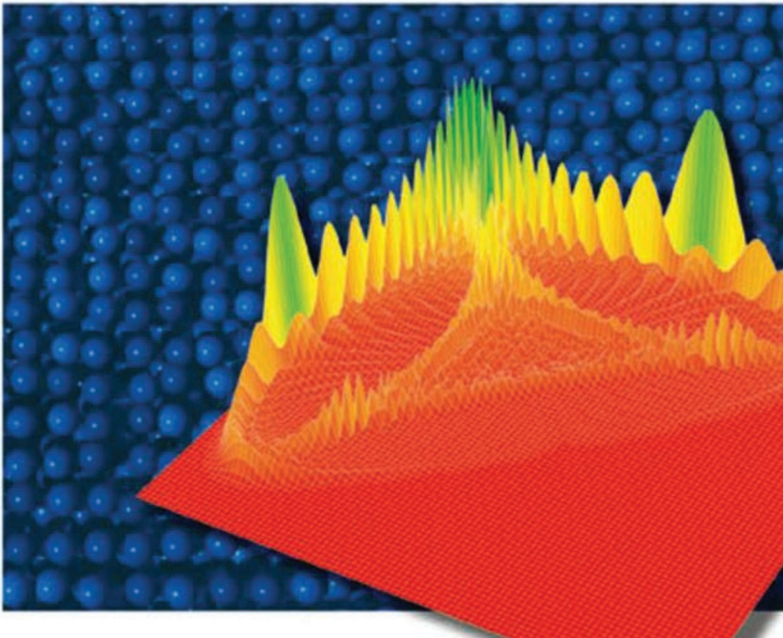
Edited by
W. Arendt and W. P. Schleich

WILEY-VCH

Mathematical Analysis of Evolution, Information, and Complexity



Mathematical Analysis of Evolution, Information, and Complexity

Edited by
Wolfgang Arendt and
Wolfgang P. Schleich

For additional information reagarding this topic, please refer also to the following publications

Bruß, D., Leuchs, G. (eds.)

Lectures on Quantum Information

2007
ISBN 978-3-527-40527-5

Audretsch, J. (ed.)

Entangled World
The Fascination of Quantum Information and Computation

2006
ISBN 978-3-527-40470-4

Stolze, J., Suter, D.

Quantum Computing
A Short Course from Theory to Experiment

2004
ISBN 978-3-527-40438-4

Mathematical Analysis of Evolution, Information, and Complexity

Edited by
Wolfgang Arendt and Wolfgang P. Schleich

WILEY-VCH Verlag GmbH & Co. KGaA

The Editors

Prof. Dr. Wolfgang Arendt
University of Ulm
Institute of Applied Analysis
wolfgang.arendt@uni-ulm.de

Prof. Dr. Wolfgang P. Schleich
University of UIm
Dept. of Quantum Physics
Wolfgang.Schleich@uni-ulm.de

Cover picture
Taken from: Bernd Mohring, Coherent
Manipulation and Transport of Matter Waves
Verlag Dr. Hut
ISBN-10: 3899634810
By cortesy of Bernd Mohring

All books published by Wiley-VCH are carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained in these books, including this book, to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de .

Typesetting le-tex publishing services oHG, Leipzig
Printing betz-druck GmbH, Darmstadt
Binding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN: 978-3-527-40830-6

Contents

Traveller, there are no paths.
Paths are made by walking.
Antonio Machado (1875–1939)

Preface

The present book is devoted to the mathematical analysis of evolution, information and complexity. The time evolution of systems or processes is a central question in science and covers a broad range of problems including diffusion processes, neural networks, quantum theory and cosmology. Analysis of information is needed in data compression, channel encoding, cryptography and often in the analysis of information processing in the computer or in the brain. Finally, the analysis of complexity is important for computer science, in particular algorithms, but more generally also for the investigation of complex and chaotic systems.

Since 2004 the University of Ulm has operated a graduate school dedicated to the field of *Mathematical Analysis of Evolution, Information and Complexity*. This program brings together scientists from the disciplines of mathematics, electrical engineering, computer science and physics. Our rather unique school addresses topics that need a unified and highly interdisciplinary approach. The common thread of these problems is mathematical analysis demonstrating once more the newly emerging notion of mathematics as technology.

Our book highlights some of the scientific achievements of our school and therefore bears its name *Mathematical Analysis of Evolution, Information and Complexity*. In order to introduce the reader to the subject we give elementary and thus accessible introductions to timely themes taken from different parts of science and technology such as information theory, neuro-informatics and mathematical physics.

Each article in the book was prepared by a team in which at least two different disciplines were represented. In this way mathematicians have collaborated on a chapter with physicists, or physicists have worked with electrical engineers and so on. Moreover, we have installed the rule that with every senior scientist there would be a graduate student working on this article. We hope that this rule has led to easily understandable contributions.

Mathematical Analysis of Evolution, Information and Complexity does not only represent the program of our school and has become the title of the book but has also served as the guiding principle for its organization. Indeed, we have chosen the three pillars “evolution”, “information” and “complexity” of the school as titles for the three parts of the book. For each one we have identified one or two major themes as shown in Table 0.1.

Table 0.1 Organization of the book outlining its three pillars with their themes. The number above each topic indicates the chapter.

Evolution		Information		Complexity
Spectral analysis	**Networks**	**Pattern recognition**	**Signal analysis**	**Algorithms**
1 Weyl's law	4 biological neural networks	7 speech recognition: remote access	12 Shannon's theorem	15 Shor algorithm
2 differential equations	5 gene regulation	8 speech recognition: machine learning	13 codes	16 quantum and classical algorithms
3 cosmology	6 quantum graphs	9 cluster analysis in genomics	14 signal processing in the brain	17 sorting algorithms
		10 image analysis in computer science and cosmology		
		11 data analysis and learning		

We have taken the liberty to assign each article to one of these themes. However, in many instances the contributions could have also been attributed to another theme. This feature is certainly a trade mark of an interdisciplinary article. These articles form the individual chapters of the book.

The topics addressed in each pillar range from quantum physics via bioinformatics to computer science and electrical engineering. The common element linking all chapters is mathematical analysis. To quote Galileo Galilei:

> *"Egli [il libro che è l'universo] è scritto in lingua matematica."*
> *("The book which is the universe is written in mathematical language.")*

In order to bring out most clearly the interconnections between the chapters of the book, we now briefly summarize the essential ideas of each contribution. We start with the pillar "evolution" consisting of the two themes of spectral analysis and networks.

Weyl's law describes the asymptotic distribution of the eigenvalues of the Laplacian. It has played an important role in the development of quantum theory. Chapter 1 gives a comprehensive summary of the history of Weyl's law, its generalization based on trace formulae, its application in quantum chaos, as well as a modern proof. A review and comparison of different methods of solving systems of linear ordinary differential equations is the topic of Chapter 2. Applications and extensions to partial differential equations such as the heat equation or the Schrödinger equation are given. The theme on evolution concludes in Chapter 3 with an introduction into general relativity with an alternative approach based on the scalar-tensor theory and the Higgs potential.

The theme of evolution in networks addresses biological neural networks, gene regulation and quantum graphs. For example, Chapter 4 provides an overview over models describing biological and computational neural networks. Here the central topic is the specific model developed in Ulm using neural populations as associative memories. Another example of a network of signalling compounds within the kernel of a cell is summarized in Chapter 5. The mathematical model of Boolean networks describes the gene regulation in living organisms. Quantum graphs, the topic of Chapter 6, represent yet another network. They are a toy model for a Schrödinger operator on a thin, quasi-one-dimensional network. The article studies the symmetries that emerge in such quantum networks.

A major portion of the book is dedicated to the mathematical analysis of information. Here the topics range from speech recognition via cluster analysis to signal processing in the brain. Usually speech recognition is implemented on powerful computers. In Chapter 7 tools are developed which allow remote access, for example, using cellular phones. Here, the Ulm technology of associative memories plays an important role. Spoken language dialogue systems are interactive, voice-based interfaces between humans and computers. They allow humans to carry out tasks of diverse complexity such as the reservation of tickets or the performance of bank transactions. Chapter 8 describes different appproaches for the categorization of caller utterances in the framescope of a technical support dialog system, with special focus on categorizers using small amounts of labeled examples. Functional genomics aim to provide links between genomic information and biological functions. One example is the connection between gene patterns and tumor status. Cluster analysis generates a structure of data solely exploring distances or similarities. A special feature of Chapter 9 is the demonstration that already sparse additional information leads to stable structures which are less susceptible to minor changes. Image analysis tries to detect complex structures in high-dimensional data. Chapter 10 compares and contrasts approaches developed in computer vision and cosmology. We conclude the theme of pattern recognition by discussing in Chapter 11 the fundamental method of classification in data analysis. Unfortunately, the true concept of classification is often not known. A method of combining several such concepts, called boosting, which leads to highly accurate classifiers, is described here.

