Smoothing produces two loops. If the single loop is labeled X, then we must label each of the two loops by X in order to make λ decrease by 1. If the single loop is labeled 1, then we can label the two loops by X and 1 in either order. In this second case we take the partial differential of s to be the sum of these two labeled states. This structure can be described by taking a coproduct structure with $\Delta(X) = X \otimes X$ and $\Delta(1) = 1 \otimes X + X \otimes 1$. We now have the algebra $\mathcal{A} = k[X]/(X^2)$ with product $m : \mathcal{A} \otimes \mathcal{A} \longrightarrow \mathcal{A}$ and coproduct $\Delta : \mathcal{A} \longrightarrow \mathcal{A} \otimes \mathcal{A}$, describing the differential completely. This completes the proof. ♠

Partial differentials are defined on each enhanced state s and a site τ of type A in that state. We consider states obtained from the given state by smoothing the given site τ. The result of smoothing τ is to produce a new state s' with one more site of type B than s. Forming s' from s we either amalgamate two loops to a single loop at τ, or we divide a loop at τ into two distinct loops. In the case of amalgamation, the new state s acquires the label on the amalgamated circle that is the product of the labels on the two circles that are its ancestors in s. This case of the partial differential is described by the multiplication in the algebra. If one circle becomes two circles, then we apply the coproduct. Thus if the circle is labeled X, then the resultant two circles are each labeled X corresponding to $\Delta(X) = X \otimes X$. If the orginal circle is labeled 1 then we take the partial boundary to be a sum of two enhanced states with labels 1 and X in one case, and labels X and 1 in the other case, on the respective circles. This corresponds to $\Delta(1) = 1 \otimes X + X \otimes 1$. Modulo two, the boundary of an enhanced state is the sum, over all sites of type A in the state, of the partial boundaries at these sites. It is not hard to verify directly that the square of the boundary mapping is zero (this is the identity of mixed partials!) and that it behaves as advertised, keeping $j(s)$ constant. There is more to say about the nature of this construction with respect to Frobenius algebras and tangle cobordisms. In Figs. 3.16 and 3.17 we illustrate how the partial boundaries can be conceptualized in terms of surface cobordisms. The equality of mixed partials corresponds to topological equivalence of the corresponding surface cobordisms, and to the relationships between Frobenius algebras and the surface cobordism category. The proof of invariance of Khovanov homology with respect to the Reidemeister moves (respecting grading changes) will not be given here. See [1, 2, 19]. It is remarkable that this version of Khovanov homology is uniquely specified by natural ideas about adjacency of states in the bracket polynomial.

3.6 Homology and the Potts Model

We first analyze how our heuristics leading to the Khovanov homology looks when generalized to the context of the dichromatic polynomial. (This is a different approach to the question than the method of Stosic [22] or [9, 10], but see the next section for a discussion of Stosic's approach to categorifying the dichromatic polynomial.) We ask questions about the relationship of Khovanov homology and

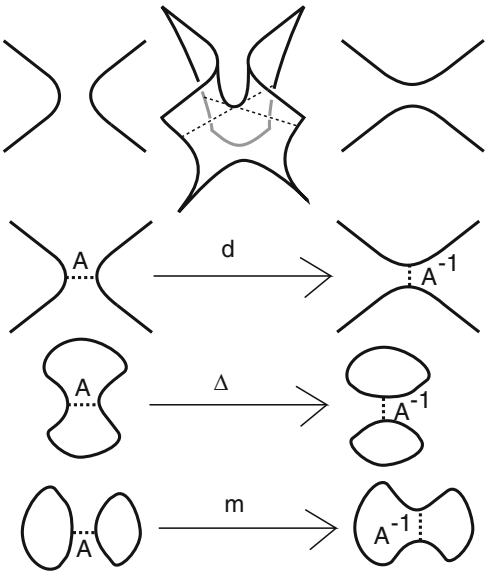


Fig. 3.16 SaddlePoints and state smoothings

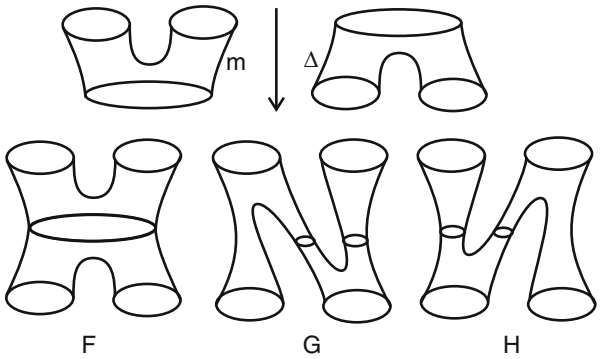


Fig. 3.17 Surface cobordisms

the Potts model. The adjacency of states in the Khovanov homology corresponds to an adjacency for energetic states of the physical system described by the Potts model.

Recall the Khovanov bracket.

$$[\times] = [)(] - q[\,)\,(\,]$$

with

$$[\bigcirc] = q + q^{-1}.$$

This expansion is an intermediary between the Potts model (dichromatic polynomial) and the topological bracket. When

$$-q = Q^{-\frac{1}{2}} v$$

and

$$q + q^{-1} = Q^{\frac{1}{2}},$$

we have the Potts model up to the multiple of $Q^{N/2}$. That is,we have

$$Z[G](Q, v) = Q^{N/2}[K(G)]$$

and so

$$[K(G)] = Q^{-N/2} Z[G](Q, v).$$

We know that the coefficients of $[K(G)](q)$ are Euler characteristics of the Khovanov homology of the link $K(G)$. These topological characteristics will not directly impinge on the Potts model except at certain specific choices of N (the number of nodes of the graph G) and certain temperatures. We will detail this correspondence below, but first we make a more general analysis. We view the dichromatic polynomial as a polynomial in v and Q, in its graph theoretic framework.

Then we can write

$$Z[G](Q, v) = \sum Q^i v^j \pi_{ij}$$

where the π_{ij} are the coefficients of the terms in the Potts model for G. These can be compared with the terms in $[K(G)](Q)$ by using the equations above. We have

$$-q Q^{\frac{1}{2}} = v$$

and hence

$$-q(q + q^{-1}) = v.$$

Hence

$$[K(G)] = q^{-N/2} Z[G](Q, v) = q^{-N/2} \sum_{i,j} (q + q^{-1})^{2i} (-1)^j q^j (q + q^{-1})^j \pi_{ij}$$

$$= q^{-N/2} \sum_{i,j} (-1)^j (q + q^{-1})^{2i+j} q^j \pi_{ij}$$

$$= q^{-N/2} \sum_{i,j} (-1)^j \Big(\sum_{k \le 2i+j} C_k^{2i+j} q^{2i+j-k} \Big) q^j \pi_{ij}$$

$$= \sum_{i,j,k \le 2i+j} (-1)^j q^{2i+2j-k-N/2} C_k^{2i+j} \pi_{ij}$$

$$= \sum_{\lambda} q^{\lambda} \sum_{\lambda = 2i+2j-k-N/2, k \le 2i+j} (-1)^j C_k^{2i+j} \pi_{ij}$$

Thus we have shown

Theorem 1

$$\chi(KH^{\bullet,\lambda}(K(G)) = \sum_{2i+2j-k-N/2=\lambda, k \le 2i+j} (-1)^j C_k^{2i+j} \pi_{ij}$$

where π_{ij} are the coefficients of the Potts partition function for the graph G and C_r^n is the binomial coefficient $n!/(r!((n-r)!)$

This shows that certain specific combinations of Potts partition function coefficients are equal to the Euler characteristic terms in the Khovanov homology. It remains to be seen if there is an interpretation of the ranks of the individual homology groups in terms of the physics of the Potts model.

In order to analyse when the Potts model has direct relationship with the topological bracket, we adopt yet another bracket expansion as indicated below. We call this two-variable bracket expansion the *ρ-bracket.* It reduces to the Khovanov version of the bracket as a function of q when ρ is equal to one.



with



We can regard this expansion as an intermediary between the Potts model (dichromatic polynomial) and the topological bracket. When $\rho = 1$ we have the topological

bracket expansion in Khovanov form. When

$$-q\rho = Q^{-\frac{1}{2}}v$$

and

$$q + q^{-1} = Q^{\frac{1}{2}},$$

we have the Potts model. We shall return to these parametrizations shortly.

Just as in the last section, we have

$$[K] = \sum_s (-\rho)^{n_B(s)} q^{j(s)}$$

where $n_B(s)$ is the number of B-type smoothings in s, $\lambda(s)$ is the number of loops in s labeled 1 minus the number of loops labeled X, and $j(s) = n_B(s) + \lambda(s)$. This can be rewritten in the following form:

$$[K] = \sum_{i,j} (-\rho)^i q^j \dim(C^{ij})$$

$$= \sum_j q^j \sum_i (-\rho)^i \dim(C^{ij}) = \sum_j q^j \chi_\rho(C^{\bullet j}),$$

where we define C^{ij} to be the linear span of the set of enhanced states with $n_B(s) = i$ and $j(s) = j$. Then the number of such states is the dimension $\dim(C^{ij})$. Now we have expressed this general bracket expansion in terms of generalized Euler characteristics of the complexes:

$$\chi_\rho(C^{\bullet j}) = \sum_i (-\rho)^i \dim(C^{ij}).$$

These generalized Euler characteristics become classical Euler characteristics when $\rho = 1$, and, in that case, are the same as the Euler characteristic of the homology. With ρ not equal to 1, we do not have direct access to the homology.

Nevertheless, I believe that this raises a significant question about the relationship of $[K](q, \rho)$ with Khovanov homology. We get the Khovanov version of the bracket polynomial for $\rho = 1$ so that for $\rho = 1$ we have

$$[K](q, \rho = 1) = \sum_j q^j \chi_\rho(C^{\bullet j}) = \sum_j q^j \chi(H(C^{\bullet j})).$$

Away from $\rho = 1$ one can ask what is the influence of the homology groups on the coefficients of the expansion of $[K](q, \rho)$, and the corresponding questions about

the Potts model. This is a way to generalize questions about the relationship of the Jones polynomial with the Potts model. In the case of the Khovanov formalism, we have the same structure of the states and the same homology theory for the states in both cases, but in the case of the Jones polynomial (ρ-bracket expansion with $\rho = 1$) we have expressions for the coefficients of the Jones polynomial in terms of ranks of the Khovanov homology groups. Only the ranks of the chain complexes figure in the Potts model itself. Thus we are suggesting here that it is worth asking about the relationship of the Khovanov homology with the dichromatic polynomial, the ρ-bracket and the Potts model without changing the definition of the homology groups or chain spaces. This also raises the question of the relationship of the Khovanov homology with those constructions that have been made (e.g. [22]) where the homology has been adjusted to fit directly with the dichromatic polynomial. We will take up this comparison in the next section.

We can look more closely at the Potts model by writing a translation between the variables q, ρ and Q, v. We have

$$-q\rho = Q^{-\frac{1}{2}}v$$

and

$$q + q^{-1} = Q^{\frac{1}{2}},$$

and from this we conclude that

$$q^2 - \sqrt{Q}q + 1 = 0.$$

Whence

$$q = \frac{\sqrt{Q} \pm \sqrt{Q-4}}{2}$$

and

$$\frac{1}{q} = \frac{\sqrt{Q} \mp \sqrt{Q-4}}{2}.$$

Thus

$$\rho = -\frac{v}{\sqrt{Q}q} = v(\frac{-1 \pm \sqrt{1-4/Q}}{2})$$

For physical applications, Q is a positive integer greater than or equal to 2. Let us begin by analyzing the Potts model at criticality (see discussion above) where

$-\rho q = 1$. Then

$$\rho = -\frac{1}{q} = \frac{-\sqrt{Q} \pm \sqrt{Q-4}}{2}.$$

For the Khovanov homology (its Euler characteristics) to appear directly in the partition function we want

$$\rho = 1.$$

Thus we want

$$2 = -\sqrt{Q} \pm \sqrt{Q-4}.$$

Squaring both sides and collecting terms, we find that $4 - Q = \mp\sqrt{Q}\sqrt{Q-4}$. Squaring once more, and collecting terms, we find that the only possibility for $\rho = 1$ is $Q = 4$. Returning to the equation for ρ, we see that this will be satisfied when we take $\sqrt{4} = -2$. This can be done in the parametrization, and then the partition function will have Khovanov topological terms. However, note that with this choice, $q = -1$ and so $v/\sqrt{Q} = -\rho q = 1$ implies that $v = -2$. Thus $e^K - 1 = -2$, and so

$$e^K = -1.$$

From this we see that in order to have $\rho = 1$ at criticality, we need a 4-state Potts model with imaginary temperature variable $K = (2n+1)i\pi$. It is worthwhile considering the Potts models at imaginary temperature values. For example the Lee-Yang Theorem [20] shows that under certain circumstances the zeros on the partition function are on the unit circle in the complex plane. We take the present calculation as an indication of the need for further investigation of the Potts model with real and complex values for its parameters.

Now we go back and consider $\rho = 1$ without insisting on criticality. Then we have $1 = -v/(q\sqrt{Q})$ so that

$$v = -q\sqrt{Q} = \frac{-Q \mp \sqrt{Q}\sqrt{Q-4}}{2}.$$

From this we see that

$$e^K = 1 + v = \frac{2 - Q \mp \sqrt{Q}\sqrt{Q-4}}{2}.$$

From this we get the following formulas for e^K : For $Q = 2$ we have $e^K = \pm i$. For Q = 3, we have $e^K = \frac{-1 \pm \sqrt{3}i}{2}$. For $Q = 4$ we have $e^K = -1$. For $Q > 4$ it is easy

to verify that e^K is real and negative. Thus in all cases of $\rho = 1$ we find that the Potts model has complex temperature values.

3.7 The Potts Model and Stosic's Categorification of the Dichromatic Polynomial

In [22] Stosic gives a categorification for certain specializations of the dichromatic polynomial. In this section we describe this categorification, and discuss its relation with the Potts model. The reader should also note that the relations between dichromatic (and chromatic) homology and Khovanov homology are observed in [9] Theorem 24. In this paper we use the Stosic formulation for our analysis.

For this purpose, we define, as in the previous section, the dichromatic polynomial through the formulas:

$$Z[G](v, Q) = Z[G'](v, Q) + vZ[G''](v, Q)$$

$$Z[\bullet \sqcup G] = QZ[G].$$

where G' is the result of deleting an edge from G, while G'' is the result of contracting that same edge so that its end-nodes have been collapsed to a single node. In the second equation, $\bullet$ represents a graph with one node and no edges, and $\bullet \sqcup G$ represents the disjoint union of the single-node graph with the graph G. The graph G is an arbitrary finite (multi-)graph.

We have, in the section on the Potts model and dichromatic polynomial, reformulated this recursion as follows: Instead of contracting an edge of the graph to a point in the second term of the formula, we label that edge so that we know that it has been used in the recursion. We rewrote the recursion in the form:

$$Z[\bullet\!\!-\!\!\bullet] = Z[\bullet\!\!\times\!\!\bullet] = Z[\bullet\!\!\supset\!\!\subset\!\!\bullet] + vZ[\bullet\!\!\smile\!\!\frown\!\!\bullet].$$

where a deletion is labeled $[\bullet\!\!\supset\!\!\subset\!\!\bullet]$ and a contraction is labeled $[\bullet\!\!\smile\!\!\frown\!\!\bullet]$. The crossing label $[\bullet\!\!\times\!\!\bullet]$ reminds us that these two labels stand for two local smoothings of the medial link diagram $K(G)$. For thinking of colorings from the set $\{1, 2, \cdots, Q\}$ when Q is a positive integer, regard an edge marked with $[\bullet\!\!\smile\!\!\frown\!\!\bullet]$ as indicating that the colors on its two nodes are the same. We then see that we can write the following formula for the dichromaatic polynomial:

$$Z[G] = \sum_{H \subset G} Q^{|H|} v^{e(H)}$$

where H is a subgraph of G, $|H|$ is the number of components of H, and $e(H)$ is the number of edges of H. The subgraph H corresponds to the edges of G with a contraction label [●—◡◠—●].

Moving now in the direction of Euler characteristics, we let $w = -v$ so that

$$Z[G] = Z[G'] - wZ[G'']$$

$$Z[\bullet \sqcup G] = QZ[G]$$

and

$$Z[G] = \sum_{H \subset G} (-1)^{e(H)} Q^{|H|} w^{e(H)}.$$

This suggests that differentials should increase the number of edges on the subgraphs, and that the terms $Q^{|H|}w^{e(H)}$ should not change under the application of the (partial) differentials. Stosic's solution to this requirement is to take

$$Q = q^n$$

and

$$w = 1 + q + q^2 + \cdots + q^n$$

so that

$$Z[G] = \sum_{H \subset G} (-1)^{e(H)} q^{n|H|} (1 + q + a^2 + \cdots + q^n)^{e(H)}.$$

To see how this works, we first rewrite this state sum (over states H that are subgraphs of G) as a sum over *enhanced states* h where we define enhanced states h for a graph G to be *labeled subgraphs* h where a labeling of h consists in an assignment of one of the elements of the set $S = \{1, X, X^2, \cdots, X^n\}$ to each component of h. Note that the components of h are in $1-1$ correspondence with the shaded regions in the medial link diagram as we have previously described it. This comes closer to the structure of the Potts model, since a spin state of the Potts model has one spin assigned to each component of the graph H. The enhanced state h corresponds to such an assignment.

Regard the X^i as generators of the ring $R = Z[X]/(X^{n+1})$. Define the *degree* of X^i by the formula $deg(X^i) = n - i$ and let

$$j(h) = n|h| + \sum_{\gamma \in C(h)} deg(label(\gamma))$$

where the sum goes over all γ in $C(h)$, the set of components of h. Each component is labeled from S and $|h|$ denotes the number of components in h. Thus we have

$$Z[G] = \sum_{H \subset G} (-1)^{e(H)} q^{n|H|} (1 + q + a^2 + \cdots + q^n)^{e(H)}$$

$$= \sum_{h \in S(G)} (-1)^{e(h)} q^{n|h|} q^{\lambda(h)}$$

$$= \sum_{h \in S(G)} (-1)^{e(h)} q^{j(h)}$$

where $S(G)$ denotes the set of enhanced states of G and $\lambda(h)$ denotes the sum of the degrees of the labels on the components of h.

We now define a chain complex for a corresponding homology theory. Let $C_i(G)$ be the module generated by the enhanced states of G with i edges. Partial boundaries applied to an enhanced state h simply add new edges between nodes, or from a node to itself. If A and B are components of an enhanced state that are joined by a partial boundary to form a new component C, then C is assigned the label X^{i+j} when A and B have respective labels X^i adn X^j. This partial boundary does not change the labels on other components of h. It may happen that a component A is transformed by adding an edge to itself to form a new component A' without changing the total number of components. In this case, if A has label 1 we assign label X^n to A' and otherwise take the partial boundary to be zero if the label of A is not equal to 1. It is then easy to check that the partial boundaries defined in this way preserve $j(h)$ as defined in the last paragraph, and are compatible so that the composition of boundary with itself is zero. We have described Stosic's homology theory for a specialization of the dichromatic polynomial. We have, as in Sect. 3.4 of this paper, and continuing from above:

$$Z[G] = \sum_{h \in S(G)} (-1)^{e(h)} q^{j(h)}$$

$$= \sum_j \sum_k (-1)^k \dim(C^{k,j})$$

$$= \sum_j q^j \chi(C^{\bullet,j})$$

$$Z[G] = \sum_j q^j \chi(H^{\bullet j}(G))$$

where $\chi(C^{\bullet j}(G))$ denotes the complex defined above generated by enhanced states h with $j = j(h)$, and correspondingly for the homology.

Here we have expressed the Potts model in terms of Stosic homology via the substitution in $Z[G](Q, v)$ so That

$$Z[G](q) = Z[G](q^n, -(1 + q + q^2 + \cdots + q^n)).$$

With a coefficient expansion of the dichromatic polynomial, as in the previous section, we can find the coefficients of the powers of q in $Z[G](q)$:

$$Z[G](Q, v) = \sum_{i,j} Q^i v^j \pi_{ij}$$

$$Z[G](q) = \sum_{i,j} q^{ni} (-1)^j (1 + q + \cdots + q^n)^j \pi_{ij}$$

Let $[j, k]$ be the multinomial coefficient for

$$(1 + q + \cdots + q^n)^j = \sum_{k=0}^{nj} [j, k] q^k.$$

Then we can write

$$Z[G](q) = \sum_{i,j,k} q^{ni+k} (-1)^i [j, k] \pi_{ij}$$

$$= \sum_{\lambda} q^{\lambda} (\sum_{\lambda = ni+k : i,j,k} (-1)^j [j, k] \pi_{ij})$$

Thus we conclude with

Theorem 2

$$\chi(H^{\bullet \lambda}(G)) = \sum_{\lambda = ni+k : i,j,k} (-1)^j [j, k] \pi_{ij}$$

giving the graded Euler characteristics of the Stosic homologies in terms of the coefficients of the Potts dichromatic polynomial model.

Now lets turn to a discussion of the Stosic homology in relation to the Potts model. In the Potts model we have $Q = q^n$ is the number of spins in the model. Thus we can take any q so that q^n is a natural number greater than or equal to 2. For example, we could take q to be an n-th root of 2 and then this would be a two-state Potts model. On the other hand, we have $w = -v = 1 - e^K$ as in our previous

analysis for the Potts model. Thus we have

$$-e^K = q + q^2 + \cdots + q^n.$$

With q real and positive, we can take

$$K = i\pi + ln(q + q^2 + \cdots + q^n),$$

arriving at an imaginary temperature for the values of the Potts model where the partition function is expressed in terms of the homology. If we take $q = (1 + n)^{1/n}$, then the model will have $Q = n + 1$ states, and so, in this case, we can identify the enhanced states of this model as corresponding to the spin assignments of $\{1, X, X^2, \cdots, X^n\}$ to the subgraphs, interpreted as regions of constant spin. The partial boundaries for this homology theory describe particular (global) ways to transit between spin-labeled regions where the regions themselves change locally. Usually, in thinking about the dynamics of a model in statistical physics one looks for evolutions that are strictly local. In the case of the partial differentials we change the configuration of regions at a single bond (edge in the graph) but we make a global change in the spin-labeling (for example, from X^i and X^j on two separate regions to X^{i+j} on the joined region). It is likely that the reason we see the results of this cohomology in the partition function only at imaginary temperature is related to this non-local structure. Nevertheless, the categorified homology is seen in direct relation to the Potts partition function and this connection deserves further examination.

3.8 Imaginary Temperature, Real Time and Quantum Statistics

The purpose of this section is to discuss the nature of imaginary temperature in the Potts model from the point of view of quantum mechanics. We have seen that for certain values of imaginary temperature, the Potts model can be expressed in terms of Euler characteristics of Khovanov homology. The suggests looking at the analytic continuation of the partition function to relate these complex values with real values of the temperature. However, it is also useful to consider reformulating the models so that *imaginary temperature is replaced with real time* and the context of the models is shifted to quantum mechanics. To see how this works lets recall again the general form of a partition function in statistical mechanics. The partition function is given by the formula

$$Z_G(Q, T) = \sum_{\sigma} e^{(-1/kT)E(\sigma)}$$

where the sum runs over all states σ of the physical system, T is the temperature, k is Boltzmann's constant and $E(\sigma)$ is the energy of the state σ. In the Potts model the underlying structure of the physical system is modeled by a graph G, and the energy has the combinatorial form that we have discussed in previous sections.

A quantum amplitude analogous to the partition function takes the form

$$A_G(Q,t) = \sum_{\sigma} e^{(it/\hbar)E(\sigma)}$$

where t denotes the *time* parameter in the quantum model. We shall make precise the Hilbert space for this model below. But note that the correspondence of form between the amplitude $A_G(Q,t)$ and the partition function $Z_G(Q,T)$ suggests that we make the substitution

$$-1/kT = it/\hbar$$

or equivalently that

$$t = (\hbar/k)(1/iT).$$

Time is, up to a factor of proportionality, inverse imaginary temperature. With this substitution, we see that when one evaluates the Potts model at imaginary temperature, it can be interpreted as an evaluation of a quantum amplitude at the corresponding time given by the formula above. Thus we obtain a quantum statistical interpretation of those places where the Potts model can be expressed directly in terms of Khovanov homology. In the process, we have given a quantum statistical interpretation of the Khovanov homology.

To complete this section, we define the associated states and Hilbert space for the quantum amplitude $A_G(Q,t)$. Let $\mathcal{H}$ denote the vector space over the complex numbers with orthonormal basis $\{|\sigma\rangle\}$ where σ runs over the states of the Potts model for the graph G. Define a unitary operator $U(t) = e^{(it/\hbar)H}$ by the formula on the basis elements

$$U(t)|\sigma\rangle = e^{(it/\hbar)E(\sigma)}|\sigma\rangle.$$

The operator $U(t)$ implicitly defines the Hamiltonian for this physical system. Let

$$|\psi\rangle = \sum_{\sigma} |\sigma\rangle$$

denote an initial state and note that

$$U(t)|\psi\rangle = \sum_{\sigma} e^{(it/\hbar)E(\sigma)}|\sigma\rangle,$$

and

$$\langle \psi | U(t) | \psi \rangle = \sum_{\sigma} e^{(it/\hbar)E(\sigma)} = A_G(Q, t).$$

Thus the Potts amplitude is the quantum mechanical amplitude for the state $|\psi\rangle$ to evolve to the state $U(t)|\psi\rangle$. With this we have given a quantum mechanical interpretation of the Potts model at imaginary temperature.

Note that if, in the Potts model, we write $v = e^K - 1$, then in the quantum model we would write

$$e^K = e^{it/\hbar}.$$

Thus we can take $t = -\hbar K i$, and if K is pure imaginary, then the time will be real in the quantum model.

Now, returning to our results in Sect. 3.4 we recall that in the Potts model we have the following formulas for e^K : For $Q = 2$ we have $e^K = \pm i$. For Q = 3, we have $e^K = \frac{-1 \pm \sqrt{3}i}{2}$. For $Q = 4$ we have $e^K = -1$. It is at these values that we can interpret the Potts model in terms of a quantum model at a real time value. Thus we have these interpretations for $Q = 2, t = \hbar\pi/2$; $Q = 3, t = \hbar\pi/6$ and $Q = 4, t = \hbar\pi$. At these values the amplitude for the quantum model is $A_G(Q, t) = \sum_{\sigma} e^{(it/\hbar)E(\sigma)}$ and is given by the formula

$$A_G(Q, t) = Q^{N/2}\{K(G)\}$$

where

$$\{K(G)\} = \sum_j q^j \chi(H^{\bullet\, j}(K(G)))$$

where $H^{\bullet\, j}(K(G))$ denotes the Khovanov homology of the link $K(G)$ associated with the planar graph G and $q = (1 - e^{it/\hbar})/\sqrt{Q}$. At these special values the Potts partition function in its quantum form is expressed directly in terms of the Khovanov homology and is, up to normalization, an isotopy invariant of the link $K(G)$.

Acknowledgments It gives the author of this paper great pleasure to acknowledge helpful conversations with Mario Rasetti at the beginning of this research.

References

1. D. Bar–Natan, On Khovanov's categorification of the Jones polynomial. Algebraic Geom. Topol. **2**(16), 337–370 (2002)
2. D. Bar-Natan, Khovanov's homology for tangles and cobordisms. Geom. Topol. **9–33**, 1465–1499 (2005). arXiv:mat.GT/0410495
3. R.J. Baxter, *Exactly Solved Models in Statistical Mechanics* (Acad. Press, New York, 1982)
4. B. Bollobas, *Modern Graph Theory* (Springer, Berlin 2002)
5. H.A. Dye, L.H. Kauffman, V.O. Manturov, On two categorifications of the arrow polynomial for virtual knots. arXiv:0906.3408
6. C.M. Fortuin, P.W. Kasteleyn, On the random-cluster model. I. Introduction and relation to other models. Physica **57**, 536–564 (1972)
7. S. Gukov, I. Sergei, *Lectures on Knot Homology and Quantum Curves*. Physics and Mathematics of Link Homology, 59–97, Contemp. Math., 680, Centre Rech. (Math. Proc., Amer. Math. Soc., Providence, 2016)
8. S. Gukov, Gauge theory and knot homologies. Fortschr. Phys. **55**(5–7), 473–490 (2007)
9. L. Helme-Guizon, J.H. Przytycki, Y. Rong, Torsion in graph homology. Fund. Math. **190**, 139–177 (2006)
10. E.F. Jasso-Hernandez, Y. Rong, A categorification of the Tutte polynomial. Alg. Geom. Topol. **6**, 2031–2049 (2006)
11. V.F.R. Jones, A polynomial invariant for links via von Neumann algebras. Bull. Am. Math. Soc. **129**, 103–112 (1985)
12. V.F.R. Jones. Hecke algebra representations of braid groups and link polynomials. Ann. Math. **126**, 335–338 (1987)
13. V.F.R. Jones, On knot invariants related to some statistical mechanics models. Pac. J. Math. **137**(2), 311–334 (1989)
14. L.H. Kauffman, State models and the Jones polynomial. Topology **26**, 395–407 (1987)
15. L.H. Kauffman, Statistical mechanics and the Jones polynomial. AMS Contemp. Math. Series **78**, 263–297 (1989)
16. L.H. Kauffman, *Knots and Physics* (World Scientific Publishers, Singapore, 1991), Second Edition (1993), Third Edition (2002)
17. L.H. Kauffman, Knot diagrammatics, in *Handbook of Knot Theory*, ed. by Menasco and Thistlethwaite (Elsevier B. V., Amsterdam, 2005), pp. 233–318
18. L.H. Kauffman, Remarks on Khovanov homology and the Potts model, in *Perspectives in Analysis, Geometry, and Topology: On the Occasion of the 60th Birthday of Oleg Viro*. Progress in Mathematics (Birkhuser/Springer, New York, 2012), pp. 237–262
19. M. Khovanov, A categorification of the Jones polynomial. Duke Math. J **101**(3), 359–426 (1997)
20. T.D. Lee, C.N. Yang, Statistical theory of equations of state and phase transitions II. Lattice gas and Ising model. Phys. Rev. Lett. **87**, 410–419 (1952)
21. K. Reidemeister, *Knotentheorie* (Chelsea Pub. Co., New York, 1948), Copyright 1932, Julius Springer, Berlin
22. M. Stosic, Categorification of the dichromatic polynomial for graphs. JKTR **17**(1), 31–45 (2008)
23. H.N.V. Temperley, *Graph Theory and Applications* (Ellis Horwood Ltd., 1981)
24. H.N.V. Temperley, E.H. Lieb, Relations between the percolation and coloring problems on the plane square lattice. Some exact results for the percolation problem. Proc. Roy. Soc. **A 322**, 251 (1971)
25. O. Viro, Khovanov homology, its definitions and ramifications. Fund. Math. **184**, 317–342 (2004)
26. E. Witten, Quantum field theory and the jones polynomial. Commun. Math. Phys. **121**, 351–399 (1989)

Chapter 4
The Penrose–Onsager–Yang Approach to Superconductivity and Superfluidity



Anthony Leggett

While the work of Professor C.N.Yang in condensed matter physics may not be quite as universally known as that in the area of particle physics, it includes a number of seminal papers, of which the one I like best is the 1962 Reviews of Modern Physics paper [1] entitled "Concept of off diagonal long range order and the quantum phases of liquid He and of superconductors". Written within five years of the publication [2] of a microscopic theory of superconductivity by Bardeen, Cooper and Schrieffer (BCS), this paper in effect asks the general question "what exactly is it that superfluid helium and superconductors have in common, and which is at the root of their anomalous properties?" Although this question has also been answered by others in rather different terms, I believe that the answer given by Yang is by far the clearest and most intuitive, and avoids raising the pseudo-questions (and providing the pseudo-answers) which competing formulations risk doing. In this brief essay, which is emphatically intended as a review-type rather than an original contribution, I shall first briefly review the context and content of ref.[1] in less formal terms than the original papers, discuss some of the (real) questions which it raises, and finally comment on some of the ways in which alternative approaches may lead to misleading conclusions. Some of the content of this essay, in particular that of Sect. 4.3, is elaborated further in Ch.2 of ref.[3], which I will frequently cross-reference. In the following I shall use the term "quantum condensation" to cover both the phenomenon of Bose-Einstein condensation (BEC) in a system of bosons and that of Cooper pairing in Fermi systems.

The transcription, from a manuscript, of this chapter is done by Steven Yan Xiao, Tsinghua University.

A. Leggett (✉)
Department of Physics, University of Illinois Urbana-Champaign, Champaign, IL, USA
e-mail: aleggett@illinois.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_4

4.1 Quantum Condensation: The Onsager-Penrose-Yang Approach

Consider first the case of a simple system of (effective) bosons such as 4-He atoms, which show the phenomenon of superfluidity (cf. ref.[3], ch.1 for definition) below a temperature of about 2.17K at saturated vapour pressure. F. London had suggested in 1948 that this phenomenon is a result of the onset in this system of Bose-Einstein condensation (BEC), that is, a transition below which a non-vanishing fraction (or equivalently a macroscopic number) of particles occupy a single one-particle state; and since a free gas of atoms with the mass and density of 4-He should in fact undergo this transition at around 3K, this suggestion is rather plausible. However, the atoms composing liquid 4-He are far from free, indeed one would expect their mutual interactions to be rather strong. So is it possible to generalize the definition of BEC to a general (arbitrarily strongly interacting) system? This question was addressed by Penrose and Onsager [4] on the basis of an earlier suggestion [5] by Penrose, with the following result: consider a state of N Bose particles (with no "internal" degrees of freedom, cf. below), which is described at time t by an (N-particle) density matrix of diagonalised form which says it has a probability $p_n(t)$ described by a many-body wavefunction $\Psi_n(\{\underline{r}_i\} : t)$ (not necessarily an eigenfunction of the many-body Hamiltonian) of the form

$$\Psi_n = \Psi_n(\underline{r}_1, \underline{r}_2, \ldots, \underline{r}_i, \ldots, \underline{r}_j, \ldots, \underline{r}_N : t) \tag{4.1}$$

which because of the Bose statistics is symmetric under the exchange of any arbitrary pair i, j. Then pick a definite particle (for simplicity let's say particle 1—because of the symmetry this choice is arbitrary—and to distinguish it from the $N-1$ other particles just label its coordinate $\underline{r}$ and define the single-particle density matrix

$$\rho(\underline{r}, \underline{r}' : t) =$$
$$N \sum_n p_n(t) \int \mathrm{d}\underline{r}_2 \ldots \mathrm{d}\underline{r}_N \Psi_n(\underline{r} : \underline{r}_2, \underline{r}_3, \ldots, \underline{r}_N : t) \Psi_n^*(\underline{r}' : \underline{r}_2, \underline{r}_3, \ldots, \underline{r}_N : t) \tag{4.2}$$

where we note that the integration runs over all $\underline{r}_i$ except $\underline{r} = \underline{r}_1$. At first sight the quantity $\rho(\underline{r}, \underline{r}' : t)$ looks horribly complicated, but it is greatly simplified by noting that by virtue of its Hermiticity it can always be cast in diagonal form:

$$\rho(\underline{r}, \underline{r}' : t) = \sum_k n_k(t) \chi_k(\underline{r}, t) \chi_k^*(\underline{r}', t) \tag{4.3}$$

where to help our intuition, I have used the notation n_k for the diagonal elements (eigenvalues). In the special case in which p_n is $\delta_{n,0}$ and the many-particle

wavefunction $\Psi_0(\{\underline{r}_i\} : t)$ is a simple symmetrised product of the single-particle wavefunctions $\psi_k(\underline{r} : t)$, it may be verified that $n_k(t)$ is just the occupation number of single-particle state k at time t; so the intuitive interpretation of $n_k(t)$ in the general case is as that quantity averaged over both the behaviour of the $N-1$ other particles and the many-body states n occurring in the N-particle density matrix. Similarly, the quantity $\chi_k(\underline{r}, t)$ is in some intuitive sense the "best approximation" of the "wave function" of the single particle state k similarly averaged (note that even if the system Hamiltonitan is translation invariant, for general ρ_N the "wave function" $\psi_k(\underline{r} : t)$ need not be anything like a plane wave).

While Penrose and Onsager (and subsequently Yang) were primarily interested in liquid 4-He, a system which at the relevant level of discussion has no "internal" degrees of freedom (DOFs), the generalisation is straightforward: all that is necessary is to add to the spatial coordinate $\underline{r}_i$ in Ψ_N and subsequently defined quantities, a second argument (call it σ_i) describing the internal DOF, so that e.g. the eigenvalues and eigenfunctions of ρ_1 are now $n_{k,\sigma}$ and $\chi_k(\underline{r}, \sigma)$ respectively. Such a generalisation is necessary when one wants to discuss the bosonic ultracold alkali gases, the atoms of which are usually characterised, in addition to the spatial coordinate, by a hyperfine DOF.

Up to this point the discussion is completely general. Now we ask: are any of the eigenvalues $n_k(t)$ extensive, i.e. do they increase linearly with the total number of particles N? If the answer is no, then we have a "normal" system, i.e. one without BEC. The only alternative which Penrose and Onsager explicitly consider is that just one eigenvalue is of order N, a situation which they identify as (generalized) BEC; they suggest that this is the correct identification, as London had suggested, of the superfluid phase of liquid 4-He. The latter was indeed the only plausible candidate at the time they wrote; however, following the experimental realization of BEC in some ultracold Bose alkali gases in the last decade of the twentieth century and the development of the relevant theory, we now recognize that an intermediate situation, "fragmented BEC", in which two or more eigenvalues are of order N, is possible and occasionally realized, for example in Bose systems with an internal (hyperfine) degree of freedom. I return in Sect. 4.3 to the question of why the "simple" BEC of Penrose and Onsager ("macroscopic occupation of a single one-particle state") seems to be the norm and the "fragmented" case the rare exception.

I now turn to ref.[1]. ln that paper, Yang generalizes the idea of macroscopic occupation of a single state to the case when this state describes not a single particle but $2, 3, \ldots, n$ particles, and proves a number of important theorems concerning such a possibility. The case of most interest is $n = 2$ ("Cooper pairing"), which he identifies as that occurring in a superconductor, so I confine my explicit discussion to that case. Consider then a system of N spin-1/2 fermions $i = 1, 2, \ldots, N$ with coordinates $\underline{r}_i$ and z-projection of the spin $\{\sigma_i\}$ with $\sigma_i = \pm 1/2$, then the many-body system density matrix ρ_N describes a system in which at time t has a probability $p_n(t)$ of being described by a wavefunction $\Psi_n(t)$. This is the same as in the Bose case except that (1) it has as arguments not only the coordinates$\{\underline{r}_i\}$ (and t) but also the spin projections $\{\sigma_i\}$ and (2) it is antisymmetric rather than symmetric under the

simultaneous exchange $\underset{\sim}{r} \leftrightarrows \underset{\sim}{r}_j$ and $\sigma_i \leftrightarrows \sigma_j$. It is straightforward to show [1] that in such a system the maximum eigenvalue of the single-particle density matrix is just 1 (this is just the Pauli principle), so let us consider the two-particle reduced density matrix ρ_2 by fixing our attention on two arbitrary particles i and j (again for convenience we choose them to be 1 and 2) and defining

$$\rho_2(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2 : \underset{\sim}{r}'_1, \sigma'_1, \underset{\sim}{r}'_2, \sigma'_2 : t) = \tag{4.4}$$

$$N(N-1) \sum_n p_n(t) \sum_{\sigma_3, \ldots, \sigma_N} \int \mathrm{d}\underset{\sim}{r}_3 \mathrm{d}\underset{\sim}{r}_4 \ldots \mathrm{d}\underset{\sim}{r}_N \Psi_n(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2, \underset{\sim}{r}_3, \sigma_3, \ldots, \underset{\sim}{r}_n, \sigma_n : t)$$

$$\Psi_n^*(\underset{\sim}{r}'_1, \sigma'_1, \underset{\sim}{r}'_2, \sigma'_2 : \underset{\sim}{r}_3, \sigma_3, \ldots, \underset{\sim}{r}_n, \sigma_n : t)$$

where we note that, again, all $\underset{\sim}{r}$' and σ's except 1 and 2 are integrated over. Just as in the Bose case, the expression for ρ_2 is enormously simplified by using its Hermitian character to diagonalise it:

$$\rho_2(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2 : \underset{\sim}{r}'_1, \sigma'_1, \underset{\sim}{r}'_2, \sigma'_2 : t) = \tag{4.5}$$

$$\sum_k n_k(t) \chi_k(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2) \chi_k^*(\underset{\sim}{r}'_1, \sigma'_1, \underset{\sim}{r}'_2, \sigma'_2)$$

so that the quantity $n_k(t)$ is the "number of pairs of the two-particle state $\chi_k(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2)$ averaged over the behaviours in all the other particles $3, \ldots, N$" (and the statistical distribution p_n of the many-body states), and $\chi_k(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2)$ is (the nearest we can get to) an effective wavefunction of this two-particle state. An important result proved in ref.[1] is that while in a Bose system the maximum value of any eigenvalue n_k can be of order N^2, in a system of fermions such as we are currently discussing that maximum value is, even for an arbitrarily large Hilbert space, only of order N. Like the Schroedinger wave function, the function $\chi_k(\underset{\sim}{r}_1, \sigma_1, \underset{\sim}{r}_2, \sigma_2)$ has to be antisymmetric under simultaneous exchange of coordinates and spin DOFs.

So, following Yang, let us consider a given state of our N-fermion system and ask the question: are any of the eigenvalues of the two-particle density matrix ρ_2 of order N? If the answer is no, then we are dealing with a "normal" system: note that, unlike in the Bose case, any thermal equilibrium state of N noninteracting fermions is "normal". As elsewhere in the paper, Yang considers explicitly only one other possibility, namely that one of the n_k's is of order N and all the rest of order 1, and he identifies this state of affairs with the superconducting state of metals (nowadays such a state would normally be characterised as "Cooper-paired"). Again in this case, it is now realized that in certain special circumstances we may get a "fragmented condensate" in which more than one eigenvalue of ρ_2 scales with N; I will return to this possibility in Sect. 4.3.

A very illuminating illustration of these ideas which was not available at the time of publication of ref.[1] is the so-called "BEC-BCS crossover" which has been realised in recent years in certain ultracold Fermi alkali gases which exhibit

a so-called "Feshbach resonance" (for details see e.g. Ch.4 of ref.[3]). In such a system, by varying the external magnetic field through the value at which the resonance occurs, one can tune the ratio of the so-called s-wave scattering length a_s to the interparticle separation r from positive and small through positive and large to negative and large to negative and small. The "positive and small" regime corresponds to the formation of $N/2$ spin-0 difermionic molecules, which behave as nearly noninteracting bosons and duly undergo BEC at sufficiently low temperatures, while the "negative and small" extreme describes a gas of weakly interacting fermions with a weak attraction, which is precisely the situation considered by BCS in their theory of superconductivity (and later in application to the neutral Fermi system liquid 3-He). We note that if we explicitly rewrite the ground-state wave function of the BEC molecular gas in terms of the constituent fermionic atoms, the two-atom reduced density matrix has a single eigenvalue which in the limit $a_s/r \to +0$ tends to $N/2$ (and the associated two-particle wave function χ_0 is just the spin singlet times the internal molecular ground-state wavefunction $\psi(\underline{\mathcal{L}}_1 - \underline{\mathcal{L}}_2)$ multiplied by a constant dependence on the COM variable $\underline{\mathcal{L}}_1 + \underline{\mathcal{L}}_2$). On the other hand, in the "BCS limit" $a_s/r \to -0$, the condensate number is only a fraction of order Δ/ϵ_F of N, where Δ is the superconducting energy gap and ϵ_F is the (normal-state) Fermi energy (in the "classic" superconductors discussed by BCS, this fraction is typically of order $10^{-3} - 10^{-4}$).

From now on, when dealing with the case of "simple" condensation discussed in ref.[1] (a single macroscopic value of ρ_1 for the Bose case or of ρ_2 for the Fermi case), I will denote the single macroscopic eigenvalue by N_0 and the corresponding eigenfunction (which is normalised so that $\int d\underline{\mathcal{L}} |\chi_0(\underline{\mathcal{L}})|^2 = 1$) by χ_0, and in either case I will refer, as mentioned above, to the transition in which such a macroscopic value emerges generally as "quantum condensation".

Ref.[1] explicitly declines to develop a theory of the energetics of the condensation process, so it may be regarded as complementary to more microscopic approaches, e.g. that of Bogoliubov in the Bose case and of BCS for fermions; it is fairly straightforward to show that each of these theories, and many other microscopic approaches developed over the last 60 years, are special cases of the general framework set out in ref.[1]. (In the BCS case, the condensate wave function is, up to normalization, the quantity often known as the "anomalous average" (cf. Sect. 4.3) and denoted by $F(\underline{r})$ and the quantity N_0 is the integral of $|F(\underline{r})|^2$ over $\underline{r}$.)

4.2 Some Considerations and Questions Raised by the Content of Sect. 4.1

In this section, I assume throughout that we are talking about "simple" quantum condensation (just a single eigenvalue ρ_1 or ρ_2 macroscopic). I first discuss the notion of "order parameter" and possible definitions thereof:

It is often convenient to combine our information about the quantities N_0 and χ_0 into a single quantity Ψ which we call the "order parameter", for a (spinless) Bose system this is defined by

$$\Psi(\underset{\sim}{r}, t) \equiv \sqrt{N_0(t)}\chi_0(\underset{\sim}{r}, t) \ . \tag{4.6a}$$

Note that with this definition the quantity $\int d\underset{\sim}{r}|\Psi(\underset{\sim}{r})|^2$ is the condensate number N_0. For a Fermi system, we can define an analogous quantity by

$$F(\underset{\sim}{r}\sigma, \underset{\sim}{r}'\sigma' : t) \equiv \sqrt{N_0(t)}\chi_0(\underset{\sim}{r}\sigma : \underset{\sim}{r}'\sigma' : t) \ . \tag{4.6b}$$

In the (BCS-based) superconductivity literature it is common to express F in terms of COM coordinates $\underset{\sim}{R} = (\underset{\sim}{r}+\underset{\sim}{r}')/2$ and relative coordinate $\underset{\sim}{\rho} \equiv \underset{\sim}{r}-\underset{\sim}{r}'$, fix the spins to $\sigma = +1$, $\sigma' = -1$ and define

$$\Psi(\underset{\sim}{R}, t) \equiv F(\underset{\sim}{R}, \underset{\sim}{\rho} = 0, \sigma = +, \sigma' = - : t) \tag{4.6c}$$

a quantity which turns out to be, apart possibly from normalisation, just the "macroscopic wave function" defined by Ginzburg and Landau on intuitive grounds. For cases more general than BCS this simple definition does not work and further considerations are necessary: see e.g. ch.7 of ref.[3]. It should be explicitly noted that with the above "Penrose-Onsager-Yang" definition of the order parameter, the phase of the latter is just that of the microscopic Schroedinger-like one- or two-particle function into which condensation has taken place, and thus just as in the simple Schroedinger case, while its variation in space and time has a physical meaning, its "absolute" value does not. This point will be revisited in Sect. 4.4.

Alternative definitions of the order parameter (hereafter OP) are possible, and common in the literature. One, which is actually given in ref. [1], is via the concept of the "off-diagonal long-range order" (ODLRO) which is part of the title of that paper: taking for definiteness the (spinless) Bose case and assuming that the Hamiltonian is translation-invariant, one considers the behavior of the single-particle reduced density matrix $\rho_1(\underset{\sim}{r}, \underset{\sim}{r}')$ in the limit $|\underset{\sim}{r}-\underset{\sim}{r}'|$ tends to infinity. It turns out (cf. [1]) that in this limit we have

$$\rho_1(\underset{\sim}{r}, \underset{\sim}{r}') = \Psi(\underset{\sim}{r})\Psi^*(\underset{\sim}{r}') = f(\underset{\sim}{r}, \underset{\sim}{r}') \ ,$$

where the function $f(\underset{\sim}{r}, \underset{\sim}{r}')$ falls to zero as $|\underset{\sim}{r} - \underset{\sim}{r}'|$ tends to infinity. Thus a possible alternative definition for $\Psi(\underset{\sim}{r})$, up to a multiplicative constant, is

$$\Psi(\underset{\sim}{r}, t) = \lim_{|\underset{\sim}{r}-\underset{\sim}{r}'|\to\infty} \rho_1(\underset{\sim}{r}, \underset{\sim}{r}' : t) \ . \tag{4.7}$$

For the Fermi case the above definition is modified by simply replacing $\underset{\sim}{r}$ by the COM variable while keeping the relative coordinate ρ and the spin variables constant, e.g. at $\rho = 0$, $\sigma = +$, $\sigma' = -$.

An alternative definition of the OP, which as far as I know has not been used extensively for the Bose case but following the work of Gor'kov [6] and others in the late fifties is very common in the literature for the Fermi case, is that it is the "anomalous average", that is the matrix element of the relevant (n-particle) operator between the groundstate of the N-particle and $(N-n)$-particle systems where n is the order of the quantum condensation:

$$\text{Bose}: \quad \Psi(\underset{\sim}{r}, t) \equiv \langle N-1|\widehat{\psi}(\underset{\sim}{r})|N\rangle(t) \ , \tag{4.8a}$$

$$\text{Fermi}: \quad F(\underset{\sim}{r}\sigma, \underset{\sim}{r}'\sigma' : t) \equiv \langle N-2|\widehat{\psi}(\underset{\sim}{r}, \sigma)\widehat{\psi}(\underset{\sim}{r}', \sigma')|N\rangle(t) \ , \tag{4.8b}$$

where $|N\rangle$ denotes the groundstate of the N-particle Hamiltonian (note that in the Fermi case N may be either even or odd).

Finally, there is one definition which, apparently largely under the influence of the late P.W.Anderson, has become almost the norm in much of the recent literature, in both the Bose and Fermi cases, namely that based on the idea of "spontaneously broken U(1) symmetry" (hereafter abbreviated SBU(1)'s). Taking the uniform Bose case for definiteness, the idea is to regard the particle annihilation operator $\widehat{\psi}(\underset{\sim}{r})$ as analogous to, say, the spin operator $\widehat{\underset{\sim}{S}}_i$ in a (uniform) Heisenberg ferromagnet. In the latter system, for which the Hamiltonian is invariant under the simultaneous rotation of all spins through the same angle, i.e. under the group O(3), it is generally accepted that the expectation value of the total spin vector $\underset{\sim}{S} \equiv \sum_i \underset{\sim}{S}_i$ is zero above the phase transition T_c but nonzero below T_c and in particular at $T = 0$; the O(3) symmetry is said to be "spontaneously broken" at low temperatures. (This formulation is too crude—a proper treatment needs a careful discussion of the order in which the limits external field tends to zero and volume tends to infinity are taken—but it will do for our purposes). In the analogous approach to (uniform) BEC, one notes that the Hamiltonian is invariant under the global U(1) transformation $\widehat{\psi}(\underset{\sim}{r}) \to \widehat{\psi}(\underset{\sim}{r})\exp(i\varphi)$ where φ is a $\underset{\sim}{r}$-independent phase; and one therefore posits that this U(1) symmetry is "spontaneously broken" in the low-temperature (BEC) phase, i.e. that the integral of $\widehat{\psi}(\underset{\sim}{r})$ over $\underset{\sim}{r}$ and hence the quantity $\widehat{\psi}(\underset{\sim}{r})$ itself, has a nonzero expectation value for at least some $\underset{\sim}{r}$. With this assumption, one defines simply

$$\Psi(\underset{\sim}{r}, t) \equiv \langle\widehat{\psi(\underset{\sim}{r})}|\widehat{\psi(\underset{\sim}{r})}\rangle(t) \ . \tag{4.9a}$$

By a similar argument, in the Fermi case one defines the two-particle OP as

$$F(\underset{\sim}{r}\sigma, \underset{\sim}{r}'\sigma' : t) \equiv \langle\widehat{\psi}(\underset{\sim}{r}\sigma)\widehat{\psi}(\underset{\sim}{r}'\sigma')|\widehat{\psi}(\underset{\sim}{r}\sigma)\widehat{\psi}(\underset{\sim}{r}'\sigma')\rangle(t) \ . \tag{4.9b}$$

While the definitions (4.9a) and (4.9b) are very compact, they have the following ineluctable consequence: since $\langle N|\widehat{\psi}(\underset{\sim}{r})|N\rangle = \langle N|\widehat{\psi}(\underset{\sim}{r}\sigma)\widehat{\psi}(\underset{\sim}{r}'\sigma')|N\rangle = 0$, it follows that for the quantity Ψ (F) to be nonzero the state of the system cannot be in an

eigenstate of the total particle number $\widehat{N}$, nor even in a classical mixture of such states, but rather must be of the form

$$\Psi = \sum_N c_N \Psi_N \ , \tag{4.10}$$

(with at least two different C_N nonzero), or a classical mixture of such states. It is only if we believe that the many-body wavefunction has the form Eq. (4.10) that we can ascribe to the order parameter a definite absolute phase. In conversations over the years with members of the sub-community which favours this approach, the present author has formed the impression that at least a large fraction of it believes that a state of the form Eq. (4.10), or a mixture of such states, is indeed the correct quantum-mechanical description of the relevant physical system when it is in the quantum-condensed state. In Sect. 4.4 arguments will be presented against this point of view.

Finally, I note that in the Fermi case it is often possible to write the many-body wavefunction in the form (up to normalisation)

$$\Psi_N(\{\underset{\sim}{\mathcal{L}}_i, \sigma_i\} : t) = \left(\iint \mathrm{d}r \mathrm{d}r' K(rr', \sigma\sigma', t) \widehat{\psi}^\dagger(r\sigma) \widehat{\psi}^\dagger(r'\sigma') \right)^{N/2} .$$

It may be tempting to identify the function $K(rr' : \sigma\sigma' : t)$ with the order parameter $F(rr', \sigma\sigma', t)$, but in general this is not correct. In fact in simple BCS theory with no COM-dependence and $\sigma = -\sigma'$, for which the Fourier transform of K with respect to the relative coordinate $\underset{\sim}{\varrho}$ is often written $c_{\underset{\sim}{k}}$, the two Fourier transforms are related by

$$F_{\underset{\sim}{k}} = \frac{c_k}{1 + |c_k|^2} \ .$$

For the more general relation between the two quantities see e.g. [7].

4.3 What Is Special About Quantum Condensates?

I believe that the feature of (simple) quantum condensates, whether BEC in a bosonic system such as superfluid 4-He or ultracold Bose atomic gases, Cooper pairing in Fermi systems such as superconductors or superfluid 3-He or other higher-n condensates to be possibly realized in the future, is that *a macroscopic number of the basic entities* (single particles, pairs...) *are behaving in exactly the same way*. This applies whether or not the system is in equilibrium (and indeed some of the more important manifestations of this behaviour refer to metastable or more generally nonequilibrium states). A nice intuitive representation of this feature is given in the cover illustration of the Dec. 22, 1995 issue of the journal Science,

where the Bose condensate, first realized in atomic gases that year and celebrated in the relevant issue as "molecule of the year", is represented as a platoon of soldiers identically uniformed and marching exactly in lockstep (while the "normal fraction" of uncondensed atoms are individually garbed and doing random things). This suggest the following consideration: imagine that you are on a mountain top and looking down at the main square of a city below, suppose first that it is an ordinary day, say a market day, and the citizens are each going about their own individual business; it will be very difficult to see from the distance what any one of them is doing. On the other hand, if it is the day of a military parade and the soldiers are behaving in the way depicted on the Science cover, it will be much easier to see what they are doing. A similar consideration holds in the case of quantum condensates: effects which at the level of a single atom or pair are too small to detect experimentally may become spectacularly amplified by the condensation and become readily visible.

A typical example of the above principle is the effect of the nuclear dipole-dipole interaction in liquid 3-He: for a single pair of atoms, even at the closest physically allowed separation, this is of order only 10^{-7} K, very much less than the thermal energy at any temperature yet attained for 3-He, and therefore in the normal phase, where the different pairs are each "doing their own thing", its effects are pretty much invisible. However, the moment the system undergoes the Cooper pairing transition, all the pairs are constrained to behave in the same way, and therefore the system has to choose between states in which they are all doing "right" from the point of view of the dipole-dipole interaction (i.e. realizing a configuration which minimizes it) or all behaving "wrong"; the energy difference between these two many-body states of the whole system is not 10^{-7} K but something like 10^{23} times this, many orders of magnitude larger than the thermal energy. This behaviour is not dissimilar from that of a Heisenberg ferromagnet in a weak magnetic field; in the paramagnetic phase above the Curie temperature T_c the different spins behave independently, and the Zeeman energy of any one spin in the field is much less than the thermal energy, so they point in random directions, while below T_c the exchange energy forces all spins to point in the same direction, so that the system has to decide between having them all point parallel to the field or all point in some other direction (e.g. antiparallel to it); the energy difference between these two many-body states is not the single-spin Zeeman energy; but N times this—a number enormous compared to kT. Consequently, in thermodynamic equilibrium a macroscopic fraction of the spins point along the weak field. For a more detailed account of the behaviour of the pairs in liquid 3-He and its consequences, see ref.[3, ch. 7].

Another illustration of the "amplification" effect can be found in the concept of the "superfluid velocity". Consider first a single spinless particle described by the time-dependent Schroedinger equation: for any given pure state of the particle we can define the probability density

$$\rho_{1\mathrm{P}}(\underset{\sim}{r}, t) = |\psi(\underset{\sim}{r}, t)|^2$$

and the current density

$$\underset{\sim}{j}_{1P}(\underset{\sim}{r}, t) = -\frac{i\hbar}{2m}\psi^*(\underset{\sim}{r}, t)\nabla\psi(\underset{\sim}{r}, t) + \text{c.c.}\,,$$

and in terms of these a quantity with the dimensions of velocity which we call $\underset{\sim}{v}_{1P}(\underset{\sim}{r}, t)$:

$$\underset{\sim}{v}_{1P} \equiv \underset{\sim}{j}_{1P}(\underset{\sim}{r}, t)/\rho_{1P} = \frac{\hbar}{m}\nabla\varphi_{1P}(\underset{\sim}{r}, t)\,, \tag{4.11}$$

where $\varphi_{1P}(\underset{\sim}{r}, t)$ is just the phase of the single-particle wavefunction. This quantity is irrotational, $\underset{\sim}{\nabla} \times \underset{\sim}{v}_{1P} = 0$, and satisfies the condition that its integral around any closed curve on which the wavefunction nowhere vanishes is quantised to nh/m, a condition which is at the root of the impossibility of reducing quantum mechanics to classical statistical mechanics; and it plays an important role in Bohm's interpretation of quantum mechanics. However, in most other interpretations the quantity $\underset{\sim}{v}_{1P}$ is not regarded as having much physical significance. The reason is that for a single particle any attempt to "measure $\underset{\sim}{v}_{1P}$" is liable to produce large fluctuations comparable to the mean value. E.g. if we consider the particle to be in the groundstate $\psi(x) = \sqrt{2/L}\sin(nx/L)$ of a 1D square well potential of length L, the "value" of $\underset{\sim}{v}_{1P}$ as defined by Eq. (4.11) is zero, but "measurements" of this quantity (which, remember, is just the momentum operator $-i\hbar\partial/\partial x$ divided by m) will, according to the standard rules of quantum mechanics, produce the eigenvalues $\pm h/mL$ at random. On the other hand, if we consider a many-body system and define a "hydrodynamic" velocity $\underset{\sim}{v}_h(\underset{\sim}{r}, t) \equiv \underset{\sim}{j}(\underset{\sim}{r}, t)/\rho(\underset{\sim}{r}, t)$ where $\underset{\sim}{j}$ is the total current density and ρ the density (the sums over different particles of expressions Eqs. (4.9a) and (4.9b) respectively) we find that while the fluctuations in its measured value are of order $1/N$, it does not satisfy the irrotationality condition and thus has lost the "essential quantum" nature of $\underset{\sim}{v}_{1P}$.

But now consider a set of structureless bosons like 4-He atoms, then in the quantum-condensed phase the OP $\Psi(\underset{\sim}{r}, t)$ is, like the single-particle Schroedinger wavefunction, a complex scalar object and we can define its phase $\varphi(\underset{\sim}{r})$ (modulo 2π). We then define $\underset{\sim}{v}_s$ by

$$\underset{\sim}{v}_s(\underset{\sim}{r}, t) \equiv \frac{\hbar}{m}\underset{\sim}{\nabla}\varphi(\underset{\sim}{r}, t)\,. \tag{4.12}$$

This gives us as it were the benefits of both worlds: $\underset{\sim}{v}_s$ characterises the behaviour of a macroscopic number of atoms and thus has fluctuations only of order $1/N$, while at the same time it is irrotational and satisfies the Onsager-Feynman quantisation rule

$$\oint \underset{\sim}{v}_s(\underset{\sim}{r}, t)\mathrm{d}\underset{\sim}{l} = \frac{nh}{m}\,,$$

where the integral is taken around any closed curve on which the OP does not vanish. Thus it may be regarded as a genuinely "macroscopic quantum" object.

It needs to be noted, briefly, that while the definition of a "superfluid velocity" is as outlined above very straightforward for a spinless Bose system, the question is a good deal more complicated for a system with an internal (spin, hyperfine or orbital) degree of freedom, as happens e.g. in superfluid 3-He and in the (Bose and Fermi) alkali gases; for some of these complications, see section 6.5 of ref.[3].

Various consequences of the "macroscopic amplification" property of quantum condensates for their transport and other properties are reviewed in section 2.5 of ref.[3]. Some of them (e.g. nonclassical rotational inertia) require the existence of a nonzero OP only in thermodynamic equilibrium, while others (e.g. the metastability of supercurrents) require its stability also under conditions far from equilibrium (and sometimes also that the OP be of a certain type). In the next section we consider the reasons for such stability.

4.4 Why Is Nature So Fond of "Simple" Quantum Condensation? Why Is "Fragmentation" So Rare?

It is needless to remark that not all many-body systems undergo quantum condensation even at the lowest temperatures to which they have been taken. Trivial counterexamples are most Bose elements, which at low temperature in thermodynamic equilibrium simply form crystalline solids, or electrons, which not only can form insulating solids but even when they continue to be mobile (metallic) may remain "normal" (nonsuperconducting) at the lowest temperatures examined. Nevertheless, the occurrence of quantum condensation is remarkably ubiquitous (most recently even close to room temperature, given sufficiently high pressures) and what is at first sight even more remarkable is that even under circumstances where one might at first sight think that fragmentation would be likely to occur, the "simple" version often seems to win out.

In the case of a spinless Bose system one factor which favours condensation is simply their indistinguishability. If we consider the textbook problem of distributing N distinguishable objects between m "unbiased" boxes, then the standard combinatoric argument leads to the conclusion that distributions in which a large fraction of the objects end up in one of the boxes are strongly disfavoured relative to ones in which the occupation is more uniform (there are many more different ways of getting the latter), and this result is not much affected by a small "bias" (e.g. an energetic one) favouring one or more of the boxes. However, once the distinguishability is removed, as for identical bosons, there is simply one and only one way of getting any state, uniform or not; so that condensation is no longer statistically disfavoured and can be induced by even a small bias factor. This rather general argument suggests that in a spinless Bose system the occurrence of BEC is

not likely to be confined to thermodynamic equilibrium but is more generic, at least given a sufficient energetic or other "bias".

However, one might at first think that this "statistical" effect is likely to be opposed by a different type of energetic consideration associated with the interparticle interactions, and in particular that interparticle repulsion might favour fragmentation. Oddly enough, at least in the dilute limit appropriate to the alkali gases, precisely the opposite is the case! To see this, we need to consider the result of making the standard Hartree-Fock approximation for the energy of two identical interacting particles, let us say in plane-wave states with a contact interaction. If they are in the same state, there is only a Hartree term and no Fock one. If on the other hand they are in different states, then in contrast to the better-known Fermi case the Fock term reinforces the Hartree one, so that the interaction is twice what it is for occupation of the same state. Thus a repulsive interaction will disfavour fragmentation, while an attractive one will favour it (but in fact tends to lead to collapse in position space). For quantitative details of the argument, see ref.[3] section 2.3. It indeed seems to be true that in most physical systems in which simple BEC is realized, and in particular in liquid 4-He, the interparticle interaction seems to be overall repulsive.

In the Fermi case there is no statistical argument in favour of condensation, and its occurrence is rather connected with the following consideration: it is straightforward to show that the contribution of the condensate to the interparticle potential is of the form

$$\langle V|V\rangle = \int \mathrm{d}\underset{\sim}{R}\,\mathrm{d}\underset{\sim}{\varrho}\,V(\underset{\sim}{\varrho})|F(\underset{\sim}{\varrho},\underset{\sim}{R})|^2$$

and thus if the interaction is overall attractive, as is believed to be the case in a superconductor, condensation does not cost too much in kinetic energy and (negative) entropy. It is likely to occur. While the BCS theory is of course the best known special case of this argument, it is again more general and suggests that the phenomenon of Cooper pairing is unlikely to be confined to the case of thermodynamic equilibrium.

However, there are cases where these arguments do not seem to be adequate to exclude fragmentation. Let's consider for example the theory of metastability of supercurrents in a spinless Bose-condensed liquid such as 4-He in an annular geometry. According to the Onsager-Feynman theorem, the condensate wave function is characterised by an integer "winding number" n such that the circulation (the integral of the superfluid velocity around the loop) is nh/m. If the container is at rest and the winding number is nonzero, this state cannot be the groundstate. Indeed, if we were to model the system as a non-interacting gas, we would find that the atoms of the condensate drop back individually into the $n = 0$ state; note that in this process a fragmented state would be automatically realised. That real liquid 4-He does not behave in this way is a consequence of the repulsive interactions which endow the nonzero-n state with topological (meta)stability; a straightforward argument (cf. ref.[3], section 2.5) indicates that there exists a free energy barrier

between the two states which must be overcome, if at all, by a statistical or possibly a quantum fluctuation; except very close to T_c these processes are rare, so the spectacular metastability of the circulating-current state is explained. A similar argument works in the superconducting case.

Nevertheless, we can still raise the question, which is well illustrated by these "rare" relaxation processes but is more general: suppose that we make a "natural" choice of the single-particle or two-particle basis, say the momentum eigenstates. Then it is clear that, unless the condensation vanishes altogether during the relaxation process, there must be some time interval during which both the initial state and at least one other state, e.g. the final state, must be macroscopically occupied. Now this can happen in one of at least two ways: (1) through "fragmented" condensation (1- or 2-particle reduced density matrix has more than one single-particle state macroscopically occupied), (2) simple condensation is maintained but is into a coherent quantum superposition of two single- or pair-states. We will denote these two possibilities as respectively (1), a "Fock" state, and (2), a "coherent" (or GP/GL) state. Why then is it uniformly assumed in the literature (e.g. in the seminal paper [8]) that the intermediate state is always a coherent rather than a Fock state?

The key to resolving this question is the observation that for any two given mutually orthogonal basis states χ_1, χ_2, if the relative occupations are N_1 and N_2 respectively with $N_1 + N_2$ equal to the condensate number N_0, the Fock state, which apart from normalisation, is

$$\Psi_{N,\mathrm{Fock}} \equiv (a_1^\dagger)^{N_1}(a_2^\dagger)^{N_2}\rangle\mathrm{vac}\ , \tag{4.13}$$

where $\rangle$vac denotes the vacuum state, can be resolved in the limit $N_1, N_2 \to \infty$ into a quantum superposition of coherent states with definite relative phase $\Delta\varphi$:

$$\Psi_{N,\mathrm{Fock}} = \frac{1}{2\pi}\int \mathrm{d}(\Delta\varphi)\Psi_{N,\mathrm{coh}}(\Delta\varphi)\ , \tag{4.14}$$

where

$$\Psi_{N,\mathrm{coh}}(\Delta\varphi) = \left(\alpha\, a_1^\dagger + \beta\exp(i(\Delta\varphi)a_2^\dagger\right)^N\rangle\mathrm{vac}, \tag{4.15}$$

with $|\alpha|^2 = N_1$, $|\beta|^2 = N_2$. The off-diagonal terms of the part of the reduced density matrix corresponding to the condensate is different for the Fock state (4.13) and for the specific coherent state (4.15):

$$\rho_{\mathrm{Fock},12} = \rho_{\mathrm{Fock},21} = 0 \tag{4.15a}$$

$$\rho_{\mathrm{coh},12} = \rho^*_{\mathrm{coh},21} = \sqrt{N_1 N_2}\exp(i(\Delta\varphi) \tag{4.15b}$$

Thus, if there is any single-particle pair potential which has nonzero matrix elements between the single-particle/pair states 1 and 2, there will always, except by pathology, be some value of $\Delta\varphi$ for which the corresponding coherent state wins out over the Fock state. Even if no such single-particle/pair potential exists, in the absence of pathology the two-particle terms will favour an appropriate coherent state (for details of the argument on this point, which applies only to the $n = 1$ case, see ref.[3], section 2.4). Note that this argument is independent of the sign of the single-particle/pair potential and is thus very general.

We thus conclude that in the absence of special factors, "simple" quantum condensation will always be favoured over the "fragmented" version. In the next section we will briefly note a few cases of such "special factors".

4.5 When Does Fragmentation Occur?

Perhaps the easiest case to visualise is when fragmentation arises in a superconducting system due to the phenomenon of so-called "Coulomb blockade". If we have two macroscopically large superconducting systems connected by a Josephson junction, then the Josephson coupling energy E_J is a special case of the "single-pair" potential discussed in the last subsection, and favours a coherent state. However, there is also a Coulomb potential E_C between the two macroscopic bodies which tends to favour a definite number of Cooper pairs on each body, i.e. a Fock state. Thus as we pass from the regime $E_C \langle\langle E_J$ to the opposite limit $E_C \gg E_J$ we pass continuously between "simple" and "fragmented" condensation; this situation has been much discussed by aficionados of mesoscopic superconductivity.

A second case in which fragmentation is less inhibited is that of either a Bose or a Fermi system with internal (e.g. hyperfine) degrees of freedom. It should be emphasized that not all of the arguments against it detailed in the last section apply in this case; e.g. if one is talking about fragmentation into two different hyperfine states with the same positional behaviour, the Fock term and hence the "factor of 2" may not arise. Nevertheless, fragmentation still seems to be the exception rather than the rule. A particularly interesting example where fragmentation appears to be not just allowed but necessary is that of a spin-1/2 Bose system (such as electronically spin-polarized atomic hydrogen with weak species-independent repulsion) which is cooled from high to low temperatures by processes which conserve the total spin. Since any uniform simply-condensed state must have a total spin equal to 1/2 times the total particle number N, while the high-temperature thermal state has a total spin very small compared to N, simple condensation is ruled out and the groundstate is believed [9] to be of the form (up to normalisation)

$$\Psi_N = \left(a^\dagger_{0,\uparrow} a^\dagger_{1,\downarrow} - a^\dagger_{0,\downarrow} a^\dagger_{1,\uparrow}\right)^{N/2} , \tag{4.16}$$

where 0 and 1 denote the single-particle groundstate and one of the lowest excited states (so that the many-body groundstate is degenerate). The state Eq. (4.16) is manifestly fragmented. Finally, while it may or may not, depending on the details of the definition, be an example of "fragmentation", a system in which simple condensation is forbidden to occur is a Bose system or superconductor in strictly 2 dimensions at $T \neq 0$ in this case the single particle density of states is too high for simple BEC to win out. Nevertheless, as shown by Berezinskii, Kosterlitz, Thouless and others, this system shows at low temperatures many of the characteristics of a system with condensation.

4.6 Alternative Approaches to Quantum Condensation: Some Problems

4.6.1 ODLRO

The main problems with the "ODLRO" definition of quantum condensation are rather trivial ones which were not reasonably foreseeable in the early sixties: in the first place, it cannot as it stands be easily generalized to cases like the ultracold alkali gases, since the latter are typically confined, by either magnetic or laser potentials, to a restricted region of space and thus it is meaningless to take the limit of infinite separation in the arguments of $\rho(\underset{\sim}{r}, \underset{\sim}{r}')$. Secondly this definition cannot be naturally extended to cope with the possibility of fragmentation.

4.6.2 Anomalous Averages

As to the definition (4.8b) (or (4.8a)) in terms of "anomalous averages", this is unambiguous at zero temperature, since e.g. many-body groundstates $\rangle N$ and $\rangle N-2$ are each uniquely defined, but at finite temperatures or out of thermodynamic equilibrium the question arises of the correct choice of the pair(s) of states between which the operator $\widehat{\psi}(\underset{\sim}{r}, \sigma)\widehat{\psi}(\underset{\sim}{r}', \sigma')$ has to be evaluated. Although there is usually an "intuitive" answer, the point is not entirely trivial. At first sight there is a related problem even with the zero-temperature equilibrium state: if for simplicity we consider the number-projected BCS model and a specific pair state $(\underset{\sim}{k}\uparrow, -\underset{\sim}{k}\downarrow)$, then while it is true that the overlap of the state $a_{-\underset{\sim}{k}\downarrow}a_{\underset{\sim}{k}\uparrow}$ on the $N-2$ particle groundstate is indeed the Fourier transform $F_{\underset{\sim}{k}}$ of $F(\underset{\sim}{r})$, this operator also has non-zero overlap on the "excited-pair" state in which the configuration of the pair is, in an obvious notation $\underset{\sim}{v}_{\underset{\sim}{k}}\rangle 00 - \underset{\sim}{u}_{\underset{\sim}{k}}\rangle 11$, and if we were to factor the expectation value of the four-particle operator

$$\langle a^{\dagger}_{\underset{\sim}{k}\uparrow}a^{\dagger}_{-\underset{\sim}{k}\downarrow}a_{-\underset{\sim}{k}\downarrow}a_{\underset{\sim}{k}\uparrow}|a^{\dagger}_{\underset{\sim}{k}\uparrow}a^{\dagger}_{-\underset{\sim}{k}\downarrow}a_{-\underset{\sim}{k}\downarrow}a_{\underset{\sim}{k}\uparrow}\rangle \ , \tag{4.17}$$

this will contribute a term of the same order as $|F_{\underset{\sim}{k}}|^2$. Why do the field-theoretic treatments following Gor'kov's original work never take account of this? I think the implicit argument is that as soon as one generalizes the argument to a 4-particle expectation value with the argument of the annihilation operators different from that of the creation operators it becomes irrelevant, since the many-body excited states are different for $\underset{\sim}{k}$ and $\underset{\sim}{k}'$. Since the number of terms of the form Eq. (4.8b) exceeds those of the form Eq. (4.17) by a factor of order N, we are presumably safe in ignoring the latter. Nevertheless, I believe that this rather subtle point may deserve more scrutiny than it usually gets in the literature.

4.6.3 "Spontaneously Broken U(1) Symmetry"

As remarked in Sect. 4.2, the use of this concept to define the order parameter requires the belief that the system is correctly described by a many-body wave function of the form Eq. (4.10), that is, as a quantum superposition of states with different values of the total particle number N (or by a classical mixture of such states). Now while the superposition of states of different *relative* particle number (e.g. in the two bulk superconductors connected by a Josephson junction) is both physically realistic and useful, I claim that Eq. (4.10) (or a classical mixture of states of the form Eq. (4.10)) is, as a description of a single system, say a single bulk superconductor, a physical nonsense: it violates the basic superselection rule of lepton (electron) number. This is obvious for a closed system, which cannot exchange electrons with its environment E. However, in the (very common) case of a system with leads it is perhaps not quite so immediately obvious. To see it, let us consider that the system S plus E constitutes a closed system, and consider the reduced single- (or two-) particle density matrix of S in the particle number basis N_S. The crucial observation is that while nothing precludes this density matrix from having more than one non-zero diagonal element, the conservation of N_{tot} implies that different values of N_S are correlated with orthogonal states of E, and thus the trace over E necessary to produce the reduced density matrix of S reduces the off-diagonal elements of $\widehat{\rho}_S$ to zero.

