Helge Holden
Ragni Piene
*Editors*

# The Abel Prize 2003–2007

## The First Five Years

ABEL PRISEN

DET NORSKE VIDENSKAPS-AKADEMI

Springer

The Abel Prize

Niels Henrik Abel 1802–1829
The only contemporary portrait of Abel, painted by Johan Gørbitz in 1826
© Matematisk institutt, Universitetet i Oslo

Helge Holden • Ragni Piene
Editors

# The Abel Prize

## 2003–2007 The First Five Years

ABEL PRISEN

DET NORSKE VIDENSKAPS-AKADEMI

Springer

Helge Holden
Department of Mathematical Sciences
Norwegian University
of Science and Technology
7491 Trondheim
Norway
holden@math.ntnu.no

Ragni Piene
Centre of Mathematics for Applications
and Department of Mathematics
University of Oslo
0316 Oslo
Norway
ragnip@math.uio.no

Photos on the cover of the book:

Photo of Jean-Pierre Serre by Arash. A. Nejad/Scanpix
Photos of Sir Michael Atiyah and Isadore Singer by Anne-Lise Flavik
Photo of Peter Lax by Knut Falch/Scanpix
Photo of Lennart Carleson by Knut Falch/Scanpix
Photo of S.R. Srinivasa Varadhan by Morten F. Holm/Scanpix

All photos in connection with the Abel Prize ceremonies are taken by Scanpix

ISBN 978-3-642-01372-0 e-ISBN 978-3-642-01373-7
DOI 10.1007/978-3-642-01373-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009926160

Mathematics Subject Classification (2000): 00-02; 00A15; 01A70

*Cover design*: WMXDesign, Heidelberg

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

# Preface

In 2002, the year marking the bicentennial of Abel's birth, the Norwegian Parliament established the Niels Henrik Abel Memorial Fund with the objective of creating an international prize for outstanding scientific work in the field of mathematics—the Abel Prize.

In this book we would like to present the Abel Prize and the Abel Laureates of the first five years. The book results from an initiative of the Mathematics section of the Norwegian Academy of Science and Letters. It is intended as the first volume in a series, each volume comprising five years.

The book starts with the history of the Abel Prize—a story that goes back more than a hundred years—written by Abel's biographer, Arild Stubhaug. It is followed by Nils A. Baas' biographical sketch of Atle Selberg; at the opening ceremony of the Abel Bicentennial Conference in Oslo in 2002, an Honorary Abel Prize was presented to Atle Selberg.

There is one part for each of the years 2003–2007. Each part starts with an autobiographical piece by the laureate(s). Then follows a text on the laureate's work: Pilar Bayer writes on the work of Jean-Pierre Serre, Nigel Hitchin on Atiyah–Singer's Index Theorem, Helge Holden and Peter Sarnak on the work of Peter Lax, Tom Körner on Lennart Carleson, and Terry Lyons on Srinivasa Varadhan. Each part contains a complete bibliography and a curriculum vitae, as well as photos—old and new.

The DVD included in the book contains the interviews that Martin Raussen and Christian Skau made with each laureate in connection with the Prize ceremonies in the years 2003–2007. Every year except the first, the interviews were broadcast on Norwegian national television. Transcripts of all interviews have been published in the *EMS Newsletter* and *Notices of the AMS*.

We would like to express our gratitude to the laureates for collaborating with us on this project, especially for providing the autobiographical pieces and the photos. We would like to thank the mathematicians who agreed to write about the laureates, and thus are helping us in making the laureates' work known to a broader audience.

Thanks go to Martin Raussen and Christian Skau for letting us use the interviews, to David Pauksztello for his translations, to Marius Thaule for his LaTeX expertise and the preparation of the bibliographies, and to Anne-Marie Astad of the

Norwegian Academy for Science and Letters for her help with the interviews and photos. HH gratefully acknowledges support from the Centre for Advanced Study at the Norwegian Academy of Science and Letters in Oslo during the academic year 2008–09.

The technical preparation of the manuscript was financed by the Niels Henrik Abel Memorial Fund.

Oslo Helge Holden and Ragni Piene
July 15, 2009

# Contents

**The History of the Abel Prize** . . . . . . . . . . . . . . . . . . . . . . . . . . 1
Arild Stubhaug
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

**2002—an Honorary Abel Prize to Atle Selberg** . . . . . . . . . . . . . . . 7
Nils A. Baas
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

**2003: Jean-Pierre Serre**

**Jean-Pierre Serre: Mon premier demi-siècle au Collège de France** . . . . 13
**My First Fifty Years at the Collège de France** . . . . . 13
Marc Kirsch

**Jean-Pierre Serre: An Overview of His Work** . . . . . . . . . . . . . . . . 33
Pilar Bayer
Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
1 The Beginnings . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
1.1 Homotopy Groups of Spheres . . . . . . . . . . . . . . . . . . 35
1.2 Hochschild–Serre Spectral Sequence . . . . . . . . . . . . . . 36
1.3 Sheaf Cohomology of Complex Manifolds . . . . . . . . . . . 36
1.4 The Amsterdam Congress . . . . . . . . . . . . . . . . . . . . 37
2 Sheaf Cohomology . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.1 FAC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
2.2 GAGA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 38
2.3 Cohomology of Algebraic Varieties . . . . . . . . . . . . . . . 38
3 Lie Groups and Lie Algebras . . . . . . . . . . . . . . . . . . . . 40
4 Local Algebra . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
5 Projective Modules . . . . . . . . . . . . . . . . . . . . . . . . . 41
6 Algebraic Number Fields . . . . . . . . . . . . . . . . . . . . . . 42
7 Class Field Theory . . . . . . . . . . . . . . . . . . . . . . . . . 43
7.1 Geometric Class Field Theory . . . . . . . . . . . . . . . . . . 43

7.2 Local Class Field Theory . . . . . . . . . . . . . . . . . . . . 45
7.3 A Local Mass Formula . . . . . . . . . . . . . . . . . . . . . 46
8 $p$-adic Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
9 Group Cohomology . . . . . . . . . . . . . . . . . . . . . . . . . 47
9.1 Cohomology of Profinite Groups and $p$-adic Lie Groups . . . . 47
9.2 Galois Cohomology . . . . . . . . . . . . . . . . . . . . . . . 48
9.3 Galois Cohomology of Linear Algebraic Groups . . . . . . . . 48
9.4 Self-dual Normal Basis . . . . . . . . . . . . . . . . . . . . . 50
9.5 Essential Dimension . . . . . . . . . . . . . . . . . . . . . . 50
10 Discrete Subgroups . . . . . . . . . . . . . . . . . . . . . . . . 51
10.1 Congruence Subgroups . . . . . . . . . . . . . . . . . . . . . 52
10.2 Cohomology of Arithmetic Groups . . . . . . . . . . . . . . . 53
11 Arithmetic of Algebraic Varieties . . . . . . . . . . . . . . . . . 54
11.1 Modular Curves . . . . . . . . . . . . . . . . . . . . . . . . . 54
11.2 Varieties Over Finite Fields . . . . . . . . . . . . . . . . . . 56
11.3 Number of Points of Curves Over Finite Fields . . . . . . . . . 56
11.4 Diophantine Problems . . . . . . . . . . . . . . . . . . . . . 58
12 Field Theory . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59
13 Galois Representations . . . . . . . . . . . . . . . . . . . . . . 62
13.1 Hodge–Tate Modules . . . . . . . . . . . . . . . . . . . . . . 62
13.2 Elliptic Curves and $\ell$-adic Representations . . . . . . . . . . 62
13.3 Modular Forms and $\ell$-adic Representations . . . . . . . . . . 65
13.4 Abelian Varieties and $\ell$-adic Representations . . . . . . . . . 69
13.5 Motives . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
14 Group Theory . . . . . . . . . . . . . . . . . . . . . . . . . . . 73
14.1 Representation Theory . . . . . . . . . . . . . . . . . . . . . 73
14.2 Algebraic Groups . . . . . . . . . . . . . . . . . . . . . . . . 76
14.3 Finite Subgroups of Lie Groups and of Algebraic Groups . . . 77
15 Miscellaneous Writings . . . . . . . . . . . . . . . . . . . . . . 78
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80

**List of Publications for Jean-Pierre Serre** . . . . . . . . . . . . . . . . 81

**Curriculum Vitae for Jean-Pierre Serre** . . . . . . . . . . . . . . . . . 95

**2004: Sir Michael Atiyah and Isadore M. Singer**

**Autobiography** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
Sir Michael Atiyah

**Autobiography** . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 109
Isadore M. Singer

**The Atiyah–Singer Index Theorem** . . . . . . . . . . . . . . . . . . . 115
Nigel Hitchin
1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 115

2 Background . . . 115
2.1 The Index . . . 115
2.2 Riemann–Roch . . . 116
2.3 The Beginning . . . 117
2.4 The Signature . . . 118
2.5 Hirzebruch–Riemann–Roch . . . 121
2.6 The Dirac Operator . . . 123
3 The Integer Index . . . 126
3.1 Formulation of the Theorem . . . 126
3.2 Integrality Theorems . . . 128
3.3 Positive Scalar Curvature . . . 129
3.4 Gauge-Theoretic Moduli Spaces . . . 130
4 The Equivariant Index . . . 132
4.1 K-Theory . . . 132
4.2 Fixed Point Theorems . . . 134
4.3 Rigidity Theorems . . . 137
5 The mod 2 Index . . . 139
5.1 Real K-Theory . . . 139
5.2 Theta Characteristics . . . 140
5.3 Positive Scalar Curvature . . . 140
6 The Index for Families . . . 141
6.1 Fredholm Operators . . . 141
6.2 Jumping of Dimension . . . 142
7 The Local Index Theorem . . . 144
7.1 The Heat Kernel . . . 144
7.2 The Eta Invariant . . . 146
7.3 Quantum Field Theory . . . 147
7.4 The Supersymmetric Proof . . . 148
References . . . 149

**List of Publications for Sir Michael Atiyah** . . . 151

**List of Publications for Isadore M. Singer** . . . 165

**Curriculum Vitae for Sir Michael Francis Atiyah, OM, FRS, FRSE** . . . 173

**Curriculum Vitae for Isadore Manual Singer** . . . 177

**2005: Peter D. Lax**

**Autobiography** . . . 181
Peter D. Lax

**A Survey of Peter D. Lax's Contributions to Mathematics** . . . 189
Helge Holden and Peter Sarnak
1 Introduction . . . 189

2 Partial Differential Equations—General Results . . . . . . . . . . . 191
3 Difference Approximations to Partial Differential Equations . . . . . 195
4 Hyperbolic Systems of Conservation Laws . . . . . . . . . . . . . . 196
5 Integrable Systems . . . . . . . . . . . . . . . . . . . . . . . . . 203
6 Lax–Phillips Scattering Theory . . . . . . . . . . . . . . . . . . . 208
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . 211

**List of Publications for Peter D. Lax** . . . . . . . . . . . . . . . . . 215

**Curriculum Vitae for Peter David Lax** . . . . . . . . . . . . . . . . . 229

**2006: Lennart Carleson**

**Reflections on My Life as a Mathematician** . . . . . . . . . . . . . . 233
Lennart Carleson

**Carleson for Beginners** . . . . . . . . . . . . . . . . . . . . . . . . 241
Tom Körner
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . 268

**List of Publications for Lennart Carleson** . . . . . . . . . . . . . . . 269

**Curriculum Vitae for Lennart Axel Edvard Carleson** . . . . . . . . . . 275

**2007: S.R. Srinivasa Varadhan**

**Autobiography** . . . . . . . . . . . . . . . . . . . . . . . . . . . . 279
S.R. Srinivasa Varadhan

**A Personal Perspective on Raghu Varadhan's Role in the Development of Stochastic Analysis** . . . . . . . . . . . . . . . . . . . . . . . . 287
Terry Lyons
1 A Great Day for the Coin Flippers . . . . . . . . . . . . . . . . . . 287
2 Stochastic Analysis . . . . . . . . . . . . . . . . . . . . . . . . . 287
3 Varadhan . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 290
4 Independent or Uncorrelated . . . . . . . . . . . . . . . . . . . . 290
4.1 A Simple Example . . . . . . . . . . . . . . . . . . . . . . . 291
5 Diffusion Equations and PDE . . . . . . . . . . . . . . . . . . . . 294
5.1 Elliptic PDEs in Non-divergence Form with Continuous Elliptic Coefficients—Uniqueness of Solution . . . . . . . . . 295
5.2 First-Order PDEs and Large Deviations . . . . . . . . . . . . . 297
5.3 PDEs and Probability in the 60s . . . . . . . . . . . . . . . . . 298
6 The Support Theorem—Understanding the Itô Differential Equation 299
6.1 On Extending the Wong–Zakai Theorem . . . . . . . . . . . . 299
7 The Donsker–Varadhan Theory of Large Deviations . . . . . . . . . 304
7.1 Short Time Behaviour—Geodesics . . . . . . . . . . . . . . . 305
7.2 Long Time Behaviour—Occupation Measures . . . . . . . . . 308

8 Hydrodynamical Limits, Interacting Particles and Other Questions . 309
8.1 Homogenised Hamilton–Jacobi Equations . . . . . . . . . . . 310
9 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 310
References . . . . . . . . . . . . . . . . . . . . . . . . . . . . 311

**List of Publications for S.R. Srinivasa Varadhan** . . . . . . . . . . . . . . 313

**Curriculum Vitae for Sathamangalam Rangaiyenga Srinivasa Varadhan** 321

**The Abel Committee** . . . . . . . . . . . . . . . . . . . . . . . . . 323

**The Board for the Niels Henrik Abel Memorial Fund** . . . . . . . . . . . 325

# The History of the Abel Prize

**Arild Stubhaug**

On the bicentennial of Niels Henrik Abel's birth in 2002, the Norwegian Government decided to establish a memorial fund of NOK 200 million. The chief purpose of the fund was to lay the financial groundwork for an annual international prize of NOK 6 million to one or more mathematicians for outstanding scientific work. The prize was awarded for the first time in 2003.

That is the history in brief of the Abel Prize as we know it today. Behind this government decision to commemorate and honor the country's great mathematician, however, lies a more than hundred year old wish and a short and intense period of activity.

Volumes of Abel's collected works were published in 1839 and 1881. The first was edited by Bernt Michael Holmboe (Abel's teacher), the second by Sophus Lie and Ludvig Sylow. Both editions were paid for with public funds and published to honor the famous scientist. The first time that there was a discussion in a broader context about honoring Niels Henrik Abel's memory, was at the meeting of Scandinavian natural scientists in Norway's capital in 1886. These meetings of natural scientists, which were held alternately in each of the Scandinavian capitals (with the exception of the very first meeting in 1839, which took place in Gothenburg, Sweden), were the most important fora for Scandinavian natural scientists. The meeting in 1886 in Oslo (called Christiania at the time) was the 13th in the series. At the meeting's farewell dinner, the Swedish mathematician Gösta Mittag-Leffler gave a toast in honor of Niels Henrik Abel, and he proposed starting a collection with the goal that in 16 years—in 1902, on the centennial of Abel's birth—a statue of the young genius could be erected. Money was collected during the meeting and national committees were appointed, but eventually the whole effort ran out of steam.

Mittag-Leffler, who had been publishing the Swedish mathematics journal, *Acta Mathematica*, since 1882, worked during these years to arrange and gather support for an international mathematics prize, namely King Oscar II's Mathematics Prize,

A. Stubhaug (✉)
Matematisk institutt, Universitetet i Oslo, 0316 Oslo, Norway
e-mail: arilds@math.uio.no

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_1, © Springer-Verlag Berlin Heidelberg 2010

a competition in which an answer was sought to one of four given questions. The prize was awarded on the King's 60th birthday in January 1889, and it was a tremendous success in every way. The prize winner was Henri Poincaré, who submitted a work that described chaos in space: a discovery that was only understood in its full breadth much later and that gradually developed into a major interdisciplinary research area. On the jury for the prize sat Charles Hermite and Karl Weierstrass together with Mittag-Leffler, and the latter discussed the possibility of establishing a permanent mathematics prize with King Oscar and various patrons and donors. Due to insufficient support, however, Mittag-Leffler initially tried to establish a smaller fund, and he proposed that money from this fund should be used for gold medals, which should be awarded to mathematicians who had published an exceptionally important work in *Acta Mathematica*. The gold medals were to be stamped with portraits of the greatest mathematicians, and he was of the opinion that it was suitable to begin with the greatest mathematician in the Nordic countries: Niels Henrik Abel.

These plans also came to naught. Instead, Mittag-Leffler managed to set up a fund that supported the editing of articles submitted to *Acta Mathematica* and that made it possible to invite great foreign mathematicians to Stockholm. When the content of Alfred Nobel's last will and testament became known in 1897, rumors abounded that Mittag-Leffler's financial antics and scientific plans and ideas might have dissuaded Nobel from providing funds for a prize in mathematics, in addition to those in physics, chemistry and medicine as well as literature and efforts to promote peace. It is true that Mittag-Leffler and Nobel discussed financial support for both Stockholm University College (now Stockholm University) and an extraordinary professorship for Sonja Kovalevsky and that they were in strong disagreement, but the reason why there was not any Nobel Prize in mathematics seems to clearly lie in Nobel's attitude to science and technology. He was a practical man and regarded mathematics in general as much too theoretical and having no practical applications.

The annual Nobel Prizes, awarded for the first time in 1901, quickly overshadowed other scientific prizes. At the academies of science in Paris and Berlin, mathematics prizes based on various problems, often in astronomy and navigation, had been awarded ever since the middle of the eighteenth century, and new prizes came into being in the nineteenth century [1]. Prizes were also announced in Leipzig, Göttingen and at other centers of learning. In 1897, the international Lobachevsky Prize was established at the University of Kazan. This prize was supposed to go to outstanding works in geometry, especially non-Euclidean geometry, and the first winner was Sophus Lie.

Sophus Lie, Norway's second world-class mathematician, died in 1899. One of the last things he used his international contact network for was to gather support for establishing a fund that would award an Abel Prize every fifth year for outstanding work in pure mathematics. Apparently, an inspiration in Lie's work was precisely the fact that Nobel's plans included no prize in mathematics. From leading centers of mathematical learning, Sophus Lie received overwhelming support for such an Abel Prize in the spring of 1898. From Rome and Pisa came assurances of support from Luigi Cremona and Luigi Bianchi; from Paris Émile Picard wrote that both he and Hermite would donate money to the fund, and Picard, who otherwise would like to see a more frequent awarding of the prize than once every fifth year, reported that through its universities and lyceums France would also probably be able to contribute large sums; Gaston Darboux followed up with similar positive reactions and thought that all mathematicians in the Academy of Science in Paris would support an Abel Fund; Sophus Lie also received a warm declaration of support from A.R. Forsyth at Cambridge, who thought that Lord Kelvin would certainly lend his support to the fund; Felix Klein at Göttingen reported that he would obviously support the work, and he believed that David Hilbert would do so as well; Lazarus Fuchs was also supportive. The only mathematicians who expressed skepticism were Georg Frobenius and H.A. Schwarz in Berlin; they thought prizes in general often diverted younger talents away from the true scientific path.

Sophus Lie's contacts and promises of support, however, were related to him personally. When Sophus Lie died, there was no one else who could carry on the work.

At the celebration of the centennial of Abel's birth in 1902, three main tasks were formulated in Norwegian political and scientific circles: first, to arrange a broad cultural commemoration, second, to erect a worthy monument to the genius, and third, to establish an international Abel Prize. The first two tasks were achieved. The Abel commemoration in September 1902 was held with pomp and circumstance, and students, citizens, scientists, artists, the national assembly, the government and the Royal House all took part. A number of foreign mathematicians were present and were awarded honorary doctorates. Gustav Vigeland's great Abel Monument on the Royal Palace grounds (in Oslo) was unveiled six years later, but the plans for an Abel Prize were put on ice for reasons of national politics.

Gustav Vigeland's Abel Monument, Oslo

In the Norwegian capital it was regarded as important that the commemoration of Abel should put Norway on the map as a cultural nation, not least with a view to the conflict over the union (with Sweden), which many realized was imminent. However, King Oscar still sat on the Swedish–Norwegian throne, and after his mathematics prize (in 1889) and his support for *Acta Mathematica* (in 1882), he was regarded as having a special fondness for mathematics. The King himself also took active part in the Abel celebrations and arranged a big festivity at the Palace. Just after the conclusion of the official celebrations, Norwegian politicians and scientists were informed that King Oscar was considering having a gold medal created in memory of Niels Henrik Abel. The idea was that the medal should be awarded once every three years by the University of Oslo for top-flight mathematical work.

Two Norwegian scientists, Waldemar C. Brøgger and Fridtjof Nansen, and a representative from the Royal Court were delegated to draw up statutes, and Gustav Vigeland drew sketches for an Abel medal. When the proposal was presented on the King's birthday in January 1903, it was recommended that the prize be awarded every fifth year by the Scientific Society of Christiania (now the Norwegian Academy of Science and Letters in Oslo), and that the prize should go to the best mathematical work published during the last five years. However, decisions about the procedure, the prize committee, etc. were to be announced later.

In the ongoing work, many people were consulted for advice. Mittag-Leffler, who was well-informed about the establishment of the Bolyai Prize in Budapest, sent a copy of the statutes for that prize to Brøgger [5]. (The Bolyai Prize was awarded for the first time in 1905 to Henri Poincaré and the next time five years later to David Hilbert.) At that time, Mittag-Leffler was afraid that an Abel Prize, if there were to be one, would be overshadowed by the Nobel Prize. He did not think it was possible to find a new patron who could elevate an Abel Prize to the Nobel Prize level, and he was also of the opinion that it would be easier for a jury to make an irreproachable selection if there was a prize competition focused on a given problem or question, preferably related to Abel's work.

The mathematicians Ludvig Sylow and Carl Størmer were the key members of a committee that was supposed to draw up a set of rules for an Abel Prize. In the autumn of 1904, they submitted a memo, but the work had not been completed when the dramatic political events of June 1905 resulted in the dissolution of the union between Sweden and Norway. All further plans for an Abel Prize were set aside. The realities of the matter were expressed by Nansen in a letter to the mathematician Elling Holst: "The Abel Prize that we had been promised by good King Oscar went to heaven with the union."

In international circles of mathematicians, however, the lack of a prize in mathematics on the same level as the Nobel Prize was a frequent topic of discussion. This lack was a prime motivation for John Charles Fields in his efforts to establish the prize medal that would come to bear his name. The Fields Medal was awarded for the first time at the International Congress of Mathematicians in Oslo in 1936. Even though no money is awarded with the Fields Medal, and it is only awarded every fourth year at the International Congress of Mathematicians to two to four mathematicians under age 40, the Fields Medal rapidly gained the status of the most eminent prize in mathematics, a kind of "Nobel Prize" in mathematics; a position it has held until the Abel Prize finally became a reality.

In Norway, Abel's name and memory were kept alive in various ways on into the twentieth century. On the occasion of the centennial of his death in 1929, Abel was commemorated on Norwegian stamps; aside from the royal family, only the playwright Henrik Ibsen had previously been so honored. In 1948, Norges Bank printed Abel's portrait on the obverse of the 500-kroner banknote. Abel has also been used in later banknote and stamp issues, and books have been written about his life and scientific efforts. When the International Mathematical Union, with UNESCO support, designated the year 2000 as the "World Mathematical Year", Abel was Norway's leading logo. Abel's international position and his life and work were also at the heart of the efforts leading up to the bicentennial of Abel's birth. The objective of a number of national and international efforts aimed at the profession, schools and society at large was to create a broader appreciation of the importance of mathematics and science for today's society.

In 1996, I published a biography of Niels Henrik Abel (an English edition was published in 2000 [3]), and in response to an initiative from the Department of Mathematics at the University of Oslo, I subsequently worked on a biography of Sophus Lie [4]. I was very familiar with Lie's contact network and efforts on behalf of an Abel Prize, and in lectures and conversations in academic circles of mathematicians, I brought up the old idea of such a prize. Most of the people I talked to thought the idea was fascinating, but extremely unrealistic. At a book signing in August of 2000, I met Tormod Hermansen, the President and CEO of Telenor at the time and a prominent Labor Party supporter. Hermansen showed immediate interest in an Abel Prize and argued in his political circles that funds should be allocated for such a prize. The reactions were positive, and at the Department of Mathematics at the University of Oslo, a working group was formed: the Working Group for the Abel Prize, consisting of Professors Jens Erik Fenstad, Arnfinn Laudal and Ragni Piene together with Administrative Head of Department Yngvar Reichelt, Assistant Professor Nils Voje Johansen and myself. With support from key figures in university, business and cultural circles, this working group had talks with the relevant Ministries and members of the Storting [the Norwegian Parliament]. Declarations of support were also received from the major international mathematics organizations—the *International Mathematical Union* and the *European Mathematical Society*. In May 2001, the working group submitted a proposal to the Prime Minister to establish an Abel Prize, and in August 2001, Prime Minister Jens Stoltenberg announced that the Norwegian Government would establish an Abel Fund worth NOK 200 million: a greater amount than the working group had proposed [2]. The Prime Minister emphasized the broad political consensus that the proposal had aroused and the hope that an annual Abel Prize would strengthen the research in and recruitment to mathematics and the natural sciences and raise international awareness of Norway as a knowledge-based nation.

The *Niels Henrik Abel Memorial Fund* is administered by the Norwegian Ministry of Education and Research, and the annual return on the fund is allocated to the Norwegian Academy of Science and Letters, which is entrusted with awarding the prize and the management of other matters related to the funds. The Norwegian Academy of Science and Letters has established a board and a committee of mathematicians for the Abel Prize. The Abel Board shall be responsible for distributing the return on the fund and for events associated with the award ceremony, whereas the Abel Committee is responsible for reviewing candidates for the prize and make a recommendation to the Academy. This international committee consists of five persons who are outstanding researchers in the field of mathematics; both the International Mathematical Union and the European Mathematical Society nominate committee members.

As it is laid down in the statutes, the annual Abel Prize is a recognition of a scientific contribution of exceptional depth in and significance for the field of mathematics, including mathematical aspects of information technology, mathematical physics, probability theory, numerical analysis and computational science, statistics, and applications of mathematics in other sciences. One of the objectives for the prize is that it shall be awarded over the years in a broad range of areas in the field of mathematics.

As is also laid down in the statutes, the prize should contribute towards raising the status of mathematics in society and stimulate the interest of young people and children in mathematics. This objective was a very important argument for the creation of the prize, it was explicitly mentioned by the Prime Minister when he announced the establishment of the Fund, and it was most likely decisive for the Government's and the Parliament's acceptance.

## References

1. Gray, J.: A history of prizes in mathematics. In: Carlson, J., Jaffe, A., Wiles, A. (eds.) The Millennium Prize Problems, pp. 3–27. Am. Math. Soc., Providence (2006)
2. Helsvig, K.G.: Elitisme på norsk. Det Norske Videnskaps-Akademi 1945–2007. Novus forlag, Oslo (2007), pp. 194–197
3. Stubhaug, A.: Niels Henrik Abel and His Times. Called Too Soon by Flames Afar. Springer, Berlin (2000)
4. Stubhaug, A.: The Mathematician Sophus Lie. It Was the Audacity of My Thinking. Springer, Berlin (2002)
5. Stubhaug, A.: Gösta Mittag-Leffler. Springer, Berlin (2010, to appear)

# 2003

# Jean-Pierre Serre

ABEL
PRISEN

# 2003

# Jean-Pierre Serre

ABEL
PRISEN

# Jean-Pierre Serre: Mon premier demi-siècle au Collège de France

# Jean-Pierre Serre: My First Fifty Years at the Collège de France

**Marc Kirsch**

Ce chapitre est une interview par Marc Kirsch. Publié précédemment dans *Lettre du Collège de France*, n$^{o}$ 18 (déc. 2006). Reproduit avec autorisation.

This chapter is an interview by Marc Kirsch. Previously published in *Lettre du Collège de France*, no. 18 (déc. 2006). Reprinted with permission.

M. Kirsch (✉)
Collège de France, 11, place Marcelin Berthelot, 75231 Paris Cedex 05, France
e-mail: marc.kirsch@college-de-france.fr

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_3, © Springer-Verlag Berlin Heidelberg 2010

Jean-Pierre Serre, Professeur au Collège de France, titulaire de la chaire d'*Algèbre et Géométrie* de 1956 à 1994.

*Vous avez enseigné au Collège de France de 1956 à 1994, dans la chaire d'Algèbre et Géométrie. Quel souvenir en gardez-vous?*

J'ai occupé cette chaire pendant 38 ans. C'est une longue période, mais il y a des précédents: si l'on en croit l'Annuaire du Collège de France, au XIX$^{e}$ siècle, la chaire de physique n'a été occupée que par deux professeurs: l'un est resté 60 ans, l'autre 40. Il est vrai qu'il n'y avait pas de retraite à cette époque et que les professeurs avaient des suppléants (auxquels ils versaient une partie de leur salaire).

Quant à mon enseignement, voici ce que j'en disais dans une interview de 1986[1]: "Enseigner au Collège est un privilège merveilleux et redoutable. Merveilleux à cause de la liberté dans le choix des sujets et du haut niveau de l'auditoire: chercheurs au CNRS, visiteurs étrangers, collègues de Paris et d'Orsay — beaucoup sont des habitués qui viennent régulièrement depuis cinq, dix ou même vingt ans. Redoutable aussi: il faut chaque année un sujet de cours nouveau, soit sur ses propres recherches (ce que je préfère), soit sur celles des autres; comme un cours annuel dure environ vingt heures, cela fait beaucoup!"

*Comment s'est passée votre leçon inaugurale?*

À mon arrivée au Collège, j'étais un jeune homme de trente ans. La leçon inaugurale m'apparaissait presque comme un oral d'examen, devant professeurs, famille, collègues mathématiciens, journalistes, etc. J'ai essayé de la préparer. Au bout d'un mois, j'avais réussi à en écrire une demi-page.

Arrive le jour de la leçon, un moment assez solennel. J'ai commencé par lire la demi-page en question, puis j'ai improvisé. Je ne sais plus très bien ce que j'ai dit (je me souviens seulement avoir parlé de l'Algèbre, et du rôle ancillaire qu'elle joue en Géométrie et en Théorie des Nombres). D'après le compte-rendu paru dans le journal *Combat*, j'ai passé mon temps à essuyer machinalement la table qui me séparait du public; je ne me suis senti à l'aise que lorsque j'ai pris en main un bâton de craie et que j'ai commencé à écrire sur le tableau noir, ce vieil ami des mathématiciens.

Quelques mois plus tard, le secrétariat m'a fait remarquer que toutes les leçons inaugurales étaient rédigées et que la mienne ne l'était pas. Comme elle avait été improvisée, j'ai proposé de la recommencer dans le même style, en me remettant mentalement dans la même situation. Un beau soir, on m'a ouvert un bureau du Collège et l'on m'a prêté un magnétophone. Je me suis efforcé de recréer l'atmosphère initiale, et j'ai refait une leçon sans doute à peu près semblable à l'originale. Le lendemain, j'ai apporté le magnétophone au secrétariat; on m'a dit que l'enregistrement était inaudible. J'ai estimé que j'avais fait tout mon possible et je m'en suis tenu là. Ma leçon inaugurale est restée la seule qui n'ait jamais été rédigée.

En règle générale, je n'écris pas mes exposés; je ne consulte pas mes notes (et, souvent, je n'en ai pas). J'aime réfléchir devant mes auditeurs. J'ai le sentiment, Jean-Pierre Serre, Professor at the Collège de France, held the Chair in *Algebra and Geometry* from 1956 to 1994.

*You taught at the Collège de France from 1956 to 1994, holding the Chair in Algebra and Geometry. What are you memories of your time there?*

I held the Chair for 38 years. That is a long time, but there were precedents. According to the Yearbook of the Collège de France, the Chair in Physics was held by just two professors in the 19th century: one remained in his post for 60 years, and the other for 40. It is true that there was no retirement in that era and that professors had deputies (to whom they paid part of their salaries).

As for my teaching career, this is what I said in an interview in 1986[1]: "Teaching at the Collège is both a marvelous and a challenging privilege. Marvelous because of the freedom of choice of subjects and the high level of the audience: CNRS [Centre national de la recherche scientifique] researchers, visiting foreign academics, colleagues from Paris and Orsay—many regulars who have been coming for 5, 10 or even 20 years. It is challenging too: new lectures have to be given each year, either on one's own research (which I prefer), or on the research of others. Since a series of lectures for a year's course is about 20 hours, that's quite a lot!"

*Can you tell us about your inaugural lecture?*

I was a young man, about 30, when I arrived at the Collège. The inaugural lecture was almost like an oral examination in front of professors, family, mathematician colleagues, journalists, etc. I tried to prepare it, but after a month I had only managed to write half a page.

When the day of the lecture came, it was quite a tense moment. I started by reading the half page I had prepared and then I improvised. I can no longer remember what I said (I only recall that I spoke about algebra and the ancillary role it plays in geometry and number theory). According to the report that appeared in the newspaper *Combat*, I spent most of the time mechanically wiping the table that separated me from my audience. I did not feel at ease until I had a piece of chalk in my hand and I started to write on the blackboard, the mathematician's old friend.

A few months later, the Secretary's Office told me that all inaugural lectures were written up, but they had not received the transcript of mine. As it had been improvised, I offered to repeat it in the same style, mentally putting myself back in the same situation. One evening, I was given a tape recorder and I went into an office at the Collège. I tried to recall the initial atmosphere, and to make up a lecture as close as possible to the original one. The next day I returned the tape recorder to the Secretary's Office. They told me that the recording was inaudible. I decided that I had done all I could and left it there. My inaugural lecture is still the only one that has not been written up.

As a rule, I don't write my lectures. I don't consult notes (and often I don't have any). I like to do my thinking in front of the audience. When I am explaining lorsque j'explique des mathématiques, de parler à un ami. Devant un ami, on n'a pas envie de lire un texte. Si l'on a oublié une formule, on en donne la structure; cela suffit. Pendant l'exposé j'ai en tête une quantité de choses qui me permettraient de parler bien plus longtemps que prévu. Je choisis suivant l'auditoire, et l'inspiration du moment.

Seule exception: le séminaire Bourbaki, où l'on doit fournir un texte suffisamment à l'avance pour qu'il puisse être distribué en séance. C'est d'ailleurs le seul séminaire qui applique une telle règle, très contraignante pour les conférenciers.

*Quel est la place de Bourbaki dans les mathématiques françaises d'aujourd'hui?*

C'est le séminaire qui est le plus intéressant. Il se réunit trois fois par an, en mars, mai et novembre. Il joue un rôle à la fois social (occasion de rencontres) et mathématique (exposé de résultats récents — souvent sous une forme plus claire que celle des auteurs); il couvre toutes les branches des mathématiques.

Les livres (*Topologie, Algèbre, Groupes de Lie,...*) sont encore lus, non seulement en France, mais aussi à l'étranger. Certains de ces livres sont devenus des classiques: je pense en particulier à celui sur les systèmes de racines. J'ai vu récemment (dans le *Citations Index* de l'AMS[2]) que Bourbaki venait au 6$^{e}$ rang (par nombre de citations) parmi les mathématiciens français (de plus, au niveau mondial, les n$^{os}$ 1 et 3 sont des Français, et s'appellent tous deux Lions: un bon point pour le Collège). J'ai gardé un très bon souvenir de ma collaboration à Bourbaki, entre 1949 et 1973. Elle m'a appris beaucoup de choses, à la fois sur le fond (en me forçant à rédiger des choses que je ne connaissais pas) et sur la forme (comment écrire de façon à être compris). Elle m'a appris aussi à ne pas trop me fier aux "spécialistes."

La méthode de travail de Bourbaki est bien connue: distribution des rédactions aux différents membres et critique des textes par lecture à haute voix (ligne à ligne: c'est lent mais efficace). Les réunions (les "congrès") avaient lieu 3 fois par an. Les discussions étaient très vives, parfois même passionnées. En fin de congrès, on distribuait les rédactions à de nouveaux rédacteurs. Et l'on recommençait. Le même chapitre était souvent rédigé quatre ou cinq fois. La lenteur du processus explique que Bourbaki n'ait publié finalement qu'assez peu d'ouvrages en quarante années d'existence, depuis les années 1930–1935 jusqu'à la fin des années 1970, où la production a décliné.

En ce qui concerne les livres eux-mêmes, on peut dire qu'ils ont rempli leur mission. Les gens ont souvent cru que ces livres traitaient des sujets que Bourbaki trouvait intéressants. La réalité est différente: ses livres traitent de ce qui est utile pour faire des choses intéressantes. Prenez l'exemple de la théorie des nombres. Les publications de Bourbaki en parlent très peu. Pourtant, ses membres l'appréciaient beaucoup, mais ils jugeaient que cela ne faisait pas partie des *Éléments*: il fallait d'abord avoir compris beaucoup d'algèbre, de géométrie et d'analyse.

Par ailleurs, on a souvent imputé à Bourbaki tout ce que l'on n'aimait pas en mathématiques. On lui a reproché notamment les excès des "maths modernes" dans les programmes scolaires. Il est vrai que certains responsables de ces programmes se mathematics, I feel I am speaking to a friend. You don't want to read a text out to a friend; if you have forgotten a formula, you give its structure; that's enough. During the lecture I have a lot of possible material in my mind—much more than possible in the allotted time. What I actually say depends on the audience and my inspiration.

Only exception: the Bourbaki seminar for which one has to provide a text sufficiently in advance so that it can be distributed during the meeting. This is the only seminar that applies this rule; it is very restrictive for lecturers.

*What is Bourbaki's place in French mathematics now?*

Its most interesting feature is the Bourbaki seminar. It is held three times a year, in March, May and November. It plays both a social role (an occasion for meeting other people) and a mathematical one (the presentation of recent results—often in a form that is clearer than that given by the authors). It covers all branches of mathematics.

Bourbaki's books (*Topology, Algebra, Lie Groups*, etc.) are still widely read, not just in France but also abroad. Some have become classics: I'm thinking in particular about the book on root systems. I recently saw (in the AMS *Citations Index*[2]) that Bourbaki ranked sixth (by number of citations) among French mathematicians. (What's more, at the world level, numbers 1 and 3 are French and both are called Lions: a good point for the Collège.) I have very good memories of my collaboration with Bourbaki from 1949 to 1973. Bourbaki taught me many things, both on background (making me write about things which I did not know very well) and on style (how to write in order to be understood). Bourbaki also taught me not to rely on "specialists".

Bourbaki's working method is well-known: the distribution of drafts to the various members and their criticism by reading them aloud (line by line: slow but effective). The meetings ("congrès") were held three times a year. The discussions were very lively, sometimes passionate. At the end of each congrès, the drafts were distributed to new writers. And so on. A chapter could often be written four or five times. The slow pace of the process explains why Bourbaki ended up publishing with relatively few books over the 40 years from 1930–1935 till the end of the 1970s when production faded away.

As for the books themselves, one may say that they have fulfilled their mission. People often believe that these books deal with subjects that Bourbaki found interesting. The reality is different: the books deal with what is useful in order to do interesting things. Take number theory for example. Bourbaki's publications hardly mention it. However, the Bourbaki members liked it very much it, but they considered that it was not part of the *Elements*: it needed too much algebra, geometry and analysis.

Besides, Bourbaki is often blamed for everything that people do not like about mathematics, especially the excesses of "modern math" in school curricula. It is true that some of those responsible for these curricula claimed to follow Bourbaki. But sont réclamés de Bourbaki. Mais Bourbaki n'y était pour rien: ses écrits étaient destinés aux mathématiciens, pas aux étudiants, encore moins aux adolescents. Notez que Bourbaki a évité de se prononcer sur ce sujet. Sa doctrine était simple: on fait ce que l'on choisit de faire, on le fait du mieux que l'on peut, mais on n'explique pas pourquoi on le fait. J'aime beaucoup ce point de vue qui privilégie le travail par rapport au discours — tant pis s'il prête parfois à des malentendus.

*Comment analysez-vous l'évolution de votre discipline depuis l'époque de vos débuts? Est-ce que l'on fait des mathématiques aujourd'hui comme on les faisait il y a cinquante ans?*

Bien sûr, on fait des mathématiques aujourd'hui comme il y a cinquante ans! Évidemment, on comprend davantage de choses; l'arsenal de nos méthodes a augmenté. Il y a un progrès continu. (Ou parfois un progrès par à-coups: certaines branches restent stagnantes pendant une décade ou deux, puis brusquement se réveillent quand quelqu'un introduit une idée nouvelle.)

Si l'on voulait dater les mathématiques "modernes" (un terme bien dangereux), il faudrait sans doute remonter aux environs de 1800 avec Gauss.

*Et en remontant plus loin, si vous rencontriez Euclide, qu'auriez-vous à vous dire?*

Euclide me semble être plutôt quelqu'un qui a mis en ordre les mathématiques de son époque. Il a joué un rôle analogue à celui de Bourbaki il y a cinquante ans. Ce n'est pas par hasard que Bourbaki a choisi d'intituler ses ouvrages des *Éléments de Mathématique*: c'est par référence aux *Éléments* d'Euclide. (Notez aussi que "Mathématique" est écrit au singulier. Bourbaki nous enseigne qu'il n'y a pas plusieurs mathématiques distinctes, mais une seule mathématique. Et il nous l'enseigne à sa façon habituelle: pas par de grands discours, mais par l'omission d'une lettre à la fin d'un mot.)

Pour en revenir à Euclide, je ne pense pas qu'il ait produit des contributions réellement originales. Archimède serait un interlocuteur plus indiqué. C'est lui le grand mathématicien de l'Antiquité. Il a fait des choses extraordinaires, aussi bien en mathématique qu'en physique.

*En philosophie des sciences, il y a un courant très fort en faveur d'une pensée de la rupture. N'y a-t-il pas de ruptures en mathématiques? On a décrit par exemple l'émergence de la probabilité comme une manière nouvelle de se représenter le monde. Quelle est sa signification en mathématiques?*

Les philosophes aiment bien parler de "rupture." Je suppose que cela ajoute un peu de piment à leurs discours. Je ne vois rien de tel en mathématique: ni catastrophe, ni révolution. Des progrès, oui, je l'ai déjà dit; ce n'est pas la même chose. Nous travaillons tantôt à de vieilles questions, tantôt à des questions nouvelles. Il n'y a pas de frontière entre les deux. Il y a une grande continuité entre les mathématiques

Bourbaki had nothing to do with it: its books are meant for mathematicians, not for students, and even less for teen-agers. Note that Bourbaki was careful not to write anything on this topic. Its doctrine was simple: one does what one chooses to do, one does it the best one can, but one does not explain why. I very much like this attitude which favors work over discourse—too bad if it sometimes lead to misunderstandings.

*How would you describe the development of your discipline since the time when you were starting out? Is mathematics conducted nowadays as it was 50 years ago?*

Of course you do mathematics today like 50 years ago! Clearly more things are understood; the range of our methods has increased. There is continuous progress. (Or sometimes leaps forward: some branches remain stagnant for a decade or two and then suddenly there's a reawakening as someone introduces a new idea.)

If you want to put a date on "modern" mathematics (a very dangerous term), you would have to go back to about 1800 and Gauss.

*Going back further, if you were to meet Euclid, what would you say to him?*

Euclid seems to me like someone who just put the mathematics of his era into order. He played a role similar to Bourbaki's 50 years ago. It is no coincidence that Bourbaki decided to give its treatise the title *Éléments de Mathématique*. This is a reference to Euclid's *Éléments*. (Note that "Mathématique" is written in the singular. Bourbaki tells us that rather than several different mathematics there is one single mathematics. And he tells us in his usual way: not by a long discourse, but by the omission of one letter from the end of one word.)

Coming back to Euclid, I don't think that he came up with genuinely original contributions. Archimedes would be much more interesting to talk to. He was the great mathematician of antiquity. He did extraordinary things, both in mathematics and physics.

*In the philosophy of science there is a very strong current in favor of the concept of rupture. Are there ruptures in mathematics? For example the emergence of probability as a new way in which to represent the world. What is its significance in mathematics?*

Philosophers like to talk of "rupture". I suppose it adds a bit of spice to what they say. I do not see anything like that in mathematics: no catastrophe and no revolution. Progress, yes, as I've already said; but that is not the same. We work sometimes on old questions and sometimes on new ones. There is no boundary between the two. There is a deep continuity between the mathematics of two centuries ago and that d'il y a deux siècles et celles de maintenant. Le temps des mathématiciens est la “longue durée” de feu mon collègue Braudel.

Quant aux probabilités, elles sont utiles pour leurs applications à la fois mathématiques et pratiques; d'un point de vue purement mathématique, elles constituent une branche de la théorie de la mesure. Peut-on vraiment parler à leur sujet de “manière nouvelle de se représenter le monde”? Sûrement pas en mathématique.

*Est-ce que les ordinateurs changent quelque chose à la façon de faire des mathématiques?*

On avait coutume de dire que les recherches en mathématiques étaient peu coûteuses: des crayons et du papier, et voilà nos besoins satisfaits. Aujourd'hui, il faut ajouter les ordinateurs. Cela reste peu onéreux, dans la mesure où les mathématiciens ont rarement besoin de ressources de calcul très importantes. À la différence, par exemple, de la physique des particules, dont les besoins en calcul sont à la mesure des très grands équipements nécessaires au recueil des données, les mathématiciens ne mobilisent pas de grands centres de calcul.

En pratique, l'informatique change les conditions matérielles du travail des mathématiciens: on passe beaucoup de temps devant son ordinateur. Il a différents usages. Tout d'abord, le nombre des mathématiciens a considérablement augmenté. À mes débuts, il y a 55 ou 60 ans, le nombre des mathématiciens productifs était de quelques milliers (dans le monde entier), l'équivalent de la population d'un village. À l'heure actuelle, ce nombre est d'au moins 100 000: une ville. Cet accroissement a des conséquences pour la manière de se contacter et de s'informer. L'ordinateur et Internet accélèrent les échanges. C'est d'autant plus précieux que les mathématiciens ne sont pas ralentis, comme d'autres, par le travail expérimental: nous pouvons communiquer et travailler très rapidement. Je prends un exemple. Un mathématicien a trouvé une démonstration mais il lui manque un lemme de nature technique. Au moyen d'un moteur de recherche — comme Google — il repère des collègues qui ont travaillé sur la question et leur envoie un e-mail. De cette manière, il a toutes les chances de trouver en quelques jours ou même en quelques heures la personne qui a effectivement démontré le lemme dont il a besoin. (Bien entendu, ceci ne concerne que des problèmes auxiliaires: des points de détail pour lesquels on désire renvoyer à des références existantes plutôt que de refaire soi-même les démonstrations. Sur des questions vraiment difficiles, mon mathématicien aurait peu de chances de trouver quelqu'un qui puisse lui venir en aide.)

L'ordinateur et Internet sont donc des outils d'accélération de notre travail. Ils permettent aussi de rendre les manuscrits accessibles dans le monde entier, sans attendre leur parution dans un journal. C'est très pratique. Notez que cette accélération a aussi des inconvénients. Le courrier électronique produit des correspondances informelles que l'on conserve moins volontiers que le papier. On jette rarement des lettres alors que l'on efface ou l'on perd facilement les emails (quand on change d'ordinateur, par exemple). On a publié récemment (en version bilingue: français sur une page, et anglais sur la page d'en face) ma correspondance avec A. Grothendieck entre 1955 et 1987; cela n'aurait pas été possible si elle avait été électronique.

of today. The time of mathematicians is the "longue durée" of my late colleague the historian Fernand Braudel.

As for probability theory, it is useful for its applications both to mathematics and to practical questions From a purely mathematical point of view, it is a branch of measure theory. Can one really describe it as "a new way in which to represent the world"? Surely not in mathematics.

*Have computers changed the manner in which mathematics is conducted?*

It used to be said that mathematical research was cheap: paper and pencils, that was all we needed. Nowadays, you have to add computers. It is not very expensive, since mathematicians rarely need a lot of processing power. This is different from, say, particle physics, where a lot of equipment is required.

In practice, computers have changed the material conditions of mathematicians' work: we spend a lot of time in front of our computer. It has several different uses. First of all, there are now considerably more mathematicians. When I started out, some 55 or 60 years ago, there were only a few thousand productive mathematicians (in the whole world), the equivalent of a village. Now, this number has grown to at least 100 000: a city. This growth has consequences for the way mathematicians contact each other and gain information. The computer and Internet have accelerated exchanges. This is especially important for us, since we are not slowed down, as others, by experimental work: we can communicate and work very rapidly. Let me give you an example. If a mathematician is working on a proof but needs a technical lemma, then through a search engine—such as Google—he will track down colleagues who have worked on the question and send them an e-mail. In this way, in just a few days or even hours, he may be able to find somebody who has proved the required lemma. (Of course, this only applies to easy problems: those for which you want to use a reference rather than to reconstruct a proof. For really difficult questions, a mathematician would have little chance of finding someone to help him.)

Computer and Internet are thus the tools which speed up our work. They allow us to make our manuscripts accessible to everybody without waiting for publication in a journal. That is very convenient. But this acceleration also has its disadvantages. E-mail produces informal correspondence which is less likely to be kept than the paper one. It is unusual to throw letters away but one can easily delete or lose e-mails (when one changes computers for example). Recently a bilingual version (French on one page and English on the other) of my correspondence with A. Grothendieck between 1955 and 1987 has been published. That would not have been possible if the correspondence had been by e-mail.

Par ailleurs, certaines démonstrations font appel à l'ordinateur pour vérifier une série de cas qu'il serait impraticable de traiter à la main. Deux cas classiques: le problème des 4 couleurs (coloriage des cartes avec seulement quatre couleurs) et le problème de Képler (empilement des sphères dans l'espace à 3 dimensions). Cela conduit à des démonstrations qui ne sont pas réellement vérifiables; autrement dit, ce ne sont pas de vraies "démonstrations" mais seulement des faits expérimentaux, très vraisemblables, mais que personne ne peut garantir.

*Vous avez évoqué l'augmentation du nombre des mathématiciens. Quelle est aujourd'hui la situation. Où vont les mathématiques?*

L'augmentation du nombre des mathématiciens est un fait important. On pouvait craindre que cela se fasse au détriment de la qualité. En fait, il n'y a rien eu de tel. Il y a beaucoup de très bons mathématiciens (en particulier parmi les jeunes français — un très bon augure).

Ce que je peux dire, concernant l'avenir, c'est qu'en dépit de ce grand nombre de mathématiciens, nous ne sommes pas à court de matière. Nous ne manquons pas de problèmes, alors qu'il y a un peu plus de deux siècles, à la fin du XVIII$^{\text{e}}$, Lagrange était pessimiste: il pensait que "la mine était tarie," qu'il n'y avait plus grand-chose à trouver. Lagrange a écrit cela juste avant que Gauss ne relance les mathématiques de manière extraordinaire, à lui tout seul. Aujourd'hui, il y a beaucoup de terrains à prospecter pour les jeunes mathématiciens (et aussi pour les moins jeunes, j'espère).

*Selon un lieu commun de la philosophie des sciences, les grandes découvertes mathématiques sont le fait de mathématiciens jeunes. Est-ce votre cas?*

Je ne crois pas que le terme de "grande découverte" s'applique à moi. J'ai surtout fait des choses "utiles" (pour les autres mathématiciens). En tout cas, lorsque j'ai eu le prix Abel en 2003, la plupart des travaux qui ont été cités par le jury avaient été faits avant que je n'aie 30 ans. Mais si je m'étais arrêté à ce moment-là, on ne m'aurait sans doute pas donné ce prix: j'ai fait aussi d'autres choses par la suite (ne serait-ce que des "conjectures" sur lesquelles beaucoup de gens ont travaillé et travaillent encore).

Dans ma génération, plusieurs de mes collègues ont continué au-delà de 80 ans, par exemple mes vieux amis Armand Borel et Raoul Bott, morts tous deux récemment à 82 ans. Il n'y a pas de raison de s'arrêter, tant que la santé le permet. Encore faut-il que le sujet s'y prête. Quand on a des sujets très larges, il y a toujours quelque chose à faire, mais si l'on est trop spécialisé on peut se retrouver bloqué pendant de longues périodes, soit parce que l'on a démontré tout ce qu'il y avait à démontrer, soit au contraire parce que les problèmes sont trop difficiles. C'est très frustrant.

Les découvertes mathématiques donnent de grandes joies. Poincaré, Hadamard, Littlewood[3] l'ont très bien expliqué. En ce qui me concerne, je garde surtout le souvenir d'une idée qui a contribué à débloquer la théorie de l'homotopie. Cela s'est passé une nuit de retour de vacances, en 1950, dans une couchette de train. Je cherchais un espace fibré ayant telles et telles propriétés. La réponse est venue:

On the other hand, some proofs do need a computer in order to check a series of cases that would be impossible to do by hand. Two classic examples are the four-color problem (shading maps using only four colors) and the Kepler conjecture (packing spheres into three-dimensional space). This leads to proofs which are not really verifiable; in other words, they are not genuine "proofs" but just experimental facts, very plausible, but nobody can guarantee them.

*You mentioned the increasing number of mathematicians today. But where is mathematics going?*

The increase in the number of mathematicians is an important fact. One could have feared that this increase in size was to the detriment of quality. But in fact, this is not the case. There are many very good mathematicians (in particular young French mathematicians—a good omen for us).

What I can say about the future is that, despite this huge number of mathematicians, we are not short of subject matter. There is no lack of problems, even though just two centuries ago, at the end of the 18th century, Lagrange was pessimistic: he thought that "the mine was exhausted", and that there was nothing much more to discover. Lagrange wrote this just before Gauss relaunched mathematics in an extraordinary way, all by himself. Today, there are many fields to explore for young mathematicians (and even for those who are not so young, I hope).

*It is often said in the philosophy of science that major mathematical discoveries are made by young mathematicians. Was this the case for you?*

I don't believe that the term "major discovery" applies to me. I have rather done things that are "useful" (for other mathematicians). When I was awarded the Abel prize in 2003, most of the work cited by the jury had been done before I was 30. But if I had stopped then, it would probably not have awarded me the prize. I have done other things after that (if only some conjectures that have kept many people busy).

Of my generation, several of my colleagues have continued working beyond the age of 80. For example, my old friends Armand Borel and Raoul Bott, who both recently died aged 82. There is no reason to stop, as long as health allows it. But the subject matter has to be there. When you are dealing with very broad subjects, there is always something to do, but if you are too specialized you can find yourself blocked for long periods of time, either because you have proved everything that can be proved, or, to the contrary, because the problems are too difficult. It is very frustrating.

Discoveries in mathematics can bring great joy. Poincaré, Hadamard and Littlewood[3] have explained it very well. As for myself, I still have the memory of an idea that contributed to unlocking homotopy theory. It happened one night while traveling home from vacation in 1950 in the sleeping car of a train. I had been looking for a fiber space with such and such properties. Then the answer came:

l'espace des lacets! Je n'ai pas pu m'empêcher de réveiller ma femme qui dormait dans la couchette du dessous pour lui dire: ça y est! Ma thèse est sortie de là, et bien d'autres choses encore. Bien sûr, ces découvertes soudaines sont rares: cela m'est arrivé peut-être deux fois en soixante ans. Mais ce sont des moments lumineux, vraiment exceptionnels.

*Le Collège de France est-il un endroit où l'on échange avec d'autres disciplines?*

Non, pas pour moi. Même entre les mathématiciens du Collège, il n'y a pas de travail collectif. Il faut préciser que nous travaillons dans des branches souvent très séparées. Ce n'est pas un mal: le Collège n'est pas censé être un club. Un certain nombre de lieux communs modernes — comme le *travail collectif*, l'*interdisciplinarité* et le *travail en équipe* — ne s'appliquent pas à nous.

*Qu'avez-vous pensé du dialogue entre un spécialiste de neurosciences, Jean-Pierre Changeux, et le mathématicien Alain Connes, qui est restitué dans le livre* Matière à pensée*?*

Ce livre est un bel exemple de dialogue de sourds. Changeux ne comprend pas ce que dit Connes, et inversement. C'est assez étonnant. Personnellement, je suis du côté de Connes. Les vérités mathématiques sont indépendantes de nous[4]. Notre seul choix porte sur la façon de les exprimer. Si on le désire, on peut le faire sans introduire aucune terminologie. Considérons par exemple une troupe de soldats. Leur général aime les arranger de deux façons, soit en rectangle, soit en 2 carrés. C'est au sergent de les placer. Il s'aperçoit qu'il n'a qu'à les mettre en rang par 4: s'il en reste 1 qu'il n'a pas pu placer, ou bien il arrivera à les mettre tous en rectangle, ou bien il arrivera à les répartir en deux carrés.

[Traduction technique: le nombre $n$ des soldats est de la forme $4k + 1$. Si $n$ n'est pas premier, on peut arranger les soldats en rectangle. Si $n$ est premier, un théorème dû à Fermat dit que $n$ est somme de deux carrés.]

*Quelle est la place des mathématiques par rapport aux autres sciences? Y a-t-il une demande nouvelle de mathématiques, venant de ces sciences?*

Sans doute, mais il faut séparer les choses. Il y a d'une part la physique théorique, qui est tellement théorique qu'elle est à cheval entre mathématique et physique, les physiciens considérant que ce sont des mathématiques, tandis que les mathématiciens sont d'un avis contraire. Elle est symbolisée par la théorie des cordes. Son aspect le plus positif est de fournir aux mathématiciens un grand nombre d'énoncés, qu'il leur faut démontrer (ou éventuellement démolir).

Par ailleurs, notamment en biologie, il y a tout ce qui relève de systèmes comportant un grand nombre d'éléments qu'il faut traiter collectivement. Il existe des branches des mathématiques qui s'occupent de ces questions. Cela répond à une demande. Il y a aussi des demandes qui concernent la logique: c'est le cas de the loop space! I couldn't help from waking up my wife who was sleeping in the bunk below: "I've got it!" I said. My thesis, and many other things, originated from that idea. Of course, these sudden discoveries are rare: they have only happened to me twice in sixty years. But they are illuminating moments: truly exceptional.

*Are there exchanges between the disciplines at the Collège de France?*

No, not for me. There is no collective work even between the mathematicians at the Collège. We work on quite different things. This is not a bad thing. The Collège is not supposed to be a club. Many commonplace sayings, such as *collective work*, *interdisciplinarity* and *team work*, do not apply to us.

*What do you think about the dialogue between the neuroscientist Jean-Pierre Changeux and the mathematician Alain Connes, recorded in the book "Matière à pensée"?*

This book is a good example of dialogue of the deaf. Changeux does not understand what Connes says and vice versa. It is quite astonishing. Personally, I am on Connes' side. Mathematical truths are independent of us.[4] Our only choice is in the way in which we express them. If you want, you can do this without introducing any terminology. Consider, for example, a company of soldiers. The general likes to arrange them in two ways, either in a rectangle or in two squares. It is up to the sergeant to put them in the correct positions. He realizes that he only has to put them in rows of four: if there is one left over that he cannot place, either he will manage to put them all in a rectangle, or manage to arrange them in two squares.

[Technical translation: the number $n$ of soldiers is congruent to 1 (mod 4). If $n$ is not a prime, the soldiers can be arranged in a rectangle. If $n$ is a prime, a theorem of Fermat shows that $n$ is the sum of two squares.]

*What is the place of mathematics in relation to other sciences? Is there a renewed demand for mathematics from these sciences?*

Probably, but there are different cases. Some theoretical physics is so theoretical that it is half way between mathematics and physics. Physicists consider it mathematics, while mathematicians have the opposite view. String theory is a good example. The most positive aspect is to provide mathematicians with a large number of statements which they have to prove (or maybe disprove).

On the other hand, in particular in biology, there are situations involving very many elements that have to be processed collectively. There are branches of mathematics that deal with such questions. They meet a need. Another branch, logic, is l'informatique, pour la fabrication des ordinateurs. Il faut mentionner aussi la cryptographie, qui est une source de problèmes intéressants relatifs à la théorie des nombres.

En ce qui concerne la place des mathématiques par rapport aux autres sciences, on peut voir les mathématiques comme un grand entrepôt empli de rayonnages. Les mathématiciens déposent sur les rayons des choses dont ils garantissent qu'elles sont vraies; ils en donnent aussi le mode d'emploi et la manière de les reconstituer. Les autres sciences viennent se servir en fonction de leurs besoins. Le mathématicien ne s'occupe pas de ce qu'on fait de ses produits. Cette métaphore est un peu triviale, mais elle reflète assez bien la situation. (Bien entendu, on ne choisit pas de faire des mathématiques pour mettre des choses sur les rayons: on fait des mathématiques pour le plaisir d'en faire.)

Voici un exemple personnel. Ma femme, Josiane, était spécialiste de chimie quantique. Elle avait besoin d'utiliser les représentations linéaires de certains groupes de symétries. Les ouvrages disponibles n'étaient pas satisfaisants: ils étaient corrects, mais employaient des notations très lourdes. J'ai rédigé pour elle un exposé adapté à ses besoins, et je l'ai ensuite publié dans un livre intitulé *Représentations Linéaires des Groupes Finis*. J'ai fait mon travail de mathématicien (et de mari): mis des choses sur les rayons.

*Le vrai en mathématiques a-t-il le même sens qu'ailleurs?*

Non. C'est un vrai absolu. C'est sans doute ce qui fait l'impopularité des mathématiques dans le public. L'homme de la rue veut bien tolérer l'absolu quand il s'agit de religion, mais pas quand il s'agit de mathématique. Conclusion: croire est plus facile que démontrer.

useful for the building of computers. Cryptography should also be mentioned; it is a source of interesting problems in number theory.

As for the place of mathematics in relation to other sciences, mathematics can be seen as a big warehouse full of shelves. Mathematicians put things on the shelves and guarantee that they are true. They also explain how to use them and how to reconstruct them. Other sciences come and help themselves from the shelves; mathematicians are not concerned with what they do with what they have taken. This metaphor is rather coarse, but it reflects the situation well enough. (Of course one does not choose to do mathematics just for putting things on shelves; one does mathematics for the fun of it.)

Here is a personal example. My wife, Josiane, was a specialist in quantum chemistry. She needed linear representations of certain symmetry groups. The books she was working with were not satisfactory; they were correct, but they used very clumsy notation. I wrote a text that suited her needs, and then published it in book form, as *Linear Representations of Finite Groups*. I thus did my duty as a mathematician (and as a husband): putting things on the shelves.

*Does truth in mathematics have the same meaning as elsewhere?*

No. It's an absolute truth. This is probably what makes mathematics unpopular with the public. The man in the street accepts the absolute in religion, but not in mathematics. Conclusion: to believe is easier than to prove.

[1] M. Schmidt, *Hommes de Science*, 218–227, Hermann, Paris, 1990.

[2] AMS: American Mathematical Society.

[3] J. E. Littlewood, *A Mathematician's Miscellany*, Methuen and Co, 1953. Ce livre explique bien la part inconsciente du travail créatif.

[4] Il y a quelques années, mon ami R. Bott et moi-même allions recevoir un prix israélien (le prix Wolf) remis dans la Knesset, à Jerusalem. Bott devait dire quelques mots sur les mathématiques. Il m'a demandé: que dire? Je lui ai dit "C'est bien simple; tu n'as qu'à expliquer ceci: les autres sciences cherchent à trouver les lois que Dieu a choisies; les mathématiques cherchent à trouver les lois auxquelles Dieu a dû obéir." C'est ce qu'il a dit. La Knesset a apprécié.

[1·] M. Schmidt, *Hommes de Science*, 218–227, Hermann, Paris, 1990.

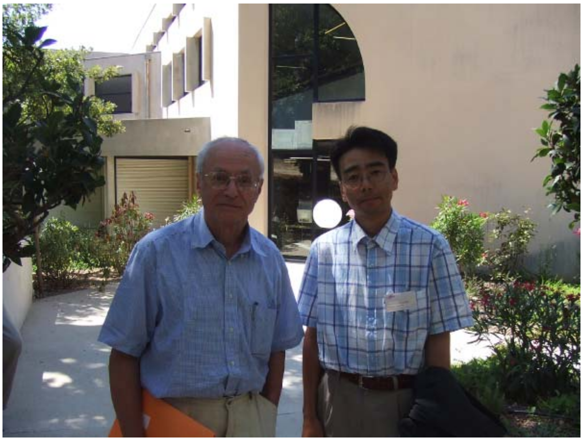
[2·] AMS: American Mathematical Society

[3·] J.E. Littlewood, *A Mathematician's Miscellany*, Methuen and Co., 1953. This book offers a very good description of the unconscious aspect of creative work.

[4·] A few years ago, my friend R. Bott and myself went to receive a prize in Israel (the Wolf prize) awarded by the Knesset in Jerusalem. Bott had to say a few words on mathematics. He asked me what he should say. I replied: "It's very simple, all you have to explain is this: other sciences seek to discover the laws that God has chosen; mathematics seeks to discover the laws which God has to obey". And that is what he said. The Knesset appreciated it.

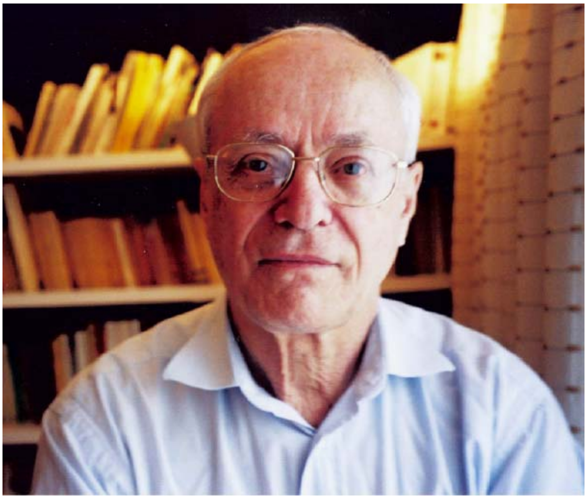
Serre and Henri Cartan, Prix Julia 1970

Anatole Abragam, Serre, and Jaques Tits

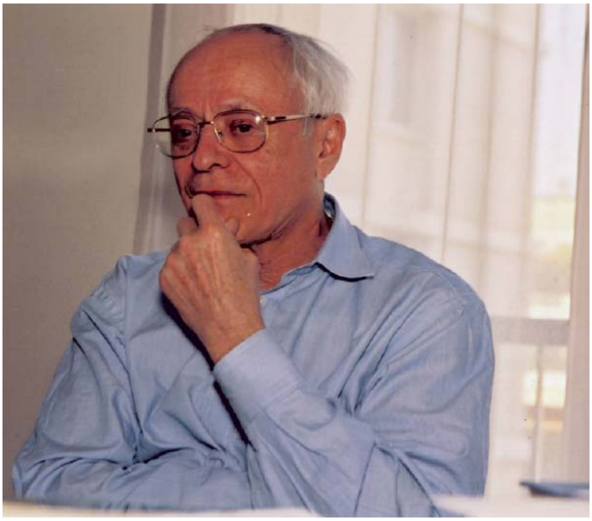


Serre and Yuichiro Taguchi

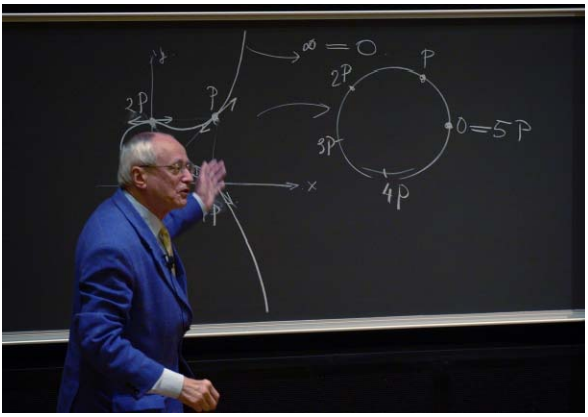
Serre



Serre, May 9, 2003 (photo by Chino Hasebe)



Serre, 2003



The Abel Lecture, Oslo 2003

# Jean-Pierre Serre: An Overview of His Work

**Pilar Bayer**

## Introduction

The work of Jean-Pierre Serre represents an important breakthrough in at least four mathematical areas: algebraic topology, algebraic geometry, algebra, and number theory. His outstanding mathematical achievements have been a source of inspiration for many mathematicians. His contributions to the field were recognized in 2003 when he was awarded the Abel Prize by the Norwegian Academy of Sciences and Letters, presented on that occasion for the first time.

To date, four volumes of Serre's work [*Œuvres, Collected Papers* I–II–III(1986); IV(2000)] ([S210], [S211], [S212]; [S261]) have been published by Springer-Verlag. These volumes include 173 papers, from 1949 to 1998, together with comments on later developments added by the author himself. Some of the papers that he coauthored with A. Borel are to be found in [A. Borel. *Œuvres*, *Collected Papers*. Springer, 1983]. Serre has written some twenty books, which have been frequently reprinted and translated into several languages (mostly English and Russian, but sometimes also Chinese, German, Japanese, Polish or Spanish). He has also delivered lectures in many seminars: Bourbaki, Cartan, Chevalley, Delange–Pisot–Poitou, Grothendieck, Sophus Lie, etc.; some of them have been gathered in the books [S262, SEM(2001; 2008)].

Summarizing his work is a difficult task—especially because his papers present a rich web of interrelationships and hence can hardly be put in linear order. Here, we have limited ourselves to a presentation of their contents, with only a brief discussion of their innovative character.

Research supported in part by MTM2006-04895.

P. Bayer (✉)
Departament d'Àlgebra i Geometria, Facultat de Matemàtiques, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain
e-mail: bayer@ub.edu
url: http://atlas.mat.ub.es/personals/bayer/

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_4, © Springer-Verlag Berlin Heidelberg 2010

Broadly speaking, the references to his publications are presented thematically and chronologically. In order to facilitate their location, the number of a paper corresponding to the present List of Publications is followed by its number in the *Œuvres* (if applicable) and by the year of its publication. Thus, a quotation of the form [S216, Œ 143(1987)] will refer to paper 216 of the List included in the *Œuvres* as number 143. Serre's books will be referred to in accordance with the List and the References at the end of this manuscript. The names of other authors, followed by a date, will denote the existence of a publication but, for the sake of simplicity, no explicit mention of it will be made.

A significant part of Serre's work was given in his annual courses at the Collège de France. When we mention one of these, it will be understood to be a course held at this institution, unless otherwise stated.

## 1 The Beginnings

The mathematical training of J.-P. Serre can be seen as coming from two (closely related) sources. On one hand, in 1948 and just after having finished his studies at the École Normale Supérieure, he starts working at the Séminaire Cartan; this was a very active collaboration that was continued for about 6 years (giving and writing lectures in homological algebra, topology, and functions of several complex variables). On the other hand, since 1949 and for about 25 years, he works with Bourbaki.

In the fifties, Serre publishes his first papers, some of them coauthored with H. Cartan, A. Borel and G.P. Hochschild, and submits a doctoral dissertation under the supervision of Cartan. In an early publication, Borel and Serre [S17, Œ 2 (1950)] prove the impossibility of fibering an Euclidean space with compact fibers (not reduced to one point).

**1.0.1.** Serre's thesis was entitled *Homologie singulière des espaces fibrés. Applications* [S25, Œ 9(1951)] and was followed by several publications: [S19, Œ 4(1950)], [S20, Œ 5(1951)], [S21, Œ 6(1951)]. Its initial purpose was to compute the cohomology groups of the Eilenberg–MacLane complexes $K(\Pi, n)$ by induction on $n$, using the fact that the loop space of $K(\Pi, n)$ is $K(\Pi, n-1)$, combined with the loop fibration (see below). The homotopy lifting property, imposed by Serre on fiber spaces, allows him to construct a spectral sequence in singular homology, analogous to the one obtained by J. Leray (1950) in the setting of Čech theory. A dual spectral sequence also exists for cohomology. The concept of fiber space that he introduces is more general than the one that was usual at the time, allowing him to deal with loop spaces, as follows: given a pathwise-connected topological space $X$ and a point $x \in X$, the loop space $\Omega$ at $x$ is viewed as the fiber of a fiber space $E$ over the base $X$. The elements of $E$ are the paths of $X$ starting at $x$. The crucial fact is that the map which assigns to each path its endpoint is a fibering $f : E \to X$ in the above sense. The space $E$ is contractible and, once Leray's theory is suitably adapted, it turns to be very useful for relating the homology of $\Omega$ to that of $X$.

Serre's thesis contains several applications. For example, by combining Morse's theory (1938) with his own results, Serre proves that, on every compact connected Riemannian manifold, there exist infinitely many geodesics connecting any two distinct points. But, undoubtedly, the most remarkable application is the one concerning the computation of the homotopy groups of spheres, $\pi_i(S_n)$.

**1.1. Homotopy Groups of Spheres.** Earlier studies by H. Hopf, H. Freudenthal (1938) and others had determined the groups $\pi_i(S_n)$ for $i < n + 2$. L. Pontrjagin and G.W. Whitehead (1950) had computed the groups $\pi_{n+2}(S_n)$ and H. Hopf had proved that the group $\pi_{2n-1}(S_n)$, for $n$ even, has $\mathbf{Z}$ as a quotient (and hence is infinite). Thanks also to Freudenthal's suspension theorem, it was known that the group $\pi_{n+k}(S_n)$ depends only on $k$ if $n > k + 1$. However, it was not even known that the $\pi_i(S_n)$ are finitely generated groups. Serre shows that they are; what is more, he shows that the groups $\pi_i(S_n)$, for $i > n$, are *finite*, except for $\pi_{2n-1}(S_n)$ when $n$ is even, which is the direct sum of $\mathbf{Z}$ and a finite group. Given a prime $p$, he also shows that the $p$-primary component of $\pi_i(S_n)$ is zero if $i < n + 2p - 3$ and $n \geqslant 3$; and that the $p$-primary component of $\pi_{n+2p-3}(S_n)$ is cyclic (he proved later that it has order $p$).