Another important theme in the part on information is represented by signal analysis covering the topics of Shannon's sampling theorem, codes and signal processing in the brain. The sampling theorem shows how a signal can be reproduced by a finite number of measurements. Chapter 12 gives a historical overview and provides two proofs. Coding theory tells us how, by adding redundancy and correcting errors, information can be transmitted accurately. An overview of codes with emphasis on algebraic geometric codes is given in Chapter 13. This section concludes with Chapter 14 describing a model of how the human cortex processes its sensor signals. Mathematically this model consists of a system of coupled nonlinear ordinary differential equations whose long time behavior is discussed.

The last part addresses the topic of complexity focusing on classical as well as quantum algorithms. A central task in computer science is to find efficient algorithms. Chapter 15 lays the foundation of this chapter by explaining the famous Shor algorithm to factor numbers from a physics point of view. In the same vein Chapter 16 describes the state of the art of two famous problems, integer factorization and the graph isomorphism problem. It points out similarities and differences between these two problems when approached by classical or quantum computing. The QuickSort algorithm is a most efficient sorting method. It relies on dividing the sequence into parts of smaller lengths. In Chapter 17 the complexity of the method is studied with a special emphasis on varying the random source which governs the division.

We would like to take the opportunity to thank the authors for their contributions, enthusiasm and reliability in producing this volume. Moreover, we are most grateful to Robin Nittka for his competent help in putting this book together. Finally, we appreciate the support of the Ministerium für Wissenschaft, Forschung und Kunst, Baden-Württemberg in the framework of the Promotionskolleg *Mathematical Analysis of Evolution, Information and Complexity*.

Ulm, August 2008

Wolfgang Arendt
Wolfgang Schleich

List of Contributors

Amparo Albalate

University of Ulm
Institute of Information Technology
89069 Ulm
Germany
amparo.albalate@uni-ulm.de

Wolfgang Arendt

University of Ulm
Institute of Applied Analysis
89069 Ulm
Germany
wolfgang.arendt@uni-ulm.de

Werner Balser

University of Ulm
Institute of Applied Analysis
89069 Ulm
Germany
werner.balser@uni-ulm.de

Nils Bezares-Roder

University of Ulm
Institute of Theoretical Physics
89069 Ulm
Germany
nils.bezares@uni-ulm.de

Jens Bolte

Department of Mathematics
Royal Holloway College,
University of London
Egham, TW20 0EX,
United Kingdom
jens.bolte@rhul.ac.uk

Martin Bossert

University of Ulm
Institute of Telecommunications
and Applied Information Theory
89069 Ulm
Germany
martin.bossert@uni-ulm.de

Stefano Cardanobile

Bernstein Center for Computational Neuroscience
Albert-Ludwigs-Universität
79104 Freiburg i. Br.
Germany
stefano.cardanobile@bccn.uni-freiburg.de

Michal Chovanec

University of Ulm, Helmholtzstraße 18
Institute of Applied Analysis
89069 Ulm
Germany
michal.chovanec@uni-ulm.de

Michael Cohen

Boston University
Department of Cognitive and Neural Systems
677 Beacon Street, Boston, MA 02215
USA
mike@cns.bu.edu

Silvia Corchs

University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
silvia.corchs@uni-ulm.de

Sebastian Dörn
University of Ulm
Institute of Theoretical Computer Science
89069 Ulm
Germany
sebastian.doern@uni-ulm.de

Daniel Haase
University of Ulm
Institute of Number Theory and Probability Theory
89069 Ulm
Germany
daniel.haase@uni-ulm.de

Holger Stefan Janzer
University of Ulm
Institute of Theoretical Physics
89069 Ulm
Germany
holger.janzer@uni-ulm.de

Zöhre Kara Kayikci
University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
zoehre.kara@uni-ulm.de

Hans A. Kestler
University of Ulm
Clinic for Internal Medicine I
and Institute of Neural Information Processing
89069 Ulm
Germany
hans.kestler@uni-ulm.de

Johann Kraus
University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
johann.kraus@uni-ulm.de

Michael Kühl
University of Ulm
Institute of Biochemistry
89069 Ulm
Germany
michael.kuehl@uni-ulm.de

Ludwig Lausser
University of Ulm
Institute of Neural Information Processing,
Internal Medicine I
89069 Ulm
Germany
ludwig.lausser@uni-ulm.de

Jürgen Lindner
University of Ulm
Institute of Information Technology
89069 Ulm
Germany
juergen.lindner@uni-ulm.de

Beatrice List
LS telcom AG
Im Gewerbegebiet 31–33
77839 Lichtenau
Germany
beatrice.list@gmx.de

Werner Lütkebohmert
University of Ulm
Institute of Pure Mathematics
89069 Ulm
Germany
werner.luetkebohmert@uni-ulm.de

Rüdiger Mack
University of Ulm
Institute of Quantum Physics
89069 Ulm
Germany
ruediger.mack@uni-ulm.de

Helmut Maier
University of Ulm
Institute of Number Theory and Probability Theory
89069 Ulm
Germany
helmut.maier@uni-ulm.de

Jörg Marhenke
University of Ulm
Institute of Pure Mathematics
89069 Ulm
Germany
joerg.marhenke@uni-ulm.de

Heiner Markert

University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
heiner.markert@uni-ulm.de

Markus Maucher

University of Ulm
Institute of Theoretical Computer Science
89069 Ulm
Germany
markus.maucher@uni-ulm.de

Wolfgang Minker

University of Ulm
Institute of Information Technology
89069 Ulm
Germany
wolfgang.minker@uni-ulm.de

Delio Mugnolo

University of Ulm, Helmholtzstraße 18
Institute of Applied Analysis
89069 Ulm
Germany
delio.mugnolo@uni-ulm.de

Heiko Neumann

University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
heiko.neumann@uni-ulm.de

Robin Nittka

University of Ulm, Helmholtzstraße 18
Institute of Applied Analysis
89069 Ulm
Germany
robin.nittka@uni-ulm.de

Günther Palm

University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
guenther.palm@uni-ulm.de

Wolfgang Peter

University of Ulm
Institute of Theoretical Physics
89069 Ulm
Germany
whmpeter@web.de

Roberto Pieraccini

SpeechCycle, Inc.
26 Broadway, 11th Floor
New York, NY 10004
USA
roberto@speechcycle.com

Florian Raudies

University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
florian.raudies@uni-ulm.de

Claudia Röscheisen

University of Ulm, Helmholtzstraße 18
Institute of Applied Analysis
89069 Ulm
Germany
claudia.roescheisen@uni-ulm.de

Wolfgang Schleich

University of Ulm
Institute of Quantum Physics
89069 Ulm
Germany
wolfgang.schleich@uni-ulm.de

Uwe Schöning

University of Ulm
Institute of Theoretical Computer Science
89069 Ulm
Germany
uwe.schoening@uni-ulm.de

Rainer Schuler

LogicLine EDV-SystemService GmbH
Planiestrasse 10
71063 Sindelfingen
Germany
rschuler@logicline.de

Friedhelm Schwenker
University of Ulm
Institute of Neural Information Processing
89069 Ulm
Germany
friedhelm.schwenker@uni-ulm.de

Frank Steiner
University of Ulm
Institute of Theoretical Physics
89069 Ulm
Germany
frank.steiner@uni-ulm.de

Eric Sträng
University of Ulm
Institute of Theoretical Physics
89069 Ulm
Germany
eric.straeng@uni-ulm.de

David Suendermann
SpeechCycle, Inc.
26 Broadway, 11th Floor
New York, NY 10004
USA
david@speechcycle.com

Jacobo Torán
University of Ulm
Institute of Theoretical Computer Science
89069 Ulm
Germany
jacobo.toran@uni-ulm.de

Fabian Wagner
University of Ulm
Institute of Theoretical Computer Science
89069 Ulm
Germany
fabian.wagner@uni-ulm.de

Christian Wawra
University of Ulm
Internal Medicine I
89069 Ulm
Germany
christian.wawra@uni-ulm.de

Dmitry Zaykovskiy
University of Ulm
Institute of Information Technology
89069 Ulm
Germany
dmitry.zaykovskiy@uni-ulm.de

Prologue

Milestones of Evolution, Information and Complexity

Wolfgang Arendt, Delio Mugnolo and Wolfgang Schleich

> *The modern world, our world of triumphant rationality, began on November 10, 1619, with a revelation and a nightmare. On that day, in a room in the small Bavarian village of Ulm, René Descartes, a Frenchman, twenty-three years old, crawled into a wall stove and, when he was well-warmed, had a vision. It was not a vision of God, or of the mother of God, or of the celestial chariots, or of the New Jerusalem. It was a vision of the unification of all science.*
>
> Philip J. Davis and Reuben Hersh, *Descartes' Dream*,
> Penguin, London 1986

René Descartes laid the foundation of modern science not only by his natural philosophy, but also by his mathematical contributions. For example, he addressed the tangent problem, which was only solved in its entirety 50 years later by Leibniz and Newton. For this purpose both of them invented mathematical calculus. In 1637 Descartes in his essay *Discours de la méthode* brought to light the physics of diffraction. He was the first to explain the most beautiful spectral phenomenon, the rainbow.