It is tempting to argue against this conclusion that the situation is really no different from that in say Heisenberg ferromagnetism, where again it can be argued that there is really no such thing as the "absolute" direction of magnetization of a single system (what do we have with which to compare it?); all we can really define is the relative direction of magnetisation of two different systems. In the ferromagnetic case we are nevertheless usually happy to talk about the direction as well as the magnitude of a single system, and if pressed would tend to justify it by invoking the notion of a "standard" of direction such as the direction of the magnetic field of a sufficiently large body such as the Earth. In the same way we could ask in e.g. the superconducting case whether it would be possible to set up a "phase standard", to be carefully preserved in e.g. the National Bureau

of Standards in Gaithersburg, Maryland, to which experimental groups could bring their superconductors to establish their supposedly "absolute" phases? This question raises a number of interesting issues, some of which are discussed in ref.[10] and earlier references cited therein, but I think one point to note is that there do not seem to be obvious analogs in the ferromagnetic case to some of the purposes for which the idea of the alleged "absolute phase" of a superconductor has been used in the literature. The most important of these is the treatment of the single-fermion degrees of freedom (Bogoliubov quasiparticles), and I now turn to this topic (which I note is of most significance in the Fermi case):

In the BCS (particle-nonconserving) theory the groundstate of a uniform superconductor containing an even number of electrons is written

$$\Psi_0 = \prod_k (u_k + v_k a^\dagger_{k\uparrow} a^\dagger_{-k\downarrow}) \rangle \mathrm{vac} \ , \tag{4.18}$$

where $\rangle$vac is the physical vacuum and the $a^\dagger$'s are free-particle creation operators. Written in the basis corresponding to the total particle number N, this wave function has the schematic form Eq. (4.10), namely

$$\Psi_0 = \sum_N c_N \Psi_N \ .$$

The elementary excitations are Bogoliubov quasiparticles defined by

$$\rangle \underline{k}, \sigma = \alpha^\dagger_{k,\sigma} \rangle \Psi_0 \ ,$$

where

$$\alpha^\dagger_{k\uparrow} = u_k a^\dagger_{k,\uparrow} + v_k a_{-k,\downarrow} \ . \tag{4.19}$$

It is clear that acting on a particular N-eigenstate the operator alpha produces a quantum superposition of states with $N + 1$ and $N - 1$ particles. The equation has been generalized by Bogoliubov and de Gennes to cover the spatially inhomogeneous case which will be of interest below: the single-fermion eigenstates are more generally of the form (for the traditional s-wave pairing)

$$\rangle \psi_{i,\uparrow} = \int \mathrm{d}\underline{r} \{ u_i(\underline{r}) \widehat{\psi}^\dagger_\uparrow(\underline{r}) + v_i(\underline{r}) \widehat{\psi}_\downarrow(\underline{r}) \} \rangle \Psi_0 \ , \tag{4.20}$$

where $u_i(\underline{r})$ and $v_i(\underline{r})$ satisfy the Bogoliubov-de Gennes equations (which we do not need to write out explicitly in the present context). At this point it is necessary to note that while it is certainly possible to use in reverse the "Anderson trick",

namely the expression of a number eigenstate Ψ_N in terms of an integral over phase eigenstates $\Psi(\varphi)$: up to normalization

$$\Psi_N = \int \mathrm{d}\varphi \exp(iN\varphi)\Psi(\varphi) \ , \tag{4.21}$$

if one does this to the BdG state it will still give a superposition of two states differing in their N-values by 2.

So far, so good: the BdG equations have been used in hundreds if not thousands of theoretical papers over the last 60-odd years, and while to the best of the present author's knowledge attempts to test them critically and explicitly against experiment have been few if any, there seems no particular reason to believe that they are unreliable in "ordinary" situations. However, over the past 20 years or so there has been a particular application of major interest, namely their application to certain types of superconductor (the so-called "topological" type[1]) to generate special types of fermion state, the so-called Majorana fermions, which are characterized by the equality

$$u(\underset{\sim}{r}) = v^*(\underset{\sim}{r}) \ , \tag{4.22}$$

a relation which, in analogy to the corresponding situation in particle physics, is often interpreted as indicating that "the Majorana fermion is its own antiparticle". These unusual particles are believed to behave in a very interesting way under "braiding" operations, and to potentially form a basis for topologically protected quantum computing; see e.g. ref. [11].

Over the past ten years or so, there have been repeated claims by experimental groups to have discovered Majorana fermions in various topological superconductors, but to the present author's knowledge none has yet received universal or even perhaps majority acceptance in the relevant community, and several of the more well-publicized ones have had to be withdrawn in the face of the failure of attempts to reproduce them. (For a rather scathing commentary on the standards currently being applied in this sub-field by an active worker in it, see [12]). I now propose a possible resolution of this problem, namely that the whole notion of Majorana fermions in topological superconductors may be a science fiction, a "unicorn" which exists only in the mind of the theorist.

Consider a closed superconducting system, and suppose that we insist on conserving total particle number N at all costs. Then the groundstate will be an eigenstate of N (by itself this does not affect the predictions much). What about the simplest fermionic excitations (which under this assumption must be produced in pairs)? They cannot be of the form Eq. (4.20), and the default modification is to leave the "particle" component (the first term in Eq. (4.20)) unchanged but to accompany

[1] Which actually needs a generalisation of Eq. (4.20) for a spin-1 condensate.

the "hole" component (the second term) by the creation of an extra Cooper pair:

$$\rangle\Psi_{i,\uparrow} = \int \mathrm{d}\underset{\sim}{r}\{u_i(\underset{\sim}{r})\widehat{\psi}^{\dagger}_{\uparrow}(\underset{\sim}{r}) + v_i(\underset{\sim}{r})\widehat{\psi}_{\downarrow}(\underset{\sim}{r})\widehat{C}^{\dagger}\}\,, \tag{4.23}$$

where the operator $\widehat{C}^{\dagger}$ creates an extra Cooper pair (which need not be exactly identical to the ones forming the N-particle groundstate, but should differ from them only by a factor $O(1/N)$). Now, provided that one is dealing with a "normal" situation involving, say, simple BCS s-wave pairing, the extra pair adds no extra quantities other than the required 2 to the particle number, and there seems no obvious reason why this consideration should modify any of the standard results of a calculation based on the BdG equations (the extra Cooper pair may be conceived as wandering off "to infinity").

However, the fundamental respect in which "topological" superconductors differ from the ordinary ("trivial") sort is precisely that the Cooper pairs carry an extra quantum number, typically angular momentum or some related property. Thus if one insists on describing the fermionic quasiparticles by Eq. (4.23) rather than Eq. (4.20), the second term (but not the first) modifies the many-body wavefunction in a non-trivial way. In particular, in the case of Majorana solutions, not only is the "particle" no longer its own antiparticle, but it is not at all clear that the accepted "braiding statistics" still hold.

It is of course entirely possible that by the time this paper is published one or more experimental groups will have produced evidence for Majorana fermions which is universally accepted in the relevant community. However, I would like to make an observation about the "sociology" of the theoretical community, which I have found to apply also to one or two other problems in condensed matter physics: suppose we have a situation in which one part of the theoretical community (the "enthusiasts") has proposed the existence of a certain novel phenomenon, while another group of theorists (the "sceptics") claims that the arguments used to support this proposal are unconvincing. There then arises an asymmetry: the enthusiasts have every motivation to submit papers developing the proposal to "high-impact" journals, and these papers will then be highly cited. On the other hand, the sceptics have little motivation to publish much: once one has given the reasons for one's scepticism, there is usually not much more to say. Thus, if as has happened in the Majorana case, experiment has at the time in question been unable to decide the issue, both the relevant experimental community, and novice theorists entering the field, are likely to get a distinctly lopsided impression of the degree of enthusiasm for the idea among theorists.

Be that as it may, there is no doubt that Prof. Yang's 1962 paper has played a major role in clarifying the conceptual basis of superconductivity and superfluidity. It is a pleasure to dedicate this essay to Professor C. N. Yang on the occasion of his 100th birthday, and to wish him yet more happy years of activity in physics and other things.

References

1. C.N. Yang, concept of off-diagonal long-range order and the quantum phases of liquid He and of superconductors. Rev. Mod. Phys. **34**, 694–704 (1962). https://doi.org/10.1103/RevModPhys.34.694
2. J. Bardeen, L.N. Cooper, J.R. Schrieffer, Phys. Rev. **108**, 1175–1204 (1957). https://doi.org/10.1103/PhysRev.108.1175
3. A.J. Leggett, *Quantum Liquids: Bose Condensation and Cooper Pairing in Condensed-Matter Systems* (OUP, Oxford, 2006)
4. O. Penrose, L. Onsager, Bose-Einstein condensation and liquid helium. Phys. Rev. **104**, 576–584 (1956). https://doi.org/10.1103/PhysRev.104.576
5. O. Penrose CXXXVI, On the quantum mechanics of helium II. Lond. Edinb. Dublin Philos. Mag. J. Sci. **42**, 1373–1377 (1951)
6. L.P. Gor'kov, Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki, **34**, 735 (1958). English translation, Soviet Phys. JETP **7**, 505 (1958)
7. V. Ambegaokar, R. Parks, *Superconductivity: Part 1 (In Two Parts)* (Taylor & Francis, London, 1969)
8. J. Langer, V. Ambegaokar, Intrinsic resistive transition in narrow superconducting channels. Phys. Rev. **164**, 498–510 (1967). https://link.aps.org/doi/10.1103/PhysRev.164.498
9. A. Kuklov, B. Svistunov, Ground states of SU(2)-symmetric confined Bose gas: quantum superposition of the phase-separated classical condensates. Phys. Rev. Lett. **89**, 170403 (2002). https://link.aps.org/doi/10.1103/PhysRevLett.89.170403
10. A.J. Leggett, Mod. Phys. Lett. **B14**(supplement), 1 (2000)
11. C. Nayak, S.H. Simon, A. Stern, M. Freedman, S. Das Sarma, Non-Abelian anyons and topological quantum computation. Rev. Mod. Phys. **80**, 1083–1159 (2008). https://doi.org/10.1103/RevModPhys.80.1083 [arXiv:0707.1889 [cond-mat.str-el]]
12. S. Frolov, Quantum computing's reproducibility crisis: Majorana fermions. Nature. **592**, 350–352 (2021)

Chapter 5
Quantum Operads



Noémie Combe, Yuri I. Manin, and Matilde Marcolli

Dedicated to C. N. Yang

Abstract The most standard description of symmetries of a mathematical structure produces *a group*. However, when the definition of this structure is motivated by physics, or information theory etc., the respective symmetry objects might become more sophisticated such as quasigroups, loops, quantum groups. In this paper, we introduce and study quantum symmetries of very general categorical structures: *operads*. Its initial motivation were spaces of probability distributions on finite sets. We also investigate here how structures of quantum information, such as quantum states and some constructions of quantum codes, are algebras over operads.

Keywords Information spaces · Moufang loops · Monoidal categories · Operads · Monoids · Magmas

... esta selva selvaggia e aspra e forte
che nel pensier rinova la paura ...
Dante Alighieri, Inferno, Canto 1

N. Combe
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Y. I. Manin (✉)
Max Planck Institute for Mathematics, Bonn, Germany
e-mail: manin@mpim-bonn.mpg.de

M. Marcolli
Mathematics Department, Caltech, Pasadena, CA, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_5

5.1 Introduction and Brief Survey

The common definition of symmetries of a structure given on a set S (in the sense of Bourbaki) is *the group of bijective maps $S \to S$ compatible with this structure*.

But in fact, symmetries of various structures related to storing and transmitting information such as *information spaces* are naturally embodied in various classes of *loops* such as *Moufang loops*, non-associative analogs of groups (cf. [4]).

Here are some representative examples: quasigroups, loops, and Moufang loops.

A *quasigroup* is a set $\mathcal{L}$ together with binary composition law

$$* : \mathcal{L} \times \mathcal{L} \to \mathcal{L} : \quad (x_1, x_2) \mapsto x_1 * x_2 =: x_3, \tag{5.1}$$

such that any two elements among $\{x_1, x_2, x_3\}$ uniquely determine the third one.

A *loop* is a quasigroup with two-sided identity $e \in \mathcal{L}$: it means that $e * x = x * e = x$ for any $x \in \mathcal{L}$.

Finally, a *Moufang loop* is a quasigroup whose compostion law satisfies the "near-associativity" condition

$$(x_1 * x_2) * (x_3 * x_4) = x_1 * ((x_2 * x_3) * x_4). \tag{5.2}$$

The idea of symmetry embodied in a group is closely related to classical physics, in a very definite sense, going back at least to Archimedes. When quantum physics started to replace classical, it turned out that classical symmetries must also be replaced by their quantum versions. For a short history of this evolution, see pp. 1–4 of [7] and also [6, 8, 9]. As a result, the mathematical theory of *quantum groups* emerged.

In this paper we suggest to apply the formalism of quantisation on the operadic level [2] to symmetries of information spaces. The motivation of our use of adjective "quantum" in [2] was sometimes too intuitive, but the tools developed in [13], furnish a very precise and well axiomatized framework for this.

The general conception of "Quantum Operad" introduced and studied here was also inspired by the introduction of quantum error-correcting codes with a Moufang loop action: see [4], Sec. 6.

We are pleased to dedicate our paper to C. N. Yang, who led the breakthrough studies of gauge symmetries in quantum field theory.

5.2 Quantum Structures in Symmetric Monoidal Categories

5.2.1 Monoidal (=Tensor) Categories V ([13], Sec. 2.2, 2.3)

Data: multiplication $\otimes$ of objects, with identity object $\mathbf{1}$ and natural isomorphisms

$$\alpha_{A,B;C} : (A \otimes B) \otimes C \to A \otimes (B \otimes C), \quad \rho_A : A \otimes \mathbf{1} \to A, \quad \lambda_A : \mathbf{1} \otimes A \to A.$$

5.2.2 Symmetric Monoidal Categories

Additional *twist* isomorphisms $\tau_{A,B} : A \otimes B \to B \otimes A$, with $\tau_{A,B}\tau_{B,A} = \mathrm{id}_{A\otimes B}$, plus many commutative diagrams.

5.2.3 Magmas, Comagmas, Bimagmas, Associativity and Commutativity for (co, bi)magmas in Symmetric Monoidal Categories ([13], Sec. 2.4)

Basic data for a magma: an object A with multiplication morphism $\nabla : A \otimes A \to A$.

Basic data for a comagma: an object A with comultiplication morhism $\Delta : A \to A \otimes A$.

Basic data for a bimagma: a triple (A, ∇, Δ) as above such that the "bimagma diagram" (2.4) ([13], p. 49) commutes.

(Co, bi)-magmas in a symmetric monoidal category $\boldsymbol{V}$ are themselves objects of respective categories. Morphisms in them are those morphisms in $\boldsymbol{V}$, which are compatible with respective basic data.

Unitality and counitality structures for a magma (A, ∇) (resp. comagma (A, Δ)) are respectively the morphisms $\eta : \mathbf{1} \to A$ or $\varepsilon : A \to \mathbf{1}$ subject to additional restrictions.

5.2.4 Monoids, Comonoids, Bimonoids, and Hopf Algebras in Symmetric MonoIdal Categories ([13], Def. 2.7)

They are essentially (co, bi)magmas with additional (co,bi)associativity restrictions.

5.2.5 Quantum Quasigroups ([13], Sec. 3.1)

A quantum quasigroup (A, ∇, Δ) is a bimagma, for which both *left composite* and *right composite* morphisms are invertible:

$$A \otimes A \xrightarrow{\Delta\otimes \mathrm{id}_A} A \otimes A \otimes A \xrightarrow{\mathrm{id}_A\otimes\nabla} A \otimes A \ ,$$

$$A \otimes A \xrightarrow{\mathrm{id}_A\otimes\Delta} A \otimes A \otimes A \xrightarrow{\nabla\otimes \mathrm{id}_A} A \otimes A \ .$$

These morphisms are sometimes called *fusion operators* or *Galois operators*.

5.2.6 Quantum Loops

A quantum loop in $\boldsymbol{V}$ is a biunital bigmagma $(A, \nabla, \Delta, \eta, \varepsilon)$ such that (A, ∇, Δ) is a quantum quasigroup.

5.2.7 Functoriality ([13], Prop. 3.4)

Any symmetric monoidal functor

$$F : (\boldsymbol{V}, \otimes, \mathbf{1}_{\boldsymbol{V}}) \to (\boldsymbol{W}, \otimes, \mathbf{1}_{\boldsymbol{W}})$$

sends quantum quasigroups (resp. quantum loops) in $\boldsymbol{V}$ to quantum quasigroups (resp. quantum loops) in $\boldsymbol{W}$.

5.2.8 Magmas etc. in the Categories of Sets with Direct Product

According to [13], beginning of Sec. 3.3, in such categories comultiplication in a counital comagma is always the respective diagonal embedding. As a corollary, we see that quantum loops and counital quantum quasigroups in such caregories are cocommutative and coassociative.

As a result, we see, that in such a category counital quantum quasigroups are equivalent to classical quasigroups, and quantum loops are equivalent to classical loops ([13], Prop. 3.11).

5.3 Monoidal Categories of Operads

5.3.1 Graphs and Their Categories

Our basic definition of graphs as quadruples (F, V, ∂, j) and their categories is explained in [2], Sec. 1.1, p. 251. There F, resp. V, are called the sets of *flags*, resp. *vertices*, and structure maps ∂, resp. j are called *boundary maps*, resp. *involutions*. Usually one flag is a pair consisting of flag as such, and a *label*, that should be defined separately. *Geometric realization* of a graph is the quotient set of the disjoint union of semi-intervals $(0, 1/2]$ labeled with flags of this graph, modulo equivalence relation, in which 0-points of a flag is glued to $1/2$ of another flag, if these flags are related by the boundary relation, or structure involution.

Depending on the context and/or type of labelling of τ, elements of F_τ might be called *flags, leaves, tails* ... In the study of magmatic operad [3] and the relevant

binary trees, vertices of the relevant corollas are called *nodes*, non-root flags are called *left child, right child* etc. We will try to attach all such "heteronyms" to our basic terminology of [2].

Below the most typical labeling of our graphs will be (see details in [2], Sec. 1.3.2 a) and 1.3.2 e), pp. 257–259):

(i) Orientation.
(ii) Cyclic labeling.

To give an orientation and cyclic labeling of corolla is essentially the same as to define it as a *planar* graph: corolla, embedded into an oriented real affine plane, with labeling compatible with its orientation.

Graphs endowed with various labelings form categories, upon which the operation of disjoint union $\sqcup$ defines a monoidal structure: see [2], Sec. 1.2.4, pp. 254–255. Our central objects of study are initially defined only for connected graphs. Therefore, introducing this monoidal product, we must first take care of "empty" (or partially empty) graphs and explain details of their functoriality. The paper [2] is interspersed with subsections directly or indirectly motivated by this necessity.

For the purposes of this paper, the most important graphs are labelled *trees* and *forests*—disjoint unions of trees, forming *"selva selvaggia e aspra e forte"*.

5.3.2 Operads and Categories of Operads (See [2], Sec. 1.6, p. 262)

We recall here the first definition of operads in [2], 1.6 (I), and morphisms of operads as in [2], Sec. 1.6.1.

First of all, we fix a symmetric monoidal category of labelled graphs Γ with disjoint union as the monoidal structure, and a symmetric monoidal *ground category* $(\mathcal{G}, \otimes)$, satisfying a part of conditions 1.4a–f in [2], p. 259.

(i) *An operad is a tensor functor between two monoidal categories* $A : (\Gamma, \sqcup) \to (\mathcal{G}, \otimes)$ *that sends any grafting morphism to an isomorphism.*
(ii) *A morphism between two operads is a functor morphism.*

Denote this category of operads by $\Gamma\mathcal{GOPER}$.

5.3.3 Operads and Collections as Symmetric Monoidal Categories

Following [2], Sec. 1.8, we will introduce now the monoidal "white product" of two operads $A, B : (\Gamma, \sqcup) \to (\mathcal{G}, \otimes)$ by the formula

$$A \circ B(\sigma) := A(\sigma) \otimes B(\sigma)$$

extended to morphisms in a straightforward way.

Clearly, $(\Gamma\mathcal{G}OPER, \circ)$ is a symmetric monoidal category.

An important related notion is that of *collection*. Starting with Γ as above, denote by ΓCOR its subcategory, whose objects are corollas in Γ, and morphisms between them are isomorphisms.

Combining it with the ground category $(\mathcal{G}, \otimes)$ as above, we can introduce the category $\Gamma\mathcal{G}COLL$ of ΓG-collections: its objects are functors $A_1 : \Gamma COR \to \mathcal{G}$, and morphisms are natural transformations between these functors.

The restriction of white product $\circ$ to $\Gamma\mathcal{G}COLL$ defines on it the structure of symmetric monoidal category. If $(\mathcal{G}, \otimes)$ has an identity object $\mathbf{1}$, then the collection $\mathbf{1}_{coll}$ sending each corolla to $\mathbf{1}$ and each isomorphism of corollas to the identical isomorphism of $\mathbf{1}$, is the identity collection.

5.3.4 Operads as Monoids

We briefly describe here a construction by B. Vallette [14], reproduced in [2], Appendix A, Subsection 5.

We will have to use here a stronger labeling of graphs in Γ than just orientation. Besides orientation, connected objects of Γ must admit a continuous real-valued function such that it decreases whenever one moves in the direction of orientation along each flag. Such graphs are called *directed ones* (see [2], Sec. 1.3.2 b).

A graph τ is called a *two-level graph*, if it is oriented, and if there exists a partition of its vertices $V_\tau = V_\tau^1 \sqcup V_\tau^2$ with the following properties:

(i) *Tails at* V_τ^1 *are all inputs of* τ, *and tails at* V_τ^2 *are all outputs of* τ.
(ii) *All edges in* E_τ *go from* V_τ^1 *to* V_τ^2.

For any two ΓG-collections A^1, A^2 define their product as

$$(A^2 \boxtimes_c A^1)(\sigma) := \mathrm{colim}(\otimes_{v \in V_\tau^1} A^1(\tau_v)) \otimes (\otimes_{v \in V_\tau^2} A^2(\tau_v)).$$

Here colim is taken over the category of morphisms from two level graphs to σ.

Theorem 5.1 *The product* $\boxtimes_c$ *is a monoidal structure on collections, and operads are monoids in the respective monoidal category.*

5.3.4.1 Freely Generated Operads

For any $\Gamma\mathcal{G}$-collection A_1 one can define another collection $\mathcal{F}(A_1)$ together with a canonical structure of operad on it, and for any operad A each morphism of collections $A_1 \to A$ extends to a morphism of operads $f_A : \mathcal{F}(A_1) \to A$.

We can imagine $\mathcal{F}(A_1)$ as the operad freely generated by the collection A_1.

5.3.5 Comonoids in Operadic Setup

We will now introduce a category OP of operads given *together with their presentations* ([2], Sec. 2.4). We start with Γ and $\mathcal{G}$ as above.

One object of OP is a family (A, A_1, i_A), where A is a $\Gamma\mathcal{G}$-operad, A_1 is a $\Gamma\mathcal{G}$-collection, such that $f_A : \mathcal{F}(A_1) \to A$ is surjective.

Define on OP a product $\odot$ by the formula

$$(A, A_1, i_A) \odot (B, B_1, i_B) = (C, C_1, i_C),$$

in which $C_1 := A_1 \circ B_1$ (cf. 2.3 above), C:= the minimal suboperad, containing the image $(i_A \circ i_B)(A_1 \circ B_1) \subset A \circ B$, and i_C is the restriction of $I_A \circ i_B$ on $A_1 \circ B_1$.

Theorem 5.2 (See [2], Sec. 2.4)

(i) The product $\odot$ defines on OP a structure of symmetric monoidal category.
(ii) The category OP is endowed with the functor of inner cohomomorphisms

$$\underline{cohom}_{OP} : OP^{op} \times OP \to OP$$

so that we can identify, functorially with respect to all arguments,

$$\mathrm{Hom}_{OP}(A, C \odot B) = \mathrm{Hom}_{OP}(\underline{cohom}_{OP}(A, B), C)$$

(iii) Therefore, one can define canonical coassociative comultiplication morphisms

$$\Delta_{A,B,C} : \underline{cohom}_{OP}(A, C) \to \underline{cohom}_{OP}(A, B) \odot \underline{cohom}_{OP}(B, C).$$

Corollary 5.1 *For any A, the coendomorphism operad*

$$\underline{coend}_{OP} A := \underline{cohom}_{OP}(A, A)$$

is a comagma in the sense of 5.2.3 above.

5.3.6 The Magmatic Operad (See [3])

Below we give a brief survey of some definitions and results from [3], sometimes slightly changing terminology and notation.

Here objects of our basic symmetric monoidal category $(\Gamma, \sqcup)$ will be disjoint unions of oriented trees with the following additional labeling: *for each tree, its outgoing flags (or leaves) are cyclically ordered*. Corollas in it are one-vertex graphs

with one root and at least two leaves. Connected objects can be obtained from a union of disjoint corollas by grafting each root of a corolla to one of leaves of another corolla. Morphisms are compatible with labeling.

An algebra over magmatic operad is a family $(\mathscr{A}, *)$ consisting of a set $\mathscr{A}$ with binary composition law $* : \mathscr{A} \times \mathscr{A} \to \mathscr{A}$.

Thus, corollas in the magmatic category correspond to products

$$(x_1 * ((x_2) * \ldots (...(x_n)))...),$$

and generally, connected graphs in it correspond to monomials of generic arguments with all possible arrangements of brackets.

5.3.7 Quasigroup Monomials and Planar Trees

Monomials that can be obtained by iteration of binary multiplication $*$ as in (5.1) correspond to *planar* trees: see 5.3.1 above for discussion of planar corollas. Below, discussing quasigroups in general, and Moufang loops in particular, we will consider connected planar trees and quasigroup monomials as encoding each other in this way.

5.4 Moufang Loops and Operads

5.4.1 Moufang Monomials and Their Encoding by Labeled Graphs

We will start with comparing mathematical structures of two types: *labeled graphs*, and *Moufang monomials*.

The words *loop monomials* will refer to the following class of objects. Let $(\mathcal{L}, *)$ be a Moufang loop in the sense of [4], Definition 5.1.1. Let $(x_1, \ldots, x_n) \in \mathcal{L}$. We can produce new elements of $\mathcal{L}$ from this sequence by applying to them iterated multiplication $*$.

The basic examples are

$$x_1 * x_2, \tag{5.3}$$

$$(x_1 * x_2) * (x_3 * x_4). \tag{5.4}$$

We will encode the monomial (5.3) by a *cyclically labeled oriented corolla* with one vertex and three flags, exactly one of which is the *output*. The bridge from (5.3) to this corolla might be imagined as an enrichment of it by additional labeling: $x_1 * x_2$

at the output, and x_1, x_2 at two other flags, such that $(x_1, x_2, x_1 * x_2)$ corresponds to the given cyclic labeling.

Since (5.4) can be obtained from (5.3) by iteration and variables change, we must explain, how such an iteration is encoded on the level of labeled graphs. The answer is obvious: it corresponds to *graftings* of certain outputs to certain inputs, so that these outputs in the enriched picture become the inputs of the respective iteration.

In this way, (5.4) becomes encoded by an *oriented and cyclically labeled tree*, with four ordered inputs, two edges, three vertices, and one output.

5.4.2 Passage to Moufang Operad: Basic Identity

According to [4], Def. 5.1.1, the Moufang loops are defined as structures $(\mathcal{L}, *)$ satisfying the "near-associativity" relations

$$(x_1 * x_2) * (x_3 * x_4) = x_1 * ((x_2 * x_3) * x_4) \tag{5.5}$$

The r.h.s. of (5.5) is, in turn, encoded by an oriented and cyclically labeled tree, with four ordered inputs (the same ones as in (5.4)), two edges, and four vertices.

5.4.3 Moufang Collections (See [2], Sec. 1.5, pp. 259–261)

Call a *Moufang corolla* an oriented cyclically ordered connected graph with one output, and morphims are *isomorphisms* between them. They are objects of category $\mathcal{MC}\,O\,R$ (particular case of categories $\Gamma C\,O\,R$ above). Clearly, $\mathcal{MC}\,O\,R$ is a groupoid.

Choose a symmetric monoidal ground category $(\mathcal{G}, \otimes)$, and define respective Moufang collections.

5.4.4 Latin Square Designs and Their Encoding by Graphs

Let $\mathcal{D} = (P, L)$ be a Latin square design as in [4], Def. 6.8.2.1.

Denote by $G^0(\mathcal{D})$ the graph, defined by the following family of data (see [2], p.251):

Vertices $V_{G^0(\mathcal{D})}$ are lines of $\mathcal{D}$:

$$V_{G^0(\mathcal{D})} := L.$$

Flags $F_{G^0(\mathcal{D})}$ are pairs $(p, l) \in P \times L$ such that $p \in l$.

The boundary map $\partial_{G^0(\mathcal{D})} : F_{G^0(\mathcal{D})} \to V_{G^0(\mathcal{D})}$ sends each (p, l) to l.

The involution $j_{G^0(\mathcal{D})} : F_{G^0(\mathcal{D})} \to F_{G^0(\mathcal{D})}$ sends (p, l) to (p', l'), if $p \neq p'$ and $l = l'$.

5.4.4.1 Simplest Examples

Using notation from [4], Def. 6.8.2.1, we see that three simplest examples correspond to cases $N := \operatorname{card} L = 0, 1, 2$.

The case $N = 0$ is degenerate: the respective designs and graphs are empty, and usually are included in consideration only for categorical reasons.

The case $N = 1$ produces a *corolla*: the graph with one vertex and three flags, and boundary map sending each flag to this vertex. The involution map is identical one.

5.4.5 From Loops to Latin Square Designs

Consider an ML $\mathcal{L}$. Produce three labelled copies of $\mathcal{L}$: $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$, with pairwise empty intersections. Define the design (P, L) by putting $P := \mathcal{L}_1 \sqcup \mathcal{L}_2 \sqcup \mathcal{L}_3$ and defining a line as such triple of points (x_1, x_2, x_3), $x_i \in \mathcal{L}_i$ that $(x_1 * x_2) * x_3 = 1 \in \mathcal{L}$. In this last formula we implicitly forget labels 1, 2, 3 and consider Moufang multiplication in $\mathcal{L}$.

5.5 Operadic Structures on Quantum States

In this section we show that the operadic structures associated to classical probabilities on finite sets, introduced in [10], extend to *non-unital operads on quantum states*.

5.5.1 Operads of Classical and Quantum Probabilities

We first recall the main operadic structures of classical probabilities, as introduced in [10], which we follow for the exposition in this subsection. It was observed in [1] and [10] that classical probabilities on finite sets come endowed with an operad structure that describes the combination of independent subsystems.

Namely, denote by $\mathcal{P}$ the operad in $Sets$ with objects $\mathcal{P}(n) = \Delta_n$, the simplex of probabilities on the finite set $\{1, \ldots, n\}$,

$$\Delta_n = \{P = (p_i)_{i=1}^n \mid p_i \geq 0,\ \sum_{i=1}^n p_i = 1\},$$

and with composition operations

$$\gamma : \mathcal{P}(n) \times \mathcal{P}(k_1) \times \cdots \times \mathcal{P}(k_n) \to \mathcal{P}(k_1 + \cdots + k_n)$$

given by the composition of probabilities of independent subsystems

$$\gamma(P; P_1, \ldots, P_n) = (p_r p_{r,j_r})_{r=1,\ldots,n,\, j_r=1,\ldots,k_r}$$

for $P_r = (p_{r,j_r})_{j_r=1}^{k_r}$ and $P = (p_r)_{r=1}^n$.

5.5.1.1 Averages as an Algebra Over the Operad $\mathcal{P}$

Making explicit in this setup the general definitions from Sects. 5.3.2–5.3.4 above, we see that an algebra A over the operad $\mathcal{P}$ in a symmetric monoidal category is a family of morphisms

$$\alpha : \mathcal{P}(n) \otimes A^{\otimes n} \to A,$$

satisfying associativity and unitality conditions, and compatibility with the symmetric groups actions.

The set of non-negative real numbers $\boldsymbol{R}_+$ can be seen as a category with a single object and morphisms $x \in \boldsymbol{R}_+$ and as an algebra (in the category of small categories) over the operad $\mathcal{P}$ with the simple operations

$$\alpha(P; x_1, \ldots, x_n) = \sum_i p_i x_i,$$

for $P = (p_i)_{i=1}^n$ and $x_i \in \boldsymbol{R}_+$.

5.5.1.2 A_∞-Operad and Entropy

As was shown in [10], there is another operadic structure in the setting of classical probability distributions over finite sets.

Let $\mathcal{T}$ be the A_∞-operad of planar rooted tress. We say that a collection of n-ary information measures S_n for $n \in \boldsymbol{N}$, satisfies the *coherence condition*, if for any $n > m$, the n-th entropy functional S_n agrees with S_m, when $n - m$ of the variables

are vanishing. In other words, assume that among the probabilities $(p_1, \ldots, p_n)$ the only non-zero ones are $(p_{i_1}, \ldots, p_{i_m})$ for some $i_1 < i_2 < \cdots < i_m$. Then coherence condition means that

$$S_n(p_1, \ldots, p_n) = S_m(p_{i_1}, \ldots, p_{i_m}).$$

We can now determine on $\boldsymbol{R}_+$ the structure of algebra over the operad $\mathcal{T}$ with the operations

$$\alpha(\tau, x_1, \ldots, x_n) = \min \left\{ \sum_{i=1}^{n} p_i x_i - \frac{1}{\beta} S_\tau(p_1, \ldots, p_n) \mid P = (p_i) \in \Delta_n \right\},$$

where $\tau \in \mathcal{T}(n)$ is a planar rooted tree with n leaves, and $\beta > 0$ is a thermodynamic parameter (inverse temperature). The n-ary entropy functional $S_\tau(p_1, \ldots, p_n)$ associated to the tree τ is uniquely determined by the branching structure of the tree τ and the collection of coherent entropies S_n.

5.5.2 *Classical Probabilities from Quantum States*

Let $\mathcal{M}^{(n)}$ denote the convex set of $n \times n$-density matrices (quantum states),

$$\mathcal{M}^{(n)} = \{\rho \in M_{n\times n}(\boldsymbol{C}) \mid \rho^* = \rho,\ \rho \geq 0,\ \mathrm{Tr}(\rho) = 1\}.$$

Positivity condition means here, that $\rho = a^* a$ for some $a \in M_{n\times n}(\boldsymbol{C})$, hence $Spec(\rho) \subset \boldsymbol{R}_+$. In a fixed basis, the diagonal density matrices form a copy of the simplex Δ_n embedded in $\mathcal{M}^{(n)}$.

There are two classical probability distributions naturally associated to a quantum state as follows.

Definition 5.1 Given $\rho \in \mathcal{M}^{(n)}$, let

$$\Lambda = (\lambda_i)_{i=1}^n \quad \text{with} \quad \lambda_i \in Spec(\rho),$$

be the set of eigenvalues of ρ, sorted in non-increasing order, and let

$$P = (p_i)_{i=1}^N \quad \text{with} \quad p_i = \rho_{ii}$$

be the list of the diagonal entries of ρ.

Definition 5.2 Given two non-increasing sequences $A = \{a_1, \ldots, a_n\}$ and $C = \{c_1, \ldots, c_n\}$ with $\sum_{i=1}^N a_i = \sum_{i=1}^N c_i$, one says that A majorises C, or $A \succ C$, if for all $1 \leq k \leq N$ one has $\sum_{i=1}^k a_i \geq \sum_{i=1}^k c_i$.

Lemma 5.1 *The Shannon information of the two classical probability distributions* $\Lambda = \Lambda(\rho)$ *and* $P = P(\rho)$ *is related by* $S(P) \geq S(\Lambda)$.

Proof By Schur lemma, the sequence Λ of eigenvalues of the hermitian matrix ρ majorises the sequence P of its diagonal entries, when both are sorted in non-increasing order. It is well known that for probabilities $\Lambda \succ P$ is equivalent to the existence a bistochastic matrix B such that $P = B\Lambda$. The Shannon entropy is monotonically non-decreasing under bistochastic matrices, so $S(B\Lambda) \geq S(\Lambda)$.

□

The eigenvalues probability $\Lambda = \Lambda(\rho)$ determines the information content of the quantum probability ρ, since in the von Neumann entropy

$$S(\rho) = \mathrm{Tr}(\rho \log \rho)$$

the term $\log \rho$ is defined via the spectral theorem, so that we have

$$S(\rho) = S(\Lambda) = -\sum_i \lambda_i \log \lambda_i,$$

the Shannon entropy of the classical probability Λ.

We will show in the next subsections that these two classical probabilities $P(\rho)$ and $\Lambda(\rho)$ determine two non-unital operad structures on the space of quantum states. The operad obtained using $P(\rho)$ has better properties and directly agrees with the operad of classical probabilities recalled in Sect. 5.5.1 above when restricted to $\Delta_n \subset \mathcal{M}^{(n)}$.

5.5.3 Non-unital Operads

In the unital case, one can equivalently describe an operad O through the composition laws

$$\gamma : O(n) \otimes O(k_1) \otimes \cdots \otimes O(k_n) \to O(k_1 + \cdots + k_n),$$

with the associativity conditions (and the symmetricity and unitality conditions in the respective cases), or else one can describe O through insertion operations

$$\circ_i : O(n) \otimes O(m) \to O(n + m - 1).$$

For $1 \leq j \leq a$ and $b, c \geq 0$, with $X \in O(a)$, $Y \in O(b)$, and $Z \in O(c)$, these insertions are subject to the conditions

$$(X \circ_j Y) \circ_i Z = \begin{cases} (X \circ_i Z) \circ_{j+c-1} Y & 1 \leq i < j \\ X \circ_j (Y \circ_{i-j+1} Z) & j \leq i < b + j \\ (X \circ_{i-b+1} Z) \circ_j Y & j + b \leq i \leq a + b - 1. \end{cases}$$

The composition laws γ satisfying the associativity condition can be obtained from the insertions $\circ_i$ through

$$\gamma(X, Y_1, \ldots, Y_n) = (\cdots (X \circ_n Y_n) \circ_{n-1} Y_{n-1}) \cdots \circ_1 Y_1).$$

While these two descriptions of operads are equivalent in the unital case, they give rise to two different versions of the notion of non-unital operad, see the discussion in [11]. Indeed, non-unital operads defined through the operations $\circ_i$ are also non-unital operads with the composition operations γ, but the converse no longer holds, so the first class of non-unital operads is more restrictive.

We will show below that the non-unital operads of quantum states belong to the more restrictive class, according to the stronger notion of non-unital operad as in [11].

5.5.4 The Q_P-Operad of Quantum States

We show here that the operad $\mathcal{P}$ of classical probabilities on finite sets extends to a compatible but non-unital operad Q_P on quantum states.

Definition 5.3 For $n \geq 1$ denote by $Q_P(n) = \mathcal{M}^{(n)}$ the convex set of density matrices, endowed with the composition laws

$$\gamma_P : Q_P(n) \times Q_P(k_1) \times \cdots \times Q_P(k_n) \to Q_P(k_1 + \cdots + k_n) :$$

$$\gamma_P(\rho; \rho_1, \ldots, \rho_n) = \gamma(P(\rho); \rho_1, \ldots, \rho_n) = \begin{pmatrix} p_1\rho_1 & & & \\ & p_2\rho_2 & & \\ \vdots & & \cdots & \vdots \\ & & & p_n\rho_n \end{pmatrix}.$$

Lemma 5.2 *The action of the symmetric group Σ_n on $\mathcal{M}^{(n)}$ given by $\sigma(\rho) = \sigma\rho\sigma^*$ is compatible with the action by permutation of the coordinates on classical probabilities. It acts on the two probability distributions $P(\rho)$ and $\Lambda(\rho)$ by*

$$P(\sigma\rho\sigma^*) = \sigma^* P(\rho) \quad \text{and} \quad \Lambda(\sigma\rho\sigma^*) = \Lambda(\rho).$$

Proof By realising the set of classical probability distributions $\Delta_n \subset \mathcal{M}^{(n)}$ as set of diagonal density matrices in a chosen basis, we see that $\sigma\rho\sigma^*$ permutes the entries by σ^*. The diagonal entries of ρ can be obtained as $\rho_{ii} = \mathrm{Tr}(\pi_i\rho)$ with π_i the i-th 1-dimensional projection in the chosen basis, and $\mathrm{Tr}(\pi_i\sigma\rho\sigma^*) = \mathrm{Tr}(\sigma^*\pi_i\sigma\rho) = \mathrm{Tr}(\pi_{\sigma^{-1}(i)}\rho) = \rho_{\sigma^{-1}(i)\sigma^{-1}(i)}$. So we have $P(\sigma\rho\sigma^*) = \sigma^* P(\rho)$. In the case of $\Lambda(\rho)$, since this distribution is defined after choosing an order in which to list the eigenvalues of ρ, such as non-increasing order, we have $\Lambda(\rho) = \Lambda(\sigma\rho\sigma^*)$, since both matrices have the same spectrum. □

Proposition 5.1 *The convex sets $Q_P(n)$ with the composition operations γ_P of Definition 5.3 determine a non-unital symmetric operad Q_P that restricts to the unital operad $\mathcal{P}$ on classical probabilities $\Delta_n \subset \mathcal{M}^{(n)}$.*

Proof It is clear by the definition of the composition operations γ that they agree with the composition operations of the operad $\mathcal{P}$ when restricted to classical probabilities $\Delta_n \subset \mathcal{M}^{(n)}$. We need to check that they satisfy the associativity and symmetry axioms on the larger set $\mathcal{M}^{(n)}$ of quantum states.

The associativity condition is given by the identities

$$\gamma(\gamma(\rho^{(m)}; \rho^{(n_1)}, \dots, \rho^{(n_m)}); \rho^{(r_{1,1})}, \dots, \rho^{(r_{1,n_1})}, \dots, \rho^{(r_{m,1})}, \dots, \rho^{(r_{m,n_m})}) =$$

$$\gamma(\rho^{(m)}; \gamma(\rho^{(n_1)}; \rho^{(r_{1,1})}, \dots, \rho^{(r_{1,n_1})}), \dots, \gamma(\rho^{(n_m)}; \rho^{(r_{m,1})}, \dots, \rho^{(r_{m,n_m})})),$$

for $\rho^{(m)} \in Q(m)$, $\rho^{(n_i)} \in Q(n_i)$, $i = 1, \dots, m$, and $\rho^{(r_{i,\ell_i})} \in Q(r_{i,\ell_i})$ with $\ell_i = 1, \dots, n_i$. The left-hand-side is

$$\gamma\left(\begin{pmatrix} \rho_{11}^{(m)}\rho^{n_1} & & & \\ & \rho_{22}^{(m)}\rho^{n_2} & & \\ \vdots & & \cdots & \vdots \\ & & & \rho_{mm}^{(m)}\rho^{n_m} \end{pmatrix}; \rho^{(r_{1,1})}, \dots, \rho^{(r_{1,n_1})}, \dots, \rho^{(r_{m,1})}, \dots, \rho^{(r_{m,n_m})}\right) =$$

$$\begin{pmatrix} \rho_{11}^{(m)}\rho_{11}^{n_1}\rho^{(r_{1,1})} & & & & & & \\ & \ddots & & & & & \\ & & \rho_{11}^{(m)}\rho_{n_1 n_1}^{n_1}\rho^{(r_{1,n_1})} & & & & \\ & & & \ddots & & & \\ & & & & \rho_{mm}^{(m)}\rho_{11}^{n_1}\rho^{(r_{m,1})} & & \\ & & & & & \ddots & \\ & & & & & & \rho_{mm}^{(m)}\rho_{n_m n_m}^{n_1}\rho^{(r_{m,n_m})} \end{pmatrix},$$

which agrees with the right-hand-side

$$\gamma\left(\rho^{(m)};\begin{pmatrix}\rho_{11}^{n_1}\rho^{(r_{1,1})} & & \\ & \ddots & \\ & & \rho_{n_1 n_1}^{n_1}\rho^{(r_{1,n_1})}\end{pmatrix},\ldots,\begin{pmatrix}\rho_{11}^{n_m}\rho^{(r_{m,1})} & & \\ & \ddots & \\ & & \rho_{n_m n_m}^{n_m}\rho^{(r_{m,n_m})}\end{pmatrix}\right).$$

The compatibility with the symmetric group action for Q_P is obtained directly from Lemma 4.4.2. Indeed, symmetric property of an operad is expressed by the following two identities, for permutations $\sigma_i \in \Sigma_{n_i}$ and $\sigma \in \Sigma_m$. The first condition is

$$\gamma_P(\sigma(\rho);\rho_{\sigma^{-1}(1)},\ldots,\rho_{\sigma^{-1}(m)}) = \widetilde{\sigma}(\gamma_P(\rho;\rho_1,\ldots,\rho_m)),$$

where on the right-hand-side $\widetilde{\sigma} \in \Sigma_{n_1+\cdots+n_m}$ is the permutation that splits the set of indices into blocks of n_i indices and permutes the blocks by σ. The second symmetric group condition is

$$\gamma_P(\rho;\sigma_1(\rho_1),\ldots,\sigma_m(\rho_m)) = \widehat{\sigma}(\gamma_P(\rho;\rho_1,\ldots,\rho_m)),$$

where on the right-hand-side $\widehat{\sigma} \in \Sigma_{n_1+\cdots+n_m}$ is the permutation that acts on the i-th block of n_i indices as the permutation σ_i.

In the first case, we have

$$\gamma_P(\sigma(\rho);\rho_{\sigma^{-1}(1)},\ldots,\rho_{\sigma^{-1}(m)}) = \gamma_P(\sigma^{-1}P(\rho);\rho_{\sigma^{-1}(1)},\ldots,\rho_{\sigma^{-1}(m)}),$$

which is the same as $\widetilde{\sigma}\gamma_P(\rho;\rho_1,\ldots,\rho_m))\widetilde{\sigma}^*$ in $\mathcal{M}^{(n_1+\cdots+n_m)}$. A similar argument proves the second relation.

The operad is non-unital. Indeed, the unit axiom is only satisfied for $\rho = 1 \in Q(1)$ with $\gamma_P(1;\rho) = \rho$, but it fails when $\rho_i = 1 \in Q(1)$, where the composition gives instead $\gamma_P(\rho;1,\ldots,1) = P(\rho)$. The unit axiom $\gamma_P(\rho;1,\ldots,1) = \rho$ is satisfied on the subset $\Delta_n \subset \mathcal{M}^{(n)}$ of classical probabilities, where the operad agrees with the unital operad $\mathcal{P}$. □

Proposition 5.2 *The composition laws γ_P of the non-unital operad Q_P are induced by insertion operations $\circ_i : Q_P(n) \times Q_P(m) \to Q_P(n+m-1)$, hence the operad Q_P is also a non-unital operad in the stronger sense.*

Proof For the composition operations γ_P to be obtained from insertions $\circ_i$, we need to have

$$\gamma_P(\rho;\rho_1,\ldots,\rho_n) = (\cdots(\rho\circ_n\rho_n)\cdots\circ_1\rho_1),$$

for $\rho \in \mathcal{M}^{(n)}$ and $\rho_i \in \mathcal{M}^{(n_i)}$. We can define morphisms

$$\circ_i : Q_P(n) \times Q_P(m) \to Q_P(m+n-1)$$

as the operations $(\rho, \rho') \mapsto \rho \circ_i \rho'$ that take a density matrix $\rho \in \mathcal{M}^{(n)}$ and replace the i-th row and columns with m rows and m columns, respectively, where all the entries outside of the $m \times m$-block around the diagonal are zero, and the diagonal block is given by the matrix $\rho_{ii}\rho'$. Clearly this implies that the repeated application of these insertions performed in the order $(\cdots(\rho \circ_n \rho_n)\cdots \circ_1 \rho_1)$ produces exactly the matrix $\gamma_P(\rho; \rho_1, \ldots, \rho_n)$. □

5.5.5 The Q_Λ-Operad of Quantum States

We can construct, in a very similar way, another operad of quantum states, using the classical probabilities $\Lambda(\rho)$ instead of $P(\rho)$. The resulting operad has slightly different properties, coming from the choice of an ordering of the eigenvalues.

Definition 5.4 For $n \geq 1$ let $Q_\Lambda(n) = \mathcal{M}^{(n)}$ be the convex set of density matrices, endowed with the composition laws

$$\gamma_\Lambda : Q_\Lambda(n) \times Q_\lambda(k_1) \times \cdots \times Q_\Lambda(k_n) \to Q_\Lambda(k_1 + \cdots + k_n)$$

$$\gamma_\Lambda(\rho; \rho_1, \ldots, \rho_n) = \gamma(\Lambda(\rho); \rho_1, \ldots, \rho_n) = \begin{pmatrix} \lambda_1\rho_1 & & & \\ & \lambda_2\rho_2 & & \\ \vdots & & \cdots & \vdots \\ & & & \lambda_n\rho_n \end{pmatrix}$$

with λ_i the eigenvalues of ρ listed in non-increasing order.

Proposition 5.3 *The convex sets $Q_\Lambda(n)$ with the composition operations γ_Λ of Definition 5.4 determine a non-unital non-symmetric operad Q_Λ. The composition laws γ_Λ are induced by insertion operations $\circ_i : Q_\Lambda(n) \times Q_\Lambda(m) \to Q(n+m-1)$.*

Proof The argument is completely analogous to Propositions 5.1 and 5.2. The associativity requirement for the composition laws γ_Λ follows as in Proposition 5.1, using the fact that

$$Spec \begin{pmatrix} \lambda_1\rho^{n_1} & & & \\ & \lambda_2\rho^{n_2} & & \\ \vdots & & \cdots & \vdots \\ & & & \lambda_m\rho^{n_m} \end{pmatrix} = \bigcup_i \lambda_i \, Spec(\rho^{n_i}).$$

The operad $\mathcal{Q}_\Lambda$ is non-unital for the same reason as $\mathcal{Q}_P$, namely $\gamma_\Lambda(\rho; 1\ldots, 1) = \Lambda(\rho) \neq \rho$. Even in the case of diagonal matrices these differ in general by a permutation.

The operad is non-symmetric, because we must choose an ordering of the eigenvalues in $\Lambda(\rho)$, for example, non-increasing ordering. This breaks the symmetric group action in the first of the two identities, making $\mathcal{Q}_\Lambda$ non-symmetric.

The insertion operations $\circ_i$ are as in Proposition 5.2, but the central $m \times m$ block is now of the form $\lambda_i \rho'$ □

5.5.6 *Trees of Projective Quantum Measurements*

We consider the range of compositions of insertion maps of the operad $\mathcal{Q}_P$ and quantum channels associated to these ranges.

First observe that the image of the insertion map

$$\circ_i : \mathcal{Q}_P(n) \times \mathcal{Q}_P(m) \to \mathcal{Q}_P(n+m-1)$$

consists of the set of those density matrices $\rho \in \mathcal{M}^{(n+m-1)}$ that are block diagonal with one $(n-1)\times(n-1)$-block and one $m \times m$-block. Moreover, all block diagonal density matrices are in the image of some composition of insertion maps. These are quantum states that decompose nontrivially into disjoint states with orthogonal ranges. On the other hand, a block diagonal density matrix can be obtained in more than one way through a composition of insertion maps.

Operators on density matrices are described by quantum channels, namely quantum measurements realized by completely positive maps. A particular class of such operators consists of *projective quantum measurements*.

In the following we work with a finite dimensional Hilbert space $\mathcal{H}$ of dimension N with a chosen orthonormal basis, and we denote as before by $\mathcal{M}^{(N)}$ the set of density matrices, written in the chosen basis.

Definition 5.5 A projective quantum measurement is a family $\Pi = \{\Pi_i\}_{i=1}^n$ of projectors $\Pi_i^* = \Pi_i = \Pi_i^2$ in a finite dimensional Hilbert space $\mathcal{H}$ of dimension N, that are mutually orthogonal, $\Pi_i \Pi_j = \delta_{ij}\Pi_i$, and satisfy the condition $\sum_i \Pi_i = 1$. The outcome of the projective measurement Π on a quantum state given by density matrices $\rho \in \mathcal{M}^{(N)}$ is

$$\rho_i = \frac{\Pi_i \rho \Pi_i}{\mathrm{Tr}(\Pi_i \rho)} \quad \text{with probability} \quad p_i = \mathrm{Tr}(\Pi_i \rho).$$

The projective quantum channel Π then maps $\rho \mapsto \Pi(\rho) = \sum_i p_i \rho_i$.

The range in $\mathcal{M}^{(N)}$ of a composition of insertion maps is specified by assigning a decomposition $N = k_1 + \cdots + k_n$ and the locus $\mathcal{M}_{k_1,\ldots,k_n} \subset \mathcal{M}^{(N)}$ of density

matrices that are block-diagonal in the chosen basis, with n blocks of size k_i. The following is immediate by construction.

Lemma 5.3 *Let Π be the projective measurement $\Pi = \{\Pi_i\}_{i=1}^n$, where Π_i is the orthogonal projection onto the span of the i-th subset of k_i basis elements, given by a quantum channel. It maps $\Pi : \mathcal{M}^{(N)} \to \mathcal{M}_{k_1,\dots,k_n}$, assigning to ρ the block-diagonal density matrix $\Pi(\rho) = \sum_i p_i \rho_i$.*

Just as elements of $\mathcal{M}_{k_1,\dots,k_n}$ can be obtained in different ways as compositions of insertion operations of the operad $\mathcal{Q}_P$ of quantum states, with different tree structures, the quantum channel $\Pi : \mathcal{M}^{(N)} \to \mathcal{M}_{k_1,\dots,k_n}$ can also be realized through different tree structures.

Let τ be a planar rooted tree with n leaves labeled by the non-negative integers k_i. We view the tree τ as oriented from the leaves toward the root. We assign to the root vertex v_0, the identity projector $\Pi^{(v_0)} = 1$. Let v be any vertex in the tree, and consider the set of incoming edges e at v, with $v_e = s(e)$ their source vertices. Let $\{\Pi^{(v_e)}\}_{t(e)=v}$ be a set of orthogonal projections with $\sum_{e:t(e)=v} \Pi^{(s(e))} = \Pi^{(v)}$. Let Π_i denote the resulting projectors at the leaves. We write, for $t(e) = v$,

$$\rho^{(s(e))} = \frac{\Pi^{(s(e))} \rho^{(v)} \Pi^{(s(e))}}{\mathrm{Tr}(\Pi^{(s(e))} \rho^{(v)})},$$

with $\rho^{(v_0)} = \rho$.

Consider the quantum measurement Π_τ that assigns to a density matrix $\rho \in \mathcal{M}^{(N)}$, with $N = k_1 + \cdots + k_n$ outcomes ρ_i^τ with probabilities p_i^τ where

$$p_i^\tau = \prod_w \mathrm{Tr}(\Pi_{i_w}^{(w)} \rho^{(w)})$$

with the product over the vertices on the directed path from the i-th leaf to the root of τ, with i_w indicating the direction at the vertex w along this path and the ρ_i^τ are obtained by repeatedly computing $\rho^{(s(e))}$ from $\rho^{(t(e))}$ along the path connecting the i-th leaf to the root.

Lemma 5.4 *All the quantum channels Π_τ obtained in this way, are just the same quantum channel $\Pi : \mathcal{M}^{(N)} \to \mathcal{M}_{k_1,\dots,k_n}$.*

Proof This can be seen very easily by writing

$$\prod_w \mathrm{Tr}(\Pi_{i_w}^{(w)} \rho^{(w)}) = \prod_\ell \frac{\mathrm{Tr}(\Pi^{(w_\ell)} \Pi^{(w_{\ell-1})} \cdots \Pi^{(w_0)} \rho)}{\mathrm{Tr}(\Pi^{(w_{\ell-1})} \cdots \Pi^{(w_0)} \rho)} = \prod_\ell \frac{\mathrm{Tr}(\Pi^{(w_\ell)} \rho)}{\mathrm{Tr}(\Pi^{(w_{\ell-1})} \rho)} = \mathrm{Tr}(\Pi_i \rho)$$

with Π_i the projection at the leaf, since we have $\Pi_j^{(s(e))} \Pi_e^{(t(e))} = \Pi_j^{(s(e))}$ as $\Pi_j^{(s(e))}$ is a projection onto a subspace of the range of $\Pi_e^{(t(e))}$. We similarly obtain $\rho_i^\tau = \rho_i = \frac{\Pi_i \rho \Pi_i}{\mathrm{Tr}(\Pi_i \rho)}$. □

5.5.7 Entropy Functionals

Consider a family of quantum entropy functionals

$$S_n : \mathcal{M}^{(n)} \to \boldsymbol{R}$$

satisfying the consistency condition that S_n restricts to S_k for $k < n$ over any copy of $\mathcal{M}^{(k)}$ embedded in $\mathcal{M}^{(n)}$ as density matrices with a set of $n - k$ vanishing eigenvalues.

Examples of such consistent collections of entropy functionals include the von Neumann entropy

$$\mathcal{N}(\rho) = -\mathrm{Tr}(\rho \log \rho),$$

or, for a real parameter $q > 0$ with $q \neq 1$, the quantum Rényi entropy

$$Ry_q(\rho) = \frac{1}{1-q} \log \mathrm{Tr}(\rho^q)$$

and the quantum Tsallis entropy

$$Ts_q(\rho) = \frac{1}{1-q}(\mathrm{Tr}(\rho^q) - 1).$$

We obtain a family of quantum entropies associated to trees in the following way, as a direct generalization of the entropy functionals S_τ for classical probabilities constructed in [10].

Proposition 5.4 *A tree τ with n leaves labeled by integers $k_i \geq 1$, together with a coherent family $\{S_n\}$ of quantum entropies, determines an entropy functional*

$$S_\tau : \mathcal{M}^{(N)} \to \boldsymbol{R}$$

for $N = k_1 + \cdots + k_n$.

Proof In the case where τ is a corolla with a single root vertex and n leaves, we set

$$S_\tau(\rho) = S(P) + \sum_i p_i S(\rho_i)$$

with $p_i = \mathrm{Tr}(\Pi_i \rho)$, resp. $\rho_i = \frac{\Pi_i \rho \Pi_i}{\mathrm{Tr}(\Pi_i \rho)}$, are the probabilities, resp. density matrices, at the leaves. In the case of the von Neumann entropy, by the extensivity property, this is the same as $\mathcal{N}(\Pi^\tau(\rho)) = \mathcal{N}(\sum_i p_i \rho_i)$. Inductively assuming that S_τ is constructed for all trees with less than n leaves, consider the subtrees τ_j, $j = 1, \ldots, m$, attached at the root vertex v_0, each with a set L_j of leaves, card $L_j < n$.

By the construction of the quantum channel Π^τ we have a system Π_j of orthogonal projections with $\sum_j \Pi_j = 1$ associated to the incoming edges e_j at the root vertex, and probabilities $p_j = \mathrm{Tr}(\Pi_j \rho)$ and density matrices $\rho_j = \frac{\Pi_j \rho \Pi_j}{\mathrm{Tr}(\Pi_j \rho)}$ associated to the root vertices v_j of the subtrees τ_j. We then set

$$S_\tau(\rho) = S(P) + \sum_j p_j S_{\tau_j}(\rho_j),$$

where $S(P)$ is the Shannon entropy of the classical probability $P = (p_j)$ and S_{τ_j} are the entropy functionals inductively constructed for the subtrees τ_j with less than n leaves. This suffices to determine S_τ uniquely. □

Remark 5.1 In the case of the von Neumann entropy, the extensivity property and the identification of all the quantum channels Π^τ with the quantum channel $\Pi : \mathcal{M}^{(N)} \to \mathcal{M}_{k_1,\dots,k_n}$, seen in Lemma 5.4 above, imply that $\mathcal{N}_\tau(\rho) = \mathcal{N}(\sum_i p_i \rho_i) = S(P) + \sum_i p_i \mathcal{N}(\rho_i)$ for all τ. This is not the case for non-extensive entropies like Rényi and Tsallis.

5.5.8 A_∞-Operad of Quantum Channels

We now investigate structures on quantum states that generalize the operations based on classical probabilities that we recalled in Sect. 5.5.1.2.

Definition 5.6 A quantum channel $\Phi : \mathcal{M}^{(N)} \to \mathcal{M}^{(N)}$ is a trace preserving completely positive map. It is well known that any such map can be represented, in a non-unique way, in Kraus form, namely as

$$\Phi(\rho) = \sum_i A_i \rho A_i^*$$

for a collection $\{A_i\}$ of operator satisfying the condition $\sum_i A_i^* A_i = 1$.

The projective quantum channels considered in the previous subsections are those for which the operators A_i are mutually orthogonal projectors.

We can construct more general quantum channels associated to rooted trees. We consider a finite dimensional complex Hilbert space $\mathcal{H}$ of dimension N. All operators here will be linear operators on $\mathcal{H}$.

Definition 5.7 Let τ be a planar rooted tree with n leaves. We consider τ oriented from the leaves to the root. A tree quantum channel C_A^τ is an assignment of operators

$A = \{A_e\}_{e \in E(\tau)}$ to the edges of τ, satisfying the condition that at each vertex v

$$\sum_{e\,:\,t(e)=v} A_e^* A_e = 1.$$

Lemma 5.5 *The tree quantum channels C_A^τ of Definition 5.7 are quantum channel as in Definition 5.6, acting on density matrices $\rho \in \mathcal{M}^{(N)}$ by*

$$C_A^\tau(\rho) = \sum_{i=1}^{n} A_{e_{i,1}} \cdots A_{e_{i,m_i}} \rho\, A_{e_{i,m_i}}^* \cdots A_{e_{i,1}}^*,$$

where the sum is over the leaves of τ and for each $i = 1, \ldots, n$ we consider the oriented path $e_{i,1}, \ldots, e_{i,m_i}$ from the i-th leaf to the root, $s(e_{i,1}) = v_i$, $t(e_{i,j}) = s(e_{i,j+1})$, $t(e_{i,m_i}) = v_0$, the root vertex.

Proof This is essentially the Kraus form of a quantum channel, since we have

$$\sum_{i=1}^{n} A_{e_{i,m_i}}^* \cdots A_{e_{i,1}}^*\, A_{e_{i,1}} \cdots A_{e_{i,m_i}} = 1.$$

Indeed, write $A_v := A_{e_1} \cdots A_{e_m}$ for the composition of the operators A_e along the oriented path from the vertex v to the root v_0. So we write the above as $\sum_i A_i^* A_i$. Starting at the leaves and considering the adjacent vertices, we can rewrite the sum as

$$\sum_v \sum_{i\,:\,t(e_i)=v} A_v^* A_{e_{i,1}}^*\, A_{e_{i,1}} A_v = \sum_v A_v^* A_v,$$

where the set $\{i \,:\, t(e_i) = v\}$ is non-empty only for the vertices v adjacent to the leaves. This reduces by one the length of the path. Thus, we obtain inductively that the normalisation $\sum_v A_v^* A_v = 1$ holds, with the condition $\sum_{e\,:\,t(e)=v} A_e^* A_e = 1$ implying that it holds for length one. □

Theorem 5.3 *The tree quantum channels C_A^τ form an A_∞-operad $\mathcal{QC}$.*

Proof Consider $\mathbf{Z}$-modules $\mathcal{QC}(n) := span_{\mathbf{Z}}\{C_A^\tau \mid \tau \in \mathcal{T}(n)\}$ where $\mathcal{T}(n)$ is the A_∞-operad of planar rooted trees. The operadic composition laws

$$\gamma_{\mathcal{QC}} : \mathcal{QC}(n) \otimes \mathcal{QC}(k_1) \otimes \cdots \otimes \mathcal{QC}(k_n) \to \mathcal{QC}(k_1 + \cdots + k_n)$$

are given by

$$\gamma_{\mathcal{QC}}(C_A^\tau; C_{A_1}^{\tau_1}, \ldots, C_{A_n}^{\tau_n}) = C_{A \cup \{A_1, \ldots, A_n\}}^{\gamma_{\mathcal{T}}(\tau; \tau_1, \ldots, \tau_n)}.$$

The associativity, unity, and symmetric properties of $\mathcal{QC}$ follow directly from the same properties of the operad $\mathcal{T}$. The DG-structure of $\mathcal{QC}$ is also inherited from the DG-structure of the A_∞-operad $\mathcal{T}$, with the differential given by edge contractions

$$dC_A^\tau = \sum_{\tau' : \tau = \tau'/e} \epsilon\, C_{A'}^{\tau'},$$

where $\epsilon = (-1)^{\ell(e)}$ with $\ell(e)$ the number of edges below and to the left of e in τ with respect to the planar structure.

The collection of operators A' on a tree τ' with $\tau = \tau'/e$ agrees with A on all edges e' with $t(e') \neq t(e), s(e)$ and is of the following form on the remaining edges. Let E_t be the set of edges e' of τ such that $t(e') = t(e)$ in τ' and E_s the set of edges e' in τ such that $t(e') = s(e)$ in τ'. The set $E_t \cup E_s$ consists of all the edges of τ with the same target vertex v in τ that is split into two vertices $s(e), t(e)$ in τ'. Thus, in C_A^τ we have the relation $\sum_{e' \in E_t \cup E_s} A_e^* A_e = 1$.

Let $B_t := \sum_{e' \in E_t} A_{e'}^* A_{e'}$ and $B_s := \sum_{e' \in E_t} A_{e'}^* A_{e'}$. These are positive operators, namely $\langle Bv, v\rangle \geq 0$ for all $v \in \mathcal{H}$. Put $N_s = \operatorname{card} E_s$ and consider the operators $\frac{1}{N_s} B_t$ and $A_{e'}^* A_{e'} + \frac{1}{N_s} B_t$ for $e' \in E_s$. These are also positive operators. Thus, we can write $B_s = A^* A$ and $A_{e'}^* A_{e'} + \frac{1}{N_s} B_t = \widetilde{A}_{e'}^* \widetilde{A}_{e'}$ for some operators A and $\widetilde{A}_{e'}$.

We then take $A'_e := A$ and $A'_{e'} := \widetilde{A}_{e'}$ for $e' \in E_s$. This completes the description of A' on τ' in a way that still satisfies the conditions at vertices

$$\sum_{e' : t(e') = s(e)} {A'_{e'}}^* A'_{e'} = \sum_{e' \in E_s} (A_{e'}^* A_{e'} + \frac{1}{N_s} B_t) = \sum_{e' \in E_s} A_{e'}^* A_{e'} + \sum_{e' \in E_t} A_{e'}^* A_{e'} = 1,$$

$$\sum_{e' : t(e') = t(e)} {A'_{e'}}^* A'_{e'} = {A'_e}^* A'_e + \sum_{e' \in E_t} A_{e'}^* A_{e'} = \sum_{e' \in E_s} A_{e'}^* A_{e'} + \sum_{e' \in E_t} A_{e'}^* A_{e'} = 1.$$

Note that, in order to make the composition operations compatible with the differential of the DG-structure, they should also include appropriate signs, as specified for instance in [15]. We will omit the details as they are exactly the same as in the original case of the A_∞-operad $\mathcal{T}$.

□

Taking formal linear combinations of quantum channels, as in the definition of $\mathcal{QC}$ above, has the advantage of being able to define the differential and DG-structure described in Theorem 5.3. However, it is somewhat unnatural, since the positivity property of quantum channels is lost in the linear combinations. It is therefore more natural in this context to consider only convex combinations. This leads to a variant of the operad $\mathcal{QC}$ of tree quantum channels.

Definition 5.8 Let $\mathcal{QC}^+$ be given by $\mathcal{QC}^+(n) = \text{convex span}\{C_A^\tau \mid \tau \in \mathcal{T}(n)\}$. The $\mathcal{QC}^+(n)$ are convex sets rather than $\mathbf{Z}$-modules (or vector spaces). The composition

operations are the same as the γ_{QC} (without signs),

$$\gamma_{QC^+}(C_A^\tau; C_{A_1}^{\tau_1}, \ldots, C_{A_n}^{\tau_n}) = C_{A\cup\{A_1,\ldots,A_n\}}^{\gamma_\mathcal{T}(\tau;\tau_1,\ldots,\tau_n)}.$$

Then QC^+ defined in this way is an operad, though it no longer has a DG-structure.

Proposition 5.5 *The convex set $\mathcal{M}^{(N)}$ has the structure of an algebra over the operad QC^+.*

Proof The operations $\alpha : QC^+(n) \otimes \mathcal{M}^{(N)^{\otimes n}} \to \mathcal{M}^{(N)}$ are given by a slight modification of the action of tree quantum channels of Lemma 5.5,

$$\alpha(C_A^\tau; \rho_1, \ldots, \rho_n) = \sum_{i=1}^n p_i \widetilde{\rho}_i,$$

$$\widetilde{\rho}_i = \frac{A_{\gamma_i}\rho_i A_{\gamma_i}^*}{\mathrm{Tr}(A_{\gamma_i}^* A_{\gamma_i}\rho_i)} \quad \text{and} \quad p_i = \mathrm{Tr}(A_{\gamma_i}^* A_{\gamma_i}),$$

where $A_{\gamma_i} = A_{e_{i,1}} \cdots A_{e_{i,m_i}}$, along the oriented path $\gamma_i = e_{i,1}, \ldots, e_{i,m_i}$ from the i-th leaf to the root. □

5.6 Operads and Almost-Symplectic Quantum Codes

In this section we continue our investigation of operadic structures in quantum information, by revisiting a construction of quantum codes from (not always Moufang) loops obtained from almost-symplectic vector spaces over finite fields that we developed in [4]. Here we need to consider a slightly different definition of almost-symplectic structure over finite fields with respect to the one used in [4]. The choice we consider here is better because it allows for an operadic composition, but the construction of the associated quantum codes is then less well behaved. We show that the space of almost-symplectic structures (in the sense we consider here) is an algebra over operad modelled on May's little square operad, and the set of data defining the associated quantum codes is a partial-algebra over the same operad.

5.6.1 Rational and Binary Little Square Operads

The little square operad [12] provides a characterization of topological spaces that are 2-fold loop spaces. Little n-cube operads [12] similarly characterize n-fold loop spaces.

We consider first a sub-operad where we impose an additional condition on the linearly scaled versions of the unit square in the operad, namely that they have corners located at rational points in the unit square, namely that the scaling is affected by linear functions with rational coefficients.

Let $\mathcal{I} = [0, 1]$ be the unit interval, with $\mathcal{J} = (0, 1)$ its interior. Let $\mathcal{I}_{\mathcal{Q}} = \mathcal{I} \cap \mathcal{Q}$ and $\mathcal{J}_{\mathcal{Q}} = \mathcal{J} \cap \mathcal{Q}$. Thus, $\mathcal{I}^2$ is the unit square with its interior $\mathcal{J}^2$ and $\mathcal{I}^2_{\mathcal{Q}}$ and $\mathcal{J}^2_{\mathcal{Q}}$ are the respective sets of rational points. A rational little square is a function

$$c : \mathcal{I}^2_{\mathcal{Q}} \to \mathcal{I}^2_{\mathcal{Q}}, \quad c = (c^1, c^2), \quad c^i(t) = (y_i - x_i)t + x_i \quad \forall t \in \mathcal{I}_{\mathcal{Q}},$$

for some $x_i, y_i \in \mathcal{I}_{\mathcal{Q}}$. An n-tuple $\langle c_1, \ldots, c_n \rangle$ of rational little squares has disjoint interiors if $c_i(\mathcal{J}^2_{\mathcal{Q}}) \cap c_j(\mathcal{J}^2_{\mathcal{Q}}) = \emptyset$ for $i \neq j$.

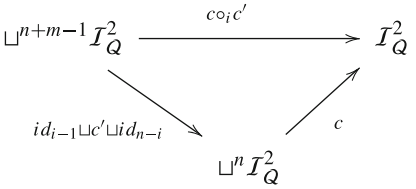
The n-th object of the rational little square operad $C^{\mathcal{Q}}_2(n)$ is the space of n-tuples $\langle c_1, \ldots, c_n \rangle$ of rational little squares with disjoint interiors. For $n = 0$, the space $C^{\mathcal{Q}}_2(0)$ consists of a unique function $\emptyset \to \mathcal{I}^2$.

Let $\sqcup^n \mathcal{I}^2$ denote the disjoint union $\mathcal{I}^2 \sqcup \cdots \sqcup \mathcal{I}^2$ of n copies of $\mathcal{I}^2$. By identifying $\langle c_1, \ldots, c_n \rangle$ with a function $c_1 \sqcup \cdots \sqcup c_n : \sqcup^n \mathcal{I}^2 \to \mathcal{I}^2$, the set $C^{\mathcal{Q}}_2(n)$ is endowed with the topology induced by the compact-open topology on the set of continuous maps from $\sqcup^n \mathcal{I}^2$ to $\mathcal{I}^2$.

The operad compositions

$$\circ_i : C^{\mathcal{Q}}_2(n) \times C^{\mathcal{Q}}_2(m) \to C^{\mathcal{Q}}_2(n + m - 1)$$

of $c = \langle c_1, \ldots, c_n \rangle \in C^{\mathcal{Q}}_2(n)$ and $c' = \langle c'_1, \ldots, c'_m \rangle \in C^{\mathcal{Q}}_2(m)$ are determined by the diagrams



The unit of the rational little square operad is the identity map $id : \mathcal{I}^2_{\mathcal{Q}} \to \mathcal{I}^2_{\mathcal{Q}}$ in $C^{\mathcal{Q}}_2(1)$. The action on $c = \langle c_1, \ldots, c_n \rangle$ of a permutation σ in the symmetric group Σ_n is given by

$$c\sigma := \langle c_{\sigma(1)}, \ldots, c_{\sigma(n)} \rangle ,$$

permuting the labels of the little squares.

5.6.1.1 Binary Little Square Operad

We then consider subspaces $C_2^{\mathcal{F}_2}(n) \subset C_2^{\mathcal{Q}}(n)$ given by those rational little squares $c = \langle c_1, \ldots, c_n \rangle$ with disjoint interiors, with the property that the endpoints of each $c_i(\mathcal{I}^2)$ are rational points of $\mathcal{I}^2$ that lie on the square grid of length 2^{-N}, for some $N \geq 0$.

We refer to the parallel grid of length 2^{-N} in the unit square $\mathcal{I}^2$ as *the N-grid*.

Lemma 5.6 *The subspaces $C_2^{\mathcal{F}_2}(n) \subset C_2^{\mathcal{Q}}(n)$, with the induced composition operations, determine an operad $C_2^{\mathcal{F}_2}$, the "binary little square operad".*

Proof Under the operad composition operations of $C_2^{\mathcal{Q}}$, the compositions $c \circ_i c'$ of two binary little squares is still a binary little square, so $C_2^{\mathcal{F}_2}$ is a sub-operad of $C_2^{\mathcal{Q}}$. □

5.6.1.2 Strict Binary Little Squares

Given a binary little square $c \in C_2^{\mathcal{F}_2}(n)$, with $c = \langle c_1, \ldots, c_n \rangle$, let $N_c \in \boldsymbol{N}$ be the smallest natural number such that the corners of all the $c_i(\mathcal{I}^2)$ are at vertices of the N_c-grid of size 2^{-N_c}.

Definition 5.9 A binary little square c is "strict" if every row and column of the N_c-grid has at least one square that is not contained in the union $\cup_{i=1}^n c_i(\mathcal{I}^2)$.

Consider the sub-spaces $C_2^{\mathcal{F}_2,s}(n) \subset C_2^{\mathcal{F}_2}(n)$ consisting of binary little squares that are strict.