**1.1.1.** The study of the homotopy groups was pursued by Serre for about two years; the results were published in several papers: [S31, Œ 12(1952)], [S32, Œ 13(1952)], [S48, Œ 18(1953)], [S43, Œ 19(1953)], [S40, Œ 22(1953)]. He also wrote two *Comptes rendus* notes, with Cartan, on the technique of "killing homotopy groups": [S29, Œ 10(1952)], [S30, Œ 11(1952)]; and two papers with Borel on the use of Steenrod operations [S26, Œ 8(1951)], [S44 (1951)], a consequence being that the only spheres which have an almost complex structure are $S_0$, $S_2$ and $S_6$ (whether $S_6$ has a complex structure or not is still a very interesting open question, despite several attempts to prove the opposite).

**1.1.2.** Soon after his thesis, Serre was invited to Princeton. During his stay (January–February 1952), he realized that some kind of "localization process" is possible in the computation of homotopy groups. More precisely, the paper [S48, Œ 18(1953)] introduces a "mod $\mathcal{C}$" terminology in which a class of objects $\mathcal{C}$ is treated as "zero", as is done in arithmetic mod $p$. For instance, he proves that the groups $\pi_i(S_n)$, $n$ even, are $\mathcal{C}$-isomorphic to the direct sum of $\pi_{i-1}(S_{n-1})$ and $\pi_i(S_{2n-1})$, where $\mathcal{C}$ denotes the class of the finite 2-groups.

The paper [S48, Œ 18(1953)] also shows that every connected compact Lie group is homotopically equivalent to a product of spheres, modulo certain exceptional prime numbers; for classical Lie groups they are those which are $\leqslant h$ where $h$ is the Coxeter number. In the paper [S43, Œ 19(1953)], Serre determines the mod 2 cohomology algebra of an Eilenberg–MacLane complex $\mathrm{K}(\Pi; q)$, in the case where the abelian group $\Pi$ is finitely generated. For this he combines results from both Borel's thesis and his own. He also determines the asymptotic behaviour of the Poincaré series of that algebra (by analytic arguments, similar to those used in the theory of partitions), and deduces that, for any given $n > 1$, there are infinitely many $i$'s such that $\pi_i(S_n)$ has even order.

**1.1.3.** In the same paper, he computes the groups $\pi_{n+i}(S_n)$ for $i \leqslant 4$; and in [S32, Œ 13(1952)] and [S40, Œ 22(1953)], he goes up to $i \leqslant 8$. (These groups are now known for larger values of $i$, but there is very little information on their asymptotic behaviour for $i \to \infty$.)

**1.2. Hochschild–Serre Spectral Sequence.** The first study of the cohomology of group extensions was R. Lyndon's thesis (1948). Serre [S18, Œ 3(1950)] and Hochschild–Serre [S46, Œ 15(1953)] go further. Given a discrete group $G$, a normal subgroup $K$ of $G$, and a $G$-module $A$, they construct a spectral sequence

$$H(G/K, H(K, A)) \Rightarrow H(G, A).$$

If $H^r(K, A) = 0$, for $0 < r < q$, the spectral sequence gives rise to the exact sequence

$$0 \to H^q(G/K, A^K) \to H^q(G, A) \to H^q(K, A)^{G/K} \to H^{q+1}(G/K, A^K)$$
$$\to H^{q+1}(G, A).$$

This sequence became a key ingredient in many proofs. Similar results hold for Lie algebras, as shown in [S47, Œ 16(1953)].

**1.3. Sheaf Cohomology of Complex Manifolds.** In his seminar at the École Normale Supérieure, Cartan showed in 1952–1953 that earlier results of K. Oka and himself can be reinterpreted and generalized in the setting of Stein manifolds by using analytic coherent sheaves and their cohomology; he thus obtained his well-known "Theorems A and B". In [S42, Œ 23(1953)] (see also the letters to Cartan reproduced in [S231 (1991)]), Serre gives several applications of Cartan's theorems; he shows for instance that the Betti numbers of a Stein manifold of complex dimension $n$ can be computed à la de Rham, using holomorphic differential forms; in particular, they vanish in dimension $> n$. But Serre soon became more interested in compact complex manifolds, and especially in algebraic ones. A first step was the theorem (obtained in collaboration with Cartan, see [S41, Œ 24(1953)]) that the cohomology groups $H^q(X, \mathcal{F})$, associated to a compact complex manifold $X$ and with values in an analytic coherent sheaf $\mathcal{F}$, are finite dimensional vector spaces; the proof is based on a result due to L. Schwartz on completely continuous maps between Fréchet spaces. This finiteness result played an essential role in "GAGA", see Sect. 2.2.

**1.3.1.** In a paper dedicated to H. Hopf, Serre [S58, Œ 28(1955)] proves a "duality theorem" in the setting of complex manifolds. The proof is based on Schwartz's theory of distributions (a distribution can be viewed either as a generalized function, or as a linear form on smooth functions; hence, distribution theory has a built-in self-duality).

**1.3.2.** Previously, in a letter [Œ 20(1953)] addressed to Borel, Serre had conjectured a generalization of the Riemann–Roch theorem to varieties of higher dimension. This generalization was soon proved by F. Hirzebruch in his well-known *Habilitationsschrift* and presented by Serre at the Séminaire Bourbaki. The more general version of the Riemann–Roch theorem, due to Grothendieck (1957), was the topic of a Princeton seminar by Borel and Serre. A detailed account appeared in [S74 (1958)], a paper included in [A. Borel. *Œuvres*, *Collected Papers*, no. 44]), and which was for many years the only reference on this topic.

**1.4. The Amsterdam Congress.** At the International Congress of Mathematicians, held in Amsterdam in 1954, K. Kodaira and J.-P. Serre were awarded the Fields Medals. With respect to Serre, the committee acknowledged the new insights he had provided in topology and algebraic geometry. At not quite 28, Serre became the youngest mathematician to receive the distinction, a record that still stands today. In his presentation of the Fields medalists, H. Weyl (perhaps a little worried by Serre's youth) recommended the two laureates to "carry on as you began!". In the following sections we shall see just how far Serre followed Weil's advice.

In his address at the International Congress [S63, Œ 27(1956)], Serre describes the extension of sheaf theory to algebraic varieties defined over a field of any characteristic (see "FAC", Sect. 2.1). One of the highlights is the algebraic analogue of the analytic "duality theorem" mentioned above (it was soon vastly generalized by Grothendieck). He also mentions the following problem: If $X$ is a non-singular projective variety, is it true that the formula

$$B_n = \sum_{p+q=n} \dim H^q(X, \Omega^p)$$

yields the $n$-th Betti numbers of $X$ occurring in Weil conjectures? This is so for most varieties, but there are counterexamples due to J. Igusa (1955).

## 2 Sheaf Cohomology

The paper FAC (1955) by Serre, the paper *Tôhoku* (1957) by Grothendieck, and the book by R. Godement on sheaf theory (1964) were publications that did the most to stimulate the emergence of a new methodology in topology and abstract algebraic geometry. This new approach emerged in the next twenty years through the momentous work [EGA (1960–1964)], [SGA (1968; 1971–1977)] accomplished by Grothendieck and his collaborators.

**2.1. FAC.** In his foundational paper entitled *Faisceaux algébriques cohérents* [S56, Œ 29(1955)], known as FAC, Serre introduces coherent sheaves in the setting of algebraic varieties over an algebraically closed field $k$ of arbitrary characteristic. There are three chapters in FAC.

Chapter I is devoted to coherent sheaves and general sheaf theory. Chapter II starts with a sheaf-style definition of what an “algebraic variety” is (with the restriction that its local rings are reduced), and then shows that the theory of affine algebraic varieties is similar to Cartan’s theory of Stein manifolds: the higher cohomology groups of a coherent sheaf are zero.

Chapter III is devoted to projective varieties. The cohomology groups of coherent sheaves are usually non-zero (but they are finite-dimensional); it is shown that they can be computed algebraically, using the “Ext” functors which had just been defined by Cartan–Eilenberg (this was their first application to algebraic geometry—there would be many others . . . ).

**2.2. GAGA.** In 1956, Serre was appointed *Professeur* at the Collège de France, in the chair *Algèbre et Géométrie*. The same year, he published the paper *Géométrie algébrique et géométrie analytique* [S57, Œ 32(1955–1956)], usually known as GAGA, in which he compares the algebraic and the analytic aspects of the complex projective varieties. The main result is the following:

> Assume that $X$ is a projective variety over $\mathbf{C}$, and let $X^h$ be the complex analytic space associated to $X$. Then the natural functor “algebraic coherent sheaves on $X$” → “analytic coherent sheaves on $X^h$” is an equivalence of categories, which preserves cohomology.

As applications, we mention the invariance of the Betti numbers under automorphisms of the complex field $\mathbf{C}$, when $X$ is non-singular, as well as the comparison of principal algebraic fiber bundles of base $X$ and principal analytic fiber bundles of base $X^h$ with the same structural group $G$.

GAGA contains an appendix introducing the notion of flatness, and applying it to compare the algebraic and the analytic local rings of $X$ and $X^h$ at a given point. Flatness was to play an important role in Grothendieck’s later work.

**2.2.1.** Let $X$ be a normal analytic space and $S$ a closed analytic subset of $X$ with $\operatorname{codim}(S) \geqslant 2$ at every point. In [S120, Œ 68(1966)], Serre studies the extendibility of coherent analytic sheaves $\mathcal{F}$ on $X - S$. He shows that it is equivalent to the coherence of the direct image $i_*(\mathcal{F})$, where $i : X - S \to X$ denotes the inclusion. When $X$ is projective, this implies that the extendible sheaves are the same as the algebraic ones.

**2.3. Cohomology of Algebraic Varieties.** The paper [S68, Œ 35(1957)] gives a cohomological characterization of affine varieties, similar to that of Stein manifolds (cf. [S42, Œ 23(1953)]).

**2.3.1.** In his lecture [S76, Œ 38(1958)] at the International Symposium on Algebraic Topology held in Mexico City, Serre associates to an algebraic variety $X$, defined over an algebraically closed field $k$ of characteristic $p > 0$, its cohomology groups $H^i(X, \mathcal{W})$ with values in a sheaf of Witt vectors $\mathcal{W}$. Although this did not provide suitable Betti numbers, the paper contains many ideas that paved the way for the birth of crystalline and $p$-adic cohomology. We stress the treatment given in this work to the Frobenius endomorphism $F$, as a semilinear endomorphism of $H^1(X, \mathcal{O})$, when $X$ is a non-singular projective curve. The space $H^1(X, \mathcal{O})$ may be identified with a space of classes of repartitions (or "adèles") over the function field of $X$ and its Frobenius endomorphism gives the Hasse–Witt matrix of $X$. By using Cartier's operator on differential forms, Serre proves that $H^1(X, \mathcal{W})$ is a free module of rank $2g - s$ over the Witt ring $W(k)$, where $g$ denotes the genus of the curve and $p^s$ is the number of divisor classes of $X$ killed by $p$.

**2.3.2.** The above results were completed in the paper [S75, Œ 40(1958)], dedicated to E. Artin. Given an abelian variety $A$, Serre shows that the cohomology algebra $H^*(A, \mathcal{O})$ is the exterior algebra of the vector space $H^1(A, \mathcal{O})$, as in the classical case. He also shows that the Bockstein's operations are zero; that is, $A$ has no cohomological "torsion". Moreover, he gives an example of an abelian variety for which $H^2(A, \mathcal{W})$ is not a finitely generated module over the Witt vectors, thus contradicting an "imprudent conjecture" (sic) he had made in [S76, Œ 38(1958)].

**2.3.3.** In [S76, Œ 38(1958)], the influence of André Weil is clear (as it is in many of Serre's other papers). The search for a good cohomology for varieties defined over finite fields was motivated by the Weil conjectures (1948) on the zeta function of these varieties. As is well known, such a cohomology was developed a few years later by Grothendieck, using étale topology.

**2.3.4.** In [S89, Œ 45(1960)], Serre shows that Weil's conjectures could be proved easily if (a big "if") some basic properties of the cohomology of complex Kähler varieties could be extended to projective varieties over a finite field. This was the starting point for Grothendieck's formulation of the so-called "standard conjectures" on motives, which are still unproved today.

**2.3.5.** In [S94, Œ 50(1961)], Serre constructs a non-singular projective variety in characteristic $p > 0$ which cannot be lifted to characteristic zero. He recently (2005) improved this result by showing that, if the variety can be lifted (as a flat scheme) to a local ring $A$, then $p \cdot A = 0$. The basic idea consists in transposing the problem to the context of finite groups.

**2.3.6.** In his lecture [S103, Œ 56(1963)] at the International Congress of Mathematicians held in Stockholm in 1962, Serre offered a summary of scheme theory. After revising Grothendieck's notion of Grassmannian, Hilbert scheme, Picard scheme, and moduli scheme (for curves of a given genus), he goes on to schemes over complete noetherian local rings, where he mentions the very interesting and, in those days, recent results of Néron. (Although the language of schemes would become usual in Serre's texts, it is worth saying that he has never overused it.)

# 3 Lie Groups and Lie Algebras

Serre's interest in Lie theory was already apparent in his complementary thesis [S28, Œ 14(1952)], which contains a presentation of the results on Hilbert's fifth problem up to 1951 (i.e. just before it was solved by A.M. Gleason, D. Montgomery and L. Zippin).

**3.0.1.** In 1953, Borel and Serre began to be interested in the finite subgroups of compact Lie groups—a topic to which they would return several times in later years, see e.g. [S250, Œ 167(1996)] and [S260, Œ 174(1999)]. In [S45 (1953)], reproduced in [A. Borel. *Œuvres*, *Collected Papers*, no. 24], they prove that every supersolvable finite subgroup of a compact Lie group $G$ is contained in the normalizer $N$ of a maximal torus $T$ of $G$; this is a generalization of a theorem of Blichfeldt relative to $G = \mathbf{U}_n$. In particular, the determination of the abelian subgroups of $G$ is reduced to that of the abelian subgroups of $N$. Borel and Serre were especially interested in the $p$-elementary abelian subgroups of $G$, for $p$ a prime. They defined the $p$-rank $\ell_p$ of $G$ as the largest integer $n$ such that $G$ contains such a subgroup with order $p^n$; the theorem above shows that $\ell \leqslant \ell_p(G) \leqslant \ell + \ell_p(W)$, where $\ell$ is the rank of $G$ and $\ell_p(W)$ the $p$-rank of its Weyl group $W = N/T$. They show that, if $G$ is connected and its $p$-rank is greater than its rank, then $G$ has homological $p$-torsion. As a corollary, the compact Lie groups of type $G_2$, $F_4$, $E_7$ and $E_8$ have homological 2-torsion. The proof uses results on the cohomology algebra modulo $p$ of the classifying space $B_G$ of $G$, which had been studied in Borel's thesis.

**3.0.2.** Serre's book *Lie Algebras and Lie Groups* [S110, LALG(1965)] was based on a course at Harvard. As indicated by its title, it consists of two parts; the first one gives the general theory of Lie algebras in characteristic zero, including the standard theorems of Lie, Engel, Cartan and Whitehead (but not including root systems). The second one is about analytic manifolds over a complete field $k$, either real, complex or ultrametric. It is in this context that Serre gives the standard Lie dictionary

$$\text{Lie groups} \to \text{Lie algebras},$$

assuming that $k$ has characteristic zero. His interest in the $p$-adic case arose when he realized around 1962 that the rather mysterious Galois groups associated to the Tate modules of abelian varieties (see Sect. 13) are $p$-adic Lie groups, so that their Lie algebras are accessible to the general Lie theory. This elementary remark opened up many possibilities, since it is much easier to classify Lie algebras than profinite groups.

**3.0.3.** The booklet *Algèbres de Lie semi-simples complexes* [S119, ALSC(1966)] reproduces a series of lectures given in Algiers in 1965. It gives a concise introduction (mostly with proofs) to complex semisimple Lie algebras, and thus supplements *Lie Algebras and Lie Groups*. The main chapters are those on Cartan subalgebras, representation theory for $\mathfrak{sl}_2$, root systems and their Weyl groups, structure theorems for semisimple Lie algebras, linear representations of semisimple Lie algebras, Weil's character formula (without proof), and the dictionary between compact Lie groups and reductive algebraic groups over **C** (without proof, but see *Gèbres* [S239, Œ 160(1993)]). The book also gives a presentation of semisimple Lie algebras by generators and relations (including the so-called "Serre relations" which, as he says, should be called "Chevalley relations" because of their earlier use by Chevalley).

## 4 Local Algebra

Serre's work in FAC and GAGA made him introduce homological methods in local algebra, such as *flatness* and the characterization of regular local rings as the only noetherian local rings of finite homological dimension (completing an earlier result of A. Auslander and D. Buchsbaum (1956), cf. [S64, Œ 33(1956)]).

**4.0.1.** The general theory of local rings was the subject of the lecture course [S79, Œ 42(1958)], which was later published in book form *Algèbre Locale. Multiplicités* [S111, ALM(1965)]. Its topics include the general theory of noetherian modules and their primary decomposition, Hilbert polynomials, integral extensions, Krull–Samuel dimension theory, the Koszul complex, Cohen–Macaulay modules, and the homological characterization of regular local rings mentioned above. The book culminates with the celebrated "Tor formula" which gives a homological definition for intersection multiplicities in algebraic geometry in terms of Euler–Poincaré characteristics. This led Serre to several conjectures on regular local rings of mixed characteristic; most of them (but not all) were later proved by P. Roberts (1985), H. Gillet–C. Soulé (1985) and O. Gabber. The book had a profound influence on a whole generation of algebraists.

## 5 Projective Modules

Given an algebraic vector bundle $E$ over an algebraic variety $V$, let $\mathcal{S}(E)$ denote its sheaf of sections. As pointed out in FAC, one gets in this way an equivalence between vector bundles and locally free coherent $\mathcal{O}$-sheaves. When $V$ is affine, with coordinate ring $A = \Gamma(V, \mathcal{O})$, this may be viewed as a correspondence between vector bundles and finitely generated projective $A$-modules; under this correspondence trivial bundles correspond to free modules.

**5.0.1.** The above considerations apply when $V$ is the affine $n$-space over a field $k$, in which case $A$ is the polynomial ring $k[X_1, \dots, X_n]$; Serre mentions in FAC that he "does not know of any finitely generated projective $k[X_1, \dots, X_n]$-module which is not free". This gave rise to the so-called *Serre conjecture*, although it had been stated as a "problem" and not as a "conjecture". Much work was done on it (see e.g. the book by T.Y. Lam called *Serre's Conjecture* in its 1977 edition and *Serre's Problem* in its 2006 one). The case $n = 2$ was solved by C.S. Seshadri (1979); see Serre's report on it in Séminaire Dubreil–Pisot [Œ 48(1960/61)]; this report also gives an interesting relation between the problem for $n = 3$ and curves in affine 3-space which are complete intersections.

Twenty years after the publication of FAC, and after partial results had been obtained by several authors (especially in dimension 3), D. Quillen (1976) and A. Suslin (1976), independently and simultaneously, solved Serre's problem in any dimension.

**5.0.2.** In his contribution [S73, Œ 39(1958)] at the Séminaire Dubreil–Dubreil-Jacotin–Pisot, Serre applies the "projective modules = vector bundles" idea to an arbitrary commutative ring $A$. Guided by transversality arguments of topology, he proves the following splitting theorem:

> Assume $A$ is commutative, noetherian, and that Spec($A$) is connected. Then every finitely generated projective $A$-module is the direct sum of a free $A$-module and of a projective $A$-module the rank of which does not exceed the dimension of the maximal spectrum of $A$.

When $\dim(A) = 1$, one recovers the theorem of Steinitz–Chevalley on the structure of the torsion-free modules over Dedekind rings.

## 6 Algebraic Number Fields

The Séminaire Bourbaki report [S70, Œ 41(1958)] contains an exposition of Iwasawa's theory for the $p$-cyclotomic towers of number fields, and the $p$-components of their ideal class groups. The main difference with Iwasawa's papers is that the structure theorems for the so-called $\Gamma$-modules are deduced from general statements on regular local rings of dimension 2; this viewpoint has now become the standard approach to such questions.

**6.0.1.** *Cours d'Arithmétique* [S146, CA(1970)] arose as a product of two lecture courses taught in 1962 and 1964 at the École Normale Supérieure. The book (which in its first edition had the format 11 cm × 18 cm and cost only 12 francs) has been frequently translated and reprinted, and has been the most accessible introduction to certain chapters of number theory for many years. The first part, which is purely algebraic, gives the classification of quadratic forms over $\mathbf{Q}$. We find there equations over finite fields, two proofs of the quadratic reciprocity law, an introduction to $p$-adic numbers, and properties of the Hilbert symbol. The quadratic forms are studied over $\mathbf{Q}_p$, over $\mathbf{Q}$, as well as over $\mathbf{Z}$ (in the case of discriminant $\pm 1$). The second part of the book uses analytic methods. It contains a chapter on $L$-functions, culminating in the standard proof of Dirichlet's theorem on primes in arithmetic progressions, and a chapter on modular forms of level 1, together with their relations with elliptic curves, Eisenstein series, Hecke operators, theta functions and

Ramanujan's $\tau$ function. In 1995, Serre was awarded the Leroy P. Steele Prize for Mathematical Exposition for this delightful text.

**6.0.2.** The notion of a *$p$-adic modular form* was introduced by Serre in the paper [S158, Œ 97(1973)], which is dedicated to C.L. Siegel. Such a form is defined as a limit of modular forms in the usual sense. By using them, together with previous results on modular forms mod $p$ due to Swinnerton-Dyer and himself, he constructs the *$p$-adic zeta function* of a totally real algebraic number field $K$. This function interpolates $p$-adically the values at the negative integers of the Dedekind zeta function $\zeta_K(s)$ (after removal of its $p$-factors); these numbers were already known to be rational, thanks to a theorem of Siegel (1937). Serre's results generalize the one obtained by Kubota–Leopoldt in the sixties when $K$ is abelian over $\mathbf{Q}$. They were completed later by D. Barsky (1978), Pierrette Cassou-Noguès (1979) and P. Deligne–K. Ribet (1980).

## 7 Class Field Theory

Class field theory describes the abelian extensions of certain fields by means of what are known as reciprocity isomorphisms. Sometimes, the reciprocity isomorphisms can be made explicit by means of a symbol computation. The first historical example is the quadratic reciprocity law of Legendre and Gauss. The cohomological treatment of class field theory started with papers of G.P. Hochschild, E. Artin, J. Tate, A. Weil and T. Nakayama.

**7.1. Geometric Class Field Theory.** *Groupes Algébriques et Corps de Classes* [S82, GACC(1959)] was the first book that Serre published. It evolved from his first course at the Collège de France [S69, Œ 37(1957)] and its content is mainly based on earlier papers by S. Lang (1956) and M. Rosenlicht (1957).

Chapter I is a résumé of the book. Chapter II gives the main theorems on algebraic curves, including Riemann–Roch and the duality theorem (with proofs). Chapters III–IV are devoted to a theorem of Rosenlicht stating that every rational function $f : X \to G$, from a non-singular irreducible projective curve $X$ to a commutative algebraic group $G$, factors through a generalized Jacobian $J_\mathfrak{m}$. A generalized Jacobian is a commutative algebraic group, which is an extension of an abelian variety (the usual Jacobian $J$) by an algebraic linear group $L_\mathfrak{m}$, depending on a "modulus" $\mathfrak{m}$. The groups $L_\mathfrak{m}$ provide the local symbols in class field theory. In Chap. V it is shown that every abelian covering of an irreducible algebraic curve is the pull-back of a separable isogeny of a generalized Jacobian. When $\mathfrak{m}$ varies, the generalized Jacobians $J_\mathfrak{m}$ form a projective system which is the geometric analogue of the idèle class group. Class field theory for function fields in one variable over finite fields is dealt with in Chap. VI. The reciprocity isomorphism is proved and explicit computations of norm residue symbols are made. Chapter VII contains a general cohomological treatment of extensions of commutative algebraic groups.

**7.1.1.** Based on the lecture course [S91, Œ 47(1960)], a theory of commutative pro-algebraic groups is developed in [S88, Œ 49(1960)]. Its application to geometric class field theory can be found in [S93, Œ 51(1961)].

Let $k$ be an algebraically closed field. A commutative quasi-algebraic group over $k$ is defined as a pure inseparable isogeny class of commutative algebraic groups over $k$. If $G$ is such a group and is connected, then it has a unique connected linear subgroup $L$ such that the quotient $G/L$ is an abelian variety. As for the group $L$, it is the product of a torus $T$ by a unipotent group $U$. The group $T$ is a product of groups isomorphic to $\mathbf{G}_m$ and the group $U$ has a composition series whose quotients are isomorphic to $\mathbf{G}_a$; it is isogenous to a product of truncated Witt vector groups. The isomorphism classes of the groups $\mathbf{G}_a$, $\mathbf{G}_m$, cyclic groups of prime order, and simple abelian varieties are called the elementary commutative quasi-algebraic groups. The commutative quasi-algebraic groups form an abelian category $\mathcal{Q}$. The finite commutative quasi-algebraic groups form a subcategory $\mathcal{Q}_0$ of $\mathcal{Q}$. If $G$ is a commutative quasi-algebraic group and $G^0$ is its connected component, the quotient $\pi_0(G) = G/G^0$ is a finite abelian group. The category of commutative pro-algebraic groups is defined as $\mathcal{P} = \mathrm{Pro}(\mathcal{Q})$. Let $\mathcal{P}_0 = \mathrm{Pro}(\mathcal{Q}_0)$ be the subcategory of abelian profinite groups. The category $\mathcal{P}$ has projective limits and enough projective objects. Every projective object of $\mathcal{P}$ is a product of indecomposable projectives and the indecomposable projective groups coincide with the projective envelopes of the elementary commutative quasi-algebraic groups. The functor $\pi_0 : \mathcal{P} \to \mathcal{P}_0$ is right exact; its left derived functors are denoted by $G \mapsto \pi_i(G)$. One of the main results of the paper is that $\pi_i(G) = 0$ if $i > 1$. The group $\pi_1(G)$ is called the fundamental group of $G$. The connected and simply connected commutative pro-algebraic groups form a subcategory $\mathcal{S}$ of $\mathcal{P}$. For each object $G$ in $\mathcal{P}$, there exists a unique group $\widetilde{G}$ in $\mathcal{S}$ and a morphism $u : \widetilde{G} \to G$, whose kernel and cokernel belong to $\mathcal{P}_0$, so that one obtains an exact sequence

$$1 \to \pi_1(G) \to \widetilde{G} \to G \to \pi_0(G) \to 1.$$

By means of the universal covering functor, the categories $\mathcal{P}/\mathcal{P}_0$ and $\mathcal{S}$ become equivalent.

After computing the homotopy groups of the elementary commutative pro-algebraic groups, it is shown in [S88, Œ 49(1960)] that every commutative pro-algebraic group has cohomological dimension $\leqslant 2$, if $k$ has positive characteristic; and has cohomological dimension $\leqslant 1$, if $k$ has characteristic zero.

**7.1.2.** The paper [S93, Œ 51(1961)] is a sequel to the one mentioned above (and is also its motivation). It deals with local class field theory in the geometric setting (in an Oberwolfach lecture, Serre once described it as *reine geometrische Klassenkörpertheorie im Kleinen*). Let $K$ be a field which is complete with respect to a discrete valuation and suppose that its residue field $k$ is algebraically closed. By using a construction of M. Greenberg, the group of units $U_K$ of $K$ may be viewed as a commutative pro-algebraic group over $k$, so that the fundamental group

$\pi_1(U_K)$ is well defined. The reciprocity isomorphism takes the simple form:

$$\pi_1(U_K) \xrightarrow{\sim} G_K^{\mathrm{ab}},$$

where $G_K^{\mathrm{ab}}$ denotes the Galois group of the maximal abelian extension of $K$. This isomorphism is compatible with the natural filtration of $\pi_1(U_K)$ and the filtration of $G_K^{\mathrm{ab}}$ given by the upper numbering of the ramification groups. Hence there is a conductor theory, related to Artin representations (see below).

**7.1.3.** An Artin representation $a$ has a $\mathbf{Z}$-valued character. In the paper [S90, Œ 46(1960)], it is shown that $a$ is rational over $\mathbf{Q}_\ell$ provided $\ell$ is different from the residue characteristic, but it is not always rational over $\mathbf{Q}$. The same paper conjectures the existence of a conductor theory for regular local rings of any dimension, analogous to the one in dimension 1; a few results have been obtained on this recently by K. Kato and his school, but the general case is still open.

**7.1.4.** An example, in the geometric case, of a separable covering of curves with a relative different whose class is not a square was given in a joint paper with A. Fröhlich and J. Tate [S100, Œ 54(1962)]. Such an example does not exist for number fields, by a well-known result of E. Hecke.

**7.1.5.** In [S150, Œ 92(1971)], Serre considers a Dedekind ring $A$ of field of fractions $K$, a finite Galois extension $L/K$ with Galois group $G$, and a real-valued virtual character $\chi$ of $G$. Under the assumption that either the extension $L/K$ is tamely ramified or that $\chi$ can be expressed as the difference of two characters of real linear representations, he proves that the Artin conductor $\mathfrak{f} = \mathfrak{f}(\chi, L/K)$ is a square in the group of ideal classes of $A$.

**7.2. Local Class Field Theory.** Group cohomology and, more specifically, Galois cohomology is the subject of the lecture course [S83, Œ 44(1959)]. The content of this course can be found in *Corps Locaux* [S98, CL(1962)].

The purpose of [CL] was to provide a cohomological presentation of local class field theory, for valued fields which are complete with respect to a discrete valuation with finite residue field. In the first part, one finds the structure theorem of complete discrete valuation rings. In the second part, Hilbert's ramification theory is given, with the inclusion of the upper numbering of the ramification groups, due to J. Herbrand, and the properties of the Artin representation, a notion due to Weil in his paper *L'avenir des mathématiques* (1947). The third part of [CL] is about group cohomology. It includes the cohomological interpretation of the Brauer group $\mathrm{Br}(k)$ of a field $k$ and class-formations à la Artin–Tate.

Local class field theory takes up the fourth part of the book. The reciprocity isomorphism is obtained from the class formation associated to the original local field; it is made explicit by means of a computation of norm residue symbols based on a theorem of B. Dwork (1958).

One also finds in [CL] the first definitions of non-abelian Galois cohomology. Given a Galois extension $K/k$ and an algebraic group $G$ defined over $k$, the elements of the set $H^1(\mathrm{Gal}(K/k), G(K))$ describe the classes of principal homogeneous $G$-spaces over $k$ which have a rational point in $K$. Easy arguments show that $H^1(\mathrm{Gal}(K/k), G(K)) = 1$ when $G$ is one of the following algebraic groups: additive $\mathbf{G}_a$, multiplicative $\mathbf{G}_m$, general linear $\mathbf{GL}_n$, and symplectic $\mathbf{Sp}_{2n}$.

**7.2.1.** Another exposition of local class field theory can be found in the lecture [S127, Œ 75(1967)]; it differs from the one given in [CL] by the use of Lubin–Tate theory of formal groups, which allows a neat proof of the "existence theorem".

**7.3. A Local Mass Formula.** Let $K$ denote a local field with finite residue field $k$ of $q$ elements and let $K_s$ be a separable closure of $K$. In [S186, Œ 115(1978)], one finds a *mass formula* for the set $\Sigma_n$ of all totally ramified extensions of $K$ of given degree $n$ contained in $K_s$, namely:

$$\sum_{L\in\Sigma_n} 1/q^{c(L)} = n,$$

where $q^{c(L)}$ is the norm of the wild component of the discriminant of $L/K$. Although the formula could (in principle) be deduced from earlier results of Krasner, Serre proves it independently in two elegant and different ways. The first proof is derived from the volume of the set of Eisenstein polynomials. The second uses the $p$-adic analogue of Weil's integration formula, applied to the multiplicative group $D^*$ of a division algebra $D$ of center $K$ such that $[D:K] = n^2$.

## 8 $p$-adic Analysis

Let $V$ be an algebraic variety over a finite field $k$ of characteristic $p$. One of Weil's conjectures is that the zeta function $Z_V(t)$ is a rational function of t. This was proved in 1960 by B. Dwork. His method involved writing $Z_V(t)$ as an alternating product of $p$-adic Fredholm determinants. This motivated Serre to study the spectral theory of completely continuous operators acting on $p$-adic Banach spaces [S99, Œ 55(1962)]. The paper, which is self-contained, provides an excellent introduction to $p$-adic analysis. Given a completely continuous endomorphism $u$ defined on a Banach space $E$ over a local field, the Fredholm determinant $\det(1-tu)$ is a power series in $t$, which has an infinite radius of convergence and thus defines an entire function of $t$. The Fredholm resolvent $P(t,u) = \det(1-tu)/(1-tu)$ of $u$ is an entire function of $t$ with values in $\mathrm{End}(E)$. Given an element $a \in K$, one shows that the endomorphism $1-au$ is invertible if and only if $\det(1-au) \neq 0$. If this is the case, then the relation $\det(1-au) = (1-au)P(a,u) = P(a,u)(1-au)$ is satisfied. If $a \in K$ is a zero of order $h$ of the function $\det(1-tu)$, then the space $E$ uniquely decomposes into a direct sum of two closed subspaces $N$, $F$

which are invariant under $u$. The endomorphism $1 - au$ is nilpotent on $N$ and invertible on $F$; the dimension of $N$ is $h$, just as in F. Riesz theory over $\mathbf{C}$. Serre proves that, given an exact sequence of Banach spaces and continuous linear mappings, $0 \to E_0 \overset{d_0}{\to} E_1 \to \cdots \overset{d_{n-1}}{\to} E_n \to 0$, and given completely continuous endomorphisms $u_i$ of $E_i$ such that $d_i \circ u_i = u_{i+1} \circ d_i$, for $0 \leqslant i < n$, then $\prod_{i=1}^{n} \det(1 - tu_i)^{(-1)^i} = 1$; this is useful for understanding some of Dwork's computations.

**8.0.1.** In [S113, Œ 65(1965)], the compact $p$-adic analytic manifolds are classified. Given a field $K$, locally compact for the topology defined by a discrete valuation, any compact analytic manifold $X$ defined over $K$, of dimension $n$ at each of its points, is isomorphic to a disjoint finite sum of copies of the ball $A^n$, where $A$ denotes the valuation ring of $K$. Two sums $rA^n$ and $r'A^n$ are isomorphic if and only if $r \equiv r' \bmod(q - 1)$, where $q$ denotes the number of elements of the residue field of $A$. The class of $r \bmod(q - 1)$ is an invariant of the manifold; two $n$-manifolds with the same invariant are isomorphic.

## 9 Group Cohomology

By definition, a profinite group is a projective limit of finite groups; special cases are the pro-$p$-groups, i.e. the projective limits of finite $p$-groups. The most interesting examples of profinite groups are provided by the Galois groups of algebraic extensions and by compact $p$-adic Lie groups.

**9.1. Cohomology of Profinite Groups and $p$-adic Lie Groups.** To each profinite group $G = \varprojlim G_i$ acting in a continuous way on a discrete abelian group $A$, one can associate cohomology groups $H^q(G, A)$ by using continuous cochains. The main properties of the cohomology of profinite groups were obtained by Tate (and also by Grothendieck) in the early 1960s, but were not published. They are collected in the first chapter of *Cohomologie Galoisienne* [S97, CG(1962)].

In the first chapter of [CG], given a prime $p$ and a profinite group $G$, the concepts of cohomological $p$-dimension, denoted by $\mathrm{cd}_p(G)$, and cohomological dimension, denoted by $\mathrm{cd}(G)$, are defined. Some pro-$p$-groups admit a duality theory; they are called Poincaré pro-$p$-groups. Those of cohomological dimension 2 are the “Demushkin groups”. They are especially interesting, since they can be described by one explicit relation (Demushkin, Serre, Labute); they appear as Galois groups of the maximal pro-$p$-extension of $p$-adic fields, cf. [S104, Œ 58(1963)] and [CG].

Chapters II and III are devoted to the study of Galois cohomology in the commutative and the non-commutative cases (most of the results of Chap. II were due to Tate, and an important part of those of Chap. III were due to Borel–Serre).

**9.1.1.** The Bourbaki report [S105, Œ 60(1964)] summarizes M. Lazard's seminal paper (1964) on $p$-adic Lie groups. One of Lazard's main results is that a profinite group is an analytic $p$-adic group if and only if it has an open subgroup $H$ which is a pro-$p$-group and is such that $(H, H) \subset H^p$, if $p \neq 2$; or $(H, H) \subset H^4$, if $p = 2$. If $G$ is a compact $p$-adic Lie group such that $\mathrm{cd}(G) = n < \infty$, then $G$ is a Poincaré pro-$p$-group of dimension $n$ and the character $\chi(x) = \det \mathrm{Ad}(x)$ is the dualizing character of $G$. Here $\mathrm{Ad}(x)$ denotes the adjoint automorphism of $\mathrm{Lie}(G)$ defined by $x$. The group $G$ has finite cohomological dimension if and only if it is torsion-free; the proof combines Lazard's results with the theorem of Serre mentioned below.

**9.1.2.** The paper [S114, Œ 66(1965)] proves that, if $G$ is a profinite group, $p$-torsion-free, then for every open subgroup $U$ of $G$ we have the equality $\mathrm{cd}_p(U) = \mathrm{cd}_p(G)$ between their respective cohomological $p$-dimensions. The proof is rather intricate. In it, Serre makes use of Steenrod powers, a tool which he had acquired during his topological days (cf. [S44 (1953)]). As a corollary, every torsion-free pro-$p$-group which contains a free open subgroup is free. Serre asked whether the discrete analogue of this statement is true, i.e. whether every torsion-free group $G$ which contains a free subgroup of finite index is free. This was proved a few years later by J. Stallings (1968) and R. Swan (1969).

**9.1.3.** More than thirty years later, Serre dedicated [S255, Œ 173(1998)] to John Tate. The paper deals with the Euler characteristic of profinite groups. Given a profinite group $G$ of finite cohomological $p$-dimension and a discrete $G$-module $A$ which is a vector space of finite dimension over the finite field $\mathbf{F}_p$, the Euler characteristic

$$e(G, A) = \sum (-1)^i \dim H^i(G, A)$$

is defined under the assumption that $\dim H^i(G, A) < \infty$, for all $i$.

Let $G_{\mathrm{reg}}$ be the subset of $G$ made up by the regular elements. Serre proves that there exists a distribution $\mu_G$ over $G_{\mathrm{reg}}$ with values in $\mathbf{Q}_p$ such that $e(G, A) = \langle \varphi_A, \mu_G \rangle$, where $\varphi_A : G_{\mathrm{reg}} \to \mathbf{Z}_p$ denotes the Brauer character of the $G$-module A. This distribution can be described explicitly in several cases, e.g. when $G$ is a $p$-torsion-free $p$-adic Lie group, thanks to Lazard's theory.

**9.2. Galois Cohomology.** Let $G = \mathrm{Gal}(K/k)$ be the Galois group of a field extension and suppose that $A$ is a discrete $G$-module. The abelian Galois cohomology groups $H^q(\mathrm{Gal}(K/k), A)$ are usually denoted by $H^q(K/k, A)$, or simply by $H^q(k, A)$ when $K = k_s$ is a separable closure of $k$.

Abelian Galois cohomology, with special emphasis on the results of Tate, was the content of the course [S104, Œ 59(1963)] and of the second chapter of [CG], while the third chapter of [CG] is about non abelian cohomology. After thirty years, Serre returned to both topics in a series of three courses [S234, Œ 153(1991)], [S236, Œ 156(1992)], [S247, Œ 165(1994)].

**9.3. Galois Cohomology of Linear Algebraic Groups.** In his lecture delivered at Brussels in the Colloquium on Algebraic Groups [S96, Œ 53(1962)], Serre presented two conjectures on the cohomology of linear algebraic groups, known as Conjecture I (CI) and Conjecture II (CII).

Given an algebraic group $G$ defined over a field $k$, we may consider the cohomology group $H^0(k,G) = G(k)$, and the cohomology set $H^1(k,G)$ = isomorphism classes of $G$-$k$-torsors.

In what follows, we will suppose that the ground field $k$ is perfect, and we will denote by $\mathrm{cd}(k)$ the cohomological dimension of $\mathrm{Gal}(k_s/k)$. The above conjectures state:

(CI) If $\mathrm{cd}(k) \leqslant 1$ and $G$ is a connected linear group, then $H^1(k,G) = 0$.
(CII) If $\mathrm{cd}(k) \leqslant 2$ and $G$ is a semisimple, simply connected linear group, then $H^1(k,G) = 0$.

At the time, the truth of Conjecture I was only known in the following cases:

- when $k$ is a finite field (S. Lang);
- when $k$ is of characteristic zero and has property "$C_1$" (T. Springer);
- for $G$ solvable, connected and linear;
- for $G$ a classical semisimple group.

Conjecture I was proved a few years later in a beautiful paper by R. Steinberg (1965), which Serre included in the English translation of [CG].

M. Kneser (1965) proved Conjecture II when $k$ is a $p$-adic field, and G. Harder (1965) did the same when $k$ is a totally imaginary algebraic number field and $G$ does not have any factor of type $E_8$; this restriction was removed more than 20 years later by V.I. Chernousov (1989). More generally, for any algebraic number field $k$, and any semisimple, simply connected linear algebraic group $G$, the natural mapping $H^1(k,G) \to \prod_{v\mathrm{real}} H^1(k_v,G)$ is bijective (Hasse's principle), in agreement with a conjecture of Kneser. A detailed presentation of this fact can be found in a book by V. Platonov and A. Rapinchuk (1991).

**9.3.1.** The Galois cohomology of semisimple linear groups was taken up again by Serre in the course [S234, Œ 153(1991)]. One of his objectives was to discuss the "cohomological invariants" of $H^1(k,G)$, i.e. (see below) the relations which connect the non-abelian cohomology set $H^1(k,G)$ and certain Galois cohomology groups $H^i(k,C)$, where $C$ is commutative (e.g. $C = \mathbf{Z}/2\mathbf{Z}$).

More precisely, let us consider a smooth linear algebraic group $G$, defined over a field $k_0$, an integer $i \geqslant 0$, and a finite Galois module $C$ over $k_0$ whose order is coprime to the characteristic.

By definition, a cohomological invariant of type $H^i(-,C)$ is a morphism of the functor $k \mapsto H^1(k,G)$ into the functor $k \mapsto H^i(k,C)$, defined over the category of field extensions $k$ of $k_0$. Suppose that the characteristic of $k$ is not 2. Then, examples of cohomological invariants are provided by

- the Stiefel–Whitney classes $w_i : H^1(k,\mathbf{O}(q)) \to H^i(k,\mathbf{Z}/2\mathbf{Z})$;
- Arason's invariant $a : H^1(k,\mathbf{Spin}(q)) \to H^3(k,\mathbf{Z}/2\mathbf{Z})$;
- Merkurjev–Suslin's invariant $ms : H^1(k,\mathbf{SL}_D) \to H^3(k,\mu_n^{\otimes 2})$;
- Rost's invariants.

**9.3.2.** The presentation of recent work on Galois cohomology and the formulation of some open problems was the purpose of the exposé [S246, Œ 166(1994)] at the Séminaire Bourbaki. To every connected semisimple group $G$ whose root system over $k$ is irreducible, one associates a set of prime numbers $S(G)$ which plays a special role in the study of the cohomology set $H^1(k,G)$. For example, all the divisors of the order of the centre of the universal covering $\widetilde{G}$ of $G$ are included in $S(G)$. Tits' theorem (1992) proves that, given a class $x \in H^1(k,G)$, there exists an extension $k_x/k$ of $S(G)$-primary degree that kills $x$ (that is to say, such that $x$ maps to zero in $H^1(k_x,G)$). Serre asks whether it is true that, given finite extensions $k_i/k$ whose degrees are coprime to $S(G)$, the mapping $H^1(k,G) \to \prod H^1(k_i,G)$ is injective (assuming that $G$ is connected). In this lecture, he also gives extensions and variants of Conjectures I and II which deal with an imperfect ground field or for which one merely assumes that $\mathrm{cd}_p(G) \leqslant 1$ for every $p \in S(G)$. He gives a list of cases in which Conjecture II has been proved, namely:

– groups of type $\mathbf{SL}_D$ associated to elements of norm 1 of a central simple $k$-algebra $D$, of rank $n^2$, by A.S. Merkurjev and A. Suslin (1983, 1985);
– Spin groups (in particular, all those of type $B_n$), by A.S. Merkurjev;
– classical groups (except those of triality type $D_4$), by Eva Bayer and Raman Parimala (1995);
– groups of type $G_2$ and $F_4$.

In conclusion, Conjecture II remains open for the types $E_6, E_7, E_8$ and triality type $D_4$.

**9.4. Self-dual Normal Basis.** E. Bayer-Fluckiger and H.W. Lenstra (1990) defined the notion of a "self-dual normal base" in a $G$-Galois algebra $L/K$, and proved the existence of such a base when $G$ has odd order. When $G$ has even order, existence criteria were given by E. Bayer and Serre [S244, Œ 163(1994)] in the special case where the 2-Sylow subgroups of $G$ are elementary abelian: if $2^d$ is the order of such a Sylow group, they associate to $L/K$ a $d$-Pfister form $q_L$ and show that a self-dual normal base exists if and only if $q_L$ is hyperbolic. (Thanks to Voevodsky's proof of Milnor's conjecture, this criterion can also be stated as the vanishing of a specific element of $H^d(K,\mathbf{Z}/2\mathbf{Z})$.)

**9.4.1.** In the Oberwolfach announcement [S275 (2005)], Serre gives a criterion for the existence of a self-dual normal base for a finite Galois extension $L/K$ of a field $K$ of characteristic 2. He proves that such a base exists if and only if the Galois group of $L/K$ is generated by squares and by elements of order 2. Note that the criterion does not depend on $K$, nor on the extension $L$, but only on the structure of $G$. The proof uses unpublished results of his own on the cohomology of unitary groups in characteristic 2.

**9.5. Essential Dimension.** Let $G$ be a simple algebraic group of adjoint type defined over a field $k$. The essential dimension of $G$ is, by definition, the essential dimension of the functor of $G$-torsors $F \to H^1(F,G)$, which is defined over the category of field extensions $F$ of $k$ (in loose terms, it is the minimal number of "parameters" one needs in order to write a generic $G$-torsor). Here $H^1(F, G)$ denotes the non-abelian Galois cohomology set of $G$. In [S277 (2006)], Serre and V. Chernousov give a lower bound for the essential dimension at a prime $p = 2$, $\mathrm{ed}(G, 2)$, and for the essential dimension $\mathrm{ed}(G)$ of $G$. It is proved in the paper that $\mathrm{ed}(G, 2) \geqslant r + 1$ and, thus, $\mathrm{ed}(G) \geqslant r + 1$, with $r = \mathrm{rank} G$. Lower bounds for $\mathrm{ed}(G, p)$ had been obtained earlier by Z. Reichstein and B. Youssin (2000). In their proof, these authors made use of resolution of singularities, so that their results were only valid for fields $k$ of characteristic zero (however a recent paper of P. Gille and Z. Reichstein has removed this restriction). The proof of Chernousov–Serre for $p = 2$ is valid in every characteristic different from 2. It makes use of the existence of suitable orthogonal representations of $G$ attached to quadratic forms. The quadratic forms involved turn out to be the normalized Killing form. (We should point out that the bound they obtain has now been superseded—especially for the Spin groups.)

## 10 Discrete Subgroups

The study of discrete subgroups of Lie groups goes back to F. Klein and H. Poincaré.

Let us consider a global field $k$ and a finite set $S$ of places of $k$ containing the set $S_\infty$ of all the archimedean places. Let $O$ be the ring of $S$-integers of $k$ and let us denote by $\mathbf{A}_k$ and $\mathbf{A}_k^S$ the ring of adeles and of $S$-adeles of $k$, respectively. We write $\mathbf{A}_k^f$ for the ring of finite adeles of $k$, obtained by taking $S = S_\infty$. Given a linear algebraic group $G$ defined over $k$, we shall consider a fixed faithful representation $G \to \mathbf{GL}_n$. Let $\Gamma := G(k) \cap \mathbf{GL}_n(O)$.

In $G(k)$ we may distinguish two types of subgroup, namely, the $S$-arithmetic subgroups and the $S$-congruence subgroups. A subgroup $\Gamma'$ of $G(k)$ is said to be $S$-arithmetic if $\Gamma \cap \Gamma'$ is of finite index in both $\Gamma$ and $\Gamma'$.

Let $\mathfrak{q} \subset O$ be an ideal and $\mathrm{GL}_n(O, \mathfrak{q}) := \ker(\mathbf{GL}_n(O) \to \mathbf{GL}_n(O/\mathfrak{q}))$. We define $\Gamma_\mathfrak{q} = \Gamma \cap \mathrm{GL}_n(O, \mathfrak{q})$. A subgroup $\Gamma''$ of $G(k)$ is said to be an $S$-congruence subgroup if it is $S$-arithmetic and it contains a subgroup $\Gamma_\mathfrak{q}$, for some non-zero ideal $\mathfrak{q}$. The "$S$-congruence subgroup problem" is the question: is every $S$-arithmetic subgroup an $S$-congruence subgroup? If $S = S_\infty$, one just refers to the "congruence subgroup problem".

Since an $S$-congruence subgroup is $S$-arithmetic, there is a homomorphism of topological groups $\pi : \widehat{G(k)} \to \overline{G(k)}$, where $\widehat{G(k)}$ denotes the completion of $G(k)$ in the topology defined by the $S$-congruence subgroups and $\overline{G(k)}$, the completion in that of the $S$-arithmetic subgroups. The group $\overline{G(k)}$ can be identified with the closure of $G(k)$ in $G(\mathbf{A}_k^f)$. Let $C^S(G)$ denote the kernel of $\pi$. The group $C^S(G)$, which coincides with the $\ker(\pi)$ restricted to $\widehat{\Gamma}$, is profinite and $\pi$ is an epimorphism.

The $S$-congruence subgroup problem has a positive answer if and only if the "congruence kernel" $C^S(G)$ is trivial, i.e. $\pi$ is an isomorphism. It is so when $G$ is a torus (Chevalley, 1951), or is unipotent. When $G$ is semisimple and not simply connected, the problem has a negative answer. Hence the most interesting case is when $G$ is semisimple and simply connected.

**10.1. Congruence Subgroups.** Recall that a semisimple group over $k$ is said to be split (or to be a "Chevalley group") if it has a maximal torus which splits over $k$.

Split groups provide a suitable framework for the study of the congruence subgroup problem. In [S126, Œ 74(1967), Œ 103(1975)], H. Bass, J. Milnor and Serre formulate the $S$-congruence subgroups conjecture precisely in the following form: if $G$ is split, of rank $\geqslant 2$, simply connected and quasi simple, then the group extension $1 \to C^S(G(k)) \to \widehat{G(k)} \to \overline{G(k)} \to 1$ is central and, moreover, $C^S(G)$ is trivial unless $k$ is totally imaginary; in the latter case $C^S(G) = \mu(k)$ is the finite subgroup consisting of the roots of unity of $k^*$.

**10.1.1.** Previously, Bass, Lazard and Serre [S108, Œ 61(1964)] had proved the congruence subgroup conjecture for $\mathbf{SL}_n(\mathbf{Z})$, $n \geqslant 3$, and $\mathbf{Sp}_{2n}(\mathbf{Z})$ for $n \geqslant 2$: every arithmetic subgroup is a congruence subgroup. The same result had been obtained independently by J. Mennicke. Bass–Lazard–Serre's proof is by induction on $n \geqslant 3$. It relies on a computation of the cohomology of the profinite groups $\mathbf{SL}_2(\mathbf{Z}_p)$, $\mathbf{Sp}_{2n}(\mathbf{Z}_p)$ with coefficients in $\mathbf{Q}/\mathbf{Z}$ and in $\mathbf{Q}_p/\mathbf{Z}_p$; this computation is made possible by Lazard's results (see above) on the cohomology of $p$-adic Lie groups.

**10.1.2.** In [S126, Œ 74(1967), Œ 103(1975)], Bass, Milnor and Serre prove the congruence subgroup conjecture when $k$ is an algebraic number field, $G = \mathbf{SL}_n$ for $n \geqslant 3$, and $G = \mathbf{Sp}_{2n}$ for $n \geqslant 2$. In order to do this, they determine the corresponding universal Mennicke symbols associated to these groups and to the ring of integers of a totally imaginary algebraic number field $k$.

**10.1.3.** The solution of the congruence subgroup problem in the case where $G$ is a split simply connected simple group of rank $> 1$ was obtained by H. Matsumoto (1966, 1969), by using the known cases $\mathbf{SL}_3$ and $\mathbf{Sp}_4$. The congruence subgroup problem, as well as its connection to Moore's theory of universal coverings of $G(\mathbf{A}_k^f)$, is discussed in Séminaire Bourbaki [S123, Œ 77(1967)].

**10.1.4.** The paper [S143, Œ 86(1970)] is about the $S$-congruence subgroup problem for $\mathbf{SL}_2$. If $\#S \geqslant 2$, the answer is almost positive: the congruence kernel $C^S$ is a finite cyclic group whose order is at most equal to the number of roots of unity in $k$; if $k$ is not totally imaginary one has $C^S = 1$: the problem has a positive answer. If $\#S = 1$, the problem has a quite negative answer: $C^S$ is an infinite group. The proof is very interesting. In the case $\#S = 1$, it uses number theory, while in the case $\#S > 1$ it uses topology. We shall now provide some of the details of this proof.

In the case $\#S \geqslant 2$, Serre shows that $C^S$ is contained in the centre of $\widehat{G(k)}$ and then makes use of a theory of C. Moore in order to determine it, and in particular to show that it is finite and cyclic. The finiteness of $C^S$ has some important consequences. For instance, given an $S$-arithmetic subgroup $N \subset \mathbf{SL}_2(O)$, a field

$K$ of characteristic zero and a linear representation $\rho : N \to G(K)$, there exists a subgroup $N_1 \subset N$, of finite index, such that the restriction of $\rho$ to $N_1$ is algebraic. This implies that $\rho$ is semisimple. Moreover, for every $k[N]$-module $V$ of finite rank over $K$, we have $H^1(N, V) = 0$. In particular, when taking for $V$ the adjoint representation, one sees that $N$ is rigid.

If $\#S = 1$, Serre shows that, for most $S$-arithmetic subgroups $N$, the group $N^{ab}$ is infinite (this is enough to show that the $S$-congruence problem has a negative answer). There are three cases: $\mathrm{char}(k) = p > 0$; $k = \mathbf{Q}$; and $k$ an imaginary quadratic field. In each case, there is a contractible "symmetric space" $X$ on which $N$ acts properly, and a study of $X/N$ shows that $N^{ab}$ is infinite (with a few exceptions).

In the case of characteristic $p > 0$, $X$ is the Bruhat–Tits tree. In the case $k = \mathbf{Q}$, $X$ is Poincaré's half-plane. In the case where $k$ is an imaginary quadratic field, $X$ is the hyperbolic 3-space, and the quotient manifold $X/N$ can be compactified by adding to it a finite set of 2-tori (which correspond to elliptic curves with complex multiplication by $k$); this is a special case of the general compactifications introduced a few years later by Borel and Serre, see Sect. 10.2.2.

**10.2. Cohomology of Arithmetic Groups.** A locally algebraic group $A$ over a perfect field $k$ is called by Borel–Serre a $k$-group of type (ALA) if it is an extension of an arithmetic $\mathrm{Gal}(\overline{k}/k)$-group by a linear algebraic group over $k$. In a joint paper with Borel [S106 (1964)], included in [A. Borel. *Œuvres*, *Collected Papers*], it is proved that for $k$ a number field and $S$ a finite set of places of $k$, the mapping $H^1(k, A) \to \prod_{v \notin S} H^1(k_v, A)$ is proper, i.e., the inverse image of any point is finite. This result, applied to $A = \mathrm{Aut}(G)$, implies the finiteness of the number of classes of $k$-torsors of a linear group $G$ which are isomorphic locally everywhere to a given $k$-torsor.

**10.2.1.** Let $k$ be a global field and $S$ a finite set of places of $k$. Let $L$ be a linear, reductive, algebraic group defined over $k$. In [S139, Œ 83(1969)], [S148 (1971)], and [S149, Œ 88(1971)], Serre undertakes the study of the cohomology of the $S$-arithmetic subgroups $\Gamma$ which are contained in $L(k)$. For this purpose, the group $\Gamma$ is viewed as a discrete subgroup of a finite product $G = \prod G_\alpha$ of (real or ultrametric) Lie groups. The main tool is provided by the Bruhat–Tits buildings associated to the $v$-adic Lie groups $L(k_v)$, for $v \in S \backslash S_\infty$. The most important contributions include bounds for the cohomological dimension $\mathrm{cd}(\Gamma)$, finiteness properties, and several results relating to the Euler–Poincaré characteristic $\chi(\Gamma)$ and its relations with the values of zeta functions at negative integers (generalizing the well-known formula $\chi(\mathbf{SL}_2(\mathbf{Z})) = \zeta(-1) = -1/12$).

**10.2.2.** Borel and Serre [S144, Œ 90(1970)] prove that if $G$ is a connected, reductive, linear algebraic group defined over $\mathbf{Q}$, which does not have non-trivial characters, it is possible to associate to $G$ a contractible manifold with corners $\overline{X}$, whose interior, $X$, is a homogeneous space of $G(\mathbf{R})$ isomorphic to a quotient $G(\mathbf{R})/K$ for a maximal compact subgroup $K$ of $G(\mathbf{R})$. Its boundary $\partial\overline{X}$ has the same homotopy type as the Tits building $X$ of $G$ (i.e. the simplicial complex whose faces correspond to the $k$-parabolic subgroups of $G$). An arithmetic subgroup $\Gamma \subset G(\mathbf{Q})$ acts properly on $\overline{X}$ and the quotient $\overline{X}/\Gamma$ is compact; this gives a compactification of $X/\Gamma$ which is often called the "Borel–Serre compactification". If $\Gamma$ is torsion-free, then the cohomology of $\Gamma$ is isomorphic to that of $\overline{X}/\Gamma$ and certain duality relations are fulfilled. In particular, the cohomological dimension of $\Gamma$ is given by $\mathrm{cd}(\Gamma) = \dim(X) - \mathrm{rg}_{\mathbf{Q}}(G)$. If $G = \mathbf{SL}_n$, the space $\overline{X}$ is, essentially, a space already defined by C.L. Siegel; it is obtained by attaching boundary points and ideal points to $X$ by means of the reduction theory of quadratic forms.

**10.2.3.** Borel and Serre investigated in [S152, Œ 91(1971)] the cohomology of $S$-arithmetic groups. Let $G$ be a semisimple algebraic group over an algebraic number field $k$ and let $\Gamma \subset G(k)$ be an $S$-arithmetic subgroup. The group $\Gamma$ is a discrete subgroup of $\prod_{v \in S} G(k_v)$. Let $X_S$ be the space defined by:

$$X_S = \overline{X}_\infty \times \prod_{v \in S \setminus S_\infty} X_v.$$

Here $X_v$ is the Bruhat–Tits building of $G$ over $k_v$ and $\overline{X}_\infty$ is the variety with corners associated to the algebraic group $\mathrm{Res}_{k/\mathbf{Q}}(G)$, obtained by restriction of scalars, see above. The group $G(k)$ and, *a fortiori*, the group $\Gamma$, acts on $X_S$. Moreover $\Gamma$ acts properly and the quotient $X_S/\Gamma$ is compact. The study of the cohomology of $\Gamma$ is thus reduced to that of $X_S/\Gamma$. In order to go further, Borel and Serre need some information on the cohomology with compact support of each $X_v$; they obtain it by compactifying $X_v$, the boundary being the Tits building of $G(k_v)$, endowed with a suitable topology. Their main results may be summarized as follows:

> Let $d = \dim(X_\infty)$, $m = \dim(X_S) - \ell = d - \ell + \sum_{v \in S \setminus S_\infty} \ell_v$, where $\ell, \ell_v$ denote the rank of $G$ over $k$ and $k_v$, respectively. Assume that $\Gamma$ is torsion-free. Then
>
> $$H^q(\Gamma, M) \simeq H_{m-q}(\Gamma, I_S \otimes M),$$
>
> for every $\Gamma$-module $M$ and for every integer $q$, the dualizing module $I_S = H_c^m(X_S, \mathbf{Z})$ being free over $\mathbf{Z}$. Moreover, $\mathrm{cd}(\Gamma) = m$ and the group $H^q(\Gamma, \mathbf{Z}(\Gamma))$ is equal to 0 if $q \neq m$ and is equal to $I_S$ if $q = m$.

The proofs can be found in the two papers by Borel–Serre [S159 (1973)], [S172 (1976)], which are reproduced in [A. Borel. *Œuvres*, *Collected Papers*, no. 98 and no. 105]; see also the survey [S190, Œ 120(1979)].

## 11 Arithmetic of Algebraic Varieties

**11.1. Modular Curves.** The three publications [S142 (1970)], [S170 (1975)], and [S183 (1977)] correspond to lectures delivered by Serre in the Séminaire Bourbaki. They were very popular in the seventies as introductory texts for the study of the arithmetic of modular curves.

**11.1.1.** The first lecture [S142 (1970)] deals with a theorem of Y. Manin (1969) according to which, given a number field $K$, an elliptic curve $E$ defined over $K$, and a prime number $p$, the order of the $p$-component of the torsion group $E_{\text{tor}}(K)$ is bounded by an integer depending only on $K$ and $p$. The proof relies on a previous result by V.A. Demjanenko and Manin on the finiteness of the number of rational points of certain algebraic curves; this is applied afterwards to the modular curve $X_0(N)$, where $N = N(p, K)$.

**11.1.2.** The second lecture [S170 (1975)] was written jointly with B. Mazur. Its purpose is to present results of A. Ogg on the cuspidal group of the Jacobian of the modular curve $X_0(N)$ and some of the results of Mazur on the rational points of this curve. In it, we find the modular interpretation of the modular curve, the definition of the Hecke operators as correspondences acting on it, the study of the Eisenstein ideal, a study of the Néron model of the Jacobian of the modular curve, a study of the regular model of the modular curve, and so on.

**11.1.3.** The third lecture [S183 (1977)] explains the results of Mazur on the Eisenstein ideal and the rational points of modular curves (1978), and on the rational isogenies of prime degree (1978). The main theorem is that, if $N$ is a prime not belonging to the set $\{2, 3, 5, 7, 11, 13, 17, 19, 37, 43, 67, 163\}$, then the modular curve $X_0(N)$ has no rational points other than the cusps. As an application, one obtains the possible structures for the rational torsion groups $E_{tor}(\mathbf{Q})$ of the elliptic curves defined over $\mathbf{Q}$. In many aspects, the above work paved the way for G. Faltings' proof of the Mordell–Weil theorem (1983).

**11.1.4.** The paper [S237 Œ 159(1993)], written with T. Ekedahl, gives a long list of curves of high genus whose Jacobian is completely decomposable, i.e., isogenous to a product of elliptic curves. They ask whether it is true that, for every genus $g > 0$, there exists a curve of genus $g$ whose Jacobian is completely decomposed or, on the contrary, whether the genus of a curve whose Jacobian is completely decomposable is bounded. (The second question has a negative answer in characteristic $p > 0$.) The examples are constructed either by means of modular curves or as coverings of curves of genus 2 or 3. The highest genus obtained is 1297.

**11.1.5.** Let $p$ be a prime number. In the paper [S254 Œ 170(1997)], Serre determines the asymptotic distribution of the eigenvalues of the Hecke operators $T_p$ on spaces of modular forms when the weight or the level varies. More precisely, let $T_p$ denote the Hecke operator associated to $p$ acting on the space $S(N, k)$ of cusp forms of weight $k$ for the congruence group $\Gamma_0(N)$, with $\gcd(N, p) = 1$, and let $T'_p = T_p/p^{(k-1)/2}$. By Deligne's theorem on the Ramanujan–Petersson conjecture, the eigenvalues of the operator $T'_p$ belong to the interval $\Omega = [-2, 2]$. Let us consider sequences of pairs of integers $(N_\lambda, k_\lambda)$ such that $k_\lambda$ is even, $k_\lambda + N_\lambda \to \infty$ as $\lambda \to \infty$, and $p$ does not divide $N_\lambda$. The main theorem proved in the paper states that the family $x_\lambda = x(N_\lambda, k_\lambda)$ of eigenvalues of the $T'_p(N, k)$ is equidistributed in the interval $\Omega$ with respect to a measure $\mu_p$ which is given by an explicit formula, similar (but not identical) to the Sato–Tate measure. In fact, in the paper, we find measures $\mu_q$ which are defined for every $q \geqslant 1$ and have the property that $\lim_{q\to\infty} \mu_q = \mu_\infty$, where $\mu_\infty$ is the Sato–Tate measure. In order to give an interpretation of $\mu_q$, Serre identifies $\Omega$ with a subset of the spectrum of the automorphism group $G = \mathrm{Aut}(A)$ of a regular tree of valency $q + 1$, which is a locally compact group with respect to the topology of simple convergence. Then, $\mu_q$ is the restriction to $\Omega$ of the Plancherel measure of $G$. Several interesting consequences are derived from the equidistribution theorem. For instance, it is shown that the maximum of the dimension of the $\mathbf{Q}$-simple factors of the Jacobian $J_0(N)$ of the modular curve $X_0(N)$ tends to infinity as $N \to \infty$. In particular, there are only finitely many integers $N \geqslant 1$ such that $J_0(N)$ is isogenous over $\mathbf{Q}$ to a product of elliptic curves, as was already stated in [S237, Œ 159(1993)].

**11.2. Varieties Over Finite Fields.** Let $q = p^e$, with $p$ a prime number and $e \geqslant 1$, and let $\mathbf{F}_q$ be a finite field with $q$ elements. The numbers $N_{q^r}$, $r \geqslant 1$, of rational points over $\mathbf{F}_{q^r}$ of non-singular projective varieties defined over $\mathbf{F}_q$ are encapsulated in their zeta function. One of the major achievements of A. Grothendieck and his school was to provide the tools for the proof of the Weil conjectures (1949), relative to the nature of these functions: cohomological interpretation, rationality, functional equation and the so-called Riemann hypothesis. Serre had a profound influence on the process. An account of the landmark paper by P. Deligne (1974) on the proof of the Riemann hypothesis for the zeta function of a non-singular variety defined over a finite field can be found in [S166 (1974)].

The paper [Œ 117(1978)] contains a report by Serre on Deligne's work, upon request by the Fields Medal Committee. Deligne was awarded the Fields Medal in 1978.

A. Grothendieck was awarded the Fields Medal in 1966, together with M. Atiyah, P. Cohen and S. Smale. Serre's original report, written in 1965, on the work of Grothendieck and addressed to the Fields Medal Committee was reproduced much later, in [S220 (1989)].

Among many other applications, Weil conjectures imply good estimations for certain exponential sums, since these sums can be viewed as traces of Frobenius endomorphisms acting on the cohomology of varieties over finite fields. The results of Deligne on this subject were explained by Serre in [S180, Œ 111(1977)].

**11.3. Number of Points of Curves Over Finite Fields.** Let $C$ be an absolutely irreducible non-singular projective curve of genus $g$ defined over $\mathbf{F}_q$. After the proof of the Riemann hypothesis for curves, due to Weil (1940–1948), it was known that the number $N = N(C)$ of the rational points of $C$ over $\mathbf{F}_q$ satisfies the inequality $|N - (q + 1)| \leqslant 2gq^{1/2}$. Several results due to H. Stark (1973), Y. Ihara (1981), and V.G. Drinfeld and S.G. Vladut (1983) showed that Weil's bound can often be improved. On the other hand, it was of interest for coding theory to have curves of low genus with many points.

In the papers [S201, Œ 128(1983)] and [S200, Œ 129(1983)], Serre expands the results of the previous authors and introduces a systematic method to obtain more precise bounds, based on Weil's "explicit formula".

Let $N_q(g)$ be the maximum value of $N(C)$ as $C$ runs through all curves of genus $g$ defined over $\mathbf{F}_q$. The value $N_q(1)$ was already known; for most $q$'s it is equal to $q + 1 + \lfloor 2q^{1/2} \rfloor$; for the others it is $q + \lfloor 2q^{1/2} \rfloor$. Serre obtains the exact value of $N_q(2)$ for every $q$. It is not very different from Weil's bound.

If $A(q) = \limsup_{g\to\infty} N_g(q)/q$, then Drinfeld–Vladut proved that

$$A(q) \leqslant q^{1/2} - 1$$

for every $q$, and Ihara showed that $A(q) = q^{1/2} - 1$ if $q$ is a square. Serre proves that $A(q) > 0$ for all $q$, more precisely $A(q) \geqslant c \cdot \log(q)$ for some absolute $c > 0$. His proof uses class field towers, like in Golod–Shafarevich.

These papers generated considerable interest in determining the actual maximum and minimum of the number of points for a given pair $(g, q)$.

**11.3.1.** Kristin Lauter obtained improvements on the bounds for the number of rational points of curves over finite fields, along the lines of [S201, Œ 128(1983)] and [S200, Œ 129(1983)]. In her papers [S265 (2001)], [S267 (2002)], we find appendices written by Serre. The appendix in [S267 (2002)] is particularly appealing. It gives an equivalence between the category of abelian varieties over $\mathbf{F}_q$ which are isogenous over $\mathbf{F}_q$ to a product of copies of an ordinary elliptic curve $E$ defined over $\mathbf{F}_q$ and the category of torsion-free $R_d$-modules of finite type, where $R_d$ denotes the ring of integers in the quadratic field of discriminant $d$, being $\#E(\mathbf{F}_q) = q + 1 - a$, $d = a^2 - 4q$, and under the assumption that $d$ is the discriminant of an imaginary quadratic field. Polarizations on these abelian varieties correspond to positive definite hermitian forms on $R$-modules. Thus, in the cases where there is no indecomposable positive definite hermitian module of discriminant 1, one obtains the non-existence of curves whose Jacobian is of that type. And, conversely, if such a hermitian module exists, one obtains a principally polarized abelian variety; if furthermore its dimension is 2, this abelian variety is a Jacobian and one gets a curve whose number of points is $q + 1 - 2a$; a similar (but less precise) result holds for genus 3: one finds a curve with either $q + 1 - 3a$ or $q + 1 + 3a$ points. Particular results on the classification of these modules in dimensions 2 and 3, due to D.W. Hoffmann (1991), and a procedure for gluing isogenies are used to determine the existence or non-existence of certain polarized abelian varieties, useful in their turn to show that for all finite fields $\mathbf{F}_q$ there exists a genus 3 curve over $\mathbf{F}_q$ such that its number of rational points is within 3 of the Serre–Weil upper or lower bound.

**11.3.2.** In a letter published in [S230, Œ 155(1991)], Serre answered a problem posed by M. Tsfasman at Luminy on the maximal number of points of a hypersurface defined over a finite field. On the hypersurface, no hypothesis of irreducibility or of non-singularity is made. Serre shows that the number $N$ of zeros of a homogeneous polynomial $f = f(X_0, \ldots, X_n)$ in $\mathbf{P}_n(\mathbf{F}_q)$ of degree $d \leqslant q + 1$ is at most $dq^{n-1} + p_{n-2}$, where $p_n = q^n + q^{n-1} + \cdots + 1$ is the number of points in the projective space $\mathbf{P}_n(\mathbf{F}_q)$. The result has been widely used in coding theory.

**11.4. Diophantine Problems.** Certain classical methods of transcendence based on the study of the solutions of differential equations, mainly due to Th. Schneider, were transposed to the $p$-adic setting by S. Lang (1962, 1965). In the Séminaire Delange–Pisot–Poitou [S124 (1967)], Serre studies the dependence of $p$-adic exponentials avoiding the use of differential equations.

**11.4.1.** The book *Lectures on the Mordell–Weil Theorem* [S203, MW(1984)] arose from the notes taken by M. Waldschmidt of a course taught by Serre (1980–1981), which were translated and revised with the help of M. Brown.

The content of the lectures was: heights, Néron–Tate heights on abelian varieties, the Mordell–Weil theorem on the finiteness generation of the rational points of any abelian variety defined over a number field, Belyi theorem characterizing the non-singular projective complex curves definable over $\overline{\mathbf{Q}}$, Chabauty and Manin–Demjanenko theorems on the Mordell conjecture (previous to Faltings' theorem (1983)), Siegel's theorem on the integral points on affine curves, Baker's effective forms of Siegel's theorem, Hilbert's irreducibility theorem and its applications to the inverse Galois problem, construction of elliptic curves of large rank, sieve methods, Davenport–Halberstam's theorem, asymptotic formulas for the number of integral points on affine varieties defined over number fields, and the solution to the class number 1 problem by using integral points on modular curves.

**11.4.2.** The paper [S189, Œ 122(1979)] is an appendix to a text by M. Waldschmidt on transcendental numbers (1970). It contains several useful properties of connected commutative algebraic groups, defined over a field $k$ of characteristic zero. They concern the following: the existence of smooth projective compactifications; quadratic growth, at most, of the height function (when $k$ is an algebraic extension of $\mathbf{Q}$), and uniformization by entire functions of order $\leqslant 2$ when $k = \mathbf{C}$.