Δ Δ Δ

Since then, spectral analysis has come a long way. It has developed into one of the most fruitful concepts in natural sciences and technology. The overtones of an instrument, such as a tambourine or an organ pipe, and even the bodywork of a Mercedes limousine, exhibit a spectrum. Also in the microscopic world the concept of spectrum is useful. For example, the energy levels of the electron in a hydrogen atom form a spectrum. This fact turned out to be a crucial stepping stone for the development of quantum theory (Chapter 1).

Cosmic microwave background radiation was discovered in 1965 by Arno Penzias and Robert Wilson. In accordance with the Big Bang Theory, it fills the whole universe and is currently considered to be the major evidence for an expanding universe. For this discovery, Penzias and Wilson received the Nobel Prize in Physics in 1978 (Chapter 10). Such a microwave radiation possesses a spectrum which is characteristic of a so-called *black body*. Black bodies have been first considered by Gustav Kirchhoff in 1859 when he laid the foundation of the theory of thermal radiation. Fourteen years earlier he had already introduced two famous laws describing the time evolution of voltages and currents in electric circuits. In this way Kirchhoff single-handedly established modern electrical engineering.

Black bodies have also played a central role in the creation of quantum mechanics. Indeed, motivated by the problem of designing efficient lightbulbs, Max Planck in 1900 discovered that the energy of oscillators is quantized. Building on Planck's insights, Albert Einstein, born in Ulm in 1879, could explain the photoelectric effect. This explanation together with his discovery of the momentum of the light quantum opened the door to the development of quantum mechanics. For these achievements he was awarded the Nobel Prize in 1921. Moreover, his groundbreaking work on relativity, deeply rooted in Riemann's geometric theory, completely changed our understanding of the time evolution of the universe and marked the birth of modern cosmology (Chapter 3).

Δ Δ Δ

To a large degree today's electrical engineering lives off the spectral analysis of signals, for example, making cell phones work. It was Claude Shannon who in 1949 discovered that a finite number of samples suffices to capture a wave (Chapter 12). Here he could build on the concept of the Fourier transform, which was introduced by Joseph Fourier in 1822 in his *Théorie analytique de la chaleur*. Shannon's sampling theorem provides the basis for the technology of digitalising and eventually perfectly reconstructing a signal. Moreover, in the very same paper entitled *Communication theory of secrecy systems,* Shannon laid the mathematical foundation of cryptography.

Still, signal processing faces a major theoretical limitation: The shorter a pulse in time, the less well defined the frequency. Bounds of this kind are intimately related to Shannon's investigations collected in his seminal paper *A mathematical theory of communication* from 1948. In this article he introduced the concept of *information entropy*, a measure of the information contained in a random message. Today this article is commonly considered to have initiated information theory.

Δ Δ Δ

Also at the end of the 1940s, Donald Hebb was completing his most influential study, *The organization of behavior*. Therein he proposed a quantitative approach to the process of learning, thus giving birth to modern neuropsychology. Hebb was the first to analytically investigate the dual nature of the brain – biological tissue as well as source of perception – combining traditional behavioral studies and modern electrophysiology. His theory of learning suggested that synaptic connections are strengthened or weakened in order to achieve more efficient apprehension. Hebb's work introduced the notion of *synaptic plasticity* and paved the road for the interpretation of the brain as an ever-changing computing system.

Shortly before, in 1943, the first artificial *neural networks* had been introduced by Warren McCulloch and Walter Pitts in their article entitled *A logical calculus of the ideas immanent in nervous activity*. It soon became clear that boolean logic could be implemented in these theoretical devices, thereby enabling them to perform complex computations. Unfortunately, the early neural networks lacked any form of adaptation or learning features. It was Hebb's research that filled this gap. Even today Hebb's laws in their mathematical formulation are among the favorite theoretical tools when setting up and tuning an artificial neural network (Chapters 4, 7 and 14). They allow one to translate learning phenomena into the time evolution of a system of differential or difference equations (Chapter 2).

Δ Δ Δ

The spectrum of a wave can be viewed as a band of eigenfrequencies determined by the Helmholtz equation. This distribution of eigenvalues provides us with a deeper insight into the behaviour of light and matter and is described by Weyl's law, proven by Hermann Weyl in 1911. Surprisingly, there is a close analogue in number theory. The prime numbers are intimately related to Riemann's ζ-function whose nontrivial zeros look very much like a spectrum encountered in atomic physics. In this context Marcus de Sautoy talks about *the music of primes*, which is the title of his book on the Riemann ζ-function. Much is known about the distribution of the primes but the related Riemann's hypothesis on the zeros of the ζ-function is still a mystery. First formulated by Bernhard Riemann in 1859, it is probably the biggest open problem in mathematics today – in fact, it has been dubbed a Millennium Problem, whose solution would be rewarded with a \$1 000 000 prize by the Clay Mathematics Institute (Chapter 1).

Prime numbers and their distribution have fascinated mathematicians for generations. Today they serve us as a mathematical technology in cryptography. A modern life necessity is to transmit secret data, for example, for online banking purposes. It is counterintuitive that encryption can be made safer and more efficient by the use of public keys. In fact, cryptographic keys had to be kept strictly secret until the 1970s. However, even codes based on secret keys are not secure. The most prominent example is the Enigma code used by the Germany military in World War II and broken by Alan Turing in 1943. *Public keys* constituted a breakthrough and a radical change of the paradigm of secrecy. They were first proposed in 1976 in a famous paper by Whitfield Diffie and Martin Hellman entitled *New directions in cryptography*. In Diffie's and Hellman's words, "each user of the network can, therefore, place his enciphering key in a public directory. This enables any user of the system to send a message to any other user in such a way that only the intended receiver is able to decipher it." The actual realization of their project is due to Ronald Rivest, Adi Shamir, and Leonard Adleman, who in 1978 developed the RSA cryptographic system. This work won them the Turing Award in 2002. It is surprising that the long awaited solution of the most famous and originally thought to be useless problem, the proof of Fermat's Last Theorem by Andrew Wiles in the 1990s, also provided us with new tools for cryptography such as *elliptic curves*. In fact, their use for enciphering and deciphering was already implicitly contained in the work of Diffie and Hellman (Chapter 16).

Efficient cryptography is just one problem of modern signal theory. Another one is to find a language which permits error-free transmission of information in a process called coding/encoding. And it is again number theory, but also Fourier analysis, that gives us the right tools to perform this task with enormous efficiency (Chapter 13). Again, Shannon's sampling theorem plays a decisive role in this context.

Δ Δ Δ

Quantum theory, even though formulated in a quite abstract mathematical language, has reached a state of broad technological applications. In 1965, Richard Feynman received the Nobel Prize in Physics for his work on quantum electrodynamics. Seventeen years later he was one of the first to recognize the potential of a *quantum computer*. In all digital computers, starting from the Zuse Z3 and the ENIAC of the early 1940s to modern miniaturised devices, the logic is based on memory devices which store either a 0 or a 1, forming a bit of information. In a quantum computer this on/off dichotomy is replaced by the possibility of a *quantum bit* being in a superposition state. Abecedarian forms of such a quantum device exist to date in research labs only. Nevertheless, they have already been studied extensively at a theoretical level.

Once available, a quantum computer would substantially simplify large data analysis. Applications known today include, but are not limited to, the determination of shortest paths in networks or even the factorization of large numbers, for instance, by means of the algorithm discovered by Peter Shor in 1994 (Chapter 15). For this work he was awarded the Gödel Prize in 1999. It is remarkable that Shor's algorithm, implemented on a reasonably large quantum computer, could easily break common cryptographic techniques, including both RSA and methods based on elliptic curves. Possible remedies are random number generators based on Riemann's ζ-function. A fascinating relation between elliptic curves and the ζ-function is suggested by the Birch and Swinnerton–Dyer conjecture, formulated in the 1960s. It is still open and represents another of the Millennium Problems named by the Clay Institute.