Lemma 5.7 *The $C_2^{\mathcal{F}_2,s}(n)$ with the induced composition operations, determine a sub-operad of $C_2^{\mathcal{F}_2}$.*

Proof We need to check that the operad compositions preserve the strict property of little squares. Given $c \in C_2^{\mathcal{F}_2,s}(n)$ and $c' \in C_2^{\mathcal{F}_2,s}(m)$, the endpoints of the regions $(c \circ_i c')_j(\mathcal{I}^2)$, $j = 1, \ldots, n+m-1$ are on a grid of size $2^{-N_{c \circ_i c'}}$, with $N_{c \circ_i c'} \geq N_c$. Suppose there is a row R (or column) of the $N_{c \circ_i c'}$-grid that is completely contained in the region $\cup_{j=1}^{n+m-1}(c \circ_i c')_j(\mathcal{I}^2)$. If R does not intersect the region $c_i(\mathcal{I}^2)$, then it is contained in the union of the $c_j(\mathcal{I}^2)$ with $j \neq i$. Since this region has all sides along the N_c-grid, this implies that there must be in fact a row of the N_c-grid, containing R, that is contained in $\cup_{j \neq i} c_j(\mathcal{I}^2)$, but this is not possible because c is strict. Thus, R must intersect the region $c_i(\mathcal{I}^2)$. This means that, within the region $c_i(\mathcal{I}^2)$ a row of the $N_{c \circ_i c'}$-grid is completely contained in the union of the linearly scaled images of the $c'_j(\mathcal{I}^2)$, but this in turn implies that in $\mathcal{I}^2$ a row of the $N_{c'}$-grid must be contained in the union of the $c'_j(\mathcal{I}^2)$, which cannot happen because c' is strict. □

5.6.2 Binary Little Square Operads and Almost Symplectic Spaces

We consider here a class of (not always Moufang) loops that we previously introduced and investigated in [4]. These are obtained from almost-symplectic structures on vector spaces over a finite field of characteristic 2. We recall the basic setting from [4]. However, the notion of almost-symplectic form we consider here is somewhat different from [4]: this will allow for better properties with respect to operadic composition, but will in turn have worse properties with respect to representations of the resulting loops, hence the construction of quantum codes considered in [4] will not directly extend to this setting.

Let $\mathscr{F} = \mathscr{F}_{2^r}$ be a finite field of characteristic two and let V be a finite dimensional vector space over $\mathscr{F}$. Let K be an unramified extension of Q_2 with residue field $O_K/\boldsymbol{m}_K = \mathscr{F}$. Consider the ring $R = O_K/\boldsymbol{m}_K^2$ and a free R-module $\widetilde{V}$ with $V = \widetilde{V}/\boldsymbol{m}_K$. Consider functions $\widetilde{\omega} : \widetilde{V} \times \widetilde{V} \to R$ with $\omega = 2\widetilde{\omega}$ the induced function $\omega : V \times V \to R$. Note that these functions are not linear as they do not satisfy the Hochschild cocycle condition of symplectic forms, namely we require that

$$d\omega(u, v, w) = \omega(v, w) - \omega(u + v, w) + \omega(u, v + w) - \omega(u, v) \neq 0 .$$

A pair (V, ω) as above is an almost-symplectic space if ω is non-degenerate, in the sense that for any $u \in V$ there is some $v \in V$ with $(u, v) \neq (0, 0)$, such that $\omega(u, v) \neq 0$. (Note that with this definition the almost-symplectic structure ω does *not* satisfy $\omega(0, v) = \omega(u, 0) = 0$, for all $u, v \in V$.) The almost-symplectic structure ω has a polarization $\beta : V \times V \to R$ satisfying $\beta(u, v) - \beta(v, u) = \omega(u, v)$. The loop $\mathcal{L}(V, \beta)$ associated to the data (V, β) is the extension

$$0 \to R \to \mathcal{L}(V, \beta) \to V \to 0$$

with non-associative multiplication

$$(x, u) \star (y, v) = (x + y + \beta(u, v), u + v) .$$

The non-associativity is a consequence of the fact that (V, ω) is almost-symplectic and not symplectic, hence the Hochshild coboundary of β is nonzero,

$$d\beta(u, v, w) = \beta(v, w) - \beta(u + v, w) + \beta(u, v + w) - \beta(u, v) = \gamma(u, v, w) \not\equiv 0 .$$

The Moufang condition for the loop $\mathcal{L}(V, \beta)$ is equivalent to an identity for $\gamma = d\beta$, see [4]. We do not necessarily require here the condition that the loop $\mathcal{L}(V, \beta)$ is Moufang.

We focus in particular on the special case where $\mathscr{F} = \mathscr{F}_2$ and $R = \boldsymbol{Z}/4\boldsymbol{Z}$. We can identify the map $\omega = 2\widetilde{\omega} : V \times V \to R$, where $V = \mathscr{F}_2^N$, as a subdivision

of the square $\mathcal{I}^2 = [0, 1] \times [0, 1]$ into the N-grid of $2^N \times 2^N$ sub-squares of side 2^{-N}, which we can label by the pairs $(u, v) \in V \times V$. Each square is colored white or black according to whether $\omega(u, v) = 0$ or not. The property that ω is almost-symplectic (that is, non-degenerate) is equivalent to the fact that in every row and column of the subdivided square there is at least one black sub-square.

Let $\mathcal{V}^{\mathcal{F}_2}$ be the space of almost symplectic finite dimensional vector spaces over $\mathcal{F}_2$, topologized as a subset of the space of maps $\cup_N Maps(\mathcal{F}_2^N \times \mathcal{F}_2^N, R)$, where $\mathcal{V}^{\mathcal{F}_2}$ consists of those maps that satisfy the non-degenerate condition above, with range $2R \subset R$, since $\omega = 2\widetilde{\omega}$.

Theorem 5.4 *The space $\mathcal{V}^{\mathcal{F}_2}$ of almost symplectic finite dimensional vector spaces over $\mathcal{F}_2$ is an algebra over the operad $C_2^{\mathcal{F}_2,s}$.*

Proof To realize $\mathcal{V}^{\mathcal{F}_2}$ as an algebra over $C_2^{\mathcal{F}_2,s}$ we need operations

$$C_2^{\mathcal{F}_2,s}(n) \times \underbrace{\mathcal{V}^{\mathcal{F}_2} \times \cdots \times \mathcal{V}^{\mathcal{F}_2}}_{n-\text{times}} \to \mathcal{V}^{\mathcal{F}_2},$$

that assign to an n-tuple $(V_1, \omega_1), \ldots, (V_n, \omega_n)$ and a strict binary little square $c = \langle c_1, \ldots, c_n \rangle$ a new almost-symplectic space $(V, \omega) = \gamma(c; (V_1, \omega_1), \ldots, (V_n, \omega_n))$, compatibly with the composition operations in the operad $C_2^{\mathcal{F}_2,s}$. We construct (V, ω) in the following way. For $V_i = \mathcal{F}_2^{N_i}$, we consider the regions $\omega_i^{-1}(0)$ with sides on the N_i-grid in $\mathcal{I}^2$. We take each of these copies of $\mathcal{I}^2$ subdivided in the subsquares of the N_i-grid, with those in $\omega_i^{-1}(0)$ colored white and the others colored black, and we scale it linearly so as to fit, respectively, into the $c_i(\mathcal{I}^2)$ regions of $\mathcal{I}^2$ determined by the binary little square $c = \langle c_1, \ldots, c_n \rangle$. We color black the outside of $\cup_i c_i(\mathcal{I}^2)$ in $\mathcal{I}^2$. Let $N \in \mathbb{N}$ be the smallest integer such that all the resulting contours in $\mathcal{I}^2$ separating the black and the white colored areas are on the N-grid. We then set $V = \mathcal{F}_2^N$ and we assign the values of $\omega(v, w)$ according to the color of the corresponding square. The fact that the binary little square is strict implies that ω is non-degenerate. □

5.6.3 Colored p-ary Little Squares

Let now $q = p^r$ be some prime power with $p > 2$. We construct an operad that generalizes the strict binary little squares operad considered in the case of characteristic 2.

For simplicity, as in the case of characteristic 2, we restrict to the case of $\mathcal{F}_p$. We then define the p-ary N-grid in the unit square $\mathcal{I}^2$ to the parallel grid of size p^{-N}, with $p^N \times p^N$ sub-squares. We just refer to it as the N-grid when the choice of p is understood.

Definition 5.10 A colored p-ary little square is a decomposition of the unit square $\mathcal{I}^2$ into p regions $\mathcal{R}_0, \dots, \mathcal{R}_{p-1}$ with disjoint interiors $\mathcal{J}(\mathcal{R}_i) \cap \mathcal{J}(\mathcal{R}_j) = \emptyset$, for $i \neq j$, with $\mathcal{J}(\mathcal{R})$ denoting the interior of a region $\mathcal{R}$. With the property that, for each $i = 0, \dots, p-1$, and for some integers $n_i, N_i \in \mathbb{N}$, there is a rational little square $c^{(i)} = \langle c_1^{(i)}, \dots, c_{n_i}^{(i)} \rangle$ in $C_2^{\mathcal{Q}}(n_i)$, with endpoints on a p-ary N_i-grid, such that $\mathcal{R}_i = \cup_{j=1}^{n_i} c_j^{(i)}(\mathcal{I}^2)$. The colored p-ary little square is strict if, for $i = 0$, the little square $c^{(0)}$ is strict, namely no row or column of the p-ary N_0-grid is completely contained in $\cup_{j=1}^{n_0} c_j^{(0)}(\mathcal{I}^2)$. We denote by $C_2^{\mathcal{F}_p}(n_0, \dots, n_{p-1})$ the set of colored p-ary little squares as above and we take

$$C_2^{\mathcal{F}_p}(n) := \bigcup_{n_1, \dots, n_{p-1}} C_2^{\mathcal{F}_p}(n, n_1, \dots, n_{p-1}) .$$

The set of strict colored p-ary little squares $C_2^{\mathcal{F}_{p,s}}(n)$ is similarly defined.

We denote the (strict) colored p-ary little squares in $C_2^{\mathcal{F}_p}(n)$ (or $C_2^{\mathcal{F}_{p,s}}(n)$, respectively) with the notation

$$\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}) := \left\{ (\mathcal{R}_0, c^{(0)} = \langle c_1^{(0)}, \dots, c_n^{(0)} \rangle), (\mathcal{R}_j, c^{(j)})_{j=1,\dots,p-1} \right\} ,$$

with $\mathcal{R} := \{(\mathcal{R}_j, c^{(j)})\}_{j=1,\dots,p-1}$. The composition operations

$$\circ_i : C_2^{\mathcal{F}_p}(n) \times C_2^{\mathcal{F}_p}(m) \to C_2^{\mathcal{F}_p}(n + m - 1)$$

are determined by taking

$$\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}) \circ_i \mathcal{I}^2(\mathcal{R}'_0, c'^{(0)}; \mathcal{R}')$$

to be given by the decomposition of the unit square $\mathcal{I}^2$ into the regions

$$\mathcal{I}^2 = \mathcal{B} \cup \bigcup_{j \neq 0} \mathcal{R}_j \cup \bigcup_{\ell \neq i} c_\ell^{(0)}(\mathcal{I}^2) ,$$

where the region $\mathcal{B}$ is obtained by linearly scaling the square $\mathcal{I}^2 = \cup_j \mathcal{R}'_j$ subdivided into the regions of the second colored p-ary little square and placing it in place of the region $c_i^{(0)}(\mathcal{I}^2)$.

Lemma 5.8 *With the composition operations above the $C_2^{\mathcal{F}_p}(n)$ (respectively, $C_2^{\mathcal{F}_{p,s}}(n)$) form an operad.*

Proof The new regions $\mathcal{R}''_\ell$ with $\ell = 1, \ldots, p-1$ of the composed $\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}) \circ_i \mathcal{I}^2(\mathcal{R}'_0, c'^{(0)}; \mathcal{R}')$ are given by

$$\mathcal{R}''_\ell = \mathcal{R}_\ell \cup c_i'^{(0)}(\mathcal{R}'_\ell)$$

while the $\mathcal{R}''_0$ region of the composed $\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}) \circ_i \mathcal{I}^2(\mathcal{R}'_0, c'^{(0)}; \mathcal{R}')$ is given by

$$\mathcal{R}''_0 = \bigcup_{j \neq i} c_j^{(0)}(\mathcal{I}^2) \cup c_i'^{(0)}(\mathcal{R}'_0)\,.$$

Thus we see that these composition operations are still the same composition operations of the operad $\mathcal{C}_2^{\mathcal{Q}}$, acting on the $c^{(0)}$ little squares while maintaining all the rest of the data unaffected. The strict condition is preserved under composition by the same argument as in Lemma 5.7. □

5.6.3.1 Operads and Almost-Symplectic Structures Over $\mathscr{F}_p$

An almost-symplectic vector space over $\mathscr{F}_p$ is a pair (V, ω) of a finite dimensional vector space over $\mathscr{F}_p$ and a function $\omega : V \times V \to \mathscr{F}_p$ which is non-degenerate, in the sense that for all $u \in V$ there is some v with $(u, v) \neq (0, 0)$ such that $\omega(u, v) \neq 0$. and with nontrivial Hochschild coboundary

$$d\omega(u, v, w) = \omega(v, w) - \omega(u + v, w) + \omega(u, v + w) - \omega(u, v) \neq 0\,.$$

In this case again the non-vanishing of the Hochschild coboundary implies that ω cannot be a linear map.

Note that here also the notion of almost symplectic structure we are considering is different from [4], hence the construction of representations and quantum codes described there does not apply directly to this case.

In this case, the (not necessarily Moufang) loop $\mathcal{L}(V, \omega)$ associated to the almost symplectic structure (V, ω) is obtained as the extension

$$0 \to \mathscr{F}_p \to \mathcal{L}(V, \omega) \to V \to 0$$

with the non-associative multiplication given by

$$(x, u) \star (y, v) = (x + y + \frac{1}{2}\omega(u, v), u + v)\,.$$

As shown in [4], the Moufang condition for $\mathcal{L}(V, \omega)$ is expressed as an identity satisfied by the function $d\omega$.

Let $\mathcal{V}^{\mathscr{F}_p}$ denote the space of almost symplectic structures (V, ω) over $\mathscr{F}_p$.

Theorem 5.5 *The space $\mathcal{V}^{\mathscr{F}_p}$ is an algebra over the operad $C_2^{\mathscr{F}_{p,s}}$ of strict colored p-ary little squares.*

Proof The argument is similar to Theorem 5.4, except for the coloring of the regions. Given an n-tuple of almost symplectic spaces (V_i, ω_i) and a strict colored p-ary little square $\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}) \in C_2^{\mathscr{F}_{p,s}}(n)$, we form a new

$$(V, \omega) = \gamma(\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}); (V_1, \omega_1), \dots, (V_n, \omega_n))$$

by associating to each (V_i, ω_i), with $V_i = \mathscr{F}_p^{N_i}$, a p-ary N_i-grid in $\mathcal{I}^2$ subdivided into regions $\mathcal{R}_\ell = \omega_i^{-1}(\ell)$ for $\ell \in \mathscr{F}_p$. This determines a colored p-ary little square $\mathcal{I}^2(\mathcal{R}_{i,0}, c^{i,(0)}; \mathcal{R}_i)$ associated to each (V_i, ω_i), which is strict because ω_i is non-degenerate. We then compose these according to the composition

$$\gamma(\mathcal{I}^2(\mathcal{R}_0, c^{(0)}; \mathcal{R}); \mathcal{I}^2(\mathcal{R}_{1,0}, c^{1,(0)}; \mathcal{R}_1), \dots, \mathcal{I}^2(\mathcal{R}_{n,0}, c^{n,(0)}; \mathcal{R}_n))$$

$$\gamma : C_2^{\mathscr{F}_{p,s}}(n) \times C_2^{\mathscr{F}_{p,s}}(k_1) \times \cdots \times C_2^{\mathscr{F}_{p,s}}(k_n)) \to C_2^{\mathscr{F}_{p,s}}(k_1 + \cdots + k_n)$$

obtained from repeated application of the compositions $\circ_i$ in the operad $C_2^{\mathscr{F}_{p,s}}$. This results in a new colored p-ary little square with regions $\mathcal{R}''_\ell$, which is also strict. This in turn defines the resulting almost symplectic space (V, ω) which has $V = \mathscr{F}_p^N$, with N the smallest natural number such that the subdivision into regions $\mathcal{R}''_\ell$ of the resulting colored p-ary little square is along the p-ary N-grid, and $\omega(u, v) = \ell$ for $(u, v) \in \mathcal{R}''_\ell$. This ω is non-degenerate because the little square is strict. □

5.6.4 Operad Partial-Action on Quantum Codes

A vector space $V = \mathscr{F}_q^N$, determines a corresponding complex vector space $(\boldsymbol{C}^q)^{\otimes N}$, representing a system of N q-ary qubits, endowed with the canonical basis given by the $|v\rangle$, labelled by the vectors $v \in V$.

We consider here the case of $\mathscr{F}_p$ with $p > 2$. The argument can be adapted to the case of characteristic 2 along the lines discussed in [4]. Let $\mathcal{L} = \mathcal{L}(V, \omega)$ be the loop obtained from an almost-symplectic (V, ω) over $\mathscr{F}_p$ as recalled above. It acts on $\mathcal{H} = \boldsymbol{C}[\mathcal{L}(V, \omega)]$ by left and right multiplication $((x, u) \star f)(y, v) = f((x, u) \star (y, v))$ and $(f \star (x, u))(y, v) = f((y, v) \star (x, u))$. This gives rise to a loop representation, which is the product $\mathcal{L} \times \mathcal{H}$ endowed with the (non-associative) multiplication

$$((x, u), f) \star ((y, v), f') = ((x, u) \star (y, v), (x, u) \star f' + f \star (y, v)) .$$

Consider the following sets Let $\mathcal{S}_1^\omega = \omega^{-1}(0) \cap \omega^{-1}(0)^\tau = \{(u, v) \in V \times V \mid \omega(u, v) = \omega(v, u) = 0\}$. Given a subset $\Omega \subset V \times V$, we write

$$\Omega_{ij} = \pi_{ij}^{-1}\Omega \subset V^L = \underbrace{V \times \cdots \times V}_{L-\text{times}},$$

with π_{ij} the projection onto the product of the i-th and j-th components. For $L \geq 2$ let $\mathcal{S}_L \subset V^L$,

$$\mathcal{S}_L^\omega = \bigcap_{1 \leq i < j \leq L} (\mathcal{S}_1^\omega)_{ij} .$$

Let $\mathcal{S} = \cup_{(V,\omega)\in\mathcal{V}^{\mathscr{F}_p}} \cup_{L\geq 1} \mathcal{S}_L^\omega$.

Given a character $\chi : \mathscr{F}_p \to \boldsymbol{C}^*$, let $\mathcal{H}_\chi \subset \mathcal{H}$ be the subspace of functions $f(x, u)$ that transform by $((x', 0) \star f)(x, u) = \chi(x')f(x, u)$. An element $\underline{u} = (u_1, \ldots, u_L) \in \mathcal{S}_L$ determines a set $\{\chi(x_1)E_{u_1}, \ldots, \chi(x_L)E_{u_L}\}_{x_i\in\mathscr{F}_p}$ of mutually commuting operators on $\mathcal{H}_\chi$. This determines a quantum code $\mathcal{Q}_{\chi,\underline{u}}^\lambda \subset \mathcal{H}$ given by a common eigenspace of these operators with eigenvalue λ. We refer to the quantum codes obtained in this way as "almost-symplectic quantum codes".

The operad action on almost-symplectic structures described in Theorem 5.5 induces a partial-action of the same operad on the data that determine these quantum codes. The notion of partial-action of an operad (partial-algebra over an operad) was introduced in [5].

Proposition 5.6 *The space $\mathcal{S}$ is a partial-algebra over the operad $C_2^{\mathscr{F}_p,s}$ of strict colored p-ary little squares.*

Proof In $\mathcal{S}^n = \underbrace{\mathcal{S} \times \cdots \times \mathcal{S}}_{n-\text{times}}$ consider the subspace

$$\mathcal{S}_0^n = \{(\underline{u}^{(1)}, \ldots, \underline{u}^{(n)}) \in \mathcal{S}_{L_1}^{\omega_1} \times \cdots \times \mathcal{S}_{L_n}^{\omega_n} \mid (\widetilde{\underline{u}}^{(1)}, \ldots, \widetilde{\underline{u}}^{(n)}) \in \mathcal{S}_{L_1+\cdots+L_n}^{\omega}\}$$

for $(V, \omega) = \gamma(\mathcal{I}(\mathcal{R}_0, c^{(0)}; \mathcal{R}); (V_1, \omega_1), \ldots, (V_n, \omega_n))$ the operad action of Theorem 5.3.4, and with $(\widetilde{\underline{u}_i})$ the vector in V obtained from the vectors $\underline{u}_i \in V_i$ under this composition. Then the operad action on $\mathcal{V}^{\mathscr{F}_p}$ induces a partial-action $\gamma : C_2^{\mathscr{F}_p,s}(n) \times \mathcal{S}_0^n \to \mathcal{S}$. □

Acknowledgments NC was supported by a Minerva grant, hosted at the Max Planck Institute for Mathematics in the Sciences; MM was supported by NSF grant DMS-2104330.

References

1. J.C. Baez, T. Fritz, T. Leinster, A characterization of entropy in terms of information loss. Entropy **13**(11), 1945–1957 (2011)
2. D. Borisov, Yu. Manin, Generalized operads and their inner cohomomorhisms, in *Geometry and Dynamics of Groups and Spaces (In memory of Aleksander Reznikov)*, ed. by M. Kapranov et al. Progress in Math., vol. 265 (Birkhäuser, Boston, 2008), pp. 247–308 arXiv:math.CT/0609748.
3. C. Chenavier, C. Cordero, S. Giraudo, Quotients of the magmatic operad: lattice structures and convergent rewrite systems. arXiv:1809.05083v2, 30 pp.
4. N.C. Combe, M. Marcolli, Yu. Manin, Moufang patterns and geometry of information. To be published in Collection, dedicated to Don Zagier, in *Pure and Applied Math. Quarterly*. arXiv:2107.07486. 42 pp.
5. I. Kříž, J.P. May, *Operads, Algebras, Modules and Motives*. Astérisque No. 233 (1995), iv+145pp.
6. Yu. Manin, Some remarks on Koszul algebras and quantum groups. Ann. Inst. Fourier **XXXVII**(4), 191–205 (1987)
7. Yu. Manin, *Quantum groups and non–commutative geometry* (Publ. de CRM, Université de Montréal, 1988), 91 pp.
8. Yu. Manin, *Topics in Noncommutative Geometry* (Princeton University Press, Princeton, 1991), 163 pp.
9. Yu. Manin, Notes on quantum groups and quantum de Rham complexes. Teoreticheskaya i Matematicheskaya Fizika **92**(3), 425–450 (1992). Reprinted in Selected Papers of Yu.I. Manin (World Scientific, Singapore 1996), pp. 529–554
10. M. Marcolli, R. Thorngren, Thermodynamic semirings. J. Noncommut. Geom. **8**(2), 337–392 (2014)
11. M. Markl, *Operads and PROPs*. Handbook of Algebra, vol. 5 (Elsevier/North-Holland, 2008), pp. 87–140
12. J.P. May, *The Geometry of Iterated Loop Spaces*. Lecture Notes in Mathematics, vol. 271 (Springer, Berlin, 1972)
13. J.D.H. Smith, Quantum quasigroups and loops. J. Algebra **456**, 46–75 (2016)
14. V. Vallette, A Koszul duality for PROPs. arXiv:math0411542v3, 78 pp.
15. A. Voronov, The A_∞ operad and A_∞-algebras. Lecture Notes (2001). https://www-users.cse.umn.edu/~voronov/8390/lec9.pdf

Chapter 6
Quantum Computational Complexity with Photons and Linear Optics



Jian-Wei Pan

Abstract On this wonderful occasion to celebrate Professor Yang Chen Ning's 100th birthday, we are deeply honored to contribute a chapter on our recent study on the demonstration of quantum computational advantage, which is a dialogue among mathematics, computational complexity, and quantum optics. The concept of quantum computational advantage denotes a milestone that a quantum device can solve a specific mathematical problem that no classical computer can solve within a reasonable amount of time. Just like using the Bell experiments to refute Einstein's local hidden variable model and establish the quantum entanglement as a valuable resource with strong-than-classical correlation, quantum computational advantage experiments provide evidences to refute the Extended Church-Turing thesis and prove the potential of faster-than-classical computation using quantum mechanics, a widely believed but unproven conjecture proposed by Richard Feynman about forty years ago. The first feasible proposal to demonstrate the quantum computational advantage is boson sampling based on multi-photon interference, where non-classical light is injected into a linear optical network, and the output highly random, photon-number- and path-entangled state is measured by single-photon detectors. We perform experiments which have produced up to 113 photon detection events out of a 144-mode photonic circuit. These rudimentary photonic quantum computers, *Jiuzhang*, yield an output state space dimension of 10^{43} and a sampling rate that is 10^{10} faster than using the state-of-the-art simulation strategy and supercomputers.

6.1 Introduction

The Extended Church-Turing Thesis, which is a foundational tenet in computer science, states that all computational problems solvable efficiently by realistic physical systems, also can be efficiently simulated by a (classical) probabilistic

J.-W. Pan (✉)
Chinese Academy of Sciences (CAS) Center for Excellence in Quantum Information and Quantum Physics, University of Science and Technology of China, Hefei, China
e-mail: pan@ustc.edu.cn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_6

Turing machine [1]. However, in 1994, Peter Shor invented a quantum algorithm [2] for factoring large numbers. Shor showed that integer factoring, believed to be an NP problem for classical computers, can be efficiently solved by a universal quantum computer. If a quantum computer is built to run a sufficiently large Shor's algorithm, this would directly disprove the Extended Church-Turing Thesis.

However, running Shor's algorithm on a quantum computer to factorize, say a 300-digit large number, would require tens of millions of physical qubits with a dauntingly high state and gate fidelity and quantum error correction for fault tolerance, which is far beyond the capability of current technology.

In 2010, Aaronson and Arkhipov proposed an intermediate quantum computational model called boson sampling [3]. They showed that sampling [4–6] from the output state of a quantum machine built with single-photon sources, passive linear optical network and single photon detectors, is related to a #P-complete problem for classical computers. The best known classical algorithm suggests that a boson sampler with about 50 photons can outperform the most powerful supercomputers. Such a milestone was first termed as quantum supremacy [7, 8] by Preskill in 2012. Here we use a more mild term: quantum computational advantage.

Interestingly, while the boson sampling is much easier to implement experimentally, it can provide more compelling complexity-theoretical evidence for the quantum computational speedup than Shor's algorithm, because the mathematical problems related to the boson sampling are, as shown below, in computational complexity's language, much more difficult than factoring.

This chapter is a contribution to the Volume "Dialogues between Physics and Mathematics: C. N. Yang at 100". The chapter features a dialogue among mathematics, computational complexity, and quantum optics, and is organized as follows. First, we introduce two matrix functions, the Permanent and Hafnian, which are #P-hard mathematical problems. Second, we discuss a physical model of quantum computation with noninteracting photons—boson sampling, which links to mathematical problems of the Permanent and Hafnian. Finally, we report the quantum optical experiment, realizing large-scale boson sampling with more than 100 photons, which produce results that are far beyond the capability of classical simulations using supercomputer.

Mathematics is the beautiful language for physicists to describe the nature. This chapter shows an interestingly new path where a computationally high-complexity mathematics elegantly links to a physical model and drives the quantum computation experiments to reach an unprecedented computational capability, which, in return, refutes the Extended Church-Turing thesis.

6.2 The Mathematics: Permanent and Hafnian

6.2.1 Permanent

In linear algebra, the Permanent of an $n \times n$ matrix A is defined as:

$$\mathrm{Per}(A) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} A_{i,\sigma_i} \tag{6.1}$$

where S_n is the group of permutations, σ runs over all elements of S_n, σ_i is the i_{th} elements. According to the definition in Eq. (6.1), Permanent is exactly the sum of possibilities of all permutations. This relation was first discovered by Caianiello in 1953 [9], when he computed the Dyson's S-matrix. In 1979, Valiant proved that calculating Permanent of (0,1)-matrix is #P-complete hard [10]. In 2004, Scheel showed that the amplitude of unitary transformation of multi-mode Fock state is linked to the computation of Permanents [11].

To efficiently calculate Permanent, the most famous algorithm is Ryser's algorithm [12] based on inclusion-exclusion principle:

$$\mathrm{Per}(A) = (-1)^n \sum_{S \subseteq \{1,2,\cdots,n\}} (-1)^{|S|} \prod_{i=1}^{n} \sum_{j \in S} A_{i,j} \tag{6.2}$$

where S is a subset of $\{1, 2, \cdots, n\}$, $|S|$ is the number of elements in S. The Ryser's algorithm requires $(2^n - 1)(n - 1)$ multiplications and $(2^n - 2)(n + 1)$ additions, which is much faster than brute-force algorithm. Another algorithm as fast as Ryser's algorithm is called Glynn's algorithm, which runs as follows:

$$\mathrm{Per}(A) = \frac{1}{2^{n-1}} \sum_{\delta} \left(\prod_{k=1}^{n} \delta_k \right) \prod_{j=1}^{n} \sum_{i=1}^{n} \delta_i A_{i,j} \tag{6.3}$$

in which the outer sum runs over all 2^{n-1} vectors $\delta = (\delta_1 = 1, \delta_2, \delta_3, \cdots, \delta_n) \in \{\pm 1\}^n$. One advantage of Glynn's algorithm is that all terms are divided by a factor of 2^{n-1}, which can improve the accuracy of Permanent when n is large enough.

6.2.2 Hafnian

The definition of Hafnian of a $2n \times 2n$ symmetric matrix G is:

$$\mathrm{Haf}(G) = \frac{1}{n! 2^n} \sum_{\sigma \in S_{2n}} \prod_{i=1}^{n} G_{\sigma_{2i-1},\sigma_{2i}} \tag{6.4}$$

where S_{2n} is the perfect matching permutation group on $[2n]$, and σ is one of the permutations.

In graph theory, Hafnian corresponds to the number of perfect matchings of an adjacent matrix, while Permanent counts the number of perfect matchings in a bipartite graph. In view of this, we can easily check that

$$\mathrm{Haf}\begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} = \mathrm{Per}(A). \tag{6.5}$$

The connection between boson sampling and graph theory enables many possible applications, which will be introduced in Sect. 6.6.

6.3 The Model: Boson Sampling

In 2010, Aaronson and Arkhipov investigated the computational complexity of non-interacting single photons in a linear optical network. As shown in Fig. 6.1, n identical photons are simultaneously injected into an m-mode ($n\langle\langle m$) passive linear network. In the output, n photons are detected by m single-photon detectors. The task of this model is to get samples from the final state distribution.

The m-mode interferometer in Fig. 6.1 can be described by an $m \times m$ unitary matrix U, of which the complex element $U_{i,j}$ contains the amplitude and phase information from the i_{th} input mode to j_{th} output mode. For simplicity, let's first consider the non-collision case, meaning that n input photons show up at n specific output ports. In this case, there are $\binom{m}{n}$ output possibilities, corresponding to an S_n symmetric permutation group. Since there is nothing changed with wave function if the positions of two identical photons are exchanged, all output possibilities should

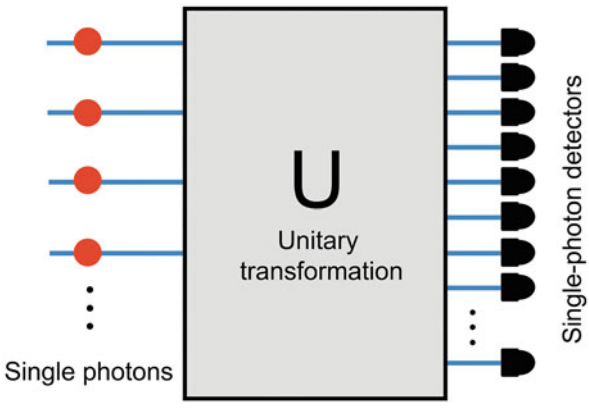


Fig. 6.1 Model of boson sampling with single photons. There are three main parts: single-photon sources, linear network, and single-photon detectors. First, n single photons pass through an m-mode ($n\langle\langle m$) linear network. At the output, photons are detected by arrays of single-photon detectors

coherently contribute to the final state, which can be written as:

$$\mathrm{P} = \left| \sum_{\sigma \in S_n} \prod_{i=1}^{n} u_{i,\sigma_i} \right|^2 . \tag{6.6}$$

The right side of Eq. (6.6) is exactly the modulus square of Permanent of a submatrix:

$$\mathrm{Per}(U) = \sum_{\sigma \in S_n} \prod_{i=1}^{n} U_{i,\sigma_i} . \tag{6.7}$$

Considering all collisions of input and ouput ports, the distribution of boson sampling can be written as

$$P(I \rightarrow O) = \frac{|\mathrm{Per}(U_{I,O}))^2}{I_1!...I_m!O_1!...O_m!} \tag{6.8}$$

in which $\sum_{i=1}^{m} I_i = n$ and $\sum_{i=1}^{m} O_i = n$, the $I = I_1, ..., I_m$ and $O = O_1, ..., O_m$ are the input and output configurations, respectively.

Aaronson and Arkhipov developed a beautiful theory connecting linear optics to computational complexity, and proved that, if there exists a polynomial-time classical algorithm that samples from the same probability distribution as a linear-optical network, then the polynomial hierarchy collapses to the third level. Note that the exact sampling is not applicable to implement an experiment, due to imperfections such as the photon losses, nonidentical photon sources, and dark counts of the detectors. Aaronson and Arkhipov further proved that with two strongly believed but unproven conjectures—Permanent-of-Gaussians Conjecture and Permanent Anti-Concentration Conjecture, approximate boson sampling also can't be solved in polynomial time, or the polynomial hierarchy will collapse to the third level.

A relevant proposal for the demonstration of quantum computational advantage is random circuit sampling, which is suitable for implementations in quantum circuit based models such as using the superconducting qubits [6, 13–15].

In 2017, Hamilton et al. [16] proposed a variant of AABS by changing the single-photon source into single-mode squeezed states (SMSS)

$$|\xi\rangle = S(\xi)|0\rangle, \tag{6.9}$$

which can be produced from strongly pumping a nonlinear crystal. The squeezing operator is

$$S(\xi) = \exp\left(\frac{1}{2}\xi^* a^2 - \frac{1}{2}\xi (a^\dagger)^2\right) \tag{6.10}$$

with $\xi = re^{i\phi}$. Unlike single-photon sources, squeezed states allow a superposition of even photon number states rather than one photon at the input, thus significantly enhance the photon generation and detection rate.

We first look at the output distribution with SMSS inputs. The unitary transformations $S(\xi)^\dagger(a^\dagger, a)^T S(\xi)$ of the creation and annihilation operators are:

$$\begin{aligned} S(\xi)^\dagger a S(\xi) &= \cosh r a - e^{i\phi} \sinh r a^\dagger \\ S(\xi)^\dagger a^\dagger S(\xi) &= \cosh r a^\dagger - e^{-i\phi} \sinh r a \end{aligned} \tag{6.11}$$

As the covariance matrix of the vacuum state is $\sigma_0 = \frac{1}{2} I$, the covariance matrix of the SMSS is

$$\begin{aligned} \sigma_{SM} &= \frac{1}{2} M \sigma_0 M^T \\ &= \frac{1}{4} \begin{pmatrix} \cosh 2r(1 + e^{-2i\phi}) + 1 - e^{-2i\phi} & -2\cos\phi \sinh 2r \\ -2\cos\phi \sinh 2r & \cosh 2r(1 + e^{2i\phi}) + 1 - e^{2i\phi} \end{pmatrix}, \end{aligned} \tag{6.12}$$

where M is the transformation matrix which can be extracted from the Eq. (6.11)

$$M = \begin{pmatrix} \cosh r & -e^{-i\phi} \sinh r \\ -e^{i\phi} \sinh r & \cosh r \end{pmatrix} \tag{6.13}$$

This result can be directly extended to m SMSS case

$$M_{2m} = \begin{pmatrix} \oplus_{i=1}^m \cosh r_i & \oplus_{i=1}^m e^{-i\phi_i} \sinh r_i \\ \oplus_{i=1}^m - e^{i\phi_i} \sinh r_i & \oplus_{i=1}^m \cosh r_i \end{pmatrix} \tag{6.14}$$

with a basis of $(a_1, a_2, ..., a_m, a_1^\dagger, a_2^\dagger, ..., a_m^\dagger)$. Thus, if these m SMSSs travel through an m-mode interferometer U, the corresponding covariance matrix becomes

$$\sigma_{out} = \frac{1}{2} \begin{pmatrix} U & 0 \\ 0 & U^* \end{pmatrix} M_{2m} M_{2m}^\dagger \begin{pmatrix} U^\dagger & 0 \\ 0 & U^T \end{pmatrix} \tag{6.15}$$

It is worth noting that the covariance matrix can fully describe the output Gaussian state.

Hamilton et al. [16, 17] considered k SMSSs propagate through an m-mode ($k < m$) linear optical network and finally sampling the output distribution at Fock state basis. Since the input state is a Gaussian state, this model is called as Gaussian boson sampling (GBS). The authors derived that: If the covariance matrix of the k SMSSs

is σ, then the possibility to detect a specific pattern $\overline{n}$ is

$$\Pr(\overline{n}) = \frac{1}{\overline{n}!\sqrt{\mathrm{Det}\left(\sigma_Q\right)}}\mathrm{Haf}\left(A_S\right) \tag{6.16}$$

where the $\sigma_Q = \sigma + I_{2m}/2$, A_S is the $2k \times 2k$ submatrix, and the matrix A is

$$A = \begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix}\left(I_{2m} - \sigma_Q^{-1}\right). \tag{6.17}$$

Here Haf is the abbreviation of Hafnian, and:

$$\mathrm{Haf}(A) = \frac{1}{n!2^n}\sum_{\sigma \in S_{2n}}\prod_{i=1}^{n} A_{\sigma_{2i-1},\sigma_{2i}} \tag{6.18}$$

In this case, photon number-resolving detectors should be employed to count the photon number for every output port.

In 2018, Quesada et al. [18] showed that if threshold detectors are used in the GBS, then the probability to detect a specific pattern should be

$$\Pr(S) = \frac{\mathrm{Tor}\begin{pmatrix} 0 & I_m \\ I_m & 0 \end{pmatrix} A_S}{\sqrt{\mathrm{Det}\left(\sigma_Q\right)}} = \frac{\mathrm{Tor}\left(I_{2m} - \sigma_Q^{-1}\right)_S}{\sqrt{\mathrm{Det}\left(\sigma_Q\right)}}, \tag{6.19}$$

where

$$\mathrm{Tor}(A) = \sum_{Z \in P([n])}\frac{(-1)^{|Z|}}{\sqrt{\mathrm{Det}(I - A_Z)}} \tag{6.20}$$

is called Torotonian of the matrix A, and $P([n])$ is power set (containing 2^n subsets) of $[n] = \{1, 2, \cdots, n\}$.

The GBS is strongly related to the AABS, since Permanent is just a special case of Hafnian. Computational complexity evidences of approximate GBS are reported [16, 17, 19, 20]. For example, Grier et al. [19] introduced bipartite GBS where the first half of modes from two-mode squeezed states are injected into an interferometer while the other half of modes pass through another interferometer. In this case, the final distribution is proportional to modulus square of Permanent rather than Hafnian. Next, using the same assumptions in the AABS, it was demonstrated that approximate bipartite GBS can not be efficiently simulated by classical computers unless the polynomial hierarchy collapses. Deshpande et al. [20] showed that, with a random-matrix theory conjecture, the GBS should be as hard as the AABS.

6.4 Single-Photon Boson Sampling Experiments

Soon after the AABS proposal, four independent proof-of-principle experiments [21–24] were demonstrated in 2013 with three photons. All these experiments are based on SPDC pseudo-single photon sources, where the source efficiencies are typically smaller than 0.01. While these experiments confirmed that the amplitudes of three-photon process are given by permanets of the 3×3 complex matrices, they don't provide any speedup, nor are the results suprising: at some level they merely confirmed quantum mechanics [25]. Several experiments followed to validate boson sampling [26, 27], to investigate generalized multi-photon interference [28].

To increase the number of photons, significant efforts have been made to develop on-demand high-efficiency single-photon sources based on semiconductor quantum dots [29–31]. In 2017, Wang et al. [32] developed two central components for high-efficiency boson sampling: an actively demultiplexed quantum-dot single photon source combining high purity, indistinguishability, and efficiency simultaneously, and a compact multiphoton interferometer with a transmission efficiency of 99%. This experiment realized the first boson sampling with the sampling rate beating the simulation capability of the early classical computers ENIAC and TRADIC. Meanwhile, He et al. [33] developed a time-bin encoding method for fully electrically programmable boson sampling.

In 2019, Wang et al. [34] demonstrated AABS with 20 input photons and a 60-mode interferometer (see Fig. 6.2). A single InAs/GaAs quantum dot, resonantly coupled to a micropillar cavity, is used to create on-demand single photons. At π pulse, $\sim$16.3 million single photons per second are directly recorded by a superconducting nanowire single-photon detector. 19 pairs of Pockels cells (PCs) and polarized beam splitters (PBSs) are utilized to actively translate photon pulses into 20 spatial modes. Optical fibers with different lengths are applied to compensate time delays. The 20 input single photons are injected into a 3D integrated, 60-

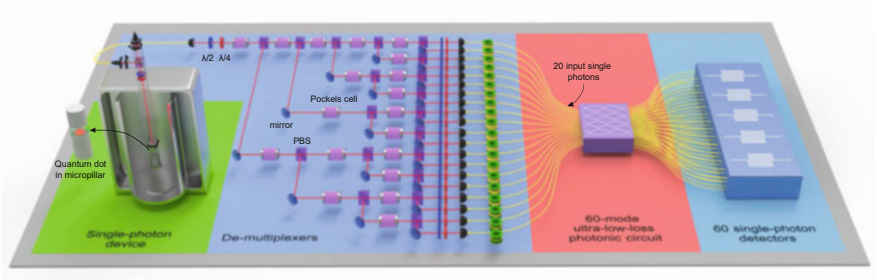


Fig. 6.2 Experimental setup of 20-photon boson sampling. A quantum dot-cavity coupled system emits high-performance single photons, which then are immediately demultiplexed into 20 spatial modes. The different time delay are compensated by fibers with specific lengths. When passing through a 60-mode 3D interferometer, the output photons are detected by 60 superconducting nanowire single-photon detectors

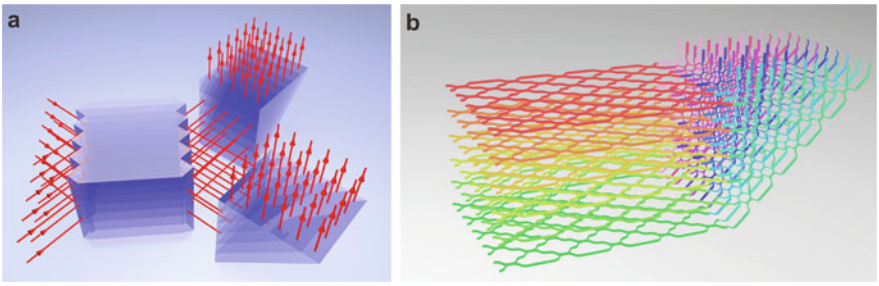


Fig. 6.3 (**a**) A schematic diagram of 3D interferometer, with 60 spatial modes. (**b**) The equivalent circuit of 3D interferometer, which indicates that this photonic circuit is fully connected

mode ultra-low-loss photonic circuit, which consists of 396 beam splitters and 108 mirrors, as shown in Fig. 6.3. Finally, the output single photons are detected by 60 superconducting nanowire single-photon detectors with efficiencies ranging from 0.6 to 0.82. All coincidences are recorded by a 64-channel coincidence count unit. Owing to the high-efficiency single-photon source and ultra-low-loss photonic circuit, at most 14 photons are detected at the output, with a coincidence rate of $\sim$6 per hour. The output state Hilbert space reaches up to 3.7×10^{14} dimensions, which is over 10 orders of magnitude larger than the previous work.

6.5 Quantum Computational Advantage with *Jiuzhang*

Realizing a large-scale GBS faced a number of challenges. It needs (i) arrays of low-noise SMSS sources with high indistinguishability and efficiency, (ii) a large interferometer with full connectivity, matrix randomness, near-perfect wave-packet overlap and phase stability, and near-unity transmission rate, (iii) phase control of the SMSSs from the source to the interferometer, (iv) high-efficiency detectors in the output, and (v) validation of the obtained sparse samples from a huge output Hilbert space, and benchmark the performance of the GBS against a supercomputer (Fig. 6.4).

These challenges have been overcome simultaneously in the experiment of photon quantum computer, *Jiuzhang*, named in honor of an ancient Chinese mathematical classic, "*The Nine Chapters on the Mathematical Art*".

We developed a stimulated squeezed light source with high brightness, and simultaneously near-unity purity and efficiency for scalable GBS [35, 36]. As shown in Fig. 6.5a, the pump pulse will double pass a PPKTP crystal, and two generated squeezed photons will interference with each other constructively. Figure 6.5b and c are measured purities and collection efficiencies of all 25 two-mode squeezed states (TMSSs), corresponding to averaged values of 0.96 and 0.92, respectively.

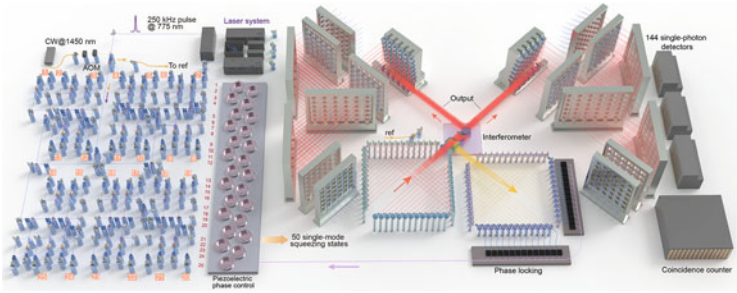


Fig. 6.4 Experimental setup of *Jiuzhang*. It is mainly composed of five parts. At the upper left region, high-intensity transform-limited pulse laser with wavelength of 775 nm are prepared to pump 25 TMSS sources (at the left region, labeled in orange). Meanwhile, continuous-wave 1450-nm laser are guided and co-propagates with the 25 TMSS sources. The 1550-nm two-mode squeezed light is collected into temperature-insensitive single-mode fiber, of which 5-m bare fiber is winded around a piezo-electric (PZT) cylinder to control the source phase (at the center region). In the center-right region, by using optical collimators and mirrors, 25 TMSSs are injected into a photonic network and 25 corresponding light beams (colored in yellow) with wavelength of 1450 nm and intensity power of about 0.5 μW are collected for phase locking. The 144 output modes are distributed into four parts using arrays of tunable periscopes and mirrors. Finally, the output modes are detected by 144 superconducting nanowire single-photon detectors and registered by a 144-channel ultra-fast electronics coincidence unit

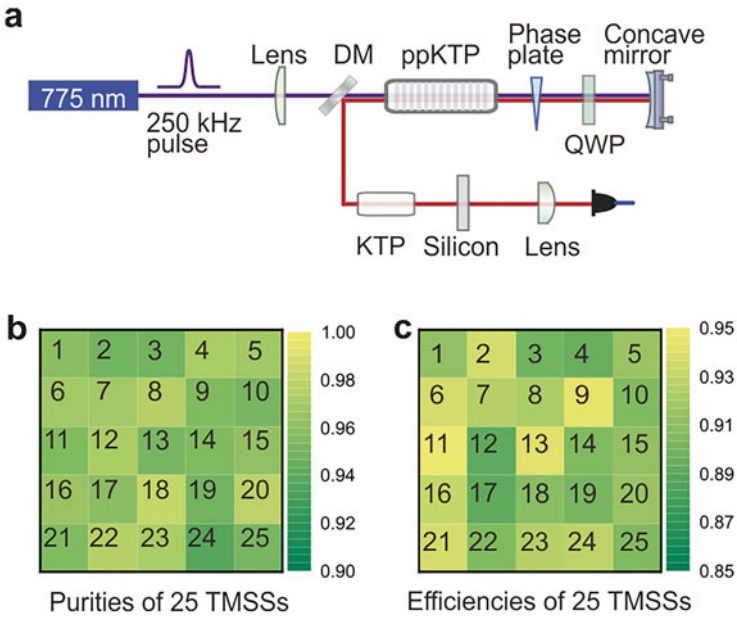


Fig. 6.5 (**a**) Squeezed light source amplified by stimulated parametric down-conversion. (**b**) Measured purities of all 25 TMSSs, the averaged value is 0.96. (**c**) Measured collection efficiency of all 25 light sources with an average of 0.92

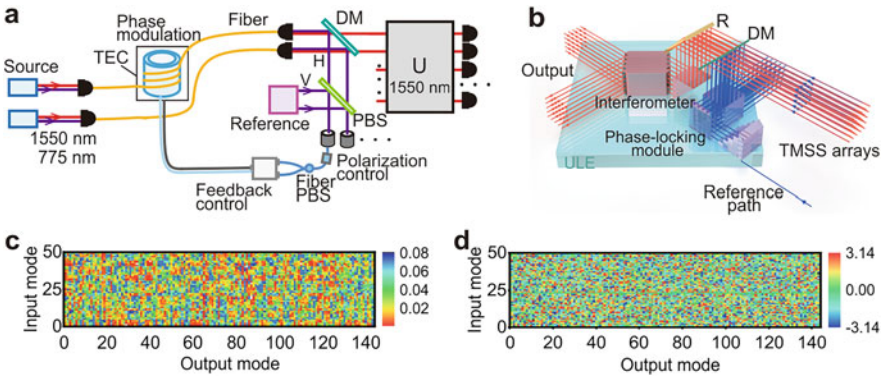


Fig. 6.6 (**a**) Schematic diagram of the active phase locking system. A pump laser beam is used as a reference for all the squeezed states. The pump laser pulses which share the same propagation path as the downconverted photons, are then combined on a beamsplitter with the reference laser pulse. A balanced detection scheme, which is insensitive to laser power fluctuation, is used to read out the phase information. (**b**) Passive phase stabilization. The interferometer is adhered onto an ultralow-expansion glass plate. The blue light paths are for the interference of the 25 pumping lasers with the reference laser. The red light paths are the input and output of the photonic network. (**c**) The measured 7200 amplitudes of the interferometer. (**d**) The measured 7200 phases of the interferometer

The whole optical set-up, from the photon generation to the interferometer, must be phase locked in the presence of various environmental perturbations. To this end, we developed active phase locking over the optical path (see Fig. 6.6a) and passive stabilization inside the interferometer (see Fig. 6.6b). The phase fluctuation of the whole system is controlled within 0.06 rad for a duration of one hour, which causes a ∼0.5% decrease of the photon interference visibility.

A compact, three-dimensional interferometer made from micro-optics is used to realize a 144 × 144 unitary transformation. The relative path length difference between different ports is calibrated within a standard deviation of 1.6 μm, indicating that the wave-packet overlap inside the interferometer is better than 0.999. The GBS matrix absorbs both the unitary transformation of the interferometer and the squeezing parameters and phases of the Gaussian input state. We reconstruct the corresponding unitary matrix of the spatial-polarization hybrid encoded 144×144 interferometer as plotted in Fig. 6.6c and d for the elements of amplitudes and phases, respectively. Further analysis show that the matrix is both unitary and Haar-random, fulfilling the hardness arguments of the boson sampling. The output state is detected by 144 high-efficiency superconducting nanowire single-photon detectors.

We categorized our experiment into three different regimes which we call as easy regime, sparse regime, and intractable regime. The easy regime is at small scales where we can obtain the full output distribution, as in most previous proof-of-principle experiments. The size of the output Hilbert spaces is a function of both the photon and mode number, which grows exponentially. When the photon click number is well above 4, then only a small fraction of output combinations

can obtain one event, while most output will have zero events, which is called the sparse regime. There it becomes infeasible to register the whole output distribution, but only sampling is possible. When the output click number exceeds ~30, even the calculations required for the validation methods become intractable for classical computers, we call it as intractable regimes. An important yet open question is how to verify the large-scale GBS.

In easy regime, two important figure of merits—fidelity and total variation distance between experimental and theoretical distributions—can quantify the performance of boson sampling. For a perfect boson sampler, the fidelity ($F = \sum_i \sqrt{q_i p_i}$) should be equal to 1 and the total variation distance ($D = 1/2 \sum_i |q_i - p_i|$) should be 0, where p_i and q_i denote the theoretical and experimental probability of the i_{th} basis, respectively. We first tested with three pairs of input TMSSs and two-photon clicks in the output. The obtained distribution is plotted in Fig. 6.7a, corresponding to an averaged fidelity of 0.990 ± 0.001, and an averaged distance of 0.103 ± 0.001. The data for all 23 different input configurations are shown in Fig. 6.7b, which further confirms that the GBS set-up is functioning as expected.

Next, we move to sparse regimes to provide supporting or circumstantial evidences for the correct operations of GBS. We performed GBS at different laser pump power (from 0.15 W to 1.6 W) and focal waists (65 μm and 125 μm), which gives rise to different photon number distributions. Figure 6.8a displays the corresponding four different photon click distributions, where the maximal output click number ranges from 32 to 113. To validate the GBS results, we first compared the distribution of total output clicks of GBS with the possible mockups. The most plausible hypotheses in this experiment include using coherent light sources (lasers) as input, distinguishable photons (owing to mode mismatch and imperfect light sources), thermal states (due to photon loss), and uniform random outcome. Figure 6.8b shows that the experimental GBS data (red) perfectly overlap with the theoretical curve (orange), while strongly deviate from the other two mockups.

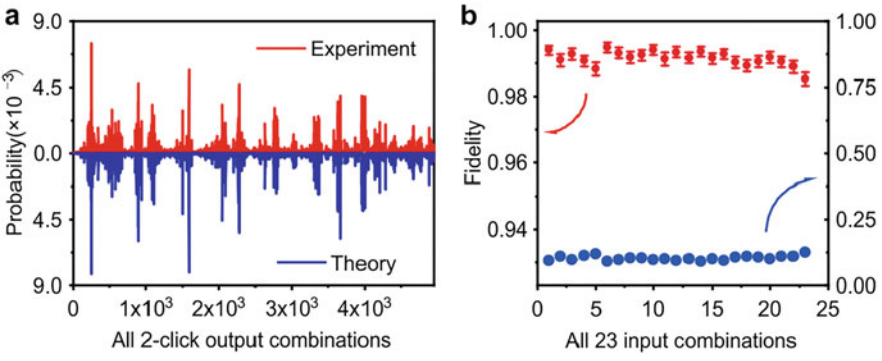


Fig. 6.7 (**a**) Experimental (red) and theoretical (blue) two-photon distribution with three TMSSs input. (**b**) Summary of statistical fidelity and total variation distance of two-photon distribution for 23 different input sets

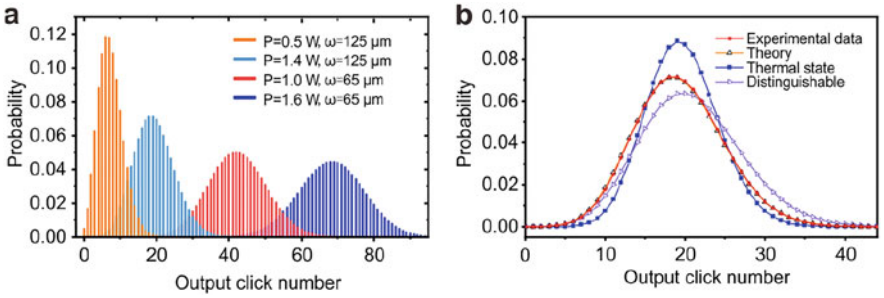


Fig. 6.8 (**a**) Photon clicks distribution at different pump power and focal waists. (**b**) Output photon-click distribution of the experiment (red), theoretical prediction (orange), thermal state (blue), and distinguishable photons (purple)

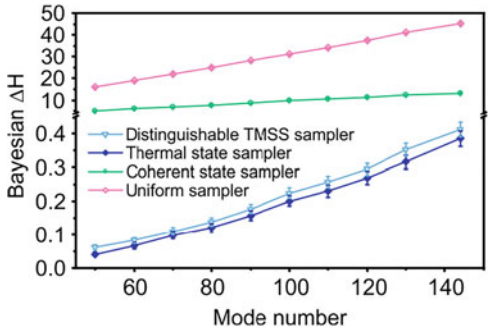


Fig. 6.9 The calculated Bayesian confidence ΔH as a function of the subsystem mode number

We apply Bayesian ΔH to measure the strength of the validation. The Bayesian ΔH is defined as

$$\Delta H = \log \chi_N / N, \tag{6.21}$$

where, N is the number of samples, and χ_N is

$$\chi_N = \prod_{k=1}^{N_{\text{event}}} \frac{q_k}{p_k}, \tag{6.22}$$

in which q_k and p_k are the probabilities associated with ideal GBS and mockup sampler, respectively. In this case, a larger value of ΔH indicates a larger deviation between the GBS and the mockups.

The low pumping intensity regime with fewer than 40 maximal photon clicks is within the computational capability of classical supercomputers. Thus, we can validate the data over the full range of photon clicks and size. Figure 6.9 shows the calculated Bayesian confidence ΔH as a function of the subsystem mode number.

For all the hypotheses, it is clear that the validation confidence is stronger for a larger mode number.

Then, we enter the computationally intractable regime under high pumping intensity, where calculating, say a submatrix Torontonian for at most 40 photons, is infeasible. To tame the validation in this regime, we propose a computationally friendly method. The idea is that, we start from a subsystem with fewer output bosonic modes, which is computationally easier to study. The Bayesian confidence ΔH are calculated and plotted with varying mode size for the validations against the uniform samplers (Fig. 6.10a), coherent state (Fig. 6.10b), distinguishable photons (Fig. 6.10c) and thermal states (Fig. 6.10d). For all the tested data, we observe that not only $\Delta H > 0$, but also it follows the same increasing trend as in the low intensity regime. This allows us to infer that the same setup with all modes in the quantum advantage regime would be validated with an even stronger confidence.

Finally, we estimate the classical computational cost to simulate the GBS. We choose the benchmarking algorithm [18] by calculating Torontonian on the Sunway TaihuLight supercomputer. The estimated run time for different photon clicks are plotted in Fig. 6.11, which indicates the sampling rate of *Jiuzhang 2.0* is $\sim 10^{24}$ times faster than classical supercomputers using brute-force simulation. A very recent improved classical algorithm by Bulmer et al. [37] narrows this gap to $\sim 10^{10}$.

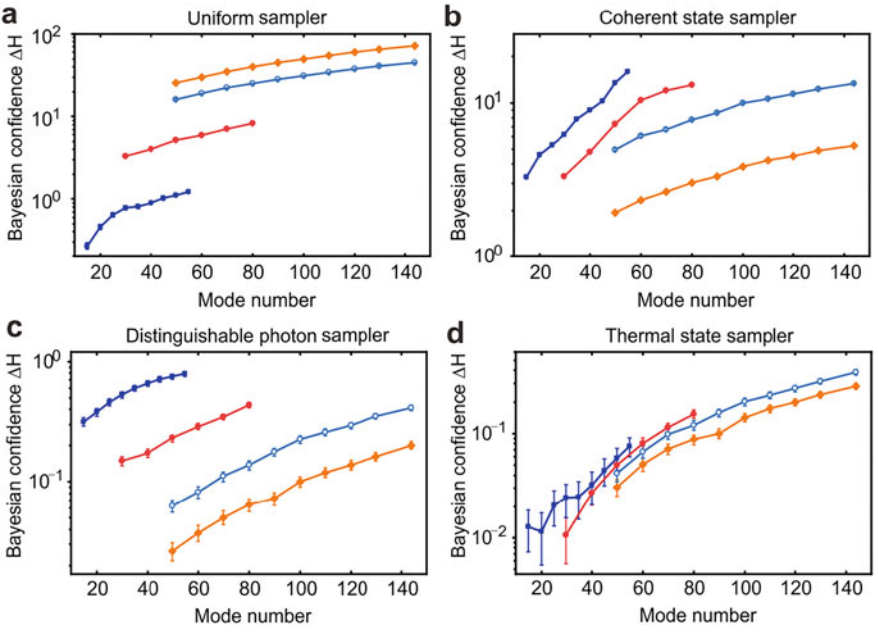


Fig. 6.10 The experimental results of Bayesian confidence ΔH at different laser pump power and focal waists for the validations against the uniform samplers (**a**), coherent state (**b**), distinguishable photons (**c**), and thermal state (**d**)

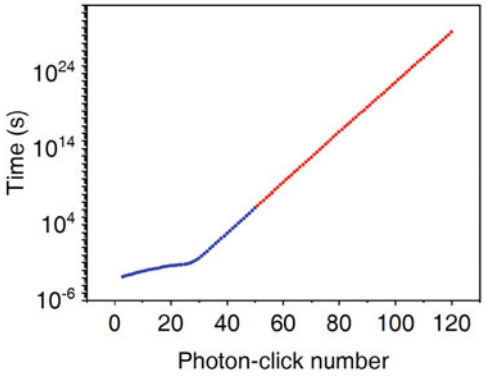


Fig. 6.11 The computational time for classical simulation on a supercomputer. The execution time of one Torontonian function on Sunway TaihuLight with input matrix of size up to N = 50 (the blue points). We extrapolate the execution time to N=120 (the red points). In the regime of $N < 30$ photons, the Sunway TaihuLight system is not fully loaded. For $N > 30$, the system becomes fully loaded and the scaling is linear in the logarithmic time

6.6 Applications

One potential application of boson sampling is to generate certified random number [38]. Because the probability distribution of GBS is related to Hafnian, which corresponds to the number of perfect matching, the GBS links to several potentially practical applications including certifiable random numbers, graph optimization [39–41], graph similarity [42], point process [43], molecular docking [44], quantum chemistry [45, 46], and quantum machine learning [47]. A natural next step would be to use the GBS quantum computer developed in our experiment as a special-purpose photonic platform to investigate these real-world applications, as a step toward noisy intermediate-scale quantum processing [48]. Finally, the quantum optical set-up consisting of squeezed states fed into a linear optical network and followed by post-selection measurement outcome, can be used as a factory for heralded production of different family of quantum states, such as Fock state, cat state, NOON state, and the Gottesman-Kitaev-Preskill code for quantum error correction [49].

Acknowledgments I would like to thank Chao-Yang Lu and Hui Wang for assistance in preparing this article.

References

1. E. Bernstein, U. Vazirani, in *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, San Diego, CA, 1993
2. P.W. Shor, Algorithms for quantum computation: discrete logarithms and factoring, in *Proceedings 35th Annual Symposium on Foundations of Computer Science* (IEEE, 1994), pp. 124–134
3. S. Aaronson, A. Arkhipov, The computational complexity of linear optics, in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, pp. 333–342 (2011)
4. B.M. Terhal, D.P. DiVincenzo, Adaptive quantum computation, constant depth quantum circuits and arthur-merlin games. Preprint (2002). arXiv:quant-ph/0205133
5. M.J. Bremner, A. Montanaro, D.J. Shepherd, Average-case complexity versus approximate simulation of commuting quantum computations. Phys. Rev. Lett. **117**(8), 080501 (2016)
6. S. Aaronson, L. Chen, Complexity-theoretic foundations of quantum supremacy experiments, in *Proceedings of the 32nd Computational Complexity Conference*, pp. 1–67 (2017)
7. J. Preskill, Quantum computing and the entanglement frontier-rapporteur talk at the 25th solvay conference, (2012)
8. A.W. Harrow, A. Montanaro, Quantum computational supremacy. Nature **549**(7671), 203–209 (2017)
9. E.R. Caianiello, On quantum field theory—i: explicit solution of dyson's equation in electrodynamics without use of feynman graphs. Il Nuovo Cimento (1943–1954) **10**(12), 1634–1652 (1953)
10. L.G. Valiant, The complexity of computing the permanent. Theor. Comput. Sci. **8**(2), 189–201 (1979)
11. S. Scheel, Permanents in linear optical networks (2004). arXiv:Quantum Physics
12. H.J. Ryser, *Combinatorial Mathematics*, vol. 14 (American Mathematical Soc., 1963)
13. S. Boixo, S.V. Isakov, V.N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M.J. Bremner, J.M. Martinis, H. Neven, Characterizing quantum supremacy in near-term devices. Nature Physics **14**(6), 595–600 (2018)
14. F. Arute, K. Arya, R. Babbush, D. Bacon, J.C. Bardin, R. Barends, R. Biswas, S. Boixo, F.G.S.L. Brandao, D.A. Buell, et al., Quantum supremacy using a programmable superconducting processor. Nature **574**(7779), 505–510 (2019)
15. Y. Wu, W.-S. Bao, S. Cao, F. Chen, M.-C. Chen, X. Chen, T.-H. Chung, H. Deng, Y. Du, D. Fan, et al., Strong quantum computational advantage using a superconducting quantum processor. Phys. Rev. Lett. **127**(18), 180501 (2021)
16. C.S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, I. Jex, Gaussian boson sampling. Phys. Rev. Lett. **119**(17), 170501 (2017)
17. R. Kruse, C.S. Hamilton, L. Sansoni, S. Barkhofen, C. Silberhorn, I. Jex, Detailed study of gaussian boson sampling. Phys. Rev. A **100**(3), 032326 (2019)
18. N. Quesada, J.M. Arrazola, N. Killoran, Gaussian boson sampling using threshold detectors. Phys. Rev. A **98**(6), 062322 (2018)
19. D. Grier, D.J. Brod, J.M. Arrazola, M.B. de Andrade Alonso, N. Quesada, The complexity of bipartite gaussian boson sampling. Preprint (2021). arXiv:2110.06964
20. A. Deshpande, A. Mehta, T. Vincent, N. Quesada, M. Hinsche, M. Ioannou, L. Madsen, J. Lavoie, H. Qi, J. Eisert, et al., Quantum computational supremacy via high-dimensional gaussian boson sampling. Preprint (2021). arXiv:2102.12474
21. M.A. Broome, A. Fedrizzi, S. Rahimikeshari, J. Dove, S. Aaronson, T.C. Ralph, A. White, Photonic boson sampling in a tunable circuit. Science **339**(6121), 794–798 (2013)
22. J.B. Spring, B.J. Metcalf, P.C. Humphreys, W.S. Kolthammer, X. Jin, M. Barbieri, A. Datta, N. Thomaspeter, N.K. Langford, D. Kundys, et al., Boson sampling on a photonic chip. Science **339**(6121), 798–801 (2013)
23. M. Tillmann, B. Dakic, R. Heilmann, S. Nolte, A. Szameit, P. Walther, Experimental boson sampling. Nature Photonics **7**(7), 540–544 (2013)

24. A. Crespi, R. Osellame, R. Ramponi, D.J. Brod, E.F. Galvao, N. Spagnolo, C. Vitelli, E. Maiorino, P. Mataloni, F. Sciarrino, Integrated multimode interferometers with arbitrary designs for photonic boson sampling. Nature Photonics **7**(7), 545–549 (2013)
25. S. Aaronson, A. Arkhipov, Bosonsampling is far from uniform. Quantum Inf. Comput. **14**(15), 1383–1423 (2014)
26. J. Carolan, J. Meinecke, P. Shadbolt, N. Russell, N. Ismail, K. Worhoff, T. Rudolph, M.G. Thompson, J.L. Obrien, J.C.F. Matthews, et al., On the experimental verification of quantum complexity in linear optics. Nature Photonics **8**(8), 621–626 (2014)
27. N. Spagnolo, C. Vitelli, M. Bentivegna, D.J. Brod, A. Crespi, F. Flamini, S. Giacomini, G. Milani, R. Ramponi, P. Mataloni, et al., Experimental validation of photonic boson sampling. Nature Photonics **8**(8), 615–620 (2014)
28. M. Tillmann, S.-H. Tan, S.E. Stoeckl, B.C. Sanders, H. De Guise, R. Heilmann, S. Nolte, A. Szameit, P. Walther, Generalized multiphoton quantum interference. Phys. Rev. X **5**(4), 041015 (2015)
29. Y.-M. He, Y. He, Y.-J. Wei, D. Wu, M. Atatüre, C. Schneider, S. Höfling, M. Kamp, C.-Y. Lu, J.-W. Pan, On-demand semiconductor single-photon source with near-unity indistinguishability. Nature Nanotechnology **8**(3), 213 (2013)
30. X. Ding, Y. He, Z. Duan, N. Gregersen, M. Chen, S. Unsleber, S. Maier, C. Schneider, M. Kamp, S. Hofling, et al., On-demand single photons with high extraction efficiency and near-unity indistinguishability from a resonantly driven quantum dot in a micropillar. Phys. Rev. Lett. **116**(2), 020401 (2016)
31. H. Wang, Y.-M. He, T.-H. Chung, H. Hu, Y. Yu, S. Chen, X. Ding, M.-C. Chen, J. Qin, X. Yang, et al., Towards optimal single-photon sources from polarized microcavities. Nature Photonics **13**(11), 770–775 (2019)
32. H. Wang, Y. He, Y.-H. Li, Z.-E. Su, B. Li, H.-L. Huang, X. Ding, M.-C. Chen, C. Liu, J. Qin, et al., High-efficiency multiphoton boson sampling. Nature Photonics **11**(6), 361–365 (2017)
33. Y. He, X. Ding, Z.-E. Su, H.-L. Huang, J. Qin, C. Wang, S. Unsleber, C. Chen, H. Wang, Y.-M. He, et al., Time-bin-encoded boson sampling with a single-photon device. Phys. Rev. Lett. **118**(19), 190501 (2017)
34. H. Wang, J. Qin, X. Ding, M. Chen, S. Chen, X. You, Y. He, L. You, Z.M. Wang, C. Schneider, et al., Boson sampling with 20 input photons and a 60-mode interferometer in a 10^{14}-dimensional hilbert space. Phys. Rev. Lett. **123**(25), 250503 (2019)
35. H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, et al., Quantum computational advantage using photons. Science **370**(6523), 1460–1463 (2020)
36. H.-S. Zhong, Y.-H. Deng, J. Qin, H. Wang, M.-C. Chen, L.-C. Peng, Y.-H. Luo, D. Wu, S.-Q. Gong, H. Su, et al., Phase-programmable gaussian boson sampling using stimulated squeezed light. Phys. Rev. Lett. **127**(18), 180502 (2021)
37. J.F.F. Bulmer, B.A. Bell, R.S. Chadwick, A.E. Jones, D. Moise, A. Rigazzi, J. Thorbecke, U.-U. Haus, T. Van Vaerenbergh, R.B. Patel, et al., The boundary for quantum advantage in gaussian boson sampling. Science Advances **8**(4), eabl9236 (2022)
38. S. Aaronson, Certified randomness from quantum supremacy, personal communication (2018)
39. K. Brádler, P.-L. Dallaire-Demers, P. Rebentrost, D. Su, C. Weedbrook, Gaussian boson sampling for perfect matchings of arbitrary graphs. Phys. Rev. A **98**(3), 032310 (2018)
40. J.M. Arrazola, T.R. Bromley, P. Rebentrost, Quantum approximate optimization with gaussian boson sampling. Phys. Rev. A **98**(1), 012322 (2018)
41. J.M. Arrazola, T.R. Bromley, Using gaussian boson sampling to find dense subgraphs. Phys. Rev. Lett. **121**(3), 030503 (2018)
42. M. Schuld, K. Brádler, R. Israel, D. Su, B. Gupt, A quantum hardware-induced graph kernel based on gaussian boson sampling. Preprint (2019). arXiv:1905.12646
43. S. Jahangiri, J.M. Arrazola, N. Quesada, N. Killoran, Point processes with gaussian boson sampling. Phys. Rev. E **101**(2), 022134 (2020)
44. L. Banchi, M. Fingerhuth, T. Babej, C. Ing, J.M. Arrazola, Molecular docking with gaussian boson sampling (2019). arXiv:Quantum Physics

45. J. Huh, G.G. Guerreschi, B. Peropadre, J.R. McClean, A. Aspuru-Guzik, Boson sampling for molecular vibronic spectra. Nature Photonics **9**(9), 615 (2015)
46. J. Huh, M.-H. Yung, Vibronic boson sampling: generalized gaussian boson sampling for molecular vibronic spectra at finite temperature. Scientific Reports **7**(1), 1–10 (2017)
47. L. Banchi, N. Quesada, J.M. Arrazola, Training gaussian boson sampling distributions. Phys. Rev. A **102**(1), 012417 (2020)
48. J. Preskill, Quantum computing in the nisq era and beyond. Quantum **2**, 79 (2018)
49. D. Gottesman, A. Kitaev, J. Preskill, Encoding a qubit in an oscillator. Phys. Rev. A **64**(1), 012310 (2001)

Chapter 7
Quantized Twistors, G_2^*, and the Split Octonions



Roger Penrose

Abstract The basic ideas of twistor theory are outlined, explaining their role in describing classical and quantum massless particles/fields with spin (-helicity), relating this fundamentally to the twistor canonical commutation relations. It is shown that this quantum twistor algebra carries a previously un-noticed split-octonion algebra and G_2^* structure.

7.1 A Key Motivation for the Formulation of Twistor Theory

The basic physical motivation underlying the introduction of twistor theory was to find a dedicated "spinorial" algebraic/geometric structure and formalism that could serve as a basis for combining foundational quantum-mechanical principles with relativistic space-time, this having to be very much tuned to the dimension and $(+, -, -, -)$ signature of the universe we actually inhabit. The key original ideas date back to late 1963, arising from various interrelated motivations. One of the most influential of these was a key aspect of quantum field theory (QFT) that could be understood geometrically, having come to my attention in early 1961. I had hoped that this geometry could be directly related to space-time structure, and so might, in some way, thereby provide a clue to a novel way of understanding how Nature brings together the foundations of quantum mechanics and relativistic space-time.

This basic QFT concept that had intrigued and impressed me in 1961, was an aspect of the framework of QFT which, though not often stressed at the time, was made clear to me by Engelbert Schücking, who had then shared an office with me, in Syracuse University, New York State. This was that field amplitudes have to be split into their positive and negative frequency parts, only the positive frequency parts being retained. This splitting is normally understood in terms of a Fourier decomposition, but I preferred a more geometrical way of looking at this issue. I had been just beginning, at around that time, to become deeply interested in the power of

R. Penrose (✉)
Mathematical Institute, University of Oxford, Oxford, UK

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_7

conformal invariance (conformal geometry being the geometry of angles and small shapes, the metric being relevant only up to proportionality) in the absence of mass, most particularly with regard to the Maxwell free-field equations (whose conformal invariance I had also learned from Engelbert). My familiarity, at that time, with the 2-spinor formalism (to be discussed explicitly here in Sect. 7.2), was greatly helpful to me, and the conformal invariance of all the massless free-field equations, for each different spin value (described here in Sect. 7.4) became clear to me fairly soon afterwards.

However, the Fourier decomposition is not, in itself, conformally invariant, whereas the splitting into positive and negative frequency parts actually turns out to be, as can be seen to be the case if we examine the issue in appropriately geometrical terms. The relevant way to think of this is by means of the *Riemann sphere*, a conformal structure which can be related to the standard complex-number plane (i.e., Wessel plane) $\mathbb{C}$ by stereographic projection. I am here thinking of the Riemann sphere $\mathcal{S}$ as a *unit sphere*, having its centre at the coordinate origin 0 of $\mathbb{C}$ and its equator as the *unit circle* in $\mathbb{C}$, where the point

$$\zeta = X + \mathrm{i}Y \tag{7.1}$$

in $\mathbb{C}$ is projected stereographically from the south pole $(0, 0, -1)$ of $\mathcal{S}$ in standard Cartesian coordinates (X, Y, Z) for the Euclidean 3-space in which the sphere $\mathcal{S}$ is taken to reside, as a unit-radius sphere centred at the coordinate origin $(0, 0, 0)$, so that ζ is represented on $\mathcal{S}$ by the point

$$(X, Y, Z) = \left(1 + \zeta\bar{\zeta}\right)^{-1} \left(\zeta + \bar{\zeta}, \mathrm{i}\bar{\zeta} - \mathrm{i}\zeta, 1 - \zeta\bar{\zeta}\right). \tag{7.2}$$

The map back from $\mathbb{C}$ to $\mathcal{S}$ is *conformal*, but now there is the additional point provided by the south pole $(0, 0, -1)$ of $\mathcal{S}$, representing $\zeta = \infty$, that is adjoined to the Wessel plane $\mathbb{C}$ to provide us with the entire Riemann sphere $\mathcal{S}$, which now is *compact*, having a smooth conformal structure at all of its points.