**11.4.3.** The publication [S194 (1980)] reproduces two letters addressed to D. Masser. The questions concern some of Masser's results on the linear independence of periods and pseudo-periods of elliptic functions (1977). In the first letter, Serre studies independence properties of the fields of $\ell$-division points of elliptic curves defined over an algebraic number field and with complex multiplication by different quadratic imaginary number fields. In the second letter, Serre proves that, under some reasonable assumptions, the degree of the field generated by the $\ell$-division points of the product of such elliptic curves is as large as possible for almost all the primes $\ell$.

## 12 Field Theory

The paper [S195, Œ 123(1980)] reproduces a letter of Serre answering a question raised by J.D. Gray about Klein's lectures on the icosahedron. The icosahedral group $G = A_5$ acts on a curve $X$ of genus zero; by extending the field of definition $k$ of $X$ and $G$ to an algebraic closure, one obtains an embedding of $G$ in the projective linear group $\mathbf{PGL}_2$. Moreover, the field $k$ must contain $\sqrt{5}$. The quotient $X/G$ is isomorphic to $\mathbf{P}_1$. If $z$ is a $k$-point of $X/G$, its lifting to $X$ generates a Galois extension $k'$ of $k$ whose Galois group is a subgroup of $G$. Serre explains that the main question posed by Hermite and Klein turns out to be whether one obtains all Galois extensions of $k$ with Galois group $G$ in this way. He then shows that the answer to this question is "almost" yes. Suppose that $k'/k$ is a Galois extension with Galois group $G$. Serre uses a descent method and works with twisted curves $X_{k'}$. The curves $X_{k'}$ are controlled by a quaternion algebra $H_{k'}$. To go from $X_{k'}$ to $H_{k'}$, Serre follows two procedures: either using the non-trivial element of $H^2(G, \mathbf{Z}/2\mathbf{Z})$, which corresponds to the binary icosahedral group, or considering the trace form $\mathrm{Tr}(z^2)$ in a quintic extension $k_1/k$ defining $k'/k$. Then he shows that $k'/k$ comes from a covering $X \to X/G$ if and only if $X_{k'}$ has a rational point over $k$, if and only if the class of $H_k'$ in the Brauer group of $k$ equals the sum of $(-1, -1)$ and the Witt invariant of the quadratic form $\mathrm{Tr}(z^2)$, on the subspace of $k_1$ of elements of trace zero. Moreover, these conditions are equivalent to the fact that $k'$ can be generated by the roots of a quintic equation of the form $X^5 + aX^2 + bX + c = 0$, which is consistent with old results of Hermite and Klein.

**12.0.1.** The obstruction associated to a Galois embedding problem, defined by a Galois extension $L/K$ and by a central extension of the group $\mathrm{Gal}(L/K)$, is given by a cohomology class, the vanishing of which characterizes the solvability of the problem. When the kernel of the central extension is the cyclic group $C_2$ of order 2, the cohomology class can be identified with an element of $\mathrm{Br}_2(K) \simeq H^2(K, C_2)$ (assuming that the characteristic is not 2). In a paper dedicated to J.C. Moore [S204, Œ 131(1984)], Serre gives a formula relating the obstruction to certain Galois embedding problems to the second Stiefel–Whitney class of the trace form $\mathrm{Tr}(x^2)$. Through the use of Serre's formula, N. Vila (1984, 1985) proved that the non-trivial double covering $\widetilde{A}_n = 2 \cdot A_n$ of the alternating group $A_n$ is the Galois group of a regular extension of $\mathbf{Q}(T)$, for infinitely many values of $n \geqslant 4$. The result was extended to all $n \geqslant 4$ by J.-F. Mestre (1990) cf. [S235, TGT(1992)]. Explicit solutions to solvable embedding problems of this type were later obtained by T. Crespo.

**12.0.2.** In his report on Galois groups over $\mathbf{Q}$ presented to the Séminaire Bourbaki [S217, Œ 147(1988)], Serre provides a summary of the status of the inverse Galois problem; that is, of the question of whether all finite groups are Galois groups of an equation with rational coefficients. He mentions the solution of the problem in the solvable case due to I. Shafarevich (1954) and its improvements by J. Neukirch (1979). He gives Hilbert's realizations of the symmetric and alternating groups as Galois groups over $\mathbf{Q}$ by means of Hilbert's irreducibility theorem. And, in the most detailed part of the exposition, he explains the rigidity methods of H. Matzat (1980) and J.G. Thompson (1984), and presents a list of the simple groups known at that moment to be Galois over $\mathbf{Q}$. As another type of example, he considers the realization of certain central extensions of simple groups, which had recently been obtained thanks to his $\mathrm{Tr}(x^2)$ formula [S204, Œ 131(1984)].

**12.0.3.** The papers [S226, Œ 151(1990)] and [S227, Œ 152(1990)] are related to the results of Mestre mentioned above. The first one is about lifting elements of odd order from $A_n$ to $\widetilde{A}_n$; if one has several elements and their product is equal to 1, what is the product of their liftings: 1 or $-1$? In the second paper (which may be viewed as a geometrization of the first one), Serre considers a ramified covering $\pi : Y \to X$ of curves, in which all the ramification indices are odd. He gives a formula relating several cohomological invariants of this covering; here the behaviour of the theta characteristics of $X$ under $\pi^*$ plays an essential role. He also asks whether there is a general formula including those in [S204, Œ 131(1984)] and [S227, Œ 152(1990)]. This was done later by H. Esnault, B. Kahn and E. Viehweg (1993).

**12.0.4.** The course given by Serre [S236, Œ 156(1992)] focuses on the Galois cohomology of pure transcendental extensions. Suppose that $K$ is a field endowed with a discrete valuation, $v$, of residue field $k$. Let $C$ denote a discrete $\mathrm{Gal}(K_s/K)$-module, unramified at $v$ and such that $nC = 0$ for some integer $n > 0$, coprime to the characteristic of $K$. Given a cohomology class $\alpha \in H^i(K, C)$, one defines the notions of a residue of $\alpha$ at $v$, a pole of $\alpha$ at $v$, and a value $\alpha(v)$. When $K = k(X)$ is the function field of a smooth, connected projective curve defined over $k$, there is a residue formula and an analogue of Abel's theorem. The theory is applied to the solution of specialization problems of the Brauer group of $K$ in the Brauer group of $k$. If $x \in X(K)$, and $\alpha \in \mathrm{Br}_n(K)$, then $\alpha(x) \in \mathrm{Br}_n(k)$, whenever $x$ is not a pole of $\alpha$. Serre looks at the function $\alpha(x)$ and, in particular, at its vanishing set $V(\alpha)$. In [S225, Œ 150(1990)], he deals with the case $K = \mathbf{Q}(T_1, \dots, T_r)$, for $n = 2$. The results are completed with asymptotic estimations on the number of zeros of $\alpha$ obtained by sieving arguments; they depend on the number of $\mathbf{Q}$-irreducible components of the polar divisor of $\alpha$. One of the questions raised in this paper ("how often does a conic have a rational point?") was solved later by C. Hooley (1993) and C.R. Guo (1995): the upper bound given by the sieve method has the right order of magnitude.

**12.0.5.** *Cohomological Invariants in Galois Cohomology* [S269, CI(2003)] is a book co-authored by S. Garibaldi, A. Merkurjev and J.-P. Serre. The algebraic invariants discussed in it are the Galois cohomology analogues of the characteristic classes of topology, but here the topological spaces are replaced by schemes $\mathrm{Spec}(k)$, for $k$ a field.

The text is divided in two parts. The first one consists of an expanded version of a series of lectures given by Serre at UCLA in 2001, with notes by S. Garibaldi. The second part is due to Merkurjev with a section by Garibaldi; we shall not discuss it here. In Chap. I, Serre defines a quite general notion of invariant to be applied throughout the book to several apparently disparate situations. Given a ground field $k_0$ and two functors

$$A : \mathrm{Fields}_{/k_0} \to \mathrm{Sets} \quad \text{and} \quad H : \mathrm{Fields}_{/k_0} \to \text{Abelian Groups},$$

an $H$-invariant of $A$ is defined as a morphism of functors

$$a : A \to H.$$

Here $\mathrm{Fields}_{/k_0}$ denotes the category of field extensions $k$ of $k_0$. Examples of functors $A$ are:

- $k \mapsto \mathrm{Et}_n(k)$, the isomorphism classes of étale $k$-algebras of rank $n$;
- $k \mapsto \mathrm{Quad}_n(k)$, the isomorphism classes of non-degenerate quadratic forms over $k$ of rank $n$;
- $k \mapsto \mathrm{Pfister}_n(k)$, the isomorphism classes of $n$-Pfister forms over $k$ of rank equal to $n$.

Examples of functors $H$ are provided by the abelian Galois cohomology groups $H^i(k, C)$ and their direct sum, $H(k, C)$, where $C$ denotes a discrete $\mathrm{Gal}(k_0^s/k_0)$-module; or by the functor $H(k) = W(k)$, where $W(k)$ stands for the Witt ring of non-degenerate quadratic forms on $k$.

The aim of the lectures is to determine the group of invariants $\mathrm{Inv}(A, H)$. In the background material of the book concerning Galois cohomology, we find the notion of the residue of a cohomology class of $H^i(K, C)$ at a discrete valuation $v$ of a field $K$, as well as the value at $v$ for those cohomology classes with residue equal to zero. Basic properties of restriction and corestriction of invariants are obtained, in perfect analogy with the case of group cohomology. An important tool is the notion of versal torsor, which plays an analogous role to that of the universal bundle in topology: an invariant is completely determined by its value on a versal $G$-torsor. These techniques allow the determination of the mod 2 invariants for quadratic forms, hermitian forms, rank $n$ étale algebras, octonions or Albert algebras, when $\mathrm{char}(k_0) \neq 2$. In particular, the mod 2 invariants of rank $n$ étale algebras make up a free $H(k_0)$-module whose basis consists of the Stiefel–Whitney classes $w_i$, for $0 \leqslant i \leqslant [n/2]$. This gives a new proof of Serre's earlier formula on this subject [S204, Œ 131(1984)], as well as its generalization by B. Kahn (1984). Similarly, the $W(k_0)$-module $\mathrm{Inv}(S_n, W)$ is free of finite rank, with basis given by the Witt classes of the first $[n/2]$ exterior powers of the trace form. Among other results, one finds an explicit description of all possible trace forms of rank $\leqslant 7$ and an application of trace forms to the study of Noether's problem, which we recall in what follows.

Given a finite group $G$, the property $\mathrm{Noe}(G/k_0)$ means that there exists an embedding $\rho : G \to \mathbf{GL}_n(k_0)$ such that, if $K_\rho$ is the subfield of $k_0(X_1, \dots, X_n)$ fixed by $G$, then $K_\rho$ is a pure transcendental extension of $k_0$. Deciding whether $\mathrm{Noe}(G/k_0)$ is true is the Noether problem for $G$ and $k_0$. Serre proves that Noether's problem has a negative answer for $\mathbf{SL}(2, \mathbf{F}_7)$, $2 \cdot A_6$ or the quaternion group $Q_{16}$ of order 16. The book includes also several letters; one of these is a letter from Serre to R.S. Garibaldi, dated in 2002, in which he explains his motivations.

**12.0.6.** Let $k$ be a field of characteristic different from 2. The norm form of any quaternion algebra defined over $k$ is a 2-fold Pfister form. In [S278 (2006)], M. Rost, J.-P. Serre and J.-P. Tignol study the trace form $q_A(x) = \mathrm{Trd}_A(x^2)$ of a central simple algebra $A$ of degree 4 over $k$, under the assumption that $k$ contains a primitive 4th root of unity. They prove that $q_A = q_2 + q_4$, in the Witt ring of quadratic forms over $k$, were $q_2$ and $q_4$ are uniquely determined 2-fold and 4-fold Pfister forms, respectively. The form $q_2$ corresponds to the norm form of the quaternion algebra which is equivalent to $A \otimes_k A$ in the Brauer group of $k$. Moreover, $A$ is cyclic if and only if $q_4$ is hyperbolic. The images of the forms $q_j$ in $H^j(k, \mathbf{Z}/2\mathbf{Z})$ yield cohomological invariants of $\mathbf{PGL}_4$, since the set $H^1(k, \mathbf{PGL}_4)$ classifies the central simple $k$-algebras of degree 4.

## 13 Galois Representations

Serre has studied Galois representations (especially $\ell$-adic representations) in several books and papers. His pioneering contributions to these topics have broken new ground and have profoundly influenced their research in the last decades.

**13.1. Hodge–Tate Modules.** The paper [S129, Œ 72(1967)] is based on a lecture delivered by Serre at a Conference on Local Fields held in Driebergen (The Netherlands). Take as the ground field a local field of characteristic zero whose residue field is of characteristic $p > 0$, and let $\mathbf{C}_p$ be the completion of an algebraic closure $\overline{K}$ of $K$. If $T$ is the Tate module associated to a $p$-divisible group, defined over the ring of integers of $K$, a deep result of Tate states that $\mathbf{C}_p \otimes T$ has a decomposition analogous to the Hodge decomposition for complex cohomology. This gives strong restrictions on the image $G$ of $\mathrm{Gal}(\overline{K}/K)$ in $\mathrm{Aut}(T)$. For instance, if the action of $G$ is semisimple, then the Zariski closure of $G$ contains a $p$-adic Mumford–Tate group. Under some additional hypotheses, Serre shows that the group $G$ is open in $\mathrm{Aut}(T)$. This applies in particular to formal groups of dimension 1, without formal complex multiplication.

**13.1.1.** The topic of Hodge–Tate decompositions was also considered in the book [S133, McGill(1968)], which we will discuss in a moment.

**13.1.2.** In [S191, Œ 119(1979)], it is shown that the inertia subgroup of a Galois group acting on a Hodge–Tate module $V$ over a local field is almost algebraic, in the sense that it is open in a certain algebraic subgroup $H_V$ of the general linear group $\mathbf{GL}_V$. In two important cases, Serre determines the structure of the connected component $H_V^0$ of $H_V$. In the commutative case, $H_V^0$ is a torus. If the weights of $V$ are reduced to 0 and 1, the simple factors of $H_V^0$ are of classical type: $A_n, B_n, C_n, D_n$.

**13.2. Elliptic Curves and $\ell$-adic Representations.** Over the years, Serre has given several courses on elliptic curves. Three of these were at the Collège de France

[S115, Œ 67(1965)], [S122, Œ 71(1966)], [S153, Œ 93(1971)] and one at McGill University (Montreal), in 1967. Abundant material was presented in these lectures. Most of it was published soon after in the papers [S118, Œ 70(1966)], [S154, Œ 94(1972)] and in the book *Abelian $\ell$-adic representations and elliptic curves* [S133, McGill(1968)].

The course [S115, Œ 67(1965)] covered general properties of elliptic curves, theorems on the structure of their endomorphism ring, reduction of elliptic curves, Tate modules, complex multiplication, and so on. Let $E$ be an elliptic curve defined over a field $k$ and let $\ell$ be a prime different from char$(k)$. The Tate modules $T_\ell(E) = \varprojlim E[\ell^n](k_s)$ are special cases of the $\ell$-adic homology groups associated to algebraic varieties. The Galois group $\mathrm{Gal}(k_s/k)$ acts on $T_\ell(E)$ and on the $\mathbf{Q}_\ell$-vector space $V_\ell(E) = \mathbf{Q}_\ell \otimes T_\ell(E)$. We may consider the associated Galois $\ell$-adic representation $\rho_\ell : \mathrm{Gal}(k_s/k) \to \mathrm{Aut}(T_\ell) \simeq \mathbf{GL}_2(\mathbf{Z}_\ell)$. The image $G_\ell$ of $\rho_\ell$ is an $\ell$-adic Lie subgroup of $\mathrm{Aut}(T_\ell)$. We shall denote by $\mathfrak{g}_\ell$ the Lie algebra of $G_\ell$. The Galois extension associated to $G_\ell$ is obtained by adding the coordinates of the points of $E(\overline{k})$ of order a power of $\ell$ to the field $k$.

Suppose that $k$ is an algebraic number field and that the elliptic curve $E$ has complex multiplication. Thus, there exist an imaginary quadratic field $F$ and a ring homomorphism $F \to \mathbf{Q} \otimes \mathrm{End}_k(E)$. Then the Galois group $G_\ell$ is abelian whenever $F \subset k$, and is non-abelian otherwise. If $F \subset k$, the action of $\mathrm{Gal}(\overline{k}/k)$ on $T_\ell(E)$ is given by a Grössencharakter whose conductor has its support in the set of places of bad reduction of $E$; this result is due to M. Deuring. The usefulness of elliptic curves with complex multiplication consists in the fact that they provide an explicit class field theory for imaginary quadratic fields.

**13.2.1.** A short account of the classical theory of complex multiplication can be found in [S128, Œ 76(1967)].

**13.2.2.** By using fiber spaces whose fibers are products of elliptic curves with complex multiplication, Serre [S107, Œ 63(1964)] constructed examples of non-singular projective varieties defined over an algebraic number field $K$ which are Galois conjugate but have non-isomorphic fundamental groups. In particular, although they have the same Betti numbers, they are not homeomorphic.

**13.2.3.** The paper [S118, Œ 70(1966)] is about the $\ell$-adic Lie groups and the $\ell$-adic Lie algebras associated to elliptic curves defined over an algebraic number field $k$ and without complex multiplication. The central result is that $\mathfrak{g}_\ell$ is "as large as possible", namely, it is equal to $\mathrm{End}(V_\ell)$, when the ground field is $\mathbf{Q}$. In the proof, Serre uses a wide range of resources: Lie algebras and $\ell$-adic Lie groups; Hasse–Witt invariants of elliptic curves; pro-algebraic groups; the existence of canonical liftings of ordinary curves in characteristic $p$; the Lie subalgebras of the ramification groups; Chebotarev's density theorem, as well as class field theory, Hodge–Tate theory, and so on. Serre also observes that, if a conjecture of Tate on Galois actions on the Tate modules is true, then the determination of the Lie algebra $\mathfrak{g}_\ell$ can be carried out for any algebraic number field. Tate's conjecture was proved almost two decades later by G. Faltings (1983) in a celebrated paper in which he also proved two more conjectures, one due to L.J. Mordell and the other due to I. Shafarevich. For these results, Faltings was awarded the Fields Medal in 1986.

In the same paper [S118, Œ 70(1966)], Serre shows that the set of places of $k$ at which a curve $E$, without complex multiplication, has a supersingular reduction is of density zero in the set of all the places of $k$. This does not preclude the set of these places being infinite. On the contrary, Serre thought that this could well be the case. Indeed, N.D. Elkies (1987) proved that, for every elliptic curve $E$ defined over a real number field, there exist infinitely many primes of supersingular reduction, in agreement with Serre's opinion. (Note that the case of a totally imaginary ground field remains open.) S. Lang and H. Trotter (1976) conjectured an asymptotic formula (which is still unproved) for the frequency of the supersingular primes in the reduction of an elliptic curve $E$ without complex multiplication and defined over $\mathbf{Q}$.

**13.2.4.** The results of [S118, Œ 70(1966)] were completed in the lecture course [S122, Œ 71(1966)] and in [S133, McGill(1968)].

In Chap. I of [McGill], Serre considers $\ell$-adic representations of the absolute Galois group $\mathrm{Gal}(k_s/k)$ of a field $k$. For $k$ an algebraic number field, he defines the concepts of a rational $\ell$-adic representation, and of a compatible system of rational $\ell$-adic representations (these notions go back to Y. Taniyama (1957)). He relates the equidistribution of conjugacy classes of Frobenius elements to the existence of some analytic properties for the $L$-functions associated to compatible systems of rational $\ell$-adic representations, a typical example being that of the Sato–Tate conjecture.

In Chap. II, Serre associates to every algebraic number field $k$ a projective family $(S_{\mathfrak{m}})$ of commutative algebraic groups defined over $\mathbf{Q}$. (From the point of view of motives, these groups are just the commutative motivic Galois groups.) For each modulus $\mathfrak{m}$ of $k$, he constructs an exact sequence of commutative algebraic groups $1 \to T_{\mathfrak{m}} \to S_{\mathfrak{m}} \to C_{\mathfrak{m}} \to 1$, in which $C_{\mathfrak{m}}$ is a finite group and $T_{\mathfrak{m}}$ is a torus. The characters of $S_{\mathfrak{m}}$ are, essentially, the Grössencharakteren of type $A_0$, in the sense of Weil, of conductor dividing $\mathfrak{m}$. They appear in the theory of complex multiplication.

In Chap. III, the concept of a locally algebraic abelian $\ell$-adic representation is defined. The main result is that such Galois representations come from linear representations, in the algebraic sense, of the family $(S_{\mathfrak{m}})$. When the number field $k$ is obtained by the composition of quadratic fields, it is shown that every semisimple abelian rational $\ell$-adic representation is locally algebraic. The proof is based upon transcendence results of C.L. Siegel and S. Lang. Serre observes that the result should also be true for any algebraic number field; this was proved later by M. Waldschmidt (1986), as a consequence of a stronger transcendence result.

In Chap. IV, the results of the previous chapters are applied to the $\ell$-adic representations associated to elliptic curves. The main theorem is that, if $E$ is an elliptic curve over an algebraic number field $k$, without complex multiplication, then $\mathfrak{g}_\ell = \mathrm{End}(V_\ell)$. The proof turns out to be a clever combination of a finiteness theorem due to Shafarevich together with the above mentioned results on abelian and locally algebraic $\ell$-adic representations of $k$.

Serre also proves that, if $E$ is an elliptic curve over an algebraic number field $k$ such that its $j$-invariant is not an algebraic integer of $k$, then the group $G := \mathrm{Im}\rho$, where $\rho = \prod \rho_\ell$, is open in $\prod \mathbf{GL}_2(\mathbf{Z}_\ell)$. Later, Serre would eliminate the condition regarding the modular invariant $j$ (see below).

It is also proved in [McGill] that, if $E$, $E'$ are elliptic curves defined over an algebraic number field $k$, whose invariants $j(E)$, $j(E')$ are not algebraic integers and whose $\mathrm{Gal}(\overline{k}/k)$-modules $V_\ell(E)$, $V_\ell(E')$ are isomorphic, then $E$ and $E'$ are isogenous over $k$. The result is a special case of the Tate conjecture proved later by G. Faltings (1983).

**13.2.5.** The above results were improved in the seminal paper *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques* [S154, Œ 94(1972)], which is dedicated to André Weil. The main theorem states that, if $E$ is an elliptic curve defined over an algebraic number field $k$, which does not have complex multiplication, then $G_\ell = \mathrm{Aut}(T_\ell(E))$, for almost all $\ell$. In particular, we have $\mathrm{Gal}(k[E_\ell]/k) \simeq \mathbf{GL}_2(\mathbf{F}_\ell)$, for almost all $\ell$. The proof is based upon local results relative to the action of the tame inertia group on the points of finite order of the elliptic curves. This action can be expressed in terms of products of fundamental characters, and the main point is that these exponents have a uniform bound (namely the ramification index of the local field). This boundedness plays a role similar to that of the local algebraicity which had been used in [McGill]. Serre conjectures that similar bounds are valid for higher dimensional cohomology; this has been proved recently, as a by-product of Fontaine's theory.

He also raised several questions concerning the effectiveness of the results. The paper includes many numerical examples in which all the prime numbers for which $\mathrm{Gal}(k[E_\ell]/k) \not\simeq \mathbf{GL}_2(\mathbf{F}_\ell)$ are computed.

In the summary of the course, Serre also mentions that if $A$ is an abelian surface such that $\mathrm{End}(A)$ is an order of a quaternion field $D$ defined over $\mathbf{Q}$ (a so-called "fake elliptic curve") then the group $\rho(\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}))$, where $\rho = \prod \rho_\ell$, is open in $D^*(\mathbf{A}^f)$. This was proved later by M. Ohta (1974), following Serre's guidelines.

**13.3. Modular Forms and $\ell$-adic Representations.** Many arithmetical functions can be recovered from the Fourier coefficients of modular functions or modular forms. In an early contribution at the Séminaire Delange–Pisot–Poitou [S138, Œ 80(1969)], one finds the remarkable conjecture that certain congruences satisfied by the Ramanujan $\tau$ function can be explained by the existence, for each prime $\ell$, of a 2-dimensional $\ell$-adic representation

$$\rho_\ell : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{Aut}(V_\ell),$$

unramified away from $\ell$, and such that $\mathrm{Tr}(\rho_\ell(F_p)) = \tau(p)$, $\det(\rho_\ell(F_p)) = p^{11}$, for each Frobenius element $F_p$, at any prime $p \neq \ell$. Assume this conjecture (which was proved a few months later by Deligne (see below)). The $\ell$-adic representation $\rho_\ell$ leaves a lattice of $V_\ell$ stable, and thus may be viewed as a representation in $\mathbf{GL}_2(\mathbf{Z}_\ell)$. When varying the different primes $\ell$, the above representations $\rho_\ell$

make up a compatible system of rational $\ell$-adic representations of $\mathbf{Q}$, in the sense of [McGill], and the images of $\rho_\ell$ are almost always the largest possible. The primes for which this does not happen are called the exceptional primes and they are finite in number. More specifically, in the case of the $\tau$ function, the exceptional primes are 2, 3, 5, 7, 23, 691 (this was proved later by Swinnerton-Dyer). For example: $\tau(p) \equiv 1 + p^{11} \pmod{691}$ is the congruence discovered by Ramanujan. As a consequence, the value of $\tau(p)$ mod $\ell$ cannot be deduced from any congruence on $p$, if $\ell$ is a non-exceptional prime.

**13.3.1.** Serre's conjecture on the existence of $\ell$-adic representations associated to modular forms was soon proved by Deligne (1971). This result has been essential for the study of modular forms modulo $p$, for that of $p$-adic modular forms, as well as for the work of H.P.F. Swinnerton-Dyer (1973) on congruences. Swinnerton-Dyer's results on this topic were presented by Serre at the Séminaire Bourbaki [S155, Œ 95(1972)].

**13.3.2.** In the papers [S161, Œ 100(1974)], [S168 (1975)] and [S173, Œ 108(1976)], it is proved that, given a modular form $f = \sum_{n=0}^{\infty} c_n e^{2\pi i n z/M}$ with respect to a congruence subgroup of the full modular group $\mathbf{SL}_2(\mathbf{Z})$, and of integral weight $k \geqslant 1$, for each integer $m \geqslant 1$, the set of integers $n$ which satisfy the congruence $c_n \equiv 0 \pmod{m}$ is of density 1. The proof uses $\ell$-adic representations combined with an analytic argument due to E. Landau. Given a cusp form $f = \sum a_n q^n$, $q = e^{2\pi i z}$, without complex multiplication, of weight $k \geqslant 2$, normalized eigenvector of all the Hecke operators and with coefficients in $\mathbf{Z}$, Serre shows that the set of integers $n$ such that $a_n \neq 0$ has a density which is $> 0$; in particular, the series $f$ is not "lacunary".

**13.3.3.** Deligne and Serre, in the paper [S162, Œ 101(1974)] dedicated to H. Cartan, prove that every cusp form of weight 1, which is an eigenfunction of the Hecke operators, corresponds by Mellin's transform to the Artin $L$-function of an irreducible complex linear representation $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}_2(\mathbf{C})$. Moreover, the Artin conductor of $\rho$ coincides with the level of the cusp form (provided it is a newform). In order to prove the theorem, Serre and Deligne construct Galois representations (mod $\ell$), for each prime $\ell$, of sufficiently small image; this allows them to lift these representations to characteristic zero and to obtain from them the desired complex representation (the proof also uses an average bound on the eigenvalues of the Hecke operators due to Rankin). Note that here the existence of $\ell$-adic representations associated to modular forms of weight $k \geqslant 2$ is used to deduce an existence theorem for complex representations associated to weight $k = 1$ modular forms!

The paper became a basic reference on the subject, since it represents a small but non-trivial step in the direction of the Langlands conjectures. In particular, it shows that certain Artin $L$-functions are entire. Another consequence is that the Ramanujan–Petersson conjecture holds for weight $k = 1$. (For weight $k \geqslant 2$, its truth follows from Deligne's results on Weil's conjectures and on the existence of

$\ell$-adic representations associated to cusp forms.) Not long after, the study of modular forms of weight $k = 1$ was illustrated by Serre in [S179, Œ 110(1977)] with many numerical examples, due to Tate.

**13.3.4.** In 1974, Serre opened the *Journées Arithmétiques* held in Bordeaux with a lecture on Hecke operators $\pmod{\ell}$ [S169, Œ 104(1975)]—in those days a fairly new subject for most of the audience. Consider the algebra $\widetilde{M}$ of modular forms $(\text{mod } \ell)$ with respect to the modular group $\mathbf{SL}_2(\mathbf{Z})$. Serre proves that the systems of eigenvalues $(a_p)$ of the Hecke operators $T_p$, $p \neq \ell$, acting on $\overline{\mathbf{F}}_\ell \otimes \widetilde{M}$ are finite in number. In particular, there exists a weight $k(\ell)$ such that each system of eigenvalues can be realized by a form of weight $\leqslant k(\ell)$; the precise value for $k(\ell)$ was found by Tate. As an illustration, Serre gives a complete list of all the systems $(a_p)$ which occur for the primes $\ell \leqslant 23$. He also raises a series of problems and conjectures which lead, twelve years later, to his own great work [S216, Œ 143(1987)] on modular Galois representations. As is well known, this became a key ingredient in the proof of Fermat's Last Theorem.

**13.3.5.** In [S178, Œ 113(1977)], Serre and H. Stark prove that each modular form of weight $1/2$ is a linear combination of theta series in one variable, thus answering a question of G. Shimura (1973). The proof relies on the "bounded denominator property" of modular forms on congruence subgroups.

**13.3.6.** In the long paper entitled *Quelques applications du théorème de densité de Chebotarev* [S197, Œ 125(1981)], one finds a number of precise estimates both for elliptic curves and for modular forms. These estimates are of two types: either they are unconditional, or they depend on the Generalized Riemann Hypothesis (GRH). The work in question is essentially analytic. It uses several different ingredients:

- explicit forms of Chebotarev theorem due to J.C. Lagarias, H.L. Montgomery, A.M. Odlyzko, with applications to infinite Galois extensions with an $\ell$-adic Lie group as Galois group;
- properties of $\ell$-adic varieties such as the following: the number of points $(\text{mod } \ell^n)$ of an $\ell$-adic analytic variety of dimension $d$ is $\mathrm{O}(\ell^{nd})$, for $n \to \infty$;
- general theorems on $\ell$-adic representations.

Let us mention two applications:

Given a non-zero modular form $f = \sum a_n q^n$, which is an eigenvalue of all the Hecke operators and is not of type CM (complex multiplication) Serre proves that the series $f$ is not lacunary; more precisely, if $M_f(x)$ denotes the number of integers $n \leqslant x$ such that $a_n \neq 0$, then there exists a constant $\alpha > 0$ such that $M_f(x) \sim \alpha x$ for $x \to \infty$. On the other hand, if $f \neq 0$ has complex multiplication, then there exists a constant $\alpha > 0$ such that $M_f(x) \sim \alpha x/(\log x)^{1/2}$, for $x \to \infty$. In concrete examples, Serre provides estimates for $\alpha$.

Furthermore, if $E/\mathbf{Q}$ is an elliptic curve without complex multiplication and if we assume (GRH), then there exists an absolute constant $c$ such that the Galois group $G_\ell$ of the points of the $\ell$-division of $E$ is isomorphic to $\mathbf{GL}_2(\mathbf{F}_\ell)$ for every prime $\ell \geqslant c(\log N_E)(\log\log 2N_E)^3$, where $N_E$ denotes the product of all the primes of bad reduction of $E$.

**13.3.7.** The Dedekind $\eta$ function is a cusp form of weight 1/2. In [S208, Œ 139(1985)], a paper which Serre dedicated to R. Rankin, he studies the lacunarity of the powers $\eta^r$, when $r$ is a positive integer. If $r$ is odd, it was known that $\eta^r$ is lacunary if $r = 1, 3$. If $r$ is even, it was known that $\eta^r$ is lacunary for $r = 2, 4, 6, 8, 10, 14, 26$. Serre proves that, if $r$ is even, the above list is complete. By one of the theorems proved in his Chebotarev paper (see above), this is equivalent to showing that $\eta^r$ is of CM type only if $r = 2, 4, 6, 8, 10, 14$ or 26. The proof consists of showing that the complex multiplication, if it exists, comes from either $\mathbf{Q}(i)$ or $\mathbf{Q}(\sqrt{-3})$.

**13.3.8.** The paper [S216, Œ 143(1987)], entitled *Sur les représentations modulaires de degré* 2 *de* Gal($\overline{\mathbf{Q}}/\mathbf{Q}$), contains one of Serre's outstanding contributions. In this profound paper, dedicated to Y.I. Manin, Serre formulates some very precise conjectures on Galois representations which extend those made twelve years before in Bordeaux [S169, Œ 104(1975)]. We shall only mention two of these conjectures: conjectures (3.2.3?) and (3.2.4?), known nowadays as Serre's modularity conjectures (or, simply, Serre's modularity conjecture). Let $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}_2(\overline{\mathbf{F}}_p)$ be a continuous irreducible representation of odd determinant.

(3.2.3?) There exists a cusp form $f$, with coefficients in $\overline{\mathbf{F}}_p$ which is an eigenfunction of the Hecke operators, whose associated representation $\rho_f$ is isomorphic to the original representation $\rho$.

(3.2.4?) The smallest possible type of the form $f$ of (3.2.3?) is equal to $(N(\rho), k(\rho), \varepsilon(\rho))$, where the level $N(\rho)$ is the Artin conductor of $\rho$ (it reflects the ramification at the primes $\ell \neq p$); the character $\varepsilon(\rho)$ is $\chi^{1-k} \cdot \det(\rho)$, where $\chi$ is the $\ell$-cyclotomic character; the weight $k(\rho)$ is given by a rather sophisticated formula, which depends only on the ramification at $p$.

The paper contains numerical examples for $p = 2, 3, 7$ in support of the conjecture; they were implemented with the help of J.-F. Mestre. Some months after the appearance of this paper, and after examining the examples more closely, Serre slightly modified the conjectures for the primes $p = 2, 3$ in the case of Galois representations of dihedral type.

Since their publication, Serre's conjectures have generated abundant literature. They imply Fermat's Last Theorem (and variants thereof) as well as the Shimura–Taniyama–Weil Conjecture (and generalizations of it). As G. Frey (1986) and Serre pointed out, a weak form of conjecture (3.2.4?), known as conjecture *epsilon*, is sufficient to prove that Fermat's Last Theorem follows from the Shimura–Taniyama–Weil conjecture in the semistable case. Conjecture *epsilon* was proved by K. Ribet (1990) in a brilliant study in which he made use of the arithmetical properties of modular curves, Shimura curves and their Jacobians. Once Ribet's theorem was proved, the task of proving Shimura–Taniyama–Weil conjecture in the semistable case would be accomplished five years later by A. Wiles (1995) and R. Taylor and A. Wiles (1995).

Serre's modularity conjecture may be viewed as a first step in the direction of a mod $p$ analogue of the Langlands program. Many people have worked on it. A general proof was presented at a Summer School held at Luminy (France), July 9–20, 2007 (a survey of this work can be found in the expository paper [Ch. Khare (2007)]).

According to the Citation Database MathSciNet, [S216, Œ 143(1987)] is Serre's second most frequently cited paper, the first one being (fittingly) the one he dedicated to A.Weil [S154, Œ 94(1972)].

**13.3.9.** A summary of Serre's lecture course on Galois representations (mod $p$) and modular forms (mod $p$) can be found in [S218, Œ 145(1988)]. In it, Serre relates modular forms modulo $p$ with quaternions. Two letters on this subject, addressed to J. Tate and D. Kazhdan, are collected in [S249, Œ 169(1996)]. In the letter to Tate, Serre formulates a quaternion approach to modular forms modulo a prime $p$ through quaternion algebras. Let $D$ be the quaternion field over $\mathbf{Q}$ ramified only at $p$ and at $\infty$, and let $D^*(A)$ be the group of the adelic points of the multiplicative group $D^*$, viewed as an algebraic group over $\mathbf{Q}$. The main result of the letter to Tate is that the systems of eigenvalues $(a_\ell)$, with $a_\ell \in \overline{\mathbf{F}}_p$, provided by the modular forms (mod $p$) coincide with those obtained under the natural Hecke action on the space of locally constant functions $f : D^*(A)/D^*_{\mathbf{Q}} \to \overline{\mathbf{F}}_p$. The result is proved by evaluating the modular forms at supersingular elliptic curves. In the letter to Kazhdan, Serre studies certain unramified representations of $\mathbf{GL}_2(\mathbf{Q}_\ell)$, in characteristic $p \neq \ell$, which are universal with the property of containing an eigenvector of the Hecke operator $T_\ell$ with a given eigenvalue $a_\ell$. In an appendix to the paper, R. Livné mentions further developments of these questions. For example, a general study of the representations of $\mathbf{GL}_2(\mathbf{Q}_\ell)$ in characteristic $p \neq \ell$ was done later by Marie-France Vignéras (1989); the case $p = \ell$ has recently been studied by several people.

**13.3.10.** The general strategy of the work of Wiles (1995) and Taylor–Wiles (1995) on modular elliptic curves and Fermat's Last Theorem was presented by Serre at the Séminaire Bourbaki [S248, Œ 168(1995)]. The proof that any semistable elliptic curve defined over $\mathbf{Q}$ is modular is long and uses results of Ribet, Mazur, Langlands, Tunnell, Diamond, among others. On this occasion, Serre said that he did not claim to have verified all the technical details of the proof, "*qui sont essentiels, bien entendu*".

**13.4. Abelian Varieties and $\ell$-adic Representations.** Let $A$ be an abelian variety of dimension $d$ defined over a field $k$. Given a prime $\ell \neq \mathrm{char}(k)$, the Tate module $T_\ell(A) = \varprojlim A[\ell^n](k_s)$ is a free $\mathbf{Z}_\ell$-module of rank $2d$. Let $V_\ell(A) = T_\ell(A) \otimes \mathbf{Q}_\ell$. The action of the absolute Galois group of $k$ on the Tate module of $A$ gives an $\ell$-adic representation $\rho_\ell : \mathrm{Gal}(k_s/k) \to \mathbf{GL}(T_\ell(A)) = \mathbf{GL}_{2d}(\mathbf{Z}_\ell)$; its image, $G_\ell$, is a compact subgroup of $\mathbf{GL}_{2d}(\mathbf{Z}_\ell)$, hence it is a Lie subgroup of the $\ell$-adic Lie group $\mathbf{GL}(T_\ell)$. The Lie algebra $\mathfrak{g}_\ell$ of $G_\ell$ is a subalgebra of $\mathfrak{gl}(V_\ell)$ and does not change under finite extensions of the ground field; it acts on $V_\ell$. When $k$ is finitely generated over $\mathbf{Q}$, it is known that the rank of $\mathfrak{g}_\ell$ is independent of $\ell$, but it is not known whether the same is true for the dimension of $\mathfrak{g}_\ell$. Serre has written several papers and letters on the properties of $G_\ell$ and $\mathfrak{g}_\ell$ (see below).

**13.4.1.** The first occurrence of the $G_\ell$ and the $\mathfrak{g}_\ell$ in Serre's papers can be found in [S109, Œ 62(1964)]; it was complemented a few years later by [S151, Œ 89(1971)]. When $k$ is a number field, the Mordell–Weil theorem says that the group $A(k)$ of the $k$-rational points of $A$ is a finitely generated abelian group. J.W.S. Cassels had asked whether it is true that every subgroup of finite index of $A(k)$ contains a congruence subgroup, at least when $A$ is an elliptic curve. In the first paper, Serre transformed the problem into another one relative to the cohomology of the $G_\ell$'s, namely the vanishing of $H^1(\mathfrak{g}_\ell, V_\ell)$ for every $\ell$ and he solved it when $\dim(A) = 1$. In the second paper, he solved the general case by proving a vanishing criterion for the cohomology of Lie algebras which implies that $H^n(\mathfrak{g}_\ell, V_\ell) = 0$, for every $n$, and every $\ell$.

**13.4.2.** In 1968, Serre and Tate published the seminal paper *Good reduction of abelian varieties* [S134, Œ 79(1968)]. Let $K$ be a field, $v$ a discrete valuation, $O_v$ the valuation ring of $v$ and $k_v$ its residue field, which is assumed to be perfect. Given an abelian variety $A$ defined over $K$, the authors start from the existence of the Néron model $A_v$ of $A$ with respect to $v$, which is a group scheme of finite type over Spec $O_v$. Serre and Tate define the concept of potential good reduction of $A$, which generalized that of good reduction. They prove that $A$ has good reduction at $v$ if and only if the Tate module $T_\ell(A)$ is unramified at $v$, where $\ell$ denotes a prime which differs from the characteristic of $k_v$. This criterion is partially due to A.P. Ogg (in the case of elliptic curves) and partially to I. Shafarevich. In their proof, the structure of the connected component $\widetilde{A}_v^0$ of the special fiber $\widetilde{A}_v$ of $A_v$ appears: it is an extension of an abelian variety $B$ by a linear group $L$, and $L$ is a product of a torus $S$ by a unipotent group $U$. The abelian variety $A$ has good reduction if and only if $L = 1$; it has potential good reduction if and only if $L = U$; and it has semistable reduction if and only if $L = S$. A second theorem says that $A$ has potential good reduction if and only if the image of the inertia group $I(\overline{v})$ for the $\ell$-adic representation $\rho_\ell : \mathrm{Gal}(K_s/K) \to \mathrm{Aut}(T_\ell)$ is finite. An appropriate use of the characters of Artin and Swan then allows the definition of the conductor of $A$. The semistable reduction theorem, conjectured by Serre in 1964 and proved later by Grothendieck in (SGA 7), would allow the definition of the conductor for every abelian variety. (The semistable reduction theorem was also proved by D. Mumford, except that his proof, based on the use of theta functions, did not include the case where the residue characteristic is equal to 2.)

Suppose that $A$ has complex multiplication by $F$ over the field $K$, where $F$ denotes an algebraic number field of degree $2d$, $d = \dim(A)$. In the same work, Serre and Tate prove that every abelian variety, defined over an algebraic number field $K$ and with complex multiplication over this field, has potential good reduction at all the places of $K$, and that it has good reduction at the places of $K$ outside the support of its Grössencharakter. This result generalizes some earlier ones of M. Deuring (1955) in the case of elliptic curves. The exponent of the conductor at $v$ is given by $2dn_v$, where $n_v$ is the smallest integer such that the Grössencharakter is zero when restricted to the ramification group $I(\overline{v})^{n_v}$, in the upper numbering.

**13.4.3.** In [S209, Œ 135(1985)], Serre explains how the theorems obtained by G. Faltings (1983) in his paper on the proof of Mordell's conjecture allow a better understanding of the properties of the $\ell$-adic representations associated to abelian varieties.

In the first part of the lectures, Serre gives an effective criterion for showing that two $\ell$-adic representations are isomorphic (the "*méthode des corps quartiques*"). This criterion was applied to prove that two elliptic curves, studied by J.-F. Mestre, of conductor 5077, are isogenous.

Let $K$ be an algebraic number field and $A$ an abelian variety defined over $K$ of dimension $d$. Let $\rho_\ell : \mathrm{Gal}(\overline{K}/K) \to \mathbf{GL}(T_\ell(A))$ be the $\ell$-adic representation defined by the Tate module. Let $G_\ell^{\mathrm{alg}}$ be the closure of $G_\ell$ under the Zariski topology, which is a $\mathbf{Q}_\ell$-algebraic subgroup of the general linear group $\mathbf{GL}_{2d}$. Mumford and Tate conjectured that, given $A$ and $K$, the group $G_\ell^{\mathrm{alg}}$ is essentially independent of $\ell$ and, more precisely, that the connected component $(G_\ell^{\mathrm{alg}})^0$ could be deduced from the Mumford–Tate group by extension of scalars of $\mathbf{Q}$ to $\mathbf{Q}_\ell$. In the second part of the course [S209, Œ 135(1985)], Serre proves a series of results in this direction. He shows for instance that the finite group $G_\ell^{\mathrm{alg}}/(G_\ell^{\mathrm{alg}})^0$ is independent of $\ell$.

**13.4.4.** In the course [S213, Œ 136(1986)], Serre studies the variation with $\ell$ of the $\ell$-adic Lie groups associated to abelian varieties. Let us keep the previous notation. Given the homomorphism

$$\rho : \mathrm{Gal}(\overline{K}/K) \to \prod_\ell G_\ell \subset \prod_\ell \mathrm{Aut}(T_\ell),$$

Serre proves that, if $K$ is sufficiently large, the image of $\rho$ is open in the product $\prod_\ell G_\ell$, i.e. the $\rho_\ell$ are "almost independent". In the case where $n$ is odd, or is equal to 2 or 6, and if $\mathrm{End}(A) = \mathbf{Z}$, he shows that the image of $\rho$ is open in the product of the groups of symplectic similitudes $\prod \mathbf{GSp}(T_\ell, e_\ell)$. Here $e_\ell$ is the alternating form over $T_\ell(A)$ deduced from a polarization $e$ of $A$. The ingredients of the proof are many: the above theorems of Faltings, Frobenius tori, McGill theory, properties of inertia groups at the places which divide $\ell$, as well as group-theoretic information regarding the subgroups of $\mathbf{GL}_N(\mathbf{F}_\ell)$ supplied by theorems of V. Nori (1985–1987).

The proofs of the above results have not been published in a formal way, but one can find an account of them in Serre's letters to K. Ribet [Œ 133(1981)] and [Œ 138(1986)], to D. Bertrand [Œ 134(1984)], and to M.-F. Vignéras [Œ 137(1986)].

**13.5. Motives.** A first lecture on zeta and $L$-functions in the setting of the theory of schemes (of finite type over $\mathrm{Spec}(\mathbf{Z})$) was given by Serre in [S112, Œ 64(1965)]. One finds in it a generalization of Chebotarev's density theorem to schemes of arbitrary dimension.

**13.5.1.** In his lecture in the Séminaire Delange–Pisot–Poitou [Œ 87(1969/70)], Serre introduces several definitions and formulates several conjectures about the local factors (gamma factors included) of the zeta function of a smooth projective variety over a number field. The local factors at the primes of good reduction do not raise any problem. The interesting cases are: (a) the primes with bad reduction; (b) the archimedean primes. In both cases Serre gives definitions based, in case (a), on the action of the local Galois group on the $\ell$-adic cohomology, and in case (b), on the Hodge type of the real cohomology. The main conjecture is that such a zeta function has an analytic continuation to the $s$-plane and a very simple functional equation.

**13.5.2.** The subject of $\ell$-adic representations had already been considered by Serre in [S177, Œ 112(1977)], in his address to the Kyoto Symposium on Algebraic Number Theory. In this paper, which is rich in problems and conjectures, we find the statement of the conjecture of Shimura–Taniyama–Weil, according to which any elliptic curve over $\mathbf{Q}$ of conductor $N$ is a quotient of the modular curve $X_0(N)$.

**13.5.3.** The paper [S229, Œ 154(1991)] is a short introduction to the theory of motives. Along these lines, we also highlight the paper [S239, Œ 160(1993)], which corresponds to a text Serre wrote for Bourbaki in 1968. The paper deals with algebraic envelopes of linear groups and their relationship with different types of algebras, coalgebras and bialgebras. Its last section contains an account of the dictionary between compact real Lie groups and complex reductive algebraic groups.

**13.5.4.** To finish this section we shall briefly summarize the paper entitled *Propriétés conjecturales des groupes de Galois motiviques et des représentations $\ell$-adiques* [S243, Œ 161(1994)]. Serre formulates a series of conjectures regarding $\ell$-adic representations which generalize many of his previous results. We denote by $M$ the category of pure motives over a subfield $k$ of $\mathbf{C}$, which we suppose to be of finite type over $\mathbf{Q}$. The motivic Galois group $G_M$ is related to the absolute Galois group of $k$ by means of an exact sequence $1 \to G_M^0 \to G_M \to \mathrm{Gal}(\overline{k}/k) \to 1$. Given a motive $E$ over $k$, let $M(E)$ be the smallest Tannakian subcategory of $M$ which contains $E$. Suppose that the standard conjectures and Hodge conjecture are true. Under these assumptions and in an optimistic vein, Serre formulates a series of conjectures and questions aimed at the description of Grothendieck's "motivic paradise". We stress the following ones:

(1?) The motivic Galois group $G_M$ is pro-reductive.
(2?) The motivic Galois group $G_{M(E)}$ is characterized by its tensor invariants.

(3?) The group $G_{M(E)/\mathbf{Q}_\ell}$ is the closure in the Zariski topology of the image of the $\ell$-adic representation $\rho_{\ell,E} : \mathrm{Gal}(\bar{k}/k) \to G_{M(E)}(\mathbf{Q}_\ell)$, associated to $E$.

(4?) The connected pro-reductive group $G^0_M$ decomposes as $G^0_M = C \cdot D$, where $C$ is a pro-torus, equal to the identity component of the centre of $G^0_M$, and $D$ is a pro-semisimple group, equal to the derived group of $G^0_M$.

(5?) If $S = (G^0_M)^{\mathrm{ab}}$, then $S$ is the projective limit of the tori $T_{\mathfrak{m}}$ defined in his McGill book.

(6?) Every homomorphism $G^0_M \to \mathbf{PGL}_2$ has a lifting to $G^0_M \to \mathbf{GL}_2$.

(7) Which connected reductive groups are realized as $G_{M(E)}$? Are $G_2$ and $E_8$ possible?

(8?) The group $G_{\ell,E} = \mathrm{Im}(\rho_{\ell,E})$ is open in $G_{M(E)}(\mathbf{Q}_\ell)$. Let

$$\rho_E = (\rho_{\ell,E}) : \mathrm{Gal}(\bar{k}/k) \to \prod_\ell G_{\ell,E} \subset G_{M(E)}(\mathbf{A}^f).$$

Suppose that $G_{M(E)}$ is connected. Then $E$ is a "maximal motive" if and only if $\mathrm{Im}(\rho_E)$ is open in the group $G_{M(E)}(\mathbf{A}^f)$, where $\mathbf{A}^f$ is the ring of the finite adeles of $\mathbf{Q}$.

The paper ends with a statement of the Sato–Tate conjecture for arbitrary motives.

## 14 Group Theory

In response to a question raised by Olga Taussky (1937) on class field towers, Serre proves in [S145, Œ 85(1970)] that for a finite $p$-group $G$, the knowledge of $G^{ab}$ does not in general imply the triviality of any term $D^n(G)$ of its derived series. More precisely, for every $n \geqslant 1$ and for every non-cyclic finite abelian $p$-group $P$ of order $\neq 4$, there exists a finite $p$-group $G$ such that $D^n(G) \neq 1$ and $G^{ab} \simeq P$.

**14.1. Representation Theory.** Serre's popular book *Représentations linéaires des groupes finis* [S130, RLGF(1967)] gives a reader-friendly introduction to representation theory. It also contains less elementary chapters on Brauer's theory of modular representations, explained in terms of Grothendieck $K$-groups, the highlight being the "cde triangle". The text is well known to physicists and chemists[1], and its first chapters are a standard reference in undergraduate or graduate courses on the subject.

**14.1.1.** The paper [S245, Œ 164(1994)] is about the semisimplicity of the tensor product of group representations. A theorem of Chevalley (1954) states that, if $k$ is a field of characteristic zero, $G$ is a group, and $V_1$ and $V_2$ are two semisimple

[1] Indeed the first part of the book was written by Serre for the use of his wife Josiane who was a quantum chemist and needed character theory in her work.

$k(G)$-modules of finite dimension, then their tensor product $V_1 \otimes V_2$ is a semisimple $k[G]$-module. Serre proves that this statement remains true in characteristic $p > 0$, provided that $p$ is large enough. More precisely, if $V_i$, $1 \leqslant i \leqslant m$, are semisimple $k[G]$-modules and $p > \sum(\dim V_i - 1)$, then the $k[G]$-module $V_1 \otimes \cdots \otimes V_m$ is also semisimple. The bound on $p$ is best possible, as the case $G = \mathbf{SL}_2(k)$ shows. In order to prove this, Serre first considers the case in which $G$ is the group of points of a simply connected quasi-simple algebraic group, and the representations $V_1$ and $V_2$ are algebraic, irreducible, and of restricted type. In this case, the proof relies on arguments on dominant weights due to J.C. Jantzen (1993). The general case is reduced to the previous one by using a "saturation process" due to V. Nori, which Serre had already used in his study of the $\ell$-adic representations associated with abelian varieties (see Sect. 13.4 above). The study of these topics is continued in [S252, Œ 171(1997)], where one finds converse theorems such as: if $V \otimes V'$ is semisimple and $\dim(V')$ is not divisible by $\mathrm{char}(k)$, then $V$ is semisimple. Here the proofs use only linear (or multilinear) algebra; they are valid in any tensor category.

**14.1.2.** In the Bourbaki report [S273 (2004), SEM(2008)], Serre extends the notion of complete reducibility (that is, semisimplicity) to subgroups $\Gamma$ not only of $\mathbf{GL}_n$ but of any reductive group $G$ over a field $k$. The main idea is to use the Tits building $T$ of $G$. A subgroup $\Gamma$ of $G$ is called completely reducible in $G$ if, for every maximal parabolic subgroup $P$ of $G$ containing $\Gamma$, there exists a maximal parabolic subgroup $P'$ of $G$ opposite to $P$ which contains $\Gamma$. There is a corresponding notion of "$G$-irreducibility": $\Gamma$ is called $G$-irreducible if it is not contained in any proper parabolic subgroup of $G$, i.e. if it does not fix any point of the building $X$. There is also a notion of "$G$-indecomposability". These different notions behave very much like the classical ones, i.e. those relative to $G = \mathbf{GL}_n$; for instance, there is an analogue of the Jordan–Hölder theorem and also of the Krull–Schmidt theorem. The proofs are based on Tits' geometric theory of spherical buildings. As one of the concrete applications given in the paper, we only mention the following: if $\Gamma \subset G(k)$, $G$ is of type $E_8$ and $V_i$, $1 \leqslant i \leqslant 8$, denote the 8 fundamental irreducible representations of $G$, and if one of them is a $\Gamma$-module semisimple, then all the others are also semisimple provided that $\mathrm{char}(k) > 270$.

**14.1.3.** The Oberwolfach report [S272 (2004)], states without proof two new results on the characters of compact Lie groups. The first one is a generalization of a theorem of Burnside for finite groups: given an irreducible complex character $\chi$ of a compact Lie group $G$, of degree $> 1$, there exists an element $x \in G$ of finite order with $\chi(x) = 0$. The second one states that $\mathrm{Tr}(\mathrm{Ad}(g)) \geqslant -\mathrm{rank}(G)$ for all $g \in G$, the bound being optimal if and only if there is an element $c \in G$ such that $ctc^{-1} = t^{-1}$ for every $t \in T$, where $T$ is a maximal torus of $G$; the proof is a case-by-case explicit computation (in the $E_6$ case, the computation was not made by Serre himself but by A. Connes).

**14.1.4.** In another Oberwolfach report [S279 (2006)], Serre defines the so-called Kac coordinates in such a way that they can be used to classify the finite subgroups of $G$ which are isomorphic to $\mu_n$, without having to assume that $n$ is prime to the characteristic.

**14.1.5.** In 1974, Serre had asked W. Feit whether, given a linear representation

$$\rho : G \to \mathbf{GL}_n(K)$$

of a finite group $G$ over a number field $K$, it could be realized over the ring of integers $O_K$. Although he did not expect a positive answer, he did not know of any counterexample. Given $\rho$, there are $O_K$-lattices which are stable under the action of $G$; but the point is that as $O_K$ is a Dedekind ring, these lattices need not be free as $O_K$-modules. There is an invariant attached to them which lies in the ideal class group $C_K = \mathrm{Pic}(O_K)$ of $K$. Feit provided the following counterexample: if $G = Q_8$ is the quaternion group of order 8 and $K = \mathbf{Q}(\sqrt{-35})$, the answer to the question is no. The paper [S281 (2008)], reproduces three letters of Serre to Feit about this question, written in 1997. Their purpose was to clarify the mysterious role of $\sqrt{-35}$ in Feit's counterexample. Let $K = \mathbf{Q}(\sqrt{-N})$, for $N$ a positive square free integer, $N \equiv 3 \pmod 8$. Let $O_K$ denote its ring of integers. The field $K$ splits the quaternion algebra $(-1,-1)$, hence there exists an irreducible representation $V$ of degree 2 over $K$ of $Q_8$. In the first letter, Serre proves that there exists an $O_K$-free lattice of $V$ which is stable under the group $Q_8$ if and only if the integer $N$ can be represented by the binary quadratic form $x^2 + 2y^2$. In order to prove this equivalence, Serre makes use of Gauss genus theory: any lattice $L \subset V$ stable under $Q_8$ yields an invariant $c(L)$ which lies in the genus group $C_K/C_K^2$ of the quadratic field $K$. It turns out that $L$ is free as $O_K$-module if and only if $c(L) = 1$. The exact evaluation of the genus characters on $c(L)$ yields the criterion above.

Another version of the computation of the invariant $c(L)$ is explained in the second letter. Serre uses the fact, due to Gauss, that for a positive square-free integer $N$, $N \equiv 3 \pmod 8$, any representation of $N$ as a sum of three squares yields an $O_K$-module of rank 1 which lies in a well defined genus and, moreover, every class in that genus is obtainable by a suitable representation of $N$ as a sum of three squares. If $D = (-1,-1)$ denotes the standard quaternion algebra over $\mathbf{Q}$ and $R$ is its Hurwitz maximal order, Serre embeds the ring of integers $O_K$ in $R$ by mapping $\sqrt{-N}$ to $ai + bj + ck$, where $a^2 + b^2 + c^2 = N$. It turns out that the invariant $c(R)$ is the same as the one obtained before. In the third letter, and more generally, given any quaternion algebra $D$ over $\mathbf{Q}$ and an imaginary quadratic field $K$ which splits $D$, if we choose an embedding $K \to D$, the $O_K$-invariant $c(O_D)$ of a maximal order $O_D$ containing $O_K$ does not depend on the choice of $O_D$. Serre determines $c(O_D) = c(D,K) \in C_K/C_K^2$ in terms of $D$ and $K$. By making use of the Hilbert symbol, the genus group $C_K/C_K^2$ can be embedded in the 2-component of the Brauer group $\mathrm{Br}_2(\mathbf{Q})$; moreover the image of $c(D,K)$ in $\mathrm{Br}_2(\mathbf{Q})$ is equal to $(D) + (d_D, -d)$. In this formula, $(D)$ denotes the element of the Brauer group defined by the quaternion algebra $D$, $d_D$ is the signed discriminant of $D$, $-d$, with $d > 0$, is the discriminant of $K$, and $(d_D, -d)$ stands for the Hilbert symbol. In the special case $D = (-1,-1)$ and $Q_8$, the formula tells us that there exists a free

$O_K$-module of rank 2 which gives the standard irreducible representation of $Q_8$ over $K = \mathbf{Q}(\sqrt{-d})$ if and only if either $(-2, d) = 0$ or $(-1, d) = 0$; that is, if and only if $d$ is representable either by $x^2 + 2y^2$ or by $x^2 + y^2$. For example, if $d = 8p$, $p \equiv 3 \pmod 8$, $p$ prime, non-free lattices exist.

**14.2. Algebraic Groups.** The lecture course [S141, Œ 84(1969)] focused on discrete groups. Some of its contents would be published in [S139, Œ 83(1969)] and [S149, Œ 88(1971)]. Another part was published in the book *Arbres, amalgames, $SL_2$*, [S176, AA(1977)], written with the help of H. Bass. In the first chapter Serre shows that it is possible to recover a group $G$ which acts on a tree $X$ from the quotient graph or fundamental domain $G\backslash X$, and the stabilizers of the vertices and of the edges. If $G\backslash X$ is a segment, then $G$ may be identified with an amalgam of two groups and, moreover, every amalgam of two groups can be obtained in this way. The study of relations between amalgams and fixed points show that groups such as $\mathbf{SL}_3(\mathbf{Z})$ and $\mathbf{Sp}_4(\mathbf{Z})$ are not amalgams, since one can show that they always have fixed points when they act on trees, see [S163 (1974)]; the method extends to all $G(\mathbf{Z})$, where $G$ is any reductive group-scheme over $\mathbf{Z}$ which is simple of rank $\geqslant 2$.

In the second chapter, the results are applied to the study of the groups $\mathbf{SL}_2(k)$, where $k$ is a local field. The group $\mathbf{SL}_2(k)$ acts on the Bruhat–Tits tree associated to the space $k^2$. The vertices of this tree are the classes of lattices of $k^2$. In this way, Serre recovers a theorem due to Y. Ihara by which every torsion-free discrete subgroup of $\mathbf{SL}_2(\mathbf{Q}_p)$ is free.

According to MathSciNet, this book is now Serre's most cited publication.

**14.2.1.** A question raised by Grothendieck concerning linear representations of group schemes was answered by Serre in [S136, Œ 81(1968)]. Suppose that $C$ is a coalgebra over a Dedekind ring $A$ which is flat. If $Com_A$ denotes the abelian category of comodules over $C$ which are of finite type as $A$-modules, one may consider the Grothendieck ring $\mathrm{R}_A$ of $Com_A$. Let $K$ be the field of fractions of $A$. Serre proves that the natural morphism $i : \mathrm{R}_A \to \mathrm{R}_K$, $E \mapsto E \otimes K$, is an isomorphism if $A$ is principal and under the assumption that all decomposition homomorphisms (defined as in Brauer's theory for finite groups) are surjective. If $M$ is an abelian group and $T_M$ denotes the $A$-group scheme whose character group is $M$, the bialgebra $C(M)$ can be identified with the group algebra $A[M]$. If $A$ is principal, one has an isomorphism $ch : \mathrm{R}_A(T_M) \xrightarrow{\sim} \mathbf{Z}[M]$, provided by the rank. Next, Serre considers a split reductive group $G$ and a split torus $T$ of $G$, which exists by hypothesis. By composing $ch$ with the restriction homomorphism $Res : \mathrm{R}_K(G) \to \mathrm{R}_K(T)$, a homomorphism $ch_G : \mathrm{R}_K(G) \to \mathbf{Z}[M]$ is obtained. Serre proves that $ch_G$ is injective and that its image equals the subgroup $\mathbf{Z}[M]^W$ of the elements of $\mathbf{Z}[M]$ which are invariant under the Weyl group $W$ of $G$ relative to $T$. As an illustration of this result, the paper gives the following example: if $G = \mathbf{GL}_n$, $M = \mathbf{Z}^n$ and $W = S_n$ is the symmetric group on $n$ elements, then $\mathbf{Z}[M] = \mathbf{Z}[X_1, \ldots, X_n, X_1^{-1}, \ldots, X_n^{-1}]$ and $\mathrm{R}_A(\mathbf{GL}_n) = \mathrm{R}_K(\mathbf{GL}_n) = \mathbf{Z}[M]^W = \mathbf{Z}[\lambda_1, \ldots, \lambda_n]_{\lambda_n}$, where $\lambda_1, \ldots, \lambda_n$ denote the elementary symmetric functions in $X_1, \ldots, X_n$, and the subscript stands for localization with respect to $\lambda_n$. This was what Grothendieck needed for his theory of $\lambda$-rings.

**14.3. Finite Subgroups of Lie Groups and of Algebraic Groups.** The problem of the determination of the finite subgroups of a Lie group has received a great amount of attention. Embedding questions of finite simple groups (and their non-split central extensions) in Lie groups of exceptional type have been solved by the work of many mathematicians. The paper [S250, Œ 167(1996)] contains embeddings of some of the groups $\mathbf{PSL}_2(\mathbf{F}_p)$ into simple Lie groups. Let $G$ denote a semisimple connected linear algebraic group over an algebraically closed field $k$, which is simple of adjoint type; let $h$ be its Coxeter number. The purpose of the paper is to prove that, if $p = h + 1$ is a prime, then the group $\mathbf{PGL}_2(\mathbf{F}_p)$ can be embedded into $G(k)$ (except if char$(k) \neq 2$ and $h = 2$), and that if $p = 2h + 1$ is a prime, then the group $\mathbf{PSL}_2(\mathbf{F}_p)$ can be embedded into $G(k)$. Since for $G = \mathbf{PGL}_2$ one has $h = 2$, the theorem generalizes the classical result that the groups $A_4 = \mathbf{PSL}_2(\mathbf{F}_3)$, $S_4 = \mathbf{PGL}_2(\mathbf{F}_3)$, and $A_5 = \mathbf{PSL}_2(\mathbf{F}_5)$ can be embedded into $\mathbf{PGL}_2(\mathbf{C})$. The result for $p = 2h + 1$ was known in the case of characteristic zero; it was part of a conjecture by B. Kostant (1983), and it had been verified case by case with the aid of computers. Moreover, the values $h + 1$ or $2h + 1$ for $p$ are maximal in the sense that if $\mathbf{PGL}_2(\mathbf{F}_p)$, respectively $\mathbf{PSL}_2(\mathbf{F}_p)$, are embedded in $G(\mathbf{C})$ then $p \leqslant h + 1$, respectively $p \leqslant 2h + 1$.

In his paper, Serre proves the two results in a unified way.

One starts from a certain principal homomorphism $\mathbf{PGL}_2(\mathbf{F}_p) \to G(\mathbf{F}_p)$ if $p \geqslant h$. If $p = h + 1$, this homomorphism can be lifted to a homomorphism $\mathbf{PGL}_2(\mathbf{F}_p) \to G(\mathbf{Z}_p)$. A key point is that the Lie algebra $L$ of $G_{/\mathbf{F}_p}$ turns out to be cohomologically trivial as a $\mathbf{PGL}_2(\mathbf{F}_p)$-module through the adjoint representation. This is not the case if $p = 2h + 1$, since then $H^2(\mathbf{PGL}_2(\mathbf{F}_p), L)$ has dimension 1; lifting to $\mathbf{Z}_p$ is not possible; one has to use a quadratic extension of $\mathbf{Z}_p$. Once this is done, the case where char$(k) = 0$ is settled. An argument based on the Bruhat–Tits theory gives the other cases.

As a corollary of the theorem, one obtains that $\mathbf{PGL}_2(\mathbf{F}_{19})$ and $\mathbf{PSL}_2(\mathbf{F}_{37})$ can be embedded in the adjoint group $E_7(\mathbf{C})$ and that $\mathbf{PGL}_2(\mathbf{F}_{31})$ and $\mathbf{PSL}_2(\mathbf{F}_{61})$ can be embedded in $E_8(\mathbf{C})$.

**14.3.1.** In his lecture [S260 (1999), SEM(2008)] delivered at the Séminaire Bourbaki (1998–1999), Serre describes the state of the art techniques in the classification of the finite subgroups of a connected reductive group $G$ over an algebraically closed field $k$ of characteristic zero. He begins by recalling several important results. For example, if $p$ is a prime which does not divide the order of the Weyl group $W$ of $G$, then every $p$-group $A$ of $G$ is contained in a torus of $G$ and hence is abelian. The torsion set Tor$(G)$ is, by definition, the set of prime numbers $p$ for which there exists an abelian $p$-subgroup of $G$ which cannot be embedded in any torus of $G$. The sets Tor$(G)$, for $G$ simply connected and quasi-simple, are well known; for instance Tor$(G) = \{2, 3, 5\}$ if $G$ is of type $E_8$; moreover Tor$(G) = \emptyset$ is equivalent to $H^1(K, G) = 0$ for every extension $K$ of $k$. For $A$ a non-abelian finite simple group, Serre reproduces a table by Griess–Ryba (1999) giving the pairs $(A, G)$ for which $G$ is of exceptional type and $A$ embeds projectively in $G$. In order to see that the table is, in fact, complete, the classification of finite simple groups is used.

**14.3.2.** Part of the material of the paper [S280 (2007)] arose from a series of three lectures at the École Polytechnique Fédérale de Lausanne in May 2005. Given a reductive group $G$ over a field $k$ and a prime $\ell$ different from char$(k)$, and $A$ a finite subgroup of $G(k)$, the purpose of the paper is to give an upper bound for $v_\ell(A)$, that is the $\ell$-adic valuation of the order of $A$, in terms of invariants of $G$, $k$ and $\ell$. Serre provides two types of such bounds, which he calls $S$-bounds and $M$-bounds, in recognition of previous work by I. Schur (1905) and H. Minkowski (1887).

The Minkowski bound, $M(n, \ell)$, applies to the situation $G = \mathbf{GL}_n$ and $k = \mathbf{Q}$ and is optimal in the sense that for every $n$ and for every $\ell$ there exists a finite $\ell$-subgroup $A$ of $\mathbf{GL}_n(\mathbf{Q})$ for which $v_\ell(A) = M(n, \ell)$. By making use of the (at the time) newly created theory of characters, due to Frobenius, Schur extended Minkowski's results to an arbitrary number field $k$: he defined a number $M_k(n, \ell)$ such that $v_\ell(A) \leqslant M_k(n, \ell)$ for any finite $\ell$-subgroup of $\mathbf{GL}_2(\mathbf{C})$ such that $\mathrm{Tr}(g)$ belongs to $k$ for any $g \in A$. As in the case $k = \mathbf{Q}$, Schur's bound is optimal. Both results were recalled by Serre in his lectures with almost complete proofs.

The $S$-bound for any reductive group and any finite subgroup of $G(k)$ is obtained in terms of $v_\ell(W)$, the $\ell$-adic valuation of the order of the Weyl group of $G$ and certain cyclotomic invariants of the field $k$, defined *ad hoc*. The Minkowski bound is more precise, but in order to obtain it, Serre needs to assume that the group $G$ is semisimple of inner type (the action of $\mathrm{Gal}(k_s/k)$ on its Dynkin diagram is trivial). If $r$ is its rank, then its Weyl group $W$ has a natural linear representation of degree $r$. The ring of invariants $\mathbf{Q}[x_1, \dots, x_r]^W$ is a polynomial algebra $\mathbf{Q}[P_1, \dots, P_r]$, where $P_i$ are homogenous polynomials of degrees $d_1 \leqslant d_2 \leqslant \cdots \leqslant d_r$. Under the assumption that $G$ is semisimple of inner type, with root system $R$, Serre gives a Minkowski-style bound $M(\ell, k, R)$ for $G$ which depends only on the $\ell$-cyclotomic invariants of the field $k$ and the degrees $d_i, i = 1, \dots, r$. Moreover, it is optimal, except when $\ell = 2$ and $-1$ does not belong to $W$. As an illustration, let us mention that, if $G$ is a $\mathbf{Q}$-group of type $E_8$, then the order of any finite subgroup of $G(\mathbf{Q})$ divides

$$M(\mathbf{Q}, E_8) = 2^{30} \cdot 3^{13} \cdot 5^5 \cdot 7^4 \cdot 11^2 \cdot 13^2 \cdot 19 \cdot 31,$$

and that this bound is sharp.

## 15 Miscellaneous Writings

Serre has written an endless number of impeccable letters over the years. They are now found as appendices of books, in papers or, simply, carefully saved in the drawers of mathematicians. As mentioned above, some of these letters are included in the *Œuvres*. Some were collected in the text edited by S.S. Chern and F. Hirzebruch in [Wolf Prize in Mathematics, vol. 2, World Scientific, 2001]: a letter to John

McCleary (1997), two letters to David Goss (1991, 2000), a letter to Pierre Deligne (1967), and a letter to Jacques Tits (1993). One finds in them comments on his thesis, on the writing of FAC, on $\ell$-adic representations, as well as historical data on the "Shimura–Taniyama–Weil" modularity conjecture. In the letter to Tits, there is an account of the themes on which Serre was working in the years around 1993: Galois representations, inverse Galois problem, Abhyankar's problem, trace forms, and Galois cohomology. The Grothendieck–Serre correspondence [S263, GRSE(2001)], published more recently, is another invaluable resource for understanding the origins of the concepts and tools of current algebraic geometry.

**15.0.1.** Serre has written essays on the work of other mathematicians: for example, a publication of historical character on a prize delivered to J.S. Smith and H. Minkowski [S238 (1993)], or publications about the life and work of A. Weil [S259 (1999)] and that of A. Borel [S270 (2004)], [S271 (2004)]. He was the editor of the Collected Works of F.G. Frobenius [S135 (1968)], in three volumes. He and R. Remmert were the editors of the Collected Works of H. Cartan [S192 (1979)], also in three volumes. He was the editor of the Collected Works of R. Steinberg [S253 (1997)].

**15.0.2.** We should also mention expository papers that Serre likes to call "mathematical entertainment", where he takes a rather simple-looking fact as a starting point for explaining a variety of deeper results.

One such paper is [S206, Œ 140(1985)], whose title is just the high school discriminant formula $\Delta = b^2 - 4ac$. Given an integer $\Delta$, one wants to classify the quadratic polynomials $ax^2 + bx + c$ with discriminant $\Delta$, up to $\mathbf{SL}_2(\mathbf{Z})$-conjugation. This is a classical problem, started by Euler, Legendre and Gauss. Serre explains the results which were obtained in the late 1980s by combining Goldfeld's ideas (1976) with a theorem of Gross–Zagier (1986) and Mestre's proof (1985) of the modularity of a certain elliptic curve of conductor 5077 and rank 3.