However, not even quantum computers have an infinite potential. They may be exponentially faster when confronted with certain tasks, but they are not inherently more powerful than today's computers. Whenever we have to solve a problem with the help of a machine, even an ideal one such as a *universal Turing machine*, we have to use an algorithm which should be optimized. To quantify the intrinsic efficiency of algorithms to be implemented on computers is the goal of *algorithmic complexity theory*, founded by Juris Hartmanis and Richard Stearns in 1965. Their paper *On the computational complexity of algorithms* earned them the Turing Award in 1993.

Δ Δ Δ

Obviously, to *determine* the explicit solution of a problem in a short time, such as factoring of a large number, or to *check* whether given data indeed solve the problem, for example whether the product of a given set of primes yields the original large number, represent two different tasks. The problems which are solvable by a fast algorithm constitute the class *P*, whereas *NP* consists of those problems which allow for a fast algorithm that is able to check whether a given possible answer is indeed a solution of the problem. Here an algorithm is called *fast* if it can be performed in a time which grows at most polynomially with the input size.

While algorithms for finding prime numbers were already known to the ancient Greek, checking whether a given number is actually a prime seems to be demanding. Nevertheless, it is only a *P*-problem as shown by Manindra Agrawal in 2002 after his post-doc stay at the University of Ulm. This stunning discovery won him the Clay Research Award in the same year and the Gödel Prize (together with his coauthors) in 2006. Another example of the *P* vs. *NP* question is the prime factorization of large numbers (Chapters 15 and 16). Today most of the cryptography devices rely upon the belief that factorization into primes, which is clearly an *NP* problem, is not a *P* problem. Still, all attempts to prove this hypothesis have remained unsuccessful. In 1956, Gödel conjectured in a letter to von Neumann that, in general, it should be possible to replace trial and error approaches by efficient algorithms – as for example done for various problems in number theory – thus implicitly suggesting that $P = NP$. Still, more than fifty years later we do not yet know the answer to the *P* vs. *NP* problem.

While it is clear that each *P* problem is also *NP*, most computer scientists firmly believe that the converse is not true, that is $P \neq NP$. This question represents a major research field of theoretical computer science and is also one of the seven Millennium Problems of the Clay Institute.

Δ Δ Δ

Consider the task of coloring a geographical map under the constraint that no two adjacent countries can have the same color. To decide for a given map whether it is possible to complete this task using only three different colors is at least as difficult as any *NP* problem, whereas it is a *P* problem to find a four-coloring of the map. The latter task is closely related to the four-color-theorem, stating that each map can in fact be colored with a maximum of four colors. This theorem was first proposed as a conjecture in 1852 and proven only in 1976 by Kenneth Appel and Wolfgang Haken. Their proof is the first one ever to rely in an essential way upon computer aid, rather than human thought, and has therefore started an everlasting debate in the philosophy of science. The four-color-theorem represents the zenith of the interplay between mathematics, logic and theoretical computer science. In contrast, Kurt Gödel's incompleteness theorem from 1931 was this interplay's nadir. It was a great enlightment to the scientific community to learn from Gödel that no automatic, computer-based mathematics will ever be possible.

Coloring problems belong to *graph theory*, a field studying properties of discrete structures. Also based on graph theoretical objects is a recent proof of the long standing Horn conjecture on the distribution of eigenvalues of a sum of matrices. This question was addressed by Weyl in 1912 but was not proven until the 1990s.

One of the key elements of the proof are surprising connections between ideas from discrete and continuous mathematics as well as from quantum mechanics. For his contributions to the solution of the Horn conjecture, the Clay Research Award was assigned to Terence Tao in 2003. He was also awarded the Fields Medal in 2006 for his results in the arithmetic of prime numbers. It is remarkable that the once abstract graph theory has also found useful applications in algorithmic computer science, especially in data clustering (Chapters 8 and 9), sequencing (Chapter 5), classification (Chapter 11), and sorting (Chapter 17).

Δ Δ Δ

In spite of its limitations in large data analysis, the most efficient system for information processing and computing is still a natural one – the brain. Hermann von Helmholtz was the first to suggest and eventually prove in 1852 that thoughts – or rather neural transmissions – have a finite speed, in fact as slow as 30 m/s. In the year of von Helmholtz's discovery Santiago Ramón y Cajal, the father of modern neurobiology, was born. To him we owe the insight that the nervous system is a fine network consisting of neurons which communicate by means of synapses. For this work, based on a biochemical technique developed by Camillo Golgi, both Golgi and Ramón y Cajal were awarded the Nobel Prize in Medicine in 1906. In 1952 Alan Hodgkin and Andrew Huxley proposed a mathematical model for the propagation of electrical pulses inside individual neurons. They were able to show that the neural transmission happens by means of an ionic current, which can also be modeled as a diffusion process and propagates as a wave. Their model won them the Nobel Prize in Medicine in 1963 (Chapter 4).

The theory of diffusion was originated in 1822 by Fourier in his studies on heat conduction. Eventually, thirty-three years later Adolf Fick formulated the law of diffusion as a partial differential equation involving time as a variable. Such laws are called *evolution equations*. Already Fick recognized that his model was not limited to thermodynamics but had many more fields of application ranging from chemistry to finance. In fact, Fick's law also agrees up to nonlinear correction terms with Hodgkin's and Huxley's differential equations. Moreover, the signalling across synapses is a chemical phenomenon that is partially based on diffusion. Further Nobel Prizes in Medicine have been awarded for related discoveries in 1936, 1963, and 1970, in a *crescendo* that was made possible by the development of electron microscopy. A hundred years after Ramón y Cajal, diffusion processes still belong to the core of computational neuroscience – in a braid of chance, linear determinism, and chaos.

Δ Δ Δ

In 1887, Henri Poincaré won a contest sponsored by the King of Sweden asking for the solution of the famous *three-body problem* in celestial mechanics. In fact, Poincaré did not present the solution, but rather indicated a major problem in the mainstream approach to celestial mechanics itself. He pointed out that even a perfect deterministic theory would not yield a useful result, since usually the initial state of a system is not known with perfect accuracy. In 1908 he founded modern chaos theory in his book *Science et méthode* by suggesting that “it may happen that small differences in the initial conditions produce very great ones in the final phenomena”. Probably Albert Einstein was the first to come into touch with the question of whether chaos plays a role in the realm of quantum mechanics, in 1917. However, a systematic approach towards the chaotic behavior in atomic and subatomic systems was only initiated by Eugene Wigner in 1951. A striking connection between the eigenvalues of the Schrödinger operator and classical dynamics had been observed by Martin Gutzwiller in 1971. Nowadays, a promising development of the study of quantum chaos rests on simplified one-dimensional models where the familiar evolution equations of quantum mechanics and Kirchhoff’s circuit laws are united to describe the time evolution of *quantum graphs* (Chapter 6).

Δ Δ Δ

What about the realization of Descartes’ vision of unification of all sciences, almost four hundred years after his nightmare and his revelation? Mathematics has become the common language of all natural sciences. Still, abstract mathematics and applied sciences have attracted and repelled each other many times over the last centuries, following the alternation of rushing theoretical developments and real-world applications. May this book contribute to this everlasting interplay.

1
Weyl's Law: Spectral Properties of the Laplacian in Mathematics and Physics

Wolfgang Arendt, Robin Nittka, Wolfgang Peter,[1] *Frank Steiner*

1.1
Introduction

Weyl's law is in its simplest version a statement on the asymptotic growth of the eigenvalues of the Laplacian on bounded domains with Dirichlet and Neumann boundary conditions. In the typical applications in physics one deals either with the Helmholtz wave equation describing the vibrations of a string, a membrane (drum), a mass of air in a concert hall, the heat radiation from a body in thermal equilibrium, the fluctuations of the gravitational field in cosmology, or the Schrödinger equation of a quantum system which may be a simple quantum billiard, an atom, a molecule or a compound nucleus. While Weyl's seminal work was provoked by the famous black body radiation problem, i.e. an electromagnetic cavity problem, in particular by a conjecture put forward independently by Sommerfeld and Lorentz in 1910, Weyl's law has its roots in music and, respectively, acoustics. Already in 1877, Lord Rayleigh had, in his famous book, "The Theory of Sound" treated the overtones of a violin or piano string and the natural notes of an organ pipe or the air contained within a room. For a room of cubical shape he derived the correct asymptotic behavior for the overtones. The trick used by Rayleigh to count the vibrational modes was to reduce the problem to a three-dimensional lattice-point problem from which he could derive that the number of overtones with frequency between ν and $\nu + \mathrm{d}\nu$ grows at high frequencies, $\nu \to \infty$, asymptotically as $V \cdot \nu^3$ (Weyl's law!), where V is the volume of the room or analogously of an organ pipe. In 1900, Rayleigh realized that the same formula can be applied to a physically completely different, but mathematically equivalent problem: the heat radiation from a body in thermal equilibrium with radiation, the importance of which had been pointed out already in 1859 by Kirchhoff. The amount of energy emitted by a body is determined by the high-frequency spectrum of standing electromagnetic waves and that spectrum should be essentially the same as for the high overtones of an organ pipe, say.