The real axis $\mathbb{R}$ of $\mathbb{C}$ is represented on $\mathcal{S}$ as the great circle $\mathcal{R}$ given by $Y = 0$ on $\mathcal{S}$, where $\mathcal{R}$ divides $\mathcal{S}$ into two hemispheres. Define $\mathcal{S}^+$ to be the "left-hand" hemisphere of $\mathcal{S}$, for which the imaginary part of ζ is negative, i.e. where Y is negative on $\mathcal{S}$, and $\mathcal{S}^-$ to be the "right-hand" hemisphere of $\mathcal{S}$, for which the imaginary part of ζ is positive, i.e. where Y is positive on $\mathcal{S}$. The basic issue is, roughly speaking, that the positive-frequency Fourier components are those that "die away" as they are extended away from $\mathcal{R}$ into $\mathcal{S}^+$ and the negative frequency ones blow up when extended into $\mathcal{S}^-$. We are interested in positive frequency in QFT, so that it is extensions into $\mathcal{S}^+$ rather than $\mathcal{S}^-$ that we are concerned with.

Both $\mathcal{S}^+$ and $\mathcal{S}^-$ are taken to be "open" as 2-dimensional manifolds; that is to say, they do not include the boundary circle $\mathcal{R}$ itself. Their respective closures $\bar{\mathcal{S}}^+$ and $\bar{\mathcal{S}}^-$ do include the circle $\mathcal{R}$, so that the restrictions of holomorphic functions on these closed sets provide us with analytic functions on $\mathcal{R}$. However, holomorphic

functions on $\mathcal{S}^+ \cup \mathcal{S}^-$ provide us with *hyperfunctions* (such as step functions or delta functions) on $\mathcal{R}$ (i.e. on $\mathbb{R}$). (See [1, Chapter 9, for example].)

This slightly perverse-seeming terminology, with regard to which hemisphere is to be called "$\mathcal{S}^+$", and which is "$\mathcal{S}^-$", arises from the standard conventions concerning the definition of "positive frequency" or, equivalently, "positive energy" according to the usual conventions of standard quantum mechanics and quantum field theory. (See [2, for example].) Somewhat more basically, we can take it that this choice of splitting of a function defined on $\mathcal{R}$ arises from the convention that in standard Heisenberg or Schrödinger conventions, $-\hbar\partial/\partial t$ refers to *energy*, when t is the *time*, which is normally taken to be positive. Thus, "positive frequency" really refers to "positive energy".

What is important for us, in this picture is that the splitting of a function defined on $\mathcal{R}$ into its positive and negative frequency parts does not depend upon this Fourier decomposition, but can be stated merely in terms of splitting into a part extending holomorphically into $\mathcal{S}^-$ (with only the constant part extending both ways). There need be no mention of a decomposition into Fourier components in this description. Moreover, there is a conformal invariance exhibited in this splitting, owing to the fact that there are conformal motions of $\mathcal{S}$ for which $\mathcal{R}$ remains a great circle dividing $\mathcal{S}$ into two hemispheres in this way—inducing fractional linear transformations of its parameter t which do not preserve the Fourier decomposition. So, we see that this positive/negative splitting indeed goes beyond the notion of Fourier decomposition.

However, in our current picture, we are actually taking what should be the "t" in Schrödinger's $-\partial/\partial t$ to be our X-coordinate along the real line $\mathbb{R}$ of $\mathbb{C}$. This does not fit in well with the prospect of extending our 3-dimensional (X, Y, Z)-picture into a 4-dimensional relativistic space-time description. The direct analogue of $\mathcal{S}^+$, in complexified 4-space-time dimensions of relativistic space-time terms is the 8-real-dimesional region referred to as the *forward tube* (which will be of relevance to us here in Sect. 7.5). However, the analogy with the Riemann sphere being divided in two hemispheres does not hold for the forward tube, as there are is an 8-dimensionional region of the complexified space-time that is contained neither in the forward tube nor its analogous backward tube. Instead, in order to relate the above discussion to the 4-complex-dimentional geometry of space-time, we shall be led to a distinctly different geometrical picture, where the Riemann sphere still plays an all-important and natural role, but a very different-looking one from that described above.

7.2 The 2-Spinor Formalism

Let us now consider the Minkowskian space-time $\mathbb{M}$ of special relativity. Using standard Minkowski coordinates $\left(x^0, x^1, x^2, x^3\right)$, where x^0 is the time coordinate

in units where the speed of light $c = 1$, we can write the metric form as

$$\mathrm{d}s^2 = \left(\mathrm{d}x^0\right)^2 - \left(\mathrm{d}x^1\right)^2 - \left(\mathrm{d}x^2\right)^2 - \left(\mathrm{d}x^3\right)^2 = g_{ab}\mathrm{d}x^a\mathrm{d}x^b. \tag{7.3}$$

The repeated indices (one upper and one lower) in the final term denote tensor contraction in the standard "physicist's notation", with an implicit summation over a and b. However, in the expressions in this article, terms appear with such repeated indices (always one upper and one lower, for each letter) are to be regarded as denoting tensor contraction operations, with no underlying coordinate frame being implicit in the expressions. Likewise, individual unrepeated indices in a term simply tell us the nature of the tensor involved (covariant for lower indices and contravariant for upper ones), with no coordinate frame being implicitly or explicitly involved. (See [3, 4].) When restricted to tensor expressions, this "abstract-index" interpretation is here of little import, but it is extremely helpful to adopt this point of view in relation to the 2-*spinor* expressions that will occur frequently in what follows.

In the 2-spinor notation, we use capital italic Latin letters $A, B, C, D, \ldots$ refer to the (complex) 2-dimensional spin-space. Primed ones $A', B', C', D', \ldots$ refer to the complex-conjugate spin-space. The canonical translation, between components of a 4-vector $\boldsymbol{V}$ in a local orthonormal 4-dimensionl Lorentzian description and in the corresponding 2-spinor form, is given by

$$\frac{1}{\sqrt{2}}\begin{pmatrix} V^0 + V^3 & V^1 + \mathrm{i}V^2 \\ V^1 - \mathrm{i}V^2 & V^0 - V^3 \end{pmatrix} = \begin{pmatrix} V^{00'} & V^{01'} \\ V^{10'} & V^{11'} \end{pmatrix}. \tag{7.4}$$

It will indeed be important, here, to adopt the *abstract-index* notation, according to which this expression simply tells us the canonical relation between *components* in a local space-time orthonormal frame and the canonically related *spin-frame* [4]. The actual *vector* $\boldsymbol{V}$ can be written in abstract-index form as V^a or, equivalently, as $V^{AA'}$, and it becomes notationally correct to write

$$V^a = V^{AA'}, \tag{7.5}$$

where the expression (7.4) simply represents the relation between the components of $\boldsymbol{V}$ expressed in an orthonormal tetrad (left-hand side of (7.4)) and in its canonically related spin-frame (right-hand side of (7.4)). In more complicated expressions we need more index letters, and it is notationally correct to identify (abstract) tensor and spinor indices according to

$$a = AA', \quad b = BB', \quad c = CC', \quad d = DD', \quad \ldots \tag{7.6}$$

this applying both to upper and lower abstract indices (see [4] for full details of the use and interpretation of this notation). Corresponding to the symmetric metric tensor (in lower or upper abstract-index form) g_{ab} or g^{ab}, the spin-space and its

complex conjugate come equipped with its *skew*-symmetric epsilon quantities ε_{AB} and ε^{AB}, together with $\varepsilon_{A'B'}$, and $\varepsilon^{A'B'}$, for which

$$g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'}, \quad g^{ab} = \varepsilon^{AB}\varepsilon^{A'B'}. \tag{7.7}$$

Each epsilon quantity has components (according to a canonically normalized spin-frame) given by

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{7.8}$$

but one must bear in mind that under various circumstances (such as with the conformal rescalings that we shall come to in Sect. 7.4 (see (7.43)) one cannot maintain the canonical nature of the spin frames under such rescaling. This is simply a matter of the choice of frames, and the abstract-index expressions do not depend upon any explicit canonical expression such as (7.8) (or (7.4)). Independently of this, spinor indices are raised and lowered according to

$$X^A = \varepsilon^{AB}X_B, \quad X_B = X^A\varepsilon_{AB}, \quad Y^{A'} = \varepsilon^{A'B'}Y_{B'}, \quad Y_{B'} = Y^{A'}\varepsilon_{A'B'}, \tag{7.9}$$

and we find (being careful about the index ordering) that $\varepsilon_A{}^B = -\varepsilon^B{}_A$ and $\varepsilon_{A'}{}^{B'} = -\varepsilon^{B'}{}_{A'}$ are "Kronecker delta" dentity symbols. It should be noted that the "norm" of any single-index spinor ξ^A (or, indeed, of any spinor with an odd number of indices) is necessarily *zero*: $\xi^A\xi_A = 0$, etc.

It is important to notice that a vector V^a is *null* (i.e., $V^aV_a = 0$) if and only if its spinor form splits

$$V^{AA'} = \xi^A\eta^{A'} \tag{7.10}$$

which we can see directly from (7.4), since $\boldsymbol{V}$ being null is the assertion that the matrix determinant is zero, so its rank is 1 (or zero), therefore splitting as asserted. If $\boldsymbol{V}$ is also real, then we can express it as

$$V^{AA'} = \xi^A\bar{\xi}^{A'} \text{ or } -\xi^A\bar{\xi}^{A'} \tag{7.11}$$

according to whether V^a is future- or past-pointing. Accordingly, any future-pointing null vector V^a is equivalent to a 2-spinor ξ^A, up to a phase freedom

$$\xi^A \to \mathrm{e}^{\mathrm{i}\theta}\xi^A \tag{7.12}$$

(where θ is real). This gives a geometrical representation of a (non-zero) single-index 2-spinor, up to a phase factor, as a future-pointing null vector. In fact, up to a

sign, the phase itself has a geometrical space-time representation, via the null (real) bivector

$$\Xi^{ab} = \xi^A \xi^B \varepsilon^{A'B'} + \varepsilon^{AB} \bar{\xi}^{A'} \bar{\xi}^{B'} \tag{7.13}$$

which describes a null 2-plane referred to as the "flag plane" of the spinor ξ^A,where the future-pointing null vector V^a is its "flagpole". (See [4, 5].) If the θ of (7.12) is increased continuously from 0 to 2π, then the flag plane rotates once about the flagpole V^a, but the spinor ξ^A ends up changing sign, to become $-\xi^A$ when $\theta = 2\pi$. A rotation of the flag plane about the flagpole through 4π is needed to restore ξ^A to its original state, which is a characteristic feature of a spinor.

If, at some space-time point $\boldsymbol{x}$, in $\mathbb{M}$, we are concerned with the space of future null directions at $\boldsymbol{x}$, i.e. with the family of directions along the future light cone at $\boldsymbol{x}$ (its "future celestial sphere"), then we can think of this explicitly as the space of non-zero spinors ξ^A up to proportionality, at $\boldsymbol{x}$ and taking a spin-frame at $\boldsymbol{x}$, we realize this space as the space of independent ratios

$$\xi^0 : \xi^1, \tag{7.14}$$

of components of ξ^A, which reveals the space to be a *Riemann sphere*. The non-reflective Lorentz group at $\boldsymbol{p}$ (the non-reflective motions of $\mathbb{M}$ which preserve the point $\boldsymbol{x}$) is thus seen as being the group of fractional linear transformations of the components $\left(\xi^0, \xi^1\right)$ of ξ^A in a spin frame, so we see that this conformal Riemann sphere has a profound local relevance at each point of the Minkowski space-time $\mathbb{M}$. (See [5, 6].)

However, at this stage, we see no connection between this role for the Riemann sphere and that described in Sect. 7.1, which had been to describe, in a conformally invariant way, the splitting of field amplitudes into positive and negative frequency parts that is needed for quantum field theory. Yet, there is indeed a deep connection, and to reveal it we shall need to look at things in a much more global way, which takes us into *twistor theory* and eventually to the new outlook that will be described in this article.

7.3 Projective Twistor Space

We have just seen, in the previous section, how the family of light rays—i.e., of null straight lines, that I shall henceforth refer to simply as *rays*—which pass through any particular point $\boldsymbol{x}$ of Minkowski space $\mathbb{M}$, constitutes a *Riemann sphere*'s worth of rays. Here, I am taking the term "Riemann sphere" to refer to a 2-sphere, with an oriented *conformal structure*. Let us try to turn this around and to think of the rays themselves as being more primitive than the points of $\mathbb{M}$, where we are now to regard "a point $\boldsymbol{x}$ of $\mathbb{M}$" to be interpreted as a certain type of sub-family of this

space of rays which is singled out as having, in the above sense, the structure of a "Riemann sphere". Let us see how this can be achieved.

To see how to represent a ray in a convenient (spinorial) way, let us choose some origin point O in $\mathbb{M}$ and let x^a be the position vector, with respect to O, of a point $\boldsymbol{x}$ on the ray $\boldsymbol{Z}$ that we wish to describe. Since the direction of $\boldsymbol{Z}$ is *null*, we can choose a 2-spinor $\pi_{A'}$ to have its flagpole direction to be that of $\boldsymbol{Z}$. (This choice of notation may seem a little odd at this stage, but it is made to fit in with some of the standard very convenient notation used elsewhere [7], and whose rationale will become clear shortly.) Put

$$\omega^A = \mathrm{i}x^{AA'}\pi_{A'}, \tag{7.15}$$

and define the twistor Z^α as being represented (with respect to the origin O) as the pair

$$Z^\alpha = \left(\omega^A, \pi_{A'}\right). \tag{7.16}$$

Here the upper twistor abstract index α is taken to be the direct sum of an upper unprimed spinor index A and a lower primed spinor index A'. Correspondingly, a lower twistor index α would be taken to be the direct sum of a lower unprimed spinor index A and an upper pried spinor index A', with a corresponding convention for other letters of the alphabet.

Various issues need to be addressed here. First, it should be pointed out that there is an overall complex scaling freedom for the twistor Z^α that does not affect the ray $\boldsymbol{Z}$ that the twistor Z^α is taken to represent. The flagpole direction of $\pi_{A'}$ is to be the direction in which $vb * Z$ points, but this still allows for this complex scaling freedom for $\pi_{A'}$, which affects ω^A in just the same way as $\pi_{A'}$. This scaling would assign a flagpole extent, flag-plane direction, and spinorial sign to $\boldsymbol{Z}$. Replacing $\pi_{A'}$ by another spinor proportional to $\pi_{A'}$, and treating ω^A correspondingly, would result only in a different overall complex scaling for Z^α, and would not affect the ray $\boldsymbol{Z}$ itself. Thus, for simply describing $\boldsymbol{Z}$, we allow for the freedom

$$Z^\alpha \to \lambda Z^\alpha, \text{ where } \mathbb{C} \ni \lambda \neq 0. \tag{7.17}$$

Once we have fixed the $\pi_{A'}$-scaling for the ray $\boldsymbol{Z}$, and having determined the direction in which $\boldsymbol{Z}$ points to be the flagpole direction of $\pi_{A'}$, we see that a change in the choice of the point $\boldsymbol{x}$ on $\boldsymbol{Z}$, will leave Z^α unaffected—i.e. $\omega^A (= \mathrm{i}x^{AA'}\pi_{A'})$ unaffected—whenever the real 4-vector q^a in

$$x^a \to x^a + q^a, \tag{7.18}$$

satisfies

$$q^{AA'}\pi_{A'} = 0. \tag{7.19}$$

This occurs whenever q^a has the form,

$$q^{AA'} = \lambda^A \pi^{A'}, \tag{7.20}$$

for some λ^A. But q^a, being *real*, must actually have the form

$$q^a = u\bar{\pi}^A \pi^{A'} \tag{7.21}$$

where u is a real number, so that (7.18) simply moves us from $\boldsymbol{x}$ to some other point on the ray $\boldsymbol{Z}$. This tells us that a (non-zero) twistor Z^α, defined according to (7.15) and (7.16), and up to the scaling freedom described in (7.17), does indeed uniquely represent a ray in Minkowsi space.

Next, we note that the complex conjugate

$$\bar{Z}_\alpha = \left(\bar{\pi}_A, \bar{\omega}^{A'}\right) \tag{7.22}$$

of the twistor $Z^\alpha = \left(\omega^A, \pi_{A'}\right)$ is a *dual* twistor, so we can form the contracted product of Z^α with $\bar{Z}_\alpha$ to form

$$Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^{A'}, \tag{7.23}$$

and writing this in terms of (7.15), $\omega^A = \mathrm{i} x^{AA'} \pi_{A'}$, where we provisionally allow x^a to be a complex vector, we find $Z^\alpha \bar{Z}_\alpha = (\mathrm{i}x^a - \mathrm{i}\bar{x}^a)\, p_a$, where

$$p_a = \pi_{A'} \bar{\pi}_A, \tag{7.24}$$

so that, for any actually *real* ray in $\mathbb{M}$, this difference vanishes, and we must have

$$Z^\alpha \bar{Z}_\alpha = 0. \tag{7.25}$$

Thus, each ray in $\mathbb{M}$ is represented by a twistor that must necessarily be *null*, in this sense.

The converse, however, is not quite true, as things stand, since the null twistors of the form

$$Z^\alpha = \left(\omega^A, 0\right) \tag{7.26}$$

do not represent actual rays in $\mathbb{M}$. These twistors represent "rays at infinity", instead, and the above twistor (7.26) may be thought of as a ray having been infinitely displaced away from O in the direction of the flagpole of ω^A of (7.26). This infinite displacement can be understood if we extend $\mathbb{M}$ to the *compactified* Minkowski space $\mathbb{M}^\#$, by including its "light cone at infinity" $\mathcal{I}$, which is a 3-cylinder of topology $S^2 \times \mathbb{R}$, this S^2 being the Riemann sphere of possible flagpole directions

for ω^A in (7.26). The 3-cylinder $\mathcal{I}$ is ruled by the rays "at infinity" [8] defined by the twistors of the form (7.26), together with this cylinder's "vertex" point $i(\in \mathbb{M}^\#)$ which represents $\mathbb{M}$'s past and future timelike infinity as well as its spacelike infinity (all at once). The entire space $\mathbb{M}^\#$ is compact and conformally smooth, with topology $S^3 \times S^1$ (see [8, 9]).

Each *point* of $\mathcal{I}$ (apart from i), provides a point at infinity for rays in $\mathbb{M}$, compactifying each ray into a topological circle. Those rays sharing the same point of $\mathcal{I}$ generate a null hyperplane in $\mathbb{M}^\#$. There is now found to be a very close relationship between $\mathbb{M}^\#$, and the space $\mathbb{PN}$ of (non-zero) null twistors, up to proportionality (7.17). $\mathbb{PN}$ is a 5-real-dimensional subspace (given by $Z^\alpha \bar{Z}_\alpha = 0$) of the 3-complex-dimensional (so 6-real-dimensional) projective 3-space $\mathbb{PT}$ of all non-zero projective twistors. The points of $\mathbb{PN}$ represent the rays in $\mathbb{M}^\#$, and the points of $\mathbb{M}^\#$ represent Riemann spheres in $\mathbb{PN}$, those Riemann spheres being the *projective lines* (i.e., $\mathbb{CP}^1$s within $\mathbb{PT}$) which happen to lie entirely within $\mathbb{PN}$.

To understand this correspondence more completely, we need to see how the *non-null* twistors are represented in $\mathbb{M}^\#$, so the question now arises of how, indeed, are we to interpret a non-null twistor Z^α in space-time terms? The easiest way of addressing this issue is to refer to the space of *dual* twistors, so let us first examine the various projective spaces that are of relevance here. The first thing to take note of is the evident fact that the (non-zero) *dual* twistors W_α, up to proportionality, also constitute a complex 3-dimentional projective space (its *dual* $\mathbb{CP}^3$) which we refer to as $\mathbb{PT}^*$. Any point W in $\mathbb{PT}^*$, determined by the dual twistor W_α, represents the family of twistors Z^α for which $Z^\alpha W_\alpha = 0$, that is, by a complex *plane* in $\mathbb{PT}$. This applies, in particular, to the complex plane $\bar{Z}$, given when $\bar{Z}_\alpha = W_\alpha$. Likewise, complex planes in $\mathbb{PT}^*$ represent the *points* of $\mathbb{PT}$. Now, each point of $\mathbb{M}^\#$ can be represented as a *line* (which is a Riemann sphere) in $\mathbb{PT}$ or, equivalently, in $\mathbb{PT}^*$.

The 3-real-dimensional family of points of $\mathbb{PN}$ lying on the intersection $\bar{Z} \cap \mathbb{PN}$ will represent a 3-real dimensional family of rays in $\mathbb{M}^\#$ that can serve to represent the projective twistor Z^α itself. When Z^α is null, then this family of rays is simply the family of rays meeting the ray Z. However, when Z^α is *not* null, then this family of rays becomes a fascinating *twisting* configuration known as a *Robinson congruence* (named after its discoverer Ivor Robinson)—whence the name *twistor*!

The rays of a Robinson congruence twist locally about each other in a right-handed or left-handed sense according to whether $Z^\alpha \bar{Z}_\alpha$ is positive or negative [7, 10]. Thus, the points of $\mathbb{PT}^+$, these defined by $Z^\alpha \bar{Z}_\alpha > 0$, represent the right-handed Robinson congruences in $\mathbb{M}^\#$. Likewise, the points of $\mathbb{PT}^-$, defined by $Z^\alpha \bar{Z}_\alpha < 0$, represent the left-handed Robinson congruences in $\mathbb{M}^\#$. The division of $\mathbb{PT}$ into three disjoint pieces according to

$$\mathbb{PT} = \mathbb{PT}^+ \cup \mathbb{PN} \cup \mathbb{PT}^- \tag{7.27}$$

may be regarded as analogous to the slitting of the Riemann sphere, given in Sect. 7.1

$$\mathcal{S} = \mathcal{S}^+ \cup \mathcal{R} \cup \mathcal{S}^-, \tag{7.28}$$

in fact, deliberately so, as we shall be seeing in Sect. 7.5.

There is a very close relationship between a Robinson congruence and the Clifford-Hopf fibration of a 3-sphere by a 2-sphere's worth of circles:

$$S^1 \hookrightarrow S^3 \to S^2, \tag{7.29}$$

where a Robinson congruence provides a similar fibration of $\mathbb{M}^{\#}$ by a 3-sphere's worth of (compactified) rays:

$$S^1 \hookrightarrow \mathbb{M}^{\#} \to S^3. \tag{7.30}$$

The S^3 here is the intersection $\bar{Z} \cap \mathbb{PN}$, each point of which represents a (compactified) ray, no two of which intersect and which, together, fill up the whole of $\mathbb{M}^{\#}$. As mentioned above, Robinson congruences can be right-handed ($Z^\alpha \bar{Z}_\alpha > 0$) or left-handed ($Z^\alpha \bar{Z}_\alpha < 0$).

The relation between (7.29) and (7.30) is more than just formal. If we choose a particular time-slice $x^0 = \text{const.}$ of $\mathbb{M}$ (see (7.3)) then each ray will meet this spacelike hyperplane in a point at which its direction projects down a particular spatial direction. These directions turn out to be tangents to a 2-parameter family of circles. These circles form a Clifford-Hopf fibration of this time-slice, when it is compactified to a 3-sphere by a point at infinity (see [7, especially Fig. 6-3 on p.62]).

7.4 Twistor Kinematics

This geometrical picture of a twistor also has a clear (though not often recognized) direct *physical* significance, in relation to the angular momentum structure of a spinning massless particle (such as a photon, when considered classically), as we shall be seeing at the end of this section. Up to this point, however, we have been thinking entirely in geometrical terms, whereas it is natural to think more *kinematically*, by regarding a null twistor to represent not simply the *history*—i.e., world-line—of a freely moving massless particle, but now regarding the massless particle actually to have a mass-energy and a momentum, i.e., a 4-*momentum* which would be given by (7.24):

$$p_a = \pi_{A'} \bar{\pi}_A. \tag{7.31}$$

This provides us also with a 6-angular momentum about the origin point O, given by

$$M_{ab} = x_a p_b - x_b p_a, \tag{7.32}$$

where, for the time being, we take the particle to have no intrinsic spin. Writing this in terms of spinor notation, incorporating (7.31), and re-ordering the terms and indices slightly (for convenience), we obtain

$$M_{ab} = \pi_{B'} x_{A'A} \bar{\pi}_B - \pi_{A'} x_{B'B} \bar{\pi}_A, \tag{7.33}$$

which can be re-written as

$$M_{ab} = 2\left\{\pi_{(B'} x_{A')[A} \bar{\pi}_{B]} - \pi_{[A'} x_{B'](B} \bar{\pi}_{A)}\right\}, \tag{7.34}$$

where round brackets around indices denote taking the symmetric part and square brackets, the anti-symmetric part (e.g. $X_{(AB)} = \frac{1}{2}(X_{AB} + X_{BA})$ and $X_{[AB]} = \frac{1}{2}(X_{AB} - X_{BA})$, half the terms cancelling, when the expression is expanded out, and the other half doubling up. Making use of $\omega^A = \mathrm{i} x^{AA'} \pi_{A'}$ (eqn. (7.15)), a brief calculation then gives us (see also [7, p.68])

$$M_{ab} = \mathrm{i}\omega_{(A} \bar{\pi}_{B)} \varepsilon_{A'B'} - \mathrm{i}\varepsilon_{AB} \bar{\omega}_{(A'} \pi_{B')}. \tag{7.35}$$

This quantity depends on the choice of origin O, and it is often useful to think of such a quantity as being a *field* on space-time (say on $\mathbb{M}$ or $\mathbb{M}^\#$), whose value at a space-time point X would be whatever we get when the origin point O is translated to X. Let us use the notation

$$Q_{\cdots}^{\{X\}\cdots} \tag{7.36}$$

for a quantity $Q_{\cdots}^{\cdots}$ originally defined with respect to the origin O, (so that $Q_{\cdots}^{\{O\}\cdots} = Q_{\cdots}^{\cdots}$) and which, like M_{ab}, has a well-defined behaviour under shift of origin. For the case of M_{ab} itself, we would have

$$M_{ab}^{\{X\}} = M_{ab} - X_a p_b + X_b p_a \tag{7.37}$$

so that, for $X = \boldsymbol{x}$, (7.37) indeed vanishes for the $M_{ab}^{\{O\}}$ defined in (7.35), as it should, since the particle indeed has no moment about the point $X = \boldsymbol{x}$ lying on the ray $\boldsymbol{Z}$. For the 4-momentum itself, of course we have

$$p_a^{\{X\}} = p_a. \tag{7.38}$$

In the case of the spinor parts of the twist $Z^a = \left(\omega^A, \pi_{A'}\right)$, we must have

$$\omega^{\{X\}A} = \omega^A - \mathrm{i} X^{AA'} \pi_{A'} \text{ and } \pi_{A'}^{\{X\}} = \pi_{A'}, \tag{7.39}$$

which is consistent with (7.15).

This way of representing a *twistor* as a field has some specific advantages. Denoting the (covariant) derivative with respect to X^a by ∇_a, we find that the origin-dependence implicit in (7.15) is satisfied by

$$\nabla_{BB'}\omega^{\{X\}A} = -\mathrm{i}\varepsilon_B{}^A\pi_{B'} \text{ and } \nabla_{BB'}\pi_{A'}^{\{X\}} = 0. \tag{7.40}$$

The equation governing $\omega^{\{X\}A}$ can be expressed, independently of $\pi_{A'}^{\{X\}}$, as

$$\nabla_{A'}^{(A}\omega^{\{X\}B)} = 0 \tag{7.41}$$

(where the symmetrizing brackets ignore the "{X}"), as follows because (7.41) implies that $\nabla_{A'}^{A}\omega^{\{X\}B}$ is skew in AB, so these two indices split off as an ε^{AB}, (7.41) thereby defining $\pi_{B'}$, whose constancy follows from taking a second derivative of (7.41) and commuting the derivative operators.

An important advantage of thinking of $\omega^{\{X\}A}$ as a field in this way, is that it allows us to think of a twistor as a conformally invariant entity, relevant to the compact conformal manifold $\mathbb{M}^{\#}$. The "twistor equation" (7.41) is, in fact, invariant under a conformal rescaling of the metric $g_{ab}\mathrm{d}x^a\mathrm{d}x^b$, given in (7.3), where we rescale according to

$$g_{ab} \to \tilde{g}_{ab} = \Omega^2 g_{ab} \text{ and } g^{ab} \to \tilde{g}^{ab} = \Omega^{-2}g^{ab} \tag{7.42}$$

which must be accompanied by

$$\begin{aligned} &\varepsilon_{AB} \to \tilde{\varepsilon}_{AB} = \Omega\varepsilon_{AB}, \qquad &&\varepsilon^{AB} \to \tilde{\varepsilon}^{AB} = \Omega^{-1}\varepsilon^{AB}, \\ &\varepsilon_{A'B'} \to \tilde{\varepsilon}_{A'B'} = \Omega\varepsilon_{A'B'}, \quad &&\varepsilon^{A'B'} \to \tilde{\varepsilon}^{A'B'} = \Omega^{-1}\varepsilon^{A'B'}. \end{aligned} \tag{7.43}$$

We need to bear in mind that these are abstract-index expressions, and that we cannot think of the epsilon quantities in (7.43) simply as skew matrices with components (see (7.8))

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{7.44}$$

Such explicit representations arise only when a *spin frame* is locally introduced, and the presence of a conformal factor would come in, in relation to the normalization of the frame. The details of this are fully explained elsewhere [4, 7] and it would be a distraction to enter into this here. Moreover, the conformal invariance of (7.41), coming about when $\omega^{\{X\}A}$ is taken to have conformal weight zero

$$\omega^{\{X\}A} \to \tilde{\omega}^{\{X\}A} = \omega^{\{X\}A}, \tag{7.45}$$

allows us to extend the definition of $\omega^{\{X\}A}$ to points on $\mathcal{I}$.

There is a technical matter of relevance here, however, concerning what is referred to as the "Grgin index" (see [7, Section 9.4]). This will not be entered into in any detail here, but such matters become relevant when we pass to a local description of twistors in terms of *fields* on $\mathbb{M}^{\#}$. As the field $\omega^{\{X\}A}$ is continued locally across $\mathcal{I}$, a "jump" by a factor of i must be involved, multiplying $\omega^{\{X\}A}$ by i, as X passes from one side of $\mathcal{I}$ (within $\mathbb{M}^{\#}$) to the other (see [7, Section 9.4]). Descriptions of the form (7.26) are adequate for our purposes here, but it is worth bearing in mind that such "Grgin jumps" are implicitly present behind the scenes.

Nevertheless, this kind of "field description" is very helpful in other respects. In particular, it helps us to understand the connection between the role of the Robinson congruence (see (7.27)), associated with a non-null twistor Z^α, and the angular momentum field $M_{ab}^{\{X\}}$ of (7.37). Let us first try to understand how the *spin* of a massless particle enters the twistor picture. Up to this point, we have been considering the angular momentum tensor M_{ab} only for a single massless particle *without* intrinsic spin. In more general situations, we can consider the *Pauli-Lubanski* 4-vector

$$S_a = \frac{1}{2} e_{abcd} p^b M^{cd}, \tag{7.46}$$

where the alternating tensor e_{abcd} is totally skew in all its indices

$$e_{abcd} = e_{[abcd]} \tag{7.47}$$

and, in the canonical orthonormal frame introduced in (7.3)

$$e_{0123} = 1. \tag{7.48}$$

It has the 2-spinor representation

$$e_{abcd} = \mathrm{i}\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'D'}\varepsilon_{B'C'} - \mathrm{i}\varepsilon_{AD}\varepsilon_{BC}\varepsilon_{A'C'}\varepsilon_{B'D'} \tag{7.49}$$

(see [4]). It follows that S_a is invariant under change of origin-point O to X,

$$S_a^{\{X\}} = S_a \tag{7.50}$$

under (7.37). This quantity need not vanish even for massless particles if we are considering a composite system consisting several independent particles, but for a single massless particle, it is a normal requirement that the Pauli-Lubanski S_a is proportional to p_a:

$$S_a = s p_a \tag{7.51}$$

where s is referred to as the *helicity* of the particle, and in standard quantum mechanics, $2s$ has to be an *integer*, which is *even* for a boson and *odd* for a fermion.

We say that the helicity is *right-handed* if $s > 0$ and *left-handed* if $s < 0$. For a photon, $s = \pm 1$, and this corresponds to circular polarization in the right- or left-handed sense.

For a single *classical* massless particle whose kinematics is described by a twistor Z^α, we find:

$$s = \frac{1}{2} Z^\alpha \bar{Z}_\alpha. \tag{7.52}$$

We can also examine the flagpole *field* for the spinor part $\omega^{\{X\}A}$ for Z^α, and we find that if $s \neq 0$, then these flagpole directions of $\omega^{\{X\}A}$ point along the *Robinson congruence* defined by Z^α. These flagpole directions, together with the 4-momentum direction given by p_a, define the angular momentum field (7.37) up to proportionality, thereby assigning a clear physical role for the notion of a Robinson congruence.

7.5 Quantized Twistor Theory and Masless Fields

In standard quantum theory, the quantum state of a particle—or its *wave-function*—can be represented as a function $\psi(x^a)$ of *position* x^a, or alternatively as a function $\tilde{\psi}(p_a)$ of *momentum* p_a. If we pass from the first such description to the second, we find that we need to adopt a translation procedure according to which "multiplying by x^a" translates to "differentiating with respect to p_a", so formally (and more specifically) we have:

$$x^a = -\mathrm{i}\hbar \frac{\partial}{\partial p_a} \tag{7.53}$$

and, in the reverse direction of passing back to $\psi(x^a)$, "multiplying by p_a" becomes (minus) "differentiating with respect to x^a":

$$p_a = \mathrm{i}\hbar \frac{\partial}{\partial x^a}, \tag{7.54}$$

either of which can be expressed (in terms of formal operations) as the commutation relation

$$p_a x^b - x^b p_a = \mathrm{i}\hbar \delta_a^b \tag{7.55}$$

together with

$$x^a x^b = x^b x^a \quad \text{and} \quad p_a p_b = p_b p_a, \tag{7.56}$$

where x^a and p_a are regarded as linear operators.

In twistor theory, we find something that is formally very similar, with regard to the description of a massless particle. Analogously to the canonical conjugacy between the quantum operators x^b and p_a, as expressed in (7.55) and (7.56), we find that the quantum twistor operator Z^α and its complex conjugate $\bar{Z}_\alpha$ are canonically conjugate operators in the sense

$$Z^\alpha \bar{Z}_\beta - \bar{Z}_\beta Z^\alpha = \hbar \delta^\alpha_\beta \tag{7.57}$$

together with

$$Z^\alpha Z^\beta = Z^\beta Z^\alpha \text{ and } \bar{Z}_\alpha \bar{Z}_\beta = \bar{Z}_\beta \bar{Z}_\alpha, \tag{7.58}$$

these being consistent with the standard quantum commutation relations between angular momentum and momentum, when translated into twistor form according to (7.31) and (7.35). In accordance with these twistor commutation relations we can also, analogously to (7.53) and (7.54) write the operator relation

$$\bar{Z}_\alpha = -\hbar \frac{\partial}{\partial Z^\alpha} \tag{7.59}$$

or, in the reverse direction,

$$Z^\alpha = \hbar \frac{\partial}{\partial \bar{Z}_\alpha}. \tag{7.60}$$

In order to appreciate the role of these rules, we shall need to see how twistor wave-functions are to be understood. Here we shall be considering what is to be regarded as a twistor wave-function of a single massless particle, of any given spin-helicity value. We find various striking similarities with the standard formalism, but also some intriguing differences that need explaining. What we do find is that there is a twistor form of a wave-function, that applies in a strikingly economical way, to massless free fields for all the various possible values of the spin. We restrict attention to Minkowski space $\mathbb{M}$, (though with appropriate attention to matters such as the the Grgin index, referred to in the previous section, we can extend the discussion to $\mathbb{M}^\#$). There is no attempt, at this stage, to develop this into a full theory of massive particles or interacting fields, or to general relativity.

An important issue here is the value of the *helicity s* when our wave-function is to be expressed in twistor terms. Repeating the calculations of Sect. 7.4, where the helicity was obtained as a proportionality factor between the Pauli-Lubanski vector S_a of (7.46) and the momentum p_a, but now incorporating the above commutation relations (7.57) and (7.58) when translating this into twistor form (from (7.31) and (7.35)), we now find that the (non-quantum) formula (7.52) is refined to

$$s = \frac{1}{4}\left(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha\right). \tag{7.61}$$

We find that when written in terms of Z^α (using (7.57)) that s is essentially a (displaced) *homogeneity operator*

$$\hbar Z^\alpha \frac{\partial}{\partial Z^\alpha} = -2s - 2. \tag{7.62}$$

It is usual to drop the "$\hbar$", in normal terminology (as will also be done in the discussion below), when referring to spin or helicity (effectively, though only temporarily, adopting the convention $\hbar = 1$) , so we may take it that the expectation of twistor theory that a twistor wave-function (in the Z^α description) for a massless field of helicity s, is described by a holomorphic function f of Z^α that is homogeneous of degree $(-2s - 2)$.

Why *holomorphic* (i.e., complex analytic)? Well, according to the rules of quantum mechanics, we are looking for a function that is independent of the canonically conjugate variable(s) $\bar{Z}_\alpha$. What does "independent" mean in the current context? It means that our twistor wave-function f should satisfy

$$\frac{\partial f}{\partial \bar{Z}_\alpha} = 0. \tag{7.63}$$

However, this equation simply asserts that the appropriate *Cauchy-Riemann equations* hold, which indeed assert the required *holomorphicity* of $f(Z^\alpha)$, as required.

Yet, the question nevertheless arises: what does this holomorphicity and homogeneity have to do with the normal space-time description of such fields? This question we come to next.

For spin zero ($s = 0$), in $\mathbb{M}$, we would be concerned with solutions of the wave equation

$$\nabla_a \nabla^a \psi = 0, \tag{7.64}$$

and for $|s| = 1$ it would be the source-free Maxwell equations

$$\nabla_{[a} F_{bc]} = 0, \quad \nabla^a F_{ab} = 0 \tag{7.65}$$

(with $F_{bc} = F_{[bc]}$); which, when we write F_{ab} in terms of its spinor parts (see [4])

$$F_{ab} = \psi_{AB} \varepsilon_{A'B'} + \varepsilon_{AB} \bar{\psi}_{A'B'} \tag{7.66}$$

with symmetry

$$\psi_{AB} = \psi_{(AB)}, \tag{7.67}$$

(7.65) can be expressed as

$$\nabla^{AA'} \psi_{AB} = 0. \tag{7.68}$$

For a photon wave-function, however, we need to replace the *real* quantity F_{ab} (as expressed in (7.66)) by a *complex* F_{ab} version

$$F_{ab} = \psi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\tilde{\psi}_{A'B'}, \tag{7.69}$$

where both parts are symmetric

$$\psi_{AB} = \psi_{(AB)} \text{ and } \tilde{\psi}_{A'B'} = \tilde{\psi}_{(A'B')}, \tag{7.70}$$

ψ_{AB} and $\tilde{\psi}_{A'B'}$ being both of *positive-frequency* (see Sect. 7.1) and independent of each other, separately satisfying

$$\nabla^{AA'}\psi_{AB} = 0 \text{ and } \nabla^{AA'}\tilde{\psi}_{A'B'} = 0. \tag{7.71}$$

Here ψ_{AB} describes the left-handed part of the photon's state, i.e., $s = -1$; whereas $\tilde{\psi}_{A'B'}$ describes the right-handed part, i.e., $s = +1$.

This generalizes to higher spin, as in weak-field gravity where $|s| = 2$ (see [4]) (and to lower spin for massless neutrinos, where $|s| = \frac{1}{2}$, the general analogues of (7.71) being

$$\nabla^{AA'}\psi_{AB\ldots E} = 0 \text{ and } \nabla^{AA'}\tilde{\psi}_{A'B'\ldots E'} = 0. \tag{7.72}$$

where each $\psi_{AB\ldots E}$ and $\tilde{\psi}_{A'B'\ldots E'}$ is symmetric:

$$\psi_{AB\ldots E} = \psi_{(AB\ldots E)} \text{ and } \tilde{\psi}_{A'B'\ldots E'} = \tilde{\psi}_{(A'B'\ldots E')}, \tag{7.73}$$

both having positive frequency.

Twistor theory provides a remarkable way of satisfying all these requirements, by consideration of a "twistor function" $f(Z^\alpha)$),which is holomorphic in Z^α and homogeneous of the required degree $-2s - 2$:

$$\frac{\partial f}{\partial \bar{Z}_\alpha} = 0 \text{ and } Z^\alpha \frac{\partial f}{\partial Z^\alpha} = (-2s - 2)f. \tag{7.74}$$

What we find is that all the required conditions can all be satisfied by the expression

$$\tilde{\psi}_{A'B'\ldots E'} = k_s \oint_{\omega = \mathrm{i}x\pi} \pi_{A'}\pi_{B'}\ldots\pi_{E'} f(Z^\alpha)\pi_{F'}\mathrm{d}\pi^{F'} \tag{7.75}$$

for $s \geq 0$, where A', B', $\ldots$, E' are $2s$ in number (see [7]), and (see [7, 11])

$$\psi_{AB\ldots E} = k_s \oint_{\omega = \mathrm{i}x\pi} \frac{\partial}{\partial\omega^A}\frac{\partial}{\partial\omega^B}\ldots\frac{\partial}{\partial\omega^E} f(Z^\alpha)\pi_{F'}\mathrm{d}\pi^{F'} \tag{7.76}$$

for $s \leq 0$, $A, B, \ldots, E$ being $-2s$ in number, where the k_s are appropriate numerical constants. The required satisfaction of the field equations (7.72) (or (7.64)) are easily confirmed by differentiation inside the integral, using

$$\nabla_a = \frac{\partial}{\partial x^a} = \mathrm{i}\pi_{A'} \frac{\partial}{\partial \omega^A} \tag{7.77}$$

from $\omega^A = \mathrm{i}x^{AA'}\pi_{A'}$ (see (7.15)).

In order to address the requirement of positive frequency, we need to arrange the *singularities* of f to be appropriately situated, so that the field quantity in question (here $\psi_{AB\ldots E}$ or $\tilde{\psi}_{A'B'\ldots E'}$) extends holomorphically into the appropriate region of complexified space-time (as initially motivated here in Sect. 7.1). In relativistic terms, this required region is what is referred to as the *forward tube* (referred to in Sect. 7.1), this being the region $\mathbb{CM}^+$ of complexified Minkowski space-time $\mathbb{CM}$ whose complex position vectors ξ^a (relative to the real origin O) have imaginary parts that are timelike and *past*-pointing. We have already seen, in Sect. 7.3, that the real points of $\mathbb{M}$ (including those at infinity $\mathcal{I}$) are represented as complex projective lines, i.e., $\mathbb{CP}^1$s, these being Riemann spheres, that lie entirely in $\mathbb{PN}$. More generally, any $\mathbb{CP}^1$ lying in the projective twistor space $\mathbb{PT}$ will represent a point in the *complexified* compactified Minkowski space $\mathbb{CM}^\#$. The points in the *forward tube* $\mathbb{CM}^+$ are represented, in projective twistor space, by the projective lines that lie entirely in the top-half region $\mathbb{PT}^+$ (see (7.27) and its accompanying text), as is readily ascertained by examining (7.15), (7.16), and (7.23) for complex x^a. (See also [7, pp. 151-2]).

To see how the singularities of f can be arranged to ensure that this positive-frequency requirement can be assured, let us first consider a very simple example. Take $s = 0$, so f has homogeneity -2:

$$f = \frac{1}{(Z^\alpha A_\alpha)(Z^\beta B_\beta)} \tag{7.78}$$

where the two planes $\boldsymbol{A}$ (given by $Z^\alpha A_\alpha = 0$) and $\boldsymbol{B}$ (given by $Z^\alpha B_\alpha = 0$) intersect in a line $\boldsymbol{L}$ lying entirely in the bottom half $\mathbb{PT}^-$ of $\mathbb{PT}$ (see (7.27)). Any line $\boldsymbol{X}$ (a Riemann sphere) lying entirely in the top half $\mathbb{PT}^+$ (thereby representing a point in the forward tube) will necessarily meet the planes $\boldsymbol{A}$ and $\boldsymbol{B}$ in two separate points, since such a line cannot meet $\boldsymbol{L}$. These are the two places where f becomes singular (simple poles) on the Riemann sphere $\boldsymbol{X}$, and we get a well-defined answer for the integral (7.76), for a contour that simply loops once around one of these poles, separating it from the other. A similar procedure applies to functions of the form

$$f = \frac{(Z^\lambda C_\lambda)^l (Z^\mu C_\mu)^m}{(Z^\alpha A_\alpha)^p (Z^\beta B_\beta)^q} \tag{7.79}$$

which allow us to produce many other positive-frequency twistor wave-functions for any helicity value.

In fact, twistor functions of the form (7.79) provide a basis for the Hilbert space of wave-functions of any particular helicity, so we could effectively stop our discussion here. But this is not too helpful for giving us an understanding of what is involved, since this description is very much tied to the particular locations of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, and $\boldsymbol{D}$. If we add such twistor functions with different choices of the planes $\boldsymbol{A}$ and $\boldsymbol{B}$, then the choices of contours need not be consistent in the different terms, so we really need a more invariant way of looking at what these contour-integrals really stand for. The answer lies in the notion of (holomorphic Čech) sheaf cohomology [7, section 6.2], which provides us with a much broader perspective on the nature of twistor functions. The details of the mathematics are not of great consequence for our particular considerations here, but in the language involved, we can refer to "holomorphic 1st cohomology elements" in $\mathbb{PT}^+$ rather than simply "holomorphic functions" on $\mathbb{PT}^+$. Quantities such as those in (7.78) and (7.79) provide merely *representatives* of this cohomology.

We may regard such a representative twistor function f to be defined (holomorphically and of appropriate homogeneity) on the *intersection* $\mathcal{A} \cap \mathcal{B}$ of two open subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathbb{PT}^+$, and whose *union* is this entire top half

$$\mathbb{PT}^+ = \mathfrak{A} \cup \mathfrak{B} \tag{7.80}$$

of projective twistor space. In the above examples (7.78) and (7.79), these open sets would be

$$\mathfrak{A} = \mathbb{PT}^+ - \boldsymbol{A} \text{ and } \mathfrak{B} = \mathbb{PT}^+ - \boldsymbol{B}. \tag{7.81}$$

However, when descriptions of this general kind are added together, with regions like $\mathfrak{A}$ and $\mathfrak{B}$ that differ from term to term, we are led to having to divide $\mathbb{PT}^+$ up into many more open sets, arising from the various intersections of the open sets involved. This leads us into more complicated descriptions of the cohomology elements, according to the rules of (Čech) 1st cohomology (see [7]), but to go into the details of this would take us much too far afield from what is required here. All that we need here, is that adequate representations for the relevant positive-frequency twistor wave functions can be obtained—such as by using twistor functions defined on the intersection of a pair of open sets that together cover $\mathbb{PT}^+$.

7.6 Split Octonions and G_2^*

Where we have been led, in our attempts to relate twistor theory to the quantum mechanics of relativistic massless particles is perhaps somewhat remarkable and, in many respects, distinctly unexpected. It is necessary to point out, however, that we appear to have been guided into a curious and inappropriate association that is not reflected in Nature. We see this particularly in the strange dual role for $\mathbb{PT}^+$, where *classically* it represents massless particles with positive helicity (i.e. right-handed

ones), whereas the *quantum* requirement of positive frequency provides a quite unconnected preference for $\mathbb{PT}^+$ (as opposed to $\mathbb{PT}^-$) in expressing the quantum need for wave-functions to have positive frequency. If, instead, we had used lower-indexed *dual* twistors (e.g. W_α), rather than upper-indexed twistors in our initial descriptions, then this association would have been reversed (i.e. negative helicity with positive frequency). This suggests that, in our attempts to find an appropriate twistorial description of Nature, we should consider what will be referred to here as *bi-twistors*.

A bi-twistor is an 8-component entity which we may regard as being composed of an upper-indexed twistor and a lower-indexed dual twistor, but it is important that we retain the feature of *quantized* twistor theory introduced in Sect. 7.5, notably the commutation relations (7.57) and (7.58) which we here rewrite as

$$\boldsymbol{Z}^\alpha \boldsymbol{W}_\beta - \boldsymbol{W}_\beta \boldsymbol{Z}^\alpha = \hbar \delta^\alpha_\beta \tag{7.82}$$

together with

$$\boldsymbol{Z}^\alpha \boldsymbol{Z}^\beta = \boldsymbol{Z}^\beta \boldsymbol{Z}^\alpha, \quad \boldsymbol{W}_\alpha \boldsymbol{W}_\beta = \boldsymbol{W}_\beta \boldsymbol{W}_\alpha, \tag{7.83}$$

which, thoug satisfied by $W_\alpha = -\hbar\partial/\partial Z^\alpha$ (see (7.57), (7.59)) removes any apparent preference for the upper-indexed Z^α, as opposed to the lower-indexed W_α.

We regard the operators $\boldsymbol{Z}^\alpha$ and W_α (now written with bold letters), subject to (7.82) and (7.83) as providing the structure of the (linear) bi-twistor system $\mathcal{O}$, whose elements can be written as

$$A_\alpha \boldsymbol{Z}^\alpha + F^\alpha \boldsymbol{W}_\alpha \tag{7.84}$$

where the "coefficients" F^α and A_α are simply ordinary "classical" *commuting* twistors and dual twistors. We refer to $A_\alpha \boldsymbol{Z}^\alpha + F^\alpha \boldsymbol{W}_\alpha$ as a *real* bi-twistor, belonging to the system $\mathcal{O}$, if

$$F^\alpha = \bar{A}^\alpha \tag{7.85}$$

and as a *complex* bi-twistor, belonging to the larger system $\mathbb{C}\mathcal{O}$, if the restriction (7.85) is not imposed.

Neither the system $\mathcal{O}$ nor $\mathbb{C}\mathcal{O}$ is an "algebra" in the normal sense of possessing a notion of "product" defined between pairs of such elements, though there is an operation of *scalar multiplication* by a number ν, where for

$$\mathcal{A} = A_\alpha \boldsymbol{Z}^\alpha + A^\alpha \boldsymbol{W}_\alpha, \tag{7.86}$$

we define

$$\nu\mathcal{A} = (\nu A_\alpha)\boldsymbol{Z}^\alpha + (\nu A^\alpha)\boldsymbol{W}_\alpha, \tag{7.87}$$

which acts (distributively) on elements of $\mathcal{O}$ if ν is real and on elements of $\mathbb{C}\mathcal{O}$ if ν is complex. What is important for, the system $\mathcal{O}$ (and also for $\mathbb{C}\mathcal{O}$), however, is that it *does* have a notion of "triple product" defined for any *three* of its elements.

To express this triple product most succinctly, it will be convenient to economize on the notation, and use the same letter for both twistor parts of a bi-twistor, such as writing $A_\alpha \boldsymbol{Z}^\alpha + A^\alpha \boldsymbol{W}_\alpha$ instead of $A_\alpha \boldsymbol{Z}^\alpha + F^\alpha \boldsymbol{W}_\alpha$, where we regard the twistor A^α to be *independent* of the dual twistor A_α despite the use of the same base letter "A" in each case, the two being distinguished by the upper/lower location of the index—and we retain the freedom to use "$A^\alpha = \bar{A}^\alpha$" to indicate the *reality* condition for a bi-twistor, i.e. membership of $\mathcal{O}$, rather than the broader $\mathbb{C}\mathcal{O}$. In accordance with this, we can use the notation

$$\begin{aligned} \mathcal{A} &= A_\alpha \boldsymbol{Z}^\alpha + A^\alpha \boldsymbol{W}_\alpha, \\ \mathcal{B} &= B_\alpha \boldsymbol{Z}^\alpha + B^\alpha \boldsymbol{W}_\alpha, \\ C &= C_\alpha \boldsymbol{Z}^\alpha + C^\alpha \boldsymbol{W}_\alpha, \end{aligned} \tag{7.88}$$

for three general elements of $\mathbb{C}\mathcal{O}$ and we write the triple product of bi-twistors $\mathcal{A}, \mathcal{B}, C$ to be

$$[\mathcal{A}, \mathcal{B}, C] = \mathrm{i} \sum_{\pm \circlearrowleft} \mathcal{A}\mathcal{B}C \tag{7.89}$$

where "$\pm \circlearrowleft$" is to assert that the sum is over 6 terms, one for each permutation of $\mathcal{A}$, $\mathcal{B}$, and C, but with a minus sign if it is an odd permutation. The product in "$\mathcal{A}\mathcal{B}C$" is, indeed, the operator product of the quantities written in as in (7.82) and (7.83), but what is important is that the application of these commutation laws affects every term, allowing all terms to result finally in an expression that is only linear in $\boldsymbol{Z}^\alpha$ and $\boldsymbol{W}_\alpha$, so the result is actually another bi-twistor, and not something of higher (or lower) order. Collecting terms in (7.89), and employing the commutator (7.82) we find that $[\mathcal{A}, \mathcal{B}, C]$ is the sum of

$$\mathrm{i}\hbar(A_\alpha \boldsymbol{Z}^\alpha + A^\alpha \boldsymbol{W}_\alpha)(B_\beta C^\beta - C_\beta B^\beta) \tag{7.90}$$

and the two other expressions obtained by cyclic permutation of A, B, and C. Using the bracket notation

$$[AB; C\boldsymbol{W}] = A^\alpha C_\alpha B^\beta \boldsymbol{W}_\beta - B^\alpha C_\alpha A^\beta \boldsymbol{W}_\beta \tag{7.91}$$

and

$$[AB; C\boldsymbol{Z}] = A_\alpha C^\alpha B_\beta \boldsymbol{Z}^\beta - B_\alpha C^\alpha A_\beta \boldsymbol{Z}^\beta, \tag{7.92}$$

we find that we can re-express (7.89) as

$$[\mathcal{A}, \mathcal{B}, C] = \mathrm{i}\hbar([AB; C\boldsymbol{W}] + [BC; A\boldsymbol{W}] + [CA; B\boldsymbol{W}] - [AB; C\boldsymbol{Z}] \\ -[BC; A\boldsymbol{Z}] - [CA; B\boldsymbol{Z}]). \tag{7.93}$$

For real bi-twistors we have

$$A^\alpha = \bar{A}^\alpha, \quad B^\alpha = \bar{B}^\alpha, \quad C^\alpha = \bar{C}^\alpha \tag{7.94}$$

in which case $[\mathcal{A}, \mathcal{B}, C]$ is also real (because of the "i" in (7.90)).

In order to define the octonionic product, we need two particular ingredients: a notion of *scalar (inner) product* and a specific "*unit*" bi-twistor $\mathcal{E}$ being of "unit length" (actually "squared length" 2) with regard to this scalar product. For the scalar product to be well defined within the bi-twistor notions, we require an operation $\mathbf{I}$ ("multiplying by i") acting on bi-twistors (and preserving bi-twistor reality), namely

$$\mathbf{I}: \ A_\alpha \boldsymbol{Z}^\alpha + A^\alpha \boldsymbol{W}_\alpha \mapsto \mathrm{i}A_\alpha \boldsymbol{Z}^\alpha - \mathrm{i}A^\alpha \boldsymbol{W}_\alpha. \tag{7.95}$$

The operation of *scalar product* between bi-twistors can then be defined by

$$\mathcal{A} \cdot \mathcal{B} = \mathrm{i}(\mathcal{A}(\mathbf{I}\mathcal{B}) - \mathrm{i}(\mathbf{I}\mathcal{B})\mathcal{A}) \tag{7.96}$$

which, when written out, becomes the intended

$$\mathcal{A} \cdot \mathcal{B} = \hbar(A_\alpha B^\alpha + A^\alpha B_\alpha). \tag{7.97}$$

This quadratic form (when expressed in terms of real numbers) is seen to have signature $(4, 4)$, which is what is required for the *split*-octonion algebra (see [12]).

We note, from the presence of $\hbar$ in both the triple product (see (7.90)) and also, as expressed here, the scalar product (7.97), that these are features of *quantized* twistor theory, so that the split-octonionic structure that we come to next *disappears* in the classical $\hbar \to 0$ limit. Having noted this fact, we henceforth adopt the convention that units are chosen so that

$$\hbar = 1. \tag{7.98}$$

We wish to see how a split-octonionic product (see [12]) can be defined in terms of the operations defined above. For this, we need our choice of $\mathcal{E}$ to be a real bi-twistor, which is to represent the *unit element* for what then becomes the *split-octonionic algebra* O:

$$\mathcal{E} = E_\alpha \boldsymbol{Z}^\alpha + E^\alpha \boldsymbol{W}_\alpha, \tag{7.99}$$

where $\mathcal{E}$ is chosen to be a real unit bi-twistor in the sense:

$$\mathcal{E} \cdot \mathcal{E} = 2 \tag{7.100}$$

(see (7.97)). A general element $\mathcal{A}$ of $\mathcal{O}$ can be written as the sum of a *scalar part* $\mathcal{A}^{\mathcal{S}}$, which is a real multiple of $\mathcal{E}$, namely

$$\mathcal{A}^{\mathcal{S}} = \frac{1}{2}(\mathcal{A} \cdot \mathcal{E})\mathcal{E} \tag{7.101}$$

and a vector part $\mathcal{A}^{\mathcal{V}} (= \mathcal{A} - \mathcal{A}^{\mathcal{S}})$, so

$$\mathcal{A} = \mathcal{A}^{\mathcal{S}} + \mathcal{A}^{\mathcal{V}} \tag{7.102}$$

with

$$\mathcal{A}^{\mathcal{V}} \cdot \mathcal{E} = 0. \tag{7.103}$$

For the (split)octonionic product, we wish to mirror the situation that occurs with ordinary (or complexified) quaternions:

$$(a + \varphi p)\,(b + \varphi q) = (ab - \varphi p \cdot \varphi q) + (a\varphi q + b\varphi p + \varphi p \times \varphi q)\,, \tag{7.104}$$

where the symbols with arrows denote vector parts, those without arrows denoting scalar parts. For the split-octonionic product, we can imitate this as soon as we have a notion of *vector product* (here denoted by "$\times$") between vector part of bi-twistors. This is achieved by use of the triple product (7.89), but specifically incorporating $\mathbf{I}$, which ensures orthogonality with $\mathcal{E}$, etc.

$$\mathcal{A} \times \mathcal{B} = [\mathbf{I}\mathcal{A}, \mathbf{I}\mathcal{B}, \mathbf{I}\mathcal{E}], \tag{7.105}$$

where the orthogonality with $\mathcal{E}$, as well as with $\mathcal{A}$ and $\mathcal{B}$, can be directly checked from (7.93), using (7.95), (7.91), and (7.92).

Accordingly, to define the (split-)octonionic product for $\mathcal{O}$ (or for $\mathbb{C}\mathcal{O}$), we can precisely mirror what happens with the product of two quaternions (or complexified quaternions), so that we can see how our desired split-octonionic product (between elements of $\mathcal{O}$ or of $\mathbb{C}\mathcal{O}$) arises. Let us consider any two elements $\mathcal{A}$ and $\mathcal{B}$, which we split into their scalar and vector parts (according to (7.102)):

$$\mathcal{A} = \mathcal{A}^{\mathcal{S}} + \mathcal{A}^{\mathcal{V}} \text{ and } \mathcal{B} = \mathcal{B}^{\mathcal{S}} + \mathcal{B}^{\mathcal{V}}, \tag{7.106}$$

then we find

$$\begin{aligned}\mathcal{AB} &= \left(\mathcal{A}^{\mathcal{S}} + \mathcal{A}^{\mathcal{V}}\right)\left(\mathcal{B}^{\mathcal{S}} + \mathcal{B}^{\mathcal{V}}\right)\\ &= \left(\mathcal{A}^{\mathcal{S}}\mathcal{B}^{\mathcal{S}} - \mathcal{A}^{\mathcal{V}}\cdot\mathcal{B}^{\mathcal{V}}\right)\mathcal{E} + \left(\mathcal{A}^{\mathcal{S}}\mathcal{B}^{\mathcal{V}} + \mathcal{B}^{\mathcal{S}}\mathcal{A}^{\mathcal{V}} + \mathcal{A}^{\mathcal{V}}\times\mathcal{B}^{\mathcal{V}}\right),\end{aligned} \tag{7.107}$$

exactly mirroring the quaternionic (7.104).

It is of interest to note that if we stay within a sub-algebra of $\mathcal{O}$ (or $\mathbb{C}\mathcal{O}$) that is generated (over $\mathcal{O}$ or $\mathbb{C}\mathcal{O}$) by only two independent elements, then we simply get a copy of the (complexified) quaternions, which is associative, the non-associativity of the octonionic algebra arising only with *three* (or more) independent generators (whence the satisfaction of the "alternative law" $\mathcal{A}(\mathcal{AB}) = (\mathcal{AA})\mathcal{B}$ (see [12]) is maintained, this involving only *two* independent elements.

The G_2^* symmetry of this system $\mathcal{O}$ arises more directly as a symmetry of the triple product (7.89). It is not the purpose of the current article to explore the properties of this remarkable exceptional simple group and its detailed relation to quantized twistor theory here, but merely to point out this striking and previously unrecognized fact. There would appear to be much that is worthy of further exploration, to be followed up in further work. One intriguing route to follow is the so-far incomplete relation of twistor theory to conformally curved space-times of general relativity (see [13]) where the left/right unbiassed bi-twistors (and their related "ambitwistors" [14, 15]) and "palatial twistors" [16] might in some sense be connected.

References

1. R. Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe* (Jonathan Cape, London, 2004). Vintage IBN:9780-679-77631-4
2. J. Bjorken, S. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, 1964). ISBN:0-07-005493-2
3. R. Penrose, Structure of space-time, in *Battelle Rencontres, 1967 Lectures in Mathematics and Physics,* eds. by C.M. DeWitt, J.A. Wheeler (Benjamin, New York, 1968)
4. R. Penrose, W. Rindler, *Spinors and Space-Time, Vol. 1: Two-Spinor Calculus and Relativistic Fields* (Cambridge University Press, Cambridge, 1984)
5. W.T. Payne, Elementary spinor theory. Am. J. Phys. **20**, 253–262 (1952)
6. R. Penrose, The apparent shape of a relativistically moving sphere. Proc. Camb. Phil. Soc. **55**, 137–139 (1959)
7. R. Penrose, W. Rindler, *Spinors and Space-Time, Vol. 2: Spinor and Twistor Methods in Space-Time Geometry* (Cambridge University Press, Cambridge, 1986)
8. R. Penrose, The light cone at infinity, in *Conférence Internationale sur les Téories Relativistes de la Gravitation*, ed. by L. Infeld (Gauthier Villars, Paris and PWN, Warsaw, 1964)
9. R. Penrose, Conformal approach to infinity, in *Relativity, Groups and Topology: The 1963 Les Houches Lectures*, eds. by B.S. DeWitt, C.M. DeWitt (Gordon and Breach, New York, 1964)
10. R. Penrose, Twistor algebra. J. Math. Phys. **8**(2), 345–366 (1967)
11. L.P. Hughston, *Twistors and Particles, Lecture Notes in Physics* No. 97 (Springer, Berlin, 1979)
12. J.C. Baez, The Octonions. Bull. Am. Math. Soc. **39**, 145–205 (2002)

13. R. Penrose, Non-linear gravitons and curved twistor theory. Gen. Rel. Grav. **7**, 31–52 (1976)
14. C.R. LeBrun, Ambi-twistors and Einstein's equations. Class. Quant. Grav. **2**, 555–563 (1985)
15. C.R. LeBrun, On four-dimensional Einstein manifolds, in *The Geometric Universe; Science, Geometry, and the Work of Roger Penrose*, eds. by S.A. Huggett, L.J. Mason, K.P. Tod, S.T. Tsou, N.M.J. Woodhouse (Oxford University Press, Oxford, 1998)
16. R. Penrose, Twistor theory: A geometric perspective for describing the physical world, in *New Spaces for Physics; Formal and Conceptual Reflections,* eds. by M. Anel, G. Catren (Cambridge University Press, Cambridge, 2021)

Chapter 8
Kronecker Anomalies and Gravitational Striction



Alexander M. Polyakov and Fedor K. Popov

Dedicated to C. N. Yang

Abstract We study quantum field theories in which the number of degrees of freedom changes discontinuously across the momentum space. This discontinuity which we call "Kronecker anomaly" leads to non-local effective actions and can be represented as a theory with the random, self-tuning coupling constants.

8.1 Introduction and General Discussion

One of the most surprising features of our world is that it is described by local quantum field theories (QFT). This means that we associate with the object we want to describe (be it the standard model or a classical magnet) a set of local operators $\{O_n(x)\}$ satisfying operator product expansion [1–3]

$$O_n \cdot O_m = \sum c^l_{nm} O_l. \tag{8.1}$$

This also means, that small perturbation can be described by the conformal perturbation theory, $V = \sum_a \lambda_a O_a$, where $\{\lambda^a\}$ are the coupling constants. In the general QFT, the operator algebra (8.1) replaces the equations of motion. But is the locality a must? We don't know the answer.

In this paper we show that there is a plausible mechanism discovered by Larkin and Pikin [4, 5], which breaks the locality. It was already mentioned long ago [6], but many related facts remain to be clarified.

A. M. Polyakov (✉)
Department of Physics, Princeton University, Princeton, NJ, USA
e-mail: polyakov@princeton.edu

F. K. Popov
CCPP, Department of Physics, New York, NY, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_8

The effect which we discuss and generalize in this paper is the following. Consider a ϕ^4 theory on a crystal lattice near a critical point. The effective action is given by

$$\mathcal{L}_0 = \frac{1}{2}\left(\partial_\mu \phi\right)^2 + \tau\phi^2 + g\phi^4, \tag{8.2}$$

where $\tau \propto (T - T_{\text{cr}})$ and ϕ is magnetization. As we account for the lattice vibrations we get another field, the displacement $u_\alpha(x)$ and strain $u_{\alpha\beta} = \partial_\alpha u_\beta + \partial_\beta u_\alpha$. The Lagrangian for this field is given by

$$\mathcal{L}_1 = \frac{\lambda}{2} u_{\alpha\alpha}^2 + \mu u_{\alpha\beta}^2, \tag{8.3}$$

where λ, μ are Lame's coefficients. Finally there is an interaction leading to magneto-striction

$$\mathcal{L}_2 \propto u_{\alpha\alpha}\phi^2. \tag{8.4}$$

The field $u_{\alpha\beta}$ is Gaussian and if we naively integrate out $u_{\alpha\beta}$, the result seems trivial. Indeed, if we use (8.3) we get

$$\langle u_{\alpha\alpha}(x) u_{\beta\beta}(y) | u_{\alpha\alpha}(x) u_{\beta\beta}(y)\rangle \propto \delta^d(x - y)$$

and the dressing of ϕ with u fluctuations results with

$$\Delta\mathcal{L}_{\text{eff}} \propto \int \langle u_{\alpha\alpha}(x) u_{\beta\beta}(y) | u_{\alpha\alpha}(x) u_{\beta\beta}(y)\rangle \, \phi^2(x)\phi^2(y) d^d x d^d y \propto \int \phi^4(x) d^d x$$

So, we obtain a trivial renormalization of the ϕ^4 term.

There is a flaw in this calculation since $u_{\alpha\beta}(x)$ are not independent fields but are subject to the Saint-Venant compatibility conditions [7]

$$R_{\alpha\beta} = \partial^2 u_{\alpha\beta} + \partial_\alpha\partial_\beta u_{\gamma\gamma} - \partial_\alpha\partial_\gamma u_{\gamma\beta} - \partial_\beta\partial_\gamma u_{\gamma\alpha} = 0 \tag{8.5}$$

which ensures that for $k \neq 0$ we can express $u_{\alpha\beta}$ in terms of some field u_α

$$u_{\alpha\beta} = k_\alpha u_\beta + k_\beta u_\alpha$$

The key point of the present paper is that the statistics of $u_{\alpha\beta}(k)$ changes discontinuously at $k = 0$, because at this point Saint-Venant's condition disappears. This jump of the numbers of the degrees of freedom leads to a non-local contribution of the elastic modes proportional to the Kronecker symbol, $\delta_{\mathbf{p},0}$, which we call the Kronecker anomaly. To be more general and precise, consider a field theory with

periodic boundary conditions or, which is the same, a theory on a D-dimensional torus. The Green functions in such a theory have the following structure

$$G\left(\mathbf{p}_1,\ldots,\mathbf{p}_N\right)=\widetilde{G}\left(\mathbf{p}_1,\ldots,\mathbf{p}_N\right)+A(0,\mathbf{p}_2,\ldots,\mathbf{p}_N)\delta_{\mathbf{p}_1,0}+\ldots, \tag{8.6}$$

where $\widetilde{G}\left(\mathbf{p}_1,\ldots,\mathbf{p}_N\right)$ is assumed to be an analytic function of its arguments[1]. Then s non-zero contribution $A(0,\mathbf{p}_2,\ldots,\mathbf{p}_N)$ is a Kronecker anomaly. We can formulate it differently—the Green functions on a torus could not be analytic functions due to the Kronecker anomaly.

To proceed, let us find the propagator for the acoustic field $u_{\alpha\alpha}$ which enters formula (8.4). For $\mathbf{k}\neq 0$ we use (8.3) and the representation $u_{\alpha\beta}=k_\alpha u_\beta+k_\beta u_\alpha$. This part of the action has the form

$$\begin{aligned} S&=\sum_{\mathbf{k}\neq 0}\left(\mu k^2\delta_{\alpha\beta}+(\lambda+\mu)\,k_\alpha k_\beta\right)u_\alpha(k)u_\beta(-k)=\\ &=\sum_{\mathbf{k}\neq 0}\left(\mu k^2\left(\delta_{\alpha\beta}-\frac{k_\alpha k_\beta}{k^2}\right)+(\lambda+2\mu)\,k^2\frac{k_\alpha k_\beta}{k^2}\right). \end{aligned} \tag{8.7}$$

These two pieces are orthogonal and the propagator, being the inverse of the above matrices is given by

$$\langle u_\alpha(k)u_\beta(-k)|u_\alpha(k)u_\beta(-k)\rangle_{k\neq 0}=\frac{1}{\mu k^2}\left[\left(\delta_{\alpha\beta}-\frac{k_\alpha k_\beta}{k^2}\right)+\frac{1}{\lambda+2\mu}\frac{k_\alpha k_\beta}{k^4}\right]. \tag{8.8}$$

So, the $k\neq 0$ part of $u_{\alpha\alpha}$ is simply

$$\langle u_{\alpha\alpha}(k)u_{\beta\beta}(-k)|u_{\alpha\alpha}(k)u_{\beta\beta}(-k)\rangle=\frac{1}{\lambda+2\mu}\quad\text{at } k\neq 0, \tag{8.9}$$

For $\mathbf{k}=0$ the propagator is given by [2]

$$\langle u_{\alpha\beta}u_{\gamma\delta}|u_{\alpha\beta}u_{\gamma\delta}\rangle\propto\int\mathcal{D}ue^{-\beta E(u)}u_{\alpha\beta}u_{\gamma\delta}, \tag{8.10}$$

[1] We also assume that these functions do not have essential singularities.

[2] We note that these modes $\mathbf{k}=0$ could not be present in the spectrum depending on the structure of the crystal [8].

where $E(u)$ is again the elastic energy (8.3). Using Ward's identity

$$\left\langle \frac{\delta E(u)}{\delta u_{\alpha\beta}} u_{\gamma\delta} \right\rangle = \delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma},$$

$$\frac{\delta E(u)}{\delta u_{\alpha\beta}} = \left(\lambda + \frac{2\mu}{d}\right) u_{\gamma\gamma}\delta_{\alpha\beta} + 2\mu\left(u_{\alpha\beta} - \frac{1}{d}\delta_{\alpha\beta}u_{\gamma\gamma}\right),$$

$$\langle u_{\alpha\alpha}u_{\beta\beta}|u_{\alpha\alpha}u_{\beta\beta}\rangle = \frac{1}{\lambda + \frac{2\mu}{d}} \tag{8.11}$$

Now we have the propagator

$$\langle u_{\alpha\alpha}(k)u_{\beta\beta}(-k)|u_{\alpha\alpha}(k)u_{\beta\beta}(-k)\rangle = C_1 + C_2\delta_{\mathbf{k},0} \tag{8.12}$$

to which in the coordinate space has the form

$$\langle u_{\alpha\alpha}(x)u_{\beta\beta}(y)|u_{\alpha\alpha}(x)u_{\beta\beta}(y)\rangle = C_1\delta(x-y) + \frac{C_2}{V}, \tag{8.13}$$

where V is the volume and

$$C_1 = \frac{1}{\lambda + 2\mu}, \quad C_2 = \frac{2\left(1 - \frac{1}{d}\right)\mu}{(\lambda + 2\mu)\left(\lambda + \frac{2}{d}\mu\right)} \tag{8.14}$$

the term C_2 represents the Kronecker's anomaly (8.6). Since the u-field is Gaussian, it is easy to calculate its effect on the partition function of the order parameter

$$\int \mathcal{D}u e^{-S[u_{\alpha\beta}]} \exp\left[\int d^d x u_{\alpha\alpha}(x)\phi^2(x)\right] =$$

$$\exp\left[\int dxdy\, \langle u_{\alpha\alpha}(x)u_{\beta\beta}(y)|u_{\alpha\alpha}(x)u_{\beta\beta}(y)\rangle\, \phi^2(x)\phi^2(y)\right]. \tag{8.15}$$

Thus, the partition function with the account of elasticity acquires an extra term coming from (8.15)

$$Z = \int \mathcal{D}\phi \exp\left\{\int d^d x\left[(\nabla\phi)^2 + \tau\phi^2 + \overline{g}\phi^4\right] + \frac{C_2}{V}\int \phi^2(x)\phi^2(y)d^d x d^d y\right\} =$$

$$= \int \mathcal{D}\phi\mathcal{D}\xi \exp\left\{\xi^2 V + C_2^{-1}\int d^d x(\tau + \xi)\phi^2\right\}. \tag{8.16}$$

So, we see that the coupling of the order parameter to acoustic phonons leads to the bilocal term in the action

$$\Delta S \sim \frac{1}{V}\int \epsilon(x_1)\epsilon(x_2)d^dx_1d^dx_2, \tag{8.17}$$

the factor $\frac{1}{V}$ is needed to make this additional term relevant in the thermodynamic limit. This term can also be cast in the form

$$\Delta S \sim \xi^2 V + \int d^dx\xi\epsilon(x), \tag{8.18}$$

where $\epsilon(x) \sim \phi^2(x)$ is the energy density of the magnet. The ξ variable plays the role of the coupling constant and we see that the coupling constants behave as random Gaussian variables.

Quite unexpectedly, the random coupling constants and bilocal operators have been discussed long ago in a total different context. In the works [9, 10] it was argued that the wormholes and baby universes generate bilocal perturbations. In our set up, one can obtain random coupling much cheaper, we need just an acoustic phonon instead of the other universe.

But this it not the end of the story. Depending on the values of the anomalous dimensions, the integral (17) may diverge. Using the OPE

$$\epsilon(x_1)\epsilon(x_2) = \frac{1}{|x_1 - x_2|^{2\Delta_\epsilon}}I + \ldots, \tag{8.19}$$

we find that if $\Delta_\epsilon < \frac{d}{2}$ we have the bilocal action and coupling are random, while for $\Delta_\epsilon > \frac{d}{2}$ we have UV-divergence and the bilocal action disappears.