Another such paper is [S268 (2002), SEM(2008)]. By an elementary theorem of C. Jordan (1872), if $G$ is a group acting transitively on a finite set of $n > 1$ elements, the subset $G_0$ of the elements of $G$ which act without fixed points is nonempty. Moreover, P.J. Cameron and A.M. Cohen (1992) have refined this result by proving that the ratio $|G_0|/|G| \geqslant 1/n$, and that it is $> 1/n$ if $n$ is not a prime power. Serre gives two applications. The first one is topological and says that if $f : T \to S$ is a finite covering of a topological space $S$, of degree $n > 1$, and with $T$ path-connected, then there exists a continuous map of the circle $\mathbf{S}_1$ in $S$ which cannot be lifted to $T$. The second application is arithmetical and concerns the number of zeros $N_p(f)$ in the finite field $\mathbf{F}_p$ of a polynomial $f \in \mathbf{Z}[X]$. Serre shows that, if the degree of $f$ is $n > 1$ and $f$ is irreducible, then the set $P_0(f)$ of the primes $p$ such that $N_p(f) > 0$ has a natural density $\geqslant 1/n$. The proof consists of combining Cameron–Cohen's theorem with Chebotarev's density theorem. The paper is illustrated with the computation of $N_p(f)$ for $f = x^n - x - 1$ and $n = 2, 3, 4$. In these three cases, it is shown how the numbers $N_p(f)$ can be read from the coefficients of suitable cusp forms of weight 1.

And, finally, let us mention the preprint *How to use finite fields for problems concerning infinite fields*, a mathematical entertainment just written by Serre [S284 (2009)] in which he discusses old results of P.A. Smith (1934), M. Lazard (1955) and A. Grothendieck (1966), and shows how to prove them (and sometimes improve them) either with elementary tools or with topological techniques.

**Acknowledgements.** The author acknowledges the valuable help provided by Jean-Pierre Serre and by the editors of the book during the preparation of the manuscript.

## References

[Œ] Serre, J.-P.: Œuvres, Collected Papers, vol. I (1949–1959), vol. II (1960–1971), vol. III (1972–1984); vol. IV (1985–1998). Springer, Berlin (1986; 2000)

[GACC] Serre, J.-P.: Groupes algébriques et corps de classes. Hermann, Paris (1959); 2nd edn. 1975 [translated into English and Russian]

[CL] Serre, J.-P.: Corps locaux. Hermann, Paris (1962); 4th edn. 2004 [translated into English]

[CG] Serre, J.-P.: Cohomologie galoisienne. LNM, vol. 5. Springer, Berlin (1964); 5th edn. revised and completed 1994 [translated into English and Russian]

[LALG] Serre, J.-P.: Lie Algebras and Lie Groups. Benjamin, New York (1965); 2nd edn. LNM, vol. 1500. Springer, Berlin, 1992 [translated into Russian]

[ALM] Serre, J.-P.: Algèbre locale. Multiplicités. LNM, vol. 11. Springer, Berlin (1965); written with the help of P. Gabriel; 3rd edn. 1975 [translated into English and Russian]

[ALSC] Serre, J.-P.: Algèbres de Lie semi-simples complexes. Benjamin, New York (1966) [translated into English and Russian]

[RLGF] Serre, J.-P.: Représentations linéaires des groupes finis. Hermann, Paris (1967) [translated into English, German, Japanese, Polish, Russian and Spanish]

[McGill] Serre, J.-P.: Abelian $l$-adic Representations and Elliptic Curves. Benjamin, New York (1968), written with the help of W. Kuyk and J. Labute; 2nd edn. A.K. Peters, 1998 [translated into Japanese and Russian]

[CA] Serre, J.-P.: Cours d'arithmétique. Presses Univ. France, Paris (1970); 4th edn. 1995 [translated into Chinese, English, Japanese and Russian]

[AA] Serre, J.-P.: Arbres, amalgames, $SL_2$. Astérisque, vol. 46. Soc. Math. France, Paris (1977), written with the help of H. Bass; 3rd edn. 1983 [translated into English and Russian]

[MW] Serre, J.-P.: Lectures on the Mordell–Weil Theorem. Vieweg, Wiesbaden (1989); 3rd edn. 1997, translated and edited by Martin Brown from notes of M. Waldschmidt. French edition: Publ. Math. Univ. Pierre et Marie Curie, 1984

[TGT] Serre, J.-P.: Topics in Galois Theory. Jones & Bartlett, Boston (1992), written with the help of H. Darmon; 2nd edn., AK Peters, 2008

[SEM] Serre, J.-P.: Exposés de séminaires (1950–1999). Documents Mathématiques (Paris), vol. 1. Soc. Math. France, Paris (2001); 2nd edn., augmented, 2008

[GRSE] Colmez, P., Serre, J.-P. (eds.): Correspondance Grothendieck–Serre. Documents Mathématiques (Paris), vol. 2. Soc. Math. France, Paris (2001); bilingual edn., AMS, 2004

[CI] Garibaldi, S., Merkurjev, A., Serre, J.-P.: Cohomological Invariants in Galois Cohomology. Univ. Lect. Ser., vol. 28. Am. Math. Soc., Providence (2003)

# 2002—an Honorary Abel Prize to Atle Selberg

**Nils A. Baas**

When the Abel Prize was established in 2002 it was decided to award an honorary prize to the renowned Norwegian mathematician Atle Selberg in recognition of his status as one of the world's leading mathematicians. His contributions to mathematics are so deep and original that his name will always be an important part of the history of mathematics. His special field in mathematics was number theory in a broad sense.

Selberg was born on June 14, 1917 in Langesund, Norway. He grew up near Bergen and went to high school at Gjøvik. His father was a high school teacher with a doctoral degree in mathematics, and two of his older brothers—Henrik and Sigmund—became professors of mathematics in Norway. He was studying mathematics at the university level at the age of 12. When he was 15 he published a little note in *Norsk Matematisk Tidsskrift*.

He studied at the University of Oslo where he obtained the *cand. real.* degree in 1939, and in the autumn of 1943 he defended his thesis which was about the Riemann Hypothesis. At that time there was little numerical evidence supporting the Riemann Hypothesis. He got the idea of studying the zeros of the Riemann zeta-function as a kind of moment problem, and this led to his famous estimate of the number of zeros. From this it followed that a positive fraction of the zeros must lie on the critical line. This result led to great international recognition.

When Carl Ludwig Siegel, who had stayed in the USA, asked Harald Bohr what had happened in mathematics in Europe during the war, Bohr answered with one word: Selberg.

During the summer of 1946 Selberg realized that his work on the Riemann zeta function could be applied to estimate the number of primes in an interval. This was the beginning of the development leading to the famous Selberg sieve method.

Selberg's collected works were published in [1], and an extensive interview appeared in [2].

N.A. Baas (✉)
Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway
e-mail: baas@math.ntnu.no

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_2, © Springer-Verlag Berlin Heidelberg 2010

Photo by Herman Landshoff. Courtesy of the Institute for Advanced Study

In 1947 Selberg went to the Institute for Advanced Study in Princeton in USA where he continued the work on his sieve method. In the spring of 1948 he proved the Selberg Fundamental Formula which later in 1948 led to an elementary proof of the Prime Number Theorem. This was a sensation since even the possibility of an elementary proof had been questioned by G.H. Hardy and other mathematicians.

For these results he was awarded the Fields Medal in 1950—at the time the highest award in mathematics.

He became a permanent member of the Institute for Advanced Study in 1949 and a professor in 1951—a position he held until he retired in 1987.

In the early 1950s Selberg again produced a new and very deep result, namely what is now called the Selberg Trace Formula. Selberg was inspired by a paper by H. Maass on differential operators, and he realized that in this connection he could use some ideas from his Master Thesis. This result has had many important implications in mathematics and theoretical physics, but Selberg was never interested in the wide range of applications. In the Trace Formula Selberg combines many mathematical areas like automorphic forms, group representations, spectral theory and harmonic analysis in an intricate and profound manner. Selberg's Trace Formula is by many mathematicians considered as one of the most important mathematical result in the 20th century. His later works on automorphic forms led to the rigidity results of lattices in higher rank Lie groups.

In his later years he continued to work on his favourite subjects: sieve methods, zeta-functions and the Trace Formula. In 2003 Selberg was asked whether he thought the Riemann Hypothesis was correct. His response was: "If anything at all in our universe is correct, it has to be the Riemann Hypothesis, if for no other reasons, so for purely esthetical reasons." He always emphasized the importance of simplicity in mathematics and that “the simple ideas are the ones that will survive”. His style was to work alone at his own pace without interference from others.

In addition to the Fields Medal in 1950, Selberg received the Wolf Prize in 1986 and then in 2002 the honorary Abel Prize prior to the regular awards. He was also a member of numerous academies.

Atle Selberg was highly respected in the international mathematical community. He possessed a natural and impressive authority that made everyone listen to him with the greatest attention.

He loved his home country Norway and always spoke affectionately about the Norwegian nature, language and literature. In 1987 he was named Commander with Star of the Royal Norwegian Order of St. Olav.

Atle Selberg died on August 6, 2007 in his home in Princeton.

## References

1. Selberg, A.: Collected Papers, Vols. I and II. Springer, Berlin (1989 and 1991)
2. Baas, N.A., Skau, C.F.: The lord of the numbers, Atle Selberg. On his life and mathematics. Bull. Am. Math. Soc. **45**, 617–649 (2008)

# 2004

# Sir Michael Atiyah
# and
# Isadore M. Singer

ABEL
PRISEN

# Autobiography

**Sir Michael Atiyah**

I was born in London on $22^{\text{nd}}$ April 1929, but in fact I lived most of my childhood in the Middle East. My father was Lebanese but he had an English education, culminating in three years at Oxford University where he met my mother, who came from a Scottish family. Both my parents were from middle class professional families, one grandfather being a minister of the church in Yorkshire and the other a doctor in Khartoum.

My father worked as a civil servant in Khartoum until 1945 when we all moved permanently to England and my father became an author and was involved in representing the Palestinian cause. During the war, after elementary schooling in Khartoum, I went to Victoria College in Cairo and (subsequently) Alexandria. This was an English boarding school with a very cosmopolitan population. I remember priding myself on being able to count to 10 in a dozen different languages, a knowledge acquired from my fellow students.

At Victoria College I got a good basic education but had to adapt to being two years younger than most others in my class. I survived by helping bigger boys with their homework and so was protected by them from the inevitable bullying of a boarding school.

In my final year, at the age of 15, I focused on mathematics and chemistry but my attraction to colourful experiments in the laboratory in due course was subdued by the large tomes which we were expected to study. I found memorizing large bodies of factual information very boring and so I gravitated towards mathematics where only principles and basic ideas matter. From this point on it seemed clear that my future lay in mathematics.

There were vague allusions to some of my older Lebanese relatives having shown mathematical talent and one of my maternal uncles had been a brilliant classical scholar, ending up as a Fellow of an Oxford college. Classics and mathematics were

M. Atiyah (✉)
School of Mathematics, University of Edinburgh, King's Buildings, Mayfield Road,
Edinburgh EH9 3JZ, Scotland, UK
e-mail: M.Atiyah@ed.ac.uk

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_5, © Springer-Verlag Berlin Heidelberg 2010

Aged 2, known as "the abbott"

the two traditional subject studied by serious scholars in England in former years, so I may have inherited some mathematical potential from both sides of my family. My younger brother Patrick became a distinguished law professor and there is a clear affinity between the legal and mathematical minds, both requiring the ability to think clearly and precisely according to prescribed rules. One of my mathematical contemporaries, and a close friend, demonstrated this by entering the legal profession and ending up as Lord Chancellor of England.

In England I completed my school education by being sent to Manchester Grammar School (MGS), widely regarded as the leading school for mathematics in the country. Here I found that I had to work very hard to keep up with the class and the competition was stiff. We had an old-fashioned but inspiring teacher who had graduated from Oxford in 1912 and from him I acquired a love of projective geometry, with its elegant synthetic proofs, which has never left me. I became, and remained, primarily a geometer though that word has been reinterpreted in different ways at different levels. I was also introduced to Hamilton's work on quarternions, whose beauty fascinated me, and still does. I have been delighted in the way that quarternions have enjoyed a new lease of life in recent years, underlying many exciting developments.

At MGS the mathematics class, a small and highly selected group, were all trained for the Cambridge scholarship examinations. In due course all the class went on to Cambridge except for one who went instead to Oxford. The students were steered to different colleges, depending on their abilities, and I was one of the top three sent to Trinity College, home of Isaac Newton, James Clerk Maxwell, Bertrand Russell and other famous names.

Having got my Trinity scholarship in 1947 I had the choice of going straight to Cambridge or of postponing my entry until I had done my two-year stint of National Service. I chose the latter for the vague idealistic reason that I should do my duty and not try to escape it (as many of my contemporaries did) by running away to university with the hope of indefinite postponement of military service. One should remember that this was just two years after the end of the war and my age group was one of the first to have escaped the harsh reality of war-time service.

In fact my military career was something of an anti-climax. I served as a clerical officer in a very routine headquarters. There were some advantages—I kept myself physically fit and even cycled home every weekend. I met a wide cross-section of humanity at a formative stage of my life and I was removed from the competitive hot-house atmosphere of mathematical competition. In my spare time I read mathematics for my own enjoyment—I remember enjoying Hardy and Wright's book on Number Theory and I even read a few articles in the Encyclopaedia where I first encountered the ideas of group theory.

The normal length of National Service at the time was two years but my tutor at Cambridge managed to persuade the authorities that I should be allowed out a few months early to attend the "Long-Vacation Term". This was the period in the summer when those with extensive practical work, such as engineers, came up for additional courses. There was no requirement for mathematics students to spend the summer in Cambridge, but scholars of the college could opt to do so at little cost, and I remember that period as quite idyllic. I enjoyed the beauty of Cambridge, played tennis and studied on my own at a leisurely pace. It all helped make the readjustment to civilian life smooth and pleasant and it gave me a head start when the academic year began in earnest.

Trinity, because of its reputation, attracted a large number of exceptionally talented mathematics students. The competition among us was friendly but fierce and it was not clear until the end of the final year where one stood in the pecking order and what ones chances of a professional mathematical career would be. In fact I came top of the whole university in the crucial examination and this gave me the confidence to plan ahead.

By the time I started in Cambridge in 1949 I was 20. Instead of being two years younger than others I was two years older (although many others had also done National Service). The additional maturity was an advantage, even if I seemed to have lost two valuable years.

After my first degree I had to make a crucial choice in picking my research supervisor. Here I made the right decision, opting for Sir William Hodge as the most famous mathematics professor in the field of geometry. He it was who steered me into the area between algebraic and differential geometry where he had himself made his name with his famous theory of "Harmonic Integrals". Although the war had interfered with his career he was still in touch with new developments and he also had wide international contacts.

When I started research the geometrical world was undergoing a revolution based on the theory of sheaves and the topological underpinning of the theory of characteristic classes. Here the leading lights were the young post-war generation, just a few years older than me. Jean-Pierre Serre and Friedrich Hirzebruch were two of these whose influence on me was decisive, and I also met them very early.

My thesis grew out of this area and my own background in classical projective geometry. There were two parts, one dealing with vector bundles on algebraic curves (which later became a very popular topic) and the other, jointly with Hodge, on "Integrals of the Second Kind". This was a modern treatment of an old subject.

My thesis in 1954 earned me one of the highly-sought Research Fellowships at Trinity, which are safe predictors of future academic success. By this time I needed larger horizons and with Hodge's help and encouragement I got a Fellowship to go to the Institute for Advanced Study in Princeton. This I did in 1955 just after

Graduation, 1952, with friends. From left to right: James Mackay Lord Chancellor, MFA, Ian Macdonald FRS mathematician, John Polkinghorne FRS physicist and theologian, John Aitcheson professor of statistics

Trinity Fellowship, 1954

Wedding, 1955

marrying Lily Brown, a fellow student, who had come down from Edinburgh to do a Ph.D. under Mary Cartwright. She already had a university position in London but, in those days, it was customary to put the husband's career first, so she resigned her post and came with me to Princeton. A few years later such a sacrifice would not have been expected, though there is never an easy solution.

Princeton was a very important part of my mathematical development. Here I met many of those who would influence or collaborate with me in the future. In addition to Serre and Hirzebruch there were Kodaira and Spencer, of the older geometers, and Bott and Singer of the younger ones. In later times I came frequently to the Institute while on sabbatical leave and for three years, 1969–72, I was a professor on the Faculty. It was my second mathematical home.

My subsequent career oscillated between Cambridge, Oxford and Princeton. In 1957 I returned to Cambridge as a University Lecturer and in 1961 I moved to Oxford, first as a Reader and then from 1963–69 as Savilian Professor of Geometry. After my stay at Princeton from 1969–72 I returned to Oxford as a Royal Society Research Professor, staying in that post until in 1990 I moved back to Cambridge as Master of Trinity College.

If these were the universities where I had permanent positions an important part was also played by Harvard and Bonn. During my close collaborations with Bott and Singer I spent two sabbatical terms at Harvard and for around twenty-five years I used to go to Bonn for the annual Arbeitstagung. These were enormously exciting events with many new results being discussed and with a stellar cast of participants.

During the early years in Bonn much attention was paid to Hirzebruch's Riemann–Roch Theorem and its subsequent generalization by Grothendieck. Almost at the same time Bott made his famous discovery of the periodicity theorem in the classical groups. By being around at the right time, having the right friends and playing around with the formulae that emerged, I soon realised the close links between the work of Bott and Grothendieck. This led to new concrete results in algebraic topology which convinced me that it would be worth developing a topological version of Grothendieck's $K$-theory. This grew into a significant enterprise and it was natural that Hirzebruch should join me in developing it. He had more experience of Lie groups and their characteristic classes and his own earlier work tied in with my developing interest.

Over the subsequent years Hirzebruch and I wrote many joint papers on various aspects and applications of $K$-theory. It was an exciting collaboration from which I learnt much, not least in how to write papers and present lectures. He was, in effect, an elder brother who continued my education.

Some of the remarkable consequences of Hirzebruch's Riemann–Roch Theorem had been the integrality of various expressions in characteristic classes. A priori, since these formulae had denominators, the answers were rational numbers. In fact, under appropriate hypotheses, they turned out to be integers. For complex algebraic manifolds this followed from their interpretation as holomorphic Euler characteristics, a consequence of the Riemann–Roch Theorem. For other manifolds Hirzebruch had been able to deduce integrality by various topological tricks, but this seemed unsatisfactory. Topological $K$-theory gave a better explanation for these integrality theorems, closer to the analytic proofs derived from sheaf theory in the case of complex manifolds.

A particularly striking case was the fact that an expression called by Hirzebruch the $\hat{A}$-genus was an integer for spin-manifolds. It was the attempt to understand this fact that eventually led Singer and me to our index theorem. Because of the comparison with analytic methods on complex manifolds it was natural to ask if there was any analytical counterpart for spin-manifolds.

A key breakthrough came with the realization that Dirac had, thirty years before, introduced the famous differential operator that bears his name. Singer, with a better background in physics and differential geometry, saw that, on a spin-manifold, one could, using a Riemannian metric, define a Dirac operator acting naturally on spinor fields. From my apprenticeship with Hirzebruch I was familiar with the character formulae for the spin representations and so it was easy to see that the index of the Dirac operator should be equal to the mysterious $\hat{A}$-genus.

All this started while Singer was spending a sabbatical term with me in Oxford. We also had a brief visit from Stephen Smale, just returned from Moscow, who told us that Gel'fand had proposed the general problem of computing the index of any elliptic differential operator.

Because of our knowledge of $K$-theory we saw that the Dirac operator was in fact the primordial elliptic operator and that, in a sense, it generated all others. Thus a proof of the conjectured index formula for the Dirac operator would yield a formula for all elliptic operators.

In retrospect it might seem surprising that the Dirac operator had not been seriously studied by differential geometers before our time. Nowadays it all seems transparently obvious to a first-year graduate student. But the reasons for this neglect of the Dirac operator are not far to seek. First the Dirac equation in space-time is hyperbolic, not elliptic, second, spinors are mysterious objects and, unlike differential forms, have no natural geometric interpretation. The first point is analogous to the difference between Maxwell's equation and Hodge theory, and it took nearly a century for this gap to be bridged. The mysterious nature of spinors is an additional reason and so a delay of thirty years is quite modest.

The road that Singer and I took to arrive at the index theorem was that of a solution looking for a problem. We knew the precise shape of the answer, but the answer to what? Such an inverse approach may not be unique but it is certainly unusual.

Having formulated our index theorem, Singer and I had to search hard for a proof. Here our many good friends in the analytical community were invaluable, and we had to master many new techniques. This was easier for Singer since his background was in functional analysis.

Over the subsequent decades the index theorem in its various forms and generalizations occupied most of our efforts. A particularly interesting strand was the Lefschetz fixed point formula which I developed with Raoul Bott, and the fuller understanding of elliptic boundary value problems which was also joint with Bott. It was during this period that I spent two sabbatical terms at Harvard and I recall this as a particularly stimulating and fruitful time.

Another important extension of the index theorem which required the collective efforts of Bott, Singer, Patodi and myself was the local form of the index theorem and the contribution of the boundary arising from the $\eta$-invariant. This was a spectral invariant, analogous to the $L$-functions of number theory and originating in fact in a beautiful conjecture of Hirzebruch on the cusps of Hilbert modular surfaces. Most of this work was done while I was a professor at Princeton, with my collaborators as visitors.

Graeme Segal, who was one of my early research students, collaborated on the equivariant version of the index theorem as well as on aspects of $K$-theory.

In 1973 I returned to Oxford and while I had no formal teaching duties I acquired, over the years, a string of talented research students who also influenced my research. Nigel Hitchin moved to Princeton with me and then returned to Oxford and we collaborated on many topics. In 1973 I also met up again with my Cambridge contemporary Roger Penrose who had now become a theoretical physicist. We interacted fruitfully as soon as we realized that the complicated contour integrals arising in his twistor theory could be reinterpreted in terms of sheaf cohomology. This established a key bridge between his group and mine.

In due course this led on to the study of instantons and monopoles and opened doors to a wider physics community. It also led to the spectacular results of Simon

With I.M. Gelfand in Oxford, 1973

The Queen opening a new building at Trinity College, 1993

Donaldson on 4-dimensional geometry, one of the highlights of 20th century mathematics.

By the late 70's the interaction between geometry and physics had expanded considerably. The index theorem became standard form for physicists working in quantum field theory, and topology was increasingly recognized as an important ingredient. Magnetic monopoles were one manifestation of this, as I had learnt from Peter Goddard. I was fortunate to get to know Edward Witten fairly early in his career while he was a Junior Fellow at Harvard. For over thirty years he has been recognised as the driving force among theoretical physicists exploring the frontiers of their subject. I learned a great deal from him and he has provided mathematicians with an entrée to theoretical physics which is remarkable in its richness and sophistication. The influence of new ideas in quantum field theory and string theory has been widespread and much more may be in store.

For most of my working life I have held research posts which left me free to concentrate on my own work. So, later in life, I felt I had a duty to take on various administrative roles which are in any case more suited to grey hairs. I have presided over the London Mathematical Society, The Royal Society, Trinity College and the Newton Institute, while currently I am President of the Royal Society of Edinburgh. I have also served on the advisory committees of mathematical institutes in many countries.

In my youth I received the Fields Medal (1966) and in old age the Abel Prize (2004). I have been fortunate in many things, the quality of my collaboration and students, the support of many centres of research and a firm family base.

# Autobiography

**Isadore M. Singer**

My mother, Freda Rosemaity, and father, Simon Singer, were born in Poland. After World War I, they immigrated to Toronto, Canada where they met and were married. My father was a printer and my mother a seamstress. In the early 1920's they moved to Detroit, Michigan. I was born there in 1924.

My parents struggled through the depression. In the mid 1930's we were able to move from a poor neighborhood to a better one with a good school system. I was an all A student who did not find my courses challenging. In the summers I played baseball during the day and read novels at night.

The periodic table, explained by my high school chemistry teacher, awakened me to science. I was enthralled by its symmetry and began devouring popular books on chemistry and physics. In my senior year I became president of the Science Club and lectured on Relativity to club members.

I won a tuition scholarship to the University of Michigan and moved to Ann Arbor in September 1941. Three months later the United States entered World War II. I enlisted in the Signal Corps; it promised not to induct me into active service until I received my Bachelor of Science degree. Nevertheless I rushed through college, graduating in January 1944. I majored in physics but still regret that I did not take advantage of the superb mathematics faculty at Michigan.

The two physics courses that intrigued and puzzled me were Relativity and Quantum Mechanics. I decided I needed a better mathematical background and was determined to get it while on active duty. When the war ended I was in charge of a Signal Corps school in the mudflats of Luzon for the Phillippino Army. Fortunately, the University of Chicago offered correspondence courses in classical Differential Geometry and in Group Theory. My evenings were spent working problems while my comrades played poker. Mail call brought letters from my family and corrected problem sets from Chicago.

I.M. Singer (✉)
Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA
e-mail: ims@math.mit.edu

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_6, © Springer-Verlag Berlin Heidelberg 2010

Central High School Graduates, 1941: Shoshana, Ben, Yitzchak, Miriam, Lillit

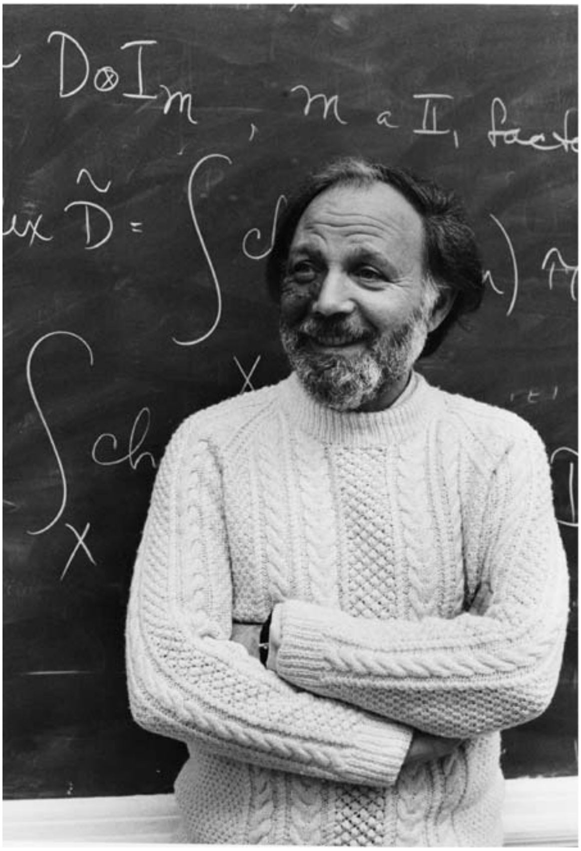
When I came home at the end of 1946, I was admitted to the graduate program in the mathematics department at the University of Chicago. I had planned to return to physics after a year, but mathematics was so elegant and exciting to learn that I stayed put. I specialized in analysis under the direction of I.E. Segal. S.S. Chern who joined the mathematics department during my last year, taught a fascinating course in Differential Geometry that described the Global Geometry of fiber bundles in terms of differential forms.

After receiving my Ph.D. in June 1950, I moved to MIT for two memorable years as an Instructor. We organized many seminars to learn the remarkable postwar developments in topology, analysis, and geometry. And it was there that I met Warren Ambrose who would become a longtime friend and collaborator. He asked me to explain Chern's course; we spent many nights drinking coffee and driving around Boston discussing geometry.

After assistant professorships at UCLA and Columbia University, I was a fellow at the Institute for Advanced Study in 1955, a very special year. I met and talked with mathematicians who were or became famous: Michael Atiyah, Raoul Bott, Fritz Hirzebruch, J.-P. Serre, and my dear graduate school friend Arnold Shapiro.

When I returned to MIT in 1956, Ambrose and I, with the help of our students, modernized and extended Chern's approach to Global Differential Geometry. We also revised many undergraduate courses to bring them up to date with the postwar advances in mathematics. It was an exciting time. Students caught our enthusiasm and instructors brought new perspectives from other places. My small office was crowded with people as I carried out simultaneous discussions on different topics, on holonomy, on commutative Banach algebras, on Eilenberg–MacLane spaces.

The Sloan Foundation awarded me a fellowship for the academic year 1961–62. I spent the fall on the Isle of Capri reviewing a manuscript on the Infinite Groups of E. Cartan, joint work with Shlomo Sternberg [21]. In December I called Michael Atiyah and asked if there was room for me at Oxford. He simply said, "Come", though he had only arrived a few months earlier himself. I came in January; Michael



At MIT

Singer and Atiyah

was a most hospitable colleague. Our collaboration of more than twenty years started then. We conjectured in the spring that the $\hat{A}$ genus of a spin manifold was the index of the Dirac operator [a generalization of Dirac's equation for the spinning

With family

electron] on that Riemannian spin manifold. We quickly extended the conjecture to include geometric elliptic differential operators and found a proof in the fall of 1962. Gelfand's insight[1] and the consequences of Seeley's MIT thesis[2] allowed us to further generalize our result, giving a topological formula for the index of any elliptic operator on a compact manifold [19]. The index theorem and its proof brought together analysis, geometry, and topology in unexpected ways. We extended it in different directions over a period of fifteen years.

My collaboration with Sir Michael has been a major part of my mathematical work. His expertise in topology and algebraic geometry and mine in analysis and differential geometry made a good match. Working with him was exciting and fun. All ideas were worth exploring at the blackboard, erased if nonsense, pursued intensively otherwise.

Our last collaboration (to date) in 1984 applied the families index theorem to the computation of chiral anomalies in gauge theory and string theory [67]. In the mid 1970's mathematicians realized that gauge fields in physics, which describe the interactions of fundamental particles, were the same as connections on principal bundles. Computing the dimension of the moduli of self dual gauge fields was an early application of the index theorem [54]. Our 1984 paper encouraged high energy theorists to apply the families' index theorem and its $K$-theory formulation to problems in string theory.

When I came to the University of California in Berkeley in 1977, I started a math/physics seminar. I wanted to know how to quantize gauge fields and why

[1] I.M. Gelfand, *On elliptic equations*, Russian Math. Surveys (1960) no 3, 113.

[2] R.T. Seeley, *Singular integral operators on compact manifolds*, Amer. J. Math. (81) 1999 658–690.

With I.M. Gelfand in Oxford, 1953

self dual fields were important in physics. Three gifted students, Dan Freed, Daniel Friedan, and John Lott helped me run the seminar. Daniel taught me much about quantum field theory on long walks at physics workshops. Orlando Alvarez joined the physics department in 1982 and became an enthusiastic participant in the seminar. We have been working together for almost fifteen years. M.J. Hopkins suggested that our last paper [102] will have some applications to elliptic cohomology. When I don't understand some physics, I call Orlando for an explanation.

I brought the seminar with me when I returned to MIT in 1984. It still flourishes. Last year was devoted to the paper by Kapustin and Witten[3] building a bridge between Electromagnetic Duality and the Geometric Langlands program in representation theory and number theory. That $S$-duality in string theory may impact number theory and/or vice versa is an exciting prospect.

Most of my academic life has been at MIT, a very fertile environment for me. My collaboration with colleagues here produced interesting mathematics—twenty papers starting with Warren Ambrose on Holonomy in 1953 [3]. My ongoing research with Richard Melrose and V. Mathai extends the index theorem to the case of twisted $K$-theory and the case where the manifold has no spin$_C$ structure [104, 106, 108].

I'm grateful to MIT for allowing me to teach what I want, the way I want, and for giving me ample time to do my own work. It has also been enthusiastic about my activities in Washington DC in support of science, a period that lasted thirty

[3] A. Kapustin and E. Witten, *Electric-magnetic duality and the geometric Langlands program*, arXiv:hep-th/0604151.

I.M. Singer (Courtesy of the MIT News Office)

years. Most interesting were membership in the White House Science Council during the Reagan administration and chairmanship of the National Academy of Sciences' "Committee on Science and Public Policy".

In 1988 J.P. Bourguignon and J.L. Gervais arranged for my appointment as Chair of Geometry and Physics, Foundation of France. I gave a weekly seminar on 'Introduction to String Theory for Mathematicians' at the École Normale and École Polytechnique. The mixture of mathematicians and physicists made a responsive audience. I fondly remember Claude Itzykson gently asked me leading questions. We became good friends. Laurent Baulieu was also a member of the audience. We talked, and soon wrote the first of several papers on cohomological quantum field theories [76, 79, 84]. I believe my seminar opened new lines of communication between different Laboratories in and around Paris.

In 1982, S.S. Chern, Calvin Moore, and I founded the Mathematical Sciences Research Institute (MSRI) in Berkeley, California, funded by the National Science Foundation. MSRI will soon celebrate its twenty-fifth anniversary with a conference reviewing its past successes. A major theme of the conference is the recognition and support of new mathematics and its applications.

Anticipating new directions is not easy. The growth and evolution of mathematics since I became a graduate student sixty years ago is astonishing. To have been a key participant in the development of index theory and its applications to physics is most gratifying. And beyond that I am fortunate to have experienced first hand the impact of ideas from high energy theoretical physics on many branches of mathematics.

# The Atiyah–Singer Index Theorem

**Nigel Hitchin**

## 1 Introduction

The Abel Prize citation for Michael Atiyah and Isadore Singer reads: "The Atiyah–Singer index theorem is one of the great landmarks of twentieth-century mathematics, influencing profoundly many of the most important later developments in topology, differential geometry and quantum field theory". This article is an attempt to describe the theorem, where it came from, its different manifestations and a collection of applications. It is clear from the citation that the theorem spans many areas. I have attempted to define in the text the most important concepts but inevitably a certain level of sophistication is needed to appreciate all of them. In the applications I have tried to indicate how one can use the theorem as a tool in a concrete fashion without necessarily retreating into the details of proof. This reflects my own appreciation of the theorem in its various forms as part of the user community. The vision and intuition that went into its proof is still a remarkable achievement and the Abel Prize is a true recognition of that fact.

## 2 Background

**2.1. The Index.** If $A : V \to V$ is a linear transformation of finite dimensional vector spaces, then as every undergraduate knows, $\dim\ker A + \dim\operatorname{im} A = \dim V$, so the dimension of the kernel of $A$ and its cokernel $V/\operatorname{im} A$ are the same. In infinite dimensions this is not true. Of course, if $V$ is a Hilbert space, then $\ker A$ may not be finite dimensional anyway, but we can restrict to the class of operators called *Fredholm* operators—bounded operators with finite-dimensional kernel, closed image and finite-dimensional cokernel. In this case the *index* is defined as

$$\operatorname{ind} A = \dim\ker A - \dim\operatorname{coker} A .$$

N. Hitchin (✉)
Mathematical Institute, Oxford University, 24–29 St. Giles, Oxford OX1 3LB, England, UK
e-mail: hitchin@maths.ox.ac.uk

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_7, © Springer-Verlag Berlin Heidelberg 2010

Using the Hilbert space structure to define the adjoint of $A$, an alternative expression for the index is

$$\operatorname{ind} A = \dim\ker A - \dim\ker A^*.$$

As an example take $V = \ell^2$, the space of square-summable sequences $(a_0, a_1, a_2 \dots)$. If $A$ is the left shift

$$A(a_0, a_1, a_2 \dots) = (a_1, a_2, a_3 \dots)$$

its image is $V$ and kernel is one-dimensional spanned by $(1, 0, 0, \dots)$ so its index is 1, and the index of $A^n$ is $n$. For the right shift we get index $-1$ with powers of it giving all negative integers. An important property of the index is that a continuous deformation of $A$ through Fredholm operators leaves it unchanged. The dimension of the kernel may jump up and down, but the index is the same and determines the different connected components of the space of all Fredholm operators.

The Atiyah–Singer index theorem concerns itself with calculating this index in the case of an elliptic operator on a differentiable manifold. With suitable boundary conditions and function spaces these are Fredholm. The challenge, to which the theorem provides an answer, is to compute this integer in terms of topological invariants of the manifold and operator.

**2.2. Riemann–Roch.** In many respects the index theorem and its uses is modelled on the Riemann–Roch theorem for compact Riemann surfaces. Riemann was attempting to understand abelian integrals and meromorphic functions on a Riemann surface described by identification of sides of a polygon.

A meromorphic function $f$ on a Riemann surface is determined up to a constant multiple by its zeros $p_i$ and poles $q_j$ which are written as

$$(f) = \sum_i m_i p_i - \sum_j n_j q_j$$

where the integer coefficients are the multiplicities. An arbitrary expression like this—a finite set of points with multiplicities—is called a divisor. Not all of them come from a meromorphic function, but given a divisor $D$ one considers the dimension $\ell(D)$ of the vector space of meromorphic functions $f$ such that all the coefficients of $(f) + D$ are non-negative. Riemann established an inequality

$$\ell(D) \geq d + 1 - g$$

where $d$ is the *degree* of the divisor $D$—the integer $\sum_i m_i - \sum_j n_j$—and $g$ is the genus of the Riemann surface. These numbers are topological invariants, unchanged under continuous deformation. In particular, the Euler characteristic $\chi$ of the surface is $2 - 2g$. So the inequality estimates something analytical by topological means.

Riemann's inequality shows that there are many meromorphic functions on a Riemann surface and helped him to prove that any two were algebraically related which showed that many features of abstract Riemann surfaces could be reduced to algebraic plane curves.

Roch was a student of Riemann who died at the age of 26 in the same year 1866 that Riemann died. He identified the difference in Riemann's inequality in terms of a similar object to $\ell(D)$. What is now called the Riemann–Roch formula for curves is

$$\ell(D) - \ell(K - D) = d + 1 - g$$

where $K$ is the divisor of a meromorphic differential—it could be the derivative of a function or more generally an abelian differential. The left hand side is a difference of two positive integers, each one of which depends in general on the divisor $D$, but the right hand side is a topological invariant. This is an example of the index theorem but it needs a more modern interpretation to make it so.

There is a differential operator here—the Cauchy–Riemann operator

$$\frac{\partial}{\partial \bar{z}} = \frac{\partial}{\partial x} + i\frac{\partial}{\partial y},$$

whose local solutions are holomorphic functions. The Riemann–Roch theorem is phrased above in terms of meromorphic functions—functions with singularities—but one gets around that by introducing the notion of a holomorphic line bundle associated to a divisor. So one considers complex-valued functions $f$ on the complement of the $p_i, q_j$ with specific behaviour near those points: near $p_i$ (with a local coordinate $z$ where $z = 0$ is $p_i$) the function $z^{-m_i} f(z, \bar{z})$ is differentiable and similarly at the points $q_j$, $z^{n_j} f(z, \bar{z})$ is differentiable. The space of all such functions is the infinite dimensional vector space of *sections* of a line bundle $L$. Because $z^{-m_i}$ and $z^{n_j}$ commute with the Cauchy-Riemann operator, there is a well-defined operator $\bar{\partial}$ on this space of sections. On suitable Sobolev spaces it defines a Fredholm operator, whose kernel has dimension $\ell(D)$. The dimension of its adjoint is $\ell(K - D)$, so the Riemann–Roch formula is

$$\dim \ker \bar{\partial} - \dim \operatorname{coker} \bar{\partial} = d + 1 - g.$$

Riemann's proof was heavily criticized because of its use of the physically inspired Dirichlet principle (not unlike some of the modern day incursions of physicists' thinking into pure mathematics) and a desire for more rigour propelled the theorem more into the algebraic domain after Riemann's death [22]. Nevertheless, its value was undeniable and indeed its use in the 19th century reflects many of the uses of the index theorem 100 years later: when $d$ is large enough, the right hand side is positive so the theorem asserts the existence of holomorphic sections of $L$, and if the degree of $K - D$ is less than zero, $\ell(K - D)$ vanishes and we get an exact formula. The theorem plus an additional vanishing theorem can be very powerful.

**2.3. The Beginning.** In 1961/62, Atiyah's first academic year in Oxford after moving from Cambridge, Singer decided to take a sabbatical from MIT. Remembering his friendship with Atiyah at the Institute for Advanced Study in Princeton in 1955, he had called to see if he could come on his own money and was of course welcome. Then, as Singer recalls, in January 1962 [28]:

> ... on my second day at the Maths Institute you walked up to the fourth floor office where I was warming myself by the electric heater. After the usual formalities, you asked "Why is the genus an integer for spin manifolds?" "What's up, Michael? You know the answer much better than I." "There's a deeper reason," you said.

And so began the Atiyah–Singer Index Theorem.

To understand the background to Atiyah's question, one has to understand the changes that had happened in geometry since Riemann's time. Riemann had invented the concept of a manifold, a higher dimensional version of a surface, but it took lifetimes for the idea to be properly understood. By 1962 however these were familiar objects and their structure was being analyzed from many different points of view. The rapid development of topology in the first half of the 20th century had provided a sophisticated algebraic setting for many of the invariants—much of it encoded in the cohomology ring. And de Rham's cohomology theory gave an analytical hold on this, representing cohomology classes by exterior differential forms. Then Hodge had showed that, with a Riemannian metric on the manifold, one could find a unique harmonic form in each cohomology class. When applied to algebraic surfaces it showed that holomorphic differentials were closed and provided a link to topology which had held up the further development of the Riemann–Roch theorem since the 19th century.

In the immediate postwar period the notion of a vector bundle—a family of vector spaces parametrized by the manifold—and in particular the tangent bundle, had come into play and the characteristic cohomology classes named after Pontryagin and Chern were the subject of great study. Most notably, Friedrich Hirzebruch, who had also been in Princeton in 1955 had come up with a means of describing the *signature* of a manifold in terms of particular combinations of Pontryagin classes.

**2.4. The Signature.** In the present context it is convenient to use de Rham cohomology to define the signature. The cohomology group $H^p(M,\mathbf{R})$ consists of the quotient space of the space of differential $p$-forms $\alpha$ such that $d\alpha = 0$ (*closed* forms) modulo those for which $\alpha = d\beta$ for some $(p-1)$-form $\beta$ (*exact* forms). Here a $p$-form is written in local coordinates as

$$\alpha = \sum_{i_1<i_2<\cdots<i_p} a_{i_1 i_2 \ldots i_p}(x) dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_p}$$

and then

$$d\alpha = \sum_{j, i_1<i_2<\cdots<i_p} \frac{\partial a_{i_1 i_2 \ldots i_p}}{\partial x_j} dx_j \wedge dx_{i_1} \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_p}.$$

For a compact orientable manifold of dimension $n$, $H^n(M,\mathbf{R})$ is one-dimensional and the exterior product of forms defines a dual pairing between $H^p(M,\mathbf{R})$

and $H^{n-p}(M, \mathbf{R})$:

$$([\alpha], [\beta]) = \int_M \alpha \wedge \beta. \tag{2.1}$$

If we introduce a Riemannian metric $g_{ij}$ then there is a naturally defined volume form $\omega = \sqrt{\det g_{ij}} dx_1 \wedge \cdots \wedge dx_n$ and an inner product on forms. The *Hodge star operator* is the linear map $* : \Omega^p \to \Omega^{n-p}$ from the space of all $p$-forms to $(n-p)$-forms with the property that at each point

$$(\alpha, \beta)\omega = \alpha \wedge *\beta.$$

We have $*^2 = (-1)^{p(n-p)}$ when $*$ acts on $p$-forms.

The formal adjoint $d^*$ of $d$ satisfies the condition

$$\int_M (d\alpha, \beta)\omega = \int_M (\alpha, d^*\beta)\omega$$

and can be written using the star operator as

$$d^* = (-1)^{np+n+1} * d * . \tag{2.2}$$

The Hodge theorem says that in each cohomology class there is a unique representative form which satisfies $d\alpha = 0$ and $d^*\alpha = 0$.

Hodge theory immediately implies that the pairing (2.1) is non-degenerate since if $([\alpha], [\beta]) = 0$ for all $[\beta]$ then in particular we can take $\beta = *\alpha$ where $\alpha$ is harmonic (from (2.2) $\beta$ is closed). This implies that

$$0 = \int_M (\alpha, *^2\alpha) = \pm \int_M (\alpha, \alpha)$$

and so $\alpha = 0$.

If $n = 2m$ is even, we obtain a nondegenerate bilinear form on $H^m(M, \mathbf{R})$. This is symplectic when $m$ is odd (since odd forms anticommute) and symmetric when $m$ is even. In the latter case there is a basis in which the matrix is diagonal with $p$ positive entries and $q$ negative ones. The *signature* $\tau(M)$ of the manifold $M$ is defined to be the integer $p - q$.

The signature has some very natural properties: for a product $\tau(M \times N) = \tau(M)\tau(N)$ and for a change of orientation (replacing the volume form $\omega$ by $-\omega$) we clearly get $-\tau(M)$. Most importantly, if $M$ is the boundary of another oriented manifold of one dimension higher, then $\tau(M) = 0$.

Now in the mid 1950s René Thom had developed the theory of cobordism—considering equivalence classes of closed manifolds under the relation that two $n$-dimensional manifolds are equivalent if there is an $(n+1)$-dimensional manifold whose boundary has two components $M$ and $N$. One then introduces a ring structure on the equivalence classes using the two operations of product and disjoint union. Introducing orientations, one writes $[-M]$ for the class $M$ with opposite orientation and then $[M] + [-M] = 0$ by consideration of the cylinder $M \times [0, 1]$.

By the remarks above, $\tau$ defines a homomorphism from the cobordism ring to the integers.

Over the rational numbers, Thom determined this ring: he showed that a class is determined by the Pontryagin numbers of the tangent bundle and also gave generators. The Pontryagin numbers play an important role in the index theorem, so let us look more closely at these.

The basic topological invariant of a *surface* is its Euler characteristic—for a triangulation it is $V - E + F$ where $V, E, F$ are the numbers of vertices, edges and faces respectively. It is also quite familiar that this number can be calculated by the number of zeros of a vector field, counted with sign and multiplicity. For every oriented vector bundle of rank two on a manifold $M$, there is a cohomology class in $H^2(M, \mathbf{Z})$ called the *Euler class* which when evaluated on a surface in $M$ counts the number of zeros of a section. In the case of the tangent bundle of a surface a section is a vector field and so this number is the Euler characteristic. For a Riemann surface, a holomorphic line bundle is a complex vector bundle of rank one which can be thought of as a real rank two bundle and this number is the degree which appears on the right hand side of the Riemann–Roch formula. The Euler class changes sign if we change the orientation of the bundle—evaluating it on a surface necessitates also a choice or orientation on the surface so the integer (for example the Euler characteristic itself) does not depend on orientation.

Now suppose a rank four bundle $E$ is a direct sum $E_1 \oplus E_2$ of two rank two bundles. We have two Euler classes $e_1, e_2 \in H^2(M, \mathbf{Z})$. The signs are indeterminate as is their order, but the class $e_1^2 + e_2^2 \in H^4(M, \mathbf{Z})$ is insensitive to this. If we have an overall orientation on $E$ then there is another class $e_1 e_2 \in H^4(M, \mathbf{Z})$ which is well-defined. The first example is called the (first) Pontryagin class of $E$. It makes sense even if $E$ is not a direct sum (by a trick called the splitting principle one can pass to another space over which the bundle does split as a sum without losing information; so most calculations can be performed by imagining that the bundle does split). For a vector bundle of rank $2m$ we define the Pontryagin class $p_k \in H^{4k}(M, \mathbf{Z})$ using the $k$th elementary symmetric function in $e_1^2, e_2^2, \dots, e_k^2$. There is also, with an orientation, a class $e_1 e_2 \dots e_k \in H^{2k}(M, \mathbf{Z})$ called the Euler class. A *Pontryagin number* of a compact manifold of dimension $4k$ is obtained by taking the Pontryagin classes of the tangent bundle and evaluating a degree $4k$ class

$$p_{i_1} p_{i_2} \cdots p_{i_n}$$

where $(i_1, i_2, \dots, i_n)$ is a partition of $k$. The Pontryagin number is an integer.

Let $Q(x)$ be a power series with $Q(0) = 1$ and rational coefficients then the product

$$Q(e_1^2) Q(e_2^2) \dots Q(e_k^2)$$

is a series whose terms are of degree $0, 4, 8, \dots$ and each term of a given degree is a symmetric polynomial in the $e_i^2$ and hence a polynomial in Pontryagin classes. The degree $4k$ component can be evaluated on a manifold of dimension $4k$ to give a rational combination of Pontryagin numbers. This number $q(M)$ satisfies the condition that $q(M \times N) = q(M)q(N)$. Moreover since the Pontryagin classes themselves are independent of orientation, when we evaluate on the manifold we need a choice of orientation, so the numbers $q(M)$ change sign if we change the orientation. Thom's result that the cobordism ring is determined rationally by the Pontryagin numbers means that $q$ defines a ring homomorphism to $\mathbf{Q}$.

Hirzebruch's task was to find the function $Q$ for which this homomorphism is the signature, and he discovered that it was

$$Q(x) = \frac{\sqrt{x}}{\tanh\sqrt{x}}.$$

Expanding this in symmetric polynomials and substituting for the Pontryagin classes gives

$$L = 1 + \frac{1}{3}p_1 + \frac{1}{45}(7p_2 - p_1^2) + \cdots$$

so Hirzebruch's theorem says that the signature of a $4k$-dimensional manifold is the Pontryagin number of degree $4k$ in this expansion. Because of the cobordism invariance, all one has to do is to check both sides on generators of the cobordism ring, which are standard well-known manifolds.

**2.5. Hirzebruch–Riemann–Roch.** Hirzebruch followed up his work on the signature with a version of the Riemann–Roch theorem for algebraic varieties of arbitrary dimension, not just Riemann's original one-dimensional case. This work appeared in the highly influential book of 1956 [24]. Whereas the original theorem related the dimensions of two vector spaces of holomorphic sections of a line bundle, the higher-dimensional case involves more complicated objects. These are most conveniently described by the Dolbeault approach.

On a complex manifold $M$ of complex dimension $n$, one can consider not only the exterior derivative $d : \Omega^p \to \Omega^{p+1}$ but also an analogue on $(0, p)$ forms: a $(0, p)$-form is locally written in complex coordinates $z_i$ as

$$\alpha = \sum_{i_1 < i_2 < \cdots < i_p} a_{i_1 i_2 \ldots i_p}(x) d\bar{z}_{i_1} \wedge d\bar{z}_{i_2} \wedge \cdots \wedge d\bar{z}_{i_p}$$

and then

$$\bar{\partial}\alpha = \sum_{j, i_1 < i_2 < \cdots < i_p} \frac{\partial a_{i_1 i_2 \ldots i_p}}{\partial \bar{z}_j} d\bar{z}_j \wedge d\bar{z}_{i_1} \wedge d\bar{z}_{i_2} \wedge \cdots \wedge d\bar{z}_{i_p}.$$

By analogy with de Rham cohomology one defines the Dolbeault cohomology group $H^{0,p}$ as the kernel of $\bar{\partial}$ on $(0, p)$-forms modulo the image of $\bar{\partial}$ on $(0, p-1)$-forms. One can also incorporate a holomorphic vector bundle $E$ and consider the associated operator on forms with values in $E$. When $p = 0$, the kernel of $\bar{\partial}$ is simply the space of global holomorphic sections of $E$. Hirzebruch gave a formula for the alternating sum

$$\sum_{p=0}^{n}(-1)^p \dim H^{0,p}(M,E)$$

in terms of topological invariants which are the complex analogues of Pontryagin classes, called *Chern classes.*

A complex line bundle $L$ defines a class $c(L) \in H^2(M,\mathbf{Z})$—considered as a real oriented rank two bundle this is the Euler class $e$. If a complex vector bundle $E$ of rank $m$ splits as a sum of line bundles $L_i$, then the elementary symmetric functions in $c(L_i)$ define the Chern classes $c_k(E) \in H^{2k}(M,\mathbf{Z})$, and we can form Chern numbers instead of Pontryagin numbers by evaluating products in degree $2n$ on the manifold $M$, and use power series $Q(x)$. Hirzebruch developed a method closely related to his proof of the signature theorem which allowed him to find the right combination of Chern numbers to give the value of the alternating sum, by evaluating both sides on some standard examples. In the case without the vector bundle $E$ his formula is

$$\sum_{p=0}^{n}(-1)^p \dim H^{0,p}(M) = \mathrm{td}(TM)[M]$$

where td is the *Todd polynomial* defined by evaluating the Chern numbers of the tangent bundle generated by the polynomial

$$Q(x) = \frac{x}{1-e^{-x}}$$

which gives

$$\mathrm{td} = 1 + \frac{1}{2}c_1 + \frac{1}{12}(c_1^2 + c_2) + \cdots .$$

With a vector bundle $E$, one introduces another polynomial in symmetric functions, the *Chern character*, defined by

$$\mathrm{ch}(E) = \sum_i e^{c(L_i)} = \mathrm{rk}\, E + c_1(E) + \frac{1}{2}(c_1^2 - 2c_2)(E) + \cdots$$

and then the general Riemann–Roch formula is

$$\sum_{p=0}^{n}(-1)^p \dim H^{0,p}(M,E) = \mathrm{ch}(E)\,\mathrm{td}(TM)[M].$$

When $M$ is one-dimensional, a Riemann surface, and $E$ is rank one, a line bundle $L$, the right hand side is

$$(1 + c(L))\left(1 + \frac{1}{2}c_1(TM)\right)[M] = \deg L + \frac{1}{2}(2-2g) = d + 1 - g$$

which is the right hand side of the classical Riemann–Roch theorem. The left hand side is

$$\dim H^{0,0}(M, L) - \dim H^{0,1}(M, L).$$

To link this with the traditional formulation one needs the Serre duality theorem which in general asserts that

$$H^{0,p}(M, E)^* \cong H^{0,n-p}(M, E^* \otimes K)$$

where $K$ is the canonical bundle of holomorphic $n$-forms. The struggles of the 19th century geometers to obtain a Riemann–Roch theorem for algebraic surfaces may well have been reflected by the inability to come to terms with higher cohomology—Serre duality does not convert the $H^{0,1}$ term into anything more amenable.

Hirzebruch showed that the Todd polynomial was closely related to Pontryagin classes. He introduced the $\hat{A}$ polynomials in Pontryagin classes defined by the power series

$$Q(x) = \frac{\sqrt{x}/2}{\sinh(\sqrt{x}/2)}$$

giving

$$\hat{A} = 1 - \frac{1}{24}p_1 + \frac{1}{2^7 3^2 5}(-4p_2 + 7p_1^2) + \cdots \tag{2.3}$$

and he showed that

$$\mathrm{td}(TM) = e^{c_1(TM)/2}\hat{A}(TM). \tag{2.4}$$

All of these formulae provoke an obvious question—the right hand side is a *rational* combination of Pontryagin numbers and so a priori doesn't give an integer, though the interpretation of the left hand side—either the signature or an alternating sum of dimensions—clearly is. Algebraic topologists were explaining this by quite sophisticated methods in the early 1960s, and the question that Atiyah asked Singer in January 1962 was motivated by one of these, relating precisely to the $\hat{A}$ polynomial above. On an algebraic variety with $c_1(TM) = 0$ the Hirzebruch–Riemann–Roch formula shows that $\hat{A}(TM)[M]$ is an integer. Hirzebruch had also shown that a weaker result holds. The mod 2 reduction of $c_1(TM) \in H^2(M, \mathbf{Z})$ is an invariant called the second Stiefel–Whitney class $w_2(TM) \in H^2(M, \mathbf{Z}_2)$ which exists on any manifold, complex or not. It was known that for any oriented manifold with $w_2 = 0$, the $\hat{A}$-genus was an integer. Why? As Atiyah commented: "We had the answer: we didn't know what the problem was" [2].

**2.6. The Dirac Operator.** By March 1962 Atiyah and Singer had found a candidate for the problem—determine the index of the Dirac operator. In a way they had rediscovered this operator since physicists were already familiar with it, but there was a huge difference between the Euclidean signature of Riemannian geometry which was needed here and the Lorentzian signature of relativity.

The construction of the Dirac operator begins with the Clifford algebra: the algebra generated by the vectors in a real vector space $V$ with positive definite inner product and the single relation

$$v^2 = -(v,v)1.$$

When $V = \mathbf{R}$ this gives the complex numbers, when $V = \mathbf{R}^2$ the quaternions. If $e_1, \dots, e_n$ is an orthonormal basis in $\mathbf{R}^n$ then the Dirac operator

$$D = \sum_{i=1}^{n} e_i \frac{\partial}{\partial x_i}$$

has the property

$$D^2 = -\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}.$$

What does $D$ act on? The complexified Clifford algebra is isomorphic to a matrix algebra in even dimensions $n = 2m$ and so $D$ acts on functions with values in this $2^m$ dimensional space of *spinors*.

On a Riemannian manifold each tangent space has an inner product and so one gets a bundle of Clifford algebras. But finding a global rank $2^m$ bundle $S$ on which this acts requires a topological constraint, satisfied if the second Stiefel–Whitney class $w_2(X) = 0$. This condition is therefore necessary for the existence of a global Dirac operator. A manifold satisfying this condition is called a *spin manifold*. If the manifold is not simply-connected there is a finite choice to be made of spin structures and Dirac operators.

Atiyah and Singer had been made aware of some of the results of Gelfand and his coworkers on the homotopy invariance of indices of elliptic boundary value problems [19, 20] and so a potential link with differential operators was already in the air. They conjectured that the $\hat{A}$ polynomial should give the index of a Dirac operator, to explain the integrality puzzle. The Dirac operator on its own is self-adjoint and so has zero index but the bundle $S$ can be broken up further according to the two half-spin representations. The volume form $\omega$ represents in the Clifford algebra an element such that $\omega^2 = (-1)^m$ and its two eigenspaces define a decomposition $S = S_+ \oplus S_-$ of the spinor bundle. For a vector $v \in V$, $v\omega = -\omega v$ in even dimensions, so the Dirac operator can be viewed as

$$D : C^\infty(S_+) \to C^\infty(S_-).$$

The index of this operator should be the $\hat{A}$-genus.

The other integrality questions were also amenable to an index interpretation. In fact the simplest is the Euler characteristic itself:

$$\chi(M) = \sum_{p=0}^{n} (-1)^p \dim H^p(M) = \dim H^{even} - \dim H^{odd}.$$

The operator

$$d + d^* \colon \Omega^{even} \to \Omega^{odd}$$

has by Hodge theory a kernel isomorphic to $H^{even}$ and a cokernel isomorphic to $H^{odd}$ so the index is the Euler characteristic.

Similarly

$$\sum_{p=0}^{n} (-1)^p \dim H^{0,p}(M, E) = \dim H^{0,even}(M, E) - \dim H^{0,odd}(M, E)$$

is the index of

$$\bar{\partial} + \bar{\partial}^* : \Omega^{0,even}(E) \to \Omega^{0,odd}(E)$$

and the Hirzebruch–Riemann–Roch theorem was explained as an index. Moreover, the theorem now held for an arbitrary complex manifold.

Finally Hirzebruch's signature theorem could be explained by considering, on a manifold of dimension $n = 2m$, the involution $\alpha : \Omega^p \to \Omega^{2m-p}$ defined by

$$\alpha = i^{p(p+1)} * .$$

This operator anticommutes with $d + d^*$ and, just as in the case of the Dirac operator, if we consider the ± eigenspaces we get an operator

$$d + d^* : \Omega_+ \to \Omega_-$$

and an index

$$\dim H_+(M) - \dim H_-(M).$$

When $p < m$, $\alpha$ preserves $H^p(M) \oplus H^{2m-p}(M)$ and identifying $H^p(M)$ and $H^{2m-p}(M)$ using $*$, $\alpha$ is

$$\pm \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

so that the number of $+1$ and $-1$ eigenvalues is the same. Hence the index is the difference of these dimensions just for $p = m$, in the middle degree $H^m(M)$, which is the signature.

The integrality results could be explained by indices of operators, but the passage from the operator to the Pontryagin number required a theorem which became the index theorem. A proof was finally completed in the autumn of 1962 as Atiyah visited Harvard, and the results were presented in a seminar run by Bott and Singer, subsequently expounded in detail in the Princeton seminar [27]. The proof was based on Hirzebruch's 1953 proof of the signature theorem: the index theorem can be reduced to the evaluation of special cases which generate the cobordism classes. For the index problem one has to describe its change under cobordism and this required an extension of elliptic boundary value techniques to singular integral operators.

## 3 The Integer Index

**3.1. Formulation of the Theorem.** There are many variants of the Atiyah–Singer index theorem. We begin with the standard version, where the index is simply an integer. We start with a linear elliptic differential operator

$$D : C^{\infty}(V_{+}) \to C^{\infty}(V_{-})$$

on a compact manifold $M$ with vector bundles $V_{+},\ V_{-}$. Ellipticity is a property of the highest order term, the principal *symbol.*

A differential operator of order $r$ is locally expressible (using multi-index notation) as

$$Df = \sum_{|\alpha| \leq r} a_{\alpha} D^{\alpha} f$$

on vector-valued functions $f$. This means we have locally trivialized the bundle to write a section as the function $f$. In another trivialization $f$ is changed to $Pf$ for some invertible matrix-valued function $P$. The coefficients $a_{\alpha}$ of order less that $r$ transform involving a derivative of $P$ but the highest order terms do not. The symbol is invariantly defined as a section of

$$S^{r}T \otimes \mathrm{Hom}(V_{+}, V_{-})$$

where $S^{r}T$ is the bundle of symmetric rank $r$ tensors. We can also think of it as a section $\sigma$ of $p^{*}\,\mathrm{Hom}(V_{+}, V_{-})$ on the cotangent bundle $p : T^{*}M \to M$, homogeneous of degree $r$ along the fibres. The operator is *elliptic* if $\sigma(\xi)$ is invertible for $\xi \neq 0$. For example, the Dirac operator is elliptic because its symbol is the Clifford product $\sigma(\xi)\psi = \xi \cdot \psi$ and

$$\sigma(\xi)^{2} \cdot \psi = -(\xi, \xi)\psi$$

where $(\xi, \xi) \geq 0$ is the Riemannian inner product on one-forms.

Ellipticity depends only on the principal symbol as does the index, which also is homotopy invariant. To obtain a topological object from it, we use the fact that it gives an isomorphism $V_{+} \cong V_{-}$ outside the zero-section of $T^{*}M$ and defines a class in the cohomology with compact supports $H^{*}_{c}(T^{*}M)$.

We explain further: if $X$ is a non-compact manifold, we can consider the de Rham cohomology of differential forms with compact supports. So, for example although $H^{n}(\mathbf{R}^{n}) = 0$ for $n > 0$, $H^{n}_{c}(\mathbf{R}^{n}) \cong \mathbf{R}$, represented by $\varphi dx_{1} \wedge dx_{2} \wedge \cdots \wedge dx_{n}$ where $\varphi \geq 0$ vanishes outside a compact set.

A *connection* $A$ on a vector bundle $V$ is a first order differential operator $d_{A}$ : $C^{\infty}(V) \to C^{\infty}(V \otimes T^{*})$ whose symbol is $\sigma(\xi)s = s \otimes \xi$. It extends to an "exterior derivative" operator on $p$-forms with values in $V$:

$$d_{A} : \Omega^{p}(V) \to \Omega^{p+1}(V)$$

but $d_{A}^{2}$ is no longer zero, and instead defines a *curvature* form $F_{A} \in \Omega^{2}(\mathrm{End}(V))$. We meet here the Chern–Weil theory, descendant of the classical Gauss–Bonnet theorem. In de Rham cohomology, the Pontryagin and Chern classes are represented by closed differential forms obtained by evaluating certain polynomials of matrices on the curvature form.

In our case we take the Chern character of $E$, which is represented in de Rham cohomology by the closed form

$$\sum_k \frac{1}{(2\pi i)^k} \operatorname{tr}(F_A^k).$$

In the situation above we choose connections $A, B$ on $V_+$ and $V_-$ and pull them back to $T^*M$. Outside some neighbourhood of the zero section of $T^*M$, $\sigma^*B = A + a$ where $a \in \Omega^1(\operatorname{End} V)$. Extending $a$ to the whole of $T^*M$ we have connections $A_+, A_-$ on $p^*V_+, p^*V_-$ which are equivalent by the isomorphism $\sigma$ outside of a compact set, hence $\operatorname{ch}(p^*V_+) - \operatorname{ch}(p^*V_-)$ defines a compactly supported cohomology class in $H_c^*(T^*M)$. This is the topological data derived from the operator $D$.

Now the cotangent bundle $T^*M$ is naturally a symplectic manifold. Using a Riemannian metric on $M$ its tangent bundle becomes a complex vector bundle and we can take its Todd class $\operatorname{td}(T)$. Then the Atiyah–Singer index theorem can be formulated as:

**Theorem 3.1** *Let $D : C^\infty(V_+) \to C^\infty(V_-)$ be an elliptic operator on a compact manifold. Then*

$$\operatorname{ind} D = \dim\ker D - \dim\operatorname{coker} D = \int_{T^*M} (\operatorname{ch}(V_+) - \operatorname{ch}(V_-))\operatorname{td}(T).$$

In many applications it is the Dirac operator coupled to a vector bundle which is the relevant operator. For the Dirac operator alone the formula is

$$\operatorname{ind} D = \hat{A}(TM)[M].$$

If $E$ is an auxiliary bundle with connection $d_A$ we define the Dirac operator with coefficient bundle $E$ as the composition of

$$\nabla \otimes 1 + 1 \otimes d_A : C^\infty(S_+ \otimes E) \to C^\infty(S_+ \otimes E \otimes T^*)$$

(where $\nabla$ is the Levi-Civita connection) with the Clifford multiplication map

$$S_+ \otimes E \otimes T^* \to S_- \otimes E$$

defined by $\varphi \otimes \xi \otimes e \mapsto \xi\varphi \otimes e$. The index formula for this operator is

$$\operatorname{ind} D_E = \operatorname{ch}(E)\hat{A}(TM)[M].$$

The example of the elliptic operator

$$d + d^* : \Omega^{even} \to \Omega^{odd}$$

arising from the de Rham complex

$$\cdots \xrightarrow{d} \Omega^p \xrightarrow{d} \Omega^{p+1} \xrightarrow{d} \cdots$$

gives rise to the associated idea of an *elliptic complex*

$$\cdots C^\infty(V^{p-1}) \xrightarrow{D_{p-1}} C^\infty(V^p) \xrightarrow{D_p} C^\infty(V^{p+1}) \xrightarrow{D_{p+1}} \cdots$$

where ellipticity means that if $\xi \neq 0$ and $\sigma_p$ is the symbol of $D_p$ then $\sigma_p(\xi)v = 0$ implies that $v = \sigma_{p-1}(\xi)w$. By choosing inner products on the $V^p$, the elliptic complex generates an elliptic operator

$$D + D^* \colon C^\infty(V^{even}) \to C^\infty(V^{odd})$$

just as in the de Rham complex and the index is the alternating sum of the dimensions of the cohomology spaces.

**3.2. Integrality Theorems.** The index theorem provided an explanation for many of the previously known and slightly puzzling integrality theorems, especially when the central role of the Dirac operator was appreciated. In the first place, there was the link with Riemann–Roch. If the vector space $V$ has a Hermitian inner product then the space $\oplus_0^m \Lambda^{0,p} V$ is a module over the Clifford algebra: given $v \in V$, take its (0, 1) part $v^{0,1}$ and define

$$v\varphi = \frac{1}{\sqrt{2}}(e(v) - e(v)^*)\varphi$$

where $e(v)$ is the exterior product by $v^{0,1}$ and $e(v)^*$ its adjoint. Taking $V$ to be the tangent space to a complex manifold with a Hermitian metric, this shows that the symbol of the Dirac operator on the bundle $\oplus_0^m \Lambda^{0,p} T^*$ is essentially the same as the symbol of $\bar\partial + \bar\partial^*$ and so they have the same index.

But then we don't need complex coordinates to get integrality of the Todd genus $\mathrm{td}(TM)[M]$ because it is just the index of a Dirac operator. This means (as was known) that the Todd genus of an *almost complex* manifold is an integer.

The space of $(0, p)$ forms considered as a Clifford module is not the standard one—for that we need to choose a square root $L$ of the canonical bundle, a line bundle $L$ such that $L^2 \cong K$. The topological obstruction to finding that is the condition

$$c_1(T) \bmod 2 = c_1(K) \bmod 2 = w_2 = 0$$

and then the standard Dirac operator is equivalent to the $\bar\partial + \bar\partial^*$ operator with values in the line bundle $L = K^{1/2}$. This explains Hirzebruch's link between the Todd polynomials and the $\hat{A}$ polynomials (2.4).

The Dirac operator when the coefficient bundle is the spin bundle itself is the $d + d^*$ bundle on exterior forms. In this case the signature theorem and the formula for the Euler characteristic can both be seen to be examples of index theorems for Dirac operators.

A particularly nice example of integrality is a proof of Rochlin's 1952 result that if a compact 4-manifold is oriented and has $w_2 = 0$, then its signature is divisible by 16. The index theory proof goes as follows: since $w_2 = 0$ the manifold has a Dirac operator whose index from (2.3) is

$$\hat{A} = \left(1 - \frac{1}{24}p_1 + \cdots\right)[M] = -\frac{1}{24}p_1(TM)[M].$$

But in four dimensions we have the isomorphism $Spin(4) \cong Sp(1) \times Sp(1)$ where $Sp(1)$ is the group of unit quaternions. This means that the Dirac operator is quaternionic hence its kernel and the kernel of its adjoint are quaternionic vector spaces. In particular as complex vector spaces they are even-dimensional. It follows that $p_1(TM)[M]/24$ is an *even* integer. On the other hand the signature is

$$\tau(M) = \left(1 + \frac{1}{3}p_1 + \cdots\right)[M] = \frac{1}{3}p_1(TM)[M]$$

and so is divisible by $48/3 = 16$.

**3.3. Positive Scalar Curvature.** We noted in Sect. 2.6 that in $\mathbf{R}^n$ if

$$D = \sum_{i=1}^{n} e_i \frac{\partial}{\partial x_i}$$

then

$$D^2 = -\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$$

so that the Dirac operator is a sort of square root of the Laplacian. On a curved Riemannian manifold, the corresponding formula involves an extra zero-order term:

$$D^2 = \nabla^*\nabla + \frac{1}{4}R$$

where $R$ is the *scalar curvature* of the metric and $\nabla^*$ is the formal adjoint of the covariant derivative $\nabla\colon C^\infty(S) \to C^\infty(S \otimes T^*)$. The operator $\nabla^*\nabla$ is non-negative.

This formula, originally due to Schrödinger in 1932 in the Lorentzian setting, was introduced in the Riemannian case by Lichnerowicz in 1963 [26] as an early application of the index theorem. If $D\varphi = 0$ then taking global inner products

$$0 = (D^2\varphi, \varphi) = (\nabla^*\nabla\varphi, \varphi) + \frac{1}{4}(R\varphi, \varphi).$$

Thus if $R > 0$, since $\nabla^*\nabla$ is non-negative, we must have $\varphi = 0$, whether $\varphi$ is a section of $S_+$ or $S_-$. It follows that the index of the Dirac operator $\hat{A}(TM)[M] = 0$.

The simplest example of this is the four dimensional K3 surface (for example the quartic $z_0^4+z_1^4+z_2^4+z_3^4=0$ in $\mathbf{C}P^3$) which has $w_2=0$ and signature $-16$ hence $\hat{A}(TM)[M]=2$. This manifold cannot admit a metric of positive scalar curvature. In fact Yau's proof of the Calabi conjecture showed that it does admit a metric of zero scalar curvature.

**3.4. Gauge-Theoretic Moduli Spaces.** One of the most useful applications of the index theorem is to the calculation of the dimension of certain moduli spaces of solutions to nonlinear equations. These include the anti-self-dual Yang–Mills equations on a four-manifold, the Seiberg–Witten equations, equations for Higgs bundles, magnetic monopoles, pseudo-holomorphic curves and others. These are all nonlinear partial differential equations whose linearization can be made elliptic. The index theorem then produces the expected dimension of the space of solutions and with a little further information Banach space implicit function theorems give manifold structures on the moduli space. As an example we shall take the anti-self-duality equations on a compact 4-manifold $M$ (see for example [17]).

Let $E$ be a complex vector bundle of rank $r$ over $M$ with a Hermitian metric. A connection $A$ on $E$ has a curvature form $F_A \in \Omega^2(\operatorname{End} E)$. The connection is called *anti-self-dual* if

$$*F_A=-F_A.$$

If $A$ preserves the Hermitian metric, then $F_A$ is skew-adjoint.

A gauge transformation $g$ is a unitary automorphism of $E$ and $g$ acts on the curvature by conjugation hence preserves the notion of anti-self-duality. We want to understand the moduli space—the space of all such connections modulo gauge equivalence. Elliptic operators appear when we look at the linearization of the problem. The derivative at $A$ of a one-parameter family of connections is given by $\dot{A} \in \Omega^1(\operatorname{End} E)$. If this arises from a one-parameter family of gauge equivalent connections then $\dot{A}=d_A\psi$ where $\psi \in \Omega^0(\operatorname{End} E)$. The derivative of the curvature form at $A$ is $d_A\dot{A} \in \Omega^2(\operatorname{End} E)$ so if this arises from a one-parameter family of *anti-self-dual* connections then $*d_A\dot{A}=-d_A\dot{A}$, or

$$d_A^+\dot{A}=0 \in \Omega^2_+(\operatorname{End} E)$$

where the $+$ subscript indicates orthogonal projection onto the $+1$ eigenspace of $*$ on $\Omega^2(\operatorname{End} E)$.

We thus have a sequence of first order operators

$$\Omega^0(\operatorname{End} E) \xrightarrow{d_A} \Omega^1(\operatorname{End} E) \xrightarrow{d_A^+} \Omega^2_+(\operatorname{End} E).$$

Moreover since $d_A^2=F_A$ and $*F_A=-F_A$, it follows that $d_A^+d_A=0$ so this is a complex.

The linearization of our problem (the tangent space of the moduli space) is thus the kernel of $d_A^+$ (the infinitesimal deformations of anti-self-dual connections) modulo the image of $d_A$ (the deformations arising from gauge transformations). Harmonic theory for this complex tells us that this is isomorphic to the kernel of

$$d_A^* + d_A^+ : \Omega^1(\mathrm{End}\, E) \to \Omega^0(\mathrm{End}\, E) \oplus \Omega^2_+(\mathrm{End}\, E)$$

and this is an elliptic operator.