[1] Corresponding author.

In the crucial black body radiation experiments carried out in the 1890s, which led Planck, in 1900, to the famous radiation law named after him and to the discovery of quantum theory, one measures the energy density emitted rather than the energy itself, i.e. the energy divided by the volume V. Thus it follows from Rayleigh's asymptotic result $V \cdot \nu^3$, derived for a cubical geometry, that the volume factor is canceled if one considers the energy density, in accordance with the expectations using physical arguments and, very importantly, in complete agreement with the experimental findings. It was realized, however, and emphasized by Sommerfeld and Lorentz in 1910 that there arises the *mathematical problem* to prove that the number of sufficiently high overtones which lie between ν and $\nu + \mathrm{d}\nu$ is *independent of the shape* of the enclosure and is simply *proportional to its volume*. It was a great achievement when Weyl proved in 1911 that, by applying the Fredholm–Hilbert theory of integral equations, the Sommerfeld–Lorentz conjecture holds! From then on, Weyl himself and many other mathematicians and physicists have studied and generalized Weyl's law by deriving corrections or even complete expressions for the remainder term.

The Weyl asymptotics as discussed above in the three-dimensional case is particularly striking if considered as an inverse spectral problem, which became quite popular after Kac's talk in 1966 entitled "Can one hear the shape of a drum?".

Subsequently, several partially affirmative answers to this question have been given. But, on the other hand, a particularly striking counterexample by Gordon, Webb and Wolpert from 1992 shows that not all geometric information about the domain is contained in the spectrum.

This chapter is organized as follows. In Section 1.2 we give a historical account of Weyl's law. The following two chapters are devoted to Weyl's law with remainder term and the statistical behavior of the latter using trace formulae. We discuss the Laplacian on the torus in Section 1.3 and the Laplace–Beltrami operator on Riemann surfaces in Section 1.4. Then two generalizations of Weyl's law to Robin boundary conditions and for unbounded quantum billiards are presented in Section 1.5. In Section 1.6 we provide a self-contained proof of Weyl's law for bounded Euclidean domains and Dirichlet boundary conditions; the case Weyl himself treated in his first article on this topic. However, we follow a different, very fruitful, approach based on heat kernels. In Section 1.7 we give an account on what is known today about Kac's question. In particular we show under which precise regularity assumptions one can hear whether a body is a ball.

1.2
A Brief History of Weyl's Law

1.2.1
Weyl's Seminal Work in 1911–1915

In February 1911, David Hilbert presented to a meeting of the Royal Academy of Sciences of Göttingen a short note [1] written by Hermann Weyl. This note contains for the first time a rigorous proof of the asymptotic behavior of the eigenvalues λ_n of the two-dimensional (scalar) Helmholtz wave equation

$$(\Delta + \lambda)\, u(x) = 0 \qquad (x \in \Omega) \tag{1.1}$$

satisfying the *Dirichlet boundary condition*

$$u(x) = 0 \qquad (x \in \partial\Omega)\ , \tag{1.2}$$

where $\Omega \in \mathbb{R}^2$ is an arbitrary bounded domain with area $|\Omega|$, boundary $\partial\Omega$, and

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$$

denotes the *Laplacian* on Ω. The "membrane problem" (1.1, 1.2) has nontrivial solutions u only for a discrete set of eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$. The corresponding eigenfunctions $\{u_n\}_{n\in\mathbb{N}}$ provide an orthonormal basis of $L^2(\Omega)$, and we may enumerate the eigenvalues in increasing order $0 < \lambda_1 \le \lambda_2 \le \ldots$

Note that the eigenvalues λ_n can have different physical interpretations. In the case of a vibrating membrane with clamped edge, where u describes the transversal vibrations of the membrane, one has $\lambda_n = k_n^2$, where $k_n = (2\pi/c)\,\nu_n$ is the wave number which is proportional to the eigenfrequency ν_n, i.e. to the pure tones which the membrane is capable of producing. The constant c is the sound velocity depending on the physical properties of the membrane, i.e. on the mass density and the tension under which the membrane is held. In the case of quantum mechanics, where u is the wave function having the meaning of a probability amplitude, Equation (1.1) is the time independent Schrödinger equation of a freely moving particle with mass m, and $\lambda_n = \left(2m/\hbar^2\right) E_n$ is proportional to the quantal energy levels E_n. ($\hbar$ denotes Planck's constant.)

Since explicit analytical expressions for the eigenvalues are known only for a few membranes with simple shape (for example equilateral triangles, rectangles, circles) and their numerical computation for large n is very difficult for general domains, it is natural to study their asymptotic distribution as $n \to \infty$. Applying the Fredholm–Hilbert theory of linear integral equations, Weyl proved that

$$\lim_{n\to\infty} \frac{n}{\lambda_n} = \frac{|\Omega|}{4\pi}\ . \tag{1.3}$$

Defining the *counting function* $N(\lambda) := \#\{\lambda_n \le \lambda\}$, (1.3) is equivalent to the asymptotic behavior

$$N(\lambda) = \frac{|\Omega|}{4\pi}\lambda + o(\lambda) \qquad (\lambda \to \infty)\,. \tag{1.4}$$

These results are now called *Weyl's law*. Shortly afterwards, Weyl submitted three papers [2–4] which contain the details of his proof, a generalization of (1.4) to the three-dimensional scalar wave equation ($\Omega \subset \mathbb{R}^3$),

$$N(\lambda) = \frac{|\Omega|}{6\pi^2}\lambda^{3/2} + o(\lambda^{3/2}) \qquad (\lambda \to \infty)\ , \tag{1.5}$$

and the extension to the vector Helmholtz wave equation describing the vibrations of the electric field $\boldsymbol{E}$ in an empty cavity Ω with perfectly reflecting walls $\partial\Omega$. As we shall discuss in more detail in Sections 1.2.3–1.2.8, it is exactly this electrodynamic cavity problem, studied extensively in those years by theoretical physicists, which was one of the open problems that provoked Weyl to start his seminal work.

The electromagnetic cavity problem requires of the electric field vector boundary conditions which are more involved than the simple boundary condition (1.2). In his first papers [2,4] on this problem, Weyl considered some nonphysical boundary conditions; following a suggestion of Levi–Civita, in his subsequent paper [5] the correct boundary conditions $\boldsymbol{E} \times \boldsymbol{n} = 0$ and $\nabla \boldsymbol{E} = 0$ on $\partial\Omega$ were taken into account. However, the Gauss law $\nabla \boldsymbol{E} = 0$ on Ω i.e. throughout the interior of the cavity, was still discarded.

In his paper [5] Weyl went even one step further and conjectured the existence of a second asymptotic term

$$N(\lambda) = \frac{|\Omega|}{4\pi}\lambda \mp \frac{|\partial\Omega|}{4\pi}\sqrt{\lambda} + o\left(\sqrt{\lambda}\right) \qquad (\lambda \to \infty) \tag{1.6}$$

for the two-dimensional problem (1.1), where $|\partial\Omega|$ denotes the length of the circumference of the membrane and the (–) sign refers to the Dirichlet boundary condition (1.2) and the (+) sign to the *Neumann boundary condition* ($\partial u/\partial n = 0$, $x \in \partial\Omega$), and

$$N(\lambda) = \frac{|\Omega|}{6\pi^2}\lambda^{3/2} \mp \frac{|\partial\Omega|}{16\pi}\lambda + o\left(\lambda\right) \qquad (\lambda \to \infty) \tag{1.7}$$

for the three-dimensional case, where $|\partial\Omega|$ now denotes the surface area of $\partial\Omega$. The formulae (1.6) and (1.7) became known as *Weyl's conjecture*. It was justified (under certain conditions on Ω) by Ivrii [6] and Melrose [7] only in 1980.

In 1915, Weyl concluded his work on the asymptotic behavior of eigenvalues with a study [8] of the elastic vibrations $\boldsymbol{u}$ of a homogeneous body with volume $|\Omega|$ which are determined by the solutions of the differential equation

$$B\Delta\boldsymbol{u} + A \operatorname{grad} \operatorname{div} \boldsymbol{u} + \lambda\boldsymbol{u} = 0\ . \tag{1.8}$$

Here λ is related to the frequency ν by $\lambda = (2\pi\nu)^2$, and A, B are positive constants (related to the Lamé coefficients characterizing the elastomechanical properties of the body). Imposing the boundary conditions $\nabla\boldsymbol{u} = 0$ and $\boldsymbol{n} \times \boldsymbol{u} = 0$ on the boundary $\partial\Omega$ of the body, Weyl proved for arbitrary shapes of the body

$$N(\lambda) = \frac{|\Omega|}{6\pi^2}F\lambda^{3/2} + o\left(\lambda^{3/2}\right) \qquad (\lambda \to \infty)\ , \tag{1.9}$$

where F is a function of the elastic constants, $F = 2/c_\mathrm{T}^3 + 1/c_\mathrm{L}^3$ with $c_\mathrm{T} = \sqrt{B}$ the transverse and $c_\mathrm{L} = \sqrt{A+B}$ the longitudinal sound velocity.