The above discussion shows that the essence of bilocality lies in Kronecker's terms. An important example of this mechanism is the cosmological constant problem. In this case the role of the acoustic phonons is played by gravitons. We have a metric tensor, $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$, with the gauge condition

$$k_\mu h_{\mu\nu} = \frac{1}{2}k_\nu h_{\lambda\lambda} \tag{8.20}$$

and it is obvious that we have extra degrees of freedom at $k = 0$, where the gauge condition degenerates. However, a more subtle analysis of gauge invariance is needed for the cosmological constant problem. We leave this problem for the future work and present here only a simpler case of the 2d string theory in the dilaton background [11–13]. In this case we have two sets of oscillators satisfying the relation

$$\left[a_n^\mu, a_m^\nu\right] = n\eta^{\mu\nu}\delta_{n,-m},$$

and all states could be built by applying these operators at vacuum state $\rangle 0$. Naively, there are no photon-like degrees of freedom in two dimensions. However, due to the Kronecker anomaly, infinite number of the so called discrete states appear in the theory. As discussed in [11–14], the first excited state of the open string has the usual form $e_\mu(p)a^\dagger_{1\mu}\rangle 0$ and satisfy the gauge condition

$$p(p+2b)=0, \quad \left(p_\mu+2b_\mu\right)e_\mu(p)=0, \quad e_\mu \sim e_\mu+\lambda p_\mu, \tag{8.21}$$

where b_μ is proportional to the derivative of the dilaton $b_\mu \sim \partial_\mu\varphi$, φ is the dilaton. This quantity is fixed by the condition that the central charge of the string is zero. In $1+1$ dimension we can try to eliminate the "photon" $e_\mu(p)$. We choose a gauge and set $e_1=0$ and then using the constraint to get $e_0=0$. This is a textbook trick to show that in d-dimensions photon has $(d-2)$ degrees of freedom. In the present case we also can do this except of the points $p_\mu=0$ and $p_\mu=-2b_\mu$. Analogous reasoning works for the higher string excitations. As a result our string has infinite number of the so called discrete states. This was explicitly verified in [12, 13] by calculating scattering amplitudes, which contain infinite number of poles. Their very non-trivial dynamics is important for establishing string-matrix model correspondence.

We will not discuss here the dynamics of discrete states, but just present the second level string excitation, which was discussed in [12, 13, 15]. It has the form

$$\rangle\psi_2=\left(e_\mu a^\dagger_{2\mu}+\xi_{\mu\nu}a^\dagger_{1\mu}a^\dagger_{1\nu}\right)\rangle p. \tag{8.22}$$

The coefficients $e_\mu, \xi_{\mu\nu}$ satisfy the constraint

$$\left(p_\mu+2b_\mu\right)\xi_{\mu\nu}+e_\nu=0, \quad \left(p_\mu+3b_\mu\right)e_\mu+\frac{1}{2}\xi_{\lambda\lambda}=0, \tag{8.23}$$

and two gauge transformations

$$\begin{aligned} &e_\mu \sim e_\mu+\eta_\mu, \quad \xi_{\mu\nu}\sim\xi_{\mu\nu}+\frac{1}{2}\left(p_\mu\eta_\nu+p_\nu\eta_\mu\right), \quad \left(p_\mu+2b_\mu\right)\eta_\mu=0, \\ &e_\mu \sim e_\mu+\lambda\left(5p_\mu-2b_\mu\right), \quad \xi_{\mu\nu}\sim\xi_{\mu\nu}+\lambda\left(\delta_{\mu\nu}+3p_\mu p_\nu\right) \end{aligned} \tag{8.24}$$

The analysis in [12, 13] shows that the physical states exist only at $p_1=\pm\frac{1}{2}$, as a result of the degeneration of the gauge condition. Deeper dynamical understanding of this system seems highly desirable.

8.2 Kronecker Anomaly in Thermal Harmonic Oscillator

The other way to generate non-local terms in the effective action is to consider a theory at thermal state (or on a torus). For instance, we can study a simple harmonic

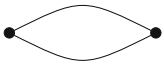
oscillator at temperature β with the action

$$S = \int_0^\beta d\tau \left[\frac{1}{2}\dot{x}^2(\tau) + \frac{1}{2}m^2 x^2(\tau) \right], \tag{8.25}$$

where we assume periodic boundary conditions $x(\tau+\beta) = x(\tau)$ and mass, m, plays the role of the frequency of the oscillator. From this we can derive the propagator in the Euclidian signature to be

$$G(\tau) = \frac{1+n}{2m} e^{-m|\tau|} + \frac{n}{2m} e^{m|\tau|},\ n = \frac{1}{e^{\beta m}-1},\ G(\omega_m) = \frac{1}{\omega_m^2 + m^2},\ \omega_m = \frac{2\pi m}{\beta}, \tag{8.26}$$

where n was fixed by periodic boundary conditions. The action (8.25) is very simple and we can compute all physical observables, for instance, the susceptibility $\alpha = -\frac{\partial \langle x^2 | x^2 \rangle}{\partial m^2}$. It could be done by using Kubo formula $\alpha_\omega \propto \int_0^\beta d\tau e^{i\omega\tau} \langle x^2(0)x^2(\tau)|x^2(0)x^2(\tau)\rangle$. First we do this computation in the lower-half plane of the frequency range $i\Omega \in \frac{2\pi}{\beta}\mathbb{N}$

$\Pi(\Omega) =$  $= T \sum_n \frac{1}{\omega_n^2 + m^2} \frac{1}{(\Omega - \omega_n)^2 + m^2} = \frac{1}{m} \frac{\coth \frac{\beta m}{2}}{4m^2 + \Omega^2},$ (8.27)

if one naively assumes that $\Pi(\Omega)$ is an analytic function of the frequency Ω we would get a wrong answer for the $\Pi(\Omega = 0)$. To see this we make the same computation in the coordinate representation

$$\Pi(\tau) = G^2(\tau) = \frac{(1+n)^2}{4m^2} e^{-2m|\tau|} + \frac{n(1+n)}{2m^2} + \frac{n^2}{4m^2} e^{2m|\tau|},$$

$$\text{then} \quad \Pi(\Omega) = \frac{1}{m} \frac{\coth \frac{\beta m}{2}}{4m^2 + \Omega^2} + \frac{1}{4m^2} \frac{1}{\cosh m\beta - 1} \delta_{\Omega,0}, \tag{8.28}$$

where the non-analytic term appears because of the interference between two exponents in the thermal propagator $G(\tau)$. It means that our calculation of the polarization operator in the lower-half plane of the frequency range does not capture the zero mode contribution in (8.27) as compared to (8.28). It actually plays an important role, because $\Pi(0)$ is a susceptibility and could be measured at the experiment. To sum up, we have two possible values for the susceptibility α, that can be found by analytical continuation from the lower-half plane Π_A and by direct

computation of Π_M at $\Omega = 0$:

$$\Pi_A(0) = \frac{\coth\frac{\beta m}{2}}{4m^3}, \quad \text{and} \quad \Pi_M(0) = \frac{1}{m}\frac{\coth\frac{\beta m}{2}}{4m^2} + \frac{1}{4m^2}\frac{1}{\cosh m\beta - 1}. \tag{8.29}$$

To understand this discrepancy we compute the susceptibility for harmonic oscillator using the straightforward definition of α and the action (8.25),

$$\langle n|x^2|n|n|x^2|n\rangle = \frac{2n+1}{2m},$$
$$\langle\langle x^2|x^2\rangle|\,\langle x^2|x^2\rangle\rangle = \frac{\mathrm{tr}\left[x^2\, e^{-\beta H}\right]}{\mathrm{tr}\left[e^{-\beta H}\right]} = \frac{1}{2\omega}\frac{\sum_{n=0}^{\infty}(2n+1)e^{-n\beta m}}{\sum_{n=0}^{\infty} e^{-\beta m}} = \frac{1}{2m}\coth\frac{\beta m}{2}.$$

We can change the mass ,m, in two different regimes: keeping temperature or entropy constant. They lead to two different types of the susceptibilities. First, we compute the isothermal susceptibility of the system,

$$\alpha_\beta = -\left(\frac{\partial\,\langle\langle x^2|x^2\rangle|\,\langle x^2|x^2\rangle\rangle}{\partial m^2}\right)_\beta = \frac{1}{m}\frac{\coth\frac{\beta m}{2}}{4m^2} + \frac{\beta}{2m^2}\frac{1}{\cosh m\beta - 1} = \Pi_M(0), \tag{8.30}$$

while, adiabatic or isentropic susceptibility is

$$\alpha_S = -\left(\frac{\partial\,\langle\langle x^2|x^2\rangle|\,\langle x^2|x^2\rangle\rangle}{\partial m^2}\right)_S = \frac{1}{m}\frac{\coth\frac{\beta m}{2}}{4m^2} = \Pi_A(0), \tag{8.31}$$

Then from this computation we can understand the discrepancy in the quantities (8.29). The computation of $\Pi_M(0)$ corresponds to the isothermal susceptibility because in the Matsubara technique the temperature is assumed to be fixed by the boundary conditions $x(\tau+\beta) = x(\tau)$. But we can show in general that $\Pi_A(\Omega)$ gives the answer for real-time computation [16] with the use of the Keldysh-Schwinger technique

$$\Pi_{KS}(0) = \int_{-\infty}^{t} dt'\,\mathrm{Im}\left[G_\beta^2(t-t')\right]dt' = \frac{1}{m}\frac{\coth\frac{\beta m}{2}}{4m^2} = \Pi_A(0). \tag{8.32}$$

It happens because for the real time with unitary evolution the entropy $S = -\mathrm{tr}\left[\rho\log\rho\right]$ remains constant. Therefore, the discrepancy (8.29) measures, roughly speaking, the difference between these two different susceptibilities, at constant temperature or entropy. According to our definition this discrepancy is the Kronecker anomaly (8.6). The existence of the Kronecker term changes the behavior

of the system. If we couple the oscillator to an external system and integrate its degrees of freedom, the resulted effective action would contain non-local terms for the external system as in the case of acoustic phonons. The same discrepancy could arise in quantum field theories.

For example, we can consider the following action

$$S = \int_0^\beta d\tau dx \left[\frac{1}{2}(\partial_\mu a)^2 + \frac{1}{2}(\partial_\mu \phi)^2 - \frac{1}{2}m^2\phi^2 - ga\phi^2 \right]. \tag{8.33}$$

Then we can study one-loop correction to the propagator of the field a. It has the following form

$$\begin{gathered} G^{-1}(\Omega_n, p) = \Omega_n^2 + p^2 + m^2 - \Sigma(\Omega_n, p), \\ \Sigma(\Omega, p) = g^2 T \sum_n \int \frac{dq}{2\pi} \frac{1}{\omega_n^2 + m^2 + \mathbf{q}^2} \frac{1}{(\Omega - \omega_n)^2 + m^2 + (p-q)^2}. \end{gathered} \tag{8.34}$$

We expect that $\Sigma(\Omega, p)$ has a non-analytical contribution at $p = 0$ (then the particles will have the same energies and can contribute to the Kronecker anomaly (8.48)). And if we evaluate this function at $\Omega_n = 2\pi T n$ and continue to the $\Omega = 0$ and compare it with the direct computation we would get two different answers:

$$\delta\Sigma = \Sigma(\Omega = 0, \mathbf{p} = 0) - \Sigma_A(0, \mathbf{p} = 0) = \int \frac{dq}{2\pi} \frac{1}{8T(m^2 + q^2)} \frac{1}{\sinh^2\left[\frac{\sqrt{m^2+q^2}}{2T}\right]} \tag{8.35}$$

This difference would induce an additional non-local term in the effective action.

There are a lot of other interesting mechanisms that could lead to the emergence of bilocal actions. We mostly discuss them in the supplementary material. In Appendix "Kronecker Anomaly in Electromagnetic Theory", we discuss how the Kronecker anomaly could arise in the $U(1)$ gauge theory to the jump in the number of degrees of freedom for $k = 0$ and $k \neq 0$. In Appendix "Mathematical Aspects of Kronecker Anomaly", we show the physical meaning of this Kronecker anomaly that arise in a space with compact direction and give precise conditions for the emergence of it. In Appendix "Two Dimensional Scalar Electrodynamics on a Torus" we show how Kronecker anomaly could arise in quantum electrodynamics or gravity on a torus, coupled to a scalar or fermionic field. In Appendix "Kronecker Anomaly in Spaces with Constant Curvature", we that in odd dimensional de Sitter spaces quantum field theories also have Kronecker anomalies, while in even dimensional de Sitters such non-analytical contributions are absent. In Appendix "De Sitter

Lacuna", we show how discontinuous propagator in the coordinate space arise in the hyperbolic spaces.

Appendix

Kronecker Anomaly in Electromagnetic Theory

The Kronecker anomaly appears also in the electromagnetic theory

$$S = \frac{1}{4}\int d^d x F_{\mu\nu}^2, \quad F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{8.36}$$

for $k \neq 0$ we can use Landau gauge $k^\mu A_\mu = 0$ to get the propagator

$$\langle A_\mu(k) A_\nu(-k) | A_\mu(k) A_\nu(-k)\rangle = \frac{g_{\mu\nu} - \frac{k_\mu k_\nu}{k^2}}{k^2}. \tag{8.37}$$

The electromagnetic field is subjected to the following Maxwell equations

$$\delta F = j, \quad dF = 0, \quad \delta = *d*\tag{8.38}$$

The first equation is dynamical, that shows how $F_{\mu\nu}$ depends on the external current, while the second equation is just a statement, that $F_{\mu\nu}$ is a closed form and for some A_μ, $F = dA$, that is analogous to the Eq. (8.5). Namely, if we consider the constraint with some wave vector k^μ this equation states that $F_{\mu\nu}(k)$ has $d-1$ degrees of freedom (d comes from the consideration that $F = dA$ for some 1-form A and (-1) comes from gauge invariance of such a configuration $A \sim A + d\phi$ would give the same field strength F).

At $k^\mu = 0$, when constant configurations are considered we have $\frac{d(d-1)}{2}$ degrees of freedom (in 4 dimensions it corresponds to 6 constant electric and magnetic fields) and we would have the same type of the change of the number of degrees of freedom. Indeed, let us consider the following correlation

$$\begin{aligned}
&\frac{1}{2}\langle F_{\mu\nu}(k) F_{\mu\nu}(-k) | F_{\mu\nu}(k) F_{\mu\nu}(-k)\rangle = \\
&\frac{1}{k^4}\left(g_{\mu\nu}k^2 - k_\mu k_\nu\right)\left(g^{\mu\nu}k^2 - k^\mu k^\nu\right) = \\
&\qquad d-1,
\end{aligned} \tag{8.39}$$

At $k^\mu = 0$ the propagator is singular and naively it would give the same answer, $d-1$, but if we consider it more carefully we would get

$$S = \frac{1}{4} F_{\mu\nu}^2 V,$$

$$\frac{1}{2} \langle F_{\mu\nu}(k=0) F_{\mu\nu}(k=0) | F_{\mu\nu}(k=0) F_{\mu\nu}(k=0) \rangle = \frac{d(d-1)}{2} \frac{1}{V}, \tag{8.40}$$

giving us at the end

$$\frac{1}{2} \langle F_{\mu\nu}(x) F_{\mu\nu}(y) | F_{\mu\nu}(x) F_{\mu\nu}(y) \rangle = (d-1)\delta^d(x-y) + \frac{(d-1)(d-2)}{2} \frac{1}{V}, \tag{8.41}$$

that is an analog of the Eq. (8.14).

Mathematical Aspects of Kronecker Anomaly

In this section we explain how the Kronecker anomaly arises in a loop computation. Let us consider a quantum field theory at finite temperature β and try to compute any loop diagram. For brevity we will omit the variable p in the latter formulas. According to the Matsubara technique we should sum over the discrete frequencies

$$\mathcal{A}_l(\omega_i) = T \sum_n G_M(i\Omega_n) G_M(i(\Omega_n + \omega_1)) \dots G_M(i(\Omega_n + \omega_1 + \dots \omega_{l-1})). \tag{8.42}$$

Since we have a quantum system at thermal equilibrium

$$G_M(i\omega_n) = \int \frac{d\omega}{2\pi} \frac{\mathrm{Im} G_R(\omega)}{\omega - i\omega_n}. \tag{8.43}$$

Generally, $\mathrm{Im} G_R(\omega)$ depends only on the spectrum of the model but not on the concrete state itself. For instance, for a harmonic oscillator $\mathrm{Im} G_R(\omega) = \pi\delta(\omega^2 - m^2)$. We plug this representation in the amplitude (8.42)

$$\mathcal{A}_l(\omega_i) =$$

$$T \int \prod_{i=1}^{l} \frac{dx_i}{2\pi} \mathrm{Im} G_R(x_i) \sum_n \frac{1}{x_1 - i\Omega_n} \frac{1}{x_2 - i\Omega_n - i\omega_1} \dots \frac{1}{x_l - i\Omega_n - i\omega_1 - \dots - i\omega_{l-1}}$$

Considering summands of this expression as an analytic function of Ω_n we can decompose it as a sum of simple poles $\frac{C(\omega_i)}{x-i\Omega_n}$. The summation over Ω_n yields

$$T\sum_n \frac{C(\omega_i)}{x-i\Omega_n} = \frac{C(\omega_i)}{2}\coth\frac{x}{2T},$$

$$\mathcal{A}_l(\omega_i) =$$

$$\frac{1}{2}\int\prod_{i=1}^{l}\frac{dx_i}{2\pi}\mathrm{Im}G_R(x_i)\left[\coth\frac{x_1}{2T}\frac{1}{x_2-x_1-i\omega_1}\ldots\frac{1}{x_l-x_1-i\omega_1-\ldots-i\omega_{l-1}}+\right.$$
$$\left.+\frac{1}{2}\coth\frac{x_2-i\omega_1}{2T}\frac{1}{x_1-x_2+i\omega_1}\ldots\frac{1}{x_l-x_2-i\omega_2-\ldots-i\omega_{l-1}}+\ldots\right] \tag{8.44}$$

One can see that if $\sum_{i=l}^{k}\omega_i \neq 0$ for any choice of the indices the integral, the amplitude would not have divergences and the resulting function would be analytic. Otherwise there are poles if some sum of the ω's go to zero. For instance, for $l=2$ we have

$$\mathcal{A}_2(\omega_1) = \frac{1}{2}\int\prod_{i=1}^{l}\frac{dx_i}{\pi}\mathrm{Im}G_R(x_i)\frac{\coth\frac{x_1}{2T}+\coth\frac{x_2}{2T}}{x_1+x_2-i\omega_1}. \tag{8.45}$$

This formula is analogous to the susceptibility of the Fermi liquid [17], where such singularities lead to the different k and ω limits. If $\omega_1 \neq 0$ the integral does not contain any singularities and therefore an analytic function of ω_1. If $\omega_1 = 0$ there are possible divergences

$$\lim_{x_1\to -x_2}\frac{\coth\frac{x_1}{2T}+\coth\frac{x_2}{2T}}{x_1+x_2-i\omega_1} = \begin{cases} 0, & \omega_1 \neq 0 \\ -\frac{1}{2T\sinh^2\frac{x_2}{2T}}, & \omega_1 = 0 \end{cases} \tag{8.46}$$

This additional term can amount to the Kronecker anomaly at $\omega_1 = 0$. This singularity happens when $x_1 = -x_2$ and if the product of $\mathrm{Im}G_R(x_1)$ is not enough singular there would not be any contributions. But if $\mathrm{Im}G_R(x_i)$ is singular, the term (8.46) will contribute and we would get a Kronecker anomaly. In the example of the harmonic oscillator

$$\mathrm{Im}G_R(x_1)\mathrm{Im}G_R(x_2) = \pi^2\delta(x_1^2-m^2)\delta(x_2^2-m^2), \tag{8.47}$$

and two-dimensional integral reduces to a sum over the subset $x_1 = \pm x_2 = \pm m$. The additional term (8.46) amounts to the loop computation and coincides exactly

with the discrepancy (8.29). If we introduce a Kronecker function $\delta_{x,y}$

$$\Delta A_2(\omega_1) = -\frac{1}{4T}\int \prod_{i=1}^{l} \frac{dx_i}{\pi} \mathrm{Im} G_R(x_i) \frac{\delta_{x_1,x_2}\delta_{\omega_1,0}}{\sinh^2 \frac{x_2}{2T}} \tag{8.48}$$

We see only stable quasi-particles contributes to the Kronecker anomaly, when $\mathrm{Im}G_R$ has delta-functional singularities.

We present another way of understanding of Kronecker anomaly at $\omega = 0$. Thus, we consider a two-point Whitman function of two operators $A(z = t + i\tau)$ and $B(0)$

$$G_{AB}(z) = \frac{1}{Z}\mathrm{tr}\Big[A(t + i\tau)B(0)e^{-\beta H}\Big] = \tag{8.49}$$

$$\frac{1}{Z}\sum_{n,m} \langle n|A|m|n|A|m\rangle \, \langle m|B|n|m|B|n\rangle \, e^{-\beta E_n} e^{(\tau+it)(E_n-E_m)}.$$

Near the points $\tau = 0, \beta$ this function could be rewritten as

$$G_{AB}(t + i\beta) = \frac{1}{Z}\sum_{n,m} \langle n|A|m|n|A|m\rangle \, \langle m|B|n|m|B|n\rangle \, e^{-\beta E_m} e^{it(E_n-E_m)},$$

$$G_{AB}(t) = \frac{1}{Z}\sum_{n,m} \langle n|A|m|n|A|m\rangle \, \langle m|B|n|m|B|n\rangle \, e^{-\beta E_n} e^{it(E_n-E_m)} = G^*_{AB}(t + i\beta) \tag{8.50}$$

where one can notice that $G_{AB}(t + i\tau)$ is analytic function in the region $0 < \tau < \beta$ (otherwise the sum over n or m would diverge). The analytical structure of this function is drawn in the Fig. 8.1. In the limit $z \to \infty$ the highly oscillating terms

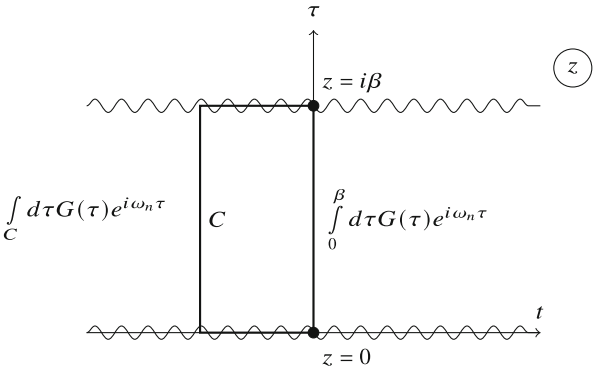


Fig. 8.1 Analytical structure of the two-point function (8.49). The function $G_{AB}(z)$ has two cuts at $\mathrm{Im}z = 0, \beta$

$e^{it(E_n-E_m)}$ would cancel out all contributions unless $E_n = E_m$. Thus we define

$$G_{AB}(\pm\infty) = g_{AB} = \frac{1}{Z}\sum_n \langle n|A|n|n|A|n\rangle \langle n|B|n|n|B|n\rangle e^{-\beta E_n} \in \mathbb{R} \tag{8.51}$$

The Fourier transform of $G_{AB}(\tau)$ is given by

$$G_{AB}(\omega_n) = \int_0^\beta d\tau e^{i\omega_n\tau} G_{AB}(i\tau). \tag{8.52}$$

Let us consider the case of $\omega_n \leq 0$. Then we can deform the contour of integration (Fig. 8.1) and get

$$G_{AB}(\omega_n) = \int_C d\tau e^{i\omega_n\tau} G_{AB}(i\tau) =$$

$$= \lim_{t_0\to-\infty} e^{\omega_n t_0}\int_C d\tau e^{i\omega_n\tau} G_{AB}(i\tau - t_0) + \int_{-\infty}^0 dt e^{\omega_n t}\mathrm{Im}G_{AB}(t) = I_n + \mathcal{F}(\omega_n), \tag{8.53}$$

one can see that the second term $\mathcal{F}(\omega_n)$ is a analytical function of its argument ω_n. The first term vanishes for $\omega_n < 0$ since it is multiplied by a bounded function. But taking into account that $G_{AB} \to g_{AB}$ at infinity we get

$$I_n = \lim_{t_0\to-\infty} e^{\omega_n t_0}\int_C d\tau e^{i\omega_n\tau} G_{AB}(i\tau - t_0) = \frac{g_{AB}}{T}\delta_{\omega_n,0}, \tag{8.54}$$

that is precisely a Kronecker anomaly.

Two Dimensional Scalar Electrodynamics on a Torus

In this section we show how Kronecker anomaly affects a dynamics of quantum electrodynamics and gravity coupled to a massive scalar and fermionic fields on torus $S^1_\beta \times S^1_{\beta'}$. It could be seen as a system at finite temperature with compact spatial coordinate. The action of this model is

$$S = \int d^2x \left[\left|\left(\partial_\mu - ieA_\mu\right)\phi\right|^2 - m^2|\phi|^2 - \frac{1}{4}F_{01}^2\right]. \tag{8.55}$$

We can read out the propagator

$$G(n,m) = \frac{1}{p_n^2 + \omega_k^2 + m^2}, \qquad p_n = \frac{2\pi n}{\beta'}, \quad \omega_k = \frac{2\pi k}{\beta}. \tag{8.56}$$

The one-loop contribution to the polarization operator, i.e. in the lowest order of the coupling constant, e^2,

$$\Pi_{00}(\Omega, q) =$$
$$\frac{e^2}{\beta'\beta} \sum_{n,m} \left[\frac{(2\omega_n + \Omega)^2}{(\omega_n^2 + p_m^2 + m^2)\left((\omega_n - \Omega)^2 + (p_m - q)^2 + m^2\right)} - \frac{2}{\omega_n^2 + p_m^2 + m^2} \right]$$

Using Ward identities we can find the following relations for the polarization operators

$$\Omega\Pi_{00}(\Omega, q) + q\,\Pi_{01}(\Omega, q) = 0, \quad \text{and} \quad \Omega\Pi_{01}(\Omega, q) + q\,\Pi_{11}(\Omega, q) = 0. \tag{8.57}$$

We note that if $\Omega \neq 0$ then we can express $\Pi_{00} = -\frac{q}{\Omega}\Pi_{01}$ and $\Pi_{00}(\Omega, q)$ must be equal to zero for $q = 0$ and $\Omega \neq 0$. The equations (8.57) do not constrain the polarization operators only when $\Omega = q = 0$.

Therefore, non-analytical contributions can come only at $\Omega = q = 0$. Thus we set $\Omega = 0$ and consider the corresponding component as a function of the spatial momentum, q,

$$\Pi_{00}(0, q) = \frac{2e^2}{\beta'\beta} \sum_{n,m} \left[\frac{2\omega_n^2}{\left(\omega_n^2 + p_m^2 + m^2\right)\left(\omega_n^2 + (p_m + q)^2 + m^2\right)} - \frac{1}{\omega_n^2 + p_m^2 + m^2} \right]. \tag{8.58}$$

Summing this series when $q = 0$ and $q \neq 0$ we get the following result,

$$\Pi_{00}(0, q) = \pi_{00}\,\delta_{q,0} + \widetilde{\Pi}_{00}(0, q), \quad \pi_{00} = \frac{e^2}{\beta} \sum_n \frac{\omega_n^2}{\omega_n^2 + m^2} \frac{1}{\sinh^2\left(\frac{\sqrt{\omega_n^2 + m^2}}{2T'}\right)}. \tag{8.59}$$

The last term brings to the effective action of the electromagnetic field an additional non-trivial interaction for $A_0(0, 0)$.

To understand this better we can study the $2d$ model with massless Dirac fermions on the same torus $S^1_\beta \times S^1_{\beta'}$ (with periodic boundary conditions over the

spatial circle $S^1_{\beta'}$). In this case the propagator for the Kronecker mode is

$$\Pi_{00}(\omega=0,q)=e^2TT'\sum_{n,m}\frac{2\omega_n^2-2p_m(p_m+q)}{\left(\omega_n^2+p_m^2\right)\left(\omega_n^2+(p_m+q)^2\right)}=$$
$$=\frac{1}{q}e^2T'\sum_m\left(\tanh\left(\frac{p_m+q}{2T}\right)-\tanh\left(\frac{p_m}{2T}\right)\right),\ \text{here } T=\frac{1}{\beta}\text{ and }T'=\frac{1}{\beta'} \tag{8.60}$$

If $q\neq 0$ we can shift the summation and compute the difference between the first and second term in the summand to find that:

$$\Pi_{00}(\omega=0,q)=\frac{e^2}{2\pi} \tag{8.61}$$

In the limit as $q\to 0$ we must replace the difference by the derivative to get that

$$\Pi_{00}(\omega=0,q=0)=e^2\frac{\beta}{\beta'}\sum_n\frac{1}{\cosh^2\frac{\pi n\beta}{2\beta'}}=e^2g\left(\frac{\beta}{\beta'}\right)+\frac{e^2}{2\pi},$$
$$\text{where}\quad g(x)=\sum_{n=-\infty}^{\infty}\frac{x}{\cosh^2\frac{\pi nx}{2}}-\frac{1}{2\pi}. \tag{8.62}$$

The sum is different again for $q=0$ and $q\neq 0$. We would like to stress that this additional term is not connected to the dynamics of the Wilson lines—it is already captured by Schwinger term. Indeed, the Wilson line is equal to $W_t(x)=\int d\tau A_0(x,\tau)$ or in Fourier space $W_t(q)=A_0(0,q)=\frac{F_{01}(0,q)}{q}$ then we can rewrite the above contribution to the action as

$$S=\frac{e^2}{2\pi}\sum_q\frac{F_{01}(0,q)F_{01}(0,-q)}{q^2}+e^2g\left(\frac{\beta}{\beta'}\right)W_t^2(q=0)= \tag{8.63}$$
$$=\sum_q W_t(q)W_t(-q)+e^2g\left(\frac{\beta}{\beta'}\right)W_t^2(q=0)=$$
$$=\int dx\,W_t^2(x)+e^2g\left(\frac{\beta}{\beta'}\right)\left(\int dx\,dt\,A_0\right)^2.$$

The action contains the local dynamics of the Wilson lines, while the last term does not have the same form and brings an additional dynamics. These non-local contributions at all orders of perturbation theory were firstly computed in [18].

An analogous calculations could be performed for the gravitational case. We will work with the linearized gravity in order to apply intuition we developed in the abelian case. Let us study a theory symmetric field h_{ab} that is invariant under the

following transformations

$$h_{ab} \to h_{ab} + \partial_a \epsilon_b + \partial_b \epsilon_a, \quad h_{ab}(k) \to h_{ab}(k) + k_a \epsilon_b + k_b \epsilon_a. \tag{8.64}$$

The Fourier transform of the Eq. (8.64) degenerates for $\mathbf{k} = 0$, meaning that additional degrees of freedom could contribute to the dynamics. We couple this gravity theory to a fermion field on the torus $\mathbb{T}^2 = S^1_\beta \times S^1_{\beta'}$. If we integrate out the fermions from the theory, these new degrees of freedom could become alive. The gravity field h_{ab} is coupled to a stress energy tensor

$$\delta S = \int d^2x h_{ab} T_{ab},$$

$$T_{00} = \overline{\psi}\gamma_0 \partial_0 \psi, \quad T_{01} = \frac{1}{2}\overline{\psi}\left[\gamma_0 \partial_1 + \gamma_1 \partial_0\right]\psi, \quad T_{11} = \overline{\psi}\gamma_1 \partial_1 \psi. \tag{8.65}$$

Due to the gauge invariance

$$\Pi_{00|00}(\Omega, q = 0) = \langle T_{00}(0, q) T_{00}(0, -q) | T_{00}(0, q) T_{00}(0, -q) \rangle = 0$$

when $\Omega \neq 0$. Hence, to get the non-analytical contributions to the polarization operator $\Pi_{00|00}$ we set $\Omega = 0$ and study it as a function of q. The computation gives the following polarization operator

$$\Pi_{00|00}(\Omega = 0, q) = \frac{1}{\beta\beta'} \sum_{n,m} \frac{\omega_n^2 \left(\omega_n^2 - p_m(p_m + q)\right)}{(p_m^2 + \omega_n^2)((p_m + q)^2 + \omega_n^2)}. \tag{8.66}$$

Summation over frequencies ω_n could be performed and it yields:

$$\Pi_{00|00}(\Omega = 0, q) = \frac{1}{\beta'} \sum_n \frac{(p_n + q)^2 \tanh \frac{\beta(p_n+q)}{2} - p_n^2 \tanh \frac{\beta p_n}{2}}{2q} \tag{8.67}$$

This sum over p_n is seemingly divergent but after regularization for $q \neq 0$:

$$\Pi_{00|00}(\Omega = 0, q \neq 0) = -\frac{q^2}{6}. \tag{8.68}$$

This term brings to the effective action the following term

$$\delta S = h_{00}(\omega = 0, q) \frac{q^2}{6} h_{00}(\omega = 0, -q) = \frac{R(\omega = 0, q) R(\omega, -q)}{6q^2}, \tag{8.69}$$

where we have used that at linear order $R(\omega = 0, q) = -q^2 h_{00}(\omega = 0, q)$. Thus, we have managed to restore Liouville action on a torus. This fact is not surprising.

But if we consider the case of $q = 0$, when the constraints trivially satisfied, we get

$$\delta\Pi_{00|00}(\Omega = 0, q = 0) = \sum_n \frac{p_n^2 + 2\frac{p_n}{\beta}\sinh[p_n\beta]}{\cosh^2\left[\frac{p_n\beta}{2}\right]} \tag{8.70}$$

This term brings a non-local interaction to the effective action. It would be interesting to understand better the origin and the meaning of this term.

Kronecker Anomaly in Spaces with Constant Curvature

Before we considered only the cases of the spaces with compact subspace $\mathcal{M} = \widetilde{\mathcal{M}} \times S^1$. The calculation was quite explicit and Kronecker anomalies appeared. It would be interesting to study the case of other compact spaces. The most direct generalization is to consider a three-dimensional sphere S^3. The problem is quite interesting and related to the problem of de Sitter stability [19, 20]. Also it has already been shown that loop corrections for principal series the loop corrections contain non-analytical contributions [21]. Let us study the following theory

$$S = \frac{1}{2}\int d^3x\sqrt{-g}\left[\left(\partial_\mu a\right)^2 - m^2a^2 + \lambda a^3\right]. \tag{8.71}$$

We pick the following metric on the sphere

$$ds^2 = d\phi^2 + \sin^2\phi\left(d\theta^2 + \sin^2\theta\, d\varphi^2\right). \tag{8.72}$$

Due to high symmetry of S^3 two point functions depend on the geodesic distance $Z = \cos\phi$ [20]

$$G_\mu(\cos\phi) = \frac{\sinh\mu\,(\pi - \phi)}{\sinh\mu\pi\,\sin\phi} = \sum_{L=0}^{\infty}\frac{(L+1)}{L(L+2)+m^2}\frac{\sin(L+1)\phi}{\sin\phi}, \tag{8.73}$$

where $\mu = \sqrt{m^2 - 1}$ and $\frac{\sin(L+1)\phi}{\sin\phi}$ is a eigenfunction on the three-dimensional sphere. Let us note that

$$G_\mu(\cos\phi) = \sum_{K=-\infty}^{\infty}\frac{K}{K^2+\mu^2}\frac{\sin K\phi}{\sin\phi}, \tag{8.74}$$

while to get Feynman propagator we should replace the summation with an integral

$$Q_\mu(\cos\phi) = \int\limits_{-\infty}^{\infty} \frac{dK}{2\pi} \frac{K}{K^2+\mu^2} \frac{\sin K\phi}{\sin\phi} \tag{8.75}$$

We want to decompose the one-loop correction $\Pi(\cos\phi) = G_\mu^2(\cos\phi)$ through the eigenfunctions of the 3d sphere

$$P_L(\cos\phi) = \frac{\sin(L+1)\phi}{\sin\phi} = \sum_{K=L,L-2,\ldots,-L} e^{iK\phi},$$

$$\int\limits_0^\pi d\phi \sin^2\phi P_L(\cos\phi) P_K(\cos\phi) = \delta_{LK},$$

$$\Pi(\cos\phi) = \sum_L \Pi_L P_L(\cos\phi), \quad \Pi_L = \int\limits_0^\pi d\phi \sin^2\phi\, \Pi(\cos\phi)\, P_L(\cos\phi), \tag{8.76}$$

we get that

$$\Pi_L = \sum_{K=L,L-2,\ldots,-L} \pi_K, \quad \pi_K = \frac{1}{\sinh^2\pi\mu} \frac{\mu\sinh 2\pi\mu}{K^2+4\mu^2} - \frac{\pi}{2\sinh^2\pi\mu}\delta_{K,0}. \tag{8.77}$$

Summing this series over we get the following expression

$$\Pi_L = -\frac{i}{4}\coth\pi\mu\left[H_{-i\mu-\frac{L}{2}-1} - H_{i\mu-\frac{L}{2}-1} - H_{-i\mu+\frac{L}{2}} + H_{i\mu+\frac{L}{2}}\right] - \frac{\pi\delta_{L,\text{even}}}{2\sinh^2\pi\mu} \tag{8.78}$$

As one can see there is a non-analytical contribution at even L's. If we sum them over, we would get that the Kronecker anomaly depends on ϕ

$$\delta\Pi(x,y) = \delta\Pi(\cos\phi) = -\frac{\pi}{2\sinh^2\pi\mu}\sum_k \frac{\sin(2k+1)\phi}{\sin\phi} = -\frac{\pi}{2\sinh^2\pi\mu}\frac{1}{\sin^2\phi} \tag{8.79}$$

Does it mean that the analytical continuation from S^3 to dS^3 is problematic? And if we change adiabatically the mass, m^2, we would get isentropic change of the $\langle a^2|a^2\rangle$?

As in the case of harmonic oscillator we can show that Π_0 corresponds to the change of the mass, either keeping temperature or entropy constant. Indeed,

$$G_\mu(\cos\phi) \approx \frac{1}{\phi} - \mu\coth\pi\mu + O(\phi), \quad -\frac{\partial G_\mu(Z=1)}{\partial\mu^2} = \frac{\coth\pi\mu - \frac{\pi\mu}{\sinh^2\pi\mu}}{2\mu} = \Pi_0 \tag{8.80}$$

On the other hand we can do this calculation in the dS^3 (that is done by analytical continuation $\phi \to i\alpha$) with the use of Keldysh-Schwinger technique

$$G_\mu(\cosh\alpha) = i\frac{\sinh\mu\,(\pi - i\alpha)}{\sinh\mu\pi\,\sinh\alpha},$$

$$-\frac{\partial G_\mu(Z=1)}{\partial\mu^2} = \int_0^\infty d\alpha\,\sinh^2\alpha\,\mathrm{Im}\left[G_\mu^2(\cosh\alpha)\right] = \frac{\coth\pi\mu}{2\mu}, \tag{8.81}$$

that is equal to the Π_0 without the Kronecker term. It means that if we change μ^2 in the dS_3 instead of going to the BD vacuum we will descend in some other vacuum. Let us stress that we integrated over global dS_3 space. Therefore this result is not applicable for the Expanding Poincare Patch. Nonetheless, we can study the Static Patch of de Sitter [22] which is related again to three-dimensional sphere by analytical continuation. It would be interesting to see whether such an non-analytical contribution to Π_L would also relate the difference in the isothermal and isentropic susceptibilities. It shows, that the quantum field theory considered in the odd-dimensional de Sitter space does not make any sense—the system is very sensitive to the way it was prepared. Any small adiabatic deformation of the theory could change its properties.

Absence of Kronecker Anomalies in Even Dimensional de Sitter Spaces

We want to generalize the previous calculation to S^2. The propagator on a 2d sphere is

$$G(Z) = \frac{\pi}{\cosh\pi\mu} P_{-\frac{1}{2}+i\mu}(-Z) \tag{8.82}$$

Namely, we are interested in the decomposition of the self-energy function $\Pi(Z) = G(Z)^2$ as a sum over 2d sphere eigenfunctions

$$\Pi(Z) = \sum_{L=0}^{\infty} (2L+1)\,P_L(Z)\Pi_L, \tag{8.83}$$

where $P_L(Z)$ are Legendre polynomials. Using orthogonality we can extract coefficients

$$\Pi_L = \frac{1}{2}\int_{-1}^{1} dZ\,\Pi(Z)P_L(Z), \tag{8.84}$$

this integral could be rewritten in the following way. First we notice that we have Legendre Q function of integer index with the following property

$$\mathrm{Im}\, Q_n(x) = \frac{\pi}{2}P_n(x), \quad x \in [-1,1]\,, \tag{8.85}$$

then the Eq. (8.84) is

$$\Pi_L = \frac{1}{\pi}\int_{-1}^{1} dZ\,\Pi(Z)\,\mathrm{Im}\, Q_L(Z) = \frac{1}{2\pi i}\int_{C} dZ\,\Pi(Z)Q_L(Z), \tag{8.86}$$

where C encircles the set $[-1,1]$, we deform this contour to $C_\infty = \lim_{R\to\infty}\{Z\in\mathbb{C}, |Z| = R\}$. Since $\Pi(Z)Q_L(Z) \sim \frac{1}{Z^{L+2}}$ for $L \geq 0$

$$\Pi_L = \frac{1}{2\pi i}\int_{C_\infty} dZ\,\Pi(Z)Q_L(Z) = \frac{1}{\pi}\int_{1}^{\infty} dZ\,\mathrm{Im}\,\Pi(Z)Q_L(Z), \tag{8.87}$$

but now Q_L is a smooth function of L, namely $Q_L = Q_\nu e^{i\pi\nu}$, so we can introduce the following function

$$F(\nu) = \frac{1}{\pi}\int_{1}^{\infty} dZ\,\mathrm{Im}\,\Pi(Z)Q_\nu(Z)e^{i\pi\nu}, \tag{8.88}$$

this function is analytic and $F(L) = \Pi_L$ for integer L. We can check numerically that it is a nice agreement with the explicit resummation. Therefore, there are no Kronecker anomalies for even dimensional quantum field theories in de Sitter spaces.

De Sitter Lacuna

In this section we will briefly discuss possible discontinuities of the propagator as a function of geodesic distance or interval in the background of non-trivial metric.

The most simple example of such a background is again a de Sitter space. It is very well-known, that in de Sitter space due to external gravity, the in-vacuum is not stable. It could be seen by directly computing the decay rate

$$\langle \text{out}|\text{in}|\text{out}|\text{in}\rangle = \int \mathcal{D}\phi e^{i\int d^d x\left[\frac{1}{2}(\partial_\mu\phi)^2-\frac{1}{2}m^2\phi^2\right]},$$

$$\frac{\partial \log\langle \text{out}|\text{in}|\text{out}|\text{in}\rangle}{\partial m^2} = \int d^d x\, \langle \text{out}|\phi^2(x)|\text{in}|\text{out}|\phi^2(x)|\text{in}\rangle \tag{8.89}$$

The in-out matrix element could be computed in the following way. For instance in $d = 2$, we would have the following in-out propagator

$$\langle \text{out}|\phi(x)\phi(y)|\text{in}|\text{out}|\phi(x)\phi(y)|\text{in}\rangle = Q_{-\frac{1}{2}+i\mu}(Z), \quad \mu = \sqrt{m^2-\frac{1}{4}}, \tag{8.90}$$

where Z is a geodesic distance between these two points x and y in the de Sitter space. Then the imaginary part of the partition function is

$$\text{Im}\left[\frac{\partial \log\langle \text{out}|\text{in}|\text{out}|\text{in}\rangle}{\partial \mu^2}\right] = \pi \tanh \pi\mu V_{dS}, \tag{8.91}$$

that suggests that in- and out-states in de Sitter space are orthogonal to each other. Therefore, we would expect if one turns on the interaction the in-state would decay to the out-state. It would be interesting to understand how the in-in propagator transforms to out-out propagator. In 2d, the in-in propagator could be obtained by a analytical continuation from sphere

$$G_{\text{in/in}}(Z) = \sum_{L=0}^{\infty} \frac{(2L+1)P_L(Z)}{L(L+1)+m^2} = \frac{\pi}{\cosh \pi\mu} P_{-\frac{1}{2}+i\mu}(-Z), \tag{8.92}$$

the out-out propagator could not be computed by some analytical continuation. Nevertheless, we can write down the expression for the out-out propagator in terms of modes

$$G_{\text{out/out}}(Z) = \int \frac{dp}{2\pi} e^{ipx} J_{i\mu}(p\eta) J_{-i\mu}(p\eta), \tag{8.93}$$

we can see that the product of two Bessel functions, $J_{i\mu}(p\eta)J_{-i\mu}(p\eta)$ is an entire function of momentum p. Then if $|x| > 2\eta$ we can change the counter of integration to infinity

$$\int \frac{dp}{2\pi} e^{ipx} J_{i\mu}(p\eta) J_{-i\mu}(p\eta) = 0, \quad |x| > 2\eta \quad \Rightarrow \quad G_{\text{out/out}}(Z) = \theta(Z+1)D(Z) \tag{8.94}$$

Plugging this assumption into the equation for propagator we can see that $D(Z)$ should satisfy the same equation for massive scalar field in de Sitter space. And since $G_{\text{out/out}}(Z)$ does not have singularities at $Z = -1$ we have

$$G^{d=2}_{\text{out/out}}(Z) = \theta(Z+1)G^{d=2}_{\text{in/in}}(Z), \tag{8.95}$$

we can find analogous formulae in higher dimensions

$$\begin{aligned} G^{d=2k+2}_{\text{out/out}}(Z) =& \frac{d^k}{dZ^k} G^{d=2}_{\text{out/out}}(Z) \\ =& \theta(Z+1)G^{d=2k+2}_{\text{in/in}}(Z) + C_1\delta(Z+1) + C_2\delta'(Z+1) + \dots \end{aligned}$$

In the case of the odd dimensions the situation a bit more strange. First, the in-out propagator does not contain imaginary contributions and therefore the in-state could not decay to the out-state [23]. Second, the out-out propagator would be different

$$\begin{aligned} G_{\text{out/out}}(Z) = \int \frac{d^2p}{(2\pi)^2} e^{ipx} J_{i\mu}(p\eta) J_{-i\mu}(p\eta) = \\ = \int \frac{pdp d\cos\theta_p}{2\pi} \frac{d\theta}{2\pi} \frac{d\theta'}{2\pi} e^{i\mu(\theta-\theta')} e^{ip[x\cos\theta_p + \eta\cos(\theta-\theta_p) + \eta\cos(\theta'-\theta_p)]}, \end{aligned} \tag{8.96}$$

from this we can see that if $|x| > 2\eta$ this integral is zero. For $|x| < 2\eta$ we can compute this integral

$$G^{d=3}_{\text{out/out}}(Z) = \theta(Z+1)\frac{e^{i\mu\theta}}{\sinh\theta} = \theta(Z+1)G_{\text{in/out}}(Z), \tag{8.97}$$

that is drastically different from the behavior in even dimensions. Again in higher dimensions

$$G^{d=2k+3}_{\text{out/out}}(Z) = \frac{d^k}{dZ^k} G^{d=3}_{\text{out/out}}(Z) = \theta(Z+1)G^{d=2k+3}_{\text{in/out}}(Z) + C_1\delta(Z+1) + \dots \tag{8.98}$$

Acknowledgments We are grateful to Boris Altshuler, Igor R. Klebanov, Juan M. Maldacena, Emil T. Akhmedov and P. Wiegmann for insightful discussions and suggestions throughout the project. F.K.P. is currently a Simons Junior Fellow at NYU and supported by a grant 855325FP from the Simons Foundation.

References

1. K.G. Wilson, Non-Lagrangian models of current algebra. Physical Review **179**(5), 1499 (1969)
2. L.P. Kadanoff, Operator algebra and the determination of critical indices. Phys. Rev. Lett. **23**(25), 1430 (1969)
3. A. Polyakov, Properties of long and short range correlations in the critical region. Zh. Eksp. Teor. Fiz **57**, 271–283 (1969)
4. A. Larkin, S. Pikin, Zh. ETF **56**, 1664 (1969); Soy. Phys. JETP **29**, 891 (1969)
5. S. Pikin, Weak first-order phase transitions. Phys. A Stat. Mech. Appl. **194**(1-4), 352–363 (1993)
6. A.M. Polyakov, Self-tuning fields and resonant correlations in 2d-gravity. Modern Phys. Lett. A **6**(07), 635–644 (1991)
7. L.D. Landau, E.M. Lifshitz, A.M. Kosevich, L.P. Pitaevskii, *Theory of Elasticity*, vol. 7 (Elsevier, 1986)
8. L. Golubović, T. Lubensky, C. O'hern, Structural properties of the sliding columnar phase in layered liquid crystalline systems. Phys. Rev. E **62**(1), 1069 (2000)
9. S. Coleman, Why there is nothing rather than something: a theory of the cosmological constant. Nucl. Phys. B **310**(3-4), 643–668 (1988)
10. S.W. Hawking, Do wormholes fix the constants of nature? Nucl. Phys. B **335**(1), 155–165 (1990)
11. J.G. Polchinski, *String Theory, Volume I: An Introduction to the Bosonic String* (Cambridge University Press, Cambridge, 1998)
12. I.R. Klebanov, A.M. Polyakov, Interaction of discrete states in two-dimensional string theory. Mod. Phys. Lett. **A6**, 3273–3281 (1991). hep-th/9109032
13. A.M. Polyakov, Singular states in 2-d quantum gravity, 9 (1991). https://doi.org/10.1142/9789814538992_0006
14. E. Witten, Ground ring of two-dimensional string theory. Nucl. Phys. B **373**(1), 187–213 (1992)
15. D.J. Gross, T. Piran, S. Weinberg, *Two Dimensional Quantum Gravity and Random Surfaces-8th Jerusalem Winter School for Theoretical Physics*, vol. 8 (World Scientific, 1991)
16. L.D. Landau, E.M. Lifshits, *Statisticheskaia fizika*, vol. 5 (Pergamon, 1980)
17. A. Abrikosov, L. Gor'kov, I. Dzyaloshinskii, *Methods of Quantum Field Theory in Statistical Physics [in Russian]* (GIFML, Moscow, 1962)
18. I. Sachs, A. Wipf, Finite temperature Schwinger model. Helv. Phys. Acta **65**, 652–678 (1992). 1005.1822
19. E. Akhmedov, Lecture notes on interacting quantum fields in de Sitter space. Int. J. Mod. Phys. D **23**, 1430001 (2014). 1309.2557
20. A. Polyakov, Infrared instability of the de Sitter space. 1209.4135
21. E. Akhmedov, U. Moschella, K. Pavlenko, F. Popov, Infrared dynamics of massive scalars from the complementary series in de Sitter space. Phys. Rev. D **96**(2), 025002 (2017). 1701.07226
22. E. Akhmedov, K. Bazarov, D. Diakonov, U. Moschella, Quantum fields in the static de Sitter universe. 2005.13952
23. E.T. Akhmedov, K.V. Bazarov, D.V. Diakonov, U. Moschella, F.K. Popov, C. Schubert, Propagators and Gaussian effective actions in various patches of de Sitter space. Phys. Rev. **D100**(10), 105011 (2019). 1905.09344

Chapter 9
Projecting Local and Global Symmetries to the Planck Scale



Gerard 't Hooft

Abstract The Standard Model of the elementary particles is controlled by more than 20 parameters, of which it is not known today how they can be linked to deeper principles. Any attempt to clean up this theory, in general results in producing more such parameters rather than less. Yet it is clear that, at distance scales of the order of the Planck length, the gravitational force presents itself in such a way that the need for new physical principles is evident.

A stand-in-the-way is then quantum mechanics, a theory that demands the occurrence of superpositions of physical states in such a way that, when combined with general relativity, space and time themselves may require new formalisms for being used as primary frames for the descriptions of events.

In previous papers the author proposed that quantum mechanics as a theory for the elementary particles should be rephrased as originating from a combination of deterministic evolution laws and discreteness at the Planck scale. This may well have a drastic effect on the symmetry structures and algebras. Local, discrete and continuous symmetries do not emerge without a cause, and we suggest that the symmetries can tell us more about fundamental constants, among which the Higgs mass is the most peculiar and the most challenging one.

9.1 Introduction

In 1954, Chen Ning Yang and Robert L. Mills published [1] their idea that Maxwell's equations for the electro-magnetic force could be generalised, such that the underlying local symmetry involving the vector field $A_\mu(\mathbf{x}, t)$ would turn into a non-Abelian one, and the quanta of this vector field would become more complex than the familiar photons, as they would interact directly with one another, and instead of a single type of electric charge, there would be an algebra of charges.

G. 't Hooft (✉)
Faculty of Science, Department of Physics, Institute for Theoretical Physics, Utrecht, The Netherlands
e-mail: g.thooft@uu.nl

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_9

The properties of the interactions are almost completely determined by this algebra, but an important exception has always been the universal strengths of these interactions; it has been generally agreed upon that the interaction strengths are given by one freely adjustable parameter for each invariant subgroup of the algebra, forming a set of constants of nature, whose actual values are freely adjustable in the theory, so that they can only be determined by meticulous measurements.

The discovery of Yang-Mills fields may be regarded as one of a few elementary milestones of our journey into the realm of elementary particle physics. It inspired other scientists to study the possibility that photons of the Yang-Mills type might exist, hidden somewhere in the world of elementary particles, and in the mean time investigate various types of currents that may be related to the Yang-Mills currents, and could play a role in other interactions.

There were other milestones in this field, such as the discovery that the Yang-Mills interactions tend to be modified by vacuum effects, leading to the existence of scalar particles that would hide the Yang-Mills symmetry [2–4], the so-called Higgs particles. Subsequently it was discovered that almost all interactions among the elementary particles can eventually be reduced to Yang-Mills interactions. Indeed, while most theories for the interactions of elementary particles had the tendency to generate infinitely strong forces and mass terms, particularly when particles had spin, the Yang-Mills theory was found to be one of the very few theories that allowed for the interaction strengths and mass terms to be renormalised so that the infinities disappeared, in spite of the presence of spin-one particles, and all this without violating important principles such as locality and unitarity.

A conclusion that should be emphasised, is: *if you believe that the Yang-Mills theory may mean something important in particle physics, you should not take just a piece of it* (like focusing only on the currents) *or mutilate it* (such as adding mass terms that violate the local symmetry), *but one should accept the theory all the way, without modifying its principles.* We physicists had to learn that the correct way to generate masses for the vector particles is to invoke the Brout-Englert-Higgs mechanism, which does not affect local gauge invariance. The deeper significance of such lessons will be emphasised when we hit upon another example shortly in this chapter.

The Lagrangian $\mathcal{L}$ most often used in elementary particle physics may be written symbolically like

$$\mathcal{L} = -\sum_k \tfrac{1}{4g_k^2} F^k_{\mu\nu} F^k_{\mu\nu} - \sum_j \overline{\psi}_j (\gamma^\mu (\partial_\mu + iA_\mu) + m_j)\psi_j \\ - \tfrac{1}{2}((\partial_\mu + iA_\mu)\phi)^2 - V(\phi) - \overline{\psi} Y(\phi)\psi, \tag{9.1}$$

where we used some abbreviations to represent the real mathematical structure. Here, the index k counts the invariant subgroups of the algebra, and the coefficients g_k are the freely adjustable Yang-Mills coupling coefficients, while the scalars ϕ come with similar interaction coefficients λ_i in the expressions for $V(\phi)$, and their Yukawa interactions with the fermions ψ come with coefficients Y_i .

Not only may these coupling parameters be chosen arbitrarily, in principle, but they also depend on the momentum- and energy scale of the interactions considered, a dependence controlled by renormalization group equations [5]. These facts actually raise important mathematical questions that can be phrased but have not been properly answered. We now propose that the mathematical basis of these theories has not yet been explored sufficiently rigorously. This paper is a plea to investigate alleys for improvement here; as yet physicists' attitude has often been that the theories work just fine up to the extent that we really need them.

It has become largely accepted that perturbation theory is reliable as long as the expansion coefficients, g_k, λ_i and Y_i are kept sufficiently small. However, what does 'sufficiently small' really mean? Indeed, nature does not always keep them small. Since these coefficients depend on the energy scales of the particles involved, there are occasions where the power expansions in these variables do not converge sufficiently well. One may try to avoid these domains, but it is not well-understood whether this can always be done, and how the entire theory should be redefined if we ask for mathematical rigour everywhere. For instance, in the $U(1)$ sector of the theory, the expansion parameter g_1 tends to blow up, producing a Landau ghost in the amplitudes in the high-energy/small distance domain, exactly where we would demand mathematical accuracy. *Can our theories be defined with mathematical rigour?* It looks as if this is not the case.

We return to this point soon, as we have ideas for improvement.

9.2 Quantum Mechanics

Elementary particles obey the laws of quantum mechanics. Most authors are convinced that these laws are fundamentally different form any classical system, as has been argued by J.S. Bell [6]. Bell considered the time evolution of a quantum system, and asked how we should define the concept of *causality,* to relate states or configurations in the future, to what they may have been in the past.

He thought that statistical configurations in the past must be fundamentally uncorrelated, and that the outcome of experimental measurements, as predicted by quantum theory and later checked accurately, show correlations that cannot be explained in any deterministic law or model. Most of these arguments, however, were verbal and not corroborated by complete sets of equations of motion, in any explicit physical model.

Present author has been suspicious of such arguments for a long time. And indeed, we recently found models that are fundamentally deterministic, and, most importantly, statistical distributions in the past are strongly affected by states a system may be observed to be in in the future. This is because there is information in the present state, and deterministic equations cannot erase such information, so, in a hidden form, knowledge of the present gives highly entangled information (in the classical sense) concerning the states in the past [7].

The way to phrase what we have found in [7] is, that quantum mechanics should be regarded as a *vector representation* of all physical states allowed in any kind of physical model. This underlying physical model may be completely classical and deterministic in the usual sense.

In a somewhat different wording, the same conclusion was arrived at by Wetterich [8], who describes fermion models in terms of deterministic evolution laws: the only reason why particles obey probabilistic laws is that the initial states are distributed probabilistically. The evolution laws themselves are classical.

We emphatically include *superdeterminism* [9], which here means that also the behavior of the observer, and the settings chosen in his/her measurement devices cannot be viewed independently from the physical states of all other participating particles.

Indeed, assuming an underlying model to be *completely* deterministic removes most of the 'quantum paradoxes' that were thought to be special to quantum mechanics alone. *If one is prepared to consider deterministic theories for elementary particles, one must accept determinism, as well as 'superdeterminism', all the way* [9].

An important example here is the way physicists assumed *statistics* should be incorporated in a causal, deterministic model. A model may only be truly deterministic, and truly causal, if its dynamical variables do not depend on real numbers but only on integers, or perhaps also some rational numbers. In this case, it will be a lot more straightforward to employ *time reversal*. If real numbers would be involved, the statistical behavior of a system such as a gas or liquid, consisting of atoms, would exhibit chaos, controlled by Lyapunov exponents. Statistical mixing would occur naturally, and assuming statistical independence would then be possible. However, one can easily imagine models where time reversal symmetries hold, but where only integers describe the relevant variables. The number of physically allowed states would then naturally be limited by thermodynamical rules. What happens can be visualised in computer simulations: motion backwards in time may seem to generate statistical mixture, but regardless how far we go to the past, we shall find that statistical mixture is only apparent, the model going backwards in time generates as many correlations, and as much entanglement, as the model forwards in time.

This is particularly of importance in the interpretation of Bell's theorem. If, in his Gedanken experiment, either Alice or Bob 'choose a different setting', for instance by switching a variable in their computer simulation, one finds that, at all times, statistical correlations between photons my have altered, in the entire past as much as in the entire future.

The key element in our classical models that overturns Bell's theorem is the observation that *superpositions* are man-made structures that describe probabilities in quantum mechanics, while probabilities in a classical theory do not include quantum superpositions. Forbidding quantum superpositions in the classical description affects the way statistical distributions propagate in time. It implies that, whenever Alice and or Bob make even the slightest modification in their settings of an experiment, quantum physics would assume the photons they measure to be

superpositions of the previous photons, while in the classical descriptions these photons are altogether different. Bell assumed that superimposed photons evolve into superimposed photons. Quantum mechanically they do, but classically they can't, while the classical probabilities can be modelled in such a way, that the quantum experiments are indeed reproduced. I refer to our latest papers [10] where these features are elaborated.

While one might accuse philosophers of science of having not gone deeper into the equations of motion for classical systems, one might also criticise physicists, in turn, for not taking the problems raised by philosophers seriously enough. Remedies of the putative shortcomings will affect the Standard Model almost certainly when we try to incorporate 'quantum gravity' effects. Usually, this is considered to be the domain of superstring theory, and philosophers are asked to get out of the way. But string theory [11] is not deterministic by construction,[1] and we now suspect that this is an essential weakness.

In short, the lesson we had learned with Yang-Mills theory was that local gauge invariance had to be respected as a primary principle all the way, instead of partly hanging on to older ideas. It is not self-evident, but we propose that here also, we should pick up this lesson right at the present point, by saying that determinism should be embraced all the way as a primary physical principle, that is, we demand it to hold without any exceptions. And now we propose to unravel the consequences of this principle in the case of the Standard Model and the question how it is to be linked with the gravitational force.

In the past, it was 'quantum mechanics' that was thought to be the fundamental basic principle for all our theories about the smallest units of mass, energy, momentum, and so on, but time has come to accept the discovery that there is no strict dividing line between quantum mechanics and classical logic. We found basic models that can be handled entirely as if they were quantum models, but that can be observed to obey the strictest form of determinism at the same time. These models are also essentially based on *discreteness*, and this may indicate that freely adjustable parameters in theories such as the Standard Model, should actually be limited to discrete subsets. What these subsets exactly are, is still beyond our ability to determine, but we hope that further research here will bring new light.

One of the most serious concerns in our understanding of quantum mechanics, has been the issue of *locality*. In Bell's experiment, as in many other quantum 'paradoxes', it seemed that 'spooky signals' may be travelling faster than light, just so as to restore information that was thought to be lost due to statistical independence. Again, these are mere appearances. The 'spooky signals' behave totally normally if we follow them backwards in time. We emphasise again that the models we start off with, are ordinary deterministic models, such as a classical cellular automaton. In terms of the variables in such models, all signals are constrained by speed limits, so that our classical systems obviously do not support signals going much faster than light. All we now have to add to this is that our

[1] But its state counting properties might suggest otherwise, see [12].

quantum systems are based on exactly such models. Spooky signals may arise in our descriptions of the vector representation that reproduces our Schrödinger equation, but since all real information of the system is in its classical variables, these spooky signals cannot be used to transmit observable signals.

Our theory predicts that 'real' quantum miracles will never occur; however, this explicitly includes Planck length physics. Therefore there may be quantum effects at the atomic scale that can only be mimicked by classical systems after scaling the latter to Planckian dimensions. Quantum features more exotic than that must be left to where they belong: in the science fiction literature.

9.3 Classical Models Underlying Quantum Mechanics

Consider a system that only has a finite number of states (later, one can always consider the infinite limit case). The evolution law dictates how the system evolves in one smallest time step δt. Such a system must be periodic, but the period may depend on the choice of the initial state. One finds that such a model may always be regarded as a combination of mechanical parts of the form of Fig. 9.1a. A generic quantum system may be much more complicated, see Fig. 9.1b. Nevertheless, as we elaborated in [10], one can approximate any generic quantum system in terms of combinations of the classical form Fig. 9.1a provided the latter be large enough. This may come as a surprise, and, at first sight, it may sound unbelievable. But then one must remember that today's experiments in physics have access only to a tiny subset of Hilbert space. There are huge numbers of states containing particles that may be heavier than all the ones presently known, or simply contain more

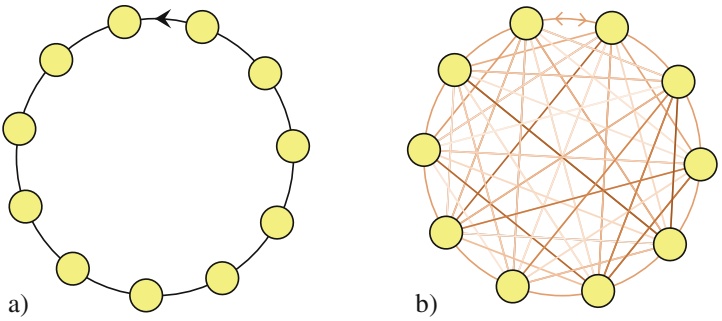


Fig. 9.1 (**a**) Essential part of a classical model. In a bounded universe the model must be periodic. Generic classical models are built by combining ingredients all of the same basic form. (**b**) Generic quantum system. Transitions consist of states that can evolve into superpositions of states. In the text it is explained how a generic quantum model can be obtained by treating (**b**) as an approximation to much more complicated classical models, even if these essentially only consist of parts of the form (**a**)

energy and momenta than one ever had the chance to observe. The basic argument is summarised as follows.

We consider a (quantum) system that allows for all sorts of transitions, as sketched in Fig. 9.1b. This means that the Hamiltonian maps every state onto any kind of superposition of states. Notably, this is assumed to be the case even if the basis elements of Hilbert space, the yellow circles in this figure, have been characterised exclusively in a local formalism. In the case of a single particle, this would be the position elements $|\mathbf{x}\rangle$, but in general we shall be more interested in quantised field theories, where we demand that the basis elements are characterised by the fields, $\{\phi(\mathbf{x})\}$ in position space, where ϕ stands symbolically for all types of (quantised) fields. Clearly, this would include the Standard Model, although we shall have to add states with constraints, invisible under ordinary circumstances, as will be made clear.

Our claim is now that, given any margin of accuracy, a classical model can be constructed that will reproduce all probability expressions of the above quantum system, provided that the initial state is one where the local observable is specified, which means that we exclude superimposed states for the initial state. One may then inspect all states that evolve from this initial state and check that our classical model generates the same probability distributions as the quantum system.

At first sight our constraint on the initial state may seem to be too restrictive, but the next assumption is that such a model may define a universe that started as a Big Bang, in which indeed the initial field values were not entangled. This assumption is all we need, to prove our theorem.

But, as the reader may have suspected, there is an other requirement in our construction: the classical model is assumed to include *fast fluctuating variables*. These are variables that evolve into different states within stretches of time short compared to $\hbar/E$, where E is the maximal energy that can be delivered to any particle in our quantum model. For example, one may think that, in accordance with most Grand Unified Field Theories, particles will exist with masses way beyond our allowed quantum energies E. Such particles will have to be omitted in our considerations. In practice this means that the fields of these fast fluctuating variables are in their *lowest energy stare*, that is, we only use their vacuum expectation values in our model, but we do not desire to include the actual particles, which are too energetic to be observable in any experiment.

The fast fluctuating variable is an extra ingredient of the form described in Fig. 9.1a, which at first sight is completely classical. The classical states evolve periodically, taking all N values on the circle within the indicated stretch of time. The period is $T = N\,\delta t$.

Central in our argument will be the concept of energy. The fact that the fast fluctuating variables move faster than we can detect, in practice means that they are assumed to be in an even distribution of all possible states on the circle. To describe this mode with all probabilities equal, we borrow the vector notation from standard quantum mechanics. If the periodicity is N steps, the evolution operator over time $T = N\delta t$, is the identity operator. Therefore the N eigen values of the evolution

operator $U(t)$ are $e^{-iE_n t}$, with $E_n = 2\pi n/T$, $T = N\,\delta t$. The integer n takes any value between 0 and N.

To describe the completely even state, we use the energy ground state, $E = 0$. The higher energy states, $n = 1, \cdots, N-1$, will hardly ever be excited, as in ordinary quantum mechanics. The available energies are too low.

Here we see the emergence of superimposed states. At the stage we are discussing now, superposition is a very simple consequence of the necessary smearing over a stretch of time equal to T. Now comes a crucial step. The fast fluctuating variables will interact, classically, with the variables that describe the (localised and observable) basis states of our model. The latter will be called 'slow' variables.

Details of this interaction are described in [10]. The interactions are classical, as are the basis states of our model, but we may use the quantum notation. Typically, they will take the shape of *exchanges of the slow states,* depending on the positions of the fast variables (one could hardly imagine other classical interactions). All interactions of our quantum fields $\phi(\mathbf{x})$ will be slow, and these interactions can be mimicked by considering a multiple of fast variables, which are now postulated to dictate exchanges among the slow states only rarely, depending on their precise positions in their respective cycles. They will mostly be found in positions where no interactions take place, so that the interactions that do take place are suppressed when we calculate them:

$$H_{\rm slow} = \langle E = 0\,|U^{\rm int}|\,E = 0\,\rangle\ , \tag{9.2}$$

where now we may use that the state $|E = 0\,\rangle$ for the fast moving objects have the effect of averaging out the fast—but rare—interactions $U^{\rm int}$. They are represented by the crossing lines in Fig. 9.1b.

Our calculation of the effective Hamiltonian for the slow variables in Eq. (9.2) is of the standard perturbative type. This is exactly how Grand Unified fields are handled traditionally; in practice, only the vacuum expectation values of the most massive fields are of any importance, and these expressions are subsequently handled in terms of perturbation expansions.

For further illuminating details we refer to [10]. The conclusion is that, we should consider classical models where some variables move very fast while dictating classical transitions among other, slow variables. The fast variables, in turn, are not affected much by the slow ones. As such they therefore violate Newton's action = reaction principle, and this is how we now explain the emergence of quantum mechanics.

Note that, in precise calculations, the flat distribution of the fast variable states will acquire low amplitude, wave-like correction terms. These are the *virtual* excited particles, just as they occur in Grand Unified Theories. Since the vector notation that we use is exactly valid, these excited states still behave as in ordinary quantum mechanics.

9.4 The Standard Model

At first sight, the considerations sketched above do not seem to have a great impact on the Standard Model of the subatomic particles and the way we handle it, but, to the contrary, the implications could be sizeable. If indeed our standard theories would be reconsidered as having their roots in a set of deterministic laws of nature, as the present author now believes, there would be at least two important changes in the way we should handle the theory.

First, local determinism would imply the existence of cellular automaton rules based on evolving data that take the shape of integers only. This would impose severe limitations on the possible constructions one can try to fathom, but the models that will emerge must be denumerable, or possibly finite in number. Consequently, each candidate theory one can construct would be based on relatively simple, rational numbers. In particular, coupling parameters would also not be allowed to range over all possible real numbers, but must be limited to rational numbers, probably limited to be within some finite intervals.

Quite a few authors have been speculating that the standard coupling parameters could run towards different values at different points in space-time. In our scheme, this would be really difficult, well-nigh impossible, to allow for. If one ever hits upon such features, our theory would be falsified.

Then there is a second observation, which involves the uses of perturbative expansions. The cellular automaton that we believe to underlie the quantum nature of the Standard Model may well allow for perturbative expansions, but there is no reason at all to expect such expansions to converge. The question about convergence of the perturbation expansion is often raised with respect to quantised field theories. Lack of convergence would seem to jeopardise the internal consistency of the model. Now, we emphasise that perturbative field theory itself is an approximation that is no longer needed for *defining* the basis of the theory. Perturbation expansions are as good as they go, but the theory itself will have to be defined in terms of the cellular automaton's (exact) evolution equations. From an axiomatic point of view this would be a great improvement of our understanding.

Unfortunately, there is an important impediment barring us from commencing to implicate our program, which is the gravitational force. At all distance scales where our understanding could be checked, we are confronted with diffeomorphism invariance (general relativity), so one of the first things to do now is to pave our way through that principle. Reconciling cellular automaton rules with diffeomorphism invariance is notoriously difficult. The present work does relieve us from the requirement to do this in terms of Hilbert space, but the cellular automaton must be based on some type of lattice. Diffeomorphism transformations could turn the lattice into a kind of glassy structure. How this should be phrased is not clear at all.

Our philosophy is to advance by one step at the time. The question that must be addressed now is what cellular automaton rules could even approximately lead to the particle spectrum and the interaction algebra known today. Let us start with the symmetries, observed to be exact or approximate features of the Standard Model.

Well-known today is the local symmetry based on $U(1) \otimes SU(2) \otimes SU(3)$. These are continuous groups, but one can try first to reduce these to a subgroup based on integers. As stated above, integer numbers are mandatory in our theory—we should not drop them lightly. It may well be that we have to replace the usual local symmetry group by a unitary transformation group based on integers. This can be done on any lattice. Vector fields, based on integers, can be defined on the links of the lattice. So far, so good.[2]

We have fermionic fields in the Standard Model. In discrete theories, these can be very well represented by Boolean variables $\theta_i(\mathbf{x}, t) \in \{0, 1\}$. Anti-commuting operators are easy to define there, so we expect Dirac fields to emerge from this, but we do emphasise that not only the Yukawa interaction terms, but also the kinetic ingredients of the Dirac equations can only be ascribed to interactions with fast variables. Presumably we have Boolean fast variables there; this is certainly also allowed.

There is one very remarkable feature of the Standard Model, which is a relatively light Higgs boson (125.3 GeV). In our usual description in terms of quantised fields, this light Higgs particle presents a problem: *renormalisation counter terms play an important role in the theory, and, if extrapolated to the energy scale where the theory must be united with gravity* (the Planck scale), *these terms are tremendously larger than this mass term, which seems to emerge as balancing in a quite unnatural way.* It is now believed that nature seems to be telling us something important here [13].

In terms of the cellular automaton, this 'unnatural' feature can be rephrased in a meaningful way. Consider the field ϕ of a scalar particle—such as the Higgs boson. At the Planck scale, let us consider a translation of the type

$$\phi(\mathbf{x}, t) \longrightarrow \phi(\mathbf{x}, t) + C^{\text{nst}} . \tag{9.3}$$

Here, ϕ is the Higgs field, and also the constant C^{nst} stands for a complex, two component $SU(2)$ spinor. We now consider the effect this transformation has on the Hamiltonian of the system. The Higgs mass is extremely small at the Planck scale. Furthermore, the effective 4-field self interaction of the Higgs field is small there as well; this is the surprising result of some renormalization group calculations [14]. Suppose that, in a sense, this interaction can also be considered to be zero. Then, we find that *the Hamiltonian H of the theory is almost invariant under the transformation* (9.3).

As is well-known, this is Goldstone's principle: the 'massless' Higgs particle is the Goldstone particle of a nearly perfect global symmetry.[3] There is no associated *local* symmetry, other than the usual local gauge symmetry $SU(2) \otimes U(1)$, because that would be broken by the kinetic term in the Higgs action. In terms of our cellular

[2] Alternatively, one could suspect that the vector representation, introduced above to represent emergent quantum mechanics, might also enhance the symmetries to become continuous transformation groups. The author has not made up his mind as to which route to take here.

[3] Note that also the tiny effects of the W and Z mass terms violate the Goldstone symmetry.

automaton interpretation, only the exponentiated Hamiltonian, in the form of the evolution operator, is to be used to describe the evolution. We see that the evolution operator is (almost) invariant under transformations of the type (9.3).

This invariance is quite non-trivial. It is not exact, but deviations will be of the order of 10^{-34} or so. This could be an indication that large integers enter in the interactions, integers that play a decisive role in violating this pseudo-Goldstone invariance [16] at such a small rate. One might suspect that, without this approximate global symmetry, not only the Higgs would have been unobservable, but, through the Yukawa interactions, also all leptons and quarks would have masses at the Planck domain.

Eventually, all known matter in the universe may owe its existence to this very tiny global symmetry violation. This is our way of phrasing what Jegerlehner [13] and Veltman [15] have been talking about.

The fact that, when compared to the Planck scale, the leptons and quarks are very light as well, may be due to similar, almost perfectly invariant, transformation rules. As we are dealing with fermions here, the Goldstone fields here will have to be fermionic as well. This may imply that, at the Planck scale, we expect to have very weakly broken supersymmetries, one supersymmetry generator for every fermionic field component in the Standard Model. This symmetry differs from conventional supersymmetry in the sense that it is a global supersymmetry, and the fact that it is not a perfect invariance, giving masses to the particles.

One complication can easily be dealt with: the constant in the transformation rule (9.3) is not invariant unde the local gauge transformations of the Standard Model; this is taken care of by splitting the Higgs field variable into a unimodular part, u^i and the modulus r:

$$\phi^i(x) = r(x)\, u^i(x)\ , \tag{9.4}$$

so that the Goldstone symmetry operators only act on r and the local $SU(2)$ symmetry only acts on u^i; this ensures that the Goldstone symmetry commutes with the local $SU(2)$ transformations.

In terms of cellular automaton interactions, the fermionic global supersymmetries will be manifest as Boolean symmetries, symmetries that square into the identity operation; by itself, this feature would not be problematic.

We are still at a pioneering stage of our general insights. What is missing is the general equivalents of the Hamilton and Lagrange formalisms, which have been auspicious for understanding quantised systems and their relations with the classical limit, $\hbar \rangle 0$. We now need to know how this works when special types of quantum systems are mapped to deterministic classical systems.