We now calculate the index of this. This is a practical example which demonstrates how indices can be computed without going into the proof of the theorem. Firstly note that it is in fact a Dirac operator, with coefficient bundle $S_+ \otimes \mathrm{End}\, E$:

$$D\colon C^\infty(S_- \otimes S_+ \otimes \mathrm{End}\, E) \to C^\infty(S_+ \otimes S_+ \otimes \mathrm{End}\, E)$$

so its index is

$$\begin{aligned} &-\mathrm{ch}(S_+ \otimes \mathrm{End}\, E)\hat{A}(TM)[M] \\ &\quad = -\big(r^2 + \mathrm{ch}_2(\mathrm{End}\, E) + \cdots\big)\big(2 + \mathrm{ch}_2(S_+) + \cdots\big)\left(1 - \frac{1}{24}p_1 + \cdots\right) \\ &\quad = -2\,\mathrm{ch}_2(\mathrm{End}\, E) - r^2\left(-\frac{1}{12}p_1 + ch_2(S_+)\right). \end{aligned}$$

The last term is $r^2$ times the index of

$$d^* + d^+ \colon \Omega^1 \to \Omega^0 \oplus \Omega^2_+$$

which by Hodge theory is

$$b_1 - 1 - b_2^+ = \frac{1}{2}(2b_1 - 2 - b_2^+ - b_2^- - b_2^+ + b_2^-) = \frac{1}{2}\big(-\chi(M) - \tau(M)\big).$$

To calculate the first term note that

$$\begin{aligned} \mathrm{ch}(\mathrm{End}\, E) &= \mathrm{ch}(E^* \otimes E) = \mathrm{ch}(E^*)\,\mathrm{ch}(E) \\ &= \left(r - c_1 + \frac{1}{2}(c_1^2 - 2c_2) + \cdots\right)\left(r + c_1 + \frac{1}{2}(c_1^2 - 2c_2) + \cdots\right) \end{aligned}$$

so that

$$\mathrm{ch}_2(\mathrm{End}\, E) = \big[r(c_1^2 - 2c_2) - c_1^2\big](E)[M].$$

The final index is

$$-2\big[r(c_1^2 - 2c_2) - c_1^2\big](E)[M] - \frac{1}{2}r^2(\chi + \tau).$$

In the case of $M = S^4$, we have $b_2 = 0$, so $c_1(E) = 0$ and $\tau = 0$, $\chi = 2$ and the formula becomes $4rc_2(E) - r^2$.

This is just the index, but, as shown by Freed and Uhlenbeck, a deformation of the metric will make $d_A^* + d_A^+$ surjective. The kernel of $d_A$ is the space of covariant constant infinitesimal gauge transformations which for an irreducible connection is just the scalars, so the final dimension of the moduli space is the index plus one. In the case $r = 2$ on $S^4$ this gives the $8k - 3$ of [12]. The global study of the moduli space on a general four-manifold is of course the content of Donaldson theory.

The study of instantons on $S^4$ began with this index theoretical approach [12]. The subsequent ADHM description in terms of matrices also uses the index theorem [13], in this case for the Dirac operator

$$D : C^\infty(S_- \otimes E) \to C^\infty(S_+ \otimes E)$$

where the index is $-\operatorname{ch}(E)\hat{A}(TM)(M) = c_2(E) = k$. In this case the Lichnerowicz formula for $S_+ \otimes E$ is still just the scalar curvature term $R/4$ (positive for $S^4$) because the anti-self-dual curvature $F_A$ acts trivially on the spinors in $S_+$. Hence the index gives the actual dimension of the kernel of $D$. By stereographic projection these become $\mathcal{L}^2$ sections on $\mathbf{R}^4$ and the $k \times k$ ADHM matrices are the global inner products

$$(x_i\varphi_\alpha, \varphi_\beta)$$

for an orthonormal basis $\varphi_1, \ldots, \varphi_k$ of solutions to $D\varphi = 0$.

# 4 The Equivariant Index

**4.1. K-Theory.** The first proof of the index theorem was not flexible enough to support the myriad applications which Atiyah and Singer had in mind—in particular to study group actions and families. Hand in hand with the development of the index theorem came the development of K-theory—a generalized cohomology theory which was naturally adapted to considering families of vector spaces and not just the integer which is their dimension. A series of papers [3–7] in *Annals of Mathematics* in the period 1968–71 became the definitive version of the index theorem and K-theory was the basic tool this time. The model for the new proof was not Hirzebruch's use of cobordism, but Grothendieck's version of the Riemann–Roch theorem, replacing the algebraic K-theory groups defined in terms of coherent sheaves by the topological theory developed by Atiyah and Hirzebruch.

In particular, suppose one has a group $G$ acting on the manifold together with an action on vector bundles $V_+$ and $V_-$, preserving an elliptic operator

$$D : C^\infty(V_+) \to C^\infty(V_-).$$

Then the kernel and cokernel of $D$ are representation spaces of $G$. The plain integer index of $D$ simply gives the differences of the degrees of these representations and no further information, whereas the natural object to consider is a formal difference of representation spaces. This lies in the *Grothendieck group* $R(G)$ generated by representations of $G$—two formal differences $U - V$, $U' - V'$ define the same element in $R(G)$ if there is a representation $W$ such that $U \oplus V' \oplus W$ is isomorphic to $U' \oplus V \oplus W$.

The same construction applied to the natural numbers gives the ring of integers, and applied to isomorphism classes of vector bundles on a space $X$ (using the direct sum operation of vector bundles) defines $K(X)$, topological K-theory. This is also a ring under the operation of tensor product. If $X$ is a point, then a vector bundle is a vector space and then $U - V \mapsto \dim U - \dim V$ gives an isomorphism from $K(pt.)$ to the integers.

In general, since $\mathrm{ch}(E \oplus F) = \mathrm{ch}(E) + \mathrm{ch}(F)$ and $\mathrm{ch}(E \otimes F) = \mathrm{ch}(E)\,\mathrm{ch}(F)$ the Chern character defines a homomorphism from $K(X)$ to the cohomology ring of $X$ but only with rational coefficients because of the denominators in

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

In pre-index theory days, integrality issues were being addressed by using K-theory, so it is not surprising that it became the natural setting for much of the development of the theory.

On a non-compact space $X$ one can define K-theory with compact supports $K_c(X)$ in terms of vector bundles together with an isomorphism outside a compact set. Then an open inclusion $i : U \subset X$ induces a natural map

$$i_! : K_c(U) \to K_c(X).$$

Clearly the symbol $\sigma$ of an elliptic operator defines an isomorphism between $p^*V_+$ and $p^*V_-$ over $T^*M$ outside the zero section so we immediately obtain a symbol class

$$[\sigma] \in K_c(T^*M).$$

For the usual index theorem we need to extract an integer out of this, and to do this one uses a consequence of the Bott periodicity theorem concerning the homotopy groups of unitary groups $U(n)$, for $n$ large. This has a rather different manifestation in K-theory—if $V$ is a complex vector bundle over a space $X$, and $v \in V$ then the exterior product $e(v)$ defines a complex

$$\cdots \xrightarrow{e(v)} \Lambda^p V \xrightarrow{e(v)} \Lambda^{p+1} V \xrightarrow{e(v)} \cdots$$

exact outside the zero section $v = 0$. This gives a class $\lambda_V \in K_c(V)$ and by tensoring with vector bundles pulled back from $X$ it defines a homomorphism

$$\varphi\colon\ K(X) \to K_c(V).$$

This is an isomorphism, called the Thom isomorphism.

To use this to define a topological invariant from the symbol class $[\sigma]$ one embeds the manifold $M$ in a large Euclidean space $\mathbf{R}^m$. If $N$ is the normal bundle of $M$ in $\mathbf{R}^m$ then $N$ can be identified with a tubular neighbourhood of $M$ in $\mathbf{R}^m$—an open subset—and hence we get an open embedding of $TN$ in $T\mathbf{R}^m$. Using the induced metric, we identify $T^*M$ and the tangent bundle $TM$ and define a complex structure on the tangent bundle of $TM$, of $T\mathbf{R}^m$ and hence the normal bundle of $TM$ in $T\mathbf{R}^m$. So we have the Thom isomorphism

$$\varphi\colon K_c(TM) \to K_c(TN),$$

the open inclusion $i : TN \subset T\mathbf{R}^m$ giving

$$i_!\colon K_c(TN) \to K_c(T\mathbf{R}^m),$$

and the Thom isomorphism theorem again giving

$$\varphi\colon K(pt.) \to K_c(\mathbf{C}^m) = K_c(T\mathbf{R}^m).$$

Using all these maps together with $TM \cong T^*M$ we get a homomorphism, called the *topological index*

$$\mathrm{t{-}ind}\colon K_c(T^*M) \to \mathbf{Z}.$$

In this formulation, the integer index theorem says that the topological index of the symbol class $[\sigma] \in K_c(T^*M)$ is the analytical index of the elliptic operator $D$.

The above set-up is perfectly adapted to deal with the equivariant case, replacing the embedding $M \subset \mathbf{R}^m$ by an equivariant embedding in a representation space, possible by the Peter–Weyl theorem. The result then is that the symbol class lies in the equivariant K-theory $K_G(T^*M)$ and the topological index is in $K_G(pt.)$ which is the Grothendieck group for the representations of $G$.

**4.2. Fixed Point Theorems.** One of the remarkable features of the equivariant index is the ability to calculate it from data at the fixed point set of elements of the group $G$. The classical example of this is the Lefschetz fixed point theorem of 1926. Here, given a map $f\colon X \to X$, the Lefschetz number

$$\sum (-1)^p \,\mathrm{tr}(f_*|H^p(X))$$

is calculated in terms of the sum of certain indices at the fixed points. In good situations one gets a count of the number of fixed points.

Clearly if $G$ is a group of isometries of a Riemannian metric, it acts naturally on $p$-forms and commutes with the operator $d + d^*$. The alternating sum in the Lefschetz formula for $g \in G$ is then the *character* at $g$ of the equivariant index of

$$d + d^*\colon \Omega^{even} \to \Omega^{odd}.$$

However, there are more possibilities by taking different elliptic operators or complexes—the Dolbeault complex or the Dirac operator for example. Many of the consequences of this localization were spelled out in the papers [8, 9] of Atiyah and Bott. We mention here two applications. Both involve the action on the elliptic complex

$$\cdots \to \Omega^{0,p}(E) \xrightarrow{\bar\partial} \Omega^{0,p+1}(E) \xrightarrow{\bar\partial} \cdots$$

for a holomorphic vector bundle on which the group $G$ acts. In this case the concrete Lefschetz formula says that the alternating sum of the characters at $g$ of the action on $H^{0,p}(M, E)$ is, if $g$ has isolated fixed points,

$$\sum_{g(x)=x} \frac{\operatorname{tr}\varphi_x}{\det(1 - dg_x)} \tag{4.1}$$

where $\varphi_x \colon E_x \to E_x$ is the action on $E$ at the fixed point and $dg_x$ the action on the holomorphic tangent space at $x$.

For the first example we take a compact simply-connected simple Lie group $G$ and its maximal torus $T$. The flag manifold $G/T$ can be given the structure of a homogeneous Kähler manifold, and a homomorphism $T \to S^1$ given by the weight $\lambda$ defines a circle bundle whose associated homogeneous line bundle $L$ over $G/T$ is holomorphic. With appropriate choices the cohomology spaces $H^{0,p}(G/T, L)$ vanish for $p > 0$ and for $p = 0$ one gets an irreducible representation of $G$. This is the content of the Borel–Weil theorem—a direct realization of a representation given the maximal weight which is the initial character.

Because of the vanishing of the higher cohomology, the equivariant index for the $\bar\partial + \bar\partial^*$ operator is precisely the representation on $H^{0,0}(G/T, L)$. The index theorem gives a formula for its character.

To do this choose $g \in T$ such that the closure of the group it generates is $T$ itself. If $ghT = hT$, then $h^{-1}gh \in T$ and so taking powers and closure, $h^{-1}Th = T$. The fixed points are therefore in one-to-one correspondence with $N(T)/T$ where $N(T)$ is the normalizer of $T$, and this quotient is the Weyl group $W$, a finite group. The index theorem thus gives an expression for the character in terms of a sum over the Weyl group.

The tangent space at a point can be identified with $\mathfrak{g}/\mathfrak{t}$ which is a sum of 2-dimensional root spaces defined by the positive roots $\alpha_1, \dots, \alpha_k$, so the denominator in the formula (4.1) at $w \in W$ is a product

$$\prod_1^k w(1 - e^{\alpha_i}).$$

The numerator is the action on the line bundle $L$ which is $w(e^\lambda)$ where $\lambda$ is the maximal weight. Since $G$ is simply connected, half the sum of the positive roots $\rho$ is a weight and $w(e^\rho) = \pm e^\rho = \operatorname{sgn}(w)e^\rho$, so we get the more familiar form for the Weyl character formula:

$$\frac{1}{e^\rho \prod_1^k (1 - e^{\alpha_i})} \sum_{w \in W} \operatorname{sgn}(w) w(e^{\lambda+\rho}).$$

A wide range of other examples can be obtained by considering the action of finite groups on algebraic surfaces. The index theorem then operates as a machine for producing identities in number theory. The article [25] gives a good survey of this. Here we take the signature operator and the holomorphic action of $g \in G$, a finite group, on an algebraic surface $M$. The fixed points are either isolated points $x \in M$ or algebraic curves $Y$. In the first case $g$ acts on the tangent space as $(e^{i\alpha} \oplus e^{i\beta})$ and in the second on the normal bundle by $e^{i\theta}$. The fixed point contribution to the equivariant index is then

$$-\cot\frac{\alpha}{2}\cot\frac{\beta}{2} \quad \text{or} \quad Y \cdot Y \operatorname{cosec}^2\frac{\theta}{2}$$

where $Y \cdot Y$ is the self-intersection number of the curve $Y$ (the degree of the normal bundle).

This is the fixed point contribution which the index theorem for the signature operator relates to the action of $G$ on $H^2(M, \mathbf{R})$. A simple example is the case where $M \subset \mathbf{C}P^3$ is the algebraic surface with equation in homogeneous coordinates

$$z_0^n + z_1^n + z_2^n + z_3^n = 0$$

and $g$ is the action of the $n$th root of unity $\omega$:

$$g(z_0, z_1, \dots, z_3) = (\omega^{-1}z_0, z_1, \dots, z_3).$$

The fixed point set of $g^k$ for $k \neq 0$ is the plane section $Y$ with equation $z_0 = 0$, which has self-intersection $n$. The equivariant index theorem then gives

$$\chi(g^k) = n \operatorname{cosec}^2\frac{\pi k}{n}.$$

Now averaging the character over the group gives the degree of the trivial representation, which is

$$\frac{1}{n}\left(\sum_{k=1}^{n-1}\chi(g^k) + \tau(M)\right).$$

On the other hand, the invariant part of $H^2(M, \mathbf{R})$ can be interpreted as the cohomology of the quotient $M/G$ and the index is the signature of the invariant quadratic form on this. In our case, the quotient is $\mathbf{C}P^2$ which has signature 1. So we get

$$1 = \frac{1}{n}\left(\sum_{k=1}^{n-1}\chi(g^k) + \tau(M)\right) = \sum_{k=1}^{n-1}\operatorname{cosec}^2\frac{\pi k}{n} + \frac{1}{n}\tau(M).$$

This offers two interpretations—the topologist would calculate the Chern classes of the surface $M$ as $c_1 = (4-n)h, c_2 = (6-4n+n^2)h^2$ and use the signature theorem to give $\tau(M) = (c_1^2 - 2c_2)/3 = n(4-n^2)/3$. Then the equivariant index theorem gives the number theoretic identity:

$$\sum_{k=1}^{n-1}\operatorname{cosec}^2\frac{\pi k}{n} = \frac{n^2-1}{3}.$$

The number theorist would give an elementary proof of this and derive the signature of $M$.

This is a simple example, but when isolated points occur as fixed points, the contributions there have appeared in the classical literature as *Dedekind symbols* and results of Rademacher and Mordell can be obtained this way as well as many more identities.

**4.3. Rigidity Theorems.** The equivariant index theorem together with its fixed-point formulation enables a character to be evaluated in two different ways. As above, this provides a route to identities which can also be proved by other means with the appropriate skills. The same idea, however, also leads to some remarkable results about the degree of symmetry, or rather lack of it, of certain manifolds. For example, Atiyah and Hirzebruch showed in [10] that a manifold $M^{4k}$ with $w_2 = 0$ and $\hat{A}[M] \neq 0$ admits no non-trivial circle action. Recall from Sect. 3.3 that the same hypotheses prohibit the existence of a metric of positive scalar curvature, so this result is identifying a range of manifolds at the opposite extreme from those with positive curvature and homogeneous, like spheres. The method has since been radically extended through ideas of Witten [16].

An $S^1$-invariant elliptic operator $D$ is called *rigid* if the equivariant index is trivial as a representation, in other words if the non-trivial representations occur with the same multiplicity in the kernel and cokernel of $D$. For the $d + d^*$ operator or the signature operator this is clear since a diffeomorphism in the circle action is connected to the identity and so acts trivially on cohomology by homotopy invariance. The Hodge theorem then implies it acts trivially on the kernel of $d + d^*$. It also becomes transparent when using the equivariant index formula for isolated fixed points. For

$$d + d^* : \Omega^{even} \to \Omega^{odd}$$

the fixed point contribution is just $+1$ (this is the original Lefschetz fixed point formula). For the signature operator on $M^{4k}$ the contribution is

$$\prod_1^{2k} \frac{1 + e^{im_j\theta}}{1 - e^{im_j\theta}}$$

where the tangent space breaks up into 2-dimensional pieces on which the circle acts as $e^{im_j\theta}$.

The equivariant index theorem says that the character for a generic $g$ in the circle is the sum over fixed points

$$\sum_{g(x)=x} \prod_1^{2k} \frac{1 + e^{im_j\theta}}{1 - e^{im_j\theta}}.$$

But the character is a finite sum of terms of the form $e^{im\theta}$ so we get an identity of meromorphic functions

$$\sum_{i=-N}^{N} a_i z^i = \sum_{g(x)=x} \prod_1^{2k} \frac{1+z^{m_j}}{1-z^{m_j}}.$$

But the left hand side is a finite Laurent series and so has poles only at $z=0$ whereas the right hand side has poles on the unit circle. Both sides must therefore equal a constant function.

When $w_2=0$ we have a Dirac operator and here the contribution is

$$\prod_1^{2k} \frac{z^{m_j/2}}{1-z^{m_j}}$$

(the factor $1/2$ involves a slightly subtle lifting of the circle action to the spin structure). In this case the right hand side vanishes when $z=0$ which shows that the equivariant index is not just constant but is zero, which is the theorem of Atiyah and Hirzebruch in this case. The general proof involves the consideration of fixed point sets of arbitrary dimension.

Witten's extension of this (given mathematical proof in [16]) introduces Dirac operators whose coefficient bundles are derived from the tangent bundle in a specific manner. If $S^kT$ and $\Lambda^kT$ denote the symmetric and exterior powers of the tangent bundle, one writes

$$S_t = \sum_0^\infty t^k S^k T \qquad \Lambda_t = \sum_0^\infty t^k \Lambda^k T$$

and

$$R_q = \sum_0^\infty q^n R_n = \bigotimes_{n=1}^\infty \Lambda_{q^n} \bigotimes_{m=1}^\infty S_{q^m}$$

and

$$R'_q = \sum_0^\infty q^{n/2} R'_n = \bigotimes_{n=1}^\infty \Lambda_{q^{(2n+1)/2}} \bigotimes_{m=1}^\infty S_{q^m}$$

The theorem is then that the Dirac operator with coefficient bundle $R'_n$ and the signature operator with bundle $R_n$ are rigid.

While the mathematical proof of this is a consequence of the equivariant index theorem, understanding the reasons for this rigidity is more demanding than finding alternative proofs for number-theoretical identities. In Witten's derivation using the loop space of $M$, the modular property of the polynomial $Q(x)$ which generates these particular genera plays a fundamental role which has not yet been fully absorbed into the mathematics.

## 5 The mod 2 Index

**5.1. Real K-Theory.** There was one development of index theory which, in Atiyah's words, "could easily have been missed out at the first step ... and trodden underfoot in the stampede later on" [2]. This involves a mod 2 invariant which is not definable in terms of cohomology classes. In fact there are manifolds homotopically equivalent to spheres for which it is non-zero.

It is K-theory which reveals its presence. We mentioned in Sect. 4.1 that Bott periodicity lay behind the isomorphism $K_c(\mathbf{C}^n) \cong \mathbf{Z}$. The point is that a vector bundle which is trivial outside a compact set in $\mathbf{C}^n$ is the same as a bundle on the sphere $S^{2n}$ with a trivialization on a ball, and this itself can be described by a map from the equatorial $S^{2n-1}$ to $GL(m, \mathbf{C})$ relating the two trivializations. Since the definition of the K-group involves isomorphism classes of bundles and adding on trivial bundles, $K_c(\mathbf{C}^n)$ is isomorphic to homotopy classes of maps $S^{2n-1} \to U(m)$ for $m$ large. This is the homotopy group $\pi_{2n-1}(U(m))$ which Bott showed was infinite cyclic (and $\pi_{2n}(U(m)) = 0$).

If we consider instead real vector bundles on a space $X$ then one defines K-groups $KO(X)$ and then $KO_c(\mathbf{R}^n)$ is defined by the homotopy group $\pi_{n-1}(O(m))$. In this case Bott had showed that $\pi_{n-1}(O(m))$ is eightfold periodic in $n$ and $\pi_i(O(m))$ for $i = 0, 1, 2, \dots, 7$ mod 8 is

$$\mathbf{Z}_2 \quad \mathbf{Z}_2 \quad 0 \quad \mathbf{Z} \quad 0 \quad 0 \quad 0 \quad \mathbf{Z}.$$

The K-theory definition of the symbol class above then shows that a real elliptic operator $D$ in dimensions $4k$ has an integer invariant (which up to a multiple is just the ordinary index) but in dimensions $8k+1$ and $8k+2$ there is an invariant in $\mathbf{Z}_2$, and the challenge is to interpret this analytically.

The answer lies with skew-adjoint real Fredholm operators. By skew-adjointness the kernel and cokernel have the same dimension so the ordinary index vanishes, but the dimension of the kernel modulo 2 is a deformation invariant. For elliptic differential operators we can see simple examples of this even on the circle (dimension 1 mod 8!). There are two real line bundles over the circle—the trivial bundle and the Möbius band. The operator

$$D = \frac{d}{d\theta}$$

with periodic boundary conditions is a skew-adjoint operator on the trivial bundle and has a 1-dimensional kernel, the constants. With anti-periodic conditions the line bundle is the Möbius band and the operator has kernel zero. Slightly more interesting is a skew adjoint third order operator on the trivial bundle over $S^1$:

$$D = \frac{d^3}{d\theta^3} + 2u\frac{d}{d\theta} + u'.$$

This has a one or three-dimensional space of solutions but never a two-dimensional one.

The mod 2 index theorem in Part V of the papers of Atiyah and Singer derives the dimension modulo 2 of the kernel of a real skew-adjoint elliptic operator in terms of its symbol class.

In $8k+1$-dimensions the spin representation is real and the Dirac operator skew-adjoint so this gives a $\mathbf{Z}_2$-invariant for spin manifolds in this dimension. In $8k+2$ dimensions the Dirac operator can be considered as a skew-adjoint complex antilinear operator and the mod 2 invariant is the complex dimension mod 2 of the kernel.

**5.2. Theta Characteristics.** Applications of the mod 2 theorem are not so numerous as the other versions but in [11] Atiyah revisits some classical results on Riemann surfaces with this new tool. On a Riemann surface $M$ a spin-structure is defined by a holomorphic square root $K^{1/2}$ of the canonical bundle and the Dirac operator is

$$\bar{\partial}: C^{\infty}(K^{1/2}) \to C^{\infty}(K^{1/2}\bar{K}).$$

A metric identifies $\bar{K}$ with $K^*$ and so

$$K^{1/2}\bar{K} \cong K^{-1/2} \cong \bar{K}^{1/2}.$$

This is the identification of the two spinor bundles which makes the Dirac operator antilinear.

The null space of $\bar{\partial}$ is the space of holomorphic sections of $K^{1/2}$ and the $\mathbf{Z}_2$-invariant is the dimension modulo 2 of this space.

Any two square roots differ by a holomorphic line bundle $L$ such that $L^2$ is holomorphically trivial so there are $2^{2g}$ such choices where $g$ is the genus of $M$. These are the different spin structures referred to in Sect. 2.6. The invariant is zero for $2^{g-1}(2^g+1)$ of these and 1 for the other $2^{g-1}(2^g-1)$. Thus for an elliptic curve where $K$ is trivial there is $2^{1-1}(2^1-1)=1$ square root, the trivial one, which has an odd (namely one) number of sections. For a quartic plane curve there are $2^{3-1}(2^3-1)=28$ square roots with an odd (one again) number of sections and these are the celebrated 28 bitangents. Classically these square roots are known as theta characteristics and they have odd or even type, but the 19th century arguments involved the zeros of the Riemann theta function whereas the derivation in [11] uses elementary properties of the group $KO(M)$. One amusing result is that a real quartic with no real points has exactly four real bitangents.

The earlier example of $d/d\theta$ on the circle is precisely the Dirac operator and the two real line bundles two spin structures where the dimension mod 2 of the kernel distinguishes them.

**5.3. Positive Scalar Curvature.** The Lichnerowicz vanishing theorem in Sect. 3.3 tells us that if a spin manifold of whatever dimension admits a metric of positive scalar curvature then the kernel of the Dirac operator is zero. In dimensions $8k+1, 8k+2$ this means the mod 2 index vanishes. Surprisingly there are exotic spheres (manifolds homotopically equivalent to a sphere) in these dimensions for which the invariant is known to be non-zero. We deduce that these spheres cannot have metrics of positive scalar curvature.

Perhaps the best way to describe the invariant is to say that it is a spin cobordism invariant. Thom's cobordism theory can be modified to put extra structure on the manifolds in question. It was oriented cobordism that gave the Pontryagin numbers that Hirzebruch used for his proof of the signature theorem. These exotic spheres have the property that, while they are themselves spin manifolds, and while they bound an oriented manifold, they do not bound a spin manifold.

The spin-cobordism interpretation of obstructions to positive scalar curvature led Gromov and Lawson [23] to ask whether these were the only obstructions. They showed that any manifold which can be obtained from one of positive scalar curvature by performing surgery in codimension greater than 2 also carries a metric of positive scalar curvature. Surgery is a process which operates within a cobordism class and as a consequence they deduced that any compact simply connected spin manifold of dimension $\geq 5$ which is spin-cobordant to a manifold of positive scalar curvature also carries a metric of positive scalar curvature. Somewhat later, by using techniques from stable homotopy theory to analyze spin cobordism in more detail, Stolz [29] succeeded in proving that these invariants—the $\hat{A}$ genus in dimension $4k$ and the mod 2 index in dimensions $8k+1, 8k+2$ are the only obstructions for a simply-connected manifold to have positive scalar curvature.

## 6 The Index for Families

**6.1. Fredholm Operators.** The ordinary integer index is a deformation invariant of a Fredholm operator. This was the starting point for the index theorem. But the space $\mathcal{F}$ of all Fredholm operators on a fixed Hilbert space contains more topological information than that. A family of Fredholm operators parametrized by a space $X$ is a continuous map $f\colon X \to \mathcal{F}$ and we can consider the set of homotopy classes $[X, \mathcal{F}]$ of such maps. A theorem of Atiyah (and independently K. Jänich) says that

$$[X, \mathcal{F}] \cong K(X).$$

When $X$ is a point $[pt., \mathcal{F}] \cong K(pt.) \cong \mathbf{Z}$ is the set of components so we learn that the components of $\mathcal{F}$ are determined by the index. If $A, B$ are two Fredholm operators then

$$\dim\ker AB \leq \dim\ker A + \dim\ker B$$

so $AB$ has finite dimensional kernel and using adjoints the same is true of cokernels. This product induces the product on $[X, \mathcal{F}]$.

If $X$ is connected and the kernel of $f(x) = A_x$ has constant rank $m$, then since the index is constant the cokernel has constant rank $(m - \operatorname{ind} A_x)$ and, as $x$ varies over $X$, we have two vector bundles over $X$, whose difference $\ker A - \operatorname{coker} A$ clearly defines a class in $K(X)$. This is the basis of the isomorphism above but the key issue is that the map extends even to the case where the dimension jumps.

Suppose now that $Z \xrightarrow{\pi} X$ is a smooth fibre bundle whose fibre is diffeomorphic to a compact manifold $M^n$, and suppose we have vector bundles $V_+$, $V_-$ over $Z$ and for each $x \in X$ a smoothly varying elliptic operator

$$D : C^\infty(Z_x, V_+|_{Z_x}) \to C^\infty(Z_x, V_-|_{Z_x}).$$

Then we can convert this into a family of Fredholm operators and get an element

$$\operatorname{ind} D \in K(X).$$

The index theorem for families, proved in Part IV of the Atiyah–Singer papers, expresses this analytical class in terms of a topological class defined by the family of symbols.

If $Z \xrightarrow{\pi} X$ is a holomorphic fibration and $E$ is a holomorphic vector bundle on $Z$, then the $\bar{\partial}$ elliptic complex along the fibres is an example. The Grothendieck–Riemann–Roch theorem for this is then an example of the index theorem. The sheaf $\mathcal{O}(E)$ of holomorphic sections of $E$ on $Z$ defines coherent sheaves over $X$ whose sections over an open set $U \subset X$ are $H^p(\pi^{-1}(U), \mathcal{O}(E))$. The alternating sum of these defines an element $\pi_! \mathcal{O}(E)$ in the Grothendieck group of coherent sheaves on $X$ which maps under the Chern character to the cohomology of $X$. Then

$$\operatorname{ch}\bigl(\pi_! \mathcal{O}(E)\bigr) \operatorname{td}(TX) = \pi_*\bigl(\operatorname{ch}(E) \operatorname{td}(TZ)\bigr)$$

where $\pi_*$ is the map defined by integration over the fibres $\pi_* : H^p(M, \mathbf{R}) \to H^{p-n}(M, \mathbf{R})$.

The cohomological version of the index theorem for families has a similar form. If $D$ is a family of elliptic operators, it has a symbol class $[\sigma] \in K_c(TZ)$ and an analytical index $\operatorname{ind} D \in K(X)$. Then

$$\operatorname{ch}(\operatorname{ind} D) = (-1)^n \pi_*\bigl(\operatorname{ch} \sigma \operatorname{td}(TZ \otimes \mathbf{C})\bigr)$$

where $\pi_* \colon H_c^*(TZ) \to H^*(X)$ is integration over the fibres. This is an important formula for calculations but the problem is still best framed in K-theoretical terms. In particular the mod 2 index has a family version.

**6.2. Jumping of Dimension.** We have noted already that in a continuous family of Fredholm operators while the integer index remains constant the dimension of the kernel may jump. The index theorem for families can sometimes detect this. The integer function $\dim \ker A$ is upper semi-continuous and so the jumps are upwards in dimension. Suppose that we have a family where the index is zero. Then if $\dim \ker A$ is always zero in a family, so is $\dim \operatorname{coker} A$ and the K-theory index in $K(X)$ vanishes. Hence if we know the index is non-zero there must be non-trivial jumps somewhere in the family.

One classical example is to take the Dirac operator on a Riemann surface $M$

$$\bar{\partial} \colon C^\infty(K^{1/2} \otimes L) \to C^\infty(K^{1/2} \otimes L\bar{K})$$

with coefficient bundle a line bundle with flat unitary connection. The index is zero here since the Riemann–Roch formula gives $g-1+1-g=0$. The flat line bundles are parametrized by the torus $X=H^1(M,\mathbf{R}/\mathbf{Z})$ and if we choose a universal line bundle $L$ over $Z=M\times X$, we have a setting to apply the index theorem and find a class in $K(X)$, or from the Chern character in $H^*(X)$. This is a holomorphic situation so it is actually the Grothendieck–Riemann–Roch theorem we use.

The universal bundle $L$ has Chern class

$$c(L)=\sum_1^g(x_i y_i'-y_i x_i')\in H^1(M,\mathbf{Z})\otimes H^1(X,\mathbf{Z})\subset H^2(M\times X,\mathbf{Z})$$

where $x_1,\dots,x_g,y_1,\dots,y_g$ is a symplectic basis of $H^1(M,\mathbf{Z})$ and we use the isomorphism $H^1(X,\mathbf{Z})\cong H^1(M,\mathbf{Z})$ to define a corresponding basis $x_1',\dots,x_g'$, $y_1',\dots,y_g'$. The G–R–R formula is then

$$\operatorname{ch}\big(\pi_!\mathcal{O}(L\otimes K^{1/2})\big)\operatorname{td}(TX)=\pi_*\big(\operatorname{ch}(L\otimes K^{1/2})\operatorname{td}(TX)\operatorname{td}(M)\big)$$

or, since $TX$ is trivial,

$$\begin{aligned}\operatorname{ch}\big(\pi_!\mathcal{O}(L)\big)&=\pi_*\left(\left(1+c(L)-\frac{1}{2}c_1(TM)+\frac{1}{2}c(L)^2\right)\left(1+\frac{1}{2}c_1(TM)\right)\right)\\&=\pi_*\left(1+c(L)+\frac{1}{2}c(L)^2\right)\end{aligned}$$

since $H^p(M)$ vanishes for $p\geq 2$. Now $c(L)$ has no component in $H^2(M)\otimes H^0(X)$ so integrating over $M$ kills this. We have

$$c(L)^2=\left(\sum_1^g(x_i y_i'-y_i x_i')\right)^2=-2\omega\theta$$

where $\omega=x_1y_1=x_2y_2=\cdots=x_ny_n$ is the generator of $H^2(M,\mathbf{Z})$ and $\theta=\sum_1^g x_i'y_i'$. It follows that

$$\operatorname{ch}\big(\pi_!\mathcal{O}(L)\big)=-\theta.$$

This is non-zero and so the dimension of the kernel jumps. It does so of course on the theta divisor in the torus $X$, which is Poincaré dual to the cohomology class $\theta$.

This is a classical example but note that $\theta$ is non-zero even when $X$ has genus one i.e. is a torus itself. So even when base and fibre of $Z$ have no non-trivial characteristic classes there is still a non-trivial index. In higher dimensions one can do the same with $M$ an even-dimensional torus $T^{2m}=\mathbf{R}^{2m}/\mathbf{Z}^{2m}$. The one-form $\sum y_i dx_i$ describes a family of flat connections on the trivial bundle over $T^{2m}$ parametrized by $(y_1,\dots,y_{2m})\in\mathbf{R}^{2m}/2\pi\mathbf{Z}^{2m}=X$. The curvature of this line bundle $L$ over $T^{2m}\times X$ is

$$F=d\sum y_i dx_i=\sum dy_i\wedge dx_i.$$

The Chern character now contributes a non-trivial term to the index by integrating $F^{2m}$ over the fibre to give a non-zero class in $H^{2m}(X)$. This means in particular that there is a non-trivial jump in the dimension of the kernel of the Dirac operator. In particular the torus cannot have a metric of positive scalar curvature since in Lichnerowicz's formula there is no contribution from the zero curvature of the line bundle. This result, due to Gromov and Lawson, has spurred a great deal of work understanding which fundamental groups are compatible with positive scalar curvature. They all involve indices of a more sophisticated nature, taking values in the K-theory of the $C^*$ algebra associated to a discrete group.

# 7 The Local Index Theorem

**7.1. The Heat Kernel.** The foundational papers of Atiyah and Singer on the index theorem expressed the analytic index as a topological invariant. It could be represented in different ways depending on which version of cohomology one used and how one represented characteristic classes. However, the main operators of interest such as the Dirac operator were expressed in terms of a Riemannian metric. In this setting there was also a natural way to represent the characteristic classes—by using the curvature of the Levi-Civita connection. In the early 1970s new proofs emerged which capitalized on this fact and provided the tools for the study of another range of index problems. The original idea came from work of McKean and Singer.

Suppose $D$ is a first order elliptic operator such as the Dirac operator, then one considers the self-adjoint operators $DD^*$ and $D^*D$. They have, on a compact manifold, a discrete spectrum $0 \leq \lambda_0 \leq \lambda_1 \ldots$ each value taken only a finite number of times. If $\phi_i$ are the eigenvectors, then the *heat kernel*

$$H(x,y,t) = \sum_{j=0}^{\infty} e^{-\lambda_j t}\phi_j(x)\phi_j(y)$$

is for $t > 0$ a well-behaved smooth function, formally written as $e^{-tD^*D}$ or $e^{-tDD^*}$. In particular it has a trace

$$\operatorname{tr} e^{-tD^*D} = \sum_0^{\infty} e^{-\lambda_j t} \qquad \operatorname{tr} e^{-tDD^*} = \sum_0^{\infty} e^{-\mu_j t}.$$

Now if $D^*D\phi = \lambda\phi$, then $DD^*D\phi = \lambda D\phi$ so that if $D\phi$ is non-zero, then it is an eigenvector for $DD^*$. It follows that the non-zero eigenvalues of $DD^*$ and $D^*D$ are the same so that

$$\operatorname{tr} e^{-tD^*D} - \operatorname{tr} e^{-tDD^*} = \dim\ker D^*D - \dim\ker DD^* = \operatorname{ind} D.$$

In particular, this expression is independent of $t$ and so one can consider the behaviour of each term on the left hand side as $t$ approaches 0.

In this case, along the diagonal one has an asymptotic expansion

$$H(x,x,t) \sim \sum_{j=0}^{\infty} a_j t^{-n/2+j}$$

where the coefficients $a_j$ are determined locally. In other words each one is an algebraic expression in a finite number of derivatives of the coefficients of the operator $D$—in the case of the Dirac operator, the Riemannian metric by which it is defined. Then

$$\begin{aligned} \operatorname{ind} D &= \int_M \left[ \sum_{j=0}^{\infty} a_j(D^*D) t^{-n/2+j} - \sum_{j=0}^{\infty} a_j(DD^*) t^{-n/2+j} \right] \\ &= \int_M \big( a_{n/2}(D^*D) - a_{n/2}(DD^*) \big). \end{aligned}$$

(Note that this already implies that the index is zero in odd dimensions.)

The first term in the asymptotic expansion involves simply the volume, but the relevant terms for the index theorem are much further along and in principle could involve many derivatives of the metric. Evidence that a proof of the index theorem could be produced like this came from the ingenious cancellations that the young Indian mathematician Vijay Patodi used to prove the Gauss–Bonnet theorem with the same approach. Then appeared the work of Gilkey which led to a completely new proof of the index theorem. This involved: characterizing certain essential features of the polynomials, using invariant theory of the orthogonal group; showing that this yielded the Pontryagin forms defined from the Levi-Civita connection; finally, much as in Hirzebruch's signature theorem, evaluating on standard examples to get the correct coefficients.

Proving the index theorem this way is enough to get the general integer index theorem because the symbols of the Dirac operator with all possible coefficient bundles generate all the homotopy classes. It raised other questions of a different nature however, in the complex case for example. The use of Riemannian methods meant that Kähler manifolds could be treated this way, but how did one get a local Riemann–Roch theorem for a general complex manifold where the Riemannian connection is not compatible with the complex structure? Bismut [15] discovered that to get a local index theorem one has to use a Riemannian connection whose torsion tensor is defined by a closed 3-form $H$, thought of as a 1-form with values in skew-adjoint endomorphisms of the tangent bundle. In other words one uses connections of the form

$$\nabla + H$$

where $\nabla$ is the usual Levi-Civita connection, which has zero torsion. The curious feature here is that the local index formula for the Dirac operator defined by the connection $\nabla + H/3$ involves the Pontryagin forms of the connection $\nabla - H$. Moreover in the Hermitian case if the connection $\nabla + H$ preserves the complex structure, the $\bar{\partial} + \bar{\partial}^*$ operator is defined using $\nabla + H/3$. In the context of Lie groups this was called by Kostant the *cubic Dirac operator* and has some special features, notably that the zero-order term in the Lichnerowicz formula is still a scalar function.

**7.2. The Eta Invariant.** One of the areas which Atiyah, Singer and Patodi developed using heat equation methods was the geometrical study of some boundary value problems. The signature theorem was one motivation for this. We saw in Sect. 2.4 how the middle dimensional cohomology of a compact oriented $4k$-dimensional manifold has a non-degenerate symmetric bilinear form, which allows the definition of the signature. For a $4k$-dimensional manifold $M$ with boundary $\partial M$, one can also define the signature, using compactly supported closed forms. The additivity theorem of Novikov asserts that when two compact oriented $4k$-manifolds are glued by an orientation reversing diffeomorphism of their boundaries, the signature of their union is the sum of their signatures.

On a compact manifold, the signature is given by the integral of a differential form given as a polynomial in Pontryagin forms by Hirzebruch's formula. If we do this for the manifold with boundary $M$, this will not necessarily be so. However, if the metric near the boundary is a product we can smoothly glue together two such manifolds to get a Riemannian manifold, and it follows from Novikov additivity that the difference between the signature of $M$ and the integral is an invariant only of the Riemannian metric on the $4k-1$-dimensional boundary $\partial M$. Identifying this invariant led to another type of index theorem. As usual, it is easiest to describe for the basic Dirac operator $D$.

In dimension $4k-1$ the Dirac operator is real and self-adjoint so it has real eigenvalues $\lambda_i$, both positive and negative since $D$ is a first-order operator. One then defines

$$\eta(s) = \sum_{\lambda_j \neq 0} (\operatorname{sgn} \lambda_j)|\lambda_j|^{-s}.$$

This is holomorphic when the real part of $s$ is large but has a meromorphic extension to the whole complex plane and is finite at $s = 0$. Formally speaking then, $\eta(0)$ is the difference between the (infinite) number of positive and negative eigenvalues of $D$—the "signature" of the quadratic form $(D\varphi, \varphi)$.

This eta invariant appears as a correction term in the formula for the index of a Fredholm operator for the manifold with boundary $M$. Atiyah, Patodi and Singer consider the Dirac operator on $M$

$$D : C^\infty(S_+) \to C^\infty(S_-)$$

and solutions to $D\varphi = 0$ with the boundary condition that the projection of $\varphi$ onto the space spanned by the positive eigenvectors of the Dirac operator on the boundary is zero. It turns out that this is Fredholm and there is an index theorem of the form

$$\operatorname{ind} D = \int_M \hat{A}(TM) - \frac{1}{2}\big(\eta(0) + h\big)$$

where $h$ is the dimension of the kernel of the Dirac operator on the boundary.

Another way of interpreting this result is to note that if the projection onto the positive part and zero eigenspace vanishes, then on the boundary $\varphi$ has an expansion

$$\varphi = \sum_{\lambda_j<0} c_j \phi_j$$

and then

$$\sum_{\lambda_j<0} e^{\lambda_j t} c_j \phi_j$$

decays exponentially and is an $\mathcal{L}^2$ solution to the Dirac equation $D\varphi = 0$ on the cylinder $\partial M \times [0, \infty)$. Thus the null-space is the space of $\mathcal{L}^2$ solutions to $D\varphi = 0$ on the non-compact manifold obtained by glueing the cylinder to $M$ at its boundary. Replacing the Dirac operator by the signature operator and linking compactly supported cohomology with $\mathcal{L}^2$-cohomology gives the signature formula:

$$\tau(M) = \int_M L(TM) + (-1)^{k+1}\eta(0).$$

**7.3. Quantum Field Theory.** It turned out that, unknown to Atiyah, Singer and Patodi, the development of the local index theorem by mathematicians coincided with an interest in the theorem from theoretical physicists. As Singer has remarked, some of this was taking place in offices around the corner from his own in MIT. The context in 1970 was the *chiral anomaly* of Roman Jackiw. An anomalous symmetry in quantum field theory is a symmetry of the action, but not of the measure. In the standard model of electroweak interactions the classical current conservation law $\partial_\mu J^B_\mu = 0$ is replaced by

$$\partial_\mu J^B_\mu = \frac{g^2 C}{32\pi^2} \epsilon_{\mu\nu\alpha\beta} F_{\mu\nu} F_{\alpha\beta}.$$

The right hand side here is essentially the second Chern form for the connection defined by the gauge theory. Moreover an important physical fact is that this term is a total derivative involving $\partial_\mu K_\mu$ where

$$K_\mu = 2\epsilon_{\mu\nu\alpha\beta}\left(A_\nu \partial_\alpha A_\beta + \frac{2}{3} ig A_\nu A_\alpha A_\beta\right).$$

Mathematically this term only makes sense having chosen a trivialization of the bundle—a choice of gauge—so that the connection is $\partial_\mu + gA_\mu$ but this so-called Chern–Simons expression appears naturally in the Atiyah–Singer–Patodi formula. On a 3-manifold where the bundle is globally trivial the integral of this expression is well-defined modulo the integers and the eta-invariant is a real lift of it.

In terms of methodology, the physicists were happy with heat kernels but at that stage knew little about the topology. The ingredients for studying anomalies were the same as for the index theorem and it turned out that certain anomalies were precisely indices. It was Singer's interest in this parallel evolution that led him to talk more to the physicists and subsequently to introduce the problem of Yang–Mills instantons to Atiyah and coworkers when he visited Oxford in 1977.

Another interface with physics came from supersymmetric field theories consisting of a physical system described by a Hamiltonian $H$ and two *supercharges* $Q$ and $Q^\dagger$ which map fermions to bosons and vice versa. They satisfy the anti-commutation relation $\{Q, Q^\dagger\} = H$ so that both supercharges commute with $H$ and satisfy $Q^2 = (Q^\dagger)^2 = 0$. Clearly there is an example given by $Q = d : \Omega^{even} \to \Omega^{odd}$, $Q^\dagger = d^*$ and $H = dd^* + d^*d$, the Hodge Laplacian. More generally any elliptic complex fits this scheme, and the index theorem becomes the problem of evaluating the so-called Witten index.

This new viewpoint led to supersymmetric proofs of the index theorem by physicists [1, 18] which were given a rigorous mathematical form by Getzler [14, 21].

**7.4. The Supersymmetric Proof.** The physics motivation for Getzler's proof is the background expansion used in the supersymmetric path integral to obtain the small fluctuation Lagrangian. One uses the Dirac operator with coefficient bundle, normal coordinates at a point and an expansion $x_i + \sqrt{t}y_i$. Then Getzler introduces a clever rescaling including that of the Clifford algebra, so that if the degree of $t$ is 2, of $x_i$ is one, then the degree of $e_i$, a generator of the Clifford algebra is $-1$. The effect is that as $t \to 0$ the Clifford algebra approaches the Grassmann algebra and the Dirac Laplacian approaches

$$\left(\partial_i - \frac{1}{4} R_{ij} x_j^\wedge\right)^2 + F\wedge$$

where $R_{ij}$ is the Riemann curvature tensor considered as a matrix of 2-forms. The fact that the Lichnerowicz formula involves just the scalar curvature and doesn't contribute other terms in the Clifford algebra is a key point here, and explains the presence of the cubic Dirac operator in Bismut's modification [15].

The heat kernel is then approximated by the heat kernel for the harmonic oscillator

$$-\Delta + \theta_{ij} x_i x_j$$

but using exterior multiplication instead of scalar multiplication of functions. The heat kernel for the $m$-dimensional harmonic oscillator is

$$\begin{aligned}&(4\pi t)^{-m/2} \det\left[\frac{2t\sqrt{\theta}}{\sinh 2t\sqrt{\theta}}\right]\\ &\quad\times \exp -\frac{1}{4t}\left[\left(\frac{2t\sqrt{\theta}}{\tanh 2t\sqrt{\theta}}\right)_{ij}(x_i x_j + y_i y_j) - 2\left(\frac{2t\sqrt{\theta}}{\sinh 2t\sqrt{\theta}}\right)_{ij} x_i y_j\right].\end{aligned}$$

Replacing $\theta_{ij}$ by the matrix of forms $R_{ij}$ leads to the index formula. The rescaling has the effect that the index term $a_{n/2}$ in the asymptotic expansion becomes the leading coefficient.

The obvious feature of this formula is the natural presence of the expression

$$\frac{\sqrt{x}/2}{\sinh(\sqrt{x}/2)}$$

—the polynomial defining the $\hat{A}$-genus which prompted Atiyah's original question to Singer in 1962. The physics thus provides some form of explanation of the role of these very special polynomials.

## References

1. Alvarez-Gaumé, L.: Supersymmetry and the Atiyah–Singer index theorem. Commun. Math. Phys. **90**, 161–173 (1983)
2. Atiyah, M.: Mathematician, http://www.peoplesarchive.com
3. Atiyah, M.F., Singer, I.M.: The index of elliptic operators I. Ann. Math. **87**, 484–530 (1968)
4. Atiyah, M.F., Singer, I.M.: The index of elliptic operators III. Ann. Math. **87**, 546–604 (1968)
5. Atiyah, M.F., Singer, I.M.: The index of elliptic operators IV. Ann. Math. **93**, 119–138 (1971)
6. Atiyah, M.F., Singer, I.M.: The index of elliptic operators V. Ann. Math. **93**, 139–149 (1971)
7. Atiyah, M.F., Singer, I.M., Segal, G.B.: The index of elliptic operators II. Ann. Math. **87**, 531–545 (1968)
8. Atiyah, M.F., Bott, R.: A Lefschetz fixed point formula for elliptic complexes I. Ann. Math. **86**, 374–407 (1967)
9. Atiyah, M.F., Bott, R.: A Lefschetz fixed point formula for elliptic complexes II Applications. Ann. Math. **88**, 451–491 (1968)
10. Atiyah, M.F., Hirzebruch, F.: Spin-manifolds and group actions. In: Haefliger, A., Narasimhan, R. (eds.) Essays on Topology and Related Topics (Mémoires dédiés à Georges de Rham), pp. 18–28. Springer, New York (1970)
11. Atiyah, M.F.: Riemann surfaces and spin structures. Ann. Sci. École Norm. Sup. **4**, 47–62 (1971)
12. Atiyah, M.F., Hitchin, N.J., Singer, I.M.: Deformations of instantons. Proc. Nat. Acad. Sci. U.S.A. **74**, 2662–2663 (1977)
13. Atiyah, M.F., Hitchin, N.J., Drinfeld, V.G., Manin, Yu.I.: Construction of instantons. Phys. Lett. A **65**, 185–187 (1978)
14. Berline, N., Getzler, E., Vergne, M.: Heat Kernels and Dirac Operators. Springer, Berlin (1992)
15. Bismut, J.-M.: A local index theorem for non-Kähler manifolds. Math. Ann. **284**, 681–699 (1989)
16. Bott, R., Taubes, C.: On the rigidity theorems of Witten. J. Am. Math. Soc. **2**, 137–186 (1989)
17. Donaldson, S.K., Kronheimer, P.B.: The Geometry of Four-Manifolds. Oxford Univ. Press, Oxford (1990)
18. Friedan, D., Windey, P.: Supersymmetric derivation of the Atiyah–Singer index and the chiral anomaly. Nucl. Phys. B **235**, 395–416 (1984)
19. Gelfand, I.M.: On elliptic equations. (Russ.) Usp. Mat. Nauk **15**, 121–132 (1960)
20. Gelfand, I.M.: On elliptic equations. Russ. Math. Surv. **15**, 113–123 (1960)
21. Getzler, E.: A short proof of the local Atiyah–Singer index theorem. Topology **25**, 111–117 (1986)
22. Gray, J.J.: The Riemann–Roch theorem and geometry, 1854–1914. In: Proceedings of the International Congress of Mathematicians, vol. III, Berlin (1998). Doc. Math. 1998, Extra vol. III, 811–822 (electronic)
23. Gromov, M., Lawson, H.B. Jr.: The classification of simply connected manifolds of positive scalar curvature. Ann. Math. **111**, 423–434 (1980)

24. Hirzebruch, F.: Neue topologische Methoden in der algebraischen Geometrie. Ergebnisse der Mathematik und ihrer Grenzgebiete, Heft 9. Springer, Berlin (1956)
25. Hirzebruch, F.: The signature theorem: reminiscences and recreation. In: Prospects in Mathematics. Ann. Math. Stud., vol. 70, pp. 3–31. Princeton Univ. Press, Princeton (1971)
26. Lichnerowicz, A.: Spineurs harmoniques. C. R. Acad. Sci. Paris **257**, 7–9 (1963)
27. Palais, R.S.: Seminar on the Atiyah–Singer Index Theorem. Ann. Math. Stud., vol. 57. Princeton Univ. Press, Princeton (1965)
28. Singer, I.M.: Letter to Michael. In: Yau, S.-T. (ed.) The Founders of Index Theory: Reminiscences of Atiyah, Bott, Hirzebruch, and Singer, pp. 296–297. International Press, Somerville (2003)
29. Stolz, S.: Simply connected manifolds of positive scalar curvature. Bull. Am. Math. Soc. **23**, 427–432 (1990)

# 2005

# Peter D. Lax

ABEL
PRISEN

# Autobiography

**Peter D. Lax**

Like most mathematicians, I became fascinated with mathematics early, about age ten. I was fortunate that my uncle could explain matters that puzzled me, such as why minus times minus is plus—it follows from the laws of algebra.

Mathematics had a deep tradition in Hungary, going back to the epoch-making invention of non-Euclidean geometry by János Bolyai, an Hungarian genius in the early 19th century. To this day, the Hungarian mathematical community seeks out mathematically talented students through contests and a journal for high school students. Winners are then nurtured intensively. I was tutored by Rose Peter, an outstanding logician and pedagogue; her popular book on mathematics, “Playing with Infinity” is still the best introduction to the subject for the general public.

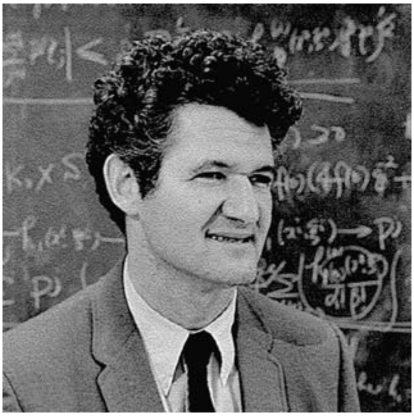
At the end of 1941 I came to the US with my family, sailing from Lisbon on December 5, 1941. It was the last boat to America; I was 15 years old. My mentors wrote to Hungarian mathematicians who had already settled in the US, asking them to take an interest in my education. They were very supportive.

I finished my secondary education at Stuyvesant High School, one of the elite public schools in New York City; its graduates include many distinguished mathematicians and physicists. For me the important thing was to learn English, and the rudiments of American history. In the meanwhile I visited from time to time Paul Erdős at the Institute for Advanced Study at Princeton. He was extremely kind and supportive; he would give me problems, some of which I managed to solve. My first publication, in 1944, was “Proof of a conjecture of P. Erdős on the derivative of a polynomial”.

At the suggestion of Gabor Szegő I enrolled at New York University in the Spring of 1943 to study under the direction of Richard Courant, widely renowned for nurturing young talent. It was the best advice I ever received. But my studies came to a temporary halt in June 1944, when I was drafted into the US Army. I became an American citizen during my basic training.

P.D. Lax (✉)
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
e-mail: lax@courant.nyu.edu

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_8, © Springer-Verlag Berlin Heidelberg 2010



(a)

(b)

(c)

I was then sent to Texas A & M to study engineering. There I passed the preliminary test in calculus with flying colors, and together with another soldier with similar background was excused from the course. Professor Klipple, a former student of R.L. Moore, generously offered to conduct just for the two of us a real variables course in the style of R.L. Moore. I learned a lot.

In June, 1945 I was posted to Los Alamos, the atomic bomb project. It was like living science fiction; upon arrival I was told that the whole town of 10,000 was engaged in an effort to build an atomic bomb out of plutonium, an element that does

(d)

With Lori Courant on the day I was notified about the Abel Prize

not exist in nature but was manufactured in a reactor at Hanford, WA. The project was led by some of the most charismatic leaders of science. I was assigned to the Theoretical Division; I joined a number of bright young soldiers, Dick Bellman, John Kemeny, Murray Peshkin and Sam Goldberg.

Von Neumann was a frequent consultant; each time he came, he gave a seminar talk on mathematics. He had little time to prepare these talks, but they were letter perfect. Only once did he get stuck; he excused himself by saying that he knew three ways of proving the theorem, but unfortunately chose a fourth.

Enrico Fermi and Niels Bohr were also consultants; since they were closely associated with nuclear physics, Security insisted that they use code names; Fermi became Henry Farmer, Bohr Nicholas Baker. At a party a woman who had spent some time in Copenhagen before the war recognised Bohr, and said “Professor Bohr, how nice to see you”. Remembering what Security had drilled into him, Bohr said “No, I am Nicholas Baker”, but then immediately added “You are Mrs Houtermans”. “No”, she replied, “I am Mrs Placzek”. She had divorced and remarried in the meantime.

I returned to New York University in the Fall of 1946 to get my undergraduate degree, and simultaneously to continue my graduate studies. I joined an outstanding class of graduate students in mathematics, Avron Douglis, Eugene Isaacson, Joe Keller, Martin Kruskal, Cathleen Morawetz, Louis Nirenberg and Anneli Kahn.

I got my PhD in 1949 under the direction of K.O. Friedrichs, a wonderful mathematician and a delightful, idiosyncratic person. He kept his life on a strict schedule; he knew that this was absurd, but it worked for him. When a graduate student of his repeatedly delayed finishing his dissertation, I explained to Friedrichs that the student was very neurotic. That seemed to him no excuse; “Am I not just as neurotic?”, he said, “yet I finish my work”.

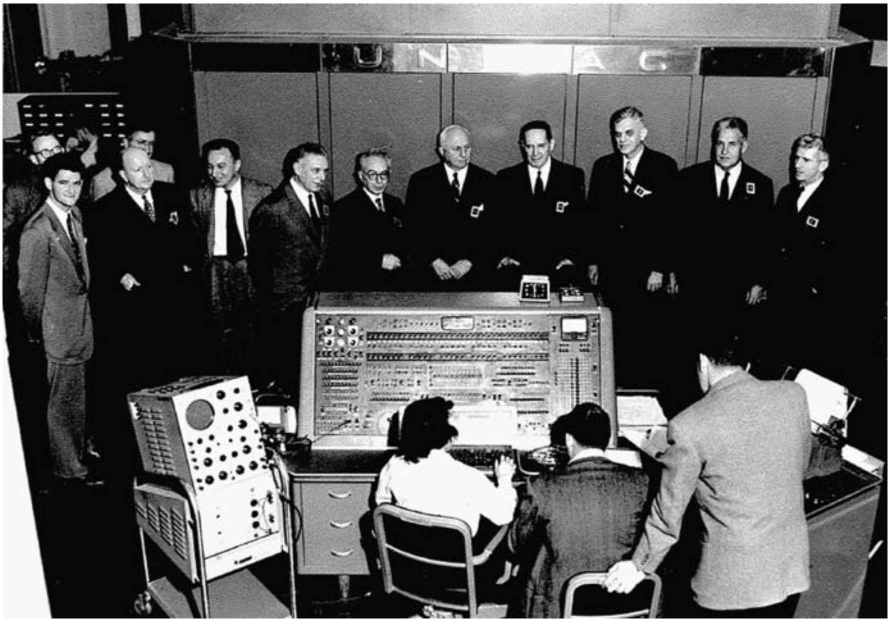
(e)

In the course of some research we did jointly, I found a reference in a Russian journal. Friedrichs said that he knew the alphabet, a couple of hundred words, and the rudiments of Russian grammar, and was willing to read the paper. I was worried about the language barrier, but Friedrichs said: “That it is in Russian is nothing; the difficulty is that it is mathematics”.

Anneli and I married in 1948; our first child, John, was born in 1950. We spent part of 1950/51 at Los Alamos. We returned to New York in the fall of ’51 to take up an appointment as Research Assistant Professor, in the mathematics department of New York University. I remained in the department for nearly fifty years, basking in the friendly collegial atmosphere of the place.

In the fifties I spent most of my summers at Los Alamos. At that time under the leadership of von Neumann, Los Alamos was the world leader in numerical computing and had the most up-to-date computers. I became, and remained, deeply involved in problems of the numerical solutions of hyperbolic equations, in particular the equations of compressible flow.

In 1954 the Atomic Energy Commission placed a Univac computer at the Courant Institute; it was the first supercomputer, with a thousand words of memory. Our first task was to calculate the flood stages on the Columbia river in case the Grand Cooley dam were destroyed by sabotage. The AEC wanted to know if the



(f) In front of the UNIVAC in 1954. From left: Peter Lax, Gen. Willoughby (adjutant of Gen. MacArthur), Lazer Bromberg (Director of Courant Computing Center), Gus Kinzel (Chairman of Courant Council), Richard Courant, James Rand (CEO of Remington-Rand), Gen. MacArthur, Henry Heald (President of NYU), Gen. Groves (Head of the Manhattan Project), Gen. Howley (Commander of the Berlin airlift, Vice President at NYU)

Hanford reactor would be flooded. Originally the Corps of Engineers was charged with this task, but it was beyond their capabilities. The team at the Courant Institute, led by Jim Stoker and Eugene Isaacson, found that the reactor would be safe.

The Univac was manufactured by the Remington–Rand Corporation. The official installation of the computer at New York University was a sufficiently important event for James Rand, the CEO, to attend and bring along some members of his Board of Directors, including the chairman, General Douglas McArthur, and General Leslie Groves. In his long career General McArthur had been Commandant of the Corps of Engineers; he was keenly interested in our calculations, and grasped the power of modern computers.

The postwar years were a heady time for mathematics, in particular for the theory of partial differential equations, one of the main lines of research at the Courant Institute as well as other institutes here and abroad. The subject is a wonderful mixture of applied and pure mathematics; most equations describe physical situations, but then take on a life of their own.

Richard Courant retired as Director of the Institute in 1958, at the age of 70. His successor was Jim Stoker who accomplished the crucial task of making the Institute part of New York University by securing tenure for its leading members.

After Stoker retired, the younger generation took over, Jürgen Moser, Louis Nirenberg, the undersigned, Raghu Varadhan, Henry McKean, Cathleen Morawetz, and others; our present Director is Leslie Greengard. Significant changes took place during these years, but the Institute adhered to the basic principle that had guided Richard Courant: not to pursue the mathematical fashion of the day ("I am against panic buying in an inflated market") but to hire promising young people. Also, for Courant mathematics was a cooperative enterprise, not competitive.

In the sixties Ralph Phillips and I embarked on a project to study scattering theory. Our cooperation lasted 30 years and led to many new results, including a reformulation of the theory. This reformulation was used by Ludvig Faddeev and Boris Pavlov to study automorphic functions; they found a connection between automorphic scattering and the Riemann hypothesis.

Also in the sixties Martin Kruskal and Norman Zabusky, guided by extensive numerical computations, found remarkable properties of solutions of the Korteweg–de Vries equation. These eventually led to the discovery that the KdV equation is completely integrable, followed by the discovery, totally unsuspected, of a whole slew of completely integrable systems. I had the pleasure and good luck to participate in this development.

In 1970 a mob protesting the Vietnam war invaded the Courant Institute and threatened to blow up our CDC computer. They left after 48 hectic hours; my colleagues and I in the Computing Center smelled smoke and rushed upstairs just in time to disconnect a burning fuse. It was a foolhardy thing to do, but we were too angry to think.

In 1980 I was appointed to a six year term on the National Science Board, the policy making body of the National Science Foundation. It was an immensely gratifying experience; I learned about issues in many parts of science, as well as about the politics of science. My colleagues were outstanding scientists and highly colorful characters.

By the time the eighties rolled around, the Government no longer placed supercomputers at universities, severely limiting the access of academic scientists to computing facilities. My position on the Science Board gave me a chance to remedy this intolerable situation. A panel I chaired recommended that the NSF set up regional Computing Centers, accessible to distant users through high capacity lines. The Arpanet Project, the precursor of the Internet, demonstrated the practicality of such an arrangement.

I have always enjoyed teaching at all levels, including introductory calculus. At the graduate level my favorite courses were linear algebra, functional analysis, and partial differential equations. The notes I have prepared while teaching formed the basis of the books I have written on these subjects.

I supervised the PhD dissertations of 55 graduate students; many have become outstanding mathematicians. Some became close personal friends.

I retired from teaching in 1999, shortly after reaching the age 70. According to a US law passed in 1994, nobody can be forced to retire on account of age in any profession, including teaching at a university. This sometimes had unwelcome consequences, such as the case of a professor at a West Coast university who stayed on the faculty well into his seventies. Eventually his colleagues petitioned the administration to retire him on the ground that he is a terrible teacher. At a hearing he had a chance to defend himself; his defense was, "I have always been a terrible teacher".

In retirement I occupy myself by writing books, and by continuing to puzzle over mathematical problems. I receive invitations to visit and lecture at mathematical centers. I attend the annual meeting of the American Mathematical Society. I spend a lot of time with my friends and my family, including three rapidly growing grandsons. Anneli died in 1999; Lori Courant and I were fortunate to find each other, and we are living happily ever after.

Mathematics is sometimes compared to music; I find a comparison with painting better. In painting there is a creative tension between depicting the shapes, colors and textures of natural objects, and making a beautiful pattern on a flat canvas. Similarly, in mathematics there is a creative tension between analyzing the laws of nature, and making beautiful logical patterns.

Mathematicians form a closely knit, world wide community. Even during the height of the Cold War, American and Soviet scientists had the most cordial relations with each other. This comradeship is one of the delights of mathematics, and should serve as an example for the rest of the world.

Added by the Editors: One can find an interview with Peter and Anneli Lax in *More Mathematical People* (D.J. Albers, G.L. Alexanderson, and C. Reid, eds.), Hartcourt Brace Jovanovich Publishers, Boston, 1990.

# A Survey of Peter D. Lax's Contributions to Mathematics

**Helge Holden and Peter Sarnak**

## 1 Introduction

Peter D. Lax has given seminal contributions to several areas of mathematics. In this paper we have decided to organize our discussion according to his *Selected Papers*, edited by P. Sarnak and A. Majda, and published in two volumes by Springer in 2005 [L215]–[L216].[1] We have benefited from the comments given there. As it is impossible to cover his entire contributions to many areas of pure and applied mathematics, we have tried to make a selection of some of the highlights of a career that spans more than six decades. His research is marked by original and concise analysis, using elementary means whenever possible (but he is never shy of using

Research supported in part by the Research Council of Norway.

[1] References of the form [L$n$] ($n$ a natural number) refer to Lax's list of publications.

H. Holden (✉)
Department of Mathematical Sciences, Norwegian University of Science
and Technology, 7491 Trondheim, Norway
e-mail: holden@math.ntnu.no
url: http://www.math.ntnu.no/~holden/

H. Holden
Centre of Mathematics for Applications, University of Oslo, P.O. Box 1053, Blindern,
0316 Oslo, Norway

P. Sarnak
School of Mathematics, Institute for Advanced Study, 1 Einstein Drive, Princeton, NJ
08540, USA
e-mail: sarnak@Math.Princeton.edu
url: http://www.mat.univie.ac.at/~jmichor/

P. Sarnak
Department of Mathematics, Princeton University, Fine Hall, Washington Road,
Princeton, NJ 08544-1000, USA

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_9, © Springer-Verlag Berlin Heidelberg 2010

advanced techniques if required) to reveal new and fundamental relations. Thus his research never goes out of vogue.

Lax has been a key figure in the development of numerical methods for partial differential equations and scientific computing since its inception in the aftermath of World War II at Los Alamos to the present prolific use of computer simulations in all areas of science and technology. Lax has always stressed the interplay between mathematical analysis and numerical experiments, as a source of mutual inspiration. Indeed, Lax once remarked [1] that "Computer simulations play a big role [in mathematics]. After all, the great mathematician G.D. Birkhoff believed all his life that the ergodic hypothesis was true and devoted much of his life to studying it. If he had been able to take one look at a computer simulation, he would have seen that it wasn't so". In [L157] he elaborated further "It is impossible to exaggerate the extent to which modern applied mathematics has been shaped and fueled by the general availability of fast computers with large memories. Their impact on mathematics, both applied and pure, is comparable to the role of telescopes in astronomy and microscopes in biology".

From 1980 to 1986, Peter Lax chaired the National Science Board which was highly instrumental in making supercomputers available to university scientists while stressing the importance of further research in the area. The *Lax Report*, or as its official name reads, "Report of the Panel on Large Scale Computing in Science and Engineering", from 1982, under the sponsorship of the Department of Defense and the National Science Foundation (NSF), was very influential in stressing the importance of the enhanced use of supercomputers in science and engineering, and by necessity, increased research in computational mathematics. Furthermore, it led to the establishment of NSF's five national computing centers as well as NSFnet.

In addition to the research papers, Peter Lax has written a number of books, including a university calculus book [L132], a book on linear algebra [L182] (two editions), a book on functional analysis [L205], two books on scattering theory [L53] (two editions), [L94], two books on hyperbolic differential equations [L83, L222], and lecture notes on partial differential equations [L5]. Furthermore, he has shown his exceptional ability as an expositor in a number of survey papers, both technical and nontechnical. For the paper [L46] on numerical solutions of partial differential equations, he received the Lester R. Ford Award of the Mathematical Association of America, and for the survey paper [L79] on shock waves, he received the Chauvenet Prize as well as the Lester R. Ford Award, both of the Mathematical Association of America. He has written several papers where he revisits old theorems from a new angle and always with an elegant twist, the latest being a new proof of Cauchy's integral theorem [L225]. In addition, he has written a number of portraits and obituaries of fellow scientists, showing a deep sense of human values, and a gift for exposition. We refer to the list of publications.

We end this introduction with a brief discussion of Peter's first paper [L1], written when he was 17 years old, and published in the country where he had newly arrived, and where he would remain for the rest of his career. In the paper he resolves a conjecture by his compatriot Erdős. If $P$ is a polynomial of degree $n$, Bernstein's inequality asserts that

$$\max_{|z|\le 1} |P'(z)| \le n \max_{|z|=1} |P(z)|. \tag{1.1}$$

The inequality turns into an equality if and only if $P(z) = az^n$. Erdős conjectured that if $P$ had no zeros in $|z| < 1$, then the $n$ could be replaced by $n/2$, with equality for $P(z) = (z^n + 1)/2$. In his first paper, Lax elegantly proved this conjecture.

It is perhaps fitting to end this introduction with Peter Lax's advice to the young generation [L157] "I heartily recommend that all young mathematicians try their skill in some branch of applied mathematics. It is a gold mine of deep problems whose solutions await conceptual as well as technical breakthroughs. It displays an enormous variety, to suit every style; it gives mathematicians a chance to be part of the larger scientific and technological enterprise. Good hunting!"

## 2 Partial Differential Equations—General Results

A substantial part of Lax's work has been in partial differential equations. In this section we collect some of his more general results. More specific results, concerning difference approximations, hyperbolic equations, and integrable systems, are treated in separate sections.

The *Lax–Milgram theorem* was proved in [L11].[2] Although only a slight extension of the Riesz representation theorem and easily proved, it has turned out to be exceptionally useful, and it is now a household theorem in textbooks. In a slightly more general version (cf. [p. 57, L205]) it reads:

**Theorem 2.1** *Let $H$ be a Hilbert space. Assume that $B\colon\ H \times H \to \mathbb{C}$ satisfies*:

(i) *$B(x, y)$ is linear in $x$ for each fixed $y$, and $B(x, y)$ is skew linear in $y$ for each fixed $x$.*
(ii) *$B$ is bounded, i.e., $|B(x, y)| \le C_1 \|x\|\,\|y\|$ for some constant $C_1$.*
(iii) *$B$ is bounded from below, i.e., $|B(x, x)| \ge C_2 \|x\|^2$ for some positive constant $C_2$.*

*Then the following assertion holds*: *For any bounded linear functional $\ell$ on $H$ there exists a unique $y \in H$ such that*

$$\ell(x) = B(x, y), \quad x \in H.$$

[2] Lax describes the background as follows [p. 116, L215]: "Arthur Milgram was an excellent topologist at the University of Minnesota. ... We became friends, and he asked me for a problem to work on. I explained to him how variational arguments can be used to extend self-adjoint operators that are bounded from below, but that there is no known method for dealing with operators that are not symmetric. After some thought he came up with this theorem".