The Weyl formulae (1.3)–(1.7) and (1.9) are very striking since they tell us that the coefficient of the leading asymptotic term is determined only by the area, resp., the volume of the domain and is independent of its shape. That is one can "hear" the area of a drum (a membrane, held fixed along its boundary) or the volume of a cavity following Marc Kac's [9] acoustic rephrasing of the problem. We refer to Sections 1.2.8, 1.3.6, 1.3.7 and 1.7 for more details on Kac's problem.

In his papers [1, 2], Weyl mentions that his asymptotic formulae "provide in particular the solution to a problem the importance of which has recently been emphasized by Sommerfeld and Lorentz" (here and in the following citations, we employ free translations from the original German).

1.2.2 The Conjecture of Sommerfeld (1910)

In September 1910, Arnold Sommerfeld delivered a talk at the "82. Naturforscher-Versammlung" in Königsberg [10]. In this talk he studied the solution of the inhomogeneous differential equation in one, two and three dimensions

$$(\Delta_\Omega + \lambda)\, v = f \tag{1.10}$$

describing forced vibrations. For this purpose, he introduced the *resolvent kernel* (called the "*Green function*")

$$G_\Omega(x, y; \lambda) := \sum_m \frac{u_m(x) u_m(y)}{\lambda - \lambda_m} \qquad (x, y \in \Omega) \ , \tag{1.11}$$

where $u_m(x)$ are the eigenfunctions of (1.1). In addition to the Dirichlet boundary condition (1.2), Sommerfeld also considered Neumann boundary conditions and, "as in the theory of heat conduction", *Robin boundary conditions* ($h_1 u + h_2\,(\partial u/\partial n) = 0$ on $\partial\Omega$, h_1, h_2 constant or arbitrary functions on $\partial\Omega$). A formal application of the operator $(\Delta_\Omega + \lambda)$ to $G_\Omega(x, y; \lambda)$ (acting on the first argument x) gives $(\Delta_\Omega + \lambda)\, G_\Omega(x, y; \lambda) = \sum_m u_m(x) u_m(y)$, and Sommerfeld remarks that this expression is zero for $x \neq y$, but is infinite for $x = y$. He calls this expression "Zackenfunktion" ("spike function"), the physical interpretation of it is a "unit source" (point source). This is, of course, an early attempt to introduce the Dirac delta distribution, since the above expression is nothing other than the *completeness relation* of the orthonormal eigenfunctions $u_m \in L^2(\Omega)$

$$\sum_m u_m(x) u_m(y) = \delta(x - y) \ . \tag{1.12}$$

The solution of the inhomogeneous problem (1.10) then reads

$$v(x) = \int_\Omega G_\Omega(x, y; \lambda) f(y)\, \mathrm{d}y \ . \tag{1.13}$$

This result is quite remarkable since it allows one to reduce the problem (1.10) of the forced vibrations on Ω to the problem (1.1) of the free vibrations on the same domain Ω. "As some material is fully characterized by its spectral lines i.e.

by its free vibrational frequencies, so is also the behavior of a domain for arbitrary vibratory motions completely determined by the spectrum of its free vibrational possibilities." [10]

Sommerfeld [10] then discusses the convergence of the series (1.11): "In the one-dimensional case the series (1.11) is absolutely convergent, in the two- and three-dimensional case only conditionally convergent. In the first case the growth of the denominator $\lambda - \lambda_m$ is sufficient for convergence, since [...] the denominator becomes infinite, as m^2. In the latter cases, will $\lambda_{m,n}$, resp. $\lambda_{m,n,l}$, equally well always approach infinity quadratically in m, n, resp. l, as I do not doubt (1). However, such a growth is not sufficient, as is well known, to render the double sum over m, n resp., the triple sum over m, n, l, convergent. Rather, here the change of sign of the nominator $u_{m,n}(x)u_{m,n}(y)$ plays an essential role as it is guaranteed in its natural ordering by the oscillatory character of the series." In the above footnote (1) Sommerfeld adds: "The general and rigorous proof of this asymptotic behavior of the eigenvalues seems to me an important and grateful mathematical problem."

Here we have *Sommerfeld's conjecture* which was one of the motives for the pioneering work of Weyl.

Sommerfeld considers, as an application of his method, the "problem of acoustics of rooms" (using Neumann boundary conditions on the walls), and he emphasizes that his "method is fundamentally different from the classical method introduced in mathematical physics by Fourier", whereby he refers to Fourier's famous work "Théorie [analytique] de la chaleur" from 1822.

Here two remarks are in order. i) In his conjecture, Sommerfeld takes for granted that the eigenvalues depend for example in the three-dimensional case on three integers ("quantum numbers") (m, n, l) i.e. $\lambda_{m,n,l}$, "which each run from 0 to ∞ and have the meaning of the number of divisions of the domain by nodal surfaces with respect to the three dimensions. (One may think of the known cases of the parallelepiped or the sphere.)" Consequently, he considers the sum in Equation (1.11) as a triple sum running over m, n, and l. It is known, however, that the situation envisaged by Sommerfeld holds in general only for domains for which the wave equation (1.1) is separable in coordinates (q_1, q_2, q_3) i.e. where the particular solutions can be written as a product $u_{m,n,l} = u_m(q_1)v_n(q_2)w_l(q_3)$. In the generic case, however, i.e. for a cavity with arbitrary shape, the eigenvalues depend on a single positive integer only, which just counts the eigenvalues in increasing order, as assumed in (1.11). ii) Sommerfeld points out that the Green function (1.11) "degenerates $(G = \infty)$" at the points $\lambda = \lambda_m$ "according to the general *resonance principle*, except at special positions of the point source, if, for example, $u_m(x) = 0$, and therefore the critical eigenvibration is not excited." In the physics literature, the Green function (1.11) is considered as a distribution by adding a small positive imaginary part $(\varepsilon > 0)$ to $\lambda \in \mathbb{R}$, i.e. one considers the kernel of the regularized resolvent operator $(\lambda + i\varepsilon + \Delta)^{-1}$. We refer also to Sections 1.3.4, 1.4 and 1.6 where expressions similar to (1.11) are given for the Green's function, for example for the heat kernel.

1.2.3 The Conjecture of Lorentz (1910)

At the end of October 1910, i.e. one month after Sommerfeld's talk, Hendrik Antoon Lorentz delivered six lectures at Göttingen under the title "Old and new problems of physics" published [11] from notes taken by Max Born and talked over with Lorentz. See also the letter dated October 28, 1929, sent by Max Born to Einstein, together with Born's comment to this letter from 1969 [12].

Lorentz, who had already received in 1902 the second Nobel prize in physics, was at that time probably the most famous living theoretical physicist. He was invited to Göttingen by the Wolfskehl commission of the Göttingen Academy which was to confer a prize for proving Fermat's last theorem. As long as the prize was not awarded, the proceeds from the principal should be used to invite eminent scientists to lecture at Göttingen. (Paul Wolfskehl (1856–1906), originally a physician, fell ill with multiple sclerosis and then became a mathematician working mainly on number theory; he taught at the Technical University of Darmstadt. The first Wolfskehl lecture was given by Poincaré in 1908 and later lectures were given, among others, by Einstein and Planck; in 1922 Niels Bohr delivered his legendary Wolfskehl lectures on his theory of the atom which later became known as the "Bohr-Festspiele". In 1997 the Wolfskehl prize was given to Andrew Wiles.)

In his last three lectures, Lorentz discussed "the phenomenon of radiating heat". The end of the fourth lecture reads as follows. "In conclusion there is a mathematical problem worth mentioning which perhaps will arouse the interest of mathematicians who are present. It originates in the radiation theory of Jeans. In an enclosure with a perfectly reflecting surface there can form standing electromagnetic waves analogous to tones of an organ pipe; we shall confine our attention only to the very high overtones. Jeans asks for the energy in the frequency interval $\mathrm{d}\nu$. To this end he first of all calculates the number of overtones which lie between the frequencies ν and $\nu + \mathrm{d}\nu$ and then multiplies this number by the energy which belongs to the frequency ν, and which according to a theorem of statistical mechanics is the same for all frequencies. In this manner he gets, indeed, the correct law of the radiation at long wavelengths."