References

1. C.N. Yang, R.L. Mills, Conservation of isotopic spin and isotopic gauge invariance. Phys. Rev. **96**, 191 (1954); see also: R. Shaw, Cambridge Ph.D. Thesis, unpublished
2. F. Englert, R. Brout, Broken symmetry and the mass of gauge vector mesons. Phys. Rev. Lett. **13**, 321 (1964)
3. P.W. Higgs, Broken symmetries, massless particles and gauge fields. Phys. Lett. **12**, 132 (1964); id. Broken symmetries and the masses of gauge bosons. Phys. Rev. Lett. **13**, 508 (1964); id., Spontaneous symmetry breakdown without massless bosons. Phys. Rev. **145**, 1156 (1966)
4. G.S. Guralnik, C.R. Hagen, T.W.B. Kibble, Global conservation laws and massless particles. Phys. Rev. Lett. **13**, 585–587 (1964). http://link.aps.org/abstract/PRL/v13/p585. https://doi.org/10.1103/PhysRevLett.13.585
5. C.G. Callan, Broken scale invariance in scalar field theory. Phys. Rev. **D2**, 1541 (1970); K. Symanzik, Small distance behaviour and power counting. Commun. Math. Phys. **18**, 227 (1970); G. 't Hooft, Dimensional regularisation and the renormalization group. Nucl. Phys. **B61**, 455 (1973)
6. J.S. Bell, On the Einstein Podolsky Rosen paradox. Physics **1**, 195 (1964)
7. G. 't Hooft, *The Cellular Automaton Interpretation of Quantum Mechanics*. Fundamental Theories of Physics, vol. 185 (Springer International Publishing, 2016). eBook ISBN:978-3-319-41285-6. https://doi.org/10.1007/978-3-319-41285-6. Hardcover ISBN:978-3-319-41284-9. Series ISSN:0168-1222, Edition Number **1**, http://arxiv.org/abs/1405.1548
8. C. Wetterich, Probabilistic cellular automata for interacting fermionic quantum field theories. Nuclear Physics **B 963**, 115296 (2021). arXiv:2007.06366 [quant-ph]; id., Fermionic quantum field theories as probabilistic cellular automata. arXiv:2111.06728v1 [hep-lat] (2021)
9. S. Hossenfelder, T.N. Palmer, Rethinking superdeterminism. Front. Phys. **8**, 139 (2020). e-Print:1912.06462 [quant-ph]; S. Hossenfelder, Superdeterminism: A guide for the perplexed. arXiv:2010.01324v2 [quant-ph] 6 Oct 2020
10. G. 't Hooft, Explicit construction of Local Hidden Variables for any quantum theory up to any desired accuracy. arxiv:2103.04335[quant-ph], to be published
11. M.B. Green, J.H. Schwarz, E. Witten, *Superstring Theory* (Cambridge Univ. Press, 1977)
12. G. 't Hooft, *Discreteness and Determinism in Superstrings*. ITP-UU-12/25; SPIN-12/23, arXiv:1207.3612v2[hep-th]
13. F. Jegerlehner, The standard model of particle physics as a conspiracy theory and the possible role of the Higgs boson in the evolution of the Early Universe. Acta Phys. Pol. **B 52**, 575 (2021)
14. F. Bezrukov, M.Y. Kalmykov, B.A. Kniehl, M. Shaposhnikov, Higgs boson mass and new physics. J. High Energy Phys. **1210**, 140 (2012). arXiv:1205.2893v2
15. M. Veltman, The Higgs and the LHC. CERN Cour. **48N9**, 28 (2008)
16. J. Goldstone, Field theories with superconductor solutions. Nuovo Cim. **19**, 154 (1961); Y. Nambu, G. Jona-Lasinio, Dynamical model of elementary particles based on an analogy with superconductivity. 1. Phys. Rev. **122**, 345 (1961)

Chapter 10
Gauge Symmetry in Shape Dynamics



Frank Wilczek

Abstract C. N. Yang's ideas about local gauge symmetry and non-integrable phases have been enormously fertile sources of inspiration in fundamental physics and in the quantum theory of matter. They also arise naturally in describing the dynamics of deformable bodies. Here I extend previous discussions of the gauge symmetry of deformable bodies in several directions, bringing in an arbitrary number of dimensions, general time-dependence, conservation laws and adiabatic residuals. I briefly indicate other potential applications of the conceptual framework.

Every theoretical physicist of my generation is a student of C. N. Yang. His insights, notably including the clusters of ideas around parity violation [1] and local gauge symmetry [2], permeate modern fundamental physics. Beyond that, his career as as a whole epitomizes the synergy of truth and beauty that is the central miracle of physics [3]. In this, as in the equations that bear his name, he is the worthy modern heir of James Clerk Maxwell.

10.1 Gauge Structure: Fundamental, Emergent, Productive

The principle of local gauge symmetry is central to our present theories of fundamental interactions. It has been a guide to discovery, not only through its constructive role in explaining the existence and properties of force-mediating particles—photons, W and Z bosons, color gluons, and gravitons—as avatars of curvature, but also through the powerful constraints it imposes on possible

F. Wilczek (✉)
Center for Theoretical Physics, MIT, Cambridge, MA, USA

T. D. Lee Institute and Wilczek Quantum Center, Shanghai Jiao Tong University, Shanghai, China
Arizona State University, Tempe, AZ, USA

Stockholm University , Stockholm, Sweden
e-mail: wilczek@mit.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_10

interactions among fundamental fields. Yet I think it is fair to say that gauge symmetry appears in relativistic field theory as a rather mysterious and even elusive gift. For example, we are led to the equations of quantum chromodynamics by imposing local $SU(3)$ color gauge symmetry, but the observables of the theory are supposed to be gauge singlets, as are the asymptotic states (confinement). The underlying gauge symmetry thereby vanishes from view, leaving behind a smile drawn in evanescent jets.

In recent years gauge symmetry has been identified in several other physical contexts, wherein it is not an irreducible axiom but rather an emergent consequence of other principles. These include the circle of ideas around Berry's phase in basic quantum mechanics [4–6], now developed to support many sophisticated applications including topological band theory [7]; the statistical gauge fields of anyon theory [8–10], recently exhibited directly in the quantum Hall effect [11, 12] and being prepared for use in quantum information processing [13, 14]; and in the low-energy theory of quantum spin ice [15], among others. This naturally raises the question whether gauge symmetries presently regarded as fundamental might emerge from more primitive structures, too. In any case, such examples bring the aforementioned mystery: Why gauge symmetry? down to earth.

Identification of gauge structure in a new problem is productive, because it brings in powerful conceptual tools—e.g., covariant derivative, curvature, holonomy, topology—that have been sharpened by decades of development and application in physics. Here, the synergy of beauty and truth is tangible.

In this note I will discuss another example of emergent gauge symmetry, in the dynamics of deformable bodies [16–19]. We would like to calculate how the internal motion of a body affects its external position and orientation. But if the shape of a body changes, its change in position is not unambiguously defined, since different parts move by different amounts. To overcome this problem, we associate with each shape a corresponding "reference shape" that has a definite position and orientation. The position and orientation of the body relative to its reference shape is then well-defined, and we can calculate it unambiguously. Of course, the introduction of reference shapes introduces an element of convention into this procedure, because we might have chosen them differently. As we will see, this element of convention is precisely a gauge structure. Moreover the associated gauge field is highly non-trivial, and it appears prominently in the dynamical equation.

10.2 Dynamical Equation for Deformable Bodies

10.2.1 Referenced Angular Momentum

We consider an assembly of point masses m^j whose configuration x^j can change with time. For each configuration we have a reference shape, as described immediately above. We define the time-dependent rotation R that relates the configuration

of the assembly to the corresponding reference shape configuration according to

$$x^j \;=\; Rs^j \tag{10.1}$$

Then we have a dynamical equation that relate change of angular momentum L^α to external torque τ^α according to

$$\tau^\alpha \;=\; \frac{dL^\alpha}{dt} \;=\; \epsilon_{\alpha\beta\gamma} \sum_j m^j \, R_{\beta\gamma} s^j_\rho \, \frac{d^2}{dt^2} \, (R_{\rho\sigma} s^j_\sigma) \tag{10.2}$$

Our goal is to work Eq. (10.2) into a form that contains fewer variables and brings out its gauge symmetry. Thus, in a broad sense, we will construct an effective theory of rotational motion that displays an emergent gauge symmetry. We will first do this in a way that is valid in any number of dimensions (see e.g., [20]) and emphasizes conceptual structure. Then we will apply some algebraic tricks special to three dimensions, to make contact with more conventional presentations. In the Appendix, as an exercise and sanity check, we get to the same dynamical equation by brute force algebra.

10.2.2 Inertia Tensor and Angular Motion

The primary object of interest, written now in a form valid in any number of dimensions, is the angular momentum

$$L_{\alpha\beta} \;\equiv\; \sum_j m^j \, (x^j_\alpha \, \dot{x}^j_\beta - x^j_\beta \, \dot{x}^j_\alpha) \tag{10.3}$$

Bringing in the reference shapes using Eq. (10.1), we are led define two shape-based contributions to L according to

$$\begin{aligned} R^{-1}_{\alpha\rho} R^{-1}_{\beta\sigma} \, L_{\rho\sigma} \;&=\; \sum_j m^j \, (s^j_\alpha \, (R^{-1} \dot{R})_{\beta\gamma} \, s^j_\gamma \;-\; s^j_\beta \, (R^{-1} \dot{R})_{\alpha\gamma} \, s^j_\gamma) \\ &+\; \sum_j m^j \, (s^j_\alpha \, \dot{s}^j_\beta \;-\; s^j_\beta \, \dot{s}^j_\alpha) \\ &\equiv\; M^S_{\alpha\beta} \;+\; L^S_{\alpha\beta} \end{aligned} \tag{10.4}$$

Here and throughout we use the superscript ^{S} to indicate objects defined using shape variables. These allow us to extend the idea of a body-fixed frame to deformable bodies.

Now let us analyze and simplify M^S. Bringing in the angular velocity

$$R^{-1}\dot{R} \equiv \omega \tag{10.5}$$

we have

$$M^S_{\alpha\beta} = \sum_j m^j \left(s^j_\alpha \omega_{\beta\gamma} s^j_\gamma - s^j_\beta \omega_{\alpha\gamma} s^j_\gamma\right) = I^S_{\alpha\beta;\gamma\eta}\omega_{\gamma\eta} \tag{10.6}$$

where

$$I^S_{\alpha\beta;\gamma\eta} \equiv \frac{1}{2}\left(\widetilde{I}^S_{\alpha\gamma}\delta_{\beta\eta} - \widetilde{I}^S_{\alpha\eta}\delta_{\beta\gamma} - \widetilde{I}^S_{\beta\gamma}\delta_{\alpha\eta} + \widetilde{I}^S_{\beta\eta}\delta_{\alpha\gamma}\right) \tag{10.7}$$

with

$$\widetilde{I}^S_{\alpha\beta} \equiv \sum_j m^j s^j_\alpha s^j_\beta \tag{10.8}$$

defines the inertia tensor.

We can assign simple labels A the antisymmetric pairings (α, β). For $D = 3$ these labels take three values, for $D = 4$ six values, and so forth. In terms of those variables $I_{A;B}$ is a symmetric matrix, ω_B is a vector, and Eq. (10.6) becomes

$$M^S_A = I^S_{A;B}\omega_B \tag{10.9}$$

or simply

$$M^S = I^S\omega \tag{10.10}$$

For our purposes we also need to know that I^S, thus defined, is generally invertible. To see that, consider the expectation value of I^S relative to a "vector" ζ. We find

$$\zeta_{\alpha\beta} I^S_{\alpha\beta;\gamma\eta}\zeta_{\gamma\delta} = 2\sum_j m^j(\zeta_{\rho\sigma}s^j_\sigma)(\zeta_{\rho\tau}s^j_\tau) \tag{10.11}$$

This is a sum of non-negative terms, each of which only vanishes when s^j lies in the subspace $\zeta_{\rho\sigma}s^j_\sigma = 0$. For assemblies of s^j in general position the expectation value is positive, and in that case $I_{a;b}$ is a symmetric positive matrix. This implies that it is not only invertible, but even diagonalizable with positive entries. (Note however that the diagonalization process involves rotations in the "index space". For $D > 3$ these might not correspond to rotations in the ambient D-dimensional space.) Expressions related to Eq. (10.11) will occur in our discussion of conservation laws, below, where they represent quantities that are generically positive on physical grounds.

10.2.3 Gauge Symmetry and Gauge Field

L^S has the form of an angular momentum in shape space. To bring out its physical significance, however, we need to transcend the element of convention (gauge symmetry!) that enters into its construction. We have the freedom to modify our choice of reference shapes according to

$$s \;\rightarrow\; U(s)^{-1}\, s \tag{10.12}$$

$$R \;\rightarrow\; R\, U(s) \tag{10.13}$$

This choice does not affect the angular momentum L (with no superscript), but it does modify everything else in Eq. (10.4). R and I^S change simply homogeneously, but ω, M^S and L^S are tricker. They bring in inhomogeneous terms, that must cancel between M^S and L^S.

Indeed, we have

$$\omega \;\rightarrow\; U^{-1}\omega\, U \;-\; U^{-1}\,\dot{U} \tag{10.14}$$

The required cancellation therefore implies that

$$\Omega dt \;\equiv\; (I^S)^{-1}\, L^S\, dt \tag{10.15}$$

transforms as an $so(3)$ gauge potential field on shape space [21]. (Of course, its transformation law can also be verified directly.) To bring out the fact that U is a function on shape space, we can introduce co-ordinates λ^κ on that space, and write

$$s(t) \;\equiv\; s\big(\lambda^\kappa(t)\big) \tag{10.16}$$

$$\dot{U} \;=\; \frac{\partial U\big(s(\lambda)\big)}{\partial \lambda^\kappa}\,\frac{d\lambda^\kappa}{dt} \tag{10.17}$$

so that

$$\omega dt \;\rightarrow\; U^{-1}\,\omega dt\, U \;-\; U^{-1}\,\frac{\partial U\big(s(\lambda)\big)}{\partial \lambda^\kappa}\, d\lambda^\kappa \tag{10.18}$$

This gauge potential naturally appears as a one-form. To make contact with conventional physics notation, we would write out components

$$\Omega\, dt \;\equiv\; \Omega_\kappa\, d\lambda^\kappa \tag{10.19}$$

This gauge structure brings in the m^j as parameters, but it is dimensionless and purely a function of the shape space geometry. It identifies the effective contribution

to the angular velocity due to shape change, as opposed to collective motion. The separation between those two contributions brings in an element of convention, i.e. a choice of gauge, but the total is free of convention, i.e. gauge covariant.

Knowing that $\Omega\, dt$ is a gauge potential, we can use the techniques of gauge theory to bring out its gauge-invariant content. We can calculate field strengths (curvature), Wilson loops (holonomy), and so forth. In the present context, this allows us to identify unambiguous, quantitative dynamical consequences of shape deformation.

Using Ω we can write

$$R^{-1}_{\alpha\mu} R^{-1}_{\beta\nu} L_{\mu\nu} \;=\; I^S_{\alpha\beta;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu}) \tag{10.20}$$

or more compactly

$$L = \mathcal{R}\, I^S(\omega + \Omega) \tag{10.21}$$

10.2.4 Dynamical Equation

It is straightforward to pass from L to the dynamical equation

$$\begin{aligned}
\frac{dL_{\alpha\beta}}{dt} \;&=\; \frac{d}{dt} R_{\alpha\rho} R_{\beta\sigma} I^S_{\rho,\sigma;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu}) \\
&=\; R_{\alpha\rho} R_{\beta\sigma} \Big(\frac{d}{dt} I^S_{\rho,\sigma;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu}) \\
&+\; \omega_{\rho\kappa} I^S_{\kappa,\sigma;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu}) \;+\; \omega_{\sigma\kappa} I^S_{\rho,\kappa;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu}) \Big) \\
&\equiv\; R_{\alpha\rho} R_{\beta\sigma} \frac{D}{Dt} I^S_{\rho,\sigma;\mu\nu}(\omega_{\mu\nu} + \Omega_{\mu\nu})
\end{aligned} \tag{10.22}$$

The main virtue of bringing R factors outside the time derivative on the right-hand side, as in the third line above, is that when the left-hand side vanishes we can simply strip them away.

We can write Eq. (10.22) more compactly as

$$\mathcal{R}^{-1} \frac{dL}{dt} \;=\; \frac{D}{Dt} I^S(\omega + \Omega) \tag{10.23}$$

For some applications it might also be useful to go to a rotating frame, where

$$x^j_{\text{rot.}}(t) \;=\; V(t) x^j(t) \tag{10.24}$$

motivates use of the variable

$$R_{\text{rot.}}(t) \equiv V(t)\, R(t)\, V^{-1}(t) \tag{10.25}$$

so that

$$V\, R\, x^j = R_{\text{rot.}}\, x^j_{\text{rot.}} \tag{10.26}$$

It is not difficult to translate the dynamical equations into this sort of frame, eliminating $R(t)$ in favor of $V(t)$ and $R_{\text{rot.}}$.

10.2.5 Three Dimensional Notation

In three dimensions a special "vector" notation for the A indices is possible (and standard). It relies on use of the invariant $\epsilon_{a\mu\nu}$ symbol to eliminate antisymmetric pairs of indices in favor of single indices. Here I will simply record the definitions and the main results. Let me remark that premature resort to this notation can tend to obscure the conceptual structure exposed above.

With

$$I^S_{ab} \equiv \frac{1}{2}\epsilon_{a\alpha\beta}\epsilon_{b\rho\sigma} I^S_{\alpha\beta;\,\rho\sigma} = \sum_j m^j\, (s^{j\,2}\delta_{ab} - s^j_a s^j_b) \tag{10.27}$$

$$\omega_a \equiv \frac{1}{2}\epsilon_{a\alpha\beta}\omega_{\alpha\beta} \tag{10.28}$$

$$\Omega_a \equiv \frac{1}{2}\epsilon_{a\alpha\beta}\Omega_{\alpha\beta} \tag{10.29}$$

we have

$$R^{-1}_{ab} L^b = I^S_{ac}(\omega_c + \Omega_c) \tag{10.30}$$

and in an evident notation (suppressing indices)

$$R^{-1} L = I^S(\omega + \Omega) \tag{10.31}$$

$$R^{-1}\frac{dL}{dt} = \frac{D}{Dt} I^S(\omega + \Omega) \tag{10.32}$$

with

$$\frac{D}{Dt} = \frac{d}{dt} + \omega\times \tag{10.33}$$

10.2.6 Specializations

Falling cats or divers can face the challenge of re-orienting their bodies in space despite starting from rest (zero angular momentum) and with no access to external torques. They can address that challenge by changing their shape. By doing so, they induce $L^S \neq 0$ and in consequence $\Omega \neq 0$. From

$$R^{-1}L \;=\; I^S(\omega + \Omega) \;=\; 0 \tag{10.34}$$

we deduce

$$\omega dt \;=\; -\,\Omega dt \;=\; -\,(I^S)^{-1}L^S dt \tag{10.35}$$

Here we see that motion in physical space reflects the gauge potential in shape space quite directly [16]. Indeed, the (ordered) time integral of angular velocity, which gives the referenced rotation in physical space, is equal in magnitude to the integral of the geometric potential, expressed as a one-form, in shape space.

Several examples illustrating this physical phenomenon have been analyzed in detail [16–19, 22–26]. These examples establish the non-triviality of the shape space gauge structure, even for very simple families of shapes. An interesting feature is the occurrence of topologically non-trivial gauge structures, including vortices and magnetic monopoles. These can occur when the allowed shape space is topologically non-trivial. Indeed, when they occur they indicate that the shape space cannot be augmented to become trivial without encountering configurations where I^S is singular (that is, non-invertible). It should be possible to analyze large classes of these singularities systematically.

It could be interesting, also, to analyze cases where the shape space gauge structure plays a dominant but not exclusive role in the dynamics, with L and $\frac{dL}{dt}$ introduced as small parameters or noise.

At the other extreme, for a rigid body we can fix the standard shape once and for all. In this case $L^S = 0$, and Eq. (10.32) reduces to the classic Euler equations for rigid body motion.

In some circumstances a specific choice of gauge suggests itself. When we have a stable ordering of the eigenvalues I_1, I_2, I_3 of the inertia tensor, we can choose the eigen-directions to be oriented along the $\widehat{x}, \widehat{y}, \widehat{z}$ directions for the corresponding reference shapes. When the eigenvalues cross, of course, we will need to patch together complementary gauge choices. Alternatively, if we are especially interested in small deformations of a specific shape, we can choose to make the gauge potential and its "radial" components along a choice of spokes emanating from that point vanish, in the style of Gaussian normal co-ordinates. (See also below.)

10.2.7 Angular Momentum and Energy

For vanishing external torques, the first line of the dynamical equation Eq. (10.22) allows us to infer, from $L_{\alpha\beta}\frac{dL_{\alpha\beta}}{dt} = 0$, the conservation law

$$(M^S_{\alpha\beta}+L^S_{\alpha\beta})(M^S_{\alpha\beta}+L^S_{\alpha\beta}) \;=\; (\omega+\Omega)_A I^S_{A;B} I^S_{B;C}(\omega+\Omega)_C \;=\; \text{constant} \qquad (10.36)$$

This is the covariant completion of the familiar conservation of $(I\omega)^2$ for untorqued rigid bodies.

The absence of external torques does not imply energy conservation, however. Indeed, in the envisaged applications to cats, divers, autonomous systems and micro-machines shape changes might be accomplished by exerting or dissipating work. It is interesting, nevertheless, to consider how the kinetic energy can be expressed.

First let us appreciate the difficulty. The kinetic energy associated with motion according to Eq. (10.1) is

$$E_{\text{kin}} = \frac{1}{2}\sum_j m^j \dot{x}^j \cdot \dot{x}^j = \frac{1}{2}\sum_j m^j \left(\omega_{\alpha\beta} s^j_\beta \omega_{\alpha\gamma} s^j_\gamma + \omega_{\alpha\beta}(s^j_\alpha \dot{s}^j_\beta - s^j_\beta \dot{s}^j_\alpha) + \dot{s}^j \cdot \dot{s}^j\right) \qquad (10.37)$$

Here the first two terms on the right-hand side match the corresponding terms of the expansion of the covariant completion of the standard expression for rigid bodies

$$\frac{1}{2} I^S_{A;B}(\omega_A+\Omega_A)(\omega_B+\Omega_B) \;=\; \frac{1}{2} I_{A;B}\omega_A\omega_B + \omega_A L^S_A + \frac{1}{2} I_{A;B}\Omega_A\Omega_B \qquad (10.38)$$

but the third terms are quite different.

We can choose fix our gauge so that $\Omega = 0$ at a given shape. *In that gauge* we have a simple separation of the kinetic energy *at that shape* into spatial motion and shape space contributions:

$$\begin{aligned} E_{\text{kin}} \;&=\; \frac{1}{2} I_{A;B}\omega_A\omega_B \;+\; E^S_{\text{kin}} \\ E^S_{\text{kin}} \;&\equiv\; \frac{1}{2}\sum_j m^j \dot{s}^j \cdot \dot{s}^j \end{aligned} \qquad (10.39)$$

This separation cannot be achieved globally, however. Gauge curvature obstructs a clean separation between rotational and deformational energy.

Since the discrepant terms are of second order in the rate of shape change, in the limit of slow shape changes we can neglect them. The term that is linear in $\dot{s}$ gives a residual correction in the adiabatic limit. Indeed, if we scale the rate of deformation by writing $t = T\tau$ inside the argument of $s(t)$, while holding its functional form

fixed, and integrate between fixed shapes $s(a)$, $s(b)$, then we have schematically

$$\int_{s(a)}^{s(b)} dt\dot{s}(t) \sim \int_{s(a)}^{s(b)} ds \tag{10.40}$$

but

$$\int_{s(a)}^{s(b)} dt\dot{s}(t)^2 \sim \int_{s(a)}^{s(b)} T d\tau \frac{1}{T^2}(\frac{ds}{d\tau})^2 \propto \frac{1}{T} \to 0 \text{ as } T \to \infty \tag{10.41}$$

10.3 Extensions

10.3.1 Blobs, Media, and Swarms

We have formulated our equations with reference to systems of particles, but of course the method carries over to continuous mass distributions as a limit. Here it is significant that our effective description needs only the inertia tensor and the shape-space angular momentum (and, for energy, the shape-space energy) as input, and is otherwise independent of the details of deformation.

The framework of reference shapes and gauge symmetry, associated above with rotations of a body in space, can be carried over to problems involving the displacement of a deformable body in a medium. The appearance of gauge structure in connection with separation of infinitesimal angular velocity (and simple velocity) into spatial and deformation pieces is quite general, though its dynamical salience will vary from case to case. It plays a central role in the description of self-propulsion at low Reynolds number [27] and in biological locomotion [28].

When one has many independently deformable bodies, each will have its own "internal" gauge potential. The dynamics of a swarm will bring in interactions that couple the gauge structures, and in some circumstances might energetically favor small space-time gradations of shape. Then we could have an emergent Yang-Mills description of the collective motion.

10.3.2 Molecules and Nuclei

Many features in the spectra of molecules and of nuclei can be interpreted in terms of motions that combine rotation and deformation [29]. When the deformations can be considered as small vibrations about a definite shape, we can choose the gauge potential at that shape to vanish. This is accomplished by introduction of "Eckart

co-ordinates" [30], which separate the motions approximately. Effects of curvature cannot be made to vanish altogether, however. Informed use of gauge theory ideas should allow more systematic and accurate treatment of this and related problems.

Acknowledgments I am happy to thank Antti Niemi and Al Shapere for stimulating discussions. This work is supported by the U.S. Department of Energy under grant Contract Number DE-SC0012567, by the European Research Council under grant 742104, and by the Swedish Research Council under Contract No. 335-2014-7424.

Appendix

Direct Calculation

We will now work Eq. (10.2) into its reduced form more straightforwardly, directly in three dimensions. We express the second time derivative $\frac{d^2}{dt^2}\,(R_{\gamma\sigma}s^j_\sigma)$ as a sum of three terms and process each of those terms into an appropriate form. We will often use the property

$$\epsilon_{\alpha\beta\gamma}R_{\beta\rho}R_{\gamma\sigma} \;=\; \epsilon_{\tau\rho\sigma}R^{-1}_{\tau\alpha} \;=\; \epsilon_{\tau\rho\sigma}R_{\alpha\tau} \tag{10.42}$$

of rotations. It expresses that they have unit determinant, and also expresses the vector character of the vector cross product.

The first term simplifies as

$$\mathrm{I} \;=\; \epsilon_{\alpha\beta\gamma}\sum_j m^j R_{\beta\rho}R_{\gamma\sigma}s^j_\rho\ddot{s}^j_\sigma \;=\; R^{-1}_{\kappa\alpha}\,\dot{L}^S_\kappa \tag{10.43}$$

where

$$L^S_\kappa \;\equiv\; \epsilon_{\kappa\rho\sigma}\sum_j m^j s^j_\rho\,\dot{s}^j_\sigma \tag{10.44}$$

has the form of angular momentum in shape space. It seems appropriate to call it geometric angular momentum, since it relates to motion in shape space rather than directly in physical space.

The second term, similarly treated, becomes

$$\mathrm{II} \;=\; 2\,\epsilon_{\alpha\beta\gamma}R_{\beta\rho}\dot{R}_{\gamma\sigma}\sum_j m^j s^j_\rho\dot{s}^j_\sigma \tag{10.45}$$

$$\;=\; 2\,\epsilon_{\kappa\rho\delta}R^{-1}_{\kappa\alpha}(R^{-1}\dot{R})_{\delta\sigma}\sum_j m^j s^j_\rho\dot{s}^j_\sigma \tag{10.46}$$

To simplify this further, we separate the parts that are symmetric and antisymmetric in ρ, σ. For the symmetric part we have

$$\frac{1}{2}\sum_j m^j (s^j_\rho \dot{s}^j_\sigma \; + \; \dot{s}^j_\rho s^j_\sigma) \; = \; \frac{1}{2}\frac{d}{dt}\,(\sum_j m^j \, s^j_\rho s^j_\sigma) \tag{10.47}$$

$$\equiv \frac{1}{2}\frac{d}{dt}\widetilde{I}^S_{\rho\sigma} \tag{10.48}$$

For the antisymmetric part we have

$$\frac{1}{2}\sum_j m^j (s^j_\rho \dot{s}^j_\sigma \; - \; \dot{s}^j_\rho s^j_\sigma) \; = \; \frac{1}{2}(\epsilon_{\mu\rho\sigma}\epsilon_{\mu\lambda\eta}) \sum_j m^j s^j_\lambda \dot{s}^j_\eta \tag{10.49}$$

$$\equiv \frac{1}{2}\epsilon_{\mu\rho\sigma}\, L^S_\mu \tag{10.50}$$

Collecting terms, we have

$$\text{II} \; = \; (\frac{d}{dt}\widetilde{I}^S_{\rho\sigma} \; + \; \epsilon_{\mu\rho\sigma} L^S_\mu)\, \epsilon_{\kappa\rho\delta}\, R^{-1}_{\kappa\alpha}\, (R^{-1}\dot{R})_{\delta\sigma} \tag{10.51}$$

To reduce the redundancy of $R^{-1}\dot{R}$, which is automatically antisymmetric, we introduce

$$(R^{-1}\dot{R})_{\delta\sigma} \; \equiv \; -\,\epsilon_{\lambda\delta\sigma}\,\omega_\lambda \; = \; -\,(\dot{R^{-1}}R)_{\delta\sigma} \tag{10.52}$$

(The conventional $-$ sign will prove convenient later, and the second equation here shows that it is related to whether we regard R or R^{-1} as primary.) Inserting this into Eq. (10.51), after some ϵ-gymnastics we find

$$\text{II} \; = \; (\frac{d}{dt}I^S_{\rho\sigma} - \epsilon_{\mu\rho\sigma} L^S_\mu)\, R^{-1}_{\sigma\alpha}\omega_\rho \tag{10.53}$$

$$= \; R^{-1}_{\sigma\alpha}\,\big(\frac{d}{dt}I^S_{\rho\sigma}\;\omega_\rho \; + \; (\omega \times L^S)_\sigma\big) \tag{10.54}$$

where

$$I^S_{\alpha\beta} \; = \; \delta_{\alpha\beta}\widetilde{I}^S_{\eta\eta} - \widetilde{I}^S_{\alpha\beta} \tag{10.55}$$

is the inertia tensor of the standard shape.

The third term, similarly treated, becomes

$$\text{III} \; = \; \epsilon_{\alpha\beta\gamma}\, R_{\beta\rho}\, \ddot{R}_{\gamma\sigma}\, \widetilde{I}^S_{\rho\sigma} \tag{10.56}$$

$$= \; \epsilon_{\kappa\rho\mu}\, R^{-1}_{\kappa\alpha}\, (R^{-1}\ddot{R})_{\mu\sigma}\widetilde{I}^S_{\rho\sigma} \tag{10.57}$$

To reduce the redundancy now we use

$$R^{-1}\ddot{R} \;=\; (R^{-1}\dot{R})^{\cdot} + (R^{-1}\dot{R})^2 \tag{10.58}$$

together with Eq. (10.52) and some ϵ-gymnastics including

$$(R^{-1}\dot{R})^2_{\mu\sigma} \;=\; \epsilon_{\mu\tau\lambda}\epsilon_{\tau\sigma\eta}\omega_\lambda\omega_\eta \;=\; \omega_\mu\omega_\sigma - \omega^2\delta_{\mu\sigma} \tag{10.59}$$

to arrive at

$$\mathrm{III} \;=\; R^{-1}_{\kappa\alpha}\,(I^S_{\lambda\kappa}\,\dot{\omega}_\lambda \;+\; \epsilon_{\kappa\mu\rho}\,\omega_\mu\,I^S_{\rho\sigma}\omega_\sigma) \tag{10.60}$$

Assembling Eqs. (10.43, 10.54, 10.60) and multiplying by R^{-1}, we arrive at

$$R^{-1}_{\beta\alpha}\frac{dL_\alpha}{dt} \;=\; \dot{L}^S_\beta \;+\; (\omega\times L^S)_\beta \;+\; \dot{M}^S_\beta \;+\; (\omega\times M^S)_\beta \tag{10.61}$$

where

$$M^S_\beta \;\equiv\; I^S_{\beta\sigma}\,\omega_\sigma \tag{10.62}$$

is the angular momentum associated directly with motion in physical space.

Finally, by introducing the covariant derivative

$$\frac{D}{Dt} \;\equiv\; \frac{d}{dt} \;+\; \omega\times \tag{10.63}$$

and adopting standard vector notation, we arrive at the reduced equation

$$R^{-1}\frac{d\mathbf{L}}{dt} \;=\; \frac{D(\mathbf{M^S}+\mathbf{L^S})}{Dt} \tag{10.64}$$

References

1. T.D. Lee, C.N. Yang, Phys. Rev. **104**, 254 (1956); E Phys. Rev. **106**, 1371 (1957)
2. C.N. Yang, R. Mills, Phys. Rev. **96**, (1954)
3. C.N. Yang, *Selected Papers with Commentary*, vols. 1 and 2 (World Scientific, Singapore, 2005; 2013)
4. M.V. Berry, Proc. Roy. Acad. Lond. **A392** 45 (1984)
5. F. Wilczek, A. Zee, Phys. Rev. Lett. **52**, 2111 (1984)
6. A. Shapere, F. Wilczek (ed.), *Geometric Phases in Physics* (World Scientific, Singapore, 1989)
7. A. Bernevig, T. Hughes, *Topological Insulators and Topological Superconductors* (Princeton University Press, Princeton, 2013)
8. J. Leinaas, J. Myrheim, Il Nuovo Cimento **B37**(1), 1 (1977)
9. F. Wilczek, Phys. Rev. Lett. **48**, 1144 (1982)

10. D. Arovas, J.R. Schrieffer, F. Wilczek, Phys. Rev. Lett. **53**, 722 (1984)
11. H. Bartolomei et al., Science **368**, 173 (2020)
12. J. Nakamura, S. Liang, G. Gardner, M.!Manfra, Nature Physics **16**, 931 (2020)
13. H.-N. Dai, B. Yang, A. Reingruber, H. Sun, X.-F. Xu, Y.-A. Chen, Z.-S. Yuan, J.-W. Pan, Nature Physics **13**, 1195 (2017)
14. C. Marcus, C. Nayak, M. Hastings, M. Freedman, P. Bonderson, R. Lutchyn, T. Karzig, Phys. Rev. **B95**, 235305 (2017); C. Nayak, Microsoft Has Demonstrated the Underlying Physics to Create a New Kind of Qubit, https://www.microsoft.com/en-us/research/blog/microsoft-has-demonstrated-the-underlying-physics-required-to-create-a-new-kind-of-qubit/
15. J. Knolle, R. Moessner, Ann. Rev. Cond. Matt. Phys. **10**, 451 (2019)
16. A. Shapere, F. Wilczek, Am. J. Phys. **57**, 514 (1989)
17. R. Montgomery, Fields Inst. Comm. **1** 193 (1993)
18. R. Littlejohn, M. Reinsch, Rev. Mod. Phys. **69**, 213 (1997)
19. A. Cabrera, J. Geo. Phys. **58**, 334 (2008)
20. M.T. Bosch, ACM Trans. Graph. **39**(4), Article 1 (2020)
21. A. Guichardet, Annales de l' I. H. P. **A40**, 329 (1984)
22. T. Slivka, *On the Kinematics of Deformable Bodies*, Unpublished Princeton University senior thesis, 1990
23. D. Alba, L. Lusanna, M. Pauri, J. Math. Phys. **43**, 373 (2002)
24. J. Wisdom, Science **299**, 1865 (2003)
25. R. Batterman, Stud. Hist. Phil. Sci. **34**, 527 (2003)
26. X. Peng, J. Dai, A. Niemi, New J. Phys. **23**, 073024 (2021)
27. A. Shapere, F. Wilczek, J. Fluid Mech. **198**, 557 (1989)
28. J. Reiser, C. Gong, H. Asley, P. Schiebel, K. Diaz, B. Chong, C. Pierce, H. Lu, R. Hatton, H. Choset, D. Goldman, Geometric phase predicts locomotion performance in undulating living systems across scales. Proc. Natl. Acad. Sci. (2022, to appear)
29. P. Bunker, P. Jensen (ed.) *Molecular Symmetry and Spectroscopy*, 2nd edn. (NRC Research Press, Ottawa, 1998)
30. C. Eckart, Phys. Rev. **47**, 552 (1935)

Chapter 11
Why Does Quantum Field Theory in Curved Spacetime Make Sense? And What Happens to the Algebra of Observables in the Thermodynamic Limit?



Edward Witten

Abstract This article aims to explain some of the basic facts about the questions raised in the title, without the technical details that are available in the literature. We provide a gentle introduction to some rather classical results about quantum field theory in curved spacetime and about the thermodynamic limit of quantum statistical mechanics. We also briefly explain that these results have an analog in the large N limit of gauge theory.

11.1 Introduction

The first goal of the present article is to provide some intuition about a basic question: Why does quantum field theory in a curved spacetime of Lorentz signature make sense? Here we will give an informal introduction, referring the reader to the literature for more detail. An inevitably partial list of references on the general topic of quantum field theory in curved spacetime is [1–14]. For some technical detail on the specific topics that will be described rather heuristically here, the reader might start with the book [8]. Another useful reference is the collection of articles [12].

We consider always a globally hyperbolic spacetime, which means a spacetime M with a complete Cauchy hypersurface S on which initial conditions for classical or quantum fields can be formulated. The sense in which quantum field theory can be formulated on M depends very much on whether S is compact. In a closed universe, in other words if S is compact, the usual Hilbert space formulation of quantum field theory is valid: to a quantum field theory on M, we can associate in a natural way a Hilbert space $\mathcal{H}$, such that the quantum dynamics of the given theory can be described in the usual way in terms of operators acting on $\mathcal{H}$. However, in contrast

E. Witten (✉)
Institute for Advanced Study, Einstein Drive, Princeton, NJ, USA
e-mail: witten@ias.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_11

to the possibly more familiar case of quantum field theory in Minkowski spacetime, there is generically no distinguished ground state or vacuum vector in $\mathcal{H}$, since there is no natural "energy" that we should try to minimize. Thus, in formulating quantum field theory in curved spacetime, one should become accustomed to the idea of a naturally defined Hilbert space that does not contain any distinguished vector.[1]

The case that S is not compact, and with no simplifying assumptions about the behavior at infinity along S, is quite different. In this case, a natural construction of a Hilbert space does not exist. Instead, roughly speaking, one has to work with density matrices for local algebras of observables, more precisely an algebra $\mathscr{A}_U$ for each bounded open set $U \subset M$. And roughly speaking, quantum field theory on M describes the evolution not of a quantum state but of a density matrix. Actually a slight generalization of the notion of a density matrix is needed, because in quantum field theory the algebras $\mathscr{A}_U$ are von Neumann algebras of Type III, not of Type I [15–19], as will be explained in due course.

To be more precise, the statement that there is no natural Hilbert space to describe quantum field theory in an open universe holds generically, that is in the absence of any special assumptions about what is happening at spatial infinity. With suitable conditions at infinity (such as asymptotic flatness), a natural Hilbert space is sometimes available. A typical example in which a natural Hilbert space does not exist is an open Big Bang cosmology—for example an expanding universe with spatial sections of zero or negative curvature. In such a case, quantum dynamics must be described by evolution of density matrices (or more precisely of the Type III generalization of density matrices), not by operators on a Hilbert space.

This difference between a closed universe and in a generic open one might come as a surprise, but actually has little to do with any details of quantum field theory. The phenomenon results purely from the fact that, even with an ultraviolet cutoff in place, the number of degrees of freedom in an open universe is infinite. Thus a rather similar phenomenon occurs in the thermodynamic limit of quantum statistical mechanics. In finite volume, a quantum field theory has a Hilbert space $\mathcal{H}$, and thermodynamic functions such as entropy, energy density, etc., can be computed by studying a density matrix $\rho = \frac{1}{Z}e^{-\beta H}$ on $\mathcal{H}$ (as usual, H, β, and Z are the Hamiltonian, inverse temperature, and partition function). Thermodynamic functions and correlation functions can be computed in finite volume and they have a large volume limit. However, if one aims to describe statistical mechanics directly in the large volume limit, one runs into the fact that at nonzero temperature, the Hilbert space $\mathcal{H}$ does not have a large volume limit, for reasons similar to what happens in quantum field theory in an open universe. There is a cure for this problem [20], which is to use the thermofield double, namely $\rho^{1/2} = \frac{1}{\sqrt{Z}}e^{-\beta H/2}$, regarded not as

[1] A corollary is that in a general spacetime, there is no useful notion of a particle: particles in quantum field theory (or quasiparticles in condensed matter physics) are defined as excitations of a distinguished ground state with certain asymptotic properties, so in the absence of any preferred state, and in the absence, for a general M, of any asymptotic region where asymptotic properties could be defined, there is no notion of a "particle."

an operator on $\mathcal{H}$ but as a vector in the tensor product of two copies of $\mathcal{H}$. Starting with the thermofield double state $\Psi_{\rm TFD} = \rho^{1/2}$, one can build a Hilbert space $\mathcal{H}_{\rm TFD}$ that has an infinite volume limit, and such that thermodynamic functions and expectation values in the infinite volume thermal ensemble can be computed as correlation functions of operators acting in $\mathcal{H}_{\rm TFD}$. A price one pays is that the operators of the original system, when interpreted as operators on $\mathcal{H}_{\rm TFD}$, naturally generate a von Neumann algebra $\mathscr{A}_{\rm TFD}$ of Type III, not a more familiar algebra of Type I. The Hamiltonian operator acting on just one of the two copies in the thermofield double is *not* well-defined and in particular is not part of $\mathscr{A}_{\rm TFD}$. Rather, the generator of time translations acts on both copies in the thermofield double and generates an outer automorphism of $\mathscr{A}_{\rm TFD}$.

The need to go to the thermofield double in the thermodynamic limit of quantum statistical mechanics is quite analogous to the need to use density matrices (or their Type III analog) for quantum fields in a generic open universe. There is also a close analogy between the reasons for the occurrence of Type III algebras. In each of the two cases, the Type III nature of the algebra results from a divergent amount of entanglement. In statistical mechanics, there is divergent entanglement between the two copies of the original system that make up the thermofield double, and in quantum field theory in curved spacetime, there is divergent entanglement between modes inside and outside a given bounded region of space. One divergence is an infrared divergence and one is an ultraviolet divergence, but nonetheless they are quite similar. Indeed in the case of a conformal field theory, one can make a direct conformal mapping between the two types of divergence.

In the course of our study of the thermofield double, we will become familiar with the basic facts about (hyperfinite) von Neumann algebras, which come in three types, Type I, Type II, and Type III. All of these algebras can be constructed from simple qubit systems. The familiar algebras are of Type I. As already mentioned, the local algebras of quantum field theory, and likewise the algebras of observables in the infinite volume limit of quantum statistical mechanics, are of Type III.

In Sect. 11.2, we provide a gentle introduction to the facts about quantum field theory in curved spacetime that were summarized earlier. The aim is to motivate the statements and help the reader gain an intuition. In Sect. 11.3, we discuss in a similar spirit the thermodynamic limit of quantum statistical mechanics and the thermofield double. This article was originally planned as an exposition of those matters, along the lines of a lecture at the 2021 Bootstrap Summer School, where the most important points are outlined [21]. However, a recent article by Leutheusser and Liu [22], with some novel observations about quantum black holes, suggested that it would be worth while to discuss from a similar point of view the large N limit of gauge theory. These issues are briefly discussed in Sect. 11.4. More detail on some aspects, including a role for algebras of Type II in the presence of gravity, will appear elsewhere [23].

In analyzing quantum field theory in Lorentz signature, we will not make use of knowledge about what happens in Euclidean signature. This is in common with the existing literature, and there is a simple reason for it: a generic Lorentz signature spacetime does not have any useful Euclidean continuation. However, Euclidean

field theory is powerful, and it is a pity not to be able to apply this power to the Lorentz signature case. A recent analysis of quantum field theory on a spacetime with a complex metric [24] offers hope that it will be possible to bridge the gap and apply Euclidean-style reasoning to the Lorentz signature case.

11.2 Quantum Field Theory in Curved Spacetime

11.2.1 The Problem

As a motivating example of the problem of quantum field theory in curved spacetime, consider a free field theory coupled to a background metric g of Lorentz signature $-++\cdots+$ on a D-manifold M. For example, a real scalar of mass m is defined by an action

$$I = -\frac{1}{2}\int \mathrm{d}^D x\sqrt{g}\left(g^{\mu\nu}\partial_\mu\phi\partial_\nu\phi + m^2\phi^2\right). \tag{11.1}$$

We assume that M is a globally hyperbolic manifold with complete Cauchy hypersurface S. The field ϕ obeys a second order wave equation

$$(D^\mu D_\mu + m^2)\phi(\mathbf{x}, t) = 0, \tag{11.2}$$

and initial data consist of the field ϕ and its normal derivative $\dot{\phi}$ along S. These initial data are supposed to satisfy the usual canonical commutation relations $[\dot{\phi}(x), \phi(y)] = -\mathrm{i}\delta(x, y)$, with other commutators vanishing.

One of the most important facts about quantum mechanics is that given any *finite* set of canonical variables p_i, x^j, $i, j = 1, \cdots, d$ with the usual commutation relations $[p_i, x^j] = -\mathrm{i}\delta_i^j$, $[p_i, p_j] = [x^i, x^j] = 0$, the Hilbert space $\mathcal{H}$ that provides an irreducible representation of this algebra exists and is unique up to isomorphism. $\mathcal{H}$ can be realized as a space of square-integrable functions (more canonically, half-densities) of $\mathbf{x}$, or as a space of square-integrable functions of $\mathbf{p}$. One can also work with holomorphic functions of $z_i = x^i + \mathrm{i}p_i$. Many slight generalizations and combinations of these constructions are also possible. These constructions can all be shown to be equivalent by Fourier transformations and more general unitary transformations. More generally, according to a theorem of von Neumann [25], the quantization of $\mathbb{R}^{2d}$ is unique, up to isomorphism, as long as one requires that $\mathbf{x}$ and $\mathbf{p}$ satisfy the usual commutation relations.[2] All of these

[2] It is essential here that $\mathbb{R}^{2d}$ is not considered as an abstract symplectic manifold, but rather we are given a preferred set of linear functions $\mathbf{x}$, $\mathbf{p}$ on $\mathbb{R}^{2d}$ (unique up to linear symplectic transformations and additive constants) whose commutators are specified. The problem of quantizing $\mathbb{R}^{2d}$ as an abstract symplectic manifold with no additional structure is completely different, and has no natural solution, because of the operator ordering problem of quantum mechanics. In our field

constructions produce a distinguished Hilbert space $\mathcal{H}$. There is no distinguished vector in $\mathcal{H}$ unless more structure is present, for example, a Hamiltonian whose ground state would be such a distinguished vector.

In infinite dimensions, this uniqueness of the irreducible representation of the canonical commutators is completely false [26]. We will see that presently with examples. Though we primarily discuss bosons in this article, the same statements hold for fermions: the canonical anticommutators have an essentially unique representation in the case of finitely many fermionic variables, but not in the case of an infinite set of fermionic variables.

Once one picks a representation of the field variables on S, satisfying the canonical commutation relations, in a Hilbert space $\mathcal{H}$, the rest of the construction of the theory is straightforward. The field ϕ obeys the second order wave equation $(D_\mu D^\mu + m^2)\phi = 0$, by virtue of which $\phi(\mathbf{x}, t)$ can be expressed for all $\mathbf{x}, t$ in terms of $\phi, \dot{\phi}$ on S. Such an expression exhibits $\phi(\mathbf{x}, t)$ as an operator acting on $\mathcal{H}$, and therefore matrix elements $\langle \Psi | \phi(\mathbf{x}_1, t_1)\phi(\mathbf{x}_2, t_2) \cdots \phi(\mathbf{x}_n, t_n) | \chi \rangle$, for states $\Psi, \chi \in \mathcal{H}$, are uniquely determined, in principle. However, if one starts this construction with the "wrong" representation of the canonical commutation relations, then the resulting matrix elements $\langle \Psi | \phi(\mathbf{x}_1, t_1)\phi(\mathbf{x}_2, t_2) \cdots \phi(\mathbf{x}_n, t_n) | \chi \rangle$ are not physically sensible: they do not have the expected short distance singularities. From this point of view, the problem of quantization is to find the right representation of the canonical commutators such that the resulting correlation functions are physically sensible.

One simple case in which it is obvious on physical grounds how to construct the appropriate representation of the canonical commutation relations is a spacetime with a Killing vector field that is everywhere timelike. In a suitable coordinate system $t, \mathbf{x}$, the metric of such a spacetime is independent of the time t. In this case, we expect the quantum theory to have a self-adjoint Hamiltonian operator H that generates time translations. In particular, such an H is diagonalizable. Because the Killing vector field is everywhere timelike, we expect H to be bounded below. A representation of the canonical commutation relations that admits such an H is unique up to isomorphism. To construct this representation, we simply expand the field in c-number modes $f_k(\mathbf{x})$, $\overline{f}_k(\mathbf{x})$ of positive and negative frequency[3]

$$\phi(\mathbf{x}, t) = \sum_k \left(a_k \, f_k(\mathbf{x}) e^{-\mathrm{i}\omega_k t} + a_k^\dagger \overline{f}_k(\mathbf{x}) e^{\mathrm{i}\omega_k t} \right), \quad \omega_k > 0. \tag{11.3}$$

The modes can be normalized so that the operators a_k, $a_k^\dagger$ obey the canonical relations $[a_k, a_l] = [a_k^\dagger, a_l^\dagger] = 0$, $[a_k, a_l^\dagger] = \delta_{kl}$. Acting with $a_k^\dagger$ increases the

theory problem, the phase space comes with a distinguished set of linear functions, namely the field variables $\phi(x)$ and $\dot{\phi}(x)$.

[3] The sum here is a discrete sum in a closed universe, and becomes a continuous integral in an open universe. We assume for the moment that all modes of ϕ have positive or negative frequency; zero-modes will be incorporated presently.

energy (the eigenvalue of H) by ω_k, and acting with a_k reduces the energy by the same amount. We introduce a "ground state" Ω of lowest energy, annihilated by the a_k. Then we define a pre-Hilbert space $\mathcal{H}_0$ of all states that can be created by acting on Ω with finitely many $a^\dagger$'s. To be more precise, we define $\mathcal{H}_0$ to have a basis of states that we denote as

$$a_{k_1}^\dagger a_{k_2}^\dagger \cdots a_{k_n}^\dagger |\Omega\rangle, \quad n = 0, 1, 2 \cdots . \tag{11.4}$$

The a's and $a^\dagger$'s are defined to act on these states in the familiar way, satisfying the commutation relations. $\mathcal{H}_0$ has all the properties of a Hilbert space except completeness; that is why we call it a pre-Hilbert space. We take the Hilbert space closure of $\mathcal{H}_0$ and this gives us the desired Hilbert space $\mathcal{H}$ for the free field ϕ in a time-independent curved spacetime. Clearly, $\mathcal{H}$ contains a distinguished vector of minimum energy, namely Ω, which is annihilated by all the a_k. Conversely, any Hilbert space $\mathcal{H}$ that represents the canonical commutators and in which H is self-adjoint and bounded below must contain a vector Ω annihilated by all the a_k (otherwise acting repeatedly with the a_k on an eigenvector of H would lower the energy indefinitely), and therefore is equivalent to the Hilbert space that was just constructed.

This discussion must be slightly modified if there are zero-modes. For a simple example with zero-modes, simply set $m = 0$ in the model we have been discussing. Then the field ϕ does have a zero-mode, namely the constant mode ϕ_0. The canonical conjugate to ϕ_0 is the constant mode π_0 of $\dot{\phi}$. If π_0 is suitably normalized, the commutation relations of these modes are the usual canonical commutators $[\pi_0, \phi_0] = -\mathrm{i}$. This is a finite set of canonical variables, so the Hilbert space $\mathcal{H}'$ obtained by quantizing these modes is unique, up to isomorphism. It can be defined, for example, as the space of L^2 functions of ϕ_0, with π_0 acting by $-\mathrm{i}\partial/\partial\phi_0$. The unique representation of the canonical commutators of the full system in which H is self-adjoint and bounded below is then given by the Hilbert space $\mathcal{H} \otimes \mathcal{H}'$, where $\mathcal{H}$ is obtained by quantizing the nonzero modes as in the last paragraph and $\mathcal{H}'$ is obtained by quantizing the zero-modes. More generally, in a time-independent closed universe, the number of zero-modes, in a free field theory with fields of any spin, is always finite,[4] so the possible existence of zero-modes never presents any problem in quantization.

Note, however, that in the example with the massless scalar field, the Hilbert space $\mathcal{H}'$ that is obtained by quantizing the zero-modes does not have any distinguished vector. The Hamiltonian H acts in $\mathcal{H}'$ as a multiple of $\pi_0^2 = -\partial^2/\partial\phi_0^2$. This operator does not have a normalizable ground state that would provide a

[4] For example, in $U(1)$ gauge theory with gauge field A, zero-modes correspond to harmonic 1-forms on S, and the number of such modes is the first Betti number of S, which is always finite for compact S. In this example, the Hilbert space has different components labeled by the first Chern class of the line bundle on which A is a connection. Each component can be analyzed as we have done in the text for the massless scalar.

distinguished vector in $\mathcal{H}$. So this is a simple example of successfully quantizing a theory in curved spacetime and not finding any distinguished vector in the resulting Hilbert space. That is somewhat exceptional in a time-independent spacetime, since it only happens in the presence of zero-modes (and not always then). But when one relaxes the assumption of time-independence, it is the typical state of affairs.

In a time-dependent situation, it is less obvious how to select an appropriate representation of the canonical commutation relations. Before discussing this question, we will practice with the simpler case of a spin system. This will also provide useful background when we come to statistical mechanics in Sect. 11.3.

11.2.2 Practicing with a Spin System

A qubit is a two-state quantum system, comprising, possibly, the two states of a spin 1/2 particle. A single qubit realizes the algebra of Pauli matrices $\sigma_x, \sigma_y, \sigma_z$. So a system of n qubits realizes the corresponding algebra of operators $\sigma_{a,k}$ where $a = x$, y or z and $k \in \{1, \cdots, n\}$ labels the choice of qubit:

$$\sigma_{a,k}\sigma_{b,k} = \delta_{ab} + \mathrm{i}\epsilon_{abc}\sigma_{c,k}, \quad [\sigma_{a,k}, \sigma_{b,k'}] = 0, \quad k \neq k'. \tag{11.5}$$

For finite n, this algebra has, up to isomorphism, a unique irreducible representation, of dimension 2^n. But for an infinite set of qubits, the representation of the algebra becomes highly non-unique. That is the point that we will explore here.[5]

As a preliminary, consider this question. Consider a countably infinite set of qubits (with no additional structure). Does the Hilbert space of such a system have a countable or uncountable dimension?

A Hilbert space $\mathcal{H}^*$ that literally describes *all* states of a countably infinite collection of qubits definitely has an uncountable dimension. But it is normally not useful to do physics in a Hilbert space as big as that. Infinite constructions in physics are normally limits of finite constructions, and it is generally possible to take the limit in such a way that all questions of interest can be answered in a Hilbert space of countably infinite dimension. Such a Hilbert space is said to be "separable." What we should do to get a separable subspace of $\mathcal{H}^*$ depends on what we are interested in. To get anywhere, we should have a class of operators on $\mathcal{H}^*$ that we regard as sensible physical observables. Then we can look for a separable subspace $\mathcal{H}$ of the ridiculously big Hilbert space $\mathcal{H}^*$ that is big enough to realize the algebra of physical observables and to describe the states that we care about. There are two ingredients here: what class of observables do we care about and what class of states do we care about? The two cannot be specified independently since the observables we care about have to be able to act on the states that we care about.

[5] The following matters are discussed, for example, in section 2.3 of [27]. That book is also a useful reference for other topics relevant to the present article.

In the case of the qubits, we might decide that we want a class of observables that contains all the single qubit operators. A basis of operators on the kth qubit (apart from the identity) are the Pauli spin operators $\sigma_{a,k}$. We define an algebra $\mathscr{A}_0$ that consists of all polynomials in these operators. Thus $\mathscr{A}_0$ consists of finite linear combinations of products of the form $\sigma_{a_1,k_1}\sigma_{a_2,k_2}\cdots\sigma_{a_r,k_r}$, satisfying the relations (11.5). The reason to call this $\mathscr{A}_0$ is that usually one wants some kind of completion of $\mathscr{A}_0$ to get the algebra $\mathscr{A}$ of physical observables. But we cannot discuss the completion without deciding what kind of states we are interested in.

For example, we might be interested in states in which the qubits almost all have "spin up" along the z-axis, or in other words in which they are almost all in the state $|\uparrow\rangle = \begin{vmatrix}1\\0\end{vmatrix}$. Then we might decide that the physical Hilbert space $\mathcal{H}$ should at least contain the vector

$$\Omega_\uparrow = \otimes_n |\uparrow\rangle_n\,. \tag{11.6}$$

If we want to have $\mathscr{A}_0$ as an algebra of observables, we have to include all states obtained by acting with $\mathscr{A}_0$ on $\Omega_\uparrow$. When we act with $\mathscr{A}_0$ on $\Omega_\uparrow$, since $\mathscr{A}_0$ consists of all the operators that act on a finite set of qubits, we get all states in which all but finitely many qubits are in the state $|\uparrow\rangle$. States of this kind make a pre-Hilbert space $\mathcal{H}_0$ with a countable basis. Its Hilbert space completion $\mathcal{H}$ is separable. Moreover, we can now actually take a completion of $\mathscr{A}_0$ by adding convergent limits of sequences of operators in $\mathscr{A}_0$. (The relevant notion of convergence is discussed momentarily.) This will give the algebra $\mathscr{A}$ of all bounded operators on $\mathcal{H}$. By definition, the algebra of all bounded operators on a Hilbert space $\mathcal{H}$ is called a von Neumann algebra of Type I. von Neumann algebras of Type I are the familiar ones, but we will meet the other types.

Obviously, instead of the state $\Omega_\uparrow$, we could have started with another state in the very large Hilbert space $\mathcal{H}^*$. For example, setting $|\downarrow\rangle = \begin{vmatrix}0\\1\end{vmatrix}$, we could have started with the state $\Omega_\downarrow = \otimes_n |\downarrow\rangle_n$ in which the spins are all down. Then we would define a pre-Hilbert space $\mathcal{H}_0'$ in which all but finitely many spins are down; it has a Hilbert space completion $\mathcal{H}'$. The algebra $\mathscr{A}_0$ acts on $\mathcal{H}'$, and by including convergent limits of operators, it can again be completed to a von Neumann algebra $\mathscr{A}'$ that in this example will be the Type I algebra of all bounded operators on $\mathcal{H}'$.

The algebras $\mathscr{A}$ and $\mathscr{A}'$ are actually different, but to understand why we need to explain what we mean by saying that a sequence $\mathsf{a}_1, \mathsf{a}_2, \cdots \in \mathscr{A}_0$ converges. If $\mathscr{A}_0$ is understood as an algebra of bounded operators on a Hilbert space $\mathcal{H}$, the relevant notion of convergence (which mathematically is called a weak limit) is to say that the sequence $\mathsf{a}_1, \mathsf{a}_2, \cdots \in \mathscr{A}_0$ converges if the limit $\lim_{n\to\infty}\langle\psi|\mathsf{a}_n|\chi\rangle$ exists, for all $\psi, \chi \in \mathcal{H}$. In that case, one defines an operator a whose matrix elements are $\langle\psi|\mathsf{a}|\chi\rangle = \lim_{n\to\infty}\langle\psi|\mathsf{a}_n|\chi\rangle$, and one says that $\lim_{n\to\infty}\mathsf{a}_n = \mathsf{a}$. The explanation given by Haag [17] for why this is the appropriate notion of convergence for physics is that any given experiment measures finitely many matrix elements to

finite precision, and therefore, if a is the weak limit of a sequence $\mathsf{a}_1, \mathsf{a}_2, \cdots$, then any given experiment cannot distinguish a_n from a if n is sufficiently large. The definition of a weak limit, however, makes it clear that to determine the limit of a sequence of elements in $\mathscr{A}_0$, we need to know the Hilbert space $\mathcal{H}$ on which $\mathscr{A}_0$ is acting. For a concrete example, let $\sigma_{+,n} = \frac{1}{2}\left(\sigma_{x,n} + \mathrm{i}\sigma_{y,n}\right)$ be the spin raising operator for the nth qubit. Then $\lim_{n\to\infty} \sigma_{+,n}^{\dagger}\sigma_{+,n} = 0$ if the operators are considered to act on $\mathcal{H}$, but $\lim_{n\to\infty} \sigma_{+,n}^{\dagger}\sigma_{+,n} = 1$ if the operators are consisted to act on $\mathcal{H}'$. These statements reflect the fact for any state in $\mathcal{H}$, the nth spin is almost certainly up (and annihilated by $\sigma_{+,n}$) for sufficiently large n, while for any state in $\mathcal{H}'$, it is almost certainly down. As a matter of terminology, an algebra of operators that is closed under weak limits is called a von Neumann algebra.

Instead of considering an abstract collection of qubits with no structure, we could arrange the qubits on a lattice of some dimension and consider a lattice Hamiltonian H, for instance a Heisenberg spin Hamiltonian. Then it would be natural to take Ψ to be the ground state of this Hamiltonian, assuming it is unique. Starting with this Ψ, we would then construct a separable Hilbert space $\mathcal{H}$ on which all of the elementary spin operators act. In this Hilbert space, we would complete the algebra $\mathscr{A}_0$ to a von Neumann algebra $\mathscr{A}$ acting on $\mathcal{H}$. This algebra will be of Type I. Our previous discussion amounted to the special case that $H = -\sum_n \sigma_{z,n}$, leading to the ground state $\Psi = \Omega_{\uparrow}$, or $H = +\sum_n \sigma_{z,n}$, leading to $\Psi = \Omega_{\downarrow}$. If H has degenerate ground states, then one would want to pick a basis of ground states consisting of states that satisfy cluster decomposition, and proceed as before, starting with any one of those states. In a moment, we will explain what is special about the states that satisfy cluster decomposition.

We can similarly construct a separable Hilbert space with an action of a completion of the algebra $\mathscr{A}_0$ starting with any vector $\Psi \in \mathcal{H}^*$. There is a certain sense, however, in which it is not true that every Ψ is an equally good starting point. To see this, consider the vector $\Psi = \frac{1}{\sqrt{2}}\Omega_{\uparrow} + \frac{1}{\sqrt{2}}\Omega_{\downarrow}$. Let $\pi_{\uparrow}$ and $\pi_{\downarrow}$ be the operators that project the first qubit onto the states $|\uparrow\rangle$ and $|\downarrow\rangle$, respectively. Then $\pi_{\uparrow}\Psi = \frac{1}{\sqrt{2}}\Omega_{\uparrow}$ and $\pi_{\downarrow}\Psi = \frac{1}{\sqrt{2}}\Omega_{\downarrow}$. Further acting with arbitrary elements of $\mathscr{A}_0$, we see that any state in either $\mathcal{H}_{\uparrow}$ or $\mathcal{H}_{\downarrow}$ can be approximated by states $\mathsf{a}\Psi$, $\mathsf{a} \in \mathscr{A}_0$, and therefore $\mathcal{H}_{\Psi} = \mathcal{H}_{\uparrow} \oplus \mathcal{H}_{\downarrow}$. Thus, while it is true that by acting with $\mathscr{A}_0$ on the state Ψ, we can build a separable Hilbert space $\mathcal{H}_{\Psi}$ that admits an action of a completion $\mathscr{A}$ of $\mathscr{A}_0$, we see that $\mathscr{A}$ does not act irreducibly on $\mathcal{H}_{\Psi}$, since $\mathcal{H}_{\Psi}$ has the $\mathscr{A}$-invariant decomposition $\mathcal{H}_{\uparrow} \oplus \mathcal{H}_{\downarrow}$. If Ψ is such that a completion of $\mathscr{A}_0$ does act irreducibly on $\mathcal{H}_{\Psi}$, we say that Ψ is a pure state[6] for the algebra $\mathscr{A}_0$. The pure states are the most interesting ones since they correspond to the irreducible

[6] There is some tension here with the usual terminology in quantum mechanics, where the phrase "pure state" usually refers to any Hilbert space state, without such a restriction. The point is that if $\mathscr{A}$ is a Type I algebra, consisting of all bounded operators on a Hilbert space $\mathcal{H}$, then every $\Psi \in \mathcal{H}$ is a pure state in the algebraic sense (the action of $\mathscr{A}$ on Ψ generates the whole Hilbert space $\mathcal{H}$, on which $\mathscr{A}$ acts irreducibly). Thus for the case that $\mathscr{A}$ is of Type I, the two notions of "pure state" are compatible.

representations of the relevant algebras of observables. In the example of a lattice spin Hamiltonian with multiple ground states, the pure states are the ones that satisfy cluster decomposition.

This description of what is meant by a pure state for the algebra $\mathscr{A}_0$ is not the most common one. Usually, one considers the function $F_\Psi(\mathsf{a}) = \langle\Psi|\mathsf{a}|\Psi\rangle$, which obeys the following properties:

- it is linear in a, namely $F(\lambda\mathsf{a}+\mu\mathsf{b}) = \lambda F(\mathsf{a})+\mu F(\mathsf{b})$, for $\mathsf{a}, \mathsf{b} \in \mathscr{A}_0$, $\lambda, \mu \in \mathbb{C}$,
- it is normalized to $F(1) = 1$,
- it is positive, in the sense that $F(\mathsf{a}^\dagger\mathsf{a}) \geq 0$ for all $\mathsf{a} \in \mathscr{A}$.

A function on an algebra, in the present example $\mathscr{A}_0$, that satisfies these properties is called a state on the algebra. Clearly then for any $\Psi \in \mathcal{H}^*$, the function $F_\Psi(\mathsf{a}) = \langle\Psi|\mathsf{a}|\Psi\rangle$ is a state on the algebra $\mathscr{A}_0$. A state on an algebra is called "pure" if it is not a convex linear combination of other states, or in other words if it is not possible to write $F(\mathsf{a}) = p_1F_1(\mathsf{a}) + p_2F_2(\mathsf{a})$ with $p_1, p_2 > 0$ and states F_1, F_2. In the present example, one readily sees that

$$F_\Psi(\mathsf{a}) = \frac{1}{2}\langle\Omega_\uparrow|\mathsf{a}|\Omega_\uparrow\rangle + \frac{1}{2}\langle\Omega_\downarrow|\mathsf{a}|\Omega_\downarrow\rangle, \quad \mathsf{a} \in \mathscr{A}_0. \tag{11.7}$$

So the state F_Ψ is not pure. The two notions of a pure state that we have described are equivalent: $\mathcal{H}_\Psi$ is irreducible for a completion of $\mathscr{A}_0$ if and only if the state $F_\Psi(\mathsf{a})$ is pure. The interested reader can verify this, possibly after reading about the GNS construction in Sect. 11.3.1.

Suppose that we construct a separable Hilbert space $\mathcal{H}_\Psi$ as before from some input state $\Psi \in \mathcal{H}^*$, and another Hilbert space $\mathcal{H}_{\Psi'}$ starting with a different input state Ψ'. $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi'}$ are both subspaces of the very big Hilbert space $\mathcal{H}^*$. Are they the same subspaces? A necessary condition is clear: Ψ' should be in $\mathcal{H}_\Psi$. If so, then Ψ' can be approximated by states $\mathsf{a}\Psi$, with $\mathsf{a} \in \mathscr{A}_0$. But in that case, any vector in $\mathcal{H}_{\Psi'}$, which by definition can be approximated by $\mathsf{a}'\Psi'$ with $\mathsf{a}' \in \mathscr{A}_0$, can in fact be approximated by $\mathsf{a}'\mathsf{a}\Psi$ and hence is in $\mathcal{H}_\Psi$. In other words, if $\Psi' \subset \mathcal{H}_\Psi$ then $\mathcal{H}_{\Psi'} \subset \mathcal{H}_\Psi$. If Ψ is pure, it actually follows that $\mathcal{H}_{\Psi'} = \mathcal{H}_\Psi$. That is because if $\mathcal{H}_{\Psi'} \subset \mathcal{H}_\Psi$, then $\mathcal{H}_\Psi$ has an $\mathscr{A}_0$-invariant decomposition $\mathcal{H}_\Psi = \mathcal{H}_{\Psi'} \oplus \mathcal{H}_{\Psi'}^\perp$. If Ψ is a pure state for the algebra $\mathscr{A}_0$, then one of the summands here must be trivial and so in this case $\mathcal{H}_\Psi = \mathcal{H}_{\Psi'}$.

11.2.3 A System of Harmonic Oscillators

Now let us discuss something that is a little closer to field theory. We will consider an infinite sequence of pairs of canonical variables, (x_1, p_1), (x_2, p_2), $\cdots$. The kth pair can be quantized to give a Hilbert space $\mathcal{H}_k$, and for this, there is no need to know anything about a preferred vector. If we want to describe all of the states in all

of the $\mathcal{H}_k$, we have to define a Hilbert space

$$\mathcal{H}^* = \otimes_k \mathcal{H}_k \tag{11.8}$$

of uncountably infinite dimension. But if we are given a state Ψ that for some reason we like, we can proceed as we did with the qubits. We define an algebra $\mathcal{A}_0$ consisting of all polynomials in the x's and p's. Then we define a pre-Hilbert space $\mathcal{H}_0$ consisting of all states that can be made from Ψ by acting with $\mathcal{A}_0$, and as usual we complete this to a Hilbert space $\mathcal{H}$. We can then also complete $\mathcal{A}_0$ to an algebra $\mathcal{A}$.

A simple example of this would be that Ψ is a tensor product of states $\psi_k \in \mathcal{H}_k$:

$$\Psi = \otimes_k \psi_k. \tag{11.9}$$

We can assume that the ψ_k and therefore also Ψ are normalized, $(\psi_k, \psi_k) = (\Psi, \Psi) = 1$. We construct a Hilbert space $\mathcal{H}_\Psi$ by acting with $\mathcal{A}_0$ on Ψ. On the other hand, consider another state of the same kind, $\Psi' = \otimes_k \psi'_k$, where again we assume the ψ'_k and Ψ' to be normalized. Is the Hilbert space $\mathcal{H}_{\Psi'}$ the same as $\mathcal{H}_\Psi$?

Let $c_k = (\psi'_k, \psi_k)$. After possibly changing the phases of the ψ'_k, we can assume that $c_k \geq 0$; clearly $c_k \leq 1$. We have

$$|\langle \Psi', \Psi \rangle| = \prod_{k=1}^{\infty} c_k. \tag{11.10}$$

We consider separately two cases: (1) $\langle \Psi', \Psi \rangle \neq 0$; (2) $\langle \Psi', \Psi \rangle = 0$. In case (1), the c_k rapidly converge to 1. By acting with the algebra $\mathcal{A}_0$ on Ψ, we can change any finite set of c_k's at will, subject to the condition $c_k \leq 1$, without changing the normalization condition $(\psi_k, \psi_k) = 1$. In particular, we can find a sequence $\mathsf{a}_1, \mathsf{a}_2, \cdots \in \mathcal{A}_0$ such that the states $\mathsf{a}_k \Psi$ are all normalized and $\lim_{n\to\infty} \langle \Psi', \mathsf{a}_k \Psi \rangle = 1$. In other words, Ψ' can be approximated by vectors $\mathsf{a}_k \Psi$; hence $\Psi' \in \mathcal{H}_\Psi$. The relation between Ψ and Ψ' was symmetrical, so likewise $\Psi \in \mathcal{H}_{\Psi'}$. As in Sect. 11.2.2, it follows that $\mathcal{H}_\Psi = \mathcal{H}_{\Psi'}$.

In case (2), where $\prod_{k=1}^{\infty} c_k$ vanishes, the interesting case is that the c_k are not individually 0 and $\prod_{k=1}^{\infty} c_k = 0$ because the c_k do not approach 1 fast enough to make this product nonzero. Then this vanishing cannot be changed by changing finitely many of the c_k. So in such a case, Ψ' is orthogonal not just to Ψ but to $\mathsf{a}\Psi$ for all $\mathsf{a} \in \mathcal{A}_0$. It follows then that $\mathsf{a}'\Psi'$ is orthogonal to $\mathsf{a}\Psi$ for all $\mathsf{a}, \mathsf{a}' \in \mathcal{A}_0$, since $\langle \mathsf{a}'\Psi' | \mathsf{a}\Psi \rangle = \langle \Psi' | \mathsf{a}'^\dagger \mathsf{a} \Psi \rangle = 0$. So $\mathcal{H}_{\Psi'}$ and $\mathcal{H}_\Psi$ are orthogonal.

If finitely many of the c_k vanish, those c_k's should just be disregarded since they could be set to 1 by replacing Ψ with some $\widetilde{\Psi} = \mathsf{a}\Psi$, $\mathsf{a} \in \mathcal{A}_0$, without changing the other c_k's. So in case finitely many of the c_k vanish, we look at the restricted product $\prod'_k c_k$ over the nonzero c_k; we will get $\mathcal{H}_{\Psi'} = \mathcal{H}_\Psi$ if this restricted product is positive, and otherwise $\mathcal{H}_{\Psi'}$ is orthogonal to $\mathcal{H}_\Psi$. If infinitely many c_k vanish, we cannot change this by acting with an element of $\mathcal{A}_0$, so $\mathcal{H}_{\Psi'}$ and $\mathcal{H}_\Psi$ are orthogonal.

It is noteworthy that there are only two outcomes here: $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi'}$ are the same, or they are orthogonal.[7] For $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi'}$ to be the same, we require no relation at all between ψ_k and ψ'_k for all but finitely many modes, but for large k, ψ_k and ψ'_k must coincide asymptotically at a sufficiently fast rate.

Now we will consider a special case of this that is useful background for field theory. For each k, pick a complex number τ_k in the upper half plane, and define

$$a_k^\dagger = \frac{1}{\sqrt{2\mathrm{Im}\,\tau_k}}\left(x_k + \tau_k p_k\right), \quad a_k = \frac{1}{\sqrt{2\mathrm{Im}\,\tau_k}}\left(x_k + \overline{\tau}_k p_k\right). \tag{11.11}$$

This has been chosen to ensure that $[a_k^\dagger, a_{k'}] = \delta_{kk'}$, with other commutators vanishing. We can choose ψ_k so that $a_k\psi_k = 0$, whereupon if $\Psi = \otimes_k \psi_k$, we have $a_k\Psi = 0$ for all k. We can view Ψ as the ground state of a harmonic oscillator Hamiltonian $H = \sum_k \epsilon_k a_k^\dagger a_k$, for arbitrary $\epsilon_k > 0$.

On the other hand, we could make the same construction with a different set of parameters τ'_k, leading to a different set of creation and annihilation operators $a'^\dagger_k$, a'_k and a different vector $\Psi' = \otimes_k \psi'_k$ that is annihilated by the new annihilation operations. Now we can ask if the Hilbert spaces $\mathcal{H}_\Psi$ and $\mathcal{H}_{\Psi'}$ are the same. From the above analysis, the condition for this is $\prod_{k=1}^\infty c_k > 0$, with $c_k = \langle\psi'_k, \psi_k\rangle$. This is equivalent to a condition that τ'_k is sufficiently close to τ_k for large k, with no condition at all on τ'_k for any finite set of values of k.

For any given k, at the classical level, a group $SL(2,\mathbb{R}) \cong Sp(2,\mathbb{R})$ of linear canonical transformations acts on the pair x_k, p_k, preserving the commutation relations. The group that acts on the quantum Hilbert space is a double cover of $Sp(2,\mathbb{R})$, called the metaplectic double cover.[8] We denote the double cover as $\widehat{Sp}(2,\mathbb{R})$. It is possible to choose for each k an element $g_k \in \widehat{Sp}(2,\mathbb{R})$ that conjugates $a_k^\dagger$ and a_k to $a'^\dagger_k$ and a'_k, and maps ψ_k to ψ'_k. Clearly, any finite product of the g_k's acts on the Hilbert space $\mathcal{H}_\Psi$ (and likewise on $\mathcal{H}_{\Psi'}$). But the infinite tensor product $\otimes_{k=1}^\infty g_k$ maps Ψ to Ψ', so it maps $\mathcal{H}_\Psi$ to $\mathcal{H}_{\Psi'}$. It only acts within $\mathcal{H}_\Psi$ if $\prod_{k=1}^\infty c_k > 0$, which is equivalent to saying that g_k approaches 1 sufficiently rapidly for $k \to \infty$.