As an example of an application of the Lax–Milgram theorem we mention the following [26, Sect. 6.2]: Define the operator

$$B(u,v)=\int_\Omega\left(\sum_{j,k=1}^n a^{jk}u_{x_j}v_{x_k}+\sum_{j=1}^n b^j u_{x_j}v+cuv\right)dx,\quad u,v\in H_0^1(\Omega),\quad (2.1)$$

where $\Omega$ is an open and bounded subset of $\mathbb{R}^n$, and $a^{jk}$, $b^j$, $c$ are bounded functions. Define $L$ by

$$Lu=-\sum_{j,k=1}^n\left(a^{jk}u_{x_j}\right)_{x_k}+\sum_{j=1}^n b^j u_{x_j}+cu. \quad (2.2)$$

Then it follows from the Lax–Milgram theorem that there exists a number $\gamma\geq 0$ such that for each $\mu\geq\gamma$ and each $f\in L^2(\Omega)$ there exists a unique weak solution $u\in H_0^1(\Omega)$ of the boundary-value problem

$$Lu+\gamma u=f \text{ in } \Omega,\quad u=0 \text{ on } \partial\Omega. \quad (2.3)$$

In the short paper [L18], Lax and Courant considered the propagation of singularities in hyperbolic systems, and they showed that discontinuities in the initial data on a spacelike manifold are propagated along characteristics. The problem was studied from a more general point of view by Lax in the pioneering paper [L23], and a major step forward was taken by his analysis of asymptotic solutions of oscillatory initial value problems for first-order hyperbolic equations. The paper represents the genesis of the theory of Fourier integral operators (see [38, 39]), microlocal analysis and semiclassical limits. In typical fashion, Lax's analysis in [L23] is novel and crisp, and uses elementary methods, unmistakably his fingerprint. To be more precise, he considers first-order hyperbolic systems on $\mathbb{R}^m\times\mathbb{R}$ given by

$$Mu=U_t+\sum_{j=1}^m A_j u_{x_j}+Bu,\quad u=u(x,t)\in\mathbb{R}^n, \quad (2.4)$$

where $A_j$ and $B$ are $n\times n$ matrices with entries that are $C^\infty$ functions of $(x,t)\in\mathbb{R}^m\times\mathbb{R}$. It is assumed that in a neighborhood of the hyperplane $t=0$ the operator $M$ is space-like in the sense that all eigenvalues of the matrix

$$\mathcal{M}(p_1,\dots,p_m)=\sum_{j=1}^m p_j A_j, \quad (2.5)$$

are real and distinct for all $p_j\in\mathbb{R}$. Lax assumes that the initial data are given as

$$u|_{t=0}(x,0)=e^{i\xi\ell(x,0)}\phi(x), \quad (2.6)$$

and he posits that the solution of $Mu = 0$ takes the form

$$u(x,t) \sim e^{i\xi\ell(x,t)} \sum_{k=0}^{\infty} v_k(x,t)\xi^{-k}. \tag{2.7}$$

The natural next step is to insert the ansatz (2.7) into the equation $Mu = 0$, and solve recursively for powers of $\xi^{-k}$. That gives the well-known eikonal equation of geometric optics. If $\lambda = \lambda(x,t,p_1,\dots,p_m)$ denotes an eigenvalue for the matrix $\mathcal{M}$, we obtain the first-order nonlinear scalar equation

$$\lambda_t = \lambda(x,t,\nabla_x \ell), \tag{2.8}$$

which can be solved for small $t$ by the standard method of characteristics, which are the bi-characteristics of the original equation. The functions $v_k$ are determined by solving the corresponding transport equations. Next he expands the $\delta$ function as in (2.6), a result that is used to construct a fundamental solution of (2.4) up to a smoothing operator. More precisely, for any natural number $n$, Lax provides a distributional kernel $K_n(t,x,y)$ which differs from the fundamental solution $K(t,x,y)$ by a function that is $C^{n+1}$ in all variables. The fundamental solution is used to derive properties regarding the propagation of singularities for (2.4). This theory was further developed 15 years later by, e.g., Hörmander and Duistermaat (see, e.g., [23, 39]).

To show the power of Lax's method one may consider the following example: Let $X$ be a smooth compact Riemannian manifold of dimension $\nu$, and let $\Delta$ be the Laplace–Beltrami operator on $X$. We can then find an orthonormal basis of eigenfunctions $\phi_j$ with corresponding eigenvalues $k_j^2$, thus,

$$-\Delta\phi_j = k_j^2\phi_j. \tag{2.9}$$

Consider next the hyperbolic wave equation

$$u_{tt} = \Delta u, \qquad u(x,0) = \delta(x,y), \quad u_t(x,0) = 0 \tag{2.10}$$

on $\mathbb{R} \times X$. The fundamental solution equals

$$K(t,x,y) = \sum_{j=0}^{\infty} \cos(k_j t)\phi_j(x)\phi_j(y). \tag{2.11}$$

The Lax construction yields the singular part of $K(t,x,y)$ for small $t$, which allows for investigation of asymptotic behavior of sums like

$$\sum_{\lambda \le k_j \le \lambda+1} \phi_j(x)\phi_j(y) \text{ as } \lambda \to \infty. \tag{2.12}$$

Applications include, e.g., Weyl's law with remainder [39]. More precisely,

$$N(\lambda) = \sum_{k_j \le \lambda} 1 = (2\pi)^{-\nu} c_\nu \mathrm{Vol}(X)\lambda^\nu + \mathcal{O}(\lambda^{\nu-1}). \tag{2.13}$$

These results are sharp for a $\nu$-dimensional sphere with its standard metric. However, if $X$ has negative curvature, the results are not sharp. One problem is to understand (2.12) with shorter sums where $\lambda \leq k_j \leq \lambda + 1$ is replaced by $\lambda \leq k_j \leq \lambda + \eta(\lambda)$ with $\eta(\lambda) \sim \lambda^{-\alpha}$ for some $\alpha$ positive. This is still a major challenge in the area. See [24] and the more recent [68].

It is by now well-known that quasilinear systems of hyperbolic equations develop singularities in finite time, even for smooth initial data,[3] but in the early days of the theory, rigorous general results were absent. While it is fairly easy to see in the scalar case, it is considerably more difficult to establish this in the case of systems. In [L41] Lax applies a simple argument that provides the first rigorous proof for breakdown of solution, by analyzing the Riemann invariants in the genuinely nonlinear case (see Sect. 4), in the case of a $2 \times 2$ system. A similar result was also obtained by Oleĭnik [55]. Klainerman and Majda extended the work by Lax to the linearly degenerate case [44].

In a joint paper with Nirenberg [L48], Lax studies Gårding's inequality in the context of difference operators in order to show stability of difference schemes. The sharp Gårding inequality reads as follows. Consider a differential operator $A = a(x, D)$ where $a(x, \xi)$ is an $n \times n$ Hermitian matrix whose elements are polynomials in $\xi$. If $a(x, \xi)$ is homogeneous in $\xi$ of degree $r$, smooth, and positive definite, then the sharp Gårding inequality states that there exists a constant $K$ such that

$$\operatorname{Re}(Au, u) \geq -K\|u\|^2_{(r-1)/2}. \tag{2.14}$$

In [L48] Lax and Nirenberg provide a new proof in the matrix case, extending the proof by Hörmander [37] in the scalar case. The discrete version considers difference operators of the form

$$P_\delta = \sum_\alpha p_\alpha(x) T^\alpha \tag{2.15}$$

where $\alpha$ is a multi-index, $T^\alpha$ denotes the shift operator $(T^\alpha u)(x) = u(x + \delta\alpha)$, and $p_\alpha(x)$ are $m \times m$ matrix functions of $x$. If the symbol $p(x, \xi)$ of $P_\delta$ is sufficiently smooth, and is an Hermitian and nonnegative matrix for every $(x, \xi)$, i.e., $p(x, \xi) \geq 0$, then

$$\operatorname{Re} P_\delta \geq -K\delta \tag{2.16}$$

for some constant $K$. The result is used to infer the stability of a class of difference schemes.

Peter Lax's name is associated with several other quantities, that we for reasons of brevity are unable to discuss in detail. We mention the following: The frequently used term *negative Lax norm* originated in the paper [L16], and the *Lax–Mizohata theorem* developed from [L23]. The so-called *Lax conjecture* originated in [L24].

[3] This is extensively discussed in Sect. 4.

# 3 Difference Approximations to Partial Differential Equations

The *Lax* or *Lax–Richtmyer stability theorem* appeared in [L17] in 1956. Those were the early days of computer simulations of difference schemes for partial differential equations. The main question was (as it still is): Does the numerical scheme converge to the solution? The quest for the true solution involves the application of finer and finer resolution to improve the approximation. Is the computed approximation stable? In setting up the scheme, a first question is whether the scheme is consistent with the underlying differential equation, i.e., does it approximate the equation? Three notions are thus involved: Consistency, stability, and convergence. The Lax stability theorem says that for linear initial value problems, stability is necessary and sufficient for convergence if the scheme is consistent. More precisely, we can formulate the theorem as follows: Let $A$ be a linear operator (involving spatial derivation, matrix multiplications, etc.) and consider the initial value problem

$$u_t = Au(t), \quad u|_{t=0} = u_0. \tag{3.1}$$

By a genuine solution of (3.1) we mean a one-parameter family $u(t)$ such that

$$\left\| \frac{1}{\tau}(u(t+\tau) - u(t)) - Au(t) \right\| \underset{\tau\to 0}{\longrightarrow} 0, \quad \text{uniformly in } t \text{ for } t \in [0,T]. \tag{3.2}$$

Assume that the problem is properly posed in the sense that there exists a uniformly bounded semigroup $E_0(t)$ such that $u(t) = E_0(t)u_0$ is the solution, thus

$$\|E_0(t)u_0\| \le K\|u_0\|, \quad t \in [0,T]. \tag{3.3}$$

If $E_0$ is defined on a dense subset of some space $\mathcal{B}$, we can extend $E_0$ to some bounded and linear extension $E$ that satisfies the same bound (3.3). Introduce now a finite difference approximation. To that end let $\Delta t$ be the time discretization parameter, and define $t_n = n\Delta t$ for $n \in \mathbb{N}$. Write $u^n$ for the approximation to $u(t_n)$, that is, $u^n \approx u(t_n)$. Derivatives are replaced by finite differences, e.g., $u_t(t) \approx (u(t+\Delta t) - u(t))/\Delta t$, and similar for spatial derivatives. This turns the differential equation into a discrete equation where the unknown function is evaluated on a lattice in space and time.[4] A finite difference scheme is then a recipe that describes how to compute the lattice-valued approximate solution. In this framework the scheme is encoded in an operator $C(\Delta t)$ such that

$$u^{n+1} = C(\Delta t)u^n. \tag{3.4}$$

We say that $C(\Delta t)$ is consistent (with $A$) if

$$\lim_{\Delta t\to 0}\left\| \left( \frac{1}{\Delta t}\big(C(\Delta t) - I\big) - A \right) u(t) \right\| = 0, \quad \text{uniformly in } t \text{ for } t \in [0,T]. \tag{3.5}$$

[4] An explicit example is given in (4.9).

We say that $C(\Delta t)$ converges to $A$ if

$$\left\| C(\Delta t)^n u_0 - E(t)u_0 \right\| \underset{\Delta t\to 0,\ n\Delta t = t}{\longrightarrow} 0, \quad t \in [0,T]. \tag{3.6}$$

Finally, we say that $C(\Delta t)$ is stable if $C(\Delta t)^n$ remains uniformly bounded for $n\Delta t \in [0,T]$ and all $\Delta t \le \tau$ for some fixed $\tau$. We have all the results we need to state the Lax–Richtmyer stability theorem.

**Theorem 3.1** *Given the properly posed initial value problem* (3.1), *and a finite difference approximation $C(\Delta t)$ to it that satisfies the consistency condition, stability is a necessary and sufficient condition that $C(\Delta t)$ be a convergent approximation.*

## 4 Hyperbolic Systems of Conservation Laws

The fundamental nature of hyperbolic conservation laws can easily be seen from the following formal derivation. Consider a conserved quantity with density $u = u(x,t)$, where $x$ denotes the space variable and $t$ denotes time. Assume that the quantity moves with velocity $v = v(x,t)$. Conservation yields that

$$\frac{d}{dt}\int_\Omega u(x,t)\,dx = -\int_{\partial\Omega} u(x,t)v(x,t)\cdot n(x,t)\,dS, \tag{4.1}$$

where $\Omega \subset \mathbb{R}^d$ is some fixed domain in space with boundary $\partial\Omega$ and outward unit normal $n(x,t)$. Gauss' theorem implies that

$$\frac{d}{dt}\int_\Omega u(x,t)\,dx = -\int_\Omega \nabla_x \cdot \big(u(x,t)v(x,t)\big)\,dx, \tag{4.2}$$

which rewrites to

$$\int_\Omega \big(u(x,t)_t + \nabla_x \cdot \big(u(x,t)v(x,t)\big)\big)\,dx = 0, \tag{4.3}$$

from which we conclude that

$$u(x,t)_t + \nabla_x \cdot \big(u(x,t)v(x,t)\big) = 0. \tag{4.4}$$

Making the assumption that the velocity $v$ depends on the density solely, we introduce the flux function $f(u) = uv(u)$ to find the hyperbolic conservation law

$$u_t + \nabla_x \cdot f(u) = 0. \tag{4.5}$$

There was nothing in the previous derivation that prevented the quantity $u$ from being a vector, $u \in \mathbb{R}^n$, in which case we have a system of hyperbolic conservation laws. A particular case of these equations is the Euler equations of gas dynamics. In spite of the fundamental nature of these equations, there is no general theory unless

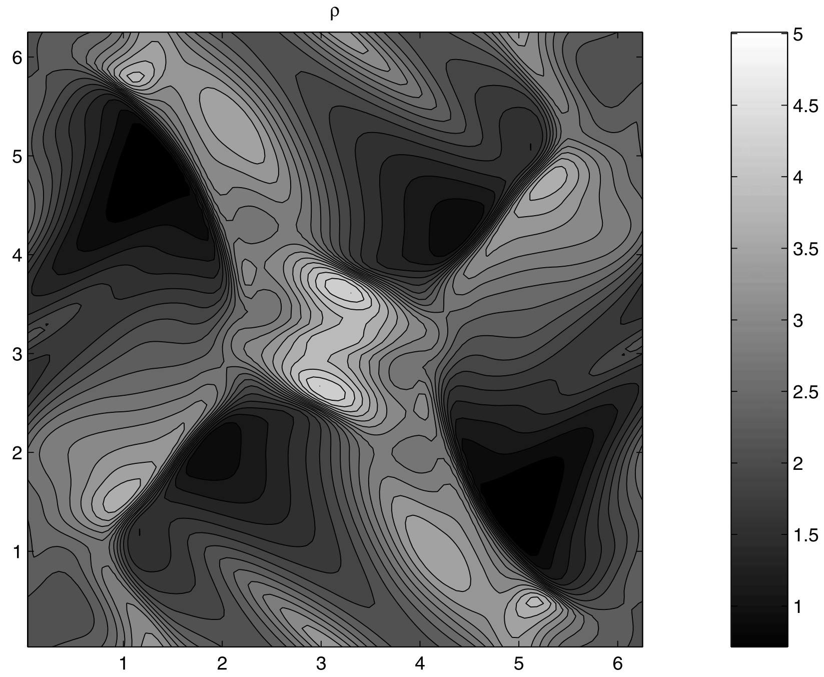


**Fig. 1** The figure shows the density in a computation in magnetohydrodynamics (MHD) using the *HLL method* [L126]. MHD can be described by a $9 \times 9$ system of hyperbolic conservation laws. The solution of the Riemann problem is not completely known, and one has to resort to numerical simulations

the spatial dimension equals one, $d = 1$, or the conservation law is scalar, $n = 1$. In Lax's words [L224] "There is no theory for the initial value problem for compressible flows in two space dimensions once shocks show up, much less in three space dimensions. This is a scientific scandal and a challenge". Numerical computations of solutions of these equations are notoriously difficult due to the intrinsic occurrence of discontinuities, denoted shocks, in the solution, even for smooth initial data. See, e.g., Fig. 1. But [L224] "Just because we cannot prove that compressible flows with prescribed initial values exist doesn't mean that we cannot compute them". Already Riemann observed that the discontinuous solutions were physical, and could not be dispensed with. The state of affairs in this area at the end of World War II, when Peter Lax entered the scene, is nicely summarized in the book [10].

Thus one has to use the notion of weak or distributional solutions, and that suddenly makes uniqueness of solutions a difficult issue. We say that $u \in L^1(\mathbb{R} \times [0, \infty))$ is a weak solution of (4.5) with given initial data $u|_{t=0} = u_0 \in L^1(\mathbb{R})$ if

$$\int_0^\infty \int_{\mathbb{R}} \big(u\phi_t + f(u) \cdot \nabla_x \phi\big)\, dx\, dt + \int_{\mathbb{R}} u_0 \phi|_{t=0}\, dx = 0 \tag{4.6}$$

for all compactly supported and smooth functions $\phi$. Criteria to identify the unique physical solution among the multitude of possible weak solutions are denoted entropy conditions.

One approach to resolve the uniqueness issue has been to consider the equation (4.5) as an approximation of a model that includes diffusion. Thus one accepts as solutions of (4.5) only those functions $u$ that are limits of solutions $u^\epsilon$, i.e., $u = \lim_{\epsilon\downarrow 0} u^\epsilon$, of the viscous regularization

$$u_t^\epsilon + \nabla_x \cdot f(u^\epsilon) = \epsilon \Delta_x u^\epsilon. \tag{4.7}$$

Let us first discuss the Cauchy problem in the scalar case on the line, i.e., $n = d = 1$, which we write as

$$u_t + f(u)_x = 0, \quad u|_{t=0} = u_0. \tag{4.8}$$

In a key paper from 1950, Hopf [36] analyzed the viscous Burgers equation, $u_t + uu_x = \epsilon u_{xx}$, i.e., the scalar one-dimensional equation (4.7) with $f(u) = u^2/2$, in the inviscid limit, that is, as $\epsilon \to 0$, and described the limit. Using the Cole–Hopf transform, the equation could be rewritten as the heat equation. The paper by Hopf represents the start of the mathematical theory of conservation laws, and the theory developed rapidly in several mathematical directions in the US, the Soviet Union, and China.

Lax introduced in [L10] what has subsequently been called the *Lax–Friedrichs difference scheme*,[5] which can be defined as follows. Let $\Delta x$, $\Delta t$ be (small) positive numbers, and denote the approximate solution to $u$ at $(j\Delta x, n\Delta t)$ by $u_j^n$, i.e., $u_j^n \approx u(j\Delta x, n\Delta t)$. Given a discretization of the initial data, i.e., given $u_j^0$ for $j \in \mathbb{Z}$, one can use the recursive definition[6]

$$u_j^{n+1} = \frac{1}{2}\big(u_{j-1}^n + u_{j+1}^n\big) - \frac{\Delta t}{2\Delta x}\big(f(u_{j+1}^n) - f(u_{j-1}^n)\big) \tag{4.9}$$

to compute $u_j^n$ for $n \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and $j \in \mathbb{Z}$. The scheme (4.9) constitutes the Lax–Friedrichs scheme,[7] and it is stable if the Courant–Friedrichs–Lewy (CFL) condition (see also [L55], [L204]), $\|f'\|_\infty \Delta t/\Delta x < 1$, holds. Being so simple to implement, and having been so well-studied over a period of more than half a century, it is still one of the first methods to be employed for simple numerical tests.

[5] Friedrichs had studied the same scheme for linear symmetric hyperbolic equations.

[6] A straightforward Taylor expansion for smooth solutions shows that the difference scheme is consistent with the differential equation. Clearly it is necessary for a difference scheme to be consistent in order to be useable. However, it is far from being sufficient for it to converge to the right solution.

[7] Peter writes [p. 227, L203]: "Numerical experiments is the life-blood of the investigation of numerical methods. My first numerical studies of the Lax–Friedrichs scheme were done in 1954, before the advent of compilers, using von Neumann's MANIAC at the Los Alamos Laboratory, built under the tutelage of Nick Metropolis. It had a huge memory of 1,024 words and used punched cards, like the voters in Palm Beach County".

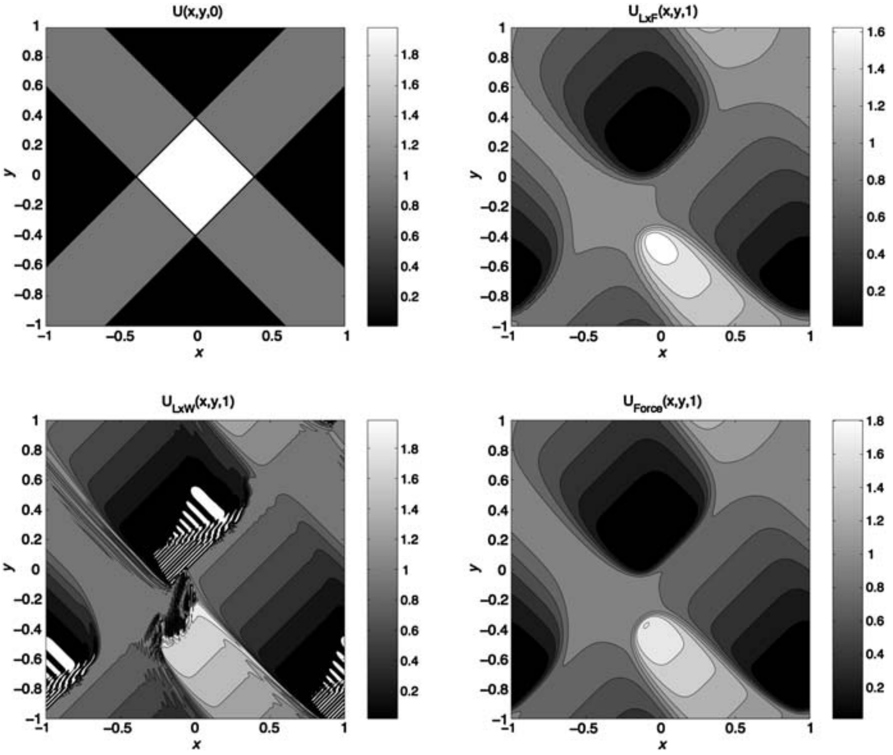


**Fig. 2** The figures show the solution of the equation $u_t + \frac{1}{2}(u^2)_x + (\sin u)_y = 0$ with initial condition as in the upper left-hand corner. In the upper right-hand corner the solution is depicted using the Lax–Friedrichs scheme. The lower left-hand corner shows the solution using the Lax–Wendroff scheme, and the oscillations are clearly visible. In the lower right-hand corner the solution using a combination of the Lax–Friedrichs and the Lax–Wendroff scheme is depicted

See, e.g., Fig. 2. Of course, the theory of numerical methods for these equations has developed dramatically over this period, and a survey of current methods can be found in [46]. The convergence of the Lax–Friedrichs scheme was proved by Oleĭnik in [55] (see also [32] and [67]), in which she also introduced the Oleĭnik entropy condition [56], which is fundamental in all further study of asymptotic behavior of solutions of scalar conservation laws. Furthermore, she proved the convergence of the vanishing viscosity method for scalar conservation laws in one space dimension. Extensions to several space dimensions have been given by Conway and Smoller [9]. The proof of convergence of the Lax–Friedrichs scheme for isentropic gas dynamics was accomplished in [6, 16–18].

The paper [L20] constitutes a cornerstone in the theory of systems of hyperbolic conservation laws. Here Lax undertakes the first comprehensive mathematical study of the case of systems. The first part addresses the scalar case, and we will start by discussing that part. Here Lax derives what has later been denoted the *Lax formula*: Assume that $f$ is strictly convex and $u_0$ is bounded. Then the weak entropy solution of (4.8) is given by $u = v_x$ where

$$v(x,t) = \sup_{y\in\mathbb{R}}\left(v_0(y) + tf^*\left(\frac{x-y}{t}\right)\right). \tag{4.10}$$

Here $v_0$ is a primitive of $u_0$ and $f^*$ is the convex conjugate function of $f$ (see [66, Thm. 2.5.1]). In [Sect. 5, L20] Lax studies the asymptotic behavior as $t \to \infty$ of solutions of scalar conservation, both on the full line and the periodic case. This paper was the starting point for the extensive generalization by Glimm and Lax in their deep analysis, [L52] and [L66], of the asymptotic behavior of solutions of $2 \times 2$ systems of conservation laws. There has always been a strong connection between the theory for hyperbolic conservation laws and Hamilton–Jacobi equations. The Lax formula played a decisive role in the work of P.L. Lions, Crandall, and Kružkov on the Hamilton–Jacobi equations.

Next, consider the case of one-dimensional systems, i.e., $n > 1$ and $d = 1$. The seminal paper [L20] sets the stage for all subsequent development in this area; the main concepts are introduced, and the paper culminates with the *Lax theorem*, i.e., the solution of the Riemann problem for systems of hyperbolic conservation laws on the line. Before we can state the theorem, we have to define the basic quantities. We consider the equation (4.5), where now $u$ and $f$ are vectors in $\mathbb{R}^n$. The Jacobian $df(u)$ is an $n \times n$ matrix assumed to have, in the strictly hyperbolic case, $n$ real and distinct eigenvalues $\lambda_j(u)$ (assumed ordered) with corresponding eigenvectors $r_j(u)$, viz.

$$df(u)r_j(u) = \lambda_j(u)r_j(u), \quad j = 1, \dots, n. \tag{4.11}$$

The $j$th wave family is said to be linearly degenerate if $\nabla\lambda_j(u) \cdot r_j(u) = 0$ and genuinely nonlinear if $\nabla\lambda_j(u) \cdot r_j(u) \neq 0$.[8] Weak solutions are defined as in the scalar case, however the question of the appropriate entropy condition is considerably more complicated for systems. Riemann analyzed in the fundamental paper [64] the following physical problem: Consider a tube with two gases separated initially by a membrane. Removing the membrane yields a Cauchy problem with initial data

$$u|_{t=0} = u_0 = \begin{cases} u_l & \text{for} \quad x < 0, \\ u_r & \text{for } x \geq 0. \end{cases} \tag{4.12}$$

Here $u_l$ and $u_r$ are constants, and the unknown vector $u$ denotes the densities of the conserved quantities. The Cauchy problem for (4.8) with initial data (4.12) is called the Riemann problem. In his study of the problem, Riemann also conserved the entropy, which, while formally derivable from the conservation laws for smooth solutions, does not hold for weak solutions, see [67, p. 264] and [L165]. It was in [L20] that a complete solution was first given in the case of small initial data. There

[8] Both cases occur for the Euler equations. Systems where a wave family changes from linearly degenerate to genuinely nonlinear, are not discussed here.

are several classes of solutions, and the full solution of the Riemann problem is composed of so-called simple waves. A shock solution has the form

$$u(x,t)=\begin{cases}u_l & \text{for } x<st,\\ u_r & \text{for } x\geq st,\end{cases}\tag{4.13}$$

where the velocity has to satisfy the Rankine–Hugoniot relation

$$s(u_r-u_l)=f(u_r)-f(u_l),\tag{4.14}$$

in order to be a weak solution.[9] However, shocks are abundant, and additional requirements are needed to eliminate unphysical solutions. Lax gave the correct entropy condition, now denoted the *Lax entropy condition*, which states that (4.13) is a $j$ entropy shock if the inequalities

$$\lambda_{j-1}(u_l)<s<\lambda_j(u_l),\qquad \lambda_j(u_r)<s<\lambda_{j+1}(u_r)\tag{4.15}$$

hold. Such shocks are denoted *Lax shocks*. Furthermore, in addition to shock solutions, the Riemann problem (4.12) may, in the genuinely nonlinear case, have continuous solutions called rarefaction waves. These are solutions of the form

$$u(x,t)=\begin{cases}u_l & \text{for } x\leq\lambda_j(u_l)t,\\ w(x/t) & \text{for } \lambda_j(u_l)t<x<\lambda_j(u_r)t,\\ u_r & \text{for } x\geq\lambda_j(u_r)t,\end{cases}\tag{4.16}$$

where the function $w=w(\xi)$ satisfies

$$\dot{w}(\xi)=r_j(w(\xi)),\qquad \nabla\lambda_j w(\xi)\cdot\dot{w}(\xi)=1.\tag{4.17}$$

In the case of a linearly degenerate wave family, we only have discontinuous solutions, denoted contact discontinuities, that equal[10]

$$u(x,t)=\begin{cases}u_l & \text{for } x<st,\\ u_r & \text{for } x\geq st,\end{cases}\tag{4.18}$$

where the speed $s$ satisfies the Rankine–Hugoniot condition, and we have $s=\lambda_j(u_l)=\lambda_j(u_r)$. Shocks, rarefaction waves, and contact discontinuities constitute the simple waves, and the solution of the Riemann problem is a concatenation of simple waves. We can now state the Lax theorem.

**Theorem 4.1** *Assume that the system* (4.8) *is strictly hyperbolic and that each wave family is either genuinely nonlinear or linearly degenerate. Consider a* $u_l\in\mathbb{R}^n$.

[9] This relation simply states that conservation holds across discontinuities.

[10] Contact discontinuities look like shocks, but they do not satisfy the Lax entropy condition.

*Then there exists a neighborhood $\Omega$ of $u_l$ such that for each $u_r \in \Omega$ the Cauchy problem with initial data* (4.12) *has a unique weak solution consisting of up to $n+1$ constant states separated by shocks*, *rarefaction waves and contact discontinuities.*

This theorem is still as central in the theory of conservation laws now as it was half a century ago. Furthermore, Lax provides a comprehensive study of the notion of Riemann invariants in [L20].

In the pioneering paper [L75], published in a conference proceedings, Lax takes a novel and penetrating look at entropies. He says that a convex function $U(u)$ is an entropy for the system (4.8) if all smooth solutions satisfy another conservation law of the form

$$U(u)_t + F(u)_x = 0. \tag{4.19}$$

A necessary and sufficient condition for this is that

$$U_u f_u = F_u, \tag{4.20}$$

which is an underdetermined system for $n = 1$, determined for $n = 2$, and overdetermined for $n > 2$ (however, for compressible flow it does have a solution). Let us introduce a dissipative regularization

$$u_t^\epsilon + f(u^\epsilon)_x = \epsilon u_{xx}^\epsilon. \tag{4.21}$$

Multiply this with $U_u$ and rearrange to find

$$U(u^\epsilon)_t + F(u^\epsilon)_x = \epsilon(U(u^\epsilon)_{xx} - u_x^\epsilon U(u^\epsilon)_{uu} u_x^\epsilon) \le \epsilon U(u^\epsilon)_{xx}, \tag{4.22}$$

using that $U$ was assumed to be convex. Let $\epsilon \to 0$, assuming $u^\epsilon \to u$ strongly, which yields

$$U(u)_t + F(u)_x \le 0 \tag{4.23}$$

weakly. We say that $u$ is a weak entropy solution if (4.23) is satisfied for all convex entropies $U$. In the scalar case in several space dimensions, Kružkov [45] had shown that it sufficed to consider the family of entropies $U(u) = |u-k|$ for all $k \in \mathbb{R}$, in which case $F(u) = \operatorname{sgn}(u-k)(f(u) - f(k))$. Observe that (4.23) introduces a direction of time, and makes the solution time irreversible, which is in contrast to the situation for linear systems. Next, Lax applies the same approach to difference schemes in a way that turned out to be instrumental in the further development of numerical schemes by, e.g., Wendroff, Harten, Hyman, Osher, Majda, Engquist, and Tadmor.

In [L29], Lax establishes with Wendroff what is now known as the *Lax–Wendroff theorem*: Consider an approximate solution $v$ to the solution $u$ of (4.8) computed by a conservative and consistent finite difference scheme with discretization parameters $\Delta t$ and $\Delta x$. Assume that as $\Delta t, \Delta x \to 0$ the approximation $v$ converges boundedly almost everywhere to some function $u$. Then $u$ is a weak solution of (4.8).

In [L29] and a subsequent paper [L43], Lax and Wendroff analyze what is now known as the *Lax–Wendroff difference scheme* (see Fig. 2), which is a second-order scheme. Convergence of the scheme is proved in [7].

There has of course been an extensive development in the mathematical theory for systems of hyperbolic conservation laws since Lax's landmark solution of the Riemann problem. Most prominent is Glimm's solution of the general Cauchy problem by the random choice method in the case of small variation in the initial data [34] (see also [L116] and [L198]). Indeed to solve the Cauchy problem, one makes a piecewise constant approximation of the initial data, turning it into a multiple Riemann problem. In order to be able to apply Lax's theorem, each jump has to be small, implying the restriction of small total variation. The restriction of small total variation is, except for special systems, a deep part of the theory, and intrinsically linked with hyperbolic conservation laws. The question of uniqueness of the weak entropy solution and the existence of a continuous semigroup of a system of hyperbolic conservation laws on the line remained open for a long time, and it was only fairly recently resolved by Bressan et al. [4] (see also [11] and [35] for a survey) using the method of (wave) front tracking. Finally, the proof of convergence of the viscous limit for systems of hyperbolic conservation laws was obtained by Bianchini and Bressan [3]. Thus one can say that with these results, the theory for solutions with small total variation of systems of hyperbolic conservation laws on the line has become mature. A comprehensive survey of the current state of affairs can be found in [11].

*Speed depends on size*
*Balanced by dispersion*
*Oh solitary splendor*
PETER D. LAX

## 5 Integrable Systems

Most surveys on integrable systems start with the story of the Scottish engineer John Scott Russell and his discovery of solitary waves in the canals outside Edinburgh in 1834. We refer to [5] for the historical background, and we enter the history at a later stage. In 1895, Korteweg and de Vries derived a nonlinear partial differential equation, now universally known as the KdV equation,[11] that models surface waves in shallow water, and that described the phenomena observed by Scott Russell. Furthermore, they showed that the equation had soliton solutions, a term coined by Zabusky and Kruskal [73] in 1965.

The landmark paper in the theory of integrable systems is the Gardner, Greene, Kruskal, and Miura paper [30] from 1967 where they solve the KdV equation

$$u_t - 6uu_x + u_{xxx} = 0, \quad u|_{t=0} = u_0 \tag{5.1}$$

[11] The KdV equation had already been derived in 1871 by Boussinesq, see [61].

by the inverse scattering transform (IST). We can describe the IST as follows. Given initial data such that $\int_{\mathbb{R}}(1+x^2)|u_0(x)|\,dx < \infty$, assume that $u = u(x,t)$ is a solution of the KdV equation. We consider the second order ordinary differential operator[12] $L(t) = -D^2 + u(x,t)$ (writing $D = \frac{d}{dx}$) where $t$ acts as a parameter, and where $u(x,t)$, called the potential, is the solution of the KdV equation. For $L(t)$ one computes the forward (direct) scattering data, that is, the reflection coefficient $R(k,t)$, transmission coefficient $T(k,t)$, the (negative) eigenvalues $\lambda_1(t), \dots, \lambda_n(t)$, and the (positive) norming constants $c_1(t), \dots, c_n(t)$ of appropriately chosen eigenvectors that are square integrable. When the potential $u(x,t)$ satisfies the KdV equation, these quantities have an amazingly simple and explicit behavior in the $t$-variable:

$$\begin{aligned} &R(k,t) = R(k,0)e^{8ik^3t}, && T(k,t) = T(k,0), \\ &\lambda_j(t) = \lambda_j(0), && c_j(t) = c_j(0)e^{4(-\lambda_j)^{3/2}t}, \quad j = 1, \dots, n. \end{aligned} \tag{5.2}$$

Thus the IST works as follows: Compute the forward scattering data for the initial data, let the scattering data evolve in $t$ according to (5.2), and use inverse scattering to determine the potential $u(x,t)$, which then is the required solution of the KdV equation. For the inverse scattering one has the very powerful machinery of Marchenko. It was known that the KdV equation possessed multi-soliton solutions, i.e., localized solutions that interacted almost particle-like, and that the KdV equation had infinitely many conserved quantities. See, e.g., [29, 31].

A big question mark with the IST was the relationship between the KdV equation and the linear ordinary differential equation. Shortly after the IST appeared in [30], Lax published [L60] where he introduced what ever after has been called the *Lax pairs* that unveiled the underlying mechanism for the series of coincidences that appeared for the KdV equation. The key derivation is so short that we can reproduce it here. Consider a one-parameter family $u(t)$ of functions in a space $\mathcal{B}$, and suppose that to each $u \in \mathcal{B}$ we can associate a self-adjoint operator $L(u)$ such that if $u$ satisfies the evolution equation

$$u_t = K(u), \tag{5.3}$$

the operator $L(u(t))$, which we for brevity denote $L(t) = L(u(t))$, remains unitarily equivalent. The unitary equivalence implies the existence of a unitary operator $U(t)$ such that

$$U(t)^{-1}L(t)U(t) \tag{5.4}$$

is time independent. Taking the time derivative in (5.4), we infer

$$-U^{-1}U_tU^{-1}LU + U^{-1}L_tU + U^{-1}LU_t = 0. \tag{5.5}$$

[12] We recognize this operator as the one particle Schrödinger Hamiltonian in nonrelativistic quantum mechanics, but this aspect plays little role in the following except for the fact that the operator has been extensively studied.

A one-parameter family $U(t)$ of unitary operators satisfies

$$U_t = PU \tag{5.6}$$

for some anti-symmetric operator $P = P(t)$. Rewriting (5.5) we find the *Lax relation*

$$L_t = [P, L] \tag{5.7}$$

where $[\cdot, \cdot]$ is the commutator and $L$, $P$ is the *Lax pair*.[13] How does this relate to the KdV equation? First of all, we have $K(u) = 6uu_x - u_{xxx}$. In [30] Gardner, Greene, Kruskal, and Miura had discovered that if $u$ satisfies the KdV equation, then the eigenvalues of the linear operator $L = -D^2 + u$ remain invariant with respect to time. Defining the operator

$$P = -4D^3 + 3uD + 3Du, \tag{5.8}$$

a simple computation reveals that

$$L_t - [P, L] = u_t - 6uu_x + u_{xxx}, \tag{5.9}$$

which is nothing but the KdV equation, rather than a complicated fifth order differential operator. Considerable guesswork was involved here: Given a nonlinear partial differential equation (the KdV equation), could one find a linear operator $L$ and an antisymmetric operator $P$ such that the Lax relation (5.7) is equivalent to the equation itself? In 1972, Zakharov and Shabat showed [75] that also the nonlinear Schrödinger equation possessed a Lax pair, and with a different associated linear operator. This was important because it showed that the Lax pair was not an artifact of the KdV equation, but could eventually be applied to other nonlinear partial differential equations as well. Shortly thereafter, they introduced the method of zero-curvature equations [76, 77], as another means to establish complete integrability. In the zero-curvature formalism one considers a vector-valued function $\Psi = \Psi(x, t, \lambda)$ (where $\lambda$ is a spectral parameter), denoted the Baker–Akhiezer function, that simultaneously satisfies two ordinary differential equations

$$\Psi_x = U\Psi, \qquad \Psi_t = V\Psi \tag{5.10}$$

for given matrices $U = U(u, \lambda)$ and $V = V(u, \lambda)$. The equality of mixed derivatives, i.e., $\Psi_{xt} = \Psi_{tx}$, implies that

$$U_t - V_x + [U, V] = 0, \tag{5.11}$$

which is the zero-curvature relation. The aim is to construct matrices $U$ and $V$ in such a way that (5.11) reduces to a given partial differential equation. In the case of

[13] We follow Gel'fand and write $P$ (Peter) and $L$ (Lax) for the Lax pair. In [L60] Lax used $B$ to denote the operator $P$.

the KdV equation, one choice is

$$U = \begin{pmatrix} 0 & 1 \\ -\frac{1}{4}\lambda + u & 0 \end{pmatrix}, \qquad V = \begin{pmatrix} -u_x & \lambda + 2u \\ (-\frac{1}{4}\lambda + u)(\lambda + 2u) - u_{xx} & u_x \end{pmatrix} \tag{5.12}$$

such that

$$U_t - V_x + [U, V] = \begin{pmatrix} 0 & 0 \\ u_t - 6uu_x + u_{xxx} & 0 \end{pmatrix} = 0, \tag{5.13}$$

which is equivalent with the KdV equation. The discovery of the Lax pair and the zero-curvature formalism started an immense race to study the key equations of mathematical physics to decide if they are completely integrable in the sense of existence of Lax pairs or zero-curvature formalism. By now many of the main equations have been determined to be completely integrable, e.g., the nonlinear Schrödinger equation, the sine-Gordon equation, the AKNS system (developed by Ablowitz, Kaup, Newell, and Segur), the modified KdV equation, the Kadomtsev–Petviashvili equation, the Boussinesq equation, the Thirring equation, the Camassa–Holm equation, the Toda lattice, the Kac–van Moerbeke lattice, and the Ablowitz–Ladik system, and an inverse scattering transform formalism has been set up for each of them.

In the very same paper [L60], Lax observes that the operator $B$ is not the unique antisymmetric operator with the property that $L_t - [B, L]$ is a pure multiplication operator (as opposed to a differential operator); by carefully constructing odd order differential operators $B_{2n+1}$, Lax could construct a hierarchy (now called the *Lax hierarchy* for the KdV equation) of nonlinear partial differential equations with similar properties to the KdV equation, for instance, the whole hierarchy possesses a Hamiltonian structure, as proved by Zakharov and Faddeev [74] for the KdV equation. The Lax hierarchy was the starting point in establishing the very rich connections between nonlinear partial differential equations and algebraic geometry. See, e.g., [2, 33, 54] for a survey. It is difficult to overestimate the importance the notion of Lax pairs;[14] hardly a paper can be written on integrable systems without describing its Lax pair or its zero-curvature formulation.[15]

In papers [L89] and [L91], Lax studies periodic solutions of the KdV equation. He constructs a large family of special solutions, periodic in $x$ and almost periodic in $t$, and study their behavior. His methods were direct, using only calculus in function spaces, properties of the Lax hierarchy, and the existence of infinitely many conserved quantities, thereby revealing many of the fundamental properties of completely integrable systems. This was in a period of very rapid development, and where the powerful machinery of algebraic geometry was applied to these equations. We refer to papers by Dubrovin and Novikov [19–22, 52, 53], Its and Matveev [41, 42], and McKean and van Moerbeke [50].

[14] In Lax's own words [p. 234, L203] "[Lax pairs] occur in surprisingly many contexts".

[15] An unscientific and informal indication of the importance is given if you google on "Lax pair", you come up with more than 25,000 hits (Jan 25, 2008).

The *Lax–Levermore theory* ([L128], [L129], and [L130], and announced in [L112]) concerns the vanishing dispersion limit of the KdV equation, and its represents a landmark in our understanding of dispersive waves and the inverse scattering transform. In the words of P. Deift [p. 611, L215] "The papers by Lax and Levermore constitute one of the earliest successes in turning the inverse scattering transform into a tool for detailed asymptotic analysis". We have seen above that if one considers the inviscid limit of the viscous Burgers equation $u_t^\epsilon + u^\epsilon u_x^\epsilon = \epsilon u_{xx}^\epsilon$, that is, consider the limit $u = \lim_{\epsilon\to 0} u^\epsilon$, one gets the correct entropy solution of $u_t + uu_x = 0$. The Lax–Levermore theory deals with the same question when one replaces the viscous regularization of the inviscid Burgers equation by the dispersive regularization (which then is the KdV equation). Thus we are interested in the weak limit $\epsilon \to 0$ of solutions of

$$u_t^\epsilon - 6u^\epsilon u_x^\epsilon + \epsilon u_{xxx}^\epsilon = 0 \tag{5.14}$$

with initial data $u|_{t=0} = u_0$.

It is rather straightforward to see that the weak limit of the vanishing dispersion equation cannot equal the weak limit of the vanishing diffusion equation. We follow the argument of [L155]: Denote by $\bar{u}$ the weak limit of $u^\epsilon$, the solution of (5.14), as $\epsilon \to 0$. If we rewrite (5.14) in conservative form, it reads

$$u_t^\epsilon - 3\big((u^\epsilon)^2\big)_x + \epsilon u_{xxx}^\epsilon = 0. \tag{5.15}$$

The first and the last term have weak limits equal to $\bar{u}_t$ and zero, respectively, and hence the middle term has a limit as well, which we write as

$$\overline{u^2} = \text{w-lim}_{\epsilon\to 0}(u^\epsilon)^2, \tag{5.16}$$

implying that

$$\bar{u}_t - 3\overline{u^2}_x = 0. \tag{5.17}$$

It is known that if the limit is weak, but not strong we have

$$\overline{u^2} > \bar{u}^2. \tag{5.18}$$

Inserting this into (5.17) we see that

$$\bar{u}_t - 3((\bar{u})^2)_x \neq 0, \tag{5.19}$$

unless $\overline{u^2}$ differs by $\bar{u}^2$ by a constant, which can be ruled out by a separate argument.

The question of the behavior of the vanishing dispersion limit turned out to have a highly nontrivial and very interesting answer using the inverse scattering transform for the KdV equation.

Recall that the scalar conservation law $u_t - 6uu_x = 0$ develops singularities, or shocks, in finite time for generic smooth initial data. The first result says that for times before singularities occur, we have that $u^\epsilon$ converges to $u$, the solution of $u_t - 6uu_x = 0$. However, for times after the classical solution ceases to exist, new phenomena occur. Lax and Levermore show that $u^\epsilon$ converges weakly to $\bar{u}$, i.e., $u^\epsilon \rightharpoonup \bar{u}$, in $L^2(\mathbb{R})$, where

$$\bar{u} = \partial_{xx} Q^* \tag{5.20}$$

with $Q^* = Q^*(x,t)$ determined by

$$Q^*(x,t) = \min_{0 \le \psi \le \phi} Q(\psi; x, t). \tag{5.21}$$

The function $\phi$ equals

$$\phi(\eta) = \mathrm{Re} \int \frac{\eta}{(-u(y) - \eta^2)^{1/2}}\, dy \tag{5.22}$$

and $Q(\psi; x, t)$ is a complicated, explicitly given, function that is linear in $x$ and $t$ while quadratic in $\psi$. The function $u_0 \in C^1(\mathbb{R})$ is assumed to be nonpositive, and have finitely many critical points. Venakides [69] has extended the theory to include positive initial data. The periodic case is treated in [70] and [71]. The fact that $u_0$ is nonpositive is used to replace it by another function $u_0^\epsilon$ with vanishing reflection coefficient and such that $u_0^\epsilon \to u_0$ in $L^2(\mathbb{R})$ as $\epsilon \to 0$. A nice survey of the theory up to 1993 can be found in [L171].

A major step forward was taken when Deift and Zhao [14] developed a novel steepest-decent method for Riemann–Hilbert problems with oscillatory coefficients. Shabat had observed that the inverse scattering transform could be viewed as a Riemann–Hilbert problem. This paved the way for the important result by Deift, Venakides, and Zhao [13] where they applied the results of [14] to determine all coefficients in the asymptotic expansion.

The Lax–Levermore–Venakides theory has also been employed in a number of problems involving asymptotic analysis, we here mention the following: the semiclassical limit of the defocusing nonlinear Schrödinger equation [43], the Toda shock problem [72], the continuum limit of the Toda lattice [12].

## 6 Lax–Phillips Scattering Theory

The collaboration of Lax and Phillips started in 1960 and lasted for over 30 years. It rivals any of the great mathematical collaborations of the 20th century. It had been suggested to both of them on different occasions that theirs was a lot like the Hardy–Littlewood collaboration and in both cases the immediate response was which of them was Hardy and which Littlewood? The primary focus of their joint works is the study of solutions to the wave equation in geometric settings, specifically unbounded domains in Euclidean spaces and on hyperbolic manifolds of finite and infinite volume. Their analysis introduced a mixture of tools from linear hyperbolic partial differential equations, functional analysis and the lovely interplay of these with the geometry of the domain. This continues today to be a central topic in geometric analysis.

The starting point of their collaboration is the paper [L32] establishing the decay of energy for solutions to the wave equation in the exterior of a smooth bounded domain $O$ in $\mathbb{R}^3$. Moreover they show that asymptotically the solutions behave like solutions to the wave equation in free space. This was followed by the joint paper [L33] with Morawetz in which they show in the exterior of a convex domain, that locally the energy decays exponentially as $t$ goes to infinity. These papers contain the elements of the abstract, time dependent, scattering theory developed by Lax and Phillips. This theory is similar in spirit to the abstract spectral theorem but it takes into account via its "incoming and outgoing" subspaces and corresponding completeness statements, the geometry at infinity which is critical to the understanding of the finer features of solutions to the wave equation. A comprehensive treatment of their Euclidean Scattering Theory is given in their well known monograph [L53]. The revised edition of [L53] from 1989 contains an up-to-date epilogue with a wealth of information about recent developments stemming from their earlier works. In their paper [L42] one finds far-reaching insights and Conjectures about the relation between the motion of rays in the exterior of the obstacle $O$ and the distribution of the scattering poles. The latter are identified in terms of the eigenvalues of the fundamental compact (not self-adjoint) operator $B$ which is the infinitesimal generator of the semi-group $Z(t)$ associated with the abstract setup of the scattering problem. $O$ is said to be non-trapping if there is a ball $C$ in $\mathbb{R}^n$ containing $O$ and a time $t_0$ such that any ray entering $C$ and obeying linear motion outside of $O$ and standard reflection on hitting $O$, exits $C$ within the time $t_0$. A basic conjecture put forth in [L42] is that the semi-group $Z(t)$ is eventually compact (which is equivalent to the imaginary parts of the poles of the scattering matrix tending to infinity) if and only if $O$ is non-trapping. If $O$ is trapping this conjecture was established by Ralston [63]. In the other direction Ludwig and Morawetz [47] established the Conjecture if $O$ is convex. After progress on some other special cases the general case of the Conjecture was proven by Melrose in [48]. His work makes use of the modern machinery of propagation of singularities for hyperbolic equations and related microlocal analysis. Not coincidentally this modern theory [39] has its roots in Lax's 1957 paper [L23] mentioned in Sect. 2.

A further study of the distribution of the poles of the scattering operator is contained in [L61], in analogy with the well studied Weyl Law for the counting of eigenvalues of the Laplacian on a compact domain. However unlike that case the poles here are not confined to lie on a line and this complicates the problem considerably. Lax and Phillips determine the order of magnitude of the number of purely imaginary poles. More recently Melrose and Zworski [49, 78] have estimated the number of poles in a large ball while Ikawa [40] gives a precise description of the location of the poles near the real axis in the case that $O$ consists of two convex bodies containing an unstable ray between them. All of these topics remain active areas of investigation even today.

Faddeev and Pavlov [27] observed that Lax and Phillips's abstract set up for scattering theory in Euclidean spaces is also well suited in the context of non-compact but finite area hyperbolic surfaces. This led Lax and Phillips to develop their theory in this context and the outcome was a series of far-reaching papers on this and related topics. The corresponding spectral theory for finite volume hyperbolic manifolds is due to Selberg [65]. His main results being the analytic continuation of Eisenstein series and the development of the trace formula. In [L77] and [L91], Lax and Phillips develop their scattering theory approach for the continuous spectrum on these manifolds. Their method is based on their semi-group $Z(t)$ and its infinitesimal generator $B$. It yields a short and conceptual proof of the analytic continuation of the Eisenstein series. It also identifies the poles of the latter in terms of the eigenvalues of the compact operator $B$. This is particularly well suited for studying the behaviour of these poles when the surface is deformed. The Lax and Phillips theory was used in [62] as a basis for such a study. The cut-off Laplacian (associated with a cusp of the manifold) was introduced in the monograph [L77] and in the hands of Colin de Verdiére [8] and Müller [51], it is a fundamental tool in the proof of the trace class conjecture for the spectrum of a general locally symmetric space.

Papers [L103] and [L115] and the series [L116], [L127], [L128], [L134] are concerned with the Radon transform in non-Euclidean spaces, and related Paley–Wiener theorems. These are developed in order to carry out their scattering theory for infinite volume but geometrically finite hyperbolic manifolds. Prior to these papers little was known about these. For the case of surfaces Patterson [57–59] developed the spectral theory but his methods were limited to two dimensions. Lax and Phillips establish the basic properties of the spectrum, that is, the finiteness of the point spectrum below the continuous spectrum and the absolute continuity of the rest of the continuous spectrum. Their series of papers on translation representations for solutions of the wave equation for these manifolds yield a complete spectral and scattering theory. They stop short of obtaining a trace formula in this context. One may view the more recent work [60] as a substitute for the trace formula in this infinite volume context.

The paper [L116] investigates the analogue of the well known problem of Gauss of counting asymptotically the number of lattice points in a large circle, in the context of Euclidean and hyperbolic spaces. It was shown by Delsarte [15] some time ago that for the case of the hyperbolic plane with a co-compact lattice acting isometrically, the spectrum of the Laplacian on the quotient can be used to obtain the leading term of the count of the number of lattice points in a large hyperbolic disk. In [L116], Lax and Phillips use the wave equation and their theory to obtain the asymptotics of lattice points in a large ball, for any geometrically finite group. The remainder term that they obtain for this count is the sharpest such result known in all cases. Before their work, Selberg in unpublished work had obtained similar results in the finite volume case. These counting problems can be generalized to other symmetric spaces and have Diophantine applications. This is an active area of current research (see, for example, [25] and [28]).

More recently in papers [L195], [L199] and [L200] with Francsics, Lax gives an explicit fundamental domain for the Picard group in $\mathrm{SU}(2,1)$. This opens the way for a numerical as well as a concrete investigation of the spectral theory of such quotients.

To end this brief survey of Lax's work in scattering theory and functional analysis we mention his 2002 text *Functional Analysis*, [L187]. There are many good texts on this subject, however this one with its many examples, enticing applications and clear development of the theory, is a classical. It gives students and researchers an excellent opportunity to learn this basic subject from one of its masters.

**Acknowledgements** HH is grateful to Nils Henrik Risebro for providing the figures, and to Constantine Dafermos, Fritz Gesztesy, Harald Hanche-Olsen, and Kenneth H. Karlsen for careful reading of the manuscript

## References

1. Albers, D.J., Alexanderson, G.L., Reid, C. (eds.): More Mathematical People. Hartcourt Brace Jovanovich, Boston (1990)
2. Belokolos, E.D., Bobenko, A.I., Enol'skii, V.Z., Its, A.R., Matveev, V.B.: Algebro-Geometric Approach to Nonlinear Integrable Equations. Springer, Berlin (1994)
3. Biachini, S., Bressan, A.: Vanishing viscosity solutions of nonlinear hyperbolic systems. Ann. Math. (2) **61**, 223–342 (2005)
4. Bressan, A.: Hyperbolic Systems of Conservation Laws. Oxford Univ. Press, Oxford (2000)
5. Bullough, R.K., Caudrey, P.J.: Solitons and the Korteweg–de Vries equation: Integrable systems in 1834–1995. Acta Appl. Math. **39**, 193–228 (1995)
6. Chen, G.-Q.: Convergence of the Lax–Friedrichs scheme for isentropic gas dynamics. III. Acta Math. Sci. **6**, 75–120 (1986)
7. Chen, G.-Q., Liu, J.-G.: Convergence of difference schemes with high resolution for conservation laws. Math. Comput. **66**, 1027–1053 (1997)
8. Colin de Verdiére, Y.: Pseudo-Laplacians. Ann. Inst. Fourier (Grenoble) **32**, 275–286 (1982)
9. Conway, E., Smoller, J.: Global solutions of the Cauchy problem for quasilinear first-order equations in several space variables. Commun. Pure Appl. Math. **19**, 95–105 (1966)
10. Courant, R., Friedrichs, K.O.: Supersonic Flow and Shock Waves. Springer, New York (1976). (First published in 1948)
11. Dafermos, C.M.: Hyperbolic Conservation Laws in Continuum Physics, 2nd edn. Springer, New York (2002)
12. Deift, P., McLaughlin, K.T.-R.: A continuum limit of the Toda lattice. Mem. Am. Math. Soc. **624**, x+216 (1998)
13. Deift, P., Venakides, S., Zhao, X.: New results in small dispersion KdV by an extension of the steepest descent method for Riemann–Hilbert problems. Internat. Math. Res. Notices **1997**, 286–299 (1997)
14. Deift, P., Zhao, X.: A steepest descent method for oscillatory Riemann–Hilbert problems. Asymptotics for the MKdV equation. Ann. Math. (2) **137**, 294–368 (1993)
15. Delsarte, J.: Sur le gitter fuchsien. C. R. Acad. Sci. Paris **214**, 147–179 (1942)
16. Ding, X.X., Chen, G.-Q., Luo, P.Z.: Convergence of the Lax–Friedrichs scheme for isentropic gas dynamics. I. Acta Math. Sci. **5**, 415–432 (1985)
17. Ding, X.X., Chen, G.-Q., Luo, P.Z.: Convergence of the Lax–Friedrichs scheme for isentropic gas dynamics. II. Acta Math. Sci. **5**, 433–472 (1985)
18. Ding, X.X., Chen, G.-Q., Luo, P.Z.: A supplement to the papers: "Convergence of the Lax–Friedrichs scheme for isentropic gas dynamics. II, III". Acta Math. Sci. **9**, 43–44 (1989)
19. Dubrovin, B.A.: Inverse problem for periodic finite-zoned potentials in the theory of scattering. Funct. Anal. Appl. **9**, 61–62 (1975)
20. Dubrovin, B.A.: Periodic problems for the Korteweg–de Vries equation in the class of finite band potentials. Funct. Anal. Appl. **9**, 215–223 (1975)

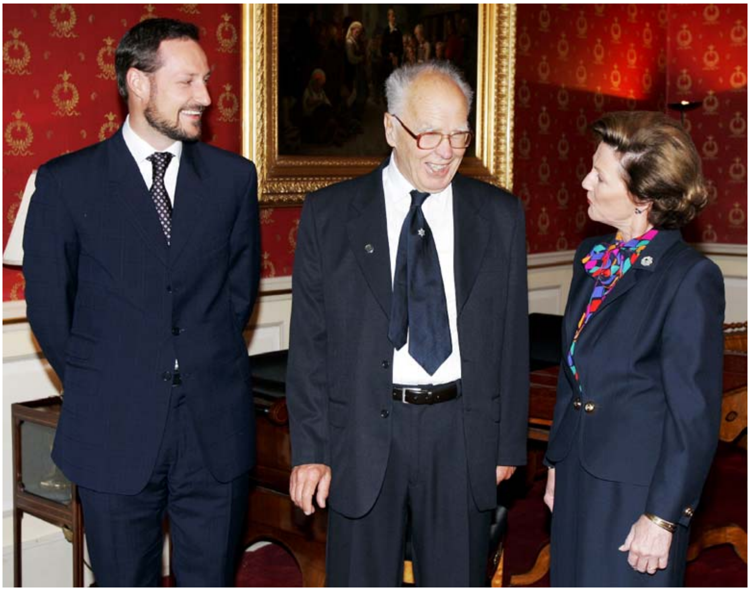
21. Dubrovin, B.A., Novikov, S.P.: A periodicity problem for the Korteweg–de Vries and Sturm–Liouville equations. Their connection with algebraic geometry. Dokl. Akad. Nauk SSSR **15**, 1597–1601 (1974)
22. Dubrovin, B.A., Novikov, S.P.: Periodic and conditionally periodic analogs of the many-soliton solutions of the Korteweg–de Vries equation. Sov. Phys. JETP **40**, 1058–1063 (1975)
23. Duistermaat, J.J.: Fourier Integral Operators. Birkhäuser, Basel (1996)
24. Duistermaat, J.J., Guillemin, V.W.: The spectrum of positive elliptic operators and periodic bicharacteristics. Invent. Math. **29**, 39–79 (1975)
25. Duke, W., Rudnick, Z., Sarnak, P.: Density of integer points on affine homogeneous varieties. Duke Math. J. **71**, 143–179 (1993)
26. Evans, L.C.: Partial Differential Equations. Am. Math. Soc., Providence (1998)
27. Faddeev, L., Pavlov, B.: Scattering theory and automorphic functions. Proc. Steklov Inst. Math. **27**, 161–193 (1972)
28. Gorodnik, A., Oh, H., Shah, N.: Integral points on symmetric varieties and Satake compatifications: Preprint, arXiv:math/0610497v2 (2008)
29. Gardner, C.S.: Korteweg–de Vries equation and generalizations. IV. The Korteweg–de Vries equation as a Hamiltonian system. J. Math. Phys. **12**, 1548–1551 (1971)
30. Gardner, C.S., Greene, J.M., Kruskal, M.D., Miura, R.M.: Method for solving the Korteweg–de Vries equation. Phys. Rev. Lett. **19**, 1095–1097 (1967)
31. Gardner, C.S., Greene, J.M., Kruskal, M.D., Miura, R.M.: Korteweg–de Vries equation and generalizations. VI. Methods for exact solution. Commun. Pure Appl. Math. **27**, 97–133 (1974)
32. Gel'fand, I.M.: Some problems in the theory of quasilinear equations. Am. Math. Soc. Transl. **29**, 295–381 (1963)
33. Gesztesy, F., Holden, H.: Soliton Equations and Their Algebro-Geometric Solutions. $(1+1)$-Dimensional Continuous Models, vol. I. Cambridge Univ. Press, Cambridge (2003)
34. Glimm, J.: Solutions in the large for nonlinear hyperbolic systems of equations. Commun. Pure Appl. Math. **18**, 697–715 (1965)
35. Holden, H., Risebro, N.H.: Front Tracking for Hyperbolic Conservation Laws. Springer, New York (2007). Corrected 2nd printing
36. Hopf, E.: The partial differential equation $u_t + uu_x = \mu u_{xx}$. Commun. Pure Appl. Math. **3**, 201–230 (1950)
37. Hörmander, L.: Pseudo-differential operators and non-elliptic boundary value problems. Ann. Math. **83**, 129–209 (1966)
38. Hörmander, L.: The spectrum of a positive elliptic operator. Acta Math. **121**, 193–218 (1968)
39. Hörmander, L.: Fourier integral operators. I. Acta Math. **127**, 79–183 (1971)
40. Ikawa, M.: On the poles of the scattering matrix for two strictly convex obstacles. J. Math. Kyoto Univ. **23**, 127–194 (1983). Addendum, *loc. sit.* **23**, 795–802 (1983)
41. Its, A.R., Matveev, V.B.: Hill's operator with finitely many gaps. Funct. Anal. Appl. **9**, 65–66 (1975)
42. Its, A.R., Matveev, V.B.: Schrödinger operators with finite-gap spectrum and $N$-soliton solutions of the Korteweg–de Vries equation. Theoret. Math. Phys. **23**, 343–355 (1975)
43. Jin, S., Levermore, C.D., McLaughlin, D.W.: The behavior of solutions of the NLS equation in the semiclassical limit. In: Ercolani, N.M., Gabitov, I.R., Levermore, C.D., Serre, D. (eds.) Singular Limits of Dispersive Waves. NATO Adv. Sci. Inst. Ser. B, Phys., vol. 320, pp. 235–255. Plenum, New York (1994)
44. Klainerman, S., Majda, A.: Formation of singularities for wave equations including the non-linear vibrating string. Commun. Pure Appl. Math. **33**, 241–263 (1980)
45. Kružkov, S.N.: First order quasi-linear equations in several independent variables. Math. USSR Sb. **10**, 217–243 (1970)
46. LeVeque, R.J.: Finite Volume Methods for Hyperbolic Problems. Cambridge Univ. Press, Cambridge (2002)
47. Ludwig, D., Morawetz, C.: The generalized Huygens' principle for reflecting bodies. Commun. Pure Appl. Math. **22**, 189–205 (1969)

48. Melrose, R.: Singularities and energy decay in acoustical scattering. Duke Math. J. **46**, 43–59 (1979)
49. Melrose, R.: Polynomial bound on the number of scattering poles. J. Funct. Anal. **53**, 287–303 (1983)
50. McKean, H.P., van Moerbeke, P.: The spectrum of Hill's equation. Invent. Math. **30**, 217–274 (1975)
51. Müller, W.: The trace class conjecture in the theory of automorphic forms. Ann. Math. **130**, 473–529 (1989)
52. Novikov, S.P.: The periodic problem for the Korteweg–de Vries equation. Funct. Anal. Appl. **8**, 236–246 (1974)
53. Novikov, S.P.: A method for solving the periodic problem for the KdV equation and its generalizations. Rocky Mountain J. Math. **8**, 83–93 (1978)
54. Novikov, S.P., Manakov, S.V., Pitaevskii, L.P., Zakharov, V.E.: Theory of Solitons. Consultants Bureau, New York (1984)
55. Oleĭnik, O.A.: Discontinuous solutions of non-linear differential equations. Am. Math. Soc. Transl. Ser. **26**, 95–172 (1963)
56. Oleĭnik, O.A.: Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation. Am. Math. Soc. Transl. Ser. **33**, 285–290 (1963)
57. Patterson, S.: The Laplacian operator on a Riemann surface. I. Compos. Math. **32**, 83–107 (1975)
58. Patterson, S.: The Laplacian operator on a Riemann surface. II. Compos. Math. **32**, 71–112 (1976)
59. Patterson, S.: The Laplacian operator on a Riemann surface. III. Compos. Math. **33**, 227–259 (1976)
60. Patterson, S.J., Perry, P.A.: The divisor of Selberg's zeta function for Kleinian groups. Duke Math. J. **106**, 321–390 (2001)
61. Pego, R.: Origin of the KdV equation. Not. Am. Math. Soc. **45**, 358 (1998)
62. Phillips, R., Sarnak, P.: Perturbation theory for the Laplacian on automorphic functions. J. Am. Math. Soc. **5**, 1–32 (1992)
63. Ralston, J.: Solutions of the wave equation with localized energy. Commun. Pure Appl. Math. **22**, 807–823 (1969)
64. Riemann, G.F.B.: Ueber die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. Abh. König. Gesell. Wiss. Göttingen **8**, 43–65 (1860)
65. Selberg, A.: Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. J. Indian Math. Soc. **20**, 47–87 (1956)
66. Serre, D.: Systems of Conservation Laws, vol. 1. Cambridge Univ. Press, Cambridge (1999)
67. Smoller, J.: Shock Waves and Reaction–Diffusion Equations, 2nd edn. Springer, New York (1994)
68. Sogge, C., Zelditch, S.: Riemannian manifolds with maximal eigenfunction growth. Duke Math. J. **114**, 387–437 (2002)
69. Venakides, S.: The zero dispersion limit of the Korteweg–de Vries equation for initials with nontrivial reflection coefficient. Commun. Pure Appl. Math. **38**, 125–155 (1985)
70. Venakides, S.: The zero dispersion limit of the Korteweg–de Vries equation for initials with periodic initial data. Trans. Am. Math. Soc. **301**, 189–226 (1987)
71. Venakides, S.: The continuum limit of theta functions. Commun. Pure Appl. Math. **42**, 711–728 (1989)
72. Venakides, S., Deift, P., Oba, R.: The Toda shock problem. Commun. Pure Appl. Math. **44**, 1171–1242 (1991)
73. Zabusky, N.J., Kruskal, M.D.: Interaction of "solitons" in a collisionless plasma and the recurrence of initial states. Phys. Rev. Lett. **15**, 240–243 (1965)
74. Zakharov, V.E., Faddeev, L.D.: Korteweg–de Vries equation: A completely integrable Hamiltonian system. Funct. Anal. Appl. **5**, 280–287 (1971)
75. Zakharov, V.E., Shabat, A.B.: Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. Sov. Phys. JETP **34**, 62–69 (1972)

76. Zakharov, V.E., Shabat, A.B.: A scheme for integrating the nonlinear equations of mathematical physics by the method of the inverse scattering problem. I. Funct. Anal. Appl. **8**, 226–235 (1974)
77. Zakharov, V.E., Shabat, A.B.: Integration of nonlinear equations of mathematical physics by the method of inverse scattering. II. Funct. Anal. Appl. **13**, 166–174 (1979)
78. Zworski, M.: Sharp polynomial bounds on the number of scattering poles for radial potentials. J. Funct. Anal. **82**, 370–403 (1989)

# 2006

# Lennart Carleson



ABEL
PRISEN

# Carleson for Beginners

**Tom Körner**

When I was just beginning as a research student, an older mathematical friend invited me to dine at Trinity. At dessert I was seated next to Littlewood, Hardy's collaborator and a legendary figure in modern British analysis. With old fashioned politeness, Littlewood set himself out to entertain me. He talked about a recent comet and recalled how fifty years earlier he had viewed Halley's comet in company with a Trinity fellow who had himself seen its previous visitation. He then spoke about Carleson's recent proof of the convergence theorem, what a marvellous result it was, how surprising it was that it turned out that Lusin's conjecture was true, how many people known to him had thought about the problem for a long time without success and how much he regretted being too old to take up the task of understanding the details of the proof.

I shall try to explain the importance of the result at the level of a well informed and able first year mathematics student. Non-mathematicians should not worry about the mathematics but just read the story. Mathematicians should note that I will make no attempt to explain the proof itself. I shall then talk more briefly about two other famous results of Carleson. Finally, in the last four or five pages, I shall address non-mathematicians directly.

Students often think of mathematics as consisting of definitions and theorems. It would be more correct to think of mathematics as consisting of examples and methods. A definition tells us nothing unless it is illuminated by concrete examples and counterexamples and can lead nowhere unless we have accompanying methods of proof.

The early 19th century saw the ultimately successful attempt to base calculus on a new type of 'epsilon, delta' definition. In 1829, Dirichlet gave one of the first and most impressive examples of the 'new analysis' by applying it to 'Fourier series'. In his work on heat conduction Fourier had given plausible arguments to show that

T. Körner (✉)
Department of Pure Mathematics and Mathematical Statistics, Centre for Mathematical Sciences, University of Cambridge, Clarkson Road, Cambridge, UK
e-mail: twk@dpmms.cam.ac.uk

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_11, © Springer-Verlag Berlin Heidelberg 2010

any reasonable periodic function (with period $L$) could be represented by its Fourier series

$$f(x)=\frac{A_0}{2}+\sum_{r=-\infty}^{\infty}\left[A_r\cos\left(\frac{2\pi rx}{L}\right)+B_r\sin\left(\frac{2\pi rx}{L}\right)\right].$$

We shall take $L=1$ and use the equivalent formulation

$$f(x)=\sum_{r=-\infty}^{\infty}a_r\exp(2\pi irx).$$

As Fourier and others before him had noted, the only reasonable choice for the $a_r$ is

$$a_r=\hat{f}(r)=\int_{-1/2}^{1/2}f(x)\exp(-2\pi irx)\,dx.$$

'Fourier sums', and the closely related Fourier integrals, occur naturally in optics, communication theory and more generally in any physical problem involving oscillation or waves. They also occur in unexpected places like number theory and methods for fast machine computation.

If we write

$$S_n(f,t)=\sum_{r=-n}^{n}\hat{f}(r)\exp(2\pi irt),$$

then the key problem facing Dirichlet and his successors was to find wide conditions under which

$$S_n(f,t)\to f(t)$$

as $n\to\infty$ and to prove that these conditions were indeed sufficient. The first observation to make is that

$$\begin{aligned}\sum_{r=-n}^{n}\hat{f}(r)\exp 2\pi irt&=\sum_{r=-n}^{n}\int_{-1/2}^{1/2}f(x)\exp(-2\pi irx)\,dx\exp(irt)\\&=\int_{-1/2}^{1/2}f(x)\sum_{r=-n}^{n}\exp\big(2\pi ir(t-x)\big)\,dx\\&=\int_{-1/2}^{1/2}f(x)K_n(t-x)\,dx,\end{aligned}$$

where (summing a geometric series)

$$K_n(s)=\sum_{r=-n}^{n}\exp(2\pi irs)=\frac{\sin(2\pi(n+\frac{1}{2})s)}{\sin\pi s}.$$

(To avoid division by zero, we set $K_n(0)=2n+1$, but this creates no problems.)

Unfortunately $K_n$ is not very well behaved. The reader who sketches the graph of $K_{10}$ will see that $K_n$ must be highly oscillatory and will not be surprised to learn that

$$\int_{-1/2}^{1/2} |K_n(s)|\, ds \to \infty$$

as $n \to \infty$ (indeed the integral grows at the same rate as $\log n$). However we do have

$$\int_{-1/2}^{1/2} K_n(s)\, ds = 1,$$

so, if $f$ is constant, then the 'oscillations cancel' and

$$S_n(f,t) = f(t).$$

We can therefore hope that, if $f$ is sufficiently well behaved, then, as $n$ gets larger and the oscillations 'crowd together', they will start to cancel and

$$S_n(f,t) = \int_{-1/2}^{1/2} f(x)K_n(t-x)\, dx \to f(t)$$

as $n \to \infty$. By careful estimation, Dirichlet was able to show that if $f$ is continuous and has only a finite number of maxima and minima, this is indeed the case.

It is possible that Dirichlet may, for a time, have thought that he could extend his result to all continuous functions. It is certain that most mathematicians and physicists thought that such a result was true. However, in 1873, du Bois-Reymond gave an example of a continuous function $f$ such that $S_n(f,0)$ fails to converge. In retrospect, the finding of such an example illustrates how, during the 19th century, mathematicians acquired a menagerie of continuous functions, came to understand the freedom of behaviour that epsilon-delta definitions allowed and developed new techniques for controlling that freedom.

Observe that it is easy to find a three times differentiable periodic function $f_n$ such that $|f_n(t)| \leq 1$ for all $t$ and

$$S_n(f_n,0) = \int_{-1/2}^{1/2} f_n(x)K_n(-x)\, dx \geq \frac{1}{2}\int_{-1/2}^{1/2} |K_n(-x)|\, dx.$$

It is plausible that, provided $N(j)$ increases fast enough,

$$f(t) = \sum_{j=1}^{\infty} 2^{-j} f_{n(j)}(t)$$

will define a continuous function such that

$$S_{n(j)}(f,0) \to \infty,$$

and this is indeed the case although the details of the proof require some care.

There are two remarks which are worth making. The first is that a natural approach to the proof is via the notion of uniform convergence. From the modern point of view, this involves treating continuous functions as points in (or if the reader prefers, elements of) a space of continuous functions and considering the 'distance' between two points (or elements) $F$ and $G$ say to be given by

$$d(F,G) = \|F - G\|_\infty = \sup_t |F(t) - G(t)|.$$

Our second remark is the following. Let the sequence $x_j$ contain each rational number infinitely often. If we modify the argument of the previous paragraph by considering

$$f(t) = \sum_{j=1}^{\infty} 2^{-j} f_{n(j)}(t - x_j),$$

then, provided we choose the $n_j$ increasing fast enough, it is both plausible and true that we will define a continuous function $f$

$$S_{n(j)}(f, x_j) \to \infty,$$

and so the Fourier sum $S_n(f,t)$ will fail to converge at every rational point $t$. A look at Dirichlet's argument shows that this means that $f$ has an infinite number of maxima and minima in each interval.