"It is here that there arises the mathematical problem to prove that the number of sufficiently high overtones which lie between ν and $\nu + \mathrm{d}\nu$ is independent of the shape of the enclosure and is simply proportional to its volume. For several simple shapes, for which the calculation can be carried out, this theorem will be verified in a Leiden dissertation. There is no doubt that it holds in general even for multiply connected spaces. Analogous theorems will also hold for other vibrating structures like elastic membranes and air masses etc."

Weyl, who was present at Lorentz's lectures, writes in a footnote of his second paper [2]: "Lorentz has stated the theorem proven here in Section 1.6 as a plausible conjecture on physical grounds. The simplest cases, for which the proof can be achieved by a direct computation of the eigenvalues, are treated in the Leiden dissertation of Fräulein Reudler." Actually Johanna Reudler verified [13] that the asymptotic number of modes depends only on the volume for three special cases, the parallelepiped, the sphere, and the cylinder.

There is an apocryphal report that Hilbert predicted that the theorem would not be proved in his lifetime [9]. Well, as we have seen, he was wrong by many, many years.

Forty years after Lorentz's lectures, Weyl came back to the "eigenvalue problem" [14]: "H.A. Lorentz had impressed upon the mathematicians the urgency for physics of a settlement of this question. For a pupil of Hilbert around 1910 it was natural to visualize the question as one concerning integral equations." In the next section of this paper [14] Weyl draws attention to a more difficult problem by saying: "The physicist will not be satisfied with a knowledge of the asymptotic behavior of the eigenvalues alone; that of the eigenfunctions should also be investigated." And Weyl mentions in this connection Carleman's law, see Section 1.4.3.

Further on in this paper we read the following sentences: "I feel that these informations about the proper oscillations of a membrane, valuable as they are, are still very incomplete. I have certain conjectures on what a complete analysis of their asymptotic behavior should aim at; but since for more than 35 years I have made no serious attempt to prove them, I think I had better keep them to myself."

1.2.4 Black Body Radiation: From Kirchhoff to Wien's Law

The study of the heat radiation from a body in thermal equilibrium with radiation has played an eminent role in the history of physics and mathematics for it led Planck in 1900 to the discovery of the quantum theory and Weyl in 1911 to a first proof of the eigenvalue asymptotics. (There are several historical studies on this subject. Here we rely on the excellent account given by Pais [15] who, however, does not discuss the aspects concerning Weyl's law.) The importance of the heat radiation problem was realized already in 1859 by Gustav Kirchhoff [16]. Let the radiation energy which a body absorbs be converted to thermal energy only, not to any other energy form, and denote by $\mathcal{E}_\nu \, \mathrm{d}\nu$ the amount of energy emitted by the body per unit time per cm^2 in the frequency interval $\mathrm{d}\nu$. (Actually, Kirchhoff uses the wavelength λ instead of the frequency ν.) Furthermore, let A_ν be its absorption coefficient for frequency ν. Kirchhoff showed that the ratio $\mathcal{E}_\nu / A_\nu$ is a universal function which depends only on ν and the equilibrium (absolute) temperature T and is independent of the shape and any other properties of the body i.e.

$$\frac{\mathcal{E}_\nu}{A_\nu} = J(\nu, T) \; . \tag{1.14}$$

A general proof of (1.14) was given much later by Hilbert using the theory of linear integral equations and his "axiomatic method" [17–19].

Kirchhoff called a body *perfectly black* or just *black* for short if $A_\nu = 1$. Thus $J(\nu, T)$ is the emitted power of a black body which can be measured if we assume (with Kirchhoff) that a perfect black body can be realized by "a space enclosed by bodies of equal temperature, through which no radiation can penetrate" [16], i.e. by an enclosure with perfectly reflecting walls. Kirchhoff challenged theorists and experimentalists alike: "It is a highly important task to find this function J. Great difficulties stand in the way of its experimental determination; nevertheless, there appear grounds for the hope that it can be found by experiments because there is no doubt that it has a simple form, as do all functions which do not depend on the properties of individual bodies and which one has become acquainted with before now." [16]

It is worthwhile to mention that Kirchhoff reports in the same paper about his experiments carried out with sunlight, interpreted as heat radiation of very high temperature produced in the atmosphere of the sun, and about his discovery of sodium there. He concludes: "Thus a way is found to ascertain the chemical nature of the atmosphere of the sun, and the same way promises also some information on the chemical nature of the brighter fixed stars." [16]

It will be seen later that Kirchhoff's statement about the shape-independence of $J(\nu, T)$ implicitly implies part of Weyl's law (1.7) stating that the leading term of the counting function is proportional to the volume $V := |\Omega|$ of the cavity by which the black body is realized. At this point it is convenient to express J in terms of the spectral energy density $\varrho(\nu, T)$ which gives the energy per unit volume of the heat radiation in thermal equilibrium at temperature T for frequency ν:

$$J(\nu, T) = \frac{c}{8\pi}\varrho(\nu, T) \tag{1.15}$$

(c is the velocity of light *in vacuo*.) It was conjectured by Josef Stefan on experimental grounds in 1879 and proved theoretically by Ludwig Boltzmann in 1884 [20] that the mean total energy $\langle E\rangle(T)$ radiated by the black body is given by the *Stefan–Boltzmann law*

$$\langle E\rangle(T) = V\int\limits_0^\infty \varrho(\nu, T)\,\mathrm{d}\nu = V\sigma T^4 \ , \tag{1.16}$$

where σ is a universal constant (now called the *Stefan–Boltzmann constant*, whose universal value could only be calculated after the discovery of Planck's law). Boltzmann's proof involves thermodynamics and the electromagnetic theory of Maxwell according to which the mean radiation pressure $\langle p\rangle$ obeys the equation of state

$$\langle p\rangle = \frac{1}{3}\frac{\langle E\rangle}{V} \ .$$

Important progress was made by Wilhelm Wien who proved in 1893 that $\varrho(\nu, T)$ has to be of the following form (*Wien's displacement law*) [21]

$$\varrho(\nu, T) = \nu^3 f(\nu/T) \ . \tag{1.17}$$

Thus the heat radiation problem was reduced to determining, instead of $J(\nu, T)$, the universal function $f(x)$ of the single scaling variable $x = \nu/T$. (Obviously, from (1.17) one immediately derives the Stefan–Boltzmann law (1.16) with $\sigma = \int_0^\infty x^3 f(x)\,\mathrm{d}x$.) Over the years, many proposals for the correct form of f have appeared, see for example the four different forms discussed in [22]. Still, 20 years later, Einstein wrote in 1913: "It would be edifying if we could weigh the brain substance which has been sacrificed by the theoretical physicists on the altar of this universal function *f*; and the end of these cruel sacrifices is not yet in sight!" [23]

In 1896 Wien proposed [24] the exponential form $f_W(x) := \alpha e^{-\beta x}$ (α, β positive constants), that is (*Wien's law*)

$$\varrho_W(\nu, T) = \alpha\nu^3 e^{-\beta\nu/T} \,. \tag{1.18}$$

At the same time, Friedrich Paschen carried out precise measurements [22, 25] in the near-infrared (for wavelengths $\lambda = c/\nu = 1$–$8\,\mu$ m, $T = 400$–1600 K) which were in such a good agreement with Wien's law (1.18) that he concluded: "It would seem very difficult to find another function of the two variables ν and T [Equation (1.18)] that represents the observations with as few constants." [25]. Thus it appeared that Wien's law was the final answer to the black-body problem.

1.2.5
Black Body Radiation: Rayleigh's Law

In June 1900, that is several months *before* Planck's revolutionary discovery, Lord Rayleigh made another proposal [26] which for the first time introduces into the black body radiation problem the *density of states* $D(\nu)$, that is the density of the vibrational modes of a black body cavity. This step played an important role since Rayleigh's proposal relies on an assumption which 10 years later led to the conjectures of Sommerfeld and Lorentz, and finally to Weyl's law (as already discussed in Sections 1.2.1–1.2.3).