The moral of the story is that in the case of an infinite set of canonical variables, a linear canonical transformation is realized in a given representation of the canonical commutation relations if and only if it is sufficiently close to the identity on all but finitely many of the variables. To motivate this statement, we have considered the simple case of a canonical transformation that, in a suitable basis, acts separately on

[7] We did not run into other possibilities encountered in Sect. 11.2.2, because the states Ψ and Ψ' are pure for the algebra $\mathscr{A}_0$. This resulted from their product structure.

[8] One way to see the occurrence of this double cover is the following. For a pair of canonical variables x, p, let H be the harmonic oscillator Hamiltonian $H = \frac{1}{2}(p^2 + x^2)$. As a canonical transformation of x and p, $\exp(2\pi\mathrm{i}H) = 1$. But quantum mechanically, as the eigenvalues of H are all half-integers, one has $\exp(2\pi\mathrm{i}H) = -1$.

each pair of variables x_k, p_k. However, the conclusion is general and is explained in generality, for example, in [8].

Another way to state the conclusion is that for an infinite set of canonical variables, to determine a specific representation of the canonical commutators in a separable Hilbert space $\mathcal{H}$, one needs to describe the desired representation to sufficient precision for all but finitely many variables. This may be done by decomposing the field variables in creation and annihilation operators, where $\mathcal{H}$ is supposed to contain a vector annihilated by the annihilation operators. To specify a particular $\mathcal{H}$ in this way, one must specify the decomposition in creation and annihilation operators in an asymptotically precise way.

11.2.4 Back to Field Theory

Now we go back to quantum field theory in curved spacetime. First we consider a closed universe, that is a globally hyperbolic spacetime M with compact Cauchy hypersurface S. Modes along S of very short wavelength compared to the radius of curvature of S can be separated to positive and negative frequency, with an uncertainty that rapidly vanishes in the limit of short wavelengths. So for very short wavelength modes, there is a natural decomposition in creation and annihilation operators to good approximation, and this approximation becomes asymptotically precise in the limit of short wavelengths. For low energy modes, we have no notion at all of how to make a decomposition in creation and annihilation operators, or how to pick a preferred state in any other way. But in a closed universe, there are only finitely many low energy modes.

Therefore, we are in a good situation. For all but finitely many low energy modes, we have a good approximate notion of how to make a separation in creation and annihilation operators. Moreover this notion is asymptotically precise for the infinitely many modes of asymptotically short wavelength. Such an asymptotic separation in creation and annihilation operators is precisely sufficient to define a distinguished Hilbert space $\mathcal{H}$ with an action of the canonical commutation relations. The Hilbert space $\mathcal{H}$ constructed this way does not really depend on the choice of S, because at asymptotically small wavelengths, evolution of the field ϕ via its equations of motion preserves the separation in positive and negative frequencies. So the asymptotic separation in positive and negative frequencies made on one surface S agrees with the asymptotic separation that one would make on another surface S'. Moreover, matrix elements of products of field operators between vectors in $\mathcal{H}$ have the standard short distance singularities, because $\mathcal{H}$ was defined using the standard separation in positive and negative frequencies at short distances.

A rigorous version of this argument is described on pp. 96-7 of [8], roughly as follows. Without changing the metric of M near S, we can modify it to be time-independent sufficiently far to the past of M and sufficiently far to the future of M. Let g_- be the time-independent metric in the past and g_+ the time-independent

metric in the future. As explained in Sect. 11.2.1, using the time-independent metric g_- in the past, we can straightforwardly construct a representation of the canonical commutation relations based on a decomposition in positive and negative frequencies. We call the resulting Hilbert space $\mathcal{H}_{g_-}$. Following the same procedure in the future, we construct a representation of the canonical commutation relations, and a Hilbert space that we call $\mathcal{H}_{g_+}$. Either choice, together with the equations of motion, determines a representation of the canonical commutators throughout all of M. However, the two representations and the two Hilbert spaces $\mathcal{H}_{g_-}$ and $\mathcal{H}_{g_+}$ are canonically the same. This is so because in a spacetime that is time-independent in both the past and the future, at asymptotically high energies, propagation from the past to the future preserves the separation in positive and negative frequencies. The error in this statement is exponentially small at high energies, so the criterion discussed in Sect. 11.2.3 for two representations of the canonical commutation relations to be the same is satisfied. Saying that $\mathcal{H}_{g_-}$ and $\mathcal{H}_{g_+}$ are canonically the same means that one can define transition amplitudes from a state in $\mathcal{H}_{g_-}$ in the past to a state in $\mathcal{H}_{g_+}$ in the future, and it also means that if we restrict to a neighborhood of S, we get the same representation of the canonical commutation relations and the same Hilbert space whether we use $\mathcal{H}_{g_-}$ or $\mathcal{H}_{g_+}$. But as g_- and g_+ can be varied independently, the fact that the representation of the commutation relations that we get in a neighborhood of S can be determined just from g_- and can also be determined just from g_+ means that actually this representation does not depend on g_- or g_+ at all. In other words, this construction produces a distinguished representation of the canonical commutators on S. Now we return to the original metric on M, whatever it was, with no assumption of time-independence anywhere, and we use the representation of the canonical commutation relations in a neighborhood of S that was just determined. As explained in Sect. 11.2.1, a choice of a representation of the canonical commutation relations near S, when combined with the equations of motion, is sufficient to completely determine the theory. This concludes the argument.

This discussion may make it clear why (as also discussed in [8]) there is not a natural Hilbert space for a quantum field in a generic open universe, for example in a Big Bang cosmology. In a generic open universe, there are infinitely many modes of moderate wavelength for which we have no candidate for a preferred state or for a separation in creation and annihilation operators. Hence there is no natural construction of a Hilbert space. To construct a Hilbert space, either by picking a separation in creation and annihilation operators for all the modes of moderate wavelength, or by otherwise specifying a seed vector Ψ that could be the starting point of the construction, one has to supply a great deal of detailed information about what is happening at spatial infinity. It is possible to define a Hilbert space that describes a quantum field in an open universe, but there are many inequivalent constructions of such a Hilbert space and there is no natural choice.

11.2.5 What Is Quantum Field Theory in an Open Universe?

If there is no natural Hilbert space in an open universe, what can it mean to do quantum field theory in one? Before answering this question, let us first discuss closed universes a little more. Consider an observer who has access only to a portion of a closed universe, or who wants to develop a formalism suitable to describe experiments in just a portion of the universe. We will be particularly interested in a region of spacetime of the following sort (Fig. 11.1). Let B be a closed, codimension 1 submanifold of a Cauchy hypersurface S, such that the complement of B is the union of two disjoint open sets V and V'. Let $\mathcal{U}$ and $\mathcal{U}'$ be the domains of dependence of V and V', respectively. We call an open set of the form $\mathcal{U}$ or $\mathcal{U}'$ a "local region," and we want to know what quantum field theory says for an observer who makes observations only in such a local region.

Classically, in a relativistic field theory, fields in $\mathcal{U}$ are determined by initial data on V. Quantum mechanically, this is also true in free field theory; we exploited this fact in Sect. 11.2.1. In a generic non-free quantum field theory, it is in general not possible to define observables on a spacelike surface such as V, and instead we thicken V slightly to an open set $\mathcal{V}$, satisfying $V \subset \mathcal{V} \subset \mathcal{U}$. Roughly, we will describe the sense in which initial data for observations in $\mathcal{U}$ can be formulated in $\mathcal{V}$. Note that $\mathcal{U}$ and $\mathcal{V}$ are both globally hyperbolic spacetimes in their own right, with a Cauchy hyperurface—namely V—which is a portion of the Cauchy hypersurface S of the full spacetime M.

An analog in ordinary quantum mechanics of restricting from a spacetime M to a local region $\mathcal{U}$ is to consider only a subsystem X of a larger bipartite system XY. In ordinary quantum mechanics, the subsystems X and Y are described by Hilbert spaces $\mathcal{H}_X$ and $\mathcal{H}_Y$, and the composite system XY is described by the tensor product $\mathcal{H}_{XY} = \mathcal{H}_X \otimes \mathcal{H}_Y$. A general state of the full system XY, possibly a pure state, when restricted to the subsystem X, is described by a density matrix ρ_X. Here ρ_X is a self-adjoint operator, positive and of trace 1, acting on the Hilbert space $\mathcal{H}_X$ of the subsystem. Let $\mathscr{A}_X$ be the algebra of all operators of subsystem X; this is the same as the algebra of all operators on $\mathcal{H}_X$. For $\mathsf{a} \in \mathscr{A}_X$, its expectation value in

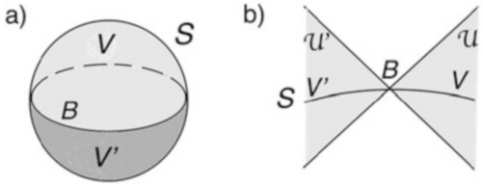


Fig. 11.1 (**a**) A Cauchy hypersurface S, drawn as a two-sphere, is divided by the equator B into an upper hemisphere V and a lower hemisphere V'. (**b**) In this view, S is drawn as a one-dimensional curve and B is depicted as a point that divides S into the two open sets V and V'. $\mathcal{U}$ and $\mathcal{U}'$ are open sets in spacetime that are the domains of dependence of V and V'. It is apparent from this picture that near B, $\mathcal{U}$ and $\mathcal{U}'$ can be modeled by opposite Rindler wedges in Minkowski space

a state characterized by the density matrix ρ_X is defined as $F(\mathsf{a}) = \mathrm{Tr}\,\rho_X \mathsf{a}$. The function $F(\mathsf{a})$ has three key properties, already introduced in Sect. 11.2.2, which characterize what is known as a "state" on an algebra:

- it is linear in a, namely $F(\lambda\mathsf{a}+\mu\mathsf{b}) = \lambda F(\mathsf{a})+\mu F(\mathsf{b})$, for $\mathsf{a}, \mathsf{b} \in \mathscr{A}_X$, $\lambda, \mu \in \mathbb{C}$,
- it is normalized to $F(1) = 1$,
- it is positive, in the sense that $F(\mathsf{a}^\dagger\mathsf{a}) \geq 0$ for all $\mathsf{a} \in \mathscr{A}_X$.

So a density matrix ρ_X for the subsystem X is a state on the algebra $\mathscr{A}_X$ consisting of all operators on the Hilbert space $\mathcal{H}_X$ of subsystem X.

In quantum field theory, there is no way to associate a Hilbert space to a local region in spacetime. The best one can do is to associate to an open set $\mathcal{W} \subset M$ a corresponding algebra of operators $\mathscr{A}_\mathcal{W}$. The algebras $\mathscr{A}_\mathcal{W}$ obey physically motivated axioms, such as a condition associated to causality: $\mathscr{A}_\mathcal{W}$ and $\mathscr{A}_{\mathcal{W}'}$ commute (in the $\mathbb{Z}_2$-graded sense if fermions are present) if the regions $\mathcal{W}$ and $\mathcal{W}'$ are spacelike separated.[9] We will discuss in more detail presently the construction of $\mathscr{A}_\mathcal{W}$. For now, we just note a difference between a local region in quantum field theory and a subsystem in ordinary quantum mechanics: the algebra $\mathscr{A}_\mathcal{W}$ is, in the von Neumann algebra language, of Type III, not Type I [15–19]. This will be explained in Sect. 11.3, in the analogous setting of quantum statistical mechanics. The Type III nature of the algebras is actually the reason that there is no way to associate a Hilbert space to a local region. If the algebra $\mathscr{A}_\mathcal{W}$ were of Type I, it would have an irreducible representation in a Hilbert space, and this Hilbert space would be naturally associated to the open set $\mathcal{W}$ in spacetime. However, a Type III algebra does not have an irreducible representation in a Hilbert space. Accordingly in quantum field theory, there is no natural way to associate a Hilbert space $\mathcal{H}_\mathcal{W}$ to an open set $\mathcal{W}$. All that we can really associate to the region $\mathcal{W}$ is the algebra $\mathscr{A}_\mathcal{W}$, not a preferred Hilbert space that it acts on.

There is also no good notion of a density matrix for the region $\mathcal{W}$, since this notion is not applicable for a Type III algebra. But the notion of a state of the algebra $\mathscr{A}_\mathcal{W}$—a linear function obeying conditions that were stated previously—does make sense. This is the appropriate analog in quantum field theory of the notion of the density matrix of a subsystem in ordinary quantum mechanics.

With this in mind, what should quantum dynamics mean for a local region $\mathcal{U} \subset M$? We assume as described earlier that $\mathcal{U}$ is the domain of dependence of some set V, which has a slight thickening to an open set $\mathcal{V} \subset \mathcal{U}$. If we could associate a Hilbert space to quantum fields on V or in $\mathcal{V}$, we could describe quantum dynamics in terms of state vectors. We would say that a quantum state that defines initial conditions on V or in $\mathcal{V}$ actually determines the probabilities for measurement outcomes in the larger region $\mathcal{U}$. Since instead we only have algebras of operators associated to $\mathcal{V}$ and $\mathcal{U}$, we have to say something similar in terms

[9] See for example section 5.3.1 of the article by Brunetti and Fredenhagen in [12] for a useful statement of axioms in the context of curved spacetime. In Minkowski space, a standard reference is [17].

of algebras. The necessary statement is just that in a quantum field theory, in this situation $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$, meaning that operators in region $\mathcal{U}$ are actually equivalent to operators in region $\mathcal{V}$ (although simple operators in region $\mathcal{U}$ might correspond to rather complicated operators in the smaller region $\mathcal{V}$). Since $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$, a state of $\mathscr{A}_{\mathcal{V}}$ is automatically a state of $\mathscr{A}_{\mathcal{U}}$: in other words, the data needed to predict measurements outcomes in region $\mathcal{V}$ (to the extent that such outcomes are predictable in quantum mechanics) also suffice to predict measurement outcomes in the larger region $\mathcal{U}$. The statement that $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$, together with general axioms about the local algebras, such as those described in [12, 17], is the content of quantum dynamics for a local region.

Now let us discuss in more detail the definition of the algebra $\mathscr{A}_{\mathcal{W}}$ associated to an open set $\mathcal{W}$. We assume to begin with that $\mathcal{W}$ is contained in a closed universe M, or at least in some spacetime (such as Minkowski space) to which the quantum field theory of interest associates a Hilbert space $\mathcal{H}$. A standard approach to defining $\mathscr{A}_{\mathcal{W}}$ is as follows. Let $O(x)$ be a local operator of the quantum field theory in question, and f a smooth function with compact support in $\mathcal{W}$. Then $K_f = \exp(\mathrm{i}\int_{\mathcal{W}} \mathrm{d}^D x \sqrt{g} f(x) O(x))$ is a bounded operator on $\mathcal{H}$. $\mathscr{A}_{\mathcal{W}}$ is then defined as the von Neumann algebra generated by such operators on $\mathcal{H}$.

The purpose of introducing the global Hilbert space $\mathcal{H}$ was to make sure that expressions such as K_f can be interpreted as Hilbert space operators so that it is possible to take weak limits and define a von Neumann algebra. The von Neumann algebra language is useful because it makes possible simple statements such as $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$ (which would not hold if we do not complete the algebras by taking weak limits). However, this language has the drawback of not directly incorporating the operator product expansion (OPE), which is an important statement of the locality of quantum field theory. One approach to incorporating the OPE is to enrich the von Neumann algebra language with a further axiom which would imply the existence of an OPE [28]. Another approach is to state axioms for quantum field theory in curved spacetime that directly incorporate the OPE [11]. In some approaches, the notion of a state has to be refined to incorporate the expected short distance singularities of the quantum field theory under study, but it is preferable if (as in [11]) the expected short distance behavior can be built into the structure of the algebra. The best treatment of such questions is not entirely clear in the author's opinion, and in this article we will not adopt any particular point of view.

Before discussing an open universe, we need a few more facts about local regions in a closed universe M. Let us specialize to the case of an open set that is a local region $\mathcal{U} \subset M$, or a smaller open set $\mathcal{V} \subset \mathcal{U}$, as described earlier. Thus $\mathcal{U}$ or $\mathcal{V}$ is globally hyperbolic, with a Cauchy hypersurface V that is an open subset of a Cauchy hypersurface S of M. There are many ways to embed $\mathcal{U}$ or $\mathcal{V}$ in some other closed universe M', such that V is an open subset of a Cauchy hypersurface S' of M'. For this, we simply modify S outside of V, or M outside of $\mathcal{U}$ or $\mathcal{V}$. However, $\mathscr{A}_{\mathcal{U}}$ and $\mathscr{A}_{\mathcal{V}}$ depend only on $\mathcal{U}$ and $\mathcal{V}$, and not on how $\mathcal{U}$ and $\mathcal{V}$ have been embedded in M. This has been called "the principle of locality" [6]. For free field theory, the statement is a theorem. In general, it is expected to hold because the algebras $\mathscr{A}_{\mathcal{U}}$ and $\mathscr{A}_{\mathcal{V}}$ are ultimately determined by operator product relations in the

regions $\mathcal{U}$ and $\mathcal{V}$, and these relations are entirely local in nature. The principle of locality is analogous to what we learned about spin systems in Sect. 11.2.2: there are many inequivalent representations of the algebra of an infinite collection of spins, but they are all equivalent for any finite set of spins.

Given the principle of locality, we can explain the meaning of quantum field theory in an open universe M', now with a noncompact initial value surface S'. Let $V \subset S'$ be an open set that is small enough that it can be embedded in a closed manifold $\widetilde{V}$. Let $\mathcal{U}$ be the domain of dependence of V in M', and $\mathcal{V} \subset \mathcal{U}$ a small thickening of V. Because V can be embedded in the compact manifold $\widetilde{V}$, $\mathcal{U}$ and $\mathcal{V}$ can be embedded in a closed universe $\widetilde{M}$, with $\widetilde{V}$ as Cauchy hypersurface. Then we can define algebras $\mathscr{A}_{\mathcal{U}}$ and $\mathscr{A}_{\mathcal{V}}$ associated to $\mathcal{U}$ and $\mathcal{V}$, and the principle of locality says that these local algebras depend only on $\mathcal{U}$ and $\mathcal{V}$ and not on the embedding in M'. The dynamical principle of a quantum field theory is now again $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$. This, along with general axioms about the algebras associated to open sets, is the content of quantum field theory in an open universe.

The principle of locality sheds light on what happens if we prefer to describe physics in an open universe using a Hilbert space. The only problem with doing so is that in an open universe there are many inequivalent choices of a Hilbert space and generically there is no preferred way to choose one. However, the principle of locality says that the algebra $\mathscr{A}_{\mathcal{U}}$ of a local region and the dynamical principle $\mathscr{A}_{\mathcal{U}} = \mathscr{A}_{\mathcal{V}}$ do not depend on the choice of a Hilbert space. Because of the principle of locality, an observer interested in physics in a bounded portion of the universe may decide that the choice of a global Hilbert space from among the myriad possibilities is irrelevant. But one can pick whichever Hilbert space one wishes without changing the predictions of a theory for local dynamics.

In an open universe that obeys special asymptotic conditions at spatial infinity, one can say more. One particularly important case is an asymptotically flat universe, that is a universe that is asymptotic at spatial infinity to Minkowski space. In the asymptotically flat case, we will assume that the theory of interest has a mass gap, for a reason explained in the next paragraph. Another important case is an asymptotically Anti de Sitter (AAdS) universe, that is, a universe that is asymptotic at spatial infinity to Anti de Sitter space. In the AAdS case, one assumes a boundary condition at spatial infinity of the sort usually considered in the AdS/CFT correspondence. An asymptotically flat or AAdS spacetime is time-independent for modes near spatial infinity, so one has a natural separation in positive and negative frequencies for modes near spatial infinity. For any mode of sufficiently short wavelength, there is a natural separation in positive and negative frequencies. In short, for the modes that either have short wavelength or are located near spatial infinity—which means for all but finitely many modes—there is a natural asymptotic separation in positive and negative frequencies. Hence, an asymptotically flat or AAdS spacetime (with a mass gap in the asymptotically flat case) is similar to a closed universe: there is a natural treatment asymptotically for all but finitely many modes, and therefore we should expect that a natural Hilbert space will exist.

Clearly, the argument as stated in the last paragraph is only heuristic, and I am not aware of precise theorems in the literature. For a quantum field theory with a mass gap m, the effects of spacetime curvature on the ground state are exponentially small in the limit that the radius of curvature is large compared to $1/m$. In such a case, in an asymptotically flat spacetime, the separation in positive and negative frequencies becomes exponentially precise near spatial infinity, where the radius of curvature diverges, so there should be no difficulty. In AAdS space with the usual sort of boundary conditions, there are no infrared issues and this case should also be safe. The delicate case is a massless theory in an asymptotically flat spacetime.[10] For a free Maxwell field in Minkowski space, there is a natural quantization that leads to a Hilbert space with a Poincaré invariant vacuum, but this quantization is not truly unique because of the possibility of "soft hair" [29]; after coupling to charged fields, usually assumed to be massive, the soft hair plays an important role in understanding infrared divergences, soft theorems, and the memory effect [30]. With massless charged fields, the non-uniqueness of quantization becomes truly essential; moreover this has a very interesting analog for General Relativity in an asymptotically flat spacetime [31]. One cannot expect quantum field theory in an asymptotically flat spacetime to be simpler than quantum field theory in Minkowski space. So in general, a straightforward Hilbert space description of quantum field theory in asymptotically flat spacetime is only available for massive theories or possibly for massless theories whose infrared behavior is such that the S-matrix can be defined in a conventional Fock space.

11.2.6 Non-Free Theories

So far we have discussed free theories. We will add a few words on non-free theories.

One of the main criteria for successful quantization of a free theory is that correlation functions such as the two-point function $\langle \Psi | \phi(\mathbf{x}, t)\phi(\mathbf{x}', t') | \chi \rangle$ should have standard short distance singularities. A two-point function with that property is the essential input to perturbation theory. So one would expect, modulo some well-understood issues of anomaly cancellation,[11] that weakly coupled theories make sense in perturbation theory when free theories do.

One would anticipate that the same is true nonperturbatively, for theories such as QCD. If a theory exists perturbatively in curved spacetime, and nonperturbatively in flat spacetime, one would expect that it works nonperturbatively in curved spacetime. Unfortunately, not much is available in terms of rigorous theorems, except for special models like two-dimensional conformal field theories. That

[10] The following was explained to me in this context by R. Wald.

[11] A theory which is well-defined in Minkowski space might have a gravitational anomaly which would cause perturbation theory in curved spacetime to fail.

reflects the general mathematical difficulty of understanding quantum field theory rigorously. One would think that rigorous results for a superrenormalizable theory in curved spacetime might be relatively accessible, but such results are not available.

11.3 Quantum Statistical Mechanics and the Thermodynamic Limit

11.3.1 The Thermofield Double

In this section, we will consider a simpler problem that is actually somewhat analogous to quantum field theory in an open universe. Instead of quantum field theory in curved spacetime, we consider quantum statistical mechanics at positive temperature $T = 1/\beta$. In finite volume, there is no problem. We just consider the density matrix $\rho = \frac{1}{Z}e^{-\beta H}$ acting on the Hilbert space $\mathcal{H}$ of the system, where $Z = \mathrm{Tr}\, e^{-\beta H}$ is the partition function. Importantly, for any conventional statistical system (possibly a relativistic quantum field theory, possibly a nonrelativistic system of some kind), the partition function converges in finite volume. Thermodynamic functions and thermal correlation functions can all be computed in finite volume, and they have a thermodynamic or large volume limit.

What happens if one wants to describe the thermodynamic limit of a system directly in terms of operators acting on a Hilbert space appropriate for an infinite volume system? Here we run into a difficulty: there is no separable Hilbert space that contains all the typical states of a system at $T \neq 0$, since fluctuations can occur throughout the whole infinite space. For example, for a spin system on an infinite lattice, since each spin has a nonzero probability to fluctuate, it takes uncountably many states to describe all the typical states at $T \neq 0$. (We discuss some examples based on lattice spin systems in Sect. 11.3.3.) In a continuum field theory, there is actually no reasonable definition of a nonseparable Hilbert space that contains all the typical thermal excitations in infinite volume.

There is a simple device that enables one to describe the thermodynamic limit of any quantum system in a separable Hilbert space [20]. This is the "thermofield double." Consider an ordinary quantum system $\mathcal{T}$ (possibly a relativistic field theory, possibly a lattice theory or something else) with Hilbert space $\mathcal{H}$ and Hamiltonian H. In finite volume, the thermofield double is defined by introducing, roughly speaking, two copies of the original system, which we will call the "right" and "left" copies, with Hilbert spaces $\mathcal{H}_r$ and $\mathcal{H}_\ell$, and Hamiltonians H_r and H_ℓ. We think of $\mathcal{H}_r$ as the Hilbert space of the system $\mathcal{T}$ and $\mathcal{H}_\ell$ as an auxiliary second copy. Actually, rather than being a second copy of $\mathcal{H}_r$, it is more canonical to define $\mathcal{H}_\ell$ to be the complex conjugate Hilbert space of $\mathcal{H}_r$. That amounts physically to saying that the left system is a time-reversal conjugate of the right system (rather than being an identical second copy). The fact that $\mathcal{H}_\ell$ is the complex conjugate of $\mathcal{H}_r$ ensures that the combined Hilbert space $\mathcal{H}_{\mathrm{TFD}} = \mathcal{H}_\ell \otimes \mathcal{H}_r$ of the two systems

can be viewed canonically as the space of linear operators acting on $\mathcal{H}_r$ (or $\mathcal{H}_\ell$). In expositions of the thermofield double, time-reversal symmetry is often assumed, in which case the left and right systems are equivalent.

In finite volume, the thermofield double $\Psi_{\rm TFD}$ is simply a vector in $\mathcal{H}_\ell \otimes \mathcal{H}_r$, defined as follows. Suppose that the Hamiltonian acting on the original system has eigenstates ψ_i with energies E_i. Then in finite volume, $\Psi_{\rm TFD}$ is defined as

$$\Psi_{\rm TFD} = \frac{1}{\sqrt{Z}} \sum_i \exp(-\beta E_i/2) |\bar{i}\rangle_\ell \otimes |i\rangle_r, \tag{11.12}$$

Here $|i\rangle_r$ is the ith energy level of the right system, with energy E_i; $|\bar{i}\rangle_\ell$ is its time-reversal conjugate in the left system; and Z is the partition function of the right or left system at temperature $1/\beta$. The point of this formula is that the density matrix of the right or left system is the usual thermal density matrix of that system:

$$\rho_r = {\rm Tr}_{\mathcal{H}_\ell} |\Psi_{\rm TFD}\rangle\langle\Psi_{\rm TFD}| = \frac{1}{Z} \sum_i e^{-\beta E_i} |i\rangle_r \langle i|_r = \frac{1}{Z} e^{-\beta H_r}. \tag{11.13}$$

So the thermofield double is a "purification" of a thermal density matrix.

An equivalent statement is the following. If we view $\mathcal{H}_\ell \otimes \mathcal{H}_r$ as the space of linear operators acting on $\mathcal{H}_r$, then $\Psi_{\rm TFD} = \rho_r^{1/2}$. To exploit this fact, let $\mathscr{A}_{r,V}$ and $\mathscr{A}_{\ell,V}$ be the algebras of all operators acting on $\mathcal{H}_r$ or $\mathcal{H}_\ell$, respectively, in finite volume V. Viewing a state vector $\Psi \in \mathcal{H}_\ell \otimes \mathcal{H}_r$ as a matrix acting on $\mathcal{H}_r$, an element $\mathsf{a}_r \in \mathscr{A}_{r,V}$ acts on Ψ by $\Psi \to \mathsf{a}_r \Psi$, and an element $\mathsf{a}_\ell \in \mathscr{A}_{\ell,V}$ acts on Ψ by $\Psi \to \Psi \mathsf{a}_\ell^{\rm tr}$ (where $\mathsf{a}_\ell^{\rm tr}$ is the transpose of a_ℓ). If Ψ, viewed as a matrix acting on $\mathcal{H}_r$, is invertible, then every matrix acting on $\mathcal{H}_r$ is of the form $\mathsf{a}_r \Psi$ and of the form $\Psi \mathsf{a}_\ell^{\rm tr}$ for some unique $\mathsf{a}_r \in \mathscr{A}_{r,V}$ and for some unique $\mathsf{a}_\ell \in \mathscr{A}_{\ell,V}$. In particular, $\Psi_{\rm TFD} = \rho_r^{1/2}$ is invertible, and therefore every vector in $\mathcal{H}_{\rm TFD}$ can be obtained from $\Psi_{\rm TFD}$ by acting with $\mathscr{A}_{r,V}$, or by acting with $\mathscr{A}_{\ell,V}$. The fancy way to say this is that if Ψ is invertible, then it corresponds to a cyclic separating vector[12] for the algebras $\mathscr{A}_{r,V}$ and $\mathscr{A}_{\ell,V}$.

We want to take the infinite volume limit of this construction. As already noted, the left and right Hilbert spaces $\mathcal{H}_r$ and $\mathcal{H}_\ell$ do not have large volume limits, and the formula (11.12) is not meaningful in the thermodynamic limit, because $Z = \infty$ in this limit and the sum would contain uncountably many terms. But it is possible to define a Hilbert space $\mathcal{H}_{\rm TFD}$ that contains $\Psi_{\rm TFD}$ and does have a large volume limit. To explain this, we first review a few standard facts about finite volume. We describe these facts in the familiar language of path integrals, though this is not strictly necessary. Let W be the spatial manifold (or lattice) on which we want to

[12] In general, if an algebra $\mathscr{A}$ acts on a Hilbert space $\mathcal{H}$, then a vector $\Psi \in \mathcal{H}$ is said to be separating for $\mathscr{A}$ if the condition $\mathsf{a}\Psi = 0$, $\mathsf{a} \in \mathscr{A}$ implies that $\mathsf{a} = 0$, and it is said to be cyclic for $\mathscr{A}$ if vectors $\mathsf{a}\Psi$, $\mathsf{a} \in \mathscr{A}$ are dense in $\mathcal{H}$.

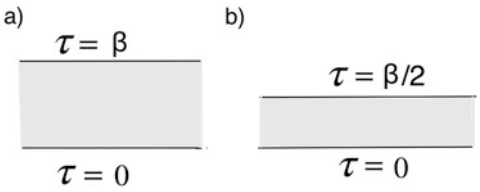


Fig. 11.2 (**a**) A thermal density matrix $\rho = \frac{1}{Z}e^{-\beta H}$ can be computed by a path integral on the strip $0 \leq \tau \leq \beta$. The horizontal direction in the picture represents the "spatial" manifold (or lattice) W. (**b**) The thermofield double state $\Psi_{\rm TFD} = \rho^{1/2}$ can be computed by a path integral on a strip of half the width, $0 \leq \tau \leq \beta/2$. Operators of the "right" or "left" copy are inserted, respectively, on the boundaries at $\tau = 0$ and $\tau = \beta/2$

formulate theory $\mathcal{T}$. In the thermodynamic limit, W will be $\mathbb{R}^{D-1}$ for a continuum field theory or an infinite lattice $\mathbb{Z}^{D-1}$ for a lattice system. The thermal density matrix $\rho_r = \frac{1}{Z}e^{-\beta H}$ can be computed in finite volume by a path integral on $W \times I$, where I is the interval $0 \leq \tau \leq \beta$ (Fig. 11.2) The factor of $1/Z$ means that from the energy density T_{00}, whose integral is the Hamiltonian $H_r = \int {\rm d}^{D-1}x\sqrt{g}T_{00}$, we should subtract a constant, the free energy density at temperature $1/\beta$. With this subtraction, the path integral, in finite volume, computes a normalized density matrix ρ_r, satisfying ${\rm Tr}\,\rho_r = 1$. To construct the thermofield double state $\Psi_{\rm TFD} = \rho_r^{1/2}$, we simply do the same path integral (with the same subtraction) on a strip of half the width, $0 \leq \tau \leq \beta/2$. Given an operator $\mathsf{a}_r \in \mathscr{A}_{r,V}$, to construct a state obtained by acting with a_r on $\Psi_{\rm TFD}$, we just insert a_r in the path integral at $\tau = 0$; similarly, to act with $\mathsf{a}_\ell \in \mathscr{A}_{\ell,V}$ on $\Psi_{\rm TFD}$, we insert $\mathsf{a}_\ell^{\rm tr}$ at $\tau = \beta/2$. Inner products between these states are computed by gluing together two path integrals on the strip of width $\beta/2$ to make a cylinder of circumference β. For example, an inner product $\langle \mathsf{b}_r\Psi_{\rm TFD}|\mathsf{a}_r\Psi_{\rm TFD}\rangle = \langle\Psi_{\rm TFD}|\mathsf{b}_r^\dagger\mathsf{a}_r\Psi_{\rm TFD}\rangle$ is computed by a path integral on the cylinder with insertion of $\mathsf{b}_r^\dagger\mathsf{a}_r$ at $\tau = 0$. This quantity is just the expectation value of $\mathsf{b}_r^\dagger\mathsf{a}_r$ in the thermal ensemble:

$$\langle \mathsf{b}_r\Psi_{\rm TFD}|\mathsf{a}_r\Psi_{\rm TFD}\rangle = \langle\Psi_{\rm TFD}|\mathsf{b}_r^\dagger\mathsf{a}_r|\Psi_{\rm TFD}\rangle = {\rm Tr}\,\rho_r\mathsf{b}_r^\dagger\mathsf{a}_r = \frac{1}{Z}{\rm Tr}_{\mathcal{H}_r}\exp(-\beta H_r)\mathsf{b}_r^\dagger\mathsf{a}_r. \tag{11.14}$$

Such thermal expectation values have a thermodynamic limit.[13] Since the density operator ρ_r is strictly positive, one has

$$\langle \mathsf{a}_r\Psi_{\rm TFD}|\mathsf{a}_r\Psi_{\rm TFD}\rangle = {\rm Tr}_{\mathcal{H}_r}\rho_r\mathsf{a}_r^\dagger\mathsf{a}_r > 0 \tag{11.15}$$

[13] In taking the thermodynamic limit, we keep the operators fixed, and hence acting on degrees of freedom in a bounded region, while the volume goes to infinity.

for all $\mathsf{a}_r \neq 0$. Likewise, for $\mathsf{a}_\ell \neq 0$, we can construct states $\mathsf{a}_\ell \Psi_{\rm TFD}$ by acting with $\mathsf{a}_\ell^{\rm tr}$ at $\tau = \beta/2$. Inner products between these states are obviously given by a formula like Eq. (11.14) which again has a thermodynamic limit:

$$\langle \mathsf{b}_\ell \Psi_{\rm TFD} | \mathsf{a}_\ell \Psi_{\rm TFD} \rangle = \frac{1}{Z} {\rm Tr}_{\mathcal{H}_\ell} \exp(-\beta H_\ell) \mathsf{b}_\ell^\dagger \mathsf{a}_\ell . \tag{11.16}$$

Inner products $\langle \mathsf{a}_\ell \Psi_{\rm TFD} | \mathsf{a}_r \Psi_{\rm TFD} \rangle$ are defined by inserting the operators a_r and $(\mathsf{a}_\ell^{\rm tr})^\dagger = \mathsf{a}_\ell^*$ (here a_ℓ^* is the complex conjugate of a_ℓ; note that because of the complex conjugation, it is naturally an operator on $\mathcal{H}_r$) at $\beta = 0$ and $\beta = \tau/2$, respectively. In operator terms, such an inner product is

$$\langle \mathsf{a}_\ell \Psi_{\rm TFD} | \mathsf{a}_r \Psi_{\rm TFD} \rangle = \frac{1}{Z} {\rm Tr}_{\mathcal{H}_r} \mathsf{a}_r e^{-\beta H_r/2} \mathsf{a}_\ell^* e^{-\beta H_r/2} . \tag{11.17}$$

Such an inner product also has a thermodynamic limit. Reflection positivity says that for any a_ℓ, there is an a_r that makes this inner product nonzero. We simply choose a_r to be obtained from a_ℓ by a reflection of the thermal circle.

In finite volume, the states that appeared in this construction are simply states in $\mathcal{H}_\ell \otimes \mathcal{H}_r$. In the infinite volume limit, this is no longer true. But now we can describe a Hilbert space $\mathcal{H}_{\rm TFD}$ that does have a thermodynamic limit. First we need to pick an algebra of operators that is supposed to act on $\mathcal{H}_{\rm TFD}$. As in Sect. 11.2.2, we at least want to include all operators that act on a bounded region of space. Ultimately we can take limits and include operators whose action is not restricted to a bounded region, but to know what limits to allow, we first have to know what Hilbert space the operators are supposed to act on. So we define algebras $\mathscr{A}_{r,0}$ and $\mathscr{A}_{\ell,0}$ consisting of bounded operators acting on a bounded region in the right system or the left system, respectively. Now at a minimum we want $\mathcal{H}_{\rm TFD}$ to contain a vector that we call $\Psi_{\rm TFD}$ which will be an infinite volume limit of the finite volume thermofield double state. In addition, we want the algebras $\mathscr{A}_{r,0}$ and $\mathscr{A}_{\ell,0}$ to act on $\mathcal{H}_{\rm TFD}$. So we say that for every $\mathsf{a}_r \in \mathscr{A}_{r,0}$, $\mathcal{H}_{\rm TFD}$ contains a vector that we call $\mathsf{a}_r \Psi_{\rm TFD}$. $\mathscr{A}_{r,0}$ acts on these states in an obvious way: $\mathsf{a}_r(\mathsf{b}_r \Psi_{\rm TFD}) = (\mathsf{a}_r \mathsf{b}_r)\Psi_{\rm TFD}$. We postulate that the inner products between these states are given by the thermodynamic limit of the thermal correlation functions in Eq. (11.14). Because of Eq. (11.15), this set of states makes up a pre-Hilbert space $\mathcal{H}_{{\rm TFD},0}$, satisfying all conditions of a Hilbert space except completeness. We take the Hilbert space completion and get the thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$. We can now also complete $\mathscr{A}_{r,0}$ to a von Neumann algebra $\mathscr{A}_r$ that acts on $\mathcal{H}_{\rm TFD}$ by including weak limits, as in Sect. 11.2.2.

Obviously, we could have carried out this process starting with $\mathscr{A}_{\ell,0}$ instead of $\mathscr{A}_{r,0}$. One might worry that this would lead to a different thermofield double Hilbert space $\widetilde{\mathcal{H}}_{\rm TFD}$, with an action of a left algebra $\mathscr{A}_\ell$ instead of $\mathscr{A}_r$. The reason that this does not happen is that the formula Eq. (11.17) for inner products between states $\mathsf{a}_\ell \Psi_{\rm TFD}$ and states $\mathsf{a}_r \Psi_{\rm TFD}$ also has a thermodynamic limit and implies that $\mathsf{a}_\ell \Psi_{\rm TFD}$ can be viewed as a vector in the same Hilbert space $\mathcal{H}_{\rm TFD}$ that we defined starting with states $\mathsf{a}_r \Psi_{\rm TFD}$.

The upshot is to define a Hilbert space $\mathcal{H}_{\rm TFD}$ with an action of completions $\mathscr{A}_\ell$ and $\mathscr{A}_r$ of $\mathscr{A}_{\ell,0}$ and $\mathscr{A}_{r,0}$. Inner products between vectors of the form $\mathsf{a}_\ell\Psi_{\rm TFD}$ and/or $\mathsf{a}_r\Psi_{\rm TFD}$ are defined via the thermodynamic limits of Eqs. (11.14), (11.16), and (11.17). All expected relations among inner products between Hilbert space vectors are satisfied, because we are simply taking the thermodynamic limits of formulas that in finite volume do represent inner products between Hilbert space vectors. Moreover, in terminology that was defined in footnote 12, $\Psi_{\rm TFD}\in\mathcal{H}_{\rm TFD}$ is always a cyclic separating vector for the algebra $\mathscr{A}_r$ (or $\mathscr{A}_\ell$).

The procedure by which we constructed the thermodynamic limit of $\mathcal{H}_{\rm TFD}$ is a special case of the Gelfand-Naimark-Segal (GNS) construction in which the input is an algebra $\mathscr{A}$ and a state on the algebra, that is a complex-valued linear function $\mathsf{a}\to F(\mathsf{a})$ such that $F(1)=1$ and $F(\mathsf{a}^\dagger\mathsf{a})>0$ for all $\mathsf{a}\neq 0$, and the output is a Hilbert space $\mathcal{H}$ with an action of $\mathscr{A}$ and a cyclic separating vector Ψ. (Closely related is the Wightman Reconstruction Theorem of axiomatic field theory, in which a Hilbert space is reconstructed from the correlation functions of a local field.) The GNS construction proceeds as follows. One formally associates a vector Ψ to the identity element $1\in\mathscr{A}$, and to every $\mathsf{a}\in\mathscr{A}$ one associates a vector $\mathsf{a}\Psi$. The action of $\mathscr{A}$ on this set of states is defined in the obvious way by $\mathsf{a}(\mathsf{b}\Psi)=(\mathsf{ab})\Psi$, and inner products are defined by $\langle\mathsf{a}\Psi|\mathsf{b}\Psi\rangle=F(\mathsf{a}^\dagger\mathsf{b})$. This gives a pre-Hilbert space $\mathcal{H}_0$ with an action of $\mathscr{A}$, and one completes it to a Hilbert space $\mathcal{H}$. From the definitions, it is immediate that $\mathscr{A}$ acts on $\mathcal{H}$ with Ψ as a cyclic separating vector. Clearly the definition of $\mathcal{H}_{\rm TFD}$ is the GNS construction applied to the case that $F(\mathsf{a})$ is the expectation value of a in the thermodynamic limit.

11.3.2 Surprises in the Thermodynamic Limit

So by going to the thermofield double, we can describe the infinite volume limit of quantum statistical mechanics in a separable Hilbert space $\mathcal{H}_{\rm TFD}$. But there are a few surprises.

First of all, evidently $\mathscr{A}_r$ does not act irreducibly on $\mathcal{H}_{\rm TFD}$, since it commutes with $\mathscr{A}_\ell$, and vice-versa. In fact $\mathscr{A}_\ell$ and $\mathscr{A}_r$ are algebras of a possibly unfamiliar kind: they are generically von Neumann algebras of Type III. What characterizes algebras of Type II or Type III, as opposed to the more familiar algebras of Type I, is roughly that entanglement is not just a property of the states that the algebra acts on but is built into the structure of the algebra. In finite volume V, the entanglement entropy of the state $\Psi_{\rm TFD}\in\mathcal{H}_\ell\otimes\mathcal{H}_r$ is the same as the thermodynamic entropy of the thermal density matrix ρ_r or ρ_ℓ on one factor. In particular, it is proportional to V and diverges in the thermodynamic limit. Thus in effect $\Psi_{\rm TFD}$ in the thermodynamic limit describes a state with an infinite entanglement between the two systems. But every state in $\mathcal{H}_{\rm TFD}$ looks at spatial infinity like $\Psi_{\rm TFD}$ and so every state has the same divergent entanglement entropy. This is different from the usual situation for a bipartite quantum system XY with Hilbert space $\mathcal{H}_{XY}=\mathcal{H}_X\otimes\mathcal{H}_Y$ where $\mathcal{H}_X$ and $\mathcal{H}_Y$ are of infinite dimension. In such a case, a vector $\Psi_{XY}\in\mathcal{H}_{XY}$ *could*

have infinite entanglement entropy, but there are also vectors in $\mathcal{H}_{XY}$ with finite or zero entanglement entropy. By contrast, in the infinite volume limit, $\mathcal{H}_{\mathrm{TFD}}$ is not a tensor product and no vector in it can be interpreted to have finite entanglement entropy; rather, the infinite entanglement entropy between the left and right systems is encoded in the structure of the algebras $\mathscr{A}_\ell$ and $\mathscr{A}_r$. We will discuss this more thoroughly in Sect. 11.3.3 with simple examples.

A second unusual property of the infinite volume thermofield double comes to light if we think about real time thermal physics. In finite volume, to study real time thermal physics, we define the time-dependence of operators in a familiar way:

$$\begin{aligned} \mathsf{a}_r(t) &= \exp(\mathrm{i}H_r t)\mathsf{a}_r \exp(-\mathrm{i}H_r t) \\ \mathsf{a}_\ell(t) &= \exp(\mathrm{i}H_\ell t)\mathsf{a}_\ell \exp(-\mathrm{i}H_\ell t). \end{aligned} \tag{11.18}$$

Then we define real time correlation functions such as

$$\langle \mathsf{a}_r(t_1)\mathsf{b}_r(t_2)\rangle = \mathrm{Tr}\,\rho_r \mathsf{a}_r(t_1)\mathsf{b}_r(t_2). \tag{11.19}$$

These functions have a thermodynamic limit. How do we describe them using the thermofield double? A subtlety arises here: the Hamiltonians H_ℓ and H_r of the left and right system do not make sense as operators on the infinite volume $\mathcal{H}_{\mathrm{TFD}}$. This is so because of thermal fluctuations. In volume V, the typical energy in the thermal ensemble is of course proportional to V, so the Hamiltonian diverges for $V \to \infty$. That in itself is not a problem, because we can "renormalize" H_ℓ and H_r by subtracting a multiple of V to get an operator whose expectation value vanishes for $V \to \infty$. The problem comes from the thermal fluctuations. The typical fluctuations in H_ℓ and H_r in volume V are of order $V^{1/2}$, which also diverges for $V \to \infty$. Even after subtracting constants from H_ℓ and H_r to set their expectations values in the thermal ensemble to zero, the divergent fluctuations remain, and imply that H_ℓ and H_r do not have limits as operators on $\mathcal{H}_{\mathrm{TFD}}$.

So how are we going to define real time correlation functions? The answer is that the operator $\widehat{H} = H_r - H_\ell$ does have a thermodynamic limit as an operator on $\mathcal{H}_{\mathrm{TFD}}$. This results from the fact that in finite volume $\widehat{H}$ annihilates Ψ_{TFD}, as is evident from the definition (11.12). With the aid of $\widehat{H}$, we can describe real time thermal physics. We define

$$\begin{aligned} \mathsf{a}_r(t) &= \exp(\mathrm{i}\widehat{H}t)\mathsf{a}_r \exp(-\mathrm{i}\widehat{H}t) \\ \mathsf{a}_\ell(t) &= \exp(-\mathrm{i}\widehat{H}t)\mathsf{a}_\ell \exp(\mathrm{i}\widehat{H}t). \end{aligned} \tag{11.20}$$

The motivation for these definitions is that in finite volume they agree with the standard definitions (11.18), since $[H_\ell, \mathsf{a}_r] = [H_r, \mathsf{a}_\ell] = 0$. In terms of these definitions, the thermodynamic limit of real time correlation functions is given by formulas such as

$$\langle \mathsf{a}_{\ell,r}(t_1)\mathsf{b}_{\ell,r}(t_2)\rangle = \langle \Psi_{\mathrm{TFD}}|a_{\ell,r}(t_1)b_{\ell,r}(t_2)|\Psi_{\mathrm{TFD}}\rangle, \tag{11.21}$$

where the subscripts ℓ, r mean that the operators can belong to either $\mathscr{A}_\ell$ or $\mathscr{A}_r$.

Since the operator $\exp(\mathrm{i}\widehat{H}t)$ is not contained in either $\mathscr{A}_\ell$ or $\mathscr{A}_r$, it follows that time translations are a group of outer automorphisms of $\mathscr{A}_\ell$ and of $\mathscr{A}_r$. Now consider the following question. Suppose that one has access to the Hilbert space $\mathcal{H}_{\rm TFD}$, the state $\Psi_{\rm TFD}$, and the algebra $\mathscr{A}_r$, but one does not know how this data arose by taking a thermodynamic limit. Just from $\mathcal{H}_{\rm TFD}$, $\Psi_{\rm TFD}$, and $\mathscr{A}_r$, how would one identify which outer automorphism group of $\mathscr{A}_r$ corresponds to time translations? The answer to this question is given by Tomita-Takesaki theory. Here we will just state a few facts, leaving the reader to look elsewhere (for example, sections 3 and 4 of [19]) for a detailed explanation. In Tomita-Takesaki theory, to a Hilbert space $\mathcal{H}$, an algebra $\mathscr{A}$ of operators on $\mathcal{H}$ and a state $\Psi \in \mathcal{H}$ that is cyclic and separating[14] for $\mathscr{A}$, one associates a modular operator Δ_Ψ. For example, if $\mathcal{H} = \mathcal{H}_{XY} = \mathcal{H}_X \otimes \mathcal{H}_Y$ is the Hilbert space of a bipartite system XY, $\mathscr{A} = \mathscr{A}_X$ is the algebra of operators on $\mathcal{H}_X$, and $\Psi = \Psi_{XY}$ is a vector in $\mathcal{H}_{XY}$, then the cyclic separating condition is that the reduced density matrices ρ_X and ρ_Y of Ψ_{XY} are invertible, and in this case the modular operator is $\Delta_\Psi = \rho_X \otimes \rho_Y^{-1}$. (See Eq. (4.26) in [19], for example.) In finite volume, these conditions are satisfied by $\mathcal{H} = \mathcal{H}_{\rm TFD} = \mathcal{H}_\ell \otimes \mathcal{H}_r$, $\Psi = \Psi_{\rm TFD}$, $\mathscr{A} = \mathscr{A}_r$, $\rho_r = \frac{1}{Z}e^{-\beta H_r}$, $\rho_\ell = \frac{1}{Z}e^{-\beta H_\ell}$. So $\Delta_{\Psi_{\rm TFD}} = \rho_r \otimes \rho_\ell^{-1} = \exp(-\beta\widehat{H})$. Though ρ_ℓ and ρ_r do not have a thermodynamic limit, $\Delta_{\Psi_{\rm TFD}}$ and $\widehat{H}$ do have such a limit and the formula $\Delta_{\Psi_{\rm TFD}} = \exp(-\beta\widehat{H})$ is valid in the thermodynamic limit. So we can express $\widehat{H}$ in terms of the modular operator: $\widehat{H} = -\frac{1}{\beta}\log\Delta_{\Psi_{\rm TFD}}$.

11.3.3 Examples from Spin Systems

Here we will consider some concrete examples of the thermofield double construction for a system with an infinite number of degrees of freedom.[15] As in Sect. 11.2.2, we consider a countably infinite set of qubits, which we label by a positive integer n, making the "right" system. To study this system via the thermofield double, we introduce another infinite collection of qubits, the "left" system. We write $\mathcal{H}_{r,N}$ for the Hilbert space of the first N qubits of the right system, and $\mathcal{H}_{\ell,N}$ for the analog for the left system.

We will consider extremely simple Hamiltonians of the form $H = \sum_{n=1}^{\infty} H_n$, where H_n is a single-qubit Hamiltonian. One may be surprised that this case is rich enough to be interesting. In fact the various von Neumann algebras[16] can all arise

[14] See footnote 12 for the meaning of "cyclic and separating."

[15] Some of the following was described from a different point of view in section 6 of [19].

[16] To be more precise, the "hyperfinite" von Neuman algebras all arise this way. A hyperfinite algebra is one that can be approximated by a finite-dimensional matrix algebra. Hyperfinite algebras are usually the ones that arise in physics, because infinite constructions in physics are usually limits of finite constructions. The classification of von Neumann algebras is much more complicated in the non-hyperfinite case.

from such a construction and some of them were first discovered in this way [32–34].

The simplest Hamiltonian of all is $H = 0$, so we will start with that case. Let us write $|i\rangle_{n,r}$, $i = 1, 2$ for the two states of the nth qubit of the "right" system, and $|\bar{i}\rangle_{n,\ell}$, $i = 1, 2$ for the time-reversed states of the nth qubit of the "left" system. The analog of defining the thermofield double state first in finite volume, as in Eq. (11.12), is to define the thermofield double state first for the first N qubits of the right system, together with their partners in the left system. With $H = 0$, the thermofield double state $\Psi_{\mathrm{TFD}} \in \mathcal{H}_{r,N} \otimes \mathcal{H}_{\ell,N}$ is just a completely entangled state of N pairs of qubits:

$$\Psi_{\mathrm{TFD}} = \frac{1}{\sqrt{Z_N}} \bigotimes_{n=1}^{N} \sum_{i=1,2} |i\rangle_{n,r} \otimes |\bar{i}\rangle_{n,\ell}, \quad Z_N = 2^N. \tag{11.22}$$

Arranging the four states $|i\rangle_{n,r} \otimes |j,\rangle_{n,\ell}$, $i, j = 1, 2$ in a 2×2 matrix in a fairly obvious way, we can write

$$\Psi_{\mathrm{TFD}} = \frac{1}{\sqrt{Z_N}} \bigotimes_{n=1}^{N} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{11.23}$$

Now let $\mathscr{A}_{r,0}$ be the algebra of all operators on the right system that act nontrivially on only finitely many qubits. If $\mathsf{a}_r \in \mathscr{A}_{r,0}$, then for sufficiently large N, we can view a_r as an operator on $\mathcal{H}_{r,N} \otimes \mathcal{H}_{\ell,N}$ and define $F(\mathsf{a}_r) = \langle \Psi_{\mathrm{TFD}} | \mathsf{a}_r | \Psi_{\mathrm{TFD}} \rangle$. For given a_r, this definition does not depend on N, once N is large enough so that $F(\mathsf{a}_r)$ is defined. So $F(\mathsf{a}_r)$ is well-defined for all $\mathsf{a}_r \in \mathscr{A}_{r,0}$. It clearly is linear in a_r and satisfies $F(\mathsf{a}_r^\dagger \mathsf{a}_r) > 0$ for $\mathsf{a}_r \neq 0$, so it defines a state on the algebra $\mathscr{A}_{r,0}$. So we can invoke the GNS construction, as described in Sect. 11.3.1, and construct the thermofield double Hilbert space $\mathcal{H}_{\mathrm{TFD}}$, with an action of $\mathscr{A}_{r,0}$, and containing the distinguished thermofield double state $\Psi_{\mathrm{TFD}} \in \mathcal{H}_{\mathrm{TFD}}$.

By including weak limits, we can take the closure of $\mathscr{A}_{r,0}$ to get an algebrs $\mathscr{A}_r$ of bounded operators that acts on $\mathcal{H}_{\mathrm{TFD}}$. This is an algebra of an unfamiliar kind—a Type II von Neumann algebra. In fact, it is a Murray-von Neumann factor[17] of Type II_1 [35]. To see that the algebra $\mathscr{A}_r$ is of an unfamiliar type, note that the function $F(\mathsf{a}_r)$ satisfies

$$F(\mathsf{a}_r \mathsf{b}_r) = F(\mathsf{b}_r \mathsf{a}_r). \tag{11.24}$$

Thus, this function has the algebraic properties of a trace, so we will denote it as one: $F(\mathsf{a}_r) = \mathrm{Tr}\, \mathsf{a}_r$. Moreover, this trace is defined (and finite) for all elements of the algebra; in particular, $\mathrm{Tr}\, 1 = 1$. By contrast, the obvious example of an infinite

[17] A factor is a von Neumann algebra whose center consists only of the complex scalars.

dimensional von Neumann algebra is the algebra $\mathscr{A}$ of all bounded operators on a separable Hilbert space $\mathcal{H}$. This algebra, which is said to be of Type I_∞, does have a trace Tr, but the trace is not defined for all elements of $\mathscr{A}$ (only for those that are "trace class") and in particular in a Type I_∞ factor, $\mathrm{Tr}\, 1 = +\infty$. So the algebra $\mathscr{A}_r$ is something new.

Intuitively, since the Type II_1 factor $\mathscr{A}_r$ was constructed by considering an infinite collection of maximally entangled qubit pairs, an infinite amount of entanglement is built into the structure of this algebra. A Type II_1 algebra does not have an irreducible representation in a Hilbert space. We constructed a representation of the Type II_1 algebra $\mathscr{A}_r$ on the Hilbert space $\mathcal{H}_{\mathrm{TFD}}$, but this representation is far from irreducible, since the action of $\mathscr{A}_r$ on $\mathcal{H}_{\mathrm{TFD}}$ commutes with the action of another Type II_1 algebra $\mathscr{A}_\ell$ that is defined similarly, starting with operators that act only on finitely many qubits of the left system and taking limits. The algebras $\mathscr{A}_r$ and $\mathscr{A}_\ell$ commute and in fact they are each other's commutants.[18] As this example illustrates, when a Type II_1 algebra acts on a Hilbert space, its commutant is another Type II algebra.

Murray and von Neumann classified the Hilbert space representations of a hyperfinite Type II_1 algebra. They are labeled up to isomorphism by a regularized "dimension" x, which is a non-negative real number or ∞. $\mathcal{H}_{\mathrm{TFD}}$ corresponds to the case $x = 1$. To construct representations with $x < 1$, we use projection operators in $\mathscr{A}_\ell$. For example, for $1 \leq k \leq 2^N$, pick a k-dimensional subspace $W \subset \mathcal{H}_{\ell,N}$, and let $\Pi : \mathcal{H}_{\ell,N} \to W$ be the orthogonal projection operator of $\mathcal{H}_{\ell,N}$ onto W (tensored with the identity operator on all other qubits). The image of Π in $\mathcal{H}_{\mathrm{TFD}}$ is a representation of $\mathscr{A}_r$ with $x = k/2^N$. All values $x < 1$ are limits of this with $k, N \to \infty$. Since x is additive when one takes the direct sum of two representations, any representation is the direct sum of a representation with $x < 1$ and a number of copies of $\mathcal{H}_{\mathrm{TFD}}$. Roughly, the value of x for a Hilbert space representation $\mathcal{R}$ of $\mathscr{A}_r$ is defined as the trace of the identity element of $\mathscr{A}_r$ in the representation $\mathcal{R}$. Some more detail is needed to make this precise, but anyway the existence of inequivalent representations of $\mathscr{A}_r$ is related to the fact that $\mathscr{A}_r$ has a trace.

The other hyperfinite von Neumann algebra of Type II is a Type II_∞ algebra. It can be obtained as the tensor product of a Type II_1 algebra with a Type I_∞ algebra of all bounded operators on a separable Hilbert space. It has a trace, which, because of the Type I_∞ factor, is not defined for all elements of the algebra.

Repeating this construction with a non-zero Hamiltonian, we can get the other hyperfinite von Neumann algebras. Again with $H = \sum_{n=1}^{\infty} H_n$, the next simplest case is to take

$$H_n = \begin{pmatrix} 0 & 0 \\ 0 & E \end{pmatrix}, \tag{11.25}$$

[18] The commutant of a von Neumann algebra $\mathscr{A}$ acting on a Hilbert space $\mathcal{H}$ is the algebra $\mathscr{A}'$ that consists of all bounded operators on $\mathcal{H}$ that commute with $\mathscr{A}$.

all with the same constant E. The thermofield double state, in the notation of (11.23), is

$$\Psi_{\rm TFD} = \frac{1}{\sqrt{Z_N}} \bigotimes_{n=1}^{N} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta E/2} \end{pmatrix}, \quad Z_N = (1 + e^{-\beta E})^N. \tag{11.26}$$

In the large N limit, we can again define a state on the algebra $\mathscr{A}_{r,0}$ by $F(\mathsf{a}_r) = \langle \Psi_{\rm TFD} | \mathsf{a}_r | \Psi_{\rm TFD} \rangle$. The GNS construction gives the thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$, with an action on $\mathcal{H}_{\rm TFD}$ of an algebra $\mathscr{A}_r$ that is a closure of $\mathscr{A}_{r,0}$, and with a cyclic separating state $\Psi_{\rm TFD}$.

The state $\Psi_{\rm TFD}$ in this example describes an infinite collection of qubit pairs all with the same non-maximal entanglement. So $\mathscr{A}_r$ is again an algebra that encodes an infinite amount of entanglement. What is different for $E \neq 0$ is that generically $F(\mathsf{a}_r b_r) \neq F(\mathsf{b}_r \mathsf{a}_r)$. So the function F is no longer a trace, and indeed for $E \neq 0$, $\mathscr{A}_r$ does not have a trace. The algebra $\mathscr{A}_r$ is called an algebra of Type III_λ with $\lambda = e^{-\beta E/2}$. An algebra of this type was originally constructed in precisely the way that we have described [33].

In almost any real problem in quantum statistical mechanics, the algebra $\mathscr{A}_r$ that acts on the thermofield double is instead a von Neumann algebra of Type III_1. The original construction of such an algebra [34] was based on again taking $H = \sum_{n=1}^{\infty} H_n$ to be a sum of single-qubit Hamiltonians, but with different H_n:

$$H_n = \begin{pmatrix} 0 & 0 \\ 0 & E_n \end{pmatrix}. \tag{11.27}$$

In this general form, we will call the model the Araki-Woods model. As long as $\lim_{n\to\infty} E_n \neq \infty$, there is an infinite amount of entanglement between the left and right systems, and the algebra $\mathscr{A}_r$ that acts on $\mathcal{H}_{\rm TFD}$ will not be of Type I. (The case $\lim_{n\to\infty} E_n = \infty$ is discussed in Sect. 11.3.5.) Except in rather special cases, $\mathscr{A}_r$ is a new type of algebra that is said to be of Type III_1. For example, if the E_n take two values E and E', each with infinite multiplicity, then, for generic E and E', the algebra $\mathscr{A}_r$ is of Type III_1. This is again an algebra without a trace.

Unlike a Type II_1 algebra, a Type III algebra only has one irreducible representation in a Hilbert space, up to isomorphism. In the construction that we have just described, the representation $\mathcal{H}_{\rm TFD}$ of the Type III_1 algebra $\mathscr{A}_r$ is far from irreducible, since the commutant of $\mathscr{A}_r$ is another Type III_1 algebra $\mathscr{A}_\ell$ that can be defined in the same way. We can again choose a projection operator $\Pi \in \mathscr{A}_\ell$, and the image of Π is a subspace $\Pi\mathcal{H}_{\rm TFD} \subset \mathcal{H}_{\rm TFD}$ that provides a representation of $\mathscr{A}_r$. But as representations of $\mathscr{A}_r$, $\mathcal{H}_{\rm TFD}$ and $\Pi\mathcal{H}_{\rm TFD}$ are equivalent, even though one might naively think that the second one is smaller. For Type II_1, the existence of the invariant x that distinguishes representations of different "size" is tied to the existence of a trace; for Type III, there is no trace and any two nontrivial representations are isomorphic.

Following is a summary of some facts about (hyperfinite) von Neumann algebras. In all of these statements, we assume that the algebra considered is a "factor" (its center consists only of complex scalars). Algebras that are not factors can be constructed in a relatively simple way from factors.

(1) A Type I algebra $\mathcal{A}$ has an irreducible representation in a Hilbert space $\mathcal{R}$. Its commutant in this representation consists only of its center, the complex scalars. This representation is essentially unique in the sense that any representation of $\mathcal{A}$ is a direct sum of copies of $\mathcal{R}$; differently put, any representation is of the form $\mathcal{R}\otimes Q$, where Q is some other Hilbert space on which $\mathcal{A}$ acts trivially. $\mathcal{A}$ is said to be of Type I_n if $\mathcal{R}$ is n-dimensional, and of Type I_∞ if $\mathcal{R}$ has countably infinite dimension.

(2) A Type II factor $\mathcal{A}$ has no irreducible representation in a Hilbert space; its commutant in any representation is another Type II factor. There are two hyperfinite Type II algebras, up to isomorphism. A hyperfinite Type II_1 algebra has a trace defined for all elements of the algebra; its representations are labeled by a positive real number x (it is also possible to have a representation with $x=\infty$; this happens if the commutant of the Type II_1 algebra is of Type II_∞). A hyperfinite Type II_∞ algebra has a trace that is defined only for some elements of the algebra; when it acts on a Hilbert space, its commutant is of Type II_1 or Type II_∞, and thus it has two distinct kinds of Hilbert space representation.

(3) A Type III algebra $\mathcal{A}$ has no irreducible representation in a Hilbert space; its commutant in any representation is another algebra isomorphic to $\mathcal{A}$. A Type III algebra has no trace, and its representations are all isomorphic. In the thermodynamic limit, the natural algebra of observables of a generic system in quantum statistical mechanics is of Type III_1. The other hyperfinite Type III algebras are the Type III_λ algebras that we have described, and a Type III_0 algebra that is a kind of $\lambda\to 0$ limit of a Type III_λ algebra.

11.3.4 Relation to Quantum Field Theory

In Sect. 11.2, in order to understand quantum field theory in a curved spacetime M, especially in the case of an open universe, it was important to consider the algebra $\mathcal{A}_\mathcal{U}$ of observables in a local region $\mathcal{U}\subset M$. Now that we have some experience with von Neumann algebras, we can return to this subject and ask what kind of algebra is $\mathcal{A}_\mathcal{U}$.

We consider again the setup discussed in Sect. 11.2.5. We take S to be a Cauchy hypersurface in M, which we assume compact for simplicity, and B to be a codimension 1 submanifold of S whose complement consists of two open sets V and V' (Fig. 11.1). We let $\mathcal{U}$ and $\mathcal{U}'$ be the domains of dependence of V and V' in M. Then $\mathcal{A}_\mathcal{U}$ and $\mathcal{A}_{\mathcal{U}'}$ are the algebras of observables in bounded regions of spacetime.

To identify the algebra $\mathscr{A}_{\mathcal{U}}$, we use the simple remarks about von Neumann algebras that were explained at the end of Sect. 11.3.3. We also use a few facts about quantum field theory. One set of facts involves entanglement. For any state $\Psi \in \mathcal{H}$, there is a universal divergence in the entanglement entropy between modes in $\mathcal{U}$ and modes in $\mathcal{U}'$. The divergence comes from short wavelength modes that are localized near B, the "entangling surface" that separates $\mathcal{U}$ and $\mathcal{U}'$. The leading divergence is universal in the sense that it does not depend on the choice of Ψ. The reason for this is that the divergence comes from modes of very short wavelength, and at short distances every state in $\mathcal{H}$ looks the same. Because it comes from modes localized near B, the leading divergence is proportional to the area of B. The divergence involves pairs of modes on opposite sides of B that are entangled. The geometry of $\mathcal{U}$ and $\mathcal{U}'$ near B can be modeled locally by two opposite Rindler wedges in Minkowski space (see Fig. 11.1b), so the entanglement structure at short distances is given by Unruh's thermal interpretation of Rindler space [36], or equivalently by the Bisognano-Wichman theorem [37] which describes the application of Tomita-Takesaki theory to Rindler space.[19] The thermal description shows that many different pairs of modes have all possible degrees of entanglement. In fact, the modular Hamiltonian of Rindler space is a Lorentz boost generator, which has a continuous spectrum consisting of all real numbers.

We also need the fact that the local algebras in quantum field theory do not have a trace. Given any $\mathsf{a} \in \mathscr{A}_{\mathcal{U}}$, its trace in $\mathcal{H}$ is divergent because of the infinity of things that could be happening in region $\mathcal{U}'$, and there is no regularized version of this trace.

Given these facts, we can guess what type of algebra is $\mathscr{A}_{\mathcal{U}}$. The algebras whose structure incorporates a universal divergence in the entanglement entropy are of Type II and Type III. The absence of a trace tells us that $\mathscr{A}_{\mathcal{U}}$ is of Type III. And the fairly generic pattern of entanglement, with many modes of all possible degrees of entanglement, analogous to a rather generic choice of E_n's in Eq. (11.27), indicates that the algebra is of Type III_1.

This is believed to be the right answer. Rigorous arguments are available [16], at least for open sets in Minkowski space. Informal summaries of these arguments with somewhat more detail than provided here can be found in [17], section V.6 and in [19], section 6.5.

In either quantum field theory or statistical mechanics, the Type III nature of the algebras is associated, as we have seen, to a divergence in the entanglement entropy. In quantum field theory, this is an ultraviolet divergence between modes inside and outside the entangling surface B; in statistical mechanics it is an infrared divergence between the modes of the left and right systems. Even though one divergence is an ultraviolet divergence and the other is an infrared divergence, they are actually quite similar.

[19] The contributions of Unruh and of Bisognano and Wichman were done independently at about the same time. The connection between them was appreciated only years afterward, by Sewell [38].

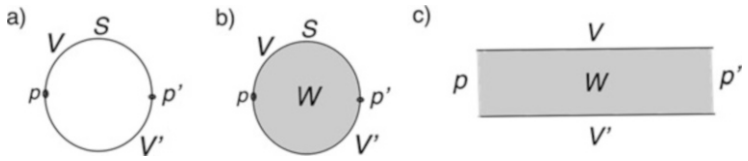


Fig. 11.3 (**a**) In a two-dimensional spacetime, we consider a Cauchy hypersurface consisting of a circle S. The entangling surface B is a pair of points p, p', whose complement in S consists of the two intervals V and V'. (**b**) A path integral on the disc W, whose boundary is S, computes the ground state of a conformal field theory. (**c**) The disc with the points p and p' removed can be conformally mapped to $\mathbb{R} \times I$, where I is a unit interval. V and V' become the two boundary components of $\mathbb{R} \times I$, and the points p, p' are projected to infinity

One aspect of the similarity is that in the case of a conformal field theory (CFT), one can actually make a conformal transformation between the two types of divergence. This is best-known in the case of a two-dimensional CFT [39, 40]. One takes the initial value surface S to be a circle, and one takes the entangling surface B to consist of a pair of points p, $p' \in S$; the complement of these points is the union of two open intervals V and V' (Fig. 11.3a). Let $\mathcal{H}$ be the Hilbert space of the given theory formulated on S. The conformally invariant ground state $\Omega \in \mathcal{H}$ can be computed by a path integral on a disc W whose boundary is S (Fig. 11.3b). The state Ω is highly entangled between regions V and V'. If one tries to quantify this entanglement by computing the entanglement entropy between V and V', one runs into an ultraviolet divergence which is related to the fact that the local algebras actually are of Type III. This ultraviolet divergence can be converted into an infrared divergence by a conformal mapping that maps the disc W, minus the two points p and p', to an infinite strip $\mathbb{R} \times I$, where I is an interval (Fig. 11.3c). After a Weyl rescaling, one can take the metric of $\mathbb{R} \times I$ to be the flat metric $\mathrm{d}x^2 + \mathrm{d}\tau^2$, $-\infty < x < \infty$, $0 \leq \tau \leq 1$. The path integral on this strip describes the thermofield double of the theory under study on the noncompact initial value surface $S' = \mathbb{R}$, and with inverse temperature 1; the two open intervals V and V' have become the boundaries of the strip at $\tau = 0, 1$, corresponding to the two copies in the thermofield double. The points p and p' have been projected to $x = \pm\infty$. In the description on the disc, there is an ultraviolet divergence in the entanglement entropy of the vacuum that arises from the behavior of modes on opposite sides of the points p or p'. In the description on the strip, this divergence is reinterpreted as an infrared divergence in the entanglement entropy of the thermofield double near $x = \pm\infty$.

All this has an analog for a conformally invariant theory in any dimension D [41, 42]. One takes the initial value surface S to be a $(D-1)$-sphere with a round metric. One further takes the entangling surface B to be an "equator" in S, and V and V' to be the northern and southern hemispheres. Thus the setup is the same as before, except that S is a $(D-1)$-sphere instead of a circle, and for $D > 2$, the equator in S is connected, rather than consisting of two isolated points. The ground state Ω of a D-dimensional CFT can be computed by a path integral on a flat unit ball W in D dimensions, whose boundary is S. After removing the equator

B of the boundary, W can be conformally mapped to a product $H \times I$, where H is hyperbolic space of dimension $D-1$, and I is again the interval $0 \leq \tau \leq 1$. A path integral on $H \times I$ again describes a thermofield double at termperature 1. The only difference from before is that the spatial manifold H on which the thermodynamics is defined is hyperbolic space of dimension $D-1$ rather than Euclidean space. But H has infinite volume just like Euclidean space, and there is an infrared divergence in the entanglement entropy between the two sides of the thermofield double. This divergence matches the ultraviolet divergence between modes on V and modes on V' in the original description on the sphere.

11.3.5 The Hagedorn Temperature

In this section, we return to the Araki-Woods model of a Hamiltonian that is a sum of single-qubit Hamiltonians $H = \sum_{n=1}^{\infty} H_n$, with

$$H_n = \begin{pmatrix} 0 & 0 \\ 0 & E_n \end{pmatrix}. \tag{11.28}$$

Now, however, we assume that $\lim_{n\to\infty} E_n = +\infty$. We also assume for convenience that the E_n are all positive, so that the ground state of the nth qubit is the state $|\uparrow\rangle_n = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$. The ground state of the whole system is $|\Omega\rangle_\uparrow = \otimes_{n=1}^{\infty} |\uparrow\rangle_n$. As in Sect. 11.2.2, a Hilbert space $\mathcal{H}^*$ that describes all states of infinitely many qubits is nonseparable. But we can define a separable Hilbert space $\mathcal{H}$ in which every state can be approximated by states in which all but finitely many qubits are in their ground state $|\uparrow\rangle_n$.

$\mathcal{H}$ has a basis consisting of states in which all qubits except some finite subset $n_1, n_2, \cdots, n_k$ are in their ground state. The energy of such a state is $E = E_{n_1} + E_{n_2} + \cdots + E_{n_k}$. Now let us ask this question: What is the probability to observe such a state in a thermal ensemble, at inverse temperature β? The probability to observe a microstate of energy E is of course $p(E,\beta) = \frac{1}{Z(\beta)} e^{-\beta E}$, where

$$Z(\beta) = \prod_{n=1}^{\infty} (1 + e^{-\beta E_n}) \tag{11.29}$$

is the partition function. In Sect. 11.3.1, we studied models in which the partition function diverges in the limit of infinitely many qubits, but now we wish to consider the case that the E_n tend to infinity quickly enough that $Z(\beta)$ converges, at least for some range of β. Clearly when $Z(\beta) < \infty$, we have $p(E,\beta) > 0$ and the probability to observe any given microstate in $\mathcal{H}$ in the thermal ensemble is positive.

This suggests that the thermal ensemble at inverse temperature β can simply be defined in the Hilbert space $\mathcal{H}$. To see that this is the case, let us ask what is the

probability that, for some N, all qubits except the first N are in their ground state. The probability for this in the thermal ensemble is $Z_N(\beta)/Z(\beta)$, where $Z_N(\beta)$ is a truncated partition function defined for the first N qubits:

$$Z_N(\beta) = \prod_{n=1}^{N}(1 + e^{-\beta E_n}). \tag{11.30}$$

If $Z(\beta) = \infty$, then $Z_N(\beta)/Z(\beta) = 0$ for all N, since $Z_N(\beta)$ is finite, but if $Z(\beta) < \infty$, then

$$\lim_{N\to\infty} \frac{Z_N(\beta)}{Z(\beta)} = 1. \tag{11.31}$$

For $Z(\beta) < \infty$, this shows that, with probability 1, any state in the thermal ensemble can be approximated arbitrarily well by states in $\mathcal{H}$.

Therefore, in this situation, it is not necessary to go to the thermofield double in order to describe the thermal ensemble of an infinite number of qubits in a separable Hilbert space. What happens if we do go to the thermofield double anyway? The answer is that we get nothing essentially new; when $Z < \infty$ the thermofield double does not combine the two copies in an interesting way.

To explain how this comes about, we consider two copies of this system,[20] the left and right copy, each consisting of countably many qubits with identical Hamiltonians. There are nonseparable Hilbert spaces $\mathcal{H}_\ell^*$ and $\mathcal{H}_r^*$ that describe all possible states of the left and right systems, and these contain separable subspaces $\mathcal{H}_\ell$ and $\mathcal{H}_r$ that have bases with finitely many qubits excited. We also write $\mathscr{A}_{\ell,0}$ or $\mathscr{A}_{r,0}$ for the algebras of operators that act on finitely many qubits of the left or right systems, respectively, and $\Omega_{\uparrow,r}$, $\Omega_{\uparrow,\ell}$ for the ground states of the right and left systems.