Of course, we can restrict ourselves to functions which satisfy Dirichlet's conditions or something similar. (If we only look at functions which are once continuously differentiable, then their Fourier sums are absolutely convergent and most of the analytic difficulties vanish.) However, in many cases, when we try to apply the results to natural problems, we find that much of our effort goes into dealing with the peripheral difficulties caused by these restrictions. It is not surprising that the subject began to stagnate. Although the point of view of this essay is very different from that of Klein, we can use his words to describe the state of Fourier Analysis at the end of the 19th century as 'like a large weapon shop in peace time. The store window is filled with showpieces whose ingenious artful and pleasing design enchants the connoisseur. The real purpose of these things, to attack and defeat the enemy, has retreated so far into the background of consciousness as to be forgotten' [2].

Fourier Analysis and analysis in general was revivified by the invention by Lebesgue and others of measure theory. One way of understanding this new idea is to consider the relation of the rational numbers to the reals. The rational numbers form a readily comprehensible algebraic system with an obvious notion of distance. Unfortunately this system behaves badly when we apply limiting processes. As a simple example consider the

$$1,\ 1+\frac{1}{1!},\ 1+\frac{1}{1!}+\frac{1}{2!},\ 1+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!},\ \dots$$

which 'ought to converge' but does not. To avoid this problem we embed the rationals in the larger system of the reals where everything that 'ought to converge' does. (More formally, every Cauchy sequence converges.) However, the new system of reals remains strongly linked to the old since every real number is the limit of a sequence of rationals. We say that the real numbers form a ‘completion’ of the rationals and that the rationals are ‘dense in the reals’. Since the rationals form a subset of the reals we can prove results about the rationals by using results about the reals, and since the rationals are dense in the reals, we can prove results about the reals by using results about the rationals.

Lebesgue showed that the notion of integration can be extended to cover a very large class (called $L^1$) of functions on the interval $[-1/2, 1/2]$. Any bounded function that the reader can write down explicitly will represent an object in $L^1$, but the chief merit of $L^1$ is that it bears much the same relation to the continuous functions as the real numbers do to the rationals. Let us measure the ‘distance’ between two functions $f$ and $g$ in $L^1$ by

$$d_1(f,g) = \|f-g\|_1 = \int_{-1/2}^{1/2} |f(t)-g(t)|\,dt.$$

Then every sequence in $L^1$ that ‘ought to converge’ does, and so it is easy to do analysis in $L^1$, but every $f \in L^1$ is the limit of continuous functions, that is to say, given $f \in L^1$ we can find continuous functions $f_n$ with

$$\|f - f_n\|_1 \to 0.$$

Since ‘the continuous functions are dense in $L^1$’, we can prove results about $L^1$ by using results about continuous functions and since every continuous function lies in $L^1$, every result on $L^1$ implies a result on continuous functions.

There is a small price to pay for extending our interest from the continuous functions to $L^1$, but it is a price well worth paying. We say that a subset $E$ of $[-1/2, 1/2]$ has measure zero if, given any $\epsilon > 0$, we can cover $E$ with a countable collection of intervals of total length less than $\epsilon$. (Another way to think of this is that, if you drop a dart at random on the interval $[-1/2, 1/2]$, the probability of hitting $E$ is zero.) It turns out that if two functions $f$ and $g$ in $L^1$ only differ on a set of measure zero then $\|f-g\|_1 = 0$ and $f$ and $g$ are indistinguishable as objects in $L^1$.

The set $L^1$ has a very important subset $L^2$ (also called the space of square integrable functions) consisting of those $f \in L^1$ such that

$$\int_{-1/2}^{1/2} |f(t)|^2\,dt < \infty.$$

A simple modification of the argument used to prove the Cauchy–Schwarz inequality shows that we can introduce a new ‘distance’ between functions $f$ and $g$ in $L^2$ given by

$$d_2(f,g) = \|f-g\|_2 = \left(\int_{-1/2}^{1/2} |f(t)-g(t)|^2\,dt\right)^{1/2}.$$

The distance $d_2$ is a natural generalisation of Euclidean distance and $L^2$ turns out to give the most natural generalisation of Euclidean ideas to infinite dimensional spaces.

Although the distances $d_1$ and $d_2$ are very different, it turns out that every sequence in $L^2$ which ought to converge under the new metric $d_2$ does indeed converge to a function in $L^2$ and that the continuous functions are dense in $L^2$ under the new metric.

It requires a fair amount of work to set up Lebesgue's theory, but, once this is done, many results which were 'almost true' in the old analysis become 'exactly true' and many arguments that 'almost worked' now work and work easily. For example Fourier analysis within $L^2$ can be condensed into four easily proved statements. (We write $e_n(t) = \exp(2\pi int)$.)

(1) If $f \in L^2$, then $\sum_{j=-\infty}^{\infty} |\hat{f}(j)|^2 < \infty$.

(2) If $\sum_{j=-\infty}^{\infty} |a_j|^2 < \infty$, then we can find an $f \in L^2$ with $\hat{f}(j) = a_j$ for all $j$.

(3) If $f, g \in L^2$ and $\hat{f}(j) = \hat{g}(j)$ for all $j$, then the set of $t$ with $f(t) \neq g(t)$ has measure zero (that is to say $f$ and $g$ are identical from the point of view of Lebesgue's theory).

(4) If $f \in L^2$, then

$$\left\| \sum_{j=-n}^{n} \hat{f}(j)e_j - f \right\|_2 \to 0$$

as $n \to \infty$.

The behaviour of Fourier series in $L^1$ is not so simple, but, here again, Lebesgue's theory made many of the old rough paths smooth. In mathematics, as in other branches of learning, the adoption of new methods must often wait until a new generation replaces the old. However, even established mathematicians like Hardy and Birkhoff adopted the new ideas with enthusiasm. For a short time it appeared that difficulties could be confined to a set of measure zero and that any set of measure zero could be ignored.

If we look at the kind of functions $f$ which appear in constructions modeled on those of du Bois-Reymond we see that, although the set $E$ on which $S_n(f,t)$ fails to converge may be very complicated, it always seems to have measure zero. It is natural to conjecture that, if $f$ is continuous, $S_n(f,t) \to f(t)$ as $n \to \infty$ except when $t$ belongs to some set of measure zero. (More briefly $S_n(f,t) \to f(t)$ except on a set of measure zero.) Lusin strengthened this conjecture to cover all $f \in L^2$. (The reader should not assign Lusin the role of a lucky bit player in the drama. He was a deep mathematician who among other things was the first to construct a sequence $a_j \to 0$ as $|j| \to \infty$ but $\sum_{j=-n}^{n} a_j \exp(2\pi ijt)$ diverges at every point $t$.)

In 1922, at the age of 19, Kolmogorov shot to international fame by showing that Lusin's conjecture is false if we replace $L^2$ by $L^1$. In its final form, his result showed that there is an $f \in L^1$ such that $S_n(f,t)$ diverges for all $t$. His proof illustrates an important property of the Dirichlet kernel $K_n$. If we fix $1/2 > \delta > 0$ then, although $\int_{|s|\geq\delta} |K_n(s)|\,ds$ remains bounded as $n \to \infty$, it does not tend to zero. The increasing oscillation of $K_n$ still means that

$$\int_{|s|\geq\delta} K_n(s)f(s)\,ds \to 0,$$

and so the limiting behaviour of $S_n(f, 0)$ does not depend on the values of $f(s)$ with $1 \geq |s| \geq \delta$. (This is the Riemann localisation principle and states that *ultimate behaviour* of the Fourier sum for $f$ at a point only depends on the value of $f$ near that point.) Kolmogorov's construction depends, in part, on the fact that for $f \in L^1$ it may take an arbitrarily long time for localisation to assert itself.

The reader should note that (though the actual behaviour is much more complicated) the divergence in Kolmogorov's example is more akin to the behaviour of $g_n$ where

$$g_{2^m+r}(t) = \begin{cases} n & \text{for } 2^{-1} + r2^{-m} \leq t \leq 2^{-1} + (r+1)2^{-m}, \text{ and } 0 \leq r \leq 2^m - 1,\\ 0 & \text{otherwise,}\end{cases}$$

(so that, at each point $t$, long periods of good behaviour are interrupted by occasional periods of bad behaviour) than $h_n$ where $h_n(t) = (-1)^n n$ and bad behaviour occurs everywhere all the time.

It seems likely that for the next forty years most mathematicians expected that some sort of tweaking of Kolmogorov's example would produce a counterexample to Lusin's conjecture. It is certainly true that, although several important theorems were obtained which proved the convergence of Fourier sums under various conditions, none of these results suggested that Lusin's conjecture was true.

However, there were advances in other parts of analysis which, we now see, shed light on the problem. Earlier, I said that it was possible to extend results from the rational numbers to the reals by a density argument. However, we must exercise caution. Consider the function $f : \mathbb{Q} \to \mathbb{Q}$ given by

$$f(x) = \begin{cases} 1 & \text{if } x^2 < 2,\\ 0 & \text{otherwise.}\end{cases}$$

Although $f$ is continuous (the reader who doubts this is asked to find a point $t \in \mathbb{Q}$ where $f$ is not continuous), we cannot find a continuous extension $\tilde{f}$ to the reals. (In other words, we cannot find a continuous $\tilde{f}\colon \mathbb{R} \to \mathbb{R}$ with $\tilde{f}(x) = f(x)$ for all $x \in \mathbb{Q}$.) On the other hand, if $g\colon \mathbb{Q} \to \mathbb{Q}$ is *uniformly* continuous, then it is easy to see that it has a continuous (indeed a uniformly continuous) extension $\tilde{g}$ to the reals.

Simple minded attempts to use this idea to tackle the kind of problem we are considering are bound to fail. The trigonometric polynomials (that is to say functions of the form $\sum_{j=-n}^{n} a_j \exp(2\pi ijt)$) are dense in $L^2$, but it is not true that every function in $L^2$ is a trigonometric polynomial.

A way forward was provided by the introduction by Hardy and Littlewood of the idea of a maximal function and the more general idea of a 'maximal inequality'. One outstanding result proved by these means was Birkhoff's ergodic theorem. Suppose that $T\colon [-1/2, 1/2] \to [-1/2, 1/2]$ is a bijection which 'preserves measure' in the sense that

$$\int_{-1/2}^{1/2} g(Tx)\,dx = \int_{-1/2}^{1/2} g(x)\,dx$$

for all $g \in L^1$. Then Birkhoff's ergodic theorem tells us that, if $f \in L^1$, then

$$\frac{f(x) + f(Tx) + \cdots + f(T^n x)}{n+1}$$

tends to a limit for all $x$ not in some set of measure zero. The standard proof of the ergodic theorem depends on identifying a set of 'good' functions for which the result is easy to prove and which are dense in $L^1$ and then using a 'maximal inequality' to extend it to all functions in $L^1$.

We illustrate these ideas in a very simple case, but, for once in this essay, instead of taking about mathematics, we shall actually do some. Our object is to prove the following theorem.

**Theorem 1** *Suppose $a_j \in \mathbb{R}$ and $\sum_{j=1}^{\infty} a_j^2$ converges. If $X_1, X_2, \ldots$ are independent identically distributed random variables with*

$$\Pr(X_j = 1) = \Pr(X_j = -1) = 1/2,$$

*then $\sum_{j=1}^{\infty} a_j X_j$ converges with probability* 1.

Informally, if we have a real sequence $a_j$ with $\sum_{j=1}^{\infty} a_j^2$ convergent, then, if we assign signs at random, $\sum_{j=1}^{\infty} \pm a_j$ will converge with probability 1.

Let us call a sequence $a_j$ 'good' if only finitely many of the $a_j$ are non-zero. It is obvious that every good sequence satisfies the conclusions of Theorem 1. Our object is to use this fact to prove Theorem 1 in general. We use the following 'reflection' lemma.

**Lemma 2** *Let $a_j$ and $X_j$ satisfy the conditions of Theorem 1. We write $\alpha = \sup_{j \geq 1} |a_j|$ and take $\lambda > \alpha$.*

(i) *If $N \geq 1$,*

$$\Pr\left(\max_{1 \leq n \leq N} \sum_{j=1}^{n} a_j X_j \geq \lambda\right) \leq 2 \Pr\left(\sum_{j=1}^{N} a_j X_j \geq \lambda - \alpha\right).$$

(ii) *If $N \geq 1$,*

$$\Pr\left(\max_{1 \leq n \leq N} \left|\sum_{j=1}^{n} a_j X_j\right| \geq \lambda\right) \leq 4 \Pr\left(\sum_{j=1}^{N} a_j X_j \geq \lambda - \alpha\right).$$

(iii) *If* $N \geq 1$,

$$\Pr\left(\max_{1\leq n\leq N}\left|\sum_{j=1}^{n} a_j X_j\right| \geq \lambda\right) \leq 2\frac{\sum_{j=1}^{N} a_j^2}{(\lambda-\alpha)^2}.$$

*Proof* Let $x_j = \pm 1$. If $\max_{1\leq n\leq N}\sum_{j=1}^{n} a_j x_j \geq \lambda$, then there exists an $1 \leq M \leq N-1$ such that

$$\sum_{j=1}^{M+1} a_j x_j \geq \lambda \quad \text{but} \quad \sum_{j=1}^{m} a_j x_j < \lambda \quad \text{for } 1 \leq m \leq M.$$

If we write $\mu = \sum_{j=1}^{M} a_j x_j$, then, automatically, $\mu \geq \lambda - \alpha$.

We now observe (and this is where the 'reflection' occurs) that

$$\sum_{j=M+1}^{N} a_j x_j \geq 0 \quad \Leftrightarrow \quad \sum_{j=M+1}^{N} a_j(-x_j) \leq 0,$$

so that

$$\sum_{j=1}^{M} a_j x_j + \sum_{j=M+1}^{N} a_j x_j \geq \mu \quad \Leftrightarrow \quad \sum_{j=1}^{M} a_j x_j + \sum_{j=M+1}^{N} a_j(-x_j) \leq \mu.$$

Thus

$$\max\left\{\sum_{j=1}^{M} a_j x_j + \sum_{j=M+1}^{N} a_j x_j, \sum_{j=1}^{M} a_j x_j + \sum_{j=M+1}^{N} a_j(-x_j)\right\} \geq \mu \geq \lambda - \alpha.$$

We have shown that at least half of the possible choices of $x_j = \pm 1$ which yield $\max_{1\leq n\leq N}\sum_{j=1}^{n} a_j x_j \geq \lambda$ also yield $\sum_{j=1}^{N} a_j x_j \geq \lambda - \alpha$. Since the $X_j$ are independent random variables with $\Pr(X_j = 1) = \Pr(X_j = -1) = 1/2$, the required result follows.

(ii) By symmetry

$$\Pr\left(\max_{1\leq n\leq N}\sum_{j=1}^{n} a_j X_j \geq \lambda\right) = \Pr\left(\min_{1\leq n\leq N}\sum_{j=1}^{n} a_j X_j \leq -\lambda\right).$$

(iii) By symmetry and Tchebychev's inequality,

$$2\,\mathrm{Pr}\left(\sum_{j=1}^{N} a_j X_j \geq \lambda - \alpha\right) = \left(\left|\sum_{j=1}^{N} a_j X_j\right| \geq \lambda - \alpha\right)$$
$$\leq \frac{\operatorname{var}\sum_{j=1}^{N} a_j X_j}{(\lambda-\alpha)^2}$$
$$= \frac{\sum_{j=1}^{N} a_j^2}{(\lambda-\alpha)^2},$$

and the result follows. □

We now introduce the 'maximal function'

$$S^*(\mathbf{X}) = \sup_{1\leq n}\left|\sum_{j=1}^{n} a_j X_j\right|.$$

(Note that $S^*$ can take the value $\infty$, but, as the next lemma shows, the probability that this occurs is zero.) Although the formal proof involves the generalisation of Lebesgue's theory called 'measure theory' the reader should have no difficulty in accepting that the result follows from Lemma 2.

**Lemma 3** *Let $a_j$ and $X_j$ satisfy the conditions of Theorem 1. If $\lambda > \alpha$, where $\alpha = \sup_{j\geq 1}|a_j|$, then*

$$\mathrm{Pr}(S^*(\mathbf{X}) \geq \lambda) \leq 2\frac{\sum_{j=1}^{\infty} a_j^2}{(\lambda-\alpha)^2}.$$

This maximal lemma now gives us a proof of Theorem 1 as follows.

**Lemma 4** *Let $a_j$ and $X_j$ satisfy the conditions of Theorem 1.*

(i) *There exists an $N(k)$ such that*

$$\mathrm{Pr}\left(\left|\sum_{j=n}^{m} a_j X_j\right| \leq 2^{-k} \quad \textit{for all } m \geq n \geq N(k)\right) \geq 1 - 2^{k}.$$

(ii) *If $r \geq 1$, then $\sum_{j=1}^{\infty} a_j X_j$ converges with probability at least $1 - 2^{-r}$.*

(iii) *$\sum_{j=1}^{\infty} a_j X_j$ converges with probability* 1.

*Proof* (i) Chose $N(k)$ so that $\sum_{j=N(k)}^{\infty} a_j^2 \leq 2^{-3k-6}$ and so, in particular $\sup_{j\geq N(k)}|a_j| \leq 2^{-k-2}$. Consider the sequence $b_j$ defined by $b_j = 0$ for $1 \leq j \leq N(k)$ and $b_j = a_j$ for $j \geq N(k)$. (Note that, if we write $c_j = a_j$ for $1 \leq j \leq N(k)$ and $c_j = 0$ for $j \geq N(k)$, then $(c_j)$ is a 'good' sequence, $(b_j)$ is a sequence which is 'close to the zero sequence' and $a_j = b_j + c_j$.)

We write

$$T^*(\mathbf{X}) = \sup_{1\le n}\left|\sum_{j=1}^{n} b_j X_j\right|$$

and observe that, if $m \ge n \ge N(k)$, we have

$$\begin{aligned}\left|\sum_{j=n}^{m} a_j X_j\right| &\le \left|\sum_{j=N(k)}^{m} a_j X_j\right| + \left|\sum_{j=N(k)}^{n} a_j X_j\right|\\ &= \left|\sum_{j=N(k)}^{m} b_j X_j\right| + \left|\sum_{j=N(k)}^{n} b_j X_j\right|\\ &\le 2T^*(\mathbf{X}).\end{aligned}$$

By the maximal inequality of Lemma 3,

$$\Pr(T^*(\mathbf{X}) \ge 2^{-k-1}) \le 2\frac{\sum_{j=1}^{\infty} b_j^2}{(2^{-(k+1)} - 2^{-k-2})^2} \le 2^{-k-1},$$

so, combining the two results obtained in this paragraph,

$$\Pr\left(\left|\sum_{j=n}^{m} a_j X_j\right| \le 2^{-k} \quad \text{for all } m \ge n \ge N(k)\right) \ge 1 - 2^{-k}.$$

(ii) By (i), we can find $N(k)$ such that

$$\Pr\left(\left|\sum_{j=n}^{m} a_j X_j\right| \le 2^{-k} \quad \text{for all } m \ge n \ge N(k)\right) \ge 1 - 2^{-k}$$

for each $k \ge r + 1$. If the probability of an event $A_k$ happening is at most $2^{-k}$, then the reader will readily accept (as can be proved formally using measure theory) that the chance of at least one of $A_{r+1}, A_{r+2}, \dots$ occurring is at most $\sum_{k=r+1}^{\infty} 2^{-k} = 2^{-r}$. Thus, with probability at least $1 - 2^{-r}$, we have

$$\left|\sum_{j=n}^{m} a_j X_j\right| \le 2^{-k} \quad \text{for all } m \ge n \ge N(k) \text{ and all } k \ge r + 1.$$

Using the general principle of convergence, it follows, with probability at least $1 - 2^{-r}$, that $\sum_{j=1}^{\infty} a_j X_j$ converges.

(iii) The result follows from the fact that (ii) holds for every $r \ge 1$. □

To see that the convergence is very different from that which we see in elementary analysis, observe that there is a strictly positive probability that all the random variables in a sequence $X_n, X_{n+1}, \dots, X_{n+q}$ take preassigned values.

**Exercise 5** Suppose that $a_j$ and $X_j$ satisfy the conditions of Theorem 1 and, in addition, that $\sum_{j=1}^{\infty}|a_j|$ diverges. Show that, given any $n$, we can find an $m \geq n$ such that

$$\Pr\left(\left|\sum_{j=n}^{m} a_j X_j\right| \geq 1\right) > 0.$$

The repeated successes of the maximal inequality technique prompted Banach to prove a theorem which showed, in effect, that any convergence theorem of the type we are discussing implies a maximal inequality. We need to introduce some notation. If $E$ is a subset of $[-1/2, 1/2]$, we say that $E$ has measure at most $K$ if, given any $\epsilon > 0$, we can cover $E$ with a countable collection of intervals of total length less than $K + \epsilon$. (Strictly speaking, we should talk about 'outer measure' but, for all the sets we talk about, the distinction does not matter. Parallelling our discussion of sets of measure zero, one way of thinking about a set of measure at most $K$ is that, if you drop a dart at random on the interval $[-1/2, 1/2]$, the probability of hitting $E$ is at most $K$.)

If we write

$$S^*(f,t) = \sup_{n \geq 0}\left|\sum_{j=-n}^{n} \hat{f}(j)\exp(2\pi i j)\right|,$$

then Banach's theorem tells us that Lusin's conjecture is true if and only if there exists a positive function $B\colon \mathbb{R} \to \mathbb{R}$ with $B(\lambda) \to 0$ as $\lambda \to \infty$ such that the measure of the set

$$\{t \in [-1/2, 1/2]\colon S^*(f,t) \geq \lambda \|f\|_2\}$$

is less than $B(\lambda)$. The proof that this inequality implies Lusin's conjecture runs along the lines set out in the proof of Lemma 3 and the proof of the converse is a 'grandchild' of the method by which we proved du Bois-Reymond's result. Although Banach's result is not very deep in itself, it confirms that, if Lusin's conjecture were true, it would be a reasonable strategy to try to prove it via a maximal inequality of the type we have just proved. It turns out that the correct form of $B$ for our particular problem is $B(\lambda) = K\lambda^2$ for some constant $K$. Rewriting the previous formula, we now know that Lusin's conjecture is true if and only if there exists a $K$ such that the set

$$\{t \in [-1/2, 1/2]\colon S^*(f,t) \geq \lambda\}$$

has measure less than $K\lambda^{-2}\|f\|_2^2$ for all $\lambda > 0$ and all $f \in L^2$.

Simple limiting arguments show that Lusin's conjecture can be restated as follows.

**Conjecture 6** *There exists a constant $K$ with the following property. Given $N$ a positive integer, $a_j \in \mathbb{C}$ $[|j| \leq N]$ and $\lambda > 0$, we can find intervals $I_1, I_2, \ldots, I_m$ of* *total length less than* $K\lambda^{-2}\sum_{j=-N}^{N}|a_j|^2$ *such that*

$$\left|\sum_{j=-r}^{r} a_j \exp 2\pi i j t\right| \leq \lambda$$

*for all* $0 \leq r \leq N$ *and all* $-1/2 \leq t \leq 1/2$ *with* $t \notin \bigcup_{k=1}^{m} I_k$.

Theorem 1 can be easily restated in a non probabilistic way. Recall that, if $x$ is real, we write

$$\operatorname{sgn} x = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x = 0. \end{cases}$$

We set $H_r(t) = \operatorname{sgn}(\sin 2\pi r t)$. The reader should sketch $H_r$ for $r = 1$, 2, 4 and for $r = 5$.

**Theorem 7** *Suppose* $a_j \in \mathbb{R}$ *and* $\sum_{j=1}^{\infty} a_j^2$ *converges. If we define* $H_r$ *as above then* $\sum_{j=1}^{\infty} a_j H_{2^j}(t)$ *converges for all* $t$ *outside a set of measure zero.*

To see why Theorem 7 is just Theorem 1 in disguise, think of the points $t$ where $H_{2^j}(t) = 1$ as corresponding to $X_j = 1$ (the $j$th throw of a fair coin comes down heads) and the points $t$ where $H_{2^j}(t) = 0$ as corresponding to $X_j = 1$ (the $j$th throw of a fair coin comes down tails). The set of points where $H_{2^j}(t) = 0$ has measure zero (equivalently have probability zero and correspond to the coin coming down on its edge) and may be ignored.

Theorem 7 suggests very strongly that, if $a_j \in \mathbb{R}$ and $\sum_{j=1}^{\infty} a_j^2$ converges, then $\sum_{j=1}^{\infty} a_j \sin(2\pi 2^j t)$ will converge for all $t$ outside a set of measure zero. Readers who try to prove this may already find the task harder than they expect, but it will be helpful to add an extra layer of complexity by considering the following almost equally plausible result: If $a_j \in \mathbb{R}$ and $\sum_{j=1}^{\infty} a_j^2$ converges, then $\sum_{j=1}^{\infty} a_j \sin(2\pi(2^j + 1)t)$ will converge for all $t$ outside a set of measure zero.

Why is the new problem harder than the old? When we considered $\sum_{j=1}^{\infty} a_j X_j$, knowing the values of $X_1, X_2, \dots, X_n$ told us nothing about the value of $X_{n+1}$ (or indeed the sequence $X_{n+1}, X_{n+2}, \dots$). In the same way when we considered $\sum_{j=1}^{\infty} a_j H_{2^j}(t)$, then, with some trivial exceptions, knowledge of $H_{2^n}(t)$ told us nothing about the value of $H_{2^{n+1}}(t)$. On the other hand, if we know $\sin(2\pi(2^n + 1)t)$ exactly, then we know that $t$ must take one of a finite number of values and so the sequence $\sin(2\pi(2^n + 1)t)$ must be one of a finite number of possible sequences.

Matters are not quite so bad as they seem. If we only know the value of $\sin(2\pi(2^n + 1)t)$ to within an accuracy $\delta$, then, speaking very roughly, we only know $t$ to within an accuracy of $2^{-n}\delta$ and so when $2^{m-n}\delta$ is substantially larger than 1, we know practically nothing about $\sin(2\pi(2^m + 1)t)$. This suggests that if $\beta > 1$, $n(r + 1)/n(r) > \beta$ the sequence of functions $s_{n(r)}$ given by $s_{n(r)}(t) = \sin(2\pi n(r)t)$

behaves like a sequence of 'almost independent random variables' with terms far apart in the sequence being 'essentially independent'. Although this heuristic argument does not point to the best proof, the result it suggests is true. If $\beta > 1$ and $n(r+1)/n(r) > \beta$ for all $j \geq 1$, then if $a_j \in \mathbb{C}$ and $\sum_{j=-\infty}^{\infty} |a_j|^2$ converges then, as $r \to \infty$, $\sum_{j=-r}^{r} a_j \sin(2\pi n(j)t)$ converges outside a set of measure zero. Similarly, the same hypotheses imply that $\sum_{j=-r}^{r} a_j \exp(in(j)t)$.

If we were very adventurous and prepared to wave our hand very wildly indeed, we could go further and, instead of talking about the $\exp(in(j)t)$ as being in some vague sense 'almost independent', we might hope that 'block sums'

$$B_r(t) = \sum_{n(r+1) \geq |j| \geq n(r)+1} a_j \exp(2\pi ijt)$$

might be almost independent in the sense that, if we work to the appropriate level of accuracy, knowledge of $B_r$ tells us very little about $B_s$ when $r$ is not close to $s$. Whether the arm waving is justified or not, Kolmogorov proved that, if $\beta > 1$ and $n(r+1)/n(r) > \beta$ for all $j \geq 1$, then if $a_j \in \mathbb{C}$ and $\sum_{j=-\infty}^{\infty} |a_j|^2$ converges then, as $r \to \infty$, $\sum_{j=-n(r)}^{n(r)} a_j \exp(2\pi ijt)$ converges outside a set of measure zero. Because of the close link between square summable sequences and $L^2$, this tells us at once that

$$\sum_{j=-n(r)}^{n(r)} \hat{f}(j) \exp(2\pi ijt) \to f(t)$$

outside a set of measure zero for all $f \in L^2$.

The example of

$$g_{2^m+r}(t) = \begin{cases} n & \text{for } 2^{-1} + r2^{-m} \leq t \leq 2^{-1} + (r+1)2^{-m} \text{ and } 0 \leq r \leq 2^m - 1, \\ 0 & \text{otherwise,} \end{cases}$$

which we discussed earlier, shows that the result we have just described can not be used to prove Lusin's conjecture.

Since we have followed what now seems the natural path from sums of random variables to Fourier sums, it is interesting to note that we have reversed the historic path. As Kolmogorov recalled 'Such topics as conditions for the validity of the law of large numbers and conditions for convergence of series of independent random variables were actually tackled by methods developed by Lusin and his pupils in the general theory of trigonometric series'.

At the end of the 1950's, two giants of the subject published treatises summarising the state of Fourier Analysis. Bari's *A Treatise on Trigonometric Series* [1] ran to 900 pages and Zygmund's second edition of *Trigonometric Series* [3] to 700. Both authors devoted a substantial part of their works to aspects of the problem of convergence of Fourier sums. Zygmund's introduction specifies 'the problem of the existence of a continuous function with everywhere divergent series' as one of the two main open problems in the Fourier Analysis but does not mention Lusin's conjecture. Bari gives a discussion of the reasons (needless to say substantially deeper than those I have discussed) which, she believed, led Lusin to his conjecture but states that '[these] arguments ... cannot have any force today'.

It came as an enormous surprise when, in 1966, Carleson announced his proof of Lusin's conjecture. Although Carleson was already known for a spectacular proof of another classical conjecture (the Corona theorem which we discuss later) many mathematicians were dubious. In my fourth year of university studies, I inherited a set of notes on Hilbert space containing the marginal notation 'A Swede called Carleson claims to have proved pointwise convergence but nobody believes him'. (The note must have been written in 1967.)

Even when it became clear that the proof was correct, the surprise remained. Carleson had essentially proved Lusin's conjecture by a frontal attack on the version given in Conjecture 6. But, although no one would have thought of stating Conjecture 6 until the 1930's, its statement would have been understood by Fourier and the proof of Lusin's conjecture (at least in so far as it concerns continuous functions) from Conjecture 6 would have been understood by Dirichlet.

In our discussion we saw that it was reasonable to divide the sum into blocks of the type

$$\sum_{j=2^n}^{2^{n+1}-1} a_j \exp(2\pi i j t) = \exp(2\pi i 2^n t) \sum_{j=0}^{2^n} a_{j-2^n} \exp(2\pi i j t),$$

which we may think of as 'frequency blocks' at scale $2^n$. Carleson's proof divides the interval $[-1/2, 1/2]$ into intervals $[r2^{-n}, (r+1)2^{-n}]$ which we may think of as space blocks at scale $2^{-n}$. It will, I hope, appear natural that we try to study the behaviour of frequency blocks at scale $2^n$ on space blocks at scale $2^{-n}$. If the behaviour on a space block at scale $2^{-n}$ is good, then we retain it. If it is bad, then it will be one of the intervals $I_k$ rejected in Conjecture 6. If we cannot decide, then we split the interval in two and examine the results at the appropriate new scale.

In his address [4] to the Nice International Congress, Hunt shows how the proof works for a model system which bears much the same relation to the trigonometric problem as the sum $\sum_{j=1}^{\infty} a_j H_{2^j}(t)$ in Theorem 7 does to the sum $\sum_{j=1}^{\infty} a_j \sin(2\pi(2^j+1)t)$. The reader who wishes to go further is recommended to consult this expository tour de force. However, even knowing how the proof works in the model case, it is entirely unclear how or if it can be made to work in a system where the 'exact' relations of the model must be replace by the 'approximate' relations of the trigonometric case. Carleson's proof overcomes these difficulties in a magnificent example of what mathematicians call 'power' and which they value far above 'cleverness'.

The difficulty of his paper is not simply due to the depth of the ideas but also to an expository problem which dogs all papers of this type. As I tried to indicate earlier the 'strategic' vision of the proof is useless without a large number of 'tactical' decisions to overcome various problems as they arise. However, there are often several ways of tackling the difficulty. As a simple example, it does not matter whether the

$C$ in Conjecture 6 is 10 or $10^{10}$, but readers may well be baffled if we use this freedom to make estimates which are weaker than they know are available. (At a trivial level, suppose that a real variable proof requires only that $(a+b)^2 \le K(a^2+b^2)$ for some constant $K$. If we write

$$(a+b)^2 \le \bigl(2\max\{|a|,|b|\}\bigr)^2 = 4\max\{a^2,b^2\} \le 4(a^2+b^2),$$

the reader who spots a better estimate may well be worried.) A paper which requires a large number of decisions, where more than one 'correct choice' is possible but no choice is 'natural', will always be difficult to read. Hunt's lecture to the Nice Conference is made much easier because the model system chosen has a fair number of 'natural choices'.

Since every continuous function is an $L^2$ function, Carleson's theorem settles the convergence problem as it would have appeared to Fourier and his successors but, in view of the difficulty of Carleson's $L^2$ proof, it may be asked whether it might not be possible to extract an easier proof which merely applies to continuous functions. Two years after Carleson's result, Kahane and Katznelson produced a construction (which may not have been an example of 'power' but was certainly one of extreme 'cleverness') which showed that given any set $E$ of measure zero there exists a continuous function $f$ whose Fourier sum diverged at every point of $E$ (and possibly others). Thus, at least from this point of view, the Fourier sums of continuous functions exhibit the same wildness as general $L^2$ functions. Although the theorems of mathematicians can, usually, be relied on, as we have seen, their opinions are just opinions. However, it is hard to see how a simpler 'continuous function proof' would work.

In theory, mathematical proofs, even those as complicated as Carleson's can be checked line by line. However, it is much more satisfactory if the underlying ideas can be tested by applying them in new situations. Hunt showed that Carleson's results applied not merely to those $L^1$ functions for which $\int_{-1/2}^{1/2} |f(t)|^2\,dt < \infty$ but also to the much larger class of $L^1$ functions for which $\int_{-1/2}^{1/2} |f(t)|^p\,dt < \infty$ for some $p > 1$ and Sjölin carried this idea even further.

Many of the ideas and methods introduced by Carleson (for example in the solution of the Corona problem) were quickly absorbed into main stream of mathematics. (Some received that rather backhanded tribute that the goddess of mathematics pays to favoured worshipers by disappearing into the unattributed common stock.) In the case of the convergence theorem it looked for some time as though no successor could be found to bend the bow of Ulysses. However a new generation of brilliant harmonic analysts have taken up Carleson's ideas and developed them further.

It is much harder to give a clear idea of the other major problems tackled by Carleson using only the ideas available to a first or second year mathematics undergraduate and from now on my discussion will be both briefer and much vaguer.

Earlier I said that, for a short time, it was possible to hope that sets of measure zero could always be ignored. For example, it was known that for many sets $E$ of measure zero the statement

$$\sum_{j=-\infty}^{\infty} a_j \exp(2\pi i j t) = 0 \text{ for } t \notin E$$

implied that $a_j = 0$ for all $j$ and it was expected that this result would turn out to be true for all sets $E$ of measure zero. In 1917, Mensov showed that this was not the case and mathematicians began to study the sets of measure zero in earnest.

Physicists have always been happy to consider 'point masses' and 'charges living on the surface of a sphere'. Lebesgue's successors created a theory of 'general measures' which included such objects as special cases. It turned out these 'general measures' were ideal tools for studying continuous functions (so that, in some sense, if we knew everything about general measures, we would know everything about continuous functions and vice-versa). Just as a point mass lives on one point, so many important general measures live on sets of measure zero (in the Lebesgue sense). Much of Carleson's early work is concerned with the study of sets of Lebesgue measure zero and the general measures that live on them. His results left a permanent mark on the subject, but he soon moved into other fields.

Much of mathematics follows the kind of pattern we have sketched out. A concrete problem turns out to require new tools for its solution. Reflection shows that these new tools are themselves concrete 'graspable' objects and their study raises new concrete problems. However, it is often profitable to seek theories which join many different concrete problems and their solutions. These abstract theories often provide more powerful tools or suggest new problems.

Consider the following problems.

*Problem A.* Given $a_j \in \mathbb{C}$ such that $\sum_{n=0}^{\infty} a_n \exp(2\pi i n t)$ converges for all real $t$, when can we find $b_j \in \mathbb{C}$ such that $\sum_{n=0}^{\infty} b_j \exp(2\pi i n t)$ converges for all real $t$ and

$$\sum_{n=0}^{\infty} a_j \exp(2\pi i n t) \sum_{n=0}^{\infty} b_n \exp(2\pi i n t) = 1$$

for all real $t$.

*Problem B.* Given $a_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |a_n|$ converges, when can we find $b_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |b_n|$ converges,

$$\sum_{j=-\infty}^{\infty} a_{r-j} b_j = 0 \quad \text{for } r \neq 0 \text{ and } \sum_{j=-\infty}^{\infty} a_{-j} b_j = 1.$$

In 1820 mathematicians would have understood both problems (though they might have considered them a bit odd) but would not have been able to resolve them. By 1850, the discovery of complex analysis had produced a simple answer to Problem A. Since $\sum_{n=0}^{\infty} a_n z^n$ converges for all $|z| = 1$, the power series $\sum_{n=0}^{\infty} a_n z^n$ converges for all $|z| \leq 1$ to a function $f$ which is analytic in the open disc $\{z\colon\ |z| < 1\}$ and continuous on $\{z\colon\ |z| < 1\}$. Conversely, if a function $g$ is analytic in the open disc $\{z\colon\ |z| < 1\}$ and continuous on the closed disc $\{z\colon\ |z| < 1\}$, then we can write

$g(z) = \sum_{n=0}^{\infty} b_n z^n$ for $|z| \leq 1$ and $\sum_{n=0}^{\infty} b_n \exp(2\pi nt)$ will converge for all $t$. With a little extra work it is now possible to restate Problem A as follows.
*Problem A′*. Given a function $f$ analytic in the open disc $\{z\colon\ |z| < 1\}$ and continuous on the closed disc $\{z\colon\ |z| \leq 1\}$, when can we find a function $g$ analytic in the open disc $\{z\colon\ |z| < 1\}$ and continuous on $\{z\colon\ |z| \leq 1\}$ such that $f(z)g(z) = 1$ for all $|z| \leq 1$.

Elementary theorems of complex analysis tell us that we can find the required $g$ if and only $f(z) \neq 0$ for all $|z| \leq 1$.

Problem B was resolved by Wiener in the 1930's. Primed by our discussion of Problem A, the reader will have no difficulty as recasting it in the following form.
*Problem B′*. Given $a_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |a_n|$ converges, when can we find $b_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |b_n|$ converges and

$$\sum_{n=0}^{\infty} a_n \exp(2\pi int) \sum_{m=0}^{\infty} b_m \exp(2\pi imt) = 1$$

for all real $t$.

Wiener showed that we can find the required $b_j$ if and only if

$$\sum_{n=-\infty}^{\infty} a_n \exp(2\pi int) \neq 0$$

for all real $t$. We may recast his result as follows.

**Theorem 8** *Given $a_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |a_n|$ converges, we can find $b_n \in \mathbb{C}$ such that $\sum_{n=-\infty}^{\infty} |b_n|$ and*

$$\sum_{n=-\infty}^{\infty} a_n z^n \sum_{m=-\infty}^{\infty} b_m z^m = 1$$

*if and only if $\sum_{n=-\infty}^{\infty} a_n z^n \neq 0$ for all $|z| = 1$.*

Here is a closely related result which reflects our answer to Problem A.

**Theorem 9** *Given $a_n \in \mathbb{C}$ such that $\sum_{n=0}^{\infty} |a_n|$ converges, we can find $b_n \in \mathbb{C}$ such that $\sum_{n=0}^{\infty} |b_n|$ converges and*

$$\sum_{n=-\infty}^{\infty} a_n z^n \sum_{m=-\infty}^{\infty} b_m z^m = 1$$

*if and only if $\sum_{n=-\infty}^{\infty} a_n z^n \neq 0$ for all $|z| \leq 1$.*

In the 1940's Gelfand and others showed that a wide range of similar problems had a similar solution. Given an appropriate collection $A$ of complex valued functions on a space $X$ we can find a space $Y \supseteq X$ and an extension of each $f \in A$

to a function $\tilde{f}$ on $Y$ such that the following result is true. Given an $f \in A$ we can find a $g \in A$ with $f(x)g(x) = 1$ for all $x \in X$ if and only if $\tilde{f}(y) \neq 0$ for all $y \in Y$. Gelfand's theory is a marvellous display of the power of abstract methods but, though the theory tells us that the appropriate $Y$ always exists, it does not always tell us how to find it.

If we look at Theorems 8 and 9, we see that they both concern similar collections of functions which live on the boundary $X = \{z\colon\ |z| = 1\}$ of the unit disc but in one case $Y = X$ whilst in the other case we have $Y = \{z\colon\ |z| \leq 1\}$, the entire closed unit disc. In more general problems the extending space $Y$ may not be readily describable in terms of $X$.

Readers are strongly advised to work through the next exercise.

**Exercise 10** Consider the collection $A$ of uniformly continuous functions $f\colon [0, 1] \cap \mathbb{Q} \to \mathbb{C}$ (that is to say, the set of uniformly continuous functions on the space $X$ of rational numbers in $[0, 1]$).

(i) Given an $f \in A$, we can find a $g \in A$ with $f(x)g(x) = 1$ for all $x \in X$ if and only if the continuous extension $\tilde{f}$ of $f$ to $Y = [0, 1]$ is nowhere zero.

(ii) Given an $f \in A$, we can find a $g \in A$ with $f(x)g(x) = 1$ for all $x \in X$ if and only if there exists a $\delta > 0$ such that $|f(x)| > \delta$ for all $x \in X$.

Here we have an example in which, although $Y \neq X$, we have $X$ dense in $Y$ and (as we have discussed before) knowledge of $X$ gives us a very strong hold on the behaviour of $Y$.

The Corona theorem concerns the space $H^\infty$ of functions $f(z) = \sum_{n=0}^{\infty} a_n z^n$ where the series converges for all $|z| < 1$ to a bounded function. In our notation, $X = \{z\colon\ |z| < 1\}$ and $A = H^\infty$. Anyone who has a slight acquaintance of the unpleasant behaviour of power series on their circle of convergence will not be surprised to learn that $A$ contains functions which cannot be extended continuously to $D = \{z\colon\ |z| \leq 1\}$ and that there is no way in which we can identify the Gelfand space $Y$ with $X$, $D$ or any other subset of $\mathbb{C}$. However, if $X$ were dense in $Y$ (in the appropriate sense), we would have very strong hold on the behaviour of $Y$.

The reader must be warned that when I speak of '$X$ being dense in $Y$ in the appropriate sense', I am probably (I hope, for the first time in this essay) overstretching analogy, but this rather abstract formulation can be replaced by the following concrete result (compare parts (i) and (ii) of Exercise 10).

**Theorem 11** *If $f_1, f_2, \ldots, f_n \in H^\infty$ and there exists a $\delta > 0$ such that $\sum_{j=1}^{n} |f_j(z)| \geq \delta$ for all $|z| < 1$ then we can find $f_1, f_2, \ldots, f_n \in H^\infty$ such that*

$$\sum_{j=1}^{n} f_j(z)g_j(z) = 1$$

*for all $|z| < 1$.*

This is Carleson's Corona Theorem. Important though the theorem is, the 'Carleson measures' that he introduced in order to prove it have turned out to be even more important and now appear in many branches of analysis.

It is characteristic of major mathematicians that they move to new fields and attack new problems throughout their career. At an age when most mathematicians have retired or, at least settled into routine teaching and research, Carleson helped establish the existence of 'strange attractors in the Hénon family of planar maps'. Incomprehensible as the phrase sounds it may be placed on the map of humanity's intellectual interests.

Mathematicians, like humanity in general, have always been interested in the long term behaviour of systems. For a mathematician this often means the study of differential equations like $\dot{x}(t) = f\big(x(t), t\big)$. Unfortunately such equations rarely have explicit solutions, so we must settle for numerical computation (which, at best, tell us about one particular case) or seek to establish general properties of the solutions of a given type of differential equation.

The natural way to seek such general principles is to look at those differential equations that we can solve and at actual physical systems described by differential equations of the appropriate type. The main class of differential equations that we can solve exactly is the class of linear differential equations with constant coefficients, that is to say, differential equations like

$$\ddot{x} + a\dot{x} + bx = 0.$$

The solutions of such equations either explode, settle down or oscillate in a periodic fashion.

If we think of a physical model, we see that, in explosions, points which start off close together separate very rapidly. Small changes in initial conditions very rapidly produce very different solutions and long term behaviour is essentially unpredictable. If you try to balance a billiard cue on its tip then, although there must be some position of balance, the slightest deviation rapidly leads to disaster. We speak of 'unpredictability', but it is better to think in terms of a soothsayer who will tell us our fortune one second into the future for 10 euros, two seconds into the future for 100 euros, three seconds into the future for 1000 euros and so on. In theory, the soothsayer will look one minute into the future, if we so desire, but in practice we cannot afford the fee.

If the system frictional, then everything settles down and small changes in initial conditions make very little difference. If we drop two steel balls fairly close together at much the same time into a deep vat of treacle (so we model an equation like $\ddot{x} + a\dot{x} + bx = c$), the two balls will remain close together as they fall through the treacle.

The development of electrical engineering greatly extended the variety of physical models which could be drawn on by mathematicians. In the 1920's and 30's this was exploited by Van der Pol who investigated the equation

$$\ddot{x} - k(1 - x^2)\dot{x} + x = E\sin(\omega t)$$

experimentally. (The term $k(1-x^2)$ can be thought of as the resistance of the circuit, but it is allowed to be *negative*.) Van der Pol observed many strange phenomena. The system could settle down to periodic behaviour but the period could be a proper integer multiple of the period of the forcing term $E\sin(\omega t)$. If the parameters of the circuit changed this, the period of the system would jump discontinuously to another period (and the new period might depend not only on the present values of the parameters but on the way in which the present values had been arrived at). The 'mathematical reality' of these observations was proved by Cartwright and Littlewood in a paper famous for its depth, length and difficulty.

Any proper discussion of these topics would need to distinguish carefully between what was known (or guessed) by particular individuals (such as Poincaré and Littlewood) or even particular mathematical schools (such as the Russian groups studying non-linearity) and what was generally known by the mathematical community. However, from the point of view of the scientific community as a whole, a turning point came with the work of Lorenz published in 1963.

Lorenz set up a system of differential equations to provide a simple model of atmospheric convection and then solved them numerically using an early desk top computer. He observed that the solutions seemed to settle down to some sort of periodic repetition only to suddenly jump to new almost periodic pattern. The new pattern would persist for some time and then the system would jump again and so on. The jumps appeared random or, more properly, showed extreme sensitivity to initial conditions.

Mathematicians were of course aware of the phenomenon of 'extreme sensitivity to initial conditions' in cases like explosions or balancing billiard cues. However Lorenz's system exhibited apparent stability (the fairly periodic patterns) combined with real instability (the essentially unpredictable jumps) in a totally unexpected way.

Differential equations have discrete analogues like $x_{n+2} = F(x_{n+1}, x_n)$. (Thus we start with $x_0$ and $x_1$, compute $x_2 = F(x_0, x_1)$ as $x_3 = F(x_1, x_2)$ and so on.) Unfortunately it proved as difficult to find explicit solutions of such equations as it is to find explicit solutions of the corresponding differential equations and much more difficult to find physical analogues. For these reasons, difference equations were very much regarded as the poor relations of differential equations.

All this changed when electronic computers were introduced. When we calculate (an approximation to) the solution of a differential equation on a computer, we do so by replacing the differential equation by a difference equation and solving the difference equation step by step. Even more importantly, it became possible for anyone with access to a hand calculator to follow the behaviour of a particular solution through many hundred steps and anyone with a computer to follow it through millions of steps. This *experimental mathematics* gave the same sort of insight that a physical experiment might give to a classical mathematician. It is a pity that the word 'chaos' is used in this connection because if the picture revealed by these numerical experiments had been truly chaotic, it would have been of little interest to mathematicians. Instead the experimenters saw a world full of unexplained patterns, and it became the job of mathematicians to find out if these patterns were truly there (they might be artifacts caused by subtle problems with the experiments) and, if so, to explain those patterns.

Mathematical progress seldom follows a unique path. Difference equations may be considered a special case of iteration in which we take a map $T$ from a set $X$ to itself and consider the *orbit* $x, Tx, T^2x, \ldots$ traced out by images of $x$ under repeated application of $T$. (In other words, we take $x_0 = x$ and consider the sequence $x_0$, $x_1$, $x_2, \ldots$, with $x_{n+1} = Tx_n$.) Birkhoff's pointwise ergodic theorem which I mentioned earlier, is an example of a very deep theorem on iteration.

At the beginning of the 20th century, the French mathematicians Julia and Fatou investigated iteration when $X = \mathbb{C}$ and $T$ took various simple forms. The following result was already known and mathematical readers may enjoy treating it as an exercise.

**Exercise 12** Let $\mathbb{C}^* = \mathbb{C} \cup \{\infty\}$. Consider the collection $\mathcal{M}$ of Möbius maps $T : \mathbb{C}^* \to \mathbb{C}^*$ given by

$$Tz = \frac{az+b}{cz+d}$$

with $ad - bc \neq 0$.

(i) Show that if a Möbius map is not the identity, then it has one or two fixed points.

(ii) Suppose $T \in \mathcal{M}$ has exactly two fixed points $z_1$ and $z_2$. If $S \in \mathcal{M}$ and $Sz_1 = 0$, $Sz_2 = \infty$, show that $STS^{-1}$ is a Möbius map which fixes 0 and $\infty$. Conclude that $STS^{-1}(z) = Az$ for some $A \in \mathbb{C}$ with $A \neq 0$.

(iii) If $|A| > 1$ show that, unless $w = z_1$, we have $T^n w \to z_2$. What happens if $|A| < 1$? What happens if $A = \exp(2\pi\alpha)$ and $\alpha$ is rational? What happens if $A = \exp(2\pi\alpha)$ and $\alpha$ is irrational?

(iv) Suppose $T \in \mathcal{M}$ has exactly one fixed point $z_1$. Show that there is a Möbius map $S$ such that $STS^{-1}$ fixes $\infty$. Show that $STS^{-1}(z) = z + B$ with $B \neq 0$. Conclude that, if $w \in \mathbb{C}^*$, we have $T^n w \to z_1$.

We might expect that similar, but slightly more complicated, results will hold for similar, but slightly more complicated, families of functions. In fact, as Julia, Fatou and their successors showed, the moment we consider slightly more complicated functions the behaviour of iterates becomes very much more complicated. A deservedly popular example (though coming from yet another intellectual source) is the logistic map.

**Exercise 13** Consider the system $x_{n+1} = rx_n(1 - x_n)$. Use a pocket calculator or a computer to trace the behaviour of the system for various values of $r$ and various initial values of $n$. (There are Internet programs that will do this for you, but it is more interesting to retrace the steps of the pioneers and do it yourself.)

(i) For $0 \leq r \leq 1$ you should find (and it is easy to prove) that $x_n \to 0$.

(ii) For $1 < r \leq 3$ you should find (and it is easy to prove) that $x_n$ tends to a unique value.

(iii) For $r$ a little bigger than 3, $x_n$ will (usually) oscillate between two values.

(iv) For $r = 3.5$, $x_n$ will (usually) oscillate between four values.

(v) As $r$ increases beyond this, matters rapidly become very complicated. (Of course, looking up the 'logistic map' on the Internet will produce clear accounts of the matter but the readers may well prefer to see how confusing things are before doing this.) Try $r = 3.9$.

The Hénon map is a map $T : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$T(x, y) = (y + 1 - ax^2, bx).$$

It was inspired by the Lorenz differential equation and also appeared to exhibit very odd properties for certain values of $a$ and $b$. For these values the iterates $\mathbf{x}_{n+1} = T\mathbf{x}_n$ either move off to infinity or move towards a well defined infinite set $A$. However, the sequence does not settle down towards one point of $A$ but moves around the entire set. The following artificial example may give some idea of what is involved.

**Exercise 14** Consider the map $S : \mathbb{C} \to \mathbb{C}$ given by

$$Sz = \frac{z}{|z|^{1/2}} \exp(i)$$

for $z \neq 0$, $S0 = 0$. If $z_0 \neq 0$ and $z_{n+1} = Sz_n$, describe the behaviour of the sequences $z_0, z_1, z_2, \ldots$.

However, the behaviour of the Hénon map is much more interesting than that of our example, since it appears to exhibit sensitive dependence on initial conditions. In other words, although we can be sure that after a large number of iterations we will be close to the set $A$, we cannot tell which part of $A$ we will be close to.

I said that the Hénon map 'appeared to exhibit very odd properties' since, as the reader will appreciate, it is very hard to distinguish between 'nearly chaotic' behaviour and 'truly chaotic' behaviour. There is a further point of difficulty. Even if we could prove that what seems to happen actually happens for some particular values of $a$ and $b$, this might be a property of the specially chosen $a$ and $b$. Since in 'real life' we can only specify $a$ and $b$ to a certain accuracy, this would not really settle the question of whether what we seem to see is actually taking place. What Carleson and Benedicks did is to show that it really does take place over a substantial range of choices of $a$ and $b$.

As the prize citation makes clear, Carleson has served mathematics in many other ways, through teaching, through encouragement of younger mathematicians, through influential books, service on committees, and much else. However, mathematicians are remembered for their theorems and not for their other services to mathematics. This is as it should be, but as one of those who benefited from Carleson's revival of the Mittag-Leffler Institute, I should add my thanks for nine gloriously happy months I spent there as a young mathematician.

So far, I have been talking to the tyro mathematician. Now let me address more general readers. In my mind's eye I see those readers in turn as engineers who wonder whether the matters considered here have any real connection with nature, politicians who wonder if they have any real use and lay people who wonder if they have any real interest.

Engineers may use the while worn jibe that they 'would never fly an aeroplane which depended on the study of sets of measure zero' (though, oddly enough, they are quite happy to fly warplanes which depend on imaginary numbers). They may claim that they use Fourier methods all the time but the kind of things we have discussed are 'pathological' and cannot bear any relation to the matters discussed here.

It is, of course, true that, at university, students 'solve' the problem of the plucked violin string by writing the form of the string as a Fourier sum with an infinite number of terms and then performing various 'formal' (that is to say 'magical') operations but the object of the exercise is to pass examinations and not to find out what actually happens to the string. In reality we must accept that we do not know the exact shape of the string at any time. Thus our task is to show that if we start with a shape which is close to our 'idealised initial shape' then our prediction will be close to the observed shape and in order to do this we must decide what it means for one function to be close to another, that is to say we must define a distance between functions.

In our previous discussion we used the fact that we could find nice functions arbitrarily close to nasty functions. Unfortunately the reverse is also true for all reasonable notions of distance. So long as we do not perform any numerical calculations, we can shut our eyes to this disagreeable fact but the moment we engage in any serious attempt to make numerical predictions (as in weather forecasting) what was merely 'some highbrow pure mathematical concern' becomes a major problem for the numerical analyst. The du Bois-Reymond example suggests that simply taking the maximum pointwise difference, that is to say, using the distance

$$d_\infty(f,g) = \sup_{-1/2\le t\le 1/2} |f(t)-g(t)|$$

may not always be the best way forward and more direct evidence from numerical analysis confirms this.

In many circumstances the best notion of distance to use is

$$d_2(f,g) = \|f-g\|_2 = \left(\int_{-1/2}^{1/2} |f(t)-g(t)|^2\,dt\right)^{1/2}$$

since, in some sense, it represents the 'power difference between two signals $f$ and $g$'. Engineers who use this metric may claim that they are only interested in the case when $f$ and $g$ are continuous. But this is exactly the same as claiming that we are only interested in rational numbers and so the real numbers are of no concern to us. Just as ordinary calculus involves all real numbers whether we like it or not, any serious calculus using $d_2$ automatically involves all $L^2$ functions whether we acknowledge this or not.

There is another way in which we can try to avoid mathematical problems. It is to assert that Nature only deals in smooth functions. Unfortunately this is not always true. It is a common experience that methods of numerical analysis designed to take advantage of high differentiability rarely work on real data, and I would certainly refuse to cross a bridge designed by an engineer who believed that wind speed varied in a nice smooth manner. The behaviour of noise in electrical circuits and the prices of stocks and shares may indeed be modeled by continuous functions, but it is notorious that those functions are nowhere differentiable. (It is worth noting that they also have an infinite number of maxima and minima in each interval so they fail to satisfy the conditions of Dirichlet's theorem.)

None of this means that engineers have to study Carleson's theorem, but it does suggest that they should be aware of the issues raised. After all, it shows a certain disrespect for Nature to think that all the problems she sets us can be resolved by the methods of the 19th century.

Next I address politicians who ask why society should pay for mathematicians to study the matters discussed here. Let me say straight away that I know no way in which Carleson's $L^2$ theorem contributes anything to the material satisfactions of mankind. For several hundred years the mathematicians of Europe pursued the goal of solving high order polynomial equations by radicals (essentially this just means finding a simple formula for the solution of standard equations). To the non-mathematical reader this must seem an obviously useful goal and so it seemed to mathematicians for the first couple of centuries. However, by the time Abel, Galois and others resolved the problem by showing that no simple formula can exist, it was clear that the answer was irrelevant for all practical purposes. In the same way, although throughout the nineteenth century most users of Fourier series thought the resolution of the pointwise convergence problem would have wide ranging practical use, the twentieth century has shown that, in fact, most practical applications involve notions of the distance between functions and appropriate approximation rather than pointwise convergence.

Jacobi wrote that Fourier reproached 'Abel and myself for not having given priority to our research in the theory of heat conduction. ... It is true that Fourier was of the opinion that the chief end of mathematics was the public good and the explanation of natural phenomena; but a philosopher such as he was should have known that the only goal is the honour of the human spirit and in this respect a question in the theory of the numbers is as valuable as a problem in physics'. With the complete, unexpected and beautiful resolution of a central problem in mathematics as well as his other successful solution of other major problems Carleson has honoured the human spirit and it is entirely appropriate that he should receive a prize named after another great upholder of the human spirit.

When we speak of great mathematics and great mathematicians, we may surely say that a society without mathematicians like a society without poets would not be aware of their absence but would, none the less, be a poorer society.

The politician may grant the truth of all I have said but remark that very few of the mathematicians supported by society reach the heights of a Carleson. Although most mathematicians do mathematics because they enjoy it and very few do it for the public good, it is possible to make a plausible defence of pure mathematics (that is to say mathematics for its own sake) on the grounds that, historically, some pure mathematics has later turned out to be useful. For example, although the problem of solution by radicals turned out to have no practical use, the methods developed to solve it gave rise to group theory (useful in Quantum Mechanics, crystallography and World War II code breaking) and the theory of fields (useful in communication theory).

One problem with such a historical defence is that it may come to resemble one of those patriotic histories in which everything is invented by an Englishman (or Russian or Frenchman according to taste). In fact ideas and methods flow in both directions across the ill defined frontiers of mathematics, physics and engineering. One can invent a history in which all the ideas of mathematics come from mathematicians but in reality mathematics has benefited from major contributions from physics and engineering and vice versa.

Having said all this, the reader may still be interested to know that Carleson's Corona theorem has fairly direct links with problems in control theory (which deals with the control of machines, electrical circuits and so on). The type of 'hard classical' Fourier analysis pursued by Carleson played a vital role in the emergence of Wavelet Theory which provides a new way of analysing complex data like photograph. (Wavelets are used in many areas of physics, engineering and biology but readers may be more interested to learn of their use in restoring the 'true sound' from old gramophone records.)

When we discussed strange attractors, we noted that Lorenz's work arose in considering weather forecasting. Although he worked with a 'toy model', the behaviour it showed, provided meteorologists with insight into why weather forecasting must have a time horizon beyond which 'sensitivity to small changes' (the so called 'butterfly effect') make accurate forecasting impractical. Developing this insight, they have realised that this time horizon is not fixed but depends on the state of the weather. Sometimes the weather permits accurate forecasting over long periods of time and sometimes it does not. Weather forecasters run their computer models many times making slight changes to the initial conditions and observe how fast the various forecasts spread out (ensemble forecasting). If they diverge rapidly, forecasts will probably only be accurate over a short time. If they diverge slowly, the forecasts can probably be relied on over a much longer period.

The outpouring of popular presentations of 'chaos theory' has obscured a fact known to aeronautical engineers since the earliest days of human flight. Insensitivity to small changes gives stability and predictability but reduces controlability. Early gliders were built for stability with the result that things went wrong only rarely but when they went wrong they stayed wrong. Controlability requires that only small changes are needed to give large effects. It is nearly impossible to balance a billiard cue on a table but very small motions allow me to keep it vertical on the tip of one finger. Provided that we do not overdo matters, 'chaotic' situations may represent an opportunity rather than a problem.

An interesting example of this comes from space travel. Since the time of Newton, mathematicians have been interested in the long term behaviour of the solar system. (The reader may say that it must be stable, otherwise we would not be here, but Newton and his followers had the option of invoking a clockmaker to reset the clock when it went badly wrong.) At the end of the 19th century, it is probable that most people who thought about the matter would have said that the solar system and similar objects were stable though mathematics could not yet prove it. As a result of the work of Poincaré, Kolmogorov and many others we might give a more nuanced answer. Most of the time, most systems of the type we are interested in show high stability (small changes in position and velocity make very little difference to the behaviour of the system even over long periods of time) but situations can arise (for example in near collisions) in which small changes can produce very different outcomes. If we deliberately place our spacecraft in such a situation, then a very small expenditure of energy can achieve changes in direction and velocity which would otherwise be impossible. This technique is routinely used for projects such as *Voyager 1*.

Modern technology has added to the number of complex systems whose long term behaviour is of obvious interest. Internet type systems which link many computers may work in satisfactory mode for several days and then freeze. Traffic on motorways can flow freely and then suddenly jam for no obvious reason. Using a mixture of advanced mathematics and ‘engineering intuition’ we can make sense of what happens to Internet systems and at least delay their collapse. Motorway traffic involves human beings and raises non-mathematical problems. (‘If the traffic jams on a motor way, the motorist blames bad luck, if we install traffic lights at the motorway entrances, the motorist blames the government’.)

Although there exist many books and articles proclaiming the imminent understanding of general complex systems, my suspicion is that we are still at the stage of examining particular properties and particular systems. Birkhoff’s ergodic theorem is a striking example of a particular type of property and Carleson’s work a deep example of the study a particular type of system. Even if the optimistic hope that we (or rather our great grand children) will be able to extract general properties which hold for general systems is not realised, the study of each particular system or property will increase our ability to deal with others. Carleson’s convergence theorem lies at the end of one research program, his work on Hénon maps is one of the markers at the beginning of another.

Finally, I should answer the general member of the public. Unlike engineers or the politicians, though on even less evidence, the general public has an even higher opinion of the cleverness and usefulness of mathematicians than they have themselves. The general public is perfectly happy to support mathematics and has no doubt that, in some undefined way, the work of mathematicians leads to a general increase in prosperity. However, the man or woman in the street is puzzled as to what satisfaction someone like Carleson can derive from doing work which can be understood by so few people.

To answer this question he or she should consider the satisfaction that a cook has in presenting a good dinner or a string quartet playing privately for their own enjoyment. It is not necessary to hear applause to feel the satisfaction of a job well done. The amateur painter or, at a lower level, the completer of a large jigsaw feels satisfaction whether or not anyone else is there to praise them. For mathematicians the satisfaction of resolving a problem is a reward in itself. (Fortunately for us, the satisfaction is only slightly dependent on the importance of the puzzle. I have seen distinguished mathematicians head over heels with delight at results which they know will interest no one else.)

The virtuoso violinist is rewarded by his or her own satisfaction, the praise of his or her peers and the applause of the multitude. The mathematician will never have the applause of the multitude (though an Abel prize comes close to it), but Carleson can enjoy the unstinted admiration of the mathematical world in which his life has been lived.

The mathematician does have one advantage over many other artists. Authors know that their books may be acclaimed by a generation of readers and critics alike and yet be condemned by the taste of the next generation. There is no sure test for poetic greatness. But someone who solves a problem that has baffled the finest mathematicians for a century and a half, knows that he has done a great thing. Names like Dirichlet, Riemann and Kolmogorov may mean nothing to the general reader but to mathematicians they are heroes. It has been given to Carleson to be the hero of a story in which the other actors are his own heroes.

## References

1. Bari, M.K.: A Treatise on Trigonometric Series. Pergamon, Elmsford (1964). (Two volumes, English translation)
2. Klein, F.: Development of Mathematics in the Nineteenth Century. Math. Sci. Press, Brookline (1979) (English translation by M. Ackerman)
3. Zygmund, A.: Trigonometric Series, vol. I, 2nd. ed. Cambridge Univ. Press, New York (1959)
4. Actes du Congrès International des Mathématiciens, 1970, Nice. Gauthier–Villars, Paris (1971)

# 2007

# S.R. Srinivasa Varadhan

ABEL
PRISEN

# Autobiography

**S.R. Srinivasa Varadhan**

According to my school records I was born on Jan 2, 1940, in the city of Madras, in the state of Madras in India, which was then a British colony. The city is now called Chennai and the state has become Tamil Nadu in the Republic of India.

My father was born in the last year of the nineteenth century, in 1899 and he married my mother in 1917, when he was eighteen and she was ten. I am an only child and my parents had been married for nearly twenty five years when I was born. Both my parents were the eldest siblings in rather large families and I have always received special attention from all my uncles, aunts, cousins, grandmothers and other assorted relatives. I was born so late that I did not really get to know either of my grandfathers.

As a child I grew up in several small towns not far from Madras. My father was a high school teacher and later principal in the county school system and was periodically transferred from one town to another within the county or district as it is called in India. That was both good and bad for me. I was treated with consideration by all my teachers. But I could not do any mischief at school without my father finding out right away.

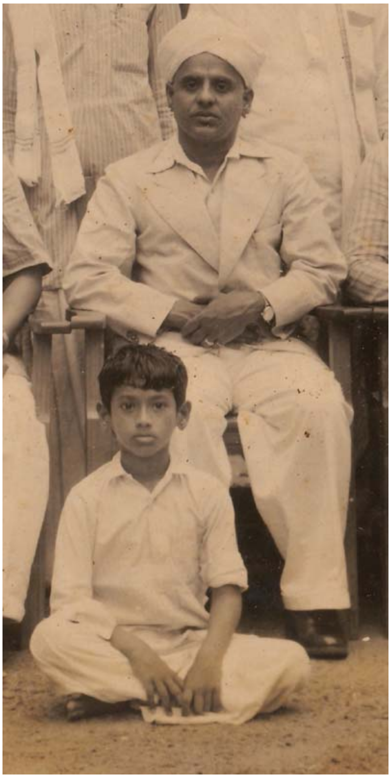
Growing up in these small towns was fun. There was plenty of time after school to play with friends on the riverbed that was mostly dry, or play indoors on rainy days. There was very little homework, and in fact there was only minimal learning at school. I did well relative to my class, but was not challenged in any sense. It was only in the last year of the high school, that my mathematics teacher took a special interest in a small group of us and would ask us to come to his house on weekends to do some mathematics problems for fun. More than anything else he taught me that solving mathematical problems or puzzles can be fun.

I had some vague ideas of becoming a doctor as a child. But once, with a group of fellow students from the high school, I went to a medical exhibition at the local

S.R.S. Varadhan (✉)
Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA
e-mail: varadhan@cims.nyu.edu

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_12, © Springer-Verlag Berlin Heidelberg 2010

(a) As a seven year old in traditional Brahmin attire



(b) As a nine year old in a group with teachers in school. I am there because my father was the principal. I am sitting at his feet. He is the one with the head dress (turban)

medical college where the medical students demonstrated their surgical skills on cadavers. That cured me of the desire to enter the medical profession.

In 1954, I moved out to live with my uncle in a suburb of Madras to attend a local college where we were required to study for two years before entering the University. This was a difficult transition. From a rather parochial school system where everything was taught in Tamil, the vernacular, I had to grasp new ideas that were being taught in English. Some of the professors were from UK and I could not quite understand their accent. I studied English and Tamil as my two languages. These were the harder subjects at college since we were expected to write critical essays, where as in school we were only required to remember facts and memorize poetry.

In addition to languages I studied Mathematics, Physics and Chemistry. I did well in all. I liked Physics the most, and I began slowly to see the connections between them.

BOARD HIGH SCHOOL, PONNERI.
S. S. L. C. Group (Tamil Section)
1953-54

(c) High school graduation picture. I am standing on the top row. My father as principal is there too, sitting in the front row wearing a turban

It was time to enter the University and the seats were limited and the competition was tough. I came from a Brahmin community, viewed by the government as privileged, and there was reverse discrimination. The goal was to get into a Honors program that led directly to a Master's degree in three years and saved you a year.

It was at this time that I heard of Statistics for the first time. I was told that there was only one college in the state that offered an Honours program in Statistics and they admitted only fourteen students each year. My father knew somebody who knew somebody who could help and luckily I was admitted. I had a choice between Physics and Statistics and I opted for Statistics.

Basically it was a three year program devoted to pure mathematics, probability and statistics. The Presidency College where the program was offered, was situated on the beach. You looked out the window and it was a sandy beach that was as nice as Ipanema. Meanwhile, my father had retired and moved out to the same town as my uncle and I stayed with my parents and commuted for a year. I wanted my parents to let me stay in the dorm for the last two years and they did, after much persuasion on my part.

I really enjoyed those two years. Long walks on the beach with a dozen or so close friends from my class. Lots of movies, arguments, discussions and a generally carefree student life. We are still a close group and see each other from time to time. Although I am the only academic, the others have gone on to do well in government, business and other areas. The studies themselves were relatively easy. I was learning a lot of new concepts and they all felt natural and I could do exceptionally well without much effort. The trouble with the system was that the expectations were not high and it was too rigid. We had only set courses and no electives.

(d) Picture with Kolmogorov during his visit to India (1962). I am third from the left behind Kolmogorov. With me are K.R. Parthasarathy (graduate student), B.P. Adhikari (Professor), me, J. Sethuraman (graduate student), C.R. Rao (Head of the Institute and my advisor) and P.K. Pathak (graduate student)

I graduated after three years having established a new record for grades obtained in the final examination where you are tested in one week on everything that you have learned in three years.

The usual career path at this point is to write a competitive civil service examination that is used to recruit high level government officials. My parents expected me to sit for this. If you succeed, you are set for life. I wanted to go for research. When my father saw that I was firm, he supported me and I went off to Calcutta to study at the Indian Statistical Institute. I had no idea what I was supposed to do. When I arrived in August, 1959, they gave me a desk and expected me to write a thesis in three years. No graduate courses were offered. There were seminars that were optional. I took one on point set topology from Varadarajan and one on measure theory from Bahadur. I did not know what to do and so learned to play bridge and I played a lot. But slowly over the year I met up with a few other graduate students who were there already for a year or two, and we organized our own seminars and programs of study. We lectured to each other, formulated our own problems and tried to solve them. It was a wonderful learning experience and I started to do "Research".

The atmosphere at ISI was very stimulating. We had a steady stream of distinguished visitors, during the pleasant winter months. Sir Ronald Fisher came every year. I had just missed Norbert Wiener, who had come the year before I went there. We had tea, twice a day, when we all met at the tea room and talked informally. J.B.S. Haldane was a regular at tea, constantly puffing on his cigar. There was always some excitement in the air and some thing or other was always happening. During my second and third year I learned a lot of functional analysis and more or less finished my dissertation. I had begun to learn Markov Processes and wanted to start working on it. During my third year Kolmogorov visited us for two months and I gave a seminar on my thesis and he was in the audience asking questions. He spoke no English, none of us knew any Russian, and so we talked through an interpreter who knew French. A group of graduate students accompanied him on a two week tour of parts of India. He was a member of my thesis committee and brought a copy of my thesis with him, to Moscow, promising to send a report from there. The report came after six months, but only after Parthasarathy, a colleague who went to Moscow on an exchange program, provided a steady daily reminder.

At this point a year had gone by after my dissertation was finished, I formally got my Ph.D., and it was time to go abroad, for a postdoctoral study. In those post Sputnik days that meant USA. Varadarajan, who had just returned to India, after three years in the US, suggested that I go to NYU, to Courant Institute. He wrote a letter on my behalf that went unacknowledged for nearly three months. In the end I was offered a postdoctoral position and I arrived in New York in the Fall of 1963.

There was an international organization that greeted foreign students upon arrival, and they offered to find a place for me to stay for my first night. They met me at the old Idlewild airport and arranged for me to stay in a hotel on Times Square. When I showed up at Courant the next day and told them where I was staying, they were horrified and found me a place to stay on Tenth and University Place. Later I found myself a studio in the West Village.

Before leaving India I became engaged to be married to Vasundara and the wedding was to take place the following June, in Madras. The year went by very fast. I met a small group of Indian academics and over the years we have remained close.

Work went well at Courant. I saw a lot of Mony Donsker, who was the principal probabilist at Courant. There were lots of visitors, seminars and other activities that kept me busy during the week and I socialized with my Indian group during weekends.

Donsker and I, discussed a lot. He had many students working on an assortment of problems. This provided me with a broad perspective and I was beginning to view things from a slightly different angle that made a lot of sense to me. Donsker was interested in specific problems that he wanted to be solved and I was always interested in developing techniques that would enable one to solve a class of problems. We complemented each other and worked closely for over twenty-five years. I went to India to get married in the summer and we found a nice apartment in Washington Square Village. My wife, who was only sixteen, started undergraduate studies at Washington Square College. When my post doctoral fellowship was renewed for two more years it proved ideal for both of us. Then, I joined the faculty in the Fall of 1966, as an assistant professor.

Donsker had a student named Schilder who did a very fine thesis on Laplace asymptotics on Wiener space. I thought a lot about it, and felt that one could develop it

(e) This picture is at the airport on my way to Courant as a postdoc (Fall 64)

(f) This picture is at the airport, returning to USA with Vasu immediately after our wedding in the summer of 1964

considerably. This has been a major thrust of my research over the years and goes by the name of Large Deviation Theory. It is the technique of estimating precisely, in a logarithmic scale, how small the probabilities of certain rare events are. I have found this to be a problem that crops up in many different contexts, and have returned to it again and again during my career.

We used to have a joint seminar with Rockefeller Institute, where Mark Kac was a professor. We would go up to 68th and York often. Once on the way back, in a taxi, someone commented on a result by Cieselski, a Polish mathematician and I immediately saw connections with large deviations and diffusions. The diffusion processes are Markov processes with continuous paths, closely related to second order elliptic partial differential operators. I was able to work out this connection and it proved to be a nice result.

There was a graduate student at Rockefeller University, Dan Stroock, who came by to talk to me about this and we became close friends. He came to Courant in 1966 and stayed for six years. We worked very closely during this period and I believe we changed the way Markov processes were viewed. We introduced systematically methods based on martingales that have become more or less standard today. We jointly wrote a book that appeared in 1979 and has been received well.

In the fall of 1973, we (the family at this time included two sons Gopal born in 69 and we lost him on 9/11/01, and Ashok born 1972) had just returned after a sabbatical from Sweden and India. Donsker as usual was interested in a specific problem about a probabilistic explanation of the variational formula for the ground state energy of the Schrödinger operator. I remember going to Durham to give a talk at Duke, and sitting in their library before my talk, I saw the connection. This led to a whole series of results on large deviation theory that Donsker and I worked on, till his untimely death in 1991.

I had become an Associate professor in 1968, won an Alfred P. Sloan Fellowship in 1969, and became a Professor in 1972. This was also a tough time at NYU, involving a financial crisis and the sale of the Bronx campus. We moved to Stanford for a year long visit in 1976–77, enjoyed the outdoor life and the open skies. But professionally I could not wait to return to New York. I also started working occasionally with George Papanicolau, who had lots of applied problems that often required new techniques. I was invited to give a talk at the International Congress of Mathematicians, in Helsinki in 1978 which was a kind of recognition.

I served for four years from 1980–84 as Director of Courant and survived the experience. I was surprised to find that the central administration consisted for the most part of talented individuals doing their best at a difficult task. I began to appreciate for the first time that we at Courant were part of a larger University. My research slowed down some due to the demands of administrative work, but I was able to continue with it. I had started to attract a steady stream of graduate students by this time and enjoyed working with them. I became recognized as an expert on Large Deviations, lectured often on it and wrote a set of lecture notes on the topic.

This was also the time when Wall Street discovered probability. The Black and Scholes model was suddenly very popular. I started some consulting for a small company that managed pension funds and worked closely with Harry Markowitz, who was also a consultant. Harry went on to win the Nobel Prize in Economics. Richard Brignoli, who was the CEO of the company was a maverick and that made the experience all the more enjoyable. It lasted several years, until the company went under. Not because it performed poorly, but the partners had a fallout, fought in court and the lawyers ended up getting everything. Our son Gopal got his first experience working for the company in the summer and later part time during his senior year at college. He chose to stay with the financial industry and his younger brother Ashok has followed him.

I took a sabbatical in 84–85, after serving for four years as Director of the Courant Institute supported partly on a Guggenheim fellowship. Around this time I went to a conference in Marseilles and we usually walked down to the sea and back up the hill, to the campus at Luminy, after lunch. During one of the walks, George explained to me a problem of establishing what he called bulk diffusion under rescaling for interacting diffusions. The problem intrigued me and I thought it would be a simple problem. I spent my sabbatical year thinking about it and found it rather difficult. I found out later that there were a whole class of such problems generally referred to as problems of hydrodynamic scaling. I worked on the problem a lot, but made no progress. Although I had some ideas, they were not sufficient to solve the problem. We had a seminar at which the speaker Josef Fritz from Budapest, proved a result of similar type for a different model. I thought about the new model and found to my surprise, that my ideas worked well for this model and could provide a better result than Fritz. George, Guo who was a student of George, and I worked feverishly on this model and completed our work. This led me to a whole set of problems requiring new methods and I worked closely with a very talented younger colleague, H.T. Yau. He has since developed the subject and has taken it in a variety of different directions.

The period from 1984–94 was personally very hard for me. My parents grew old and infirm. Since I had no siblings, I had the responsibility for their welfare. This meant frequent and prolonged trips back and forth between New York and Madras. This was hard on my wife, who had stayed in school part time, while raising the two boys, and earned a Ph.D. degree (her fourth degree at NYU) in 1985. My father died in 1990 and my mother in 1994. My wife lost her mother in 1991. The generation above us started thinning out slowly.

There was a bit of a problem at the office in 1992. George, who was to have been the director, left to go to Stanford and there was no time to find a replacement. I was drafted to serve for two years during which period a search committee was appointed and found an excellent choice of director in Dave McLaughlin.

This was also a time when I was receiving honors and recognition and that was gratifying. I was elected a Fellow of American Academy of Arts and Sciences and an Associate Fellow of the Third World Academy of Sciences in 1988, a Fellow of the Institute of Mathematical Statistics in 1991, a member of National Academy of Sciences in 1995, and a Fellow of The Royal Society in 1998. I received the George D. Birkhoff Prize from AMS and SIAM in 1994 and the Leroy P. Steele prize of the AMS that I shared with Dan Stroock in 1996. NYU gave me the Margaret and Herman Sokol award in 1995. I was invited to give a plenary address at the International Congress of Mathematicians in Zurich, in 1994. I was appointed as Frank J. Gould Professor of Science at NYU. In 2001, I was elected to serve as the President of Institute of Mathematical Sciences for 2002–03. I received honorary degrees from University of Paris in 2003 and Indian Statistical Institute in 2004.

Tragedy hit on 9/11/01 when we lost our older son Gopal who was working at the World Trade Center that day. I was in Paris, visiting IHP for a month and was told of the crash by my wife who called me from New York. It took four days to get back to New York. We miss him. Our grandson Gavin, born to our younger son Ashok and his wife Maggie helps take some of the edge off our grief.

Professionally I consider myself to have been fortunate, in that my career spanned a period, when science and mathematics were generously funded by the public. I had a stimulating ambience at Courant that has traditionally provided a very supportive environment for their younger faculty. I have enjoyed working closely with my colleagues as well as everyone of my nearly thirty doctoral students. Finally, I hope to have several more years of productive academic life.

# A Personal Perspective on Raghu Varadhan's Role in the Development of Stochastic Analysis

**Terry Lyons**

## 1 A Great Day for the Coin Flippers

I know Raghu Varadhan professionally but not personally—that is to say we have attended some of the same conferences and Oberwolfach meetings, and even the odd meal while waiting for trains home. Still, it is obvious to me, and I am sure to anyone else who comes close, that he is a person of great humanity who generates warmth and humour whenever he is in the room. A few months after the award of Fields Medals to Werner, Okounkov and Tao in Madrid, Varadhan and I were both in a group of mathematicians talking about the event. I remember clearly Varadhan's concise summary of the business as "A great day for the coin flippers". It certainly was: all three used probability in their ground-breaking work and, for the first two, Stochastic Analysis has been a decisive part of their mathematical toolbox. We were all excited that stochastic ideas were having such a substantial effect across areas as far apart as conformal field theory, geometry and number theory. We were also delighted that these achievements were recognized. To me, Varadhan's remark seemed to capture his modesty and humour rather well. Surely it was another excellent day for the coin flippers when Varadhan was awarded the Abel Prize.

## 2 Stochastic Analysis

Stochastic Analysis is an area of mathematics that, in a little over 60 years, has grown from almost nothing to a significant field. It has importance for its intrinsic

T. Lyons (✉)
Mathematical Institute and Oxford-Man Institute, University of Oxford, Oxford
OX1 3LB, UK
e-mail: tlyons@maths.ox.ac.uk

T. Lyons
Wales Institute of Mathematical and Computational Sciences, Swansea, SA2 8PP, UK

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7_13, © Springer-Verlag Berlin Heidelberg 2010

interest and for its contributions to other foundational "pure" areas of mathematics, as well as for its contributions to the applications of mathematics. These applications seem to be on an enormous and expanding scale, spanning engineering and climate modelling. It is totally clear to those that work in Stochastic Analysis that Varadhan has shaped the subject in hugely significant ways, both in his personal contributions—in the sense of direction he has given—and in the guidance he has provided to his many outstanding colleagues and students.

I am no historian, and my remarks assigning historical credit should be treated with caution. Given this caveat, I would like to try to place Varadhan's contributions in context by mentioning a few of the landmark contributions that have shaped the Stochastic Analysis scenery. At least for me, the first wave in the development of Stochastic Analysis is associated with a Russian, a Frenchman, a Japanese man, and an American. The Russian is Kolmogorov, a tower of 20th century mathematics who gave us a rigorous framework for the mathematical study of probability. The Frenchman is Lévy, who made a detailed mathematical study of Brownian motion and Lévy processes. The Japanese man is Itô, who extended differential calculus to the Brownian case and so gave us stochastic calculus. The American is Doob, who gave us the ubiquitous martingale with its optional stopping theorem and up-crossing lemma, and the tools required to provide mathematical confirmation that, in a fair world, there are no free lunches. It is impossible to convey more than a few hints as to the full significance of each of these mathematicians' contributions.

Kolmogorov's 66 page monograph, which in 1933 set out a rigorous framework of probability, and his strong law of large numbers, form essential building blocks allowing the rigorous mathematical study of infinite-dimensional probabilistic objects, such as Brownian motion (as studied by Lévy), to flourish. Modern mobile phone technology depends on the strong law of large numbers to separate the transmissions from different phones and allow robust transmission in the context of noise and interference.

Newtonian calculus and differential equations are the classical tools for expressing interactions between evolving systems. Itô's stochastic calculus (1942) extends the remit of differential equations to systems driven by random processes such as Brownian motion. Itô's theory gave a direct connection between probability theory and a wide class of second order parabolic partial differential equations. It also provided the framework for many fundamental applications, for example, the continuous time Kalman–Bucy filter which revolutionized the field of estimation and was a major breakthrough in guidance technology (the lunar landing of Apollo is a well documented example). Practical application in non-linear settings requires the construction of numerical approximations to the solutions to certain non-linear PDEs (partial differential equations). Stochastic filters are used in almost all modern military and commercial control systems.

Doob took ideas that were well known in function theory about the behaviour of harmonic functions on the boundary of the disc (Fatou's theorem, Littlewood's theorem) and showed how they had probabilistic parallels. He introduced martingales, established optional stopping and up-crossing lemmas, quantified the oscillatory behaviour, and established convergence properties. In effect, he explained how to generalise the notion of a parabolic PDE to the context of functions defined on spaces of paths. It was a remarkable achievement, demonstrating the enormous power a well chosen abstraction can have.

There is, and will remain, a substantial demand for financial intermediaries, who are able to supply trustworthy and economically priced products, that allow individuals and businesses to insure against financial risks in areas of their business where they have limited expertise and control. Financial intermediaries, who supply these products, must hedge their liabilities, and in general, this hedge will be dynamic and change from day to day, so as to ensure that, at the time an insurance policy matures, the intermediary has the appropriate resources to pay any claim against it.

The correction term (Itô's formula) to Newtonian Calculus to account for the volatility of market prices is at the heart of this hedging process. The provision of these basic insurance activities has become standard over the last 15 years, increasing transparency and forcing margins down. The need for insurance against interest rate and currency fluctuations on a huge scale will not disappear as a result of the current financial storm although one might expect providers to increase their margins somewhat.

Of course, these four mathematicians were not alone (Chung, Dynkin, Khinchine, McKean, Malliavin, and Meyer were some of the others). Nor had the full significance of their contributions been appreciated when Varadhan began to make his own critical interventions.

None the less, it was clear (to those in the area) that Stochastic Analysis had become a set of tools and techniques that could give insight into really quite high-dimensional systems. If one models the evolution of a population, one might use a parabolic PDE to describe the mean local density or intensity of the population as it evolved. Probability can inject additional insight into such deterministic systems. Stochastic analysts have developed tools that model the behaviour of the underlying stochastic population as well as its mean behaviour. This extra step requires more effort and genuinely probabilistic tools; but the value of this distinction, emphasizing the importance of sample paths, provides a clear advantage. In settings, such as control or modelling financial markets, where one has to understand how to interact with and respond to the actual—if uncertain—evolution of the world, the distinction is decisive. Average behaviour is interesting but is often far from the whole story.

Probabilists are not unhappy when the their systems are so large and internally homogeneous that on large scales they behave as if they were deterministic. They sometimes call this the fluid or hydrodynamic limit, even though fluids frequently retain very random behaviour on normal as well as microscopic scales!

Varadhan, his collaborators, and his students have made quite fundamental contributions linking the theory of probability to the theory of PDEs, in capturing the information about systems that PDEs miss (large deviations), and in proving that interesting microscopic random systems do indeed have hydrodynamic limiting behaviour. These contributions place him at the core of these developments. I am sure that in his modesty, Varadhan would be the first to note that he is not alone in making fundamental contributions to Stochastic Analysis (for example the work of Friedlin and Ventcell was around the same time and tackles similar issues to some of Varadhan's work), but I am sure that everyone in the field was delighted when Varadhan was awarded the Abel Prize.

## 3 Varadhan

Varadhan's contributions are plentiful, original, beautiful, and surprising. They have had strategic significance in many different directions, so that no-one except perhaps Varadhan could be authoritative on them all, and I doubt if Varadhan could have guessed the range of ways they would be used. I will not (and cannot) attempt a comprehensive survey. I agreed to write this article because I thought it was a great opportunity to express my (inevitably limited) understanding for some of the ways these results have had an impact on the development of Stochastic Analysis, and to explain, through some of Varadhan's striking work, why I think Stochastic Analysis remains an exciting cornerstone of modern mathematics.

Amongst Varadhan's contributions to the development of probability and analysis, I should, at the very least, draw attention to the creation and development of the theory of Large Deviations with Donsker, upon which so much of our understanding of stochastic systems depends. I must also mention the development, alone and with Stroock, of the martingale method for characterising diffusions. This work, with its roots firmly in the probability theory developed by Doob et al., showed the power of the new methods. In a deep piece of work, Stroock and Varadhan produced the first truly satisfactory treatment of elliptic second order parabolic PDEs (in non-divergence form) with continuous coefficients; through this work they demonstrated the power of the new technologies. I should also draw attention to the beautiful work on short time behaviour of diffusions, and mention Varadhan's major foundational contributions to the theory of interacting particle systems and to the development of a theory of hydrodynamic limits for these systems. Varadhan's seminal work with Lions and Papanicolou on the homogenization of the Hamilton–Jacobi equation also has to be on any list, although, unfortunately, space and time mean that I cannot include everything I might wish.

## 4 Independent or Uncorrelated

It might seem like a contradiction, but rare events happen. They can happen frequently, and the nature of the rare events that happen can have an impact in significant ways on the systems around us. I drive a car. I only do this because I believe it is really unlikely that I will have a serious accident—I see it as a rare event. On the other hand, there are many millions of drivers, and looked at across the whole population it is very likely that there will be several car accidents in a single day, and that, sadly, some will be very serious. Car accidents are rare, but they will happen because there are lots of cars.

Now, suppose that there are two classes of accidents: $A$ and $B$. Then asserting that $A$ and $B$ are rare events is essentially saying that $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are both small. However, in this context it is worth appreciating that, in general, rare events are not likely to be equally rare. For example, it may be that $\mathbb{P}(A) = 1/100$ and $\mathbb{P}(B) = 1/1000$. In this case it is trivial to see that, conditional on one of these two events occurring, then it is about 10 times more likely to be an event of type $A$

than of type $B$ which occurs. It does not take much imagination to realize that in large populations which interact in nonlinear ways, it becomes very important to understand the rare events and their relative probability. Rare events can play a key role in determining the function of the entire system.

The theory of Large Deviations provides a systematic mathematical framework for describing and estimating the probabilities of certain rare or exceptional events and predicting their consequences.

Large Deviation theory is a crucial counterbalance to the much better known Central Limit Theorem. The latter remarkable result considers the accumulated effect of repeated independent events. It is a remarkable result that says that the sample mean of the sum of independent and identically distributed variables $X_i$ with finite variance and mean zero has a distribution that is very close to Gaussian; if $A$ is open, then:

$$\mathbb{P}\left(\frac{\sum_{i=1}^{N} X_i}{\sqrt{N}} \in A\right) = \int_A \frac{e^{-\frac{x\sigma^{-1}x}{2}}dx}{|\sigma|^{1/2}(2\pi)^{d/2}} + o(1)$$

where $\sigma$ is the covariance of $X$. The strength and the implications of the Central Limit Theorem cannot be overestimated. It is a hugely valuable theorem, and it can be generalised to random paths.

The limiting distribution is simple and easily described and is entirely determined by the correlations of the underlying incremental event $X$. However, the central limit theorem only refers to the likely events. It says that, for the vast majority of the time, the sample mean behaves as if it was a multivariate normally distributed random variable; the covariance $\sigma$ of $X$ is all you have to know to understand the distribution of these typical fluctuations.

However, the Central Limit Theorem has nothing to say about rare events, those that happen on the scale of probabilities $e^{-cN}$ or for the probabilities of fluctuations of the sample mean that are $o(1)$, say. These rare events often have a rich behaviour, and at the same time they often play a decisive role in the behaviour of complex systems.

**4.1. A Simple Example.** Let $X_i \in \mathbb{Z}^d$ be a sequence of independent, identically distributed, random outcomes with mean zero and finite support. Suppose the symmetric quadratic form (known as the covariance):

$$C(u,v) := \mathbb{E}((u.X)(v.X))$$

is finite and non-degenerate, with inverse $g$. Consider the random walk on $\mathbb{Z}^d$ whose transition is $X$. That is to say, let $S_i$ represented the accumulation of these events

$$S_n = \sum_{j=0}^{n} X_j.$$

It is a basic problem to understand the macroscopic behaviour of the process $S_n$, and from a certain point of view Donsker's Functional Central Limit Theorem provides a very complete answer (see M.D. Donsker, "Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems", Annals of Mathematical Statistics, 23:277–281, 1952). The process $B_t^{(n)} := g^{1/2}n^{-1/2}S_{\lfloor nt \rfloor}$ is a very close approximation to Brownian motion; more precisely, its law as a measure on paths converges in the weak topology to Wiener measure.

It is standard practise in many areas of modelling where one accepts that there is a natural independence over disjoint time horizons to measure short term correlations, and then to model the process over longer time horizons using a renormalized Brownian motion. Donsker's theorem provides a limited justification for this. But there are dangers and, in the context of rare events, it is simply unsound.

For example, a credit agency might wish to assess the probability that a product could fail. They might believe it was safe because, although each component asset was risky, they believed that the probability of a large number of the components failing at one time was very small. This would be the case, if the components behaved independently. However, it would be inordinately difficult to create an empirical test for independence of a large number of these components. However, they may well be able to carry out accurate empirical tests which justify assuming that the values of the components are uncorrelated. It might also be reasonable to assume that the incremental behaviour of these values is independent from one time step to another. Is it reasonable for the agency to use a $d$-dimensional Brownian motion to model the evolution of these assets? In general, there is a scale on which the answer is "yes", but there is a larger scale on which this is a profoundly dangerous thing to do. The large deviations for the actual process might be quite different to the large deviations of a Brownian motion. In general, assuming that a number of events are independent leads to unrealistically small probabilities of rare events. There might well be strong empirical evidence that there is a 1 in $10^4$ probability that two of the assets lose their value, and the assumption of independence would set the probability of three losing their value at about 1 part in $10^6$. We can see how reliable this approach is by considering an example. Suppose that $X$ takes one of four values

$$\{1,1,1\},\ \{1,-1,-1\},\ \{-1,1,-1\},\ \{-1,-1,1\}$$

and each is equally likely. Then it is easy to check that $X$ has the same covariance as the increment of a 3-dimensional Brownian motion over a unit time step. Let $X_j$ be independent with distribution equal to $X$ and set

$$S_n = \sum_{j=0}^{n} X_j.$$

The central limit theorem suggests that $S_n$ should behave like a Brownian motion and the coordinates should be independent. We can see that this is not true over large scales. Name the coordinate components of $S_n$

$$S_n = (a_n, b_n, c_n).$$

What happens if $a_n$ and $b_n$ are at least $n - m$? It must be the case that

$$|\{j \mid X_j = \{1, 1, 1\}\}| > n - 2m$$

and so $c_n > n - 2m$, and for $m < n/2$ one sees that exceptional behaviour in the first two coordinates forces exceptional behaviour for the third coordinate, contradicting the intuition gained from the fact that the variables are uncorrelated, and the Central Limit Theorem.

If $\hat{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of the $X_i$ and $E$ is a Borel set with $0 \notin \bar{E}$ then standard Large Deviations theory gives asymptotic exponential decay rates for the probability that $\hat{X}_n$ lies in $E$:

$$\begin{aligned}\liminf \frac{1}{n} \log \mathbb{P}(\hat{X}_n \in E) \geq \inf_{e \in E^o} I(z),& \\ I(z) = \sup_{\lambda \in \mathbb{R}^d} (\lambda x - \log \phi(\lambda)),& \\ \phi(\lambda) = \mathbb{E}(e^{\lambda X_1})&\end{aligned}$$

where $I$ is the Legendre transform of the log of the Laplace transform

$$\phi((x, y, z)) := \frac{1}{4}(e^{2x} + e^{2y} + e^{2z} + e^{(x+y+z)})(\cosh(x + y + z) - \sinh(x + y + z))$$

of the distribution of $X$.

We can compare this with the equivalent upper bound for the case where the increment comes from Brownian motion

$$\begin{aligned}\phi(\lambda) &= \mathbb{E}(e^{\lambda B_1}) \\ &= \exp\left(\frac{1}{2}|\lambda|^2\right)\end{aligned}$$

the function $\phi$ is quite different; the difference between the two rate functions $I(z) = \sup_{\lambda \in \mathbb{R}^d}(\lambda x - \log \phi(\lambda))$ would show the differences and thus this dichotomy immediately.

Given examples like the above (where the coordinates of the increments are even pair-wise independent, as well as uncorrelated) it is a surprise to this author that, in a wide range of contexts, one observes people quoting incredibly small probabilities for the occurrence of multiple simultaneous coincidences where there cannot be empirical evidence for these probabilities (because they are too small) and where there is no reasonable justification for the joint independence of the events. We can understand that netting makes sense for a few assets in basket of bonds but systematic collapse cannot be ruled out on the basis of empirical evidence. Similar issues arise in criminal law where any claim that the probability of an alternative explanation has *extremely* small probability, typically obtained by multiplying probabilities, is more likely to be a confirmation of the stochastic illiteracy of the person making the claim than a statement of fact.

Large deviations are not predicted by the co-variance of a system.

## 5 Diffusion Equations and PDE

When pollution spreads through soil, when heat moves away from a semiconductor, or when fluid moves through a pipe, one understand that the observed phenomena represent the averaged behaviour of a huge cloud of microscopic diffusive particles. See, e.g., Fig. 1. But at least in the first two, and for systems with low Reynolds number in the third, the randomness has been so smoothed out on normal scales (by the Central Limit Theorem) that the behaviour has become steady and predictable. Second order elliptic and parabolic PDEs are basic mathematical tools used to model bulk continuum behaviour; these tools are incredibly effective in predicting the behaviour of these systems when they can be applied. However, there are settings where the diffusion is highly inhomogeneous, such as where there are layers of insulation or impervious obstacles distorting this diffusive behaviour even into a lower dimensional flow. In the case of pollution flowing through a porous media, it is easy to imagine that the diffusion is supported on a set of fractional-dimension. Today we see a division between the Stochastic Analysis methods which are well adapted to describing bulk diffusivity on the widest classes of domain, such as fractal sets, and the PDEs methods which depend on the locally smooth nature of the domain. Progress in understanding this more general setting is steady and many questions which have remained open for considerable periods, for example about uniqueness of diffusions on regular fractals, are now being solved as deeper understandings of things like Harnack inequalities and Dirichlet Forms emerge.[1]

I would not want to suggest that PDE methods are in any sense redundant; they are very effective when they can be used and powerful indicators of how to proceed

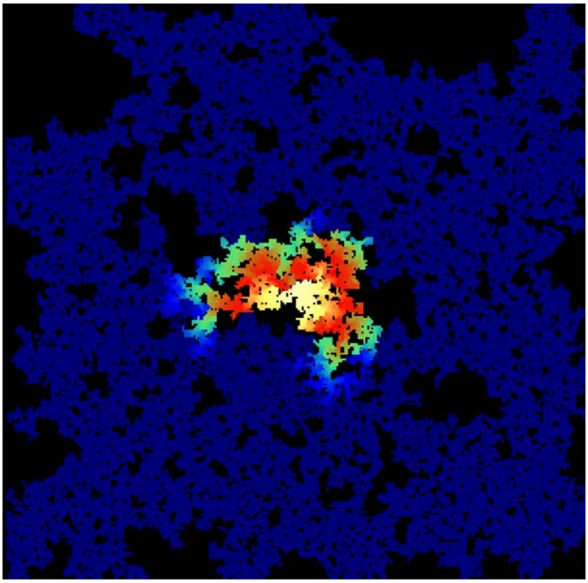


**Fig. 1** Heat diffusing in a random fractal environment—as simulated by a programme by Martin Barlow

[1] For a survey see [1, 2].

in more general settings where they cannot be appropriate. For example, it is almost a triviality that, up to scaling, there is only one translation and rotation invariant second order differential operator, and one translation and rotation invariant Markov diffusion on $\mathbb{R}^n$. This observation leads one to ask for uniqueness of diffusions respecting the symmetries of a regular fractal. The Bass–Barlow proof of uniqueness for the diffusion on the Sierpinski carpet is very hard!

The flow of understanding is not one way and Varadhan, both alone and with collaborators (particularly with Stroock) demonstrated very clearly in his early work that Stochastic Analysis is a powerful tool for tackling some of the harder and more fundamental questions associated to PDEs. If PDEs model the approximately deterministic macroscopic behaviour of large diffusive populations, then there will be cases where the diffusivity of individual elements in the population remains unchanged by the evolution of neighbouring elements, but in other cases one would expect significant interaction. In this latter case the coefficients in the PDEs modelling the macroscopic diffusion will depend on the solution for their values; a priori they are unknown and the PDEs will be non-linear; and when the environment where the diffusion takes place is rapidly fluctuating on small scales the coefficients will oscillate wildly. Even within the elliptic setting, where the diffusion sees a full ball of neighbours instantly (and is not forced into, for example, cracks or fractals), the need for a priori estimates in the study of non-linear PDEs and the need to understand the large scale behaviour of diffusion in highly fluctuating media each separately justify the extension of the mathematical theory of linear parabolic PDEs to the case where the coefficients of these equations are not smooth. The precise mathematical interpretation of a PDE with less than smooth coefficients remains subtle, even in the elliptic case.

**5.1. Elliptic PDEs in Non-divergence Form with Continuous Elliptic Coefficients—Uniqueness of Solution.** A central difficulty is to provide an interpretation of "solution" that is embracing enough to capture the actual diffusive possibilities, while at the same time have an equation that is precise enough that the initial data provided does indeed provide for the unique evolution of the system, and hidden variability does not remain. The approaches split according to whether or not the equations are in divergence form. A key question for the non-divergence form has always been to understand the linear backward operator

$$L := \frac{1}{2}\sum_{i,j\le n} a_{ij}(t,x)\frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i\le n} b_i(x)\frac{\partial}{\partial x_i} \tag{5.1}$$

where the coefficients are just bounded. Success in such an a priori approach would allow meaning in a rich and straightforward way for the widest classes of fully nonlinear equations.

A fundamental problem has been to establish the existence and uniqueness of fundamental solutions and the existence of an associated semigroup $P_t$. Under smoothness assumptions much was known. Nirenberg [10] had established a maximum principle (uniqueness) if the solution was $C^2$, while Dynkin and others had established uniqueness and existence results that went up to the Hölder continuous coefficient case. A very important contribution came from Tanaka (1964) and independently from Krylov (1966) who established compactness results and showed the existence of (possibly non-unique) semigroups associated to the operator (5.1) in the continuous coefficient case. However, neither Krylov nor Tanaka proved uniqueness, and uniqueness is every bit as much an issue as existence; without it the standard PDEs associated to (5.1) would not be well defined.

In a beautifully written pair of papers, Stroock and Varadhan completely solved this problem in the case where the coefficient $a_{ij}$ was uniformly elliptic, bounded and uniformly continuous, and $b_i$ was bounded and measurable. The papers were remarkable in so many different ways, giving us new and flexible probabilistic machinery that is still used routinely today, and at the same time demonstrating its worth by solving this basic problem in the analysis of PDEs.

Varadhan [19] had been exploring exponential martingales. Then, in the remarkable papers with Stroock, [12, 13], they assigned a Markov diffusion process to the elliptic PDE (5.1). They proved that in the case where $a$ is continuous and $b$ is bounded and measurable, then there is at most (and so exactly) one Markov process $X$ so that for each $T > 0$ and each fixed $\theta$ in $\mathbb{R}^d$

$$e^{\int_T^t (\theta \cdot dX_t - \frac{1}{2}\theta \cdot a\theta dt - -\theta \cdot b dt)}$$

is a martingale. Everything seemed to slot into place. We were given a new technology of weak solutions to stochastic differential equations, and a completely novel way, based on using exponential martingales, to prove that such solutions are unique. At the same time the results of Tanaka and Krylov were extended to show the existence of the weak solution. Stroock and Varadhan moved away from a study of solutions to a PDE (which were obscure) to looking at the action of the operator (5.1) on an exponential function (which was a local calculation and straightforward) and used stochastic integral equations to characterise the Markov process. The proofs are deep and wide ranging; in their later book on the subject they exploit techniques from topics such as singular integrals to get the estimates they need.

One can easily widen the methodology to give good local notions of $L$-supersolutions and $L$-subsolutions ([15], Remarks just after Thm 3.1), existence of solutions, and a maximum principle ([15] Thm 3.3) so solutions have the appropriate uniqueness. Viscosity-style solutions are not required for a good theory in this setting.

It seemed for many years that this intrinsic difficulty prevented these uniqueness results being extended to the uniformly elliptic bounded measurable case; but in 1997, in [9], Nadirashvili showed that the Stroock–Varadhan result is the best possible and gave the first counterexample to uniqueness in the bounded measurable case.

We emphasise that, although this is a result settling a key result in PDE theory, it, and the methodology developed around it, are important to probability. We should also not be surprised that probability seems the best route, and perhaps the only route, to studying these PDEs in depth. A Feller diffusion processes $\mathrm{X}_t$ is robustly associated to its transition semigroup defined on $C_0$ by

$$P_t(f)(x) := \mathbb{E}(f(X_t) \mid X_0 = x).$$

This semigroup characterises the diffusion, and its infinitesimal generator defines a closed operator $L$ on $C_0$. Without stochastic differential equations, any study of the process associated to it had to go in the reverse direction. Close $L$, use the fundamental solutions to the closed PDE to identify the semigroup, and use the semigroup to get the process. This is really quite problematic because PDEs locally characterise the behaviour of functions $f$. The beauty of the martingale characterisation of $X$ is that, in contrast to the semigroup description, it is also a local characterisation of $X$. With serious work and the skill of masters, this method would lead to the proof of appropriate uniqueness and existence statements for solutions to these PDEs and many other systems.

At the time of Stroock and Varadhan's work, I do not think it was at all obvious that the correct approach to proving tough questions about these PDEs was through probability. It was a tough journey they made. Now, of course, we realise that Markov diffusions are intrinsically more robust than their PDE counterparts and can be treated in a rigorous mathematical way across a wide context, including some (e.g. fractal environments and path spaces) where classical PDEs have little or no meaning.

**5.2. First-Order PDEs and Large Deviations.** It is a tautology to say that many of the most interesting features of our world and in our own personal experience evolve with time. Some evolve in a clear and predictable way. Others evolve in a less predictable and more random way. Some settle quickly to a large scale equilibrium, where the location and nature of that equilibrium is determined on a microscale by a steady incidence of rare events or large deviations. As a discipline, Stochastic Analysis is concerned with providing mathematical tools that are flexible enough to describe and study such systems, and at the same time concrete enough that they give real insight and lead one to address the real difficulties inherent in modelling them.

I would like to spend a little time discussing a beautiful paper [16] from 1966 where Varadhan is already using probabilistic methods to deeply understand a problem in non-linear first-order PDEs. The paper is fascinating for the way that, drawing on previous work of Donsker and Schilder, it forms a clear precursor result for the Large Deviations theory that would come later, and for the ease with which it brings together analytic tools that are familiar in finite-dimensional contexts (saddle point and steepest descent methods) and uses them in an infinite-dimensional setting, side-by-side with these emerging tools of Large Deviations, to create a decisive result.

Donsker had shown that the solution $u_\varepsilon$ to the initial value problem

$$u_t = uu_x + \frac{1}{2}\varepsilon^2 u_{xx} + p(t,x)$$

converges, as $\varepsilon \to 0$, to the solution to

$$u_t = uu_x + p(t,x).$$

A key step was the Hopf transformation

$$V(x) := \frac{1}{C} \exp \int^x \frac{-u(y)}{2\kappa} dy$$

which results in $V$ satisfying a third order PDE that factorises to the extent that it is obvious that if $V$ satisfies the linear equation

$$V_t = \frac{1}{2}\varepsilon^2 V_{xx} + \frac{1}{\varepsilon^2} p(t,x)V$$

then $u$ satisfies the original equation. Donsker understood that (at least for each fixed $\varepsilon > 0$) this linear equation has an easy probabilistic interpretation as an exponential path integral. Meanwhile, his student Schilder had, in his remarkable thesis, developed Laplace's method into a rigorous asymptotic expansion for certain families of exponential path integrals (M. Schilder, Some asymptotic formulas for Wiener integrals, Trans. Amer. Math. Soc. 125 (1966), 63–85).

Varadhan took these results as his backcloth and put them together in a beautiful, clean and more abstract way, with the now famous formulation where one establishes a lower bound on open sets and upper bound on closed sets.

Using this more concise overview and further demonstrating his technical power by proving the necessary bounds for this particular problem, Varadhan established a much more comprehensive result than his predecessors.

Varadhan's result in that paper was to show that if $f$ is convex then the solution to the initial value problem

$$u_t = [f(u)]_x + p(t,x)$$

can be easily constructed or approximated by successive solutions to linear problems.

The paper is remarkable and influential. Laplace's method has been used in the study of tail behaviour for iid sums since Cramer, and now it was a nontrivial tool that could be effectively used in infinite-dimensional settings—and even applied to give insight into the behaviour and origin of a class of nonlinear PDEs.

The notes from Courant [19] set out the importance of using exponential martingales to study the behaviour of general diffusion processes. This is an idea that is obviously at the heart of the work on the martingale characterization of diffusion processes and its applications to the theory of PDEs with continuous coefficients—one of his major works. Meanwhile [20] is an early example of solving a PDE with homogenization of the coefficients.

**5.3. PDEs and Probability in the 60s.** At the time of this early work described in this section of my article, I do not think it was at all obvious that the correct approach to answering outstanding questions about these PDEs was through probability; there were other pioneers for sure: Itô and McKean for example. But I am sure it was a tough journey they all made and, as I have mentioned, the results and methodologies stood the test of time.

Forty years later, as we mentioned above, we now understand that Markov diffusions and these methodologies are intrinsically more robust and general than their PDE counterparts and can be treated and exploited in a rigorous mathematical way across a wide context.

The distinction between the large deviations of the two Markov processes with the same co-variance introduced in Sect. 4 vanishes in the diffusion/PDE limit, although it might be critically important to financial regulators.

Viewing macroscopic physical quantities as local integrals of functionals defined on spaces of paths is a deep and powerful concept.

Diffusions can exist in fractal like environments where there are no local charts and classical PDEs have no meaning. They are often given meaning through a theory of Dirichlet forms and weak formulations. Pollutants evolving though soil might be modelled by such methods. At the same time these methods again allow rigorous constructions of diffusions on nonlinear path spaces and lead to rigorous treatment of such things as operators and Sobolev structures on these spaces for the first time.

## 6 The Support Theorem—Understanding the Itô Differential Equation

**6.1. On Extending the Wong–Zakai Theorem.** One of the key results in [12] and [13] associates a unique diffusion process to any PDE of type (5.1), providing that $a$ is strictly positive as a quadratic form and continuous and $b$ is bounded and measurable.

If Stroock and Varadhan had assumed their equations had Lipschitz continuous coefficients then (as was remarked in [12]) Itô's theory would have given them all the information they would have needed.

A vector field $V$ on $\mathbb{R}^d$ defines, for each $x \in \mathbb{R}^d$, a vector or direction $V(x) = (v_1(x), \ldots, v_d(x))$ in $\mathbb{R}^d$. Vector fields are useful for describing how systems evolve. For example, one might look for paths $t \to \gamma_t \in \mathbb{R}^d$ that have velocity that matches the vector field or, in more mathematical language, solve the differential equation

$$\dot{\gamma}_t = V(\gamma_t).$$

For Lipschitz vector fields there is always a unique solution with given starting point. In the example $V = (-y, x) \in \mathbb{R}^2$ the trajectories are circles centred at the origin. As we vary the initial state $\gamma_0$, the map from initial state to state at time $s$ defines an invertible map or flow $\pi_s$ by

$$\pi_s(\gamma_0) := \gamma_s.$$

In our example it rotates the plane by an angle $s$.

Now suppose that one places more interest on the interactions of the system $\gamma$ with other evolving systems $t$ and $\tau$. There are two very simple interactions, the first is replication where every increment of $\tau$ is replicated by an increment of $\gamma$

$$d\gamma_t = d\tau_t$$

and the second is the ordinary differential equation reflecting the effect of a vector field, which we now write as

$$d\gamma_t = V(\gamma_t)dt$$

for consistency.

A much richer range of interactions becomes immediately available if we mix these approaches. Suppose that $V^0, \dots, V^n$ are a family of vector fields on $\mathbb{R}^d$ and that $\tau = (\tau_1, \dots, \tau_n)$ is a (smooth enough) path in $\mathbb{R}^n$ then consider

$$d\gamma_t = \sum_{j=0}^{n} V^j(\gamma_t)d\tau_{j,t}$$

or

$$d\gamma_t = V(\gamma_t)d\tau_t \tag{6.1}$$

for short. It is helpful to view $\gamma$ as the response of the system to the control $\tau$; the map $\tau \mapsto \gamma$ is today known as the Itô map.

This equation includes the previous two via a judicious choice of $\tau$ and $V$. By adjoining variables the framework also includes the apparently more general case

$$d\gamma_t = V(\gamma_t, \tau_t)d\tau_t$$

through replicating the variable $\tau$ to $\tilde{\tau}$ and setting $\tilde{\gamma} = (\gamma, \tilde{\tau})$:

$$\begin{aligned} d\tilde{\tau}_t &= d\tau_t, \quad \tilde{\tau}_0 = \tau_0 \\ d\gamma_t &= V(\gamma_t, \tilde{\tau}_t)d\tau_t \end{aligned}$$

is a differential equation of the prescribed type. In this way, geometers might note that one class of such systems is formed by the connections on a manifold. In this case, the path on the manifold is the control and the horizontal lift to the bundle is the response. Another class of examples appears in Cartan development—in this case the control is naturally the path in the Lie algebra and the response is the path developed onto the Lie group.

In classical control theory the language is usually a little different. One can affect the evolution of a system by applying a control $\sum c_j(t)V^j$, mixing the different effects and considering the response

$$d\gamma_t = \sum_{j=0}^{n} c_j(t)V^j(\gamma_t)dt. \tag{6.2}$$

In optimal control, the discussion usually constrains $(c_j(t))$ to lie in some fixed compact convex set $K$ and optimises the choice to maximise some coordinate of the solution at the terminal time. By introducing

$$C_t = \left(\int_0^t c_1(s)ds, \dots, \int_0^t c_n(s)ds\right)$$

one sees that the classical control theory problem is also easily articulated in this language

$$d\gamma_t = V(\gamma_t)dC_t$$

where the standard optimal control problem can then be seen to maximise some coordinate of the solution (the benefit) at the terminal time subject to a constraint (the cost) on the length of the path $C_t$ in an appropriate semi-norm and parametrises this control at unit speed.

To make sense of (6.1) requires some regularity between $V$ and $\tau$. If $V$ is Lipschitz and so is $t \to \tau_t$, then, by rewriting the equation in a form similar to (6.2), one can use the classical theory of ordinary differential equations to deduce the existence of unique solutions to the initial value problem and the existence of flows and so on. However, as the replication example makes clear, (6.1) should have meaning in some contexts where $\tau$ is not differentiable and there are no $c_j(t)$. It was the remarkable achievement of Itô, now taught in most major mathematics and engineering schools world-wide, to give this equation meaning as $\gamma$ ranges over almost every Brownian or semi-martingale path. Because Brownian motion $W_t \in \mathbb{R}^n$ has independent identically distributed increments with the correct homogeneity, it is at least intuitively obvious that the solutions to the Itô–Stratonovich equation

$$dX_t = V(X_t) \circ dW_t + V^0(X_t)dt, \quad X_0 = a$$

would evolve in a way that was independent of their history, given their current position. That is to say $X_t$ is a Markov diffusion process. There are subtleties to get the exact theorems, but broadly it is easy to see from Itô's theory that if

$$Lu = \left(\sum_j v_k^j \frac{\partial}{\partial x_k} v_l^j \frac{\partial}{\partial x_l} + v_k^0 \frac{\partial}{\partial x_k}\right)u \tag{6.3}$$

and $f(x)$ is a bounded continuous function, then starting $X$ at $x$ at time $t$ and allowing $X$ to run until time $T$, and then evaluating $f$ at that location and averaging:

$$u(x,t) = \mathbb{E}[f(X_T) \mid X_t = x]$$

gives the solution to the boundary value problem

$$\frac{\partial u}{\partial t} = Lu$$
$$u(x,T) = f(x)$$

so expressing the solution $u$ as an integral of $f$ (Itô functional) over Wiener Space. One of the striking points is that the equation did not have to be elliptic or satisfy Hörmander's condition. As the example

$$V = (1, 0, \dots, 0)$$
$$L = \frac{\partial^2}{\partial x_1 \partial x_1}$$

shows clearly, even though $X$ diffuses it is not forced into any sort of spreading out in all directions; the diffusion, if it starts at a point, might well stay in a sub-manifold going through the starting point (a diffusive Hamiltonian system might still conserve momentum or energy). This explains the difficulty that one has in the subelliptic case in interpreting (6.3) in any naive way. If the process stays on a submanifold, e.g. on a circle centred on the origin, then the solutions will have no natural smoothness as one traverses from one sphere to adjacent ones. In general the curved nature of these surfaces means that $\frac{\partial^2}{\partial x_i \partial x_j} u$ will have no meaning, even though $u$ will—under reasonable hypotheses—be smooth on the circle. The fundamental solutions to (5.1) are evolving measures that, in general, do not have a density to Lebesgue measure that can be differentiated twice.

It is obviously an utterly basic question to understand, in terms of equation (5.1), the dependency of the solution at a point $(x_0, t_0)$ on this boundary data and identify the conservation laws. In probabilistic language: which sets support and contain the evolution? how does the evolution leave the domain?

"The support theorem" is the decisive answer, and today extends even to stochastic PDEs. It was motivated not only by the desire to extend Nirenberg's maximum principle to elliptic $L$ with continuous coefficients, but also to understand how to generalise Nirenberg's result to the degenerate setting where $a$ is only non-negative. The solution is a story of paths and control theory. The solution, in [15] and [14], is both intuitive, radical and technically deep. It is a story of a fundamental kind about approximation of solutions to stochastic differential equations by ordinary differential equations. The entire long last chapter of Ikeda and Watanabe's book (another ground breaking contribution to Stochastic Analysis) is dedicated to giving a second proof.

The ramifications of the methods have implications that go far beyond the fundamental problem solved. For example, the ideas underpin the standard construction of stochastic flows—a crucial concept for Bismut in [3] who gave a new proof of the index theorem.

**One-Dimension of Rough Control.** Today, the first steps in understanding this work come from two later and, in some sense, deeply misleading papers, by Doss and Sussman. Consider the controlled differential equation

$$\begin{aligned} d\tau_t &= V(\tau_t) d\gamma_t + V_0(\tau_t) dt \\ \tau_0 &= a \end{aligned} \tag{6.4}$$

where $\gamma$ is a smooth *real* (i.e., in $\mathbb{R}^n$, $n = 1$) valued path and $V$, $V_0$ are Lipschitz vector fields, then the path $\tau$ is uniquely defined using classical tools. Independently, each showed that the map $\gamma \mapsto \tau$ extends to all continuous $\gamma$ and *is continuous in $\gamma$ in the uniform topology*. It is clearly the unique extension with this property. We can thus give meaning to (6.4) for any continuous real valued path $\gamma$. In the case where $\gamma$ is a realisation of a 1-dimensional Brownian path, this definition almost surely coincides with the Itô–Stratonovich solution.

Now, if $\gamma_t \in \mathbb{R}$ is a continuous path for $t \in [0, 1]$, and $\gamma_t^{(n)}$ is its dyadic piecewise linear approximation (i.e. $\gamma_t^{(n)} = \gamma_t$ if $t \in 2^{-n}\mathbb{Z} \cap [0, 1]$, $\gamma^{(n)}$is continuous on $[0, 1]$ and linear on each interval in $[0, 1]\backslash 2^{-n}\mathbb{Z}$) then $\gamma^{(n)} \to \gamma$ uniformly. In particular, no matter how rough $\gamma$, if one defines $\tau^{(n)}$ to be the solution of (6.4) obtained by replacing the control $\gamma$ with its dyadic piecewise linear approximation $\gamma^{(n)}$ and solving the classical equation, then the $\tau^{(n)}$ converge to $\tau$. This gives a strong intuitive interpretation of (6.4) for rough one-dimensional $\gamma$. Restricting to the (Itô–Stratonovich) case where $\gamma$ is a randomly chosen Brownian path, one recovers the earlier result by Wong and Zakai that gave insight into the meaning and support of one-dimensional SDEs (stochastic differential equations).

**Multiple Dimensions of Rough Control.** The results of Doss and Sussman are misleading—not because they are wrong, but because something rather miraculous happens if $\gamma$ takes its values in a one-dimensional space. In the multi-dimensional case it is simply and spectacularly false. The Itô functional $\gamma \mapsto \tau$

$$d\tau_t = \sum_i V^i(\tau_t)d\gamma_t^i + V_0(\tau_t)dt, \quad \tau_0 = a$$

is not at all continuous in $\gamma$ in the uniform norm if the vector fields $V^i$ do not commute. It is not even closable in the space of continuous paths.

To get their support theorem, Stroock and Varadhan proved two conceptually fundamental results.

Stroock and Varadhan proved that if one approximates a Brownian path $X_t$ through its dyadic piecewise linear approximations $\gamma^{(n)}$ and solves the resulting classical ordinary differential equations then, almost surely, the $\tau_t^{(n)}$ again converge in law to the Itô–Stratonovich solution to the SDE driven by $X$. In this way they proved that the closed support of the law of $\tau_t$ is contained in the closure of the points where $\tau$ could reach when controlled by a piecewise smooth path $\gamma$.

Stroock and Varadhan also proved that if $\gamma$ was smooth, and if one conditions a Brownian path to be close to it in the uniform norm, then the solutions to the stochastic equation driven by the conditioned process converge to the solutions to the deterministic equation. From here one can see that the closed support of the $\tau_t$ controlled by piecewise smooth $\gamma$ is not bigger than the stochastic support.

From a more modern perspective, this extension of the Wong–Zakai theorem[2] also follows from work of Sipilainen [11] who proved that the dyadic polygonal

[2] At around the same time, and independently, J.M.C. Clark also tackled this problem in his doctoral thesis.

approximations to a Brownian path converge almost surely in the rough path metric.

The extension to Wong–Zakai, and the weak continuity theorem are far, far deeper results than the analogous one dimensional results. The difficulty of these results can be gauged by the fact that the statement that a path be close to $\gamma$ in the uniform norm on [0, 1] implicitly depended on a notion of distance in $\mathbb{R}^n$. Brownian motion conditioned to lie within $\varepsilon$ of $\gamma$ for one norm on $\mathbb{R}^n$ bears little statistical resemblance to the same process conditioned to lie within $\varepsilon$ of $\gamma$ for a second norm on $\mathbb{R}^n$, because these events correspond, are both rare and essentially distinct. The Stroock–Varadhan argument held for the box norm; later, Ikeda and Watanabe used the last chapter of their book to prove a similar result for the spherical norm. More recent work, presented in St Flour, allows the same result for any norm but depends on many more recent and deep developments, such as the boundary Harnack principle of Burdzy and Bass.

The conceptual content of these results is easy to take for granted today because it is so natural that the results should be true; they are basic and absorbed into our folklore. But they are also difficult. Had they not been true, then of course the whole subject would have developed differently. As an illustration of how the techniques go further than the basic results, we briefly discuss the existence of flows. Flows are solutions defined for every starting point $a$. Because there are uncountably many $a$, the Itô approach does not in itself allow such simultaneous solution. On the other hand the approximations to the solution using piecewise linear paths automatically provide flows for each approximation, and getting this sequence of diffeomorphisms to converge is possible and easier than trying to obtain the flow in other ways directly. Such flows were important to Bismut in his ground-breaking work on the Index theorem.

## 7 The Donsker–Varadhan Theory of Large Deviations

Introducing new mathematical abstraction can be dangerous and sometimes even an intellectual black hole. But it is also the life blood that allows mathematics to move forward and contribute to the wider world. At its most positive, one could think of "zero"—who needs a name for nothing? Closer to home, the introduction of filtrations of sigma algebras (adapted from measure theory) as a way of describing an evolving but partial knowledge of a system seems so dry that it could not have application.

My own score chart is positive about abstraction when it leads to simplification of arguments, and the extension of existing results to a broader class of significant examples. I get very excited when the approach leads to a completely new application. Successful innovations distinguish themselves from the black holes because they tend to be accompanied by an equally serious new approach which, although naturally proved in this level of abstraction, applies widely in settings where the abstract theory that justifies it is of little consequence. Leonardo Pisano brought us zero, but he also gave us Liber Abaci, brought the Hindu-Arabic algorithms of arithmetic to Europe, demonstrated the value of positional notation in denoting numbers, and made a non-trivial contribution to the development of (Merchant) Banking. Kolmogorov set out the axioms of probability in terms of measure theory, but immediately considered filtrations (sequences of nested $\sigma$-algebras), gave us a proper definition of independence and conditional expectation, and proved the strong law of large numbers in this context. The ideas are very abstract—it took me weeks to see any connection between the definitions and the intuitive concepts when I first came across the ideas. But the solid foundations it gave to probability, and the new theorems that came out of it, were a fundamental prerequisite to the massive development of Stochastic Analysis and probability as the toolset we have today.

In developing the theory of Large Deviations, Varadhan made a landmark achievement—he certainly proved hard theorems, demonstrating the value of the framework. Today, whether one is explaining the superconductivity associated to Josephson junctions, modelling stochastic resonance, managing communication networks, understanding the spectra of the large random matrices that occur in the analysis of mobile wireless communication channels, or understanding random surfaces, one gains substantial insight from Large Deviation theory. Rare events can be the crucial element in understanding the evolution of high dimensional (they happen) non-linear (their effects persist) systems.

**7.1. Short Time Behaviour—Geodesics.** Large Deviation theory provides the language and the basic results needed to describe and analyse these types of events, and even to model the systems when they are conditioned to have exceptional behaviour. In the papers [17] and [18] one can see the picture of large deviations theory being shaped as he proves a hard theorem. As before, Varadhan is proving theorems about non-constant coefficient elliptic and parabolic PDEs under relatively weak conditions on the coefficients. For example, consider $L$ where $a$ is Hölder and uniformly elliptic:

$$Lu = \sum_{j,k} a_{jk} \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_k} u.$$

Then one can construct (Miranda, Friedman) a unique minimal solution on $\mathbb{R}^d$:

$$\frac{\partial}{\partial t} p_\varepsilon = \frac{1}{2} L p_\varepsilon,$$

which is positive $p_t(x, y)$ on $D$ and has boundary data $\delta_x$ at $t = 0$.

In the customary way at that time, (Stroock and Varadhan came later) this is associated to a semi-group and a unique diffusion. Suppose that $D$ is open and that all boundary points are externally accessible; then we might be interested in the distribution of the time $T_D$ as the associated diffusion $X_t$, started at $x$, hits $\partial D$. From the Itô perspective it is easy to see that the Laplace transform of this random time solves the eigenfunction problem

$$\phi(x,\lambda) := \mathbb{E}_x[e^{\lambda T}]$$

$$L\phi = \lambda\phi, \quad \phi|_{\partial D} = 1.$$

Varadhan gives a bare handed proof that

$$\lim_{\lambda\to\infty}\left[-\frac{1}{\sqrt{2\lambda}}\log\phi(x,\lambda)\right] = d(x,\partial D)$$

and uses this to prove that the fundamental solution $p(t,x,y)$ has the property

$$\lim_{t\searrow 0} -2t\log p(t,x,y) = d(x,y)^2$$

where the convergence is uniform on bounded regions. It is only for solutions defined on the whole of $\mathbb{R}^d$. This result is obvious if the coefficients are constant, but far deeper in this generality. It is the core technical bound that, when combined with the Large Deviations framework, gives a series of beautiful results in the second paper and provides a rich immediate justification for the framework.

The rigorous probability starts here. It was well known by the time the paper was written that the conditions on the operator $L$ were strong enough to associate to it a unique strong Markov diffusion $(X_t)_{t\in[0,T]}$ whose Feller transition semigroup was $p_t(x,y)$. We can use this reference process to define a family of measures on $C[0,T]$ by rescaling time

$$\mathbb{P}_x^\varepsilon(A) = \mathbb{P}_x(X_{\varepsilon t}|_{t\in[0,T]} \in A).$$

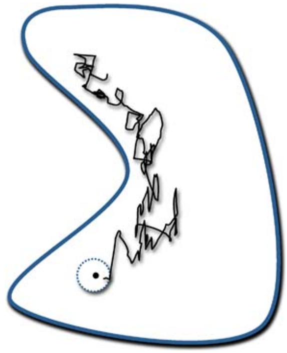
We can fix $A$ and consider in general the behaviour as $\varepsilon \to 0$ of $\mathbb{P}_x^\varepsilon(A)$. Note that $A$ is a generic set of paths and so this is truly an infinite-dimensional question. If the set of paths $A$ is separated, in the uniform metric, from the path that stays at $x$ for all time, then the probability that the reparameterised path $X_{\varepsilon t}|_{t\in[0,T]}$ started at $x$ is in $A$ becomes increasingly unlikely as $\varepsilon \to 0$. In fact one can often get a really quite precise asymptotic picture. Suppose that $x \in D$, an open set in $\mathbb{R}^d$, and $A_D$ is the set of paths $\gamma$ in $C([0,T],\mathbb{R}^d)$ that remain in $D$ until at least time $T$ and which, at time $T$, are in a ball of radius $\delta > 0$ centred on $y \neq x$

$$A_D = \{\gamma \mid \gamma([0,T]) \subset D,\ d(\gamma(T),y) < \delta\}.$$

The set $A_D$ is open in the uniform topology and, because $y \neq x$, is separated from the trivial path at $x$. It is probably best to give an example.

Can one be more precise? For example, how does the probability change if we change the shape of the region $D$ (in particular if we let it be the whole of $\mathbb{R}^d$)? How does it depend on the diffusion coefficients $a$ that determine $L$ and $X$?

We have already mentioned that, although rare events are rare, if there are several of them then it is very unusual that they are equally rare. There is an intuition that, if we aggregate a family of rare events $A_i$ into a single rare event $\cup A_i$, one of the rare



events in the aggregate will be much less rare than the others, and so dominate and determine the rareness of the aggregate. This would certainly be true if $\mathbb{P}_\varepsilon(A_i) = r_i e^{s_i/\varepsilon}$; if $r_j > 0$ and the $s_j$ are distinct then

$$\max_{j<N} \lim_{\varepsilon\to 0} \frac{\sum_{i<N} r_i e^{s_i/\varepsilon}}{r_j e^{s_j/\varepsilon}} = 1.$$

At the heart of any Large Deviations argument one has to prove that an appropriate extension of this relationship holds in a setting where the $i$ range over a continuum instead of a finite set.

Varadhan introduced a functional on paths $\gamma$ that would determine their local rareness for the diffusion process over short times. It now has a natural feel to it as the energy of the path

$$I(\gamma) := \frac{1}{2} \int_0^T \dot{\gamma}_t a(\gamma_t)^{-1} \dot{\gamma}_t dt$$

(or $+\infty$ where $\dot{\gamma}$ is not defined almost everywhere) determined by the diffusion coefficient $a$. He could then use his estimate of the probability of the reparameterised process being close to $\gamma$ to prove the following result.

**Theorem** *Let $G \subset C([0,T],\mathbb{R}^d)$ be open and $C \subset C([0,T],\mathbb{R}^d)$ be closed. Then*

$$\limsup_{\varepsilon\to 0} \varepsilon \log \mathbb{P}_x^\varepsilon[C] \leq - \inf_{\gamma\in C} I(\gamma),$$

$$\liminf_{\varepsilon\to 0} \varepsilon \log \mathbb{P}_x^\varepsilon[G] \geq - \inf_{\gamma\in C} I(\gamma).$$

In this way Varadhan established asymptotic lower bounds on the probability that a random path $X_{\varepsilon t}$, run over the interval $[0,T]$, would be in a given open set of paths. At the same time he had an asymptotic upper bound on the equivalent inclusion for closed sets of paths. Varadhan also allowed the sets $G$ and $C$, on the left hand side, to vary.

We can now look again at the example of the paths that stay in the domain $D$. As $\varepsilon \to 0$, almost all paths from $x$ to $y$ will stay close to a minimising geodesic if there is one. There is an obvious phase transition according to whether $A_D$ contains the shortest path (in the $a$ metric) from $x$ to the ball centred on $y$. In our picture it does not.

Varadhan made this all precise and, in a beautiful application, related the fundamental solutions $p_D$, $p$ to the PDE restricted to the domain and on the whole of $\mathbb{R}^d$. An immediate consequence of Theorem 4.9 in that paper is that if $p_D$ is the fundamental solution for the equation restricted to a domain $D$ with Dirichlet boundary conditions then

$$\lim_{t\searrow 0}\frac{p(t,x,y)}{p_D(t,x,y)}=1$$

for every pair $(x,y)\in D\times D$ if and only if $D$ is convex in the Riemannian sense for the metric $a^{-1}$. This is a very fine remark which is considerably more delicate than the basic large deviations result above. It symbolises how, even in 1967, Varadhan had very serious examples to test the technology and perfect the theory of Large Deviations as well as beautiful applications.[3]

**7.2. Long Time Behaviour—Occupation Measures.** Suppose a visitor to a finite neighbourhood is forced by his job to move around the neighbourhood in a Markovian way. As he moves around he interacts with the local inhabitants and would like to make friends. But they are cautious and only make friends if the visitor spends a considerable amount of time at their site and gives them significantly more than average attention. On average the visitor will spend equal time at each site, but even over long time intervals then time will not exactly equilibrate. How many friends can the visitor expect to make? How many good friends?

Mathematically, the problem is to look at a recurrent diffusion over long time periods. For example $X$ might be a Brownian motion on a compact multi-dimensional manifold. Then, from the ergodic theorem, one knows that for almost every $\omega$ the empirical occupation measure $\mu_{(T,\omega)}$ defined by

$$\mu_{(T,\omega)}(f):=\frac{1}{T}\int_0^T f(X_t(\omega))dt$$

will converge to the normalised volume measure. But in general it will never get there. One would like to describe the fluctuations and exceptional behaviour of this random(ly evolving) measure.

I suppose (but do not know) that the theory really became confirmed when, with Donsker [4, 5, 7, 8], Varadhan looked at this second intrinsically important and

[3] As matter of history, Varadhan was coming back after a seminar at Rockefeller University in a taxi with colleagues when someone remarked about a theorem of Cieselski that says the ratio $p_G(t,x,y)$ of the Dirichlet fundamental solution of Brownian Motion to the Gaussian $p(t,x,y)$ tends to 1 for all $(x,y)\in G$ if and only if $G$ is convex (modulo sets of capacity 0). Varadhan wondered what was the Diffusion analog. Varadhan tells the author "I thought the answer should be that diffusions if they go some place in a short time interval will go along geodesics. That was what led me to the short time asymptotics of diffusions."

clearly infinite-dimensional problem of convergence. Again, the dimensionality and the specificness of the problem ensured that there were hard issues to be resolved, but the Large Deviations framework pointed out the direction and focused the considerable effort required allowing much better results than a more ad hoc approach.

They identified a rate function

$$I(\mu) = \inf_{\substack{u>0 \\ u\in\mathcal{D}(L)}} \int_{x\in M} \frac{Lu}{u}(x)\mu(dx),$$

and established the Large Deviation principle, so providing a powerful family of estimates for exceptional behaviour of the occupation measure and its deviations from the invariant measure. This gives insight into what happens when one gets a large fluctuation and gives valuable information on the dynamics of the system conditional on this exceptional behaviour by minimising $I(\mu)$, subject to the constraint of the exceptional behaviour for $\mu$.

In other words, if we condition on our traveller making an unusually large number of friends, then we can expect and predict a lot of structure in how these contacts are built up. This work on occupation measures is highly applicable and recaptures Kac's use of exponential functionals to capture solutions to PDEs. As before there is always at least one basic concrete result that is needed to make it all work (no free lunch!). In this case one could cite [6], which proves under appropriate hypotheses that

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}_x\big[e^{-t\Phi(\hat{\mu})}\big] = -\inf_{\mu} \frac{1}{8}\left(\int \left|\nabla\left(\log\frac{d\mu}{d\nu}\right)\right|^2 d\mu + \Phi(\mu)\right).$$

This is determined using the diffusion metric, where $\nu$ is the invariant measure, $\hat{\mu}$ is the normalised general random occupation measure, and $\Phi$ is a lsc functional on probability measures with compact sets $\Phi \leq \lambda$.

The ramifications of the methodology established in these early papers have spread so widely that it would be impossible to do the whole subject any sort of credit and in any case others are better equipped to do this than this author.

## 8 Hydrodynamical Limits, Interacting Particles and Other Questions

We have seen that large scale non-linear phenomena can arise because of the influence of large deviations. The effect is exactly a balanced trade off between the rareness of the event, the magnitude of the event, and the influence it has on the system as a whole.

However, one needs caution. For the fact that large deviations matter throws doubt onto the framework in which they are studied, on the non-linear PDEs they justify, and on modelling in general. Most real world models have a granularity, but, like the simple random walk, sometimes there are central limit type theorems, and they can be modelled very effectively by Markov diffusion processes. However, as the introductory section in this article shows, the large deviations of the granular process may be very different to those of the limiting Markov diffusion. The macroscopic large deviations of the granular process are a function of the microstructure of the process. Microstructure that is lost on taking the diffusion limit. This is important, and also a challenge—one cannot interchange orders of limits without justification.

So, there is a very real challenge to take reasonable microscopic models of non-linear physical systems (such as fluids), systems where the non-linear and large deviation effects are likely to be highly relevant, and prove that the large scale behaviour can be successfully modelled (with PDEs or by other means). This is a major programme, proving there is a fluid limit for models with non-trivial microstructure, of current interest. Varadhan, with younger colleagues and students Kipnis, Olla (Rome), Quastel (Toronto) and H.T. Yau (Harvard) did as much as anyone to initiate this very important programme.

Almost by definition, the current state of understanding on the hydrodynamical limits and on particle systems has a huge amount of detail, and would require another long chapter to do it justice. Moreover, and sadly, I do not have enough expertise to summarise it accurately.

**8.1. Homogenised Hamilton–Jacobi Equations.** Sometimes homogenisation works, you can forget local structure, and central limit phenomena dominate! In this case, one might replace the rapidly fluctuating coefficients in the basic diffusion with a much more smoothly varying model, but still have a diffusion which, on the appropriate scales, was indistinguishable from the original. Varadhan had early work on this homogenising the coefficients of PDEs. He was also a co-author of the very influential unpublished—but widely circulated and referenced—contribution on homogenisation of the Hamilton–Jacobi equation, with Lions and Papanicolou which provided the weak viscosity solutions to the HJE equation that underpin recent work on the construction of Mather invariant sets, reconnecting PDEs and Lagrangian dynamics.

The impact of this work persists to the current time. KAM theory provides a suitable framework for understanding the dynamics of attracting sets for a system subject to a small periodic potential. Aubry and Mather, and then Mather alone, constructed invariant sets in the non-KAM non-perturbative framework. The work of Fathi and Evans, using weak viscosity solutions to the HJE developed in the seminal work with Lions and Papanicolou, constructs Mather invariants in the non-KAM non-perturbative framework.

## 9 Conclusion

The Fields Medals at Madrid illustrate how probability has come of age. Okounkov used random surfaces to connect work of Gromov and Witten with that of Donaldson and Thomas. Werner's work is motivated by statistical physics and links the evolving geometry of two-dimensional Brownian motion to conformal field theory. Tao used probabilistic motivations for many aspects of his plenary lecture.

It is inconceivable that Okounkov and Werner have not been substantially, even subconsciously, influenced by the work of Varadhan and his collaborators. Varadhan's work has influenced applications at the same level and his results and perspectives infuse over a wide horizon. Much of my work would have little interest without his contributions.

Probabilists around the world were delighted when Varadhan was awarded the Abel Prize; his deep and decisive contributions to this significant and young area of mathematics ensure that the Abel prize will get distinction from choosing Varadhan!

The author of this short article would like to offer his thanks to Professor Varadhan for his inspiration, to the editors for their patience, and to the reader for getting this far. At the same time as he hopes to have given a flavour of some of Varadhan's work, he is very conscious of, almost by the same token, the simplifications and omissions that have distorted the accuracy of the presentation; for this he asks the readers' forgiveness.

## References

1. Barlow, M.T.: Heat kernels and sets with fractal structure. Contemp. Math. **338**, 11–40 (2003)
2. Barlow, M.T., Bass, R.F., Kumagai, T., Teplyaev, A.: Uniqueness of Brownian motion on Sierpinski carpets. arXiv:0812.1802v1 (2008)
3. Bismut, J.-M.: Mécanique aléatoire. Lecture Notes in Mathematics, vol. 866. Springer, Berlin (1981). With an English summary
4. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time. I. Commun. Pure Appl. Math. **28**, 1–47 (1975)
5. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time. II. Commun. Pure Appl. Math. **28**, 279–301 (1975)
6. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Wiener integrals for large time. In: Functional Integration and Its Applications, Proc. Internat. Conf., London, 1974, pp. 15–33. Clarendon Press, Oxford (1975)
7. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time. III. Commun. Pure Appl. Math. **29**(4), 389–461 (1976)
8. Donsker, M.D., Varadhan, S.R.S.: Asymptotic evaluation of certain Markov process expectations for large time. IV. Commun. Pure Appl. Math. **36**(2), 183–212 (1983)
9. Nadirashvili, N.: Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators. Ann. Sc. Norm. Super. Pisa, Cl. Sci. (4) **24**(3), 537–549 (1997)
10. Nirenberg, L.: A strong maximum principle for parabolic equations. Commun. Pure Appl. Math. **6**, 167–177 (1953)
11. Sipilainen, E.-M.: A pathwise view of solutions of stochastic differential equations. PhD thesis, University of Edinburgh (1993)
12. Stroock, D.W., Varadhan, S.R.S.: Diffusion processes with continuous coefficients. I. Commun. Pure Appl. Math. **22**, 345–400 (1969)
13. Stroock, D.W., Varadhan, S.R.S.: Diffusion processes with continuous coefficients. II. Commun. Pure Appl. Math. **22**, 479–530 (1969)
14. Stroock, D.W., Varadhan, S.R.S.: Diffusion processes with boundary conditions. Commun. Pure Appl. Math. **24**, 147–225 (1971)

15. Stroock, D.W., Varadhan, S.R.S.: On the support of diffusion processes with applications to the strong maximum principle. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Univ. California, Berkeley, CA, 1970/1971. Probability Theory, vol. III, pp. 333–359. Univ. California Press, Berkeley (1972)
16. Varadhan, S.R.S.: Asymptotic probabilities and differential equations. Commun. Pure Appl. Math. **19**, 261–286 (1966)
17. Varadhan, S.R.S.: Diffusion processes in a small time interval. Commun. Pure Appl. Math. **20**, 659–685 (1967)
18. Varadhan, S.R.S.: On the behavior of the fundamental solution of the heat equation with variable coefficients. Commun. Pure Appl. Math. **20**, 431–455 (1967)
19. Varadhan, S.R.S.: Stochastic processes. Notes based on a course given at New York University during the year 1967/68. Courant Institute of Mathematical Sciences, New York University, New York (1968)
20. Varadhan, S.R.S.: Boundary value problems with rapidly oscillating random coefficients. In: Random Fields, Vols. I, II, Esztergom, 1979. Colloq. Math. Soc. János Bolyai, vol. 27, pp. 835–873. North-Holland, Amsterdam (1981)

# The Abel Committee

## 2003

Erling Størmer (University of Oslo, Norway), chair
John Ball (University of Oxford, UK)
Friedrich Hirzebruch (University of Bonn, Germany)
David Mumford (Brown University, USA)
Jacob Palis (National Institute for Pure and Applied Mathematics, Brazil)

## 2004

Erling Størmer (University of Oslo, Norway), chair
David Mumford (Brown University, USA)
Jacob Palis (National Institute for Pure and Applied Mathematics, Brazil)
Gilbert Strang (Massachusetts Institute of Technology, USA)
Don Zagier (Max Planck Institute for Mathematics, Germany)

## 2005

Erling Størmer (University of Oslo, Norway), chair
Ingrid Daubechies (Princeton University, USA)
László Lovász (Eötvös Loránd University, Hungary)
Gilbert Strang (Massachusetts Institute of Technology, USA)
Don Zagier (Max Planck Institute for Mathematics, Germany)

## 2006

Erling Størmer (University of Oslo, Norway), chair
Enrico Bombieri (Institute for Advanced Study, USA)
Ingrid Daubechies (Princeton University, USA)
László Lovász (Eötvös Loránd University, Hungary)
Claudio Procesi (University of Rome “La Sapienza”, Italy)

H. Holden, R. Piene (eds.), *The Abel Prize*,
DOI 10.1007/978-3-642-01373-7, © Springer-Verlag Berlin Heidelberg 2010

**2007**

Kristian Seip (Norwegian University of Science and Technology, Norway), chair
Enrico Bombieri (Institute for Advanced Study, USA)
Hans Föllmer (Humboldt University, Germany)
Dusa McDuff (Stony Brook University, USA)
Claudio Procesi (University of Rome “La Sapienza”, Italy)

# The Board for the Niels Henrik Abel Memorial Fund

### 2003

Jens Erik Fenstad (chair)
Elisabeth Grieg
Eivind Hiis Hauge
Idun Reiten
Arne B. Sletsjøe
Reidun Sirevåg (observer)

### 2004

Jens Erik Fenstad (chair)
Elisabeth Grieg
Eivind Hiis Hauge
Idun Reiten
Arne B. Sletsjøe
Reidun Sirevåg (observer)

### 2005

Ragnar Winther (chair)
Elisabeth Grieg
Eivind Hiis Hauge
Idun Reiten
Arne B. Sletsjøe
Reidun Sirevåg (observer)

### 2006

Ragnar Winther (chair)
Elisabeth Grieg
Eivind Hiis Hauge

Idun Reiten
Arne B. Sletsjøe
Reidun Sirevåg (observer)

### 2007

Ragnar Winther (chair)
Kari Gjetrang
Arne Bang Huseby
Idun Reiten
Leiv Storesletten
Reidun Sirevåg (observer)

Transcripts of parts of the interviews that Martin Raussen and Christian Skau made with each laureate in connection with the Prize ceremonies, can be found in the following publications:

**2003 Jean-Pierre Serre**

*EMS Newsletter*, issue 49 (Sep. 2003) 18–20,
*AMS Notices*, **51** (2004) 210–214,
*RMS—Revue de la Filière Mathématiques*, **114**, nr. 3 (2004) 3–9 (in French).

**2004 Sir Michael Atiyah and Isadore M. Singer**

*EMS Newsletter*, issue 53 (Sep. 2004) 24–30,
*AMS Notices*, **52** (2005) 223–231.

**2005 Peter D. Lax**

*EMS Newsletter*, issue 57 (Sep. 2005) 24–31,
*AMS Notices*, **53** (2006) 223–229.

**2006 Lennart Carleson**

*EMS Newsletter*, issue 61 (Sep. 2006) 31–36,
*AMS Notices*, **54** (2007) 223–229,
*Mathematical Advance in Translation*, **27**, nr. 1 (2008) 37–45 (in Chinese).

**2007 S.R. Srinivasa Varadhan**

*EMS Newsletter*, issue 65 (Sep. 2007) 33–40,
*AMS Notices*, **55** (2008) 238–246,
*Mathematical Advance in Translation*, **27**, nr. 2 (2008) 147–156 (in Chinese).