We can define the thermofield double state in the naive way as a vector in $\mathcal{H}_\ell^* \otimes \mathcal{H}_r^*$:

$$\Psi_{\rm TFD} = \frac{1}{\sqrt{Z(\beta)}} \bigotimes_{n=1}^{\infty} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta E_n/2} \end{pmatrix}. \tag{11.32}$$

Then we define the thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$ as the closure of the set of all vectors of the form $\mathsf{a}_r\Psi_{\rm TFD}$, $\mathsf{a}_r \in \mathscr{A}_{r,0}$.

In Sect. 11.3.3, we could not define the thermofield double state and Hilbert space of an infinite system in this simple way, because we were considering systems with $Z(\beta) = \infty$. Therefore, in Eqs. (11.23) and (11.26), we defined the thermofield double state directly by a formula analogous to (11.32) for a finite system of N

[20] The system considered has an obvious time-reversal symmetry, so the two copies are identical.

qubit pairs, and we invoked the GNS construction to define $\Psi_{\rm TFD}$ and $\mathcal{H}_{\rm TFD}$ in the limit of an infinite system.

A priori, Eq. (11.32) defines $\Psi_{\rm TFD}$ as a vector in the nonseparable Hilbert space $\mathcal{H}_\ell^* \otimes \mathcal{H}_r^*$, and likewise it defines $\mathcal{H}_{\rm TFD}$ as a subspace of that space. But actually, when $Z(\beta) < \infty$ so that the definition (11.32) makes sense, $\mathcal{H}_{\rm TFD}$ is the same as the separable Hilbert space $\mathcal{H}_\ell \otimes \mathcal{H}_r$. To see this, observe that a truncated version of $\Psi_{\rm TFD}$, namely

$$\Psi_{{\rm TFD},N} = \frac{1}{\sqrt{Z(\beta)}} \bigotimes_{n=1}^{N} \begin{pmatrix} 1 & 0 \\ 0 & e^{-\beta E_n/2} \end{pmatrix} \bigotimes_{n=N+1}^{\infty} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{11.33}$$

is manifestly contained in $\mathcal{H}_\ell \otimes \mathcal{H}_r$. But $\lim_{N\to\infty} \Psi_{{\rm TFD},N} = \Psi_{\rm TFD}$. So $\Psi_{\rm TFD} \subset \mathcal{H}_\ell \otimes \mathcal{H}_r$. This implies that $\mathcal{H}_{\rm TFD} \subset \mathcal{H}_\ell \otimes \mathcal{H}_r$, since $\mathcal{H}_{\rm TFD}$ is generated from $\Psi_{\rm TFD}$ by the action of $\mathscr{A}_{r,0}$, and $\mathcal{H}_\ell \otimes \mathcal{H}_r$ is closed under the action of $\mathscr{A}_{r,0}$. Conversely, $\mathscr{A}_{r,0}$ contains an operator Π_N that projects onto the ground state of the first N qubits, and acts trivially on others; since $|\Omega\rangle_{\uparrow,\ell} \otimes |\Omega\rangle_{\uparrow,r} = Z(\beta) \lim_{\mathcal{N}\to\infty} \Pi_N \Psi_{\rm TFD}$, we have $|\Omega\rangle_{\uparrow,\ell} \otimes |\Omega\rangle_{\uparrow,r} \in \mathcal{H}_{\rm TFD}$, implying, since $\mathcal{H}_\ell \otimes \mathcal{H}_r$ is generated from $|\Omega\rangle_{\uparrow,\ell} \otimes |\Omega\rangle_{\uparrow,r}$ by acting with $\mathscr{A}_{\ell,0} \otimes \mathscr{A}_{r,0}$, that $\mathcal{H}_\ell \otimes \mathcal{H}_r \subset \mathcal{H}_{\rm TFD}$.

In short, when $Z(\beta) < \infty$, the thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$ is just a simple tensor product $\mathcal{H}_{\rm TFD} = \mathcal{H}_\ell \otimes \mathcal{H}_r$. Likewise, if we take the closures of $\mathscr{A}_{\ell,0}$ and $\mathscr{A}_{r,0}$ to get von Neumann algebras $\mathscr{A}_\ell$ and $\mathscr{A}_r$ that act on the thermofield double, we get nothing essentially new: $\mathscr{A}_\ell$ and $\mathscr{A}_r$ are simply the Type I_∞ algebras of all bounded operators on $\mathcal{H}_\ell$ and $\mathcal{H}_r$, respectively. In this situation, the Hamiltonians H_ℓ and H_r of the left and right systems are well-defined as operators on $\mathcal{H}_{\rm TFD}$ and are elements of $\mathscr{A}_\ell$ and $\mathscr{A}_r$ respectively. All in all, when $Z(\beta) < \infty$, we gain little by going to the thermofield double; it is just a tensor product of two decoupled systems.

On the other hand, if $Z(\beta) = \infty$, we are in the situation described in Sect. 11.3.3. The thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$ is an essentially new construction, not a simple tensor product, and the corresponding algebras $\mathscr{A}_\ell$ and $\mathscr{A}_r$ are generically of Type III_1. The operators H_ℓ and H_r cannot be defined as operators on $\mathcal{H}_{\rm TFD}$. As a partial substitute one has $\widehat{H} = H_r - H_\ell$, which does have a limit for the infinite system, and generates a group of outer automorphisms of $\mathscr{A}_r$ and of $\mathscr{A}_\ell$.

A particularly interesting case, and important background for Sect. 11.4, is the borderline case in which $Z(\beta) < \infty$ at low enough temperature, but $Z(\beta) = \infty$ at sufficiently high temperature. This can happen if $E_n \sim c \log n$ for large n, with a constant c. Then at low temperatures, the thermofield double construction just describes two decoupled systems, and all algebras are of Type I. But at high termperatures, the thermofield double construction is genuinely new, combines the two systems in a subtle way, and is essential if we wish to describe the infinite qubit system in a separable Hilbert space. The algebras are of Type III, and time translations of the algebras are a group of outer automorphisms. By analogy with standard terminology in strong interaction theory, string theory, and large N gauge theory, we call the temperature at which $Z(\beta)$ ceases to converge the Hagedorn temperature T_H, and we also write $\beta_H = 1/T_H$.

11.3.6 Density Matrices and Entropy

For a quantum system with a Hilbert space $\mathcal{H}$ of physical states, one usually defines the von Neumann entropy, for a state of the system described by a density matrix $\rho : \mathcal{H} \to \mathcal{H}$, by $S(\rho) = -\mathrm{Tr}\,\rho \log \rho$. As usual, here ρ is a positive self-adjoint operator normalized to $\mathrm{Tr}\,\rho = 1$. The trace function $\mathsf{a} \to \mathrm{Tr}\,\mathsf{a}$ is a complex linear function from the algebra $\mathscr{A}$ of all operators on $\mathcal{H}$ (or all bounded operators if $\mathcal{H}$ is infinite-dimensional) to $\mathbb{C}$; it satisfies $\mathrm{Tr}\,\mathsf{a}^\dagger\mathsf{a} > 0$ for all $\mathsf{a} \neq 0$. This last condition ensures that the function $F(\mathsf{a}) = \mathrm{Tr}\,\mathsf{a}\rho$ satisfies the conditions as stated in Sect. 11.2.5 for a state on the algebra $\mathscr{A}$. Conversely, every state on the algebra $\mathscr{A}$ is of the form $F(\mathsf{a}) = \mathrm{Tr}\,\mathsf{a}\rho$ for some density matrix ρ. To prove this, one first observes that the trace is nondegenerate, in the sense that every linear function on $\mathscr{A}$ is of the form $F(\mathsf{a}) = \mathrm{Tr}\,\mathsf{ab}$ for some $\mathsf{b} \in \mathscr{A}$. The other axioms for a state (namely $F(\mathsf{a}^\dagger\mathsf{a}) \geq 0$ and $F(1) = 1$) imply that ρ must be a positive operator satisfying $\mathrm{Tr}\,\rho = 1$, in other words, a density matrix.

Now consider any von Neumann algebra $\mathscr{A}$ that acts on a Hilbert space $\mathcal{H}$. An element of $\mathscr{A}$ is called positive if it corresponds to a positive, self-adjoint operator on $\mathcal{H}$. If the algebra $\mathscr{A}$ has a nondegenerate trace, then the same arguments as before show that any state on $\mathscr{A}$ is of the form $F(\mathsf{a}) = \mathrm{Tr}\,\mathsf{a}\rho$, where ρ is a density matrix, that is, a positive element of $\mathscr{A}$ of trace 1. We can then define, for any state, the von Neumann entropy $S(\rho) = -\mathrm{Tr}\rho \log \rho$. In this article, the von Neumann algebras that we consider are factors, that is, algebras with trivial center (in other words, algebras whose center consists only of $\mathbb{C}$). A factor is the von Neumann algebra analog of a simple Lie group. In the case of a factor, any trace is nondegenerate and therefore the condition to be able to define a von Neumann entropy is only that $\mathscr{A}$ has a trace. Also, in the case of a factor, if a trace exists, it is unique up to a multiplicative constant. As we will see, in general it is important that the trace is not quite unique.

Let us discuss entropy for the different types of von Neumann algebra. The original arena for von Neumann entropy is a Type I algebra, which acts irreducibly on a Hilbert space $\mathcal{H}$. A standard argument shows that $S(\rho)$ is always nonnegative in this case. A subtlety is that in the case of a Type I_∞ algebra, that is, if $\mathcal{H}$ is infinite-dimensional, it is possible to have $S(\rho) = +\infty$. That is because the condition $\mathrm{Tr}\,\rho = 1$ does not ensure the convergence of $\mathrm{Tr}\,\rho\log\rho$. Thus for Type I_∞, the entropy takes values in $\mathbb{R}_{\geq 0} \cup +\infty$.

Now let us consider algebras of Type II and Type III. A Type III algebra has no trace and therefore no notion of von Neumann entropy. However, a Type II algebra does have a trace, as explained in Sect. 11.3.3, and therefore one can define an entropy[21] for a state of a von Neumann algebra of Type II.

Let us explore this notion, initially for the case of Type II_1. We recall that a Type II_1 algebra $\mathscr{A}$ has a trace that is defined for all elements of $\mathscr{A}$, and is conventionally

[21] This has been discussed, primarily for Type II_1, in [43, 44].

normalized so that $\mathrm{Tr}\,1 = 1$. The simplest density matrix to consider is $\rho = 1$. In this case, $\log\rho = 0$ so $S(\rho) = -\mathrm{Tr}\,\rho\log\rho = 0$. But what ρ describes is far from being a zero entropy state in any conventional sense. In fact, the density matrix $\rho = 1$ describes the maximally entangled thermofield double state Ψ_{TFD} that was defined in Eq. 11.23. Indeed, we defined the trace in a Type II_1 algebra by $\mathrm{Tr}\,\mathsf{a} = \langle\Psi_{\mathrm{TFD}}|\mathsf{a}|\Psi_{\mathrm{TFD}}\rangle$. This implies that to satisfy $\mathrm{Tr}\,\mathsf{a}\rho = \langle\Psi_{\mathrm{TFD}}|\mathsf{a}|\Psi_{\mathrm{TFD}}\rangle$, we need to take $\rho = 1$, so $\rho = 1$ is the density matrix that corresponds to the maximally entangled thermofield double state Ψ_{TFD}.

The maximally entangled state of N qubit pairs has von Neumann entropy $N\log 2$, and to get a Type II_1 algebra, one takes the limit $N\to\infty$. So the entropy diverges for $N\to\infty$. But with the Type II_1 definition of entropy, the large N limit of the thermofield double state is deemed to have entropy 0. This suggests that in general Type II_1 entropy is ordinary von Neumann entropy with the entropy of a maximally entangled state subtracted. To confirm this interpretation, recall that a general state of a Type II_1 algebra can be associated to a system of N qubit pairs, in the large N limit, in a state in which almost all the qubit pairs are almost maximally entangled. As an illustrative example, suppose that the first k qubits are definitely in a "spin up" state, and the others are maximally mixed. To describe this state by a density matrix, let Π be the orthogonal projection operator onto the subspace of states of N qubits in which the first k qubits all have spin up. For finite N, $\rho = 2^{-(N-k)}\Pi$ is a density matrix of the N qubit system, and its von Neumann entropy is $S_N = (N-k)\log 2$. But in the limit $N\to\infty$, Π becomes an element of the Type II_1 algebra with $\mathrm{Tr}\,\Pi = 2^{-k}$. So in the Type II_1 algebra, we can define a density matrix $\rho = 2^k\Pi$, with $\mathrm{Tr}\,\rho = 1$. For this density matrix, we compute $S = -\mathrm{Tr}\,\rho\log\rho = -k\log 2$. This is the same as $S = \lim_{N\to\infty}(S_N - N\log 2)$. In other words, S is the large N limit of $S_N - N\log 2$, or differently put, it is the large N limit of the entropy difference between the actual state of the first N qubits and a maximally mixed state. As this example and discussion suggest, entropy is negative definite for a Type II_1 factor.

A Type III algebra can similarly be constructed, as described in Sect. 11.3.3, from an infinite collection of (non-maximally) entangled qubit pairs. But in this case, there is no way, using only the structure of the algebra, to define a renormalized entropy. This is reflected in the fact that the divergences in the entanglement entropy of a local region in quantum field theory are model-dependent (there is a leading divergence proportional to the area of the boundary of the region, with a theory-dependent coefficient, and there are subleading divergences that depend on the spacetime dimension and the operator content of the theory). Renormalized entropies of local regions in quantum field theory can be useful and important, but defining them requires information beyond the von Neumann algebra structure. Similar remarks apply in quantum statistical mechanics if one considers the entropy of a state that is not thermal but looks thermal near spatial infinity.

Now let us consider an algebra $\mathcal{A}$ of Type II_∞. Such an algebra can be factored as[22] $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2$, where $\mathcal{A}_1$ is of Type II_1 and $\mathcal{A}_2$ is the type I_∞ algebra of bounded operators on a separable Hilbert space $\mathcal{K}$. We can define a trace Tr_1 on $\mathcal{A}_1$ normalized so that $\mathrm{Tr}_1\, 1 = 1$, and a trace Tr_2 on $\mathcal{A}_2$ normalized so that, if $\Pi \in \mathcal{A}_2$ is the orthogonal projection operator on a one-dimensional subspace $\mathcal{K}_0 \subset \mathcal{K}$, then $\mathrm{Tr}_2\, \Pi = 1$. Then we can define a trace on $\mathcal{A}$ by $\mathrm{Tr} = \mathrm{Tr}_1 \otimes \mathrm{Tr}_2$.

Since a trace is available, there is a notion of entropy for such an algebra $\mathcal{A}$. But there is a very important subtlety: there is no canonical way to normalize this trace. One immediate observation is that, since the identity element of $\mathcal{A}_2$ has a divergent trace, we cannot normalize the trace, as we did for Type II_1, by saying that the identity element has trace 1. A deeper explanation involves the fact that a Type III algebra can be factored as the tensor product of algebras of Type II_1 and Type I_∞ in many different ways.

Let $\mathcal{A}_3$ be an algebra of Type I_n, acting on an n-dimensional Hilbert space $\mathcal{H}_n$, and define the trace Tr_3 on $\mathcal{A}_3$ to be the usual trace of a linear transformation acting on $\mathcal{H}_n$. Thus $\mathrm{Tr}_3\, 1 = n$. Consider the algebra $\mathcal{A} = \mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3$, with $\mathcal{A}_1$ and $\mathcal{A}_2$ as before. The algebra $\mathcal{A}'_1 = \mathcal{A}_1 \otimes \mathcal{A}_3$ is of Type II_1, and the algebra $\mathcal{A}'_2 = \mathcal{A}_2 \otimes \mathcal{A}_3$ is of Type I_∞. So $\mathcal{A} = \mathcal{A}'_1 \otimes \mathcal{A}_2 = \mathcal{A}_1 \otimes \mathcal{A}'_2$ gives two different factorizations of the same Type II_∞ factor as the tensor product of factors of Type II_1 and Type I_∞. These two factorizations of $\mathcal{A}$ lead to definitions of the trace that differ by a factor of n. To see this, consider the element $\mathsf{a} = 1 \otimes \Pi \otimes 1 \in \mathcal{A}_1 \otimes \mathcal{A}_2 \otimes \mathcal{A}_3$, where as before Π is an element of $\mathcal{A}_2$ that projects on a one-dimensional subspace $\mathcal{K}_0 \subset \mathcal{K}$. With respect to the factorization of $\mathcal{A}$ as $\mathcal{A}'_1 \otimes \mathcal{A}_2$, a is the tensor product of $1 \in \mathcal{A}'_1$ with $\Pi \in \mathcal{A}_2$, so the normalization of the trace that is natural for this factorization gives $\mathrm{Tr}\,\mathsf{a} = \mathrm{Tr}\, 1 \otimes \Pi = 1$. On the other hand, with respect to the factorization of $\mathcal{A}$ as $\mathcal{A}_1 \otimes \mathcal{A}'_2$, a is $1 \otimes \Pi'$, where Π' projects on the n-dimensional subspace $\mathcal{K}_0 \otimes \mathcal{H}_n \subset \mathcal{K} \otimes \mathcal{H}_n$. So with the normalization of the trace that is natural for this second factorization, we would have $\mathrm{Tr}\,\mathsf{a} = n$.

Thus different choices of factorization motivate different normalizations of the trace for a Type II_∞ factor. In fact, there can be no natural normalization of the trace, because a Type II_∞ algebra has an outer automorphism group that rescales the trace by any positive real number. (This is a standard result in von Neumann algebra theory and will be explained elsewhere [23].) However, as long as we assume our algebra to be a factor, analogous to a simple Lie group, the only indeterminacy in the trace is an overall multiplicative constant.

Let us see how this indeterminacy affects the definition of entropy. If we rescale the trace by $\mathrm{Tr} \to \lambda \mathrm{Tr}$, we must compensate by rescaling the density matrix by

[22] Because of the Type I_∞ factor, entropy in a Type II_∞ factor can take the value $+\infty$. Because of the Type II_1 factor (or because of the additive indeterminacy that we discuss shortly) it can be negative. So entropy in a Type II_∞ factor is valued in $\mathbb{R} \cup +\infty$.

$\rho \to \lambda^{-1}\rho$, to preserve $\mathrm{Tr}\,\rho = 1$. Under the combined rescaling of Tr and ρ, we have

$$S(\rho) = -\mathrm{Tr}\,\rho \log \rho \to S(\rho) + \log \lambda. \tag{11.34}$$

In other words, when we rescale the trace by a factor λ, the entropy of any state is shifted by an additive constant $\log \lambda$, independent of the state.

This means that entropy in a Type II_∞ factor is somewhat analogous to entropy in classical mechanics. The concept of entropy was originally discovered in the nineteenth century by macroscopic arguments, which showed that there must be a state function S of a system in thermal equilibrium that obeys the first law of thermodynamics. If the relevant thermodynamic variables are the energy E, temperature T, pressure p, and volume V, then the first law reads

$$\mathrm{d}E = T\mathrm{d}S - p\mathrm{d}V. \tag{11.35}$$

This equation determines S up to an additive constant. Quantum mechanically, one can fix the additive constant by saying that (assuming the system under study has a unique ground state) the entropy vanishes at $T = 0$. Classically, no such statement is possible, because for example the entropy of a classical harmonic oscillator goes to $-\infty$ for $T \to 0$.

For another explanation of why it is difficult in classical physics to fix an overall additive constant in the entropy, consider a system of particles with positions $\mathbf{x}$ and momenta $\mathbf{p}$. One can describe one's state of knowledge about the system with a probability distribution function $\rho(\mathbf{p}, \mathbf{x})$. Then a classical definition of entropy is

$$S = -\int \mathrm{d}\mathbf{p}\,\mathrm{d}\mathbf{x}\rho(\mathbf{p}, \mathbf{x}) \log \rho(\mathbf{p}, \mathbf{x}). \tag{11.36}$$

The integral runs over the classical phase space of the system. But classically, there is no natural way to normalize the phase space measure, since, for any pair of canonical variables p, x, the measure $\mathrm{d}p\,\mathrm{d}x$ has dimensions of action, and classically there is no natural constant with dimensions of action that could be used to normalize it. (Quantum mechanically, one replaces $\mathrm{d}p\,\mathrm{d}x$ by $\mathrm{d}p\,\mathrm{d}x/2\pi\hbar$.) The phase space measure in classical mechanics is naturally defined up to an overall multiplicative constant, but there is no natural way to fix this constant. If we rescale the measure by $\mathrm{d}\mathbf{p}\,\mathrm{d}\mathbf{x} \to \lambda \mathrm{d}\mathbf{p}\,\mathrm{d}\mathbf{x}$, for a positive constant λ, then we have to compensate by $\rho \to \lambda^{-1}\rho$. But this then shifts S by $S \to S + \log \lambda$, just as in the discussion of a Type II_∞ algebra.

Here is an interesting example of entropy for a state of a Type II_∞ algebra $\mathscr{A}$. Pick a factorization $\mathscr{A} = \mathscr{A}_1 \otimes \mathscr{A}_2$ where $\mathscr{A}_1$ is of Type II_1 and $\mathscr{A}_2$ is of Type I_∞. Let Π_k, $k = 1, \cdots, n$, be orthogonal projection operators in $\mathscr{A}_1$ satisfying $\Pi_i \Pi_j = \delta_{ij}\Pi_j$, $\mathrm{Tr}\,\Pi_i = p_i$, $\sum_{k=1}^n \Pi_k = 1$. Let ρ_k, $k = 1, \cdots, n$ be any density matrices in $\mathscr{A}_2$.

Then $\rho = \sum_{k=1}^{n} \Pi_k \otimes \rho_k$ is a density matrix in $\mathscr{A}$. Its entropy is

$$S(\rho) = \sum_k p_k S(\rho_k) = -\sum_k p_k \mathrm{Tr}\, \rho_k \log \rho_k. \tag{11.37}$$

In verifying this, one uses the fact that $\Pi_k \log \Pi_k = 0$ for all k, since the only eigenvalues of Π_k are 0 and 1, and $x \log x = 0$ if $x = 0$ or 1. It is noteworthy that in the formula for $S(\rho)$ there is no Shannon term $-\sum_k p_k \log p_k$. Thus in this particular example, Type II_∞ entropy is the average entropy of an ensemble of density matrices.

We have phrased this discussion in terms of the von Neumann entropy, but similar remarks apply for other information theoretic measures such as the Renyi entropies $S_\alpha(\rho) = \frac{1}{1-\alpha} \log \mathrm{Tr}\, \rho^\alpha$. These entropies can be defined for a state on an algebra of Type I or Type II (but not, of course, Type III). In the Type II_∞ case, if we rescale the trace by a factor λ and compensate by rescaling the density matrix $\rho \to \lambda^{-1}\rho$, the Renyi entropies are shifted by

$$S_\alpha(\rho) \to S_\alpha(\rho) + \frac{\alpha}{\alpha - 1} \log \lambda. \tag{11.38}$$

Thus all the Renyi entropies are well-defined in Type II_∞, up to a shift that is controlled in this way by the same α-independent parameter λ that controls the indeterminacy in the von Neumann entropy.

11.4 The Large N Limit and the Thermofield Double

QCD in Minkowski space is confining at low temperatures and has a deconfinement transition at a certain positive temperature [45–47]. In finite volume, this transition is smoothed out. But if we replace QCD, which has gauge group $SU(3)$, with a theory with gauge group $SU(N)$ (or $SO(N)$ or $Sp(N)$), then it is believed that in the large N limit, there is a sharp deconfinement transition even in finite volume. This is true even for zero [48–51] or small [52] coupling. In the case of a gauge theory that has a gravitational dual, this phenomenon is relevant to the AdS/CFT correspondence. In that application, if the bulk spacetime has D noncompact dimensions, so that the conformal boundary has dimension $D-1$, the spatial manifold on which the gauge theory lives is most simply taken to be a sphere S^{D-2}. Then the deconfinement transition is dual to the Hawking-Page transition [53] between a thermal gas in Anti de Sitter space and a black hole [54]. We call the temperature at which this transition occurs T_{HP}.

Let $\mathcal{H}_N$ be the Hilbert space of a large N gauge theory, quantized on the compact manifold S^{D-2}. It is believed that $\mathcal{H}_N$ has a large N limit $\mathcal{H}_\infty$: the energy eigenvalues and multiplicities and the matrix elements of operators all have limits for $N \to \infty$ (and in particular the multiplicities do not grow with

N). At sufficiently low temperatures, the thermal ensemble can be understood in the limiting Hilbert space $\mathcal{H}_\infty$. Because the energy eigenvalues and multiplicities have limits for $N \to \infty$, the thermodynamic functions such as entropy, energy, etc., all have large N limits at sufficiently low temperatures. Thus in the low temperature phase, the energy and entropy are of order N^0 for large N. In the range of temperatures at which the thermodynamics can be described in $\mathcal{H}_\infty$, the natural algebra of observables at large N is simply the Type I_∞ algebra of all operators on $\mathcal{H}_\infty$.

It is believed that in large N gauge theories, the partition function $\mathrm{Tr}_{\mathcal{H}_\infty} e^{-\beta H}$, computed in the limiting large N Hilbert space $\mathcal{H}_\infty$, diverges at a temperature T_H, known as the Hagedorn temperature. This can be seen quite explicitly at zero coupling. At or below T_H, the description based on the Hilbert space $\mathcal{H}_\infty$ has to break down. We will loosely call the transition at which this happens the Hagedorn transition. In various weakly coupled gauge theories, it is believed that the transition occurs at a temperature slightly below T_H [52]. In gauge theories—such as $\mathcal{N} = 4$ super Yang-Mills theory in four dimensions—that participate in AdS/CFT duality, the transition away from a description in the Hilbert space $\mathcal{H}_\infty$ occurs at the Hawking-Page temperature T_{HP}, which is far below T_H (assuming that $g^2 N$ is large and the AdS theory can be studied semiclassically). In fact, T_H is at the string scale, while T_{HP} is of order the inverse radius of the AdS space. Thus in the AdS/CFT context, what we are calling the Hagedorn transition is the same as the Hawking-Page transition and occurs at a temperature far below T_H. Regardless, above this transition the physics can no longer be described in $\mathcal{H}_\infty$.

Above the transition temperature, the energy and entropy are both proportional to N^2 for large N [55] and in particular do not have large N limits. To reach the high temperature phase in the large N limit, we have to take $N \to \infty$ with the energy E proportional to N^2 (as opposed to taking $N \to \infty$ with fixed E, which will always lead to a description in the Hilbert space $\mathcal{H}_\infty$, regardless of E). In brief, we will refer to the large N limit in that regime as the large N limit above the Hagedorn transition. It seems very unlikely that there is a limiting large N Hilbert space that describes the large N limit above the Hagedorn transition. Certainly the literature does not contain any proposal for what such a large N Hilbert space would be. Since the entropy above the Hagedorn transition is of order N^2, when one increases the rank of the gauge group from N to $N + 1$, the entropy increases by a multiple of N, and the Hilbert space becomes much bigger. So most "above the Hagedorn transition" microstates with gauge group $SU(N + 1)$ are "new" and have no antecedents in the $SU(N)$ theory.

The non-existence of a large N limit of the Hilbert space above the Hagedorn transition would be analogous to the non-existence, in quantum statistical mechanics at temperature $T > 0$, of a Hilbert space that describes the infinite volume limit. In this analogy, N^2 plays the role of the volume V. There is also an analogy with the non-existence of a natural Hilbert space description of quantum field theory in an open universe.

The analogy with statistical mechanics also suggests a partial cure, which is to go to the thermofield double. This is particularly powerful in gauge theories

that participate in AdS/CFT duality. In such a case, the large N limit of the thermofield double, in the high temperature phase, is a two-sided eternal black hole [56]. By quantizing the low energy bulk fields in the black hole spacetime, one constructs a Hilbert space $\mathcal{H}_{\rm TFD}$. This Hilbert space describes the large N limit of the thermofield double of the gauge theory, in the high temperature phase.

A natural algebra of observables in the large N limit is the algebra of polynomial functions of single trace operators. Based on the analogy with quantum statistical mechanics and the discussion in Sect. 11.3.5, we can make a guess about the nature of this large N algebra. As already noted, below the Hagedorn transition, the physics can be described in the original large N Hilbert space $\mathcal{H}_\infty$, and the natural algebra acting on this Hilbert space is of Type I. In the high termperature phase, the analogy with statistical mechanics suggests that the natural algebra of single trace operators would be of Type III. Indeed, this has been argued recently by Leutheusser and Liu [22], who made this claim partly based on the AdS/CFT duality, in a paper with several novel ideas about quantum black holes. These issues and some subtle corrections to the large N limit will be further discussed elsewhere [23].

Finally let us consider the Hamiltonians H_r and H_ℓ that act on the two copies in the thermofield double. What happens in the large N limit above the Hagedorn transition is similar to what happens in the more familiar infinite volume limit. The difference $\widehat{H} = H_r - H_\ell$ annihilates the thermofield double state $\Psi_{\rm TFD}$ and converges, in the large N limit, to an operator that acts on the thermofield double Hilbert space $\mathcal{H}_{\rm TFD}$. However, because of fluctuations, H_r and H_ℓ do not separately have large N limits at $\beta < 1/T_{HP}$. For example, H_r has an expectation value $\langle H_r\rangle_\beta \sim N^2$ and so does not have a large N limit. We can of course subtract this expectation value, but the subtracted operator $H_r - \langle H_r\rangle_\beta$ has divergent thermal fluctuations $\left\langle \left(H_r - \langle H_r\rangle_\beta\right)^2\right\rangle_\beta \sim N^2$ and therefore also does not converge for large N to an operator on $\mathcal{H}_{\rm TFD}$. $\widehat{H}$ generates a group of outer automorphisms of the large N operator algebras of the left and right systems. This is a group of outer automorphisms because H_ℓ and H_r do not separately have large N limits.

Acknowledgments I thank H. Liu, R. Longo, R. Mahajan, J. Maldacena, N. Seiberg, J. Turiaci, and R. Wald for helpful and informative discussions. Research supported in part by NSF Grant PHY-1911298.

References

1. B.S. DeWitt, Quantum field theory in curved spacetime. Phys. Rep. **19**, 295–357 (1975)
2. S.A. Fulling, M. Sweeny, R.M. Wald, Singularity structure of the two-point function in quantum field theory in curved spacetime. Commun. Math. Phys. **63**, 257–264 (1978)
3. S.A. Fulling, F.J. Narcowich, R.M. Wald, Singularity structure of the two-point function in quantum field theory in curved spacetime, II. Ann. Phys. **136**, 243–272 (1981)
4. N.D. Birrell, P.C.W. Davies, *Quantum Fields In Curved Space* (Cambridge University Press, 1982)

5. S.A. Fulling, *Aspects of Quantum Field Theory in Curved Spacetime* (Cambridge University Press, 1989)
6. B.S. Kay, The principle of locality and quantum field theory on (non globally hyperbolic) curved spacetimes. Rev. Math. Phys. **4**, 167–195 (1992)
7. I.L. Buchbinder, S.D. Odintsov, I.L. Shapiro, *Effective Action in Quantum Gravity* (Taylor & Francis, 1992)
8. R.M. Wald, *Quantum Field Theory in Curved Spacetime and Black Hole Thermodynamics* (University of Chicago, 1994)
9. R. Brunetti, K. Fredenhagen, R. Verch, The generally covariant locality principle - A new paradigm for local quantum field theory. arXiv:math-ph/0112041
10. K. Fredenhagen, Locally covariant quantum field theory. arXiv:hep-th/0403007
11. S. Hollands, R.M. Wald, Axiomatic quantum field theory in curved spacetime. arXiv:0803.2003
12. C. Bar, K. Fredenhagen (eds.), *Quantum Field Theory On Curved Spacetimes* (Springer, 2009)
13. L. Parker, D. Toms, *Quantum Field Theory in Curved Spacetime* (Cambridge University Press, 2009)
14. C.J. Fewster, locally covariant quantum field theory and the problem of formulating the same physics in all spacetimes. Phil. Trans. R. Soc. **A373**, 20140238 (2015). arXiv:1502.04642
15. H. Araki, Type of von Neumann algebra associated with free field. Prog. Theor. Phys. **5**, 1–13 (1964)
16. K. Fredenhagen, On the modular structure of local algebras of observables. Commun. Math. Phys. **97**, 79 (1985)
17. R. Haag, *Local Quantum Physics* (Springer, 1992)
18. J. Yngvason, The role of Type III factors in quantum field theory. arXiv:math-ph/0411058
19. E. Witten, Some entanglement properties of quantum field theory. Rev. Mod. Phys. **90**, 045003 (2018). arXiv:1803.04993
20. R. Haag, N.M. Hugenholtz, M. Winnink, On the equilibrium states in quantum statistical mechanics. Commun. Math. Phys. **5**, 215–236 (1967)
21. E. Witten, Quantum field theory in curved spacetime. Lecture available at https://www.youtube.com/watch?v=8Wu4QhlznnY&t
22. S. Leutheusser and H. Liu, Causal connectability between quantum systems and the black hole interior in holographic duality. arXiv:2110.05497
23. E. Witten, Gravity and the crossed product. JHEP **10**, 008 (2022). https://doi.org/10.1007/JHEP10(2022)008. arXiv:2112.12828 [hep-th]
24. M. Kontsevich, G.B. Segal, Wick rotation and the positivity of energy in quantum field theory. Quart. J. Math. **72**, 673–99 (2021). arXiv:2105.10161
25. J. von Neumann, Die Eindeutigkeit der Schrödingerschen Operatoren. Math. Ann. **104**, 570–578 (1931)
26. L. Gårding, A. Wightman, Representations of the commutation relations. Proc. NAS **40**, 622–626 (1954)
27. G.L. Sewell, *Quantum Mechanics and its Emergent Macrophysics* (Princeton University Press, 2002)
28. H. Bostelmann, Operator product expansions as a consequence of phase space properties. arXiv:math-ph/0502004
29. A. Ashtekar, *Asymptotic Quantization* (Bibliopolis, 1987)
30. A. Strominger, *Lectures on the Infrared Structure of Gravity and Gauge Theory* (Princeton University Press, 2018)
31. K. Prabhu, G. Satishchandran, R. Wald, Infrared finite scattering theory in quantum field theory and quantum gravity. Phys. Rev. D **106**(6) (2022) 066005 https://doi.org/doi:10.1103/PhysRevD.106.066005. arXiv:2203.14334 [hep-th]
32. J. von Neumann, On infinite direct products. Comp. Math. **6**, 1–77 (1938)
33. R.T. Powers, Representations of uniformly hyperfinite algebras and their associated von Neumann rings. Ann. Math. **86**, 138–171 (1967)

34. H. Araki, E.J. Woods, A classification of factors. Publ. RIMS Kyoto Univ. Ser. A **3**, 51–130 (1968)
35. F.J. Murray, J. von Neumann, On rings of operators. Ann. Math. **37**, 116–229 (1936)
36. W.G. Unruh, Notes on black hole evaporation. Phys. Rev. **D14**, 870–892 (1976)
37. J. Bisognano, E.H. Wichmann, On the duality condition for quantum fields. J. Math. Phys. **17**, 303–321 (1976)
38. G.L. Sewell, Quantum fields on manifolds: PCT and gravitationally induced thermal states. Ann. Phys. **141**, 201–224 (1982)
39. C. Holzhey, F. Larsen, F. Wilczek, Geometric and renormalized entropy in conformal field theory. Nucl. Phys. **B300**, 44 (1994)
40. P. Calabrese, J. Cardy, Entanglement entropy and quantum field theory. J. Stat. Mech. **0406**, P06002 (2004). arXiv:hep-th/0405152
41. H. Casini, M. Huerta, Entanglement entropy for the n-sphere. Phys. Lett. **B694**, 167–171 (2011). arXiv:1007.1813
42. H. Casini, M. Huerta, R.C. Myers, Towards a derivation of holographic entanglement entropy. JHEP **05**, 036 (2011). arXiv:1102.0440
43. I.E. Segal, A note on the concept of entropy. J. Math. Mech. **9**, 623–629 (1960)
44. A. Connes, E. Størmer, Entropy for automorphisms of II_1 von Neumann algebras. Acta. Math. **134**, 289–306 (1975)
45. N. Cabibbo, G. Parisi, Exponential hadronic spectrum and quark liberation. Phys. Lett. **B59**, 67–69 (1975)
46. A.M. Polyakov, Thermal properties of gauge fields and quark liberation. Phys. Lett. **B72**, 477–480 (1978)
47. L. Susskind, Lattice models of quark confinement at high temperature. Phys. Rev. **D20**, 2610–2018 (1979)
48. B. Sundborg, The Hagedorn transition, deconfinement, and $\mathcal{N} = 4$ SYM theory. Nucl. Phys. **B573**, 349 (2000). arXiv:hep-th/9908001
49. J. Hallin, D. Persson, Thermal phase transition in weakly interacting, large N_c QCD. Phys. Lett. **B429**, 232 (1998). arXiv:hep-th/9803224
50. A.M. Polyakov, Gauge fields and space-time. Int. J. Mod. Phys. **A17S1**, 119 (2002)
51. O. Aharony, J. Marsano, S. Minwalla, K. Papadodimas, M. Van Raamsdonk, The Hagedorn/deconfinement phase transition in weakly coupled large N gauge theories. Phys. Rev. **D74**, 105012 (2006). arXiv:hep-th/0310285
52. O. Aharony, J. Marsano, S. Minwalla, K. Papadodimas, M. Van Raamsdonk, A first order deconfinement transition in Yang-Mills theory on a small S^3. Phys. Rev. **D71**, 125018 (2005). arXiv:hep-th/0502149
53. S.W. Hawking, D.N. Page, Thermodynamics of black holes in anti de Sitter space. Commun, Math. Phys. **87**, 577–588 (1983)
54. E. Witten, Anti de Sitter space and holography. Adv. Theor. Math. Phys. **2**, 253–291 (1998). arXiv:hep-th/9802150
55. C. Thorn, Infinite N_c QCD at finite temperature: Is there an ultimate temperature? Phys. Lett. **B99**, 458 (1981)
56. J.M. Maldacena, Eternal black holes in anti de Sitter. JHEP **04**, 021 (2003). arXiv:hep-th/0106112

Chapter 12
Quantum Anomalous Hall Effect



Qikun Xue

Abstract Professor CN Yang founded the Institute for Advanced Studies at Tsinghua University (IASTU) in 1997. It quickly became the intellectual hub of Tsinghua campus, and attracted many world-renowned scientists as faculties or visiting professors. Professor Shoucheng Zhang, who was Professor CN Yang's disciple back in Stony Brook, used to visit the institute frequently. He introduced the then emerging field of topological insulator to the Tsinghua physics community, and has mentored several extremely talented young students. Moreover, his theoretical insights played a crucial role in the discovery of quantum anomalous Hall effect (QAHE) by the joint team led by Professor Qikun Xue in the department of physics at Tsinghua. The team consisted of several world-leading groups with expertise in material growth, characterization, and transport measurements led by Profs. Yayu Wang, Xucun Ma/Ke He and Li Lu. During the experimental exploration of QAHE, this joint team had been collaborating closely with Shoucheng's theoretical group at IASTU. In this article, we will give a brief review about the QAHE, how it was discovered experimentally, and Professor CN Yang's comments about the work.

One of the most exciting fields in modern condensed matter physics is the quantum Hall effect (QHE). In 1980, Klaus von Klitzing and coworkers discovered that in strong magnetic field, the two-dimensional electron gas (2DEG) exhibits quantized Hall resistivity ρ_{yx} plateaus at h/ie^2, where i is an integer, and meanwhile the four-terminal longitudinal resistivity ρ_{xx} drops to zero. The quantity $h/e^2 \sim 25.8k\Omega$ has since been called the quantum resistance. In the QHE regime, the 2DEG forms discrete Landau levels in the presence of strong magnetic field. All the bulk carriers are localized by impurities, but the one-dimensional (1D) edge states remain extended, and each edge channel contributes a quantized Hall conductance e^2/h. Moreover, the forward and backward chiral edge states at the opposite edges are

Q. Xue (✉)
Department of Physics, Tsinghua University, Beijing, China

Southern University of Science and Technology , Shenzhen, China
e-mail: xueqk@sustech.edu.cn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_12

immune to back-scattering, thus the electrons are dissipationless with a vanishing ρ_{xx}. Two years later, Daniel Tsui and Horst Stormer discovered the fractional QHE in super high mobility 2DEG in which strong electron correlations are dominant. The discoveries of integer and fractional QHE were awarded the Nobel physics prize in 1985 and 1998 respectively.

It was realized by David Thouless and coworkers that the quantum Hall state can be viewed as a special class of insulator with topologically nontrivial electronic structure. One can define a topological character for a filled energy band, which is known as the Chern number. The Chern number is proportional to the integral of the Berry curvature of the energy band over the whole Brillouin zone, which is zero for a usual energy band but is unity for a Landau level. The Hall conductance in the QHE state is the sum of the Chern numbers of all the occupied Landau levels in the unit of e^2/h.

An intriguing question to ask is whether it is possible to realize the QHE in zero magnetic field, which can be regarded as the quantized version of the anomalous Hall effect in ferromagnetic materials with spontaneous magnetization. The expected behavior of the QAHE is schematically illustrated in the right of Fig. 12.1, in comparison with the high field QHE experimental results on the left. The QAHE is not only a new topological quantum state, but is also important for the application of the dissipationless edge states in high-speed low energy consumption electronics. In 1988, F. D. M. Haldane proposed the first model for the QHE without Landau levels. It is based on the graphene lattice with the time reversal symmetry (TRS) broken by a periodic magnetic field, but the net magnetic flux is zero. The Haldane model played a central role in the theoretical development of topological phases of matter, but it is very challenging to realize experimentally.

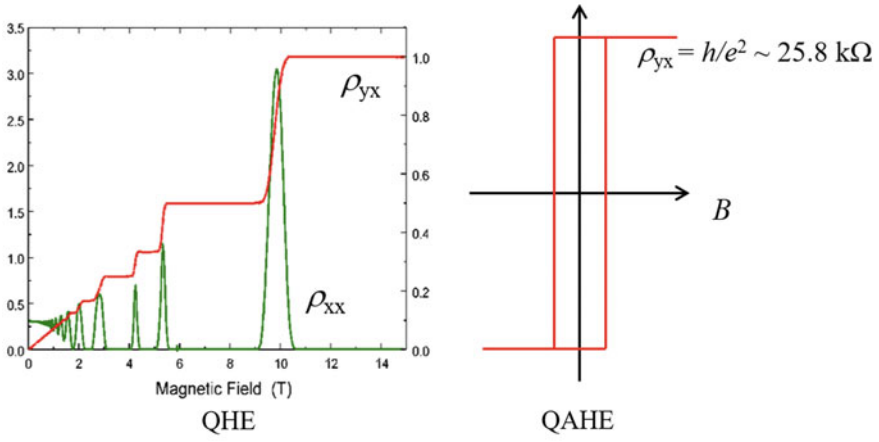


Fig. 12.1 Left: Experimental results of the QHE, in which quantized Hall plateaus form in strong magnetic field. Right: schematic illustration of the expected behavior of QAHE, in which the Hall resistivity is quantized in zero magnetic field

The discovery of topological insulator (TI) in 2005 brought new hope to the realization of QAHE. The topological property of a TI is induced by strong SOC, which causes the inversion of the bulk conduction and valence bands. In 2006, Xiaoliang Qi, who was still a graduate student at IASTU, collaborated with Yong-Shi Wu and Shoucheng Zhang to propose a general class of 2D models in which the charge Hall effect is topologically quantized in the absence of an external magnetic field. Two years later, Chaoxing Liu, another graduate student at IASTU, collaborated with Shoucheng Zhang et al. to propose a specific material, Mn doped HgTe quantum well, for realizing the QAHE. After the discovery of 3D TIs, it was proposed by Xi Dai, Zhong Fang, and Shoucheng Zhang et al. that magnetically doped Bi_2Se_3 family TIs are ideal candidates for the QAHE. Inspired by these theoretical proposals, the joint experimental team led by Professor Qikun Xue set out the journey for searching the QAHE in 2009.

The realization of QAHE requires the materials to be ferromagnetic, topological, and insulating simultaneously. The material must have topologically nontrivial bulk band structure so that there are topologically protected boundary states. Moreover, the sample must obtain long-range ferromagnetic order with its magnetization direction perpendicular to the sample plane so that the anomalous Hall effect is present. Finally, the sample must be truly insulating in the bulk so that only the chiral edge states contribute to charge transport. Apparently, finding such a magic material is a very challenging experimental task. Below we will briefly review how the necessary conditions for the QAHE were achieved step by step.

In order to realize the QAHE, ultrathin TI films with thickness in the order of a few quintuple layers (QLs) are necessary. Decreasing film thickness can reduce the bulk conduction and parallel conduction along the side surfaces. By using molecular beam epitaxy (MBE), we achieved well-controlled growth of high quality Bi_2Se_3 family TIs with thickness down to 1 QL. More importantly, the electronic structure of the TI films can be characterized *in situ* by using angle-resolved photoemission spectroscopy (ARPES) and scanning tunnelling microscopy (STM). The combination of MBE, APRES, and STM in one ultrahigh vacuum chamber was a unique setup by Professor Qikun Xue's group, and it played an indispensable role in the studies of magnetic TIs and other low dimensional quantum materials.

In order to create bulk-insulating TIs, we invent the band structure engineering technique to further eliminate the bulk carriers. Because Bi_2Te_3 and Sb_2Te_3 belong to the same TI family but with opposite types of bulk carriers, mixing them in the form of $(Bi_{1-x}Sb_x)_2Te_3$ ternary compounds can systematically tune the position of the Fermi level (E_F). We used MBE to grow a series of $(Bi_{1-x}Sb_x)_2Te_3$ films with *Sb* contents $0 \leq x \leq 1$, and ARPES electronic structure maps confirmed that the system remains topologically nontrivial for the entire x range. By combining ARPES and transport measurements, we found that the E_F lies within the bulk energy gap and the system becomes truly insulating for Sb content $x \sim 0.95$.

An important question immediately arises for the insulating TI, i.e., whether long-range ferromagnetic order can be established when it is doped with magnetic elements. Most ferromagnets are metals because itinerant electrons are essential in mediating the ferromagnetic order through the RKKY mechanism. When

charge carriers are depleted in insulating TI, it is expected that ferromagnetism will be strongly suppressed. To clarify this issue, we grew a series of 5 QL $Cr_{0.22}(Bi_xSb_{1-x})_{1.78}Te_3$ films with the same Cr doping level but different Bi:Sb ratio. Interestingly, all samples show clear anomalous Hall traces with nearly square-shaped hysteresis loops, even in the highly insulating regime. The coercive field and Curie temperature have very weak dependence on carrier type and density. The carrier independent ferromagnetism supports the van Vleck mechanism proposed by Xi Dai and Zhong Fang et al., in which the ferromagnetic order is mediated by the bulk conduction and valence bands instead of itinerant carriers.

Finally, band structure engineering by varied Bi/Sb ratio can tune the E_F into the large bulk energy gap, but to realize QAHE the E_F must precisely lie in the small energy gap at the Dirac point of the surface state opened by ferromagnetism. Because the Curie temperature of the magnetically doped TI is in the order of 10K, a rough estimate suggests that the energy gap due to broken-TRS is only in the order of 1meV. To control the E_F position to such precision, an external electrical gate is necessary. By using $SrTiO_3$ as the substrate to prepare TI thin films, we achieved carrier density variation of $\sim 3 \times 10^{13}\text{cm}^{-2}$ with a back-gate voltage (V_g) between $\pm$200V for a typical substrate thickness of 0.5mm.

The delicate control of the band structure by atomic-level control of thin film growth is not the whole story. To observe the QAHE, low-temperature high precision transport measurement is essential. While most groups in the world worked separately on either materials or on transport, this joint team had been collaborating closely on both aspects. Prof. Yayu Wang's transport group joined the efforts from the very early stage of the experiments, while Prof. Li Lu's group conducted the last-stage transport measurement at 30mK in a dilution fridge. The close-loop feedbacks between material and transport groups ensure the team to solve the most challenging problems very effectively. Through several years' efforts by many outstanding gradate students, it became clear that all the requirements for realizing the QAHE can be fulfilled in ultrathin Cr doped $(Bi_{1-x}Sb_x)_2Te_3$ TI films grown on $SrTiO_3$ substrate.

In October 2012, the QAHE was finally observed by the joint team. Figure 12.2 shows the magnetic field dependence of ρ_{yx} of the 5 QL $Cr_{0.15}(Bi_{0.1}Sb_{0.9})_{1.85}Te_3$ film at different V_{gS} measured at T=30mK. Staring from $V_g = 200$V, the AH resistance at zero magnetic field is merely $\sim 0.1h/e^2$ when the sample is strongly electron doped by a large positive gate voltage. With decreasing V_g, the AH resistivity increases rapidly, and for a range of V_g around the charge neutrality point $V_g^0 = -1.5$V, the ρ_{yx} value is perfectly quantized and forms a wide plateau, which is nearly identical as that expected for the QAHE behavior. With further decreasing of V_g to a large negative value, the ρ_{yx} value decreases again when too many holes are injected to the sample. Accompanying the formation of AH plateau, the longitudinal resistivity of the magnetic TI film decreases rapidly and approaches zero, indicating the dissipationless transport. These landmark signatures unambiguously indicate the discovery of the QAHE. The first observation of QAHE by the joint team was published in Science **340**, 167 (2013), and the results have soon been confirmed by several experimental groups around the world.

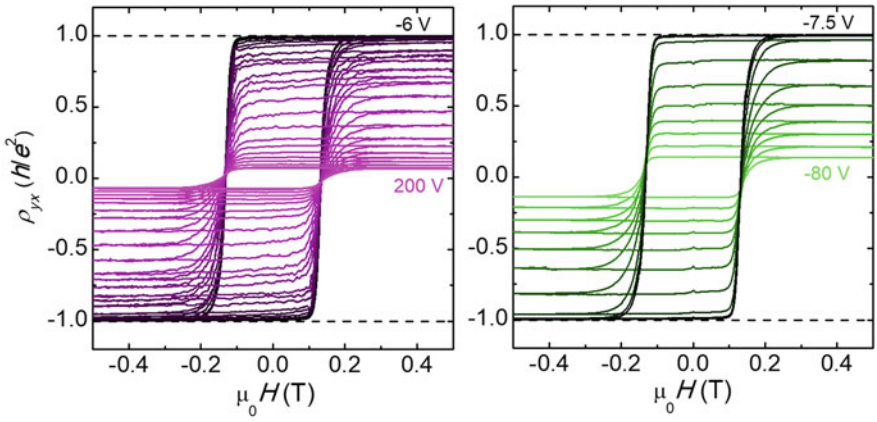


Fig. 12.2 The experimental results of the QAHE measured at T=30mK. Magnetic field dependence of ρ_{yx} at different V_gS show perfect quantization of the AH resistivity for a range of V_g near the charge neutrality point

Professor CN Yang heard about the experimental results on the QAHE, and invited us to give a talk in IASTU on this topic. After the talk, he gave a very high praise about the significance of the work, and then asked what is the most crucial factor for the success of the Tsinghua team. We answered that in addition to the technical strength, the collaborative spirit, mutual understanding and modesty among the group members are very important when addressing such a challenging task. Professor Yang remarked that our success indicates the traditional Chinese virtues could also be beneficial for scientific adventures.

The experimental realization of the QAHE effect in magnetically doped TI thin films concludes the decade-long search for the QHE without magnetic field. It also provides unambiguous evidence for the existence of the intrinsic mechanism of the AHE due to the Berry curvature of the band structure. The realization of QAHE opens new routes for many other novel quantum phenomena predicted previously, such as topological magnetoelectric effect, image magnetic monopoles, and chiral Majorana fermions. It will also facilitate the application of the dissipationless edge states in electronic devices, and the use of quantum resistance as new resistance standard.

In closing the story, we wish to emphasize again that the IASTU, founded and led by Prof. CN Yang, played a very special role in the eventual success of the QAHE experiment.

Chapter 13
Magic Superconducting States in Cuprates



Jiangping Hu, Xiaoli Dong, and Zhongxian Zhao

Abstract We show that there exists "magic" superconducting states in cuprates, where the superconductivity is stronger and more robust. These magic superconducting phases take place at the doping of $x \sim 1/n^2$ (n=3 and 4, respectively) in $La_{2-x}Sr_xCuO_4$ (LSCO). We argue that these magic states are superconducting mother states intertwined by both the pair density wave and the uniform superconductivity. Their existence not only enriches the phase diagram but also help to reveal the elusive pairing mechanism in cuprates.

One hallmark of strongly correlated electron systems is the dependence of physical properties on doping concentrations. For example, in the two dimensional fractional quantum hall (FQH) systems, discovered near 40 years ago, the doping level controls their phase diagrams. At "magic" discrete fractional numbers, such as $1/3$, robust FQH states characterized by the quantized plateaus of Hall conductivity emerge [1]. Another example is the unconventional high temperature superconductors [2], cuprates. In the typical phase diagram of cuprates, the horizontal axis is labeled by the doping concentration of carriers, either holes or electrons. One of essential research efforts in this field is to understand the phases and their physical properties varying as a function of the doping concentration.

J. Hu
Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China

Kavli Institute of Theoretical Sciences, University of Chinese Academy of Sciences, Beijing, China

X. Dong · Z. Zhao (✉)
Beijing National Laboratory for Condensed Matter Physics and Institute of Physics, Chinese Academy of Sciences, Beijing, China

School of Physical Sciences, University of Chinese Academy of Sciences, Beijing, China

Songshan Lake Materials Laboratory , Guangdong, China
e-mail: zhxzhao@iphy.ac.cn

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

M.-L. Ge, Y.-H. He (eds.), *Dialogues Between Physics and Mathematics*,
https://doi.org/10.1007/978-3-031-17523-7_13

However, the effect of the doping concentration on the electronic states of cuprates is very complicated because of strong competing between several degrees of freedoms. Depending on the doping concentration, many different states have been discovered. Without doping, there is an antiferromagnetic insulating state. At some special values of doping concentration, states with other competing long-range orders can appear. For example, at the $1/8$ hole doping, a static stripe order can be developed, where the carriers form ordered one-dimensional chains to compromise kinetic and interaction energy. Those competing orders typically take place at discrete doping values while the superconductivity(SC) seems to be continuous in the majority of doping regions. Understanding the relationship between those competing orders and superconductivity can be the key step to resolve the elusive pairing mechanism in cuprates [3, 4].

Recently, there has been reasonable experimental evidence for the existence of pair density wave (PDW) [5] in cuprates. Here the PDW order appears in the absence of external magnetic field while previously it was only known to take place under a strong magnetic field which breaks time reversal symmetry. One simple explanation about the origin of the PDW is the coexisting of a charge density wave(CDW) and a uniform superconductivity(SC), in which the PDW order is induced naturally as a descendant of these two coexisting orders. Within this explanation, the PDW is not an exotic phase so that it would be less meaningful than what we will discuss below.

In fact, the experimental evidence in cuprates is sufficiently dramatic for us to consider a different origin behind the emergence of the PDW order: the experimental results do not support a ubiquitous coexisting of CDW and SC while they strongly indicate that the PDW coexists with a homogeneous superconductivity [5]. The coexistence of the PDW and the uniform SC not only implies that the PDW is not a descendant order but also indicates that it is not a competing order but an intertwining order to the SC.

As a PDW order is featured by a certain momentum wavevector, the development of such an order is typically preferred at certain doping levels. With this preference, the different nature between intertwining and competing can lead to dramatic experimental consequences. In fact, on one hand, if the PDW is a competing order to the SC, there are certain doping levels where the PDW orders are stronger so that the SC orders are weakened [6], which will suggest that there are "magic" doping levels with much weaker SC states; on the other hand, if the PDW is an intertwining order to the SC, there could be certain "magic" doping levels with even stronger SC states because the PDW can strengthen the overall superconducting pairing. Therefore, the existence of robust SC states at "magic" doping levels can be a strong evidence for the coexisting and intertwining nature between the PDW and the uniform SC.

Here we discuss experimental evidence to demonstrate the existence of "magic" doping levels in $La_{2-x}Sr_xCuO_4$ (LSCO), in which the SC states are much stronger and more robust. There are the magic superconducting phases at the magic dopings of $1/n^2$ (n=3 and 4, respectively) [7–10].

Among the hole-doped cuprates, $La_{2-x}Sr_xCuO_4$ (LSCO) system tolerates the dopant (Sr) in a wide range of concentration, displaying a highly tunable superconducting critical T_c and unique electronic states, thus providing an ideal platform.

Although the superconducting critical temperature T_c of LSCO shows in general a dome-shaped doping dependence, a "plateau" at 30 K always appears around the doping level x = 1/9 ($\sim$ 0.11) in the underdoping regime [7, 8, 11, 12]. More important, as illustrated by the temperature-dependent magnetic susceptibilities shown in Fig. 13.1a and b, there are extremely sharp single transitions of T_{c1} = 15 K and T_{c2} = 30 K for LSCO crystals with $x \approx 0.063$ and 0.11, respectively. In contrast, in the same series of high-quality crystals, the superconducting transition becomes much broader whenever the T_c is away from T_{c1} or T_{c2}, manifested as broad field-cooling (FC) differential susceptibility ($d\chi/dT$) curves with multiple peaks structure (Fig. 13.1b). We accordingly name them as magic superconducting phases since they locate at magic dopings of $1/n^2$ (n=3 and 4, respectively).

To see the robustness of the "magic" superconducting states, we examine their transition temperature T_c under external magnetic field. The T_cs of the two magic superconducting phases are very robust under an externally applied field H and behave characteristically different from that of the other superconducting states. As

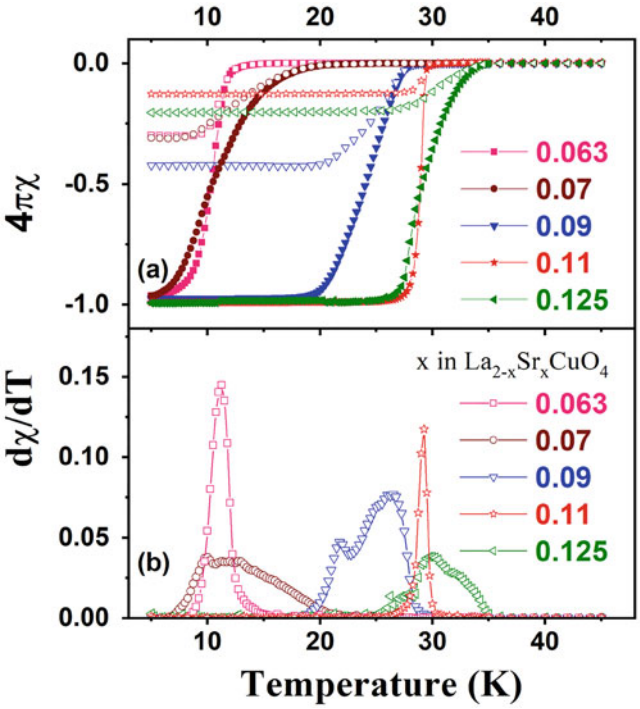


Fig. 13.1 (**a**) The magnetic susceptibilities for $La_{1-x}Sr_xCuO_4$ (LSCO) single crystals, with x = 0.063 ($\sim$ 1/16), 0.07, 0.09, 0.11 ($\sim$1/9) and 0.125 (=1/8), as functions of temperature under zero-field cooling (ZFC) and field cooling (FC) modes in 5 Oe field along c-axis. (**b**) Corresponding differential curves of FC susceptibility ($d\chi/dT$) as function of temperature. Extremely sharp single transitions of T_{c1} = 15 K and T_{c2} = 30 K are obvious for x $\sim$ 0.063 and $\sim$ 0.11 samples, respectively [7, 9]

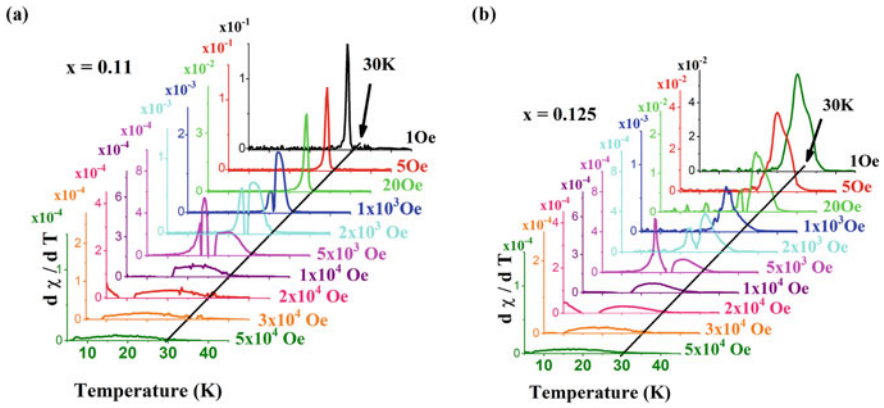


Fig. 13.2 FC dχ/dT curves under various fields for LSCO single crystals with x = (**a**) 0.11 and (**b**) 0.125. As guided by solid black lines, the 30K magic superconducting phase of 0.11 crystal is robust under fields up to 5 T, on the other hand, the apparent initial T_c (37 K) of 0.125 crystal decreases rapidly upon increasing H then stays at 30 K for H $\geq$ 1 T [9]

illustrated in Fig. 13.2a, the T-dependence of FC dχ/dT of the x = 0.11 crystal as a function of H clearly shows that while the transition width ΔT_c of the 30K magic superconducting phase of x = 0.11 crystal increases with increasing magnetic field H, the onset of the superconducting transition at 30 K is surprisingly robust with little field dependence. In contrast, as seen in Fig. 13.2b, the width of broad FC dχ/dT curve of x = 0.125 sample does not broaden with increasing H. Instead, the initial apparent T_c = 37 K is continuously suppressed until it reaches T_{c2} = 30K at H $\sim$ 1 T. Then again, ΔT_c starts to increase rapidly while the T_{c2} exhibits little field dependence for 1 T $\leq$ H $\leq$ 5 T. The extra peak-like features that appear in both crystals for H $\geq$ 0.1 T is due to the hump of the magnetization observed and attributed to the first-order vortex phase melting transition in the vicinity of the irreversibility temperature [9].

The same robustness of T_{c1} (=15 K) occurs in x = 0.063 crystal (Fig. 13.3a). However, as summarized in Fig. 13.3, whenever T_c is away from 15 K, such as that of x = 0.07 crystal (T_c = 22 K at H = 1Oe), the broad ΔT_c becomes narrower as T_c decreases with increasing H and then starts to increase at H = 0.5 T after T_c reaches 15 K from above (see inset of Fig. 13.3b). For H $\sim$ 1 T, the T_c remains at 15 K and ΔT_c increases rapidly. Similar behavior is also observed in x = 0.09 crystal (Fig. 13.3c): the transition exhibits a two-peak structure, one at T_{c2} and the other at 23 K. The ΔT_c of the second transition shrinks and disappears for H $\sim$ 0.1 T. Concomitantly, the ΔT_c of T_{c2} broadens upon increasing H with T_{c2} exhibiting little field dependence for 0.1 T $\leq$ H $\leq$ 5 T. Therefore, when tuned by magnetic field, the superconducting transitions of x = 0.063 and 0.07 crystals converge to T_{c1} = 15 K and that of x = 0.09, 0.11 and 0.125 crystals to T_{c2} = 30 K. Furthermore, the anomalous broadening and narrowing of the ΔT_c under field are closely tied to the approaching, from above, to the two robust superconducting

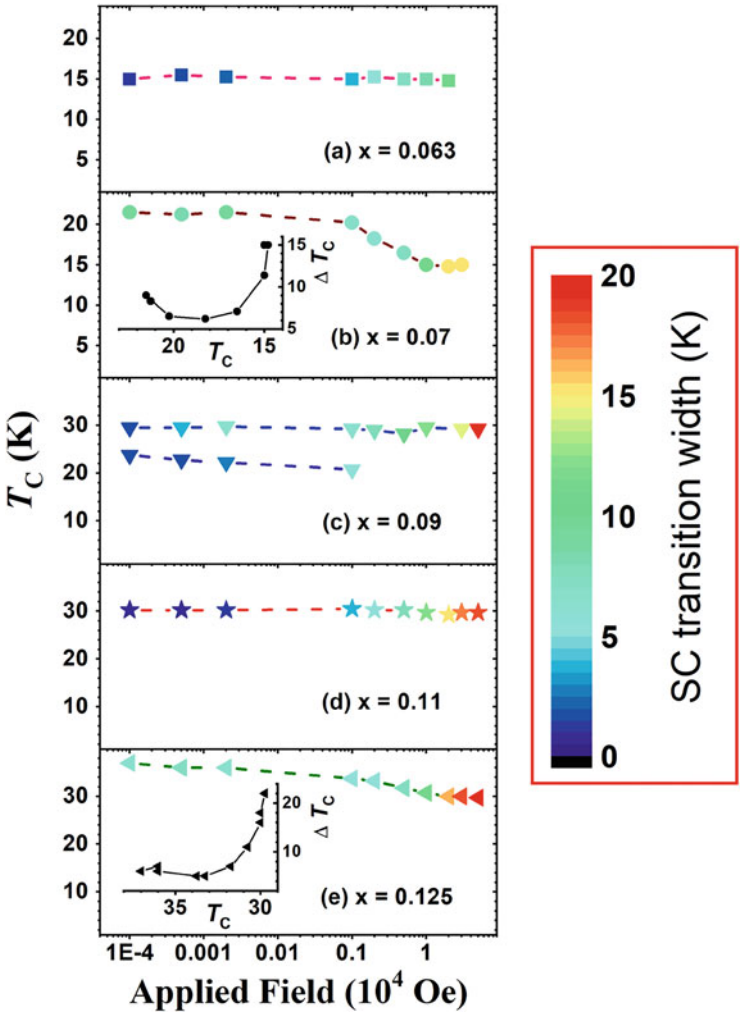


Fig. 13.3 Field dependences of T_c and transition width (color coded) for LSCO single crystals with x of (**a**) 0.063 (∼1/16), (**b**) 0.07, (**c**) 0.09, (**d**) 0.11 (∼1/9) and (**e**) 0.125 (=1/8). Insets are ΔT_c vs. T_c for 0.07 and 0.125 crystals, respectively [9]

states through H-tuning. This is characteristically different from that of conventional type II superconductors or granular superconductors where there is only one flexible superconducting state with monotonic decreasing of T_c under magnetic field.

If the "magic" superconducting phases host the intertwining PDW and SC orders, the vortex phase diagram can also be different from the other superconducting states. In the following we discuss the experimental evidence for the intertwining state from the vortex phase diagram.

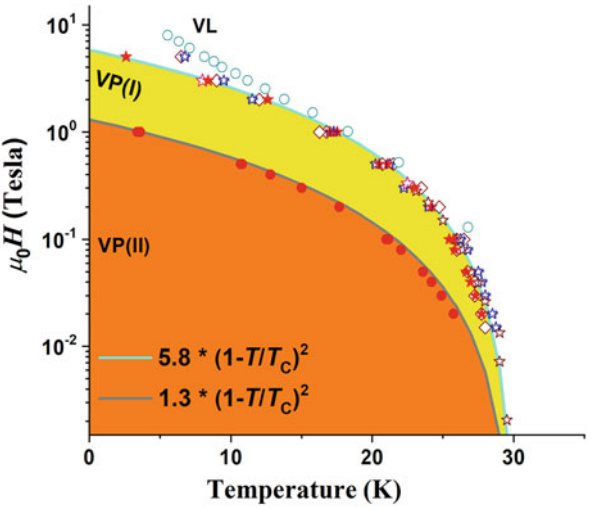


Fig. 13.4 Magnetic phase diagram of LSCO crystal with x = 1/9. Various characteristic temperatures identified in [10]. Three phases, vortex liquid(VL), vortex solid(VP(I))), vortex soild (VP(II)), are specified

In [10], it has been identified that the standard fitting to the diamagnetic signal in LSCO at the x=1/9 magic doping level results in two different phase boundaries: the standard vortex melting transition from a vortex solid to a vortex liquid marked by T_{m1} and a hidden phase transition marked by T_{m2}. In Fig. 13.4, we redraw the vortex phase diagram of the magic superconducting phase at x = 1/9. The detailed explanation of the figure can be found in [10]. The phase boundaries $T_{m1}(H)$ and $T_{m2}(H)$ are extracted from critical fittings to the reversible magnetization. They are true equilibrium phase boundaries. We label the vortex solid phase I (VP(I)) between phase boundaries $T_{m1}(H)$ and $T_{m2}(H)$ and vortex solid phase II (VP(II)) below $T_{m2}(H)$ line. It is important to note that the extraction of $T_{m1}(H)$ and $T_{m2}(H)$ works only for x = 1/9. For both x = 0.09 and 0.125 crystals, only $T_{m1}(H)$ can be obtained. Previously, this observation was interpreted as the coexistence of the SC and the spin-density wave (SDW). However, as the SDW order is typically stronger when the doping is lower, it is very difficult to understand why this phenomenon was only observed at the magic doping level.

We argue that the above observed vortex phase diagram supports the intertwining PDW and SC orders in "magic" superconducting phases. In fact, the existence of the hidden VP(I) can be easily understood in an intertwining state with both the PDW and the uniform SC. When the ordering wavevector of the PDW order commensurates with the reciprocal wavevector of the vortex lattice, the PDW order can be strongly affected, which is the reason behind the appearance of the VP(I) phase only in the intertwining states. Furthermore, unlike the uniform SC which is weakened by an external magnetic field, the PDW order can even be strengthened by the vortex lattice, which can further explain why the PDW and the SC appear to compete with each other in the VP(I) phase as discussed in [10] while in the VP(II) phase, they seem to be attractive with each other. Therefore, the VP(I) phase represents an intriguing interplay among the PDW, SC and the vortex lattice.

We can extend above analysis to argue even more remarkable consequences. If the PDW order can be strengthened by the vortex lattice, it is highly possible that an SC state without the PDW order may intertwine with the PDW order when a high magnetic field is applied. Namely, the " magic" superconducting states can be induced by a high external magnetic field. This phenomenon was observed experimentally.

It has been reported in [9] that the Meissner signals, i.e., FC diamagnetic susceptibilities χ_{FC} of samples with different doping, tend to group corresponding to the magic superconducting phases under high external magnetic field. χ_{FC} (at 5K) of 14 LSCO crystals were shown in Fig. 13.5. For H <0.1 T, there is no systematic χ_{FC} dependence on the sample size and the doping level. However, for H >0.1 T, χ_{FC}'s start to group into two distinct curves. For H > 1 T, χ_{FC}'s of x = 0.063, 0.07 and 0.22 samples have merged into one universal curve while those of x = 0.09, 0.092, 0.10, 0.11, 0.125 and 0.15 samples fall on another universal curve [9]. Similar behavior was also found in the optimal-doped $Bi_2Sr_{1.6}La_{0.4}CuO_y$ (Bi2201) and electron-doped $Pr_{0.88}LaCe_{0.12}CuO_4$ (PLCCO) as shown in Fig. 13.5. The "grouping" of χ_{FC}'s does not depend on the size, the shape or even the hole types and concentration of the crystals; it depends only on the magic superconducting phases with bulk T_{c1} = 15K or T_{c2} = 30K. Therefore, each universal χ_{FC}-H curve is an intrinsic property of the magic superconducting phases. This again indicates that the magic superconducting phases are indeed genuine thermodynamic phases, which can be induced by strong external magnetic field.

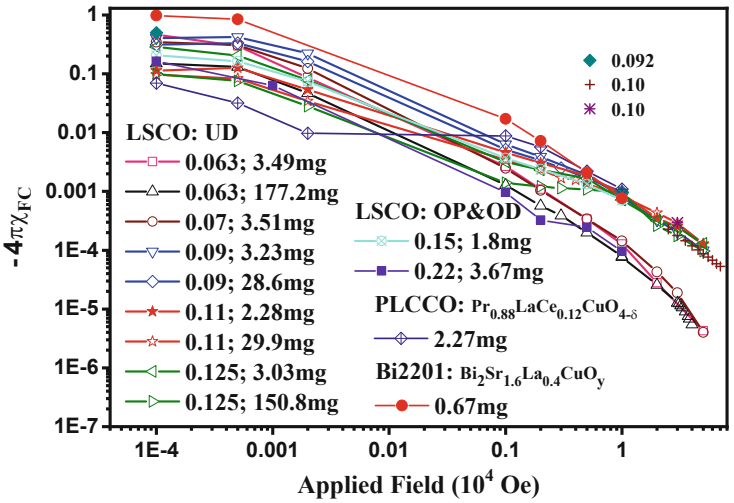


Fig. 13.5 Field dependences of χ_{FC} at 5K for 16 single-CuO-plane cuprate crystals. Doping levels and masses are indicated. Data for doping levels of 0.092 and 0.10 are from literatures [13–15]. Noted that the "grouping" of χ_{FC}(H)'s does not depend on the size, the shape or even the hole types and concentration of the crystals, it depends only on the magic superconducting phases [9]

To summarize, there are the "magic" superconducting states which can be understood as an intertwining state of the PDW and the uniform SC. This observation not only suggests that the PDW is a mother order but also explains why the PDW state remains so elusive [5] as the PDW order only appears at these "magic" doping levels without external magnetic field.

There are broader issues in cuprates which can be reexamined based on the "magic" superconducting states. First, our finding will suggest that the PDW may be the primary fluctuation order in the pseudogap regime in cuprates. The fluctuation of a PDW order can be very different from that of a uniform superconductivity as it naturally induces a charge fluctuation, which may explain the complex phenomena observed in the pseudogap regime [3]. Second, the "magic" doping suggests that the PDW orders have fixed "magic" wavevectors. This suggests the PDW order must stem from strong electron-electron interactions with very short range pairing. It can not be a Fermi surface property based on Fermi-liquid theory. Finally, the intertwining between the PDW and the uniform SC suggests that they must have a common microscopic origin. It is fascinating to construct a simple theoretical model to account for this intertwining nature in the future.

Acknowledgments We thanks Prof. F. Zhou, Prof. X.J. Zhou and Prof. P.C. Dai for valuable help.

References

1. M.E. Cage, K. Klitzing, A.M. Chang, et al., *The Quantum Hall Effect* (Springer Sci & Business Media, 2012)
2. D.M. Ginsberg, (ed.) *Physical Properties of High Tc Superconductors* (World Scientific, 1998)
3. E. Fradkin, S.A. Kivelson, J.M. Tranquada, Rev. Mod. Phys. **87**, 457 (2015)
4. E. Dagotto, Rev. Mod. Phys. **66**, 763 (1994)
5. D.F. Agterbert, et al., Annu. Rev. Condens. Matter Phys. **11**, 231 (2020)
6. H.D. Chen, O. Vafek, A. Yazdani, S.C. Zhang, Phys. Rev. Lett. **93**, 187002 (2004)
7. F. Zhou, P.H. Hor, X.L. Dong, W.X. Ti, J.W. Xiong, Z.X. Zhao, Phys. C Supercond. Appl. **408**, 430 (2004)
8. F. Zhou, W.X. Ti, J.W. Xiong, Z.X. Zhao, X.L. Dong, P.H. Hor, Z.H. Zhang, W.K. Chu, Supercond. Sci. Technol. **16**, L7 (2003)
9. X.L. Dong, P.H. Hor, F. Zhou, Z.X. Zhao, Solid State Commun. **145**, 173 (2008)
10. X.L. Dong, P.H. Hor, F. Zhou, Z.X. Zhao, Solid State Commun. **152**, 1513 (2012)
11. T. Nagano, Y. Tomioka, Y. Nakayama, K. Kishio, K. Kitazawa, Phys. Rev. B Condens Matter. **48**, 9689 (1993)
12. P.G. Radaelli, et al., Phys. Rev. B Condens Matter. **49**, 4163 (1994)
13. R. Gilardi, S. Streule, N. Momono, M. Oda, J. Mesot, Eur. Phys. J. B **47**, 231 (2005)
14. T. Sasagawa, Y. Togawa, J. Shimoyama, A. Kapitulnik, K. Kitazawa, K. Kishio, Phys. Rev. B **61**, 1610 (2000)
15. Y.M. Huh, J.E. Ostenson, F. Borsa, V.G. Kogan, D.K. Finnemore, A. Vietkin, A. Revcolevschi, M.H. Julien, Phys. Rev. B **63**, 064512 (2001)