Rayleigh's starting point is the observation that Wien's law (1.18) "viewed from the theoretical side [...] appears to me little more than a conjecture ...", and "... the law seems rather difficult of acceptance, especially the implication that as the temperature is raised, the radiation of given frequency approaches a limit." [26] Indeed, one obtains from (1.18) $\lim\limits_{T\to\infty} \varrho_W(\nu, T) = \alpha\nu^3$. He continues: "the question is one to be settled by experiment; but in the meantime I venture to suggest a modification of (1.18), which appears to me more probable *a priori*." [26]

Without further explanation, Rayleigh assumes, first of all, that the equilibrium distribution ϱ is proportional to the density of the vibrational modes of the cavity per unit volume, that is $\varrho(\nu, T) \sim D(\nu)/V$, where

$$D(\nu) := \frac{\mathrm{d}\overline{\mathcal{N}}}{\mathrm{d}\nu} \text{ with } \overline{\mathcal{N}}(\nu) := \overline{N}\Big((2\pi/c\nu)^2\Big)$$

and $\overline{N}(\lambda)$ denotes the leading asymptotic term of the counting function expressed in terms of the frequency ν. Secondly, he assumes according to the "Boltzmann–Maxwell doctrine of the partition of energy" (that is the equipartition theorem) that "every mode of vibration should be alike favored ...". Thus he assumes that "the energy should be equally divided among all the modes. ... Since the energy in each mode is proportional to *T*" (that is proportional to $k_B T$ in modern notation, where k_B denotes Boltzmann's constant, introduced actually only later by Planck!), Rayleigh's assumption amounts to $\varrho(\nu, T) \sim (D(\nu)/V) \cdot T$. As an "illustration" he first considers "the case of a stretched string vibrating transversely" and derives the correct Weyl asymptotics $\overline{N}(\lambda) \sim \sqrt{\lambda}$, that is $\overline{\mathcal{N}}(\nu) \sim \nu$ and thus $D(\nu) =$ constant ("when ν is large enough"). Then he continues: "When we pass from one dimension to three dimensions, and consider for example the vibrations of a cubic mass of air, we have ('Theory of Sound', paragraph 267) as the equation for ν^2, $\nu^2 = p^2 + q^2 + r^2$, where p, q, r are integers representing the number of subdivisions in the three directions. If we regard p, q, r as the coordinates of points forming a cubic array, ν is the distance of any point from the origin. Accordingly the number of points for which ν lies between ν and $\nu + \mathrm{d}\nu$, proportional to the volume of the corresponding spherical shell, may be represented by $\nu^2\,\mathrm{d}\nu$, and this expresses the distribution of energy according to the Boltzmann–Maxwell law, so far as regards the wavelength or frequency. If we apply this result to radiation, we shall have, since the energy in each mode is proportional to T, $T\nu^2\,\mathrm{d}\nu$ ...." [26] Thus Rayleigh obtains (apart from the numerical coefficient) the correct Weyl asymptotics for a three-dimensional cavity, that is $\overline{N}(\lambda) \sim V\lambda^{3/2}$, see (1.5), which leads to $\overline{\mathcal{N}}(\nu) \sim V\nu^3$ or $D(\nu)/V \sim \nu^2$, and thus to

$$\varrho_{\mathrm{REJ}}(\nu, T) = c_1\nu^2 T\,, \tag{1.19}$$

which is commonly known as the Rayleigh–Jeans law but which should rather be referred to as the *Rayleigh–Einstein–Jeans law*. (We shall discuss below Einstein's, Rayleigh's second, and Jeans' derivation of (1.19) which includes also the explicit expression for the coefficient c_1.)

Here several remarks are in order. i) It is obvious that Rayleigh did not worry about the fact that he used the scalar wave equation in his derivation of the mode asymptotics, by referring to his famous book on acoustics [27], instead of the vector Maxwell equations, which were studied only later by Weyl [2, 4, 5]. ii) In deriving the vibrational mode asymptotics for a cubical box, Rayleigh takes for granted that the result $\overline{\mathcal{N}}(\nu) \sim V\nu^3$ holds for any shape of the cavity and thus concludes that $D(\nu)/V$ is independent of the shape. In other words, Rayleigh *assumes* to be true what 10 years later was formulated as a *conjecture* by Sommerfeld and Lorentz. iii) Although his derivation of $D(\nu)/V \sim \nu^2$ holds only asymptotically for $\nu \to \infty$, he derives from this result the law (1.19) stating that it may have the proper form when ν/T is small! iv) Rayleigh observes that (1.19) is of the general scaling form (1.17) (with $f_{\mathrm{REJ}}(x) = c_1/x$), and he regards this "as some confirmation of the suitability of (1.19)." v) Without further comment, Rayleigh writes in his paper [26]: "If we introduce the exponential factor, the complete expression will be

$$\varrho_{\mathrm{R}} = c_1\nu^2 T e^{-c_2\nu/T}\,.\text{''} \tag{1.20}$$

It is this expression which became known as the *Rayleigh law*. vi) There is no doubt that Rayleigh must have realized that (1.19) is entirely unacceptable since the quadratic dependence of ν leads to a physically meaningless divergence (later called "ultraviolet catastrophe" by Ehrenfest) of the total radiation energy (see (1.16)). Of course, by multiplying with the exponential "convergence factor" taken over from

Wien's law (1.18), the divergence is avoided and the Stefan–Boltzmann law (1.16) holds with $\sigma = 2c_1/c_2^3$.

At the end of his paper Rayleigh writes [26]: "Whether (1.20) represents the facts of observation as well as (1.18) I am not in a position to say. It is to be hoped that the question may soon receive an answer at the hands of the distinguished experimenters who have been occupied with this subject."

We can assume that Rayleigh was well informed about the two teams working in Berlin on black body radiation experiments. The first of these, Otto Lummer and Ernst Pringsheim, had already shown in February 1900 that Wien's law (1.18) fails in the wavelength region $\lambda = 12$–$18\,\mu$m (for $T = 300$–1650 K) [28]. The second team, Heinrich Rubens and Ferdinand Kurlbaum, presented their measurements in the even further infrared ($\lambda = 30$–$60\,\mu$m, $T = -188$–$1500\,^\circ$C) to the Prussian Academy on October 25, 1900 [29]. In a famous figure, they plotted ϱ as a function of T at the fixed wavelength $\lambda = 51.2\,\mu$m and compared their data with some theoretical curves. One of these was the Wien curve, another the Rayleigh curve. Both curves did not work! But then we read in the paper [29] that they had compared their data with a "fifth formula, given by Herr M. Planck after our experiments had already been concluded ..." and which "reproduces our observation within the limits of error."

1.2.6
Black Body Radiation: Planck's Law and the Classical Limit

According to Pais [15], Planck probably discovered his law in the early evening of Sunday, October 7, 1900, Rubens and his wife had called on the Plancks on the afternoon of that day. In the course of conversation, Rubens mentioned to Planck that he had found $\varrho(\nu, T)$ to be proportional to T for small ν. Planck went to work after the visitors left and found an interpolation between his results and Wien's law, Equation (1.18). He communicated his formula by postcard to Rubens the same evening and stated it publicly [30] in a discussion remark on October 19, following the presentation of a paper by Kurlbaum. Expressed in notations introduced by Planck two months later, *Planck's law* reads:

$$\varrho_P(\nu, T) = \frac{8\pi h\nu^3}{c^3}\frac{1}{e^{h\nu/k_\mathrm{B}T} - 1}\,, \tag{1.21}$$

where h denotes Planck's constant and k_B is Boltzmann's constant.

Let us consider two limits of Planck's law. First, in the high-frequency or low-temperature regime, which is now identified as the *quantum limit* in which the photon energy $h\nu$ is much larger than the thermal energy $k_\mathrm{B}T$, that is $h\nu/k_\mathrm{B}T \gg 1$, we recover Wien's law (1.18) with $\alpha = (8\pi h)/c^3$ and $\beta = h/k_\mathrm{B}$. This explains why Paschen's experiments [22, 25], for which $h\nu/k_\mathrm{B}T \approx 15$ holds, were in such a good agreement with Wien's law, as already mentioned. At the other extreme of low frequency or high temperature, $h\nu/k_\mathrm{B}T \ll 1$, which is obtained from Planck's law in the formal limit when Planck's constant approaches zero, $h \to 0$, and is now identified as the *semiclassical limit*, we recover the Rayleigh–Einstein–Jeans law (1.19)

$$\varrho_{\mathrm{P}}(\nu, T) = \frac{8\pi\nu^2}{c^3}(k_{\mathrm{B}}T)\,[1 + O(h)] \quad (h \to 0)\ . \tag{1.22}$$

A comparison with (1.19) gives the correct value for the constant c_1 left undetermined by Rayleigh, that is

$$c_1 = \frac{8\pi k_{\mathrm{B}}}{c^3} = \frac{8\pi}{c^3}\frac{R}{N_{\mathrm{A}}}\ , \tag{1.23}$$

which does not depend on h, and where R is the gas constant and N_{A} is Avogadro's number.

Since our main interest here is to understand the role played by Weyl's law, we are not discussing at this point the arguments using "resonators" which led Planck to his formula. (Planck's original derivation does not refer to the vibrations of the cavity and thus does not involve the density of states.) Using the fact that the correct formula for the radiating heat, that is Planck's formula, in the classical limit exactly takes the form of the Rayleigh–Einstein–Jeans law, we can interpret the latter in purely classical terms by making the general ansatz (valid only for $h = 0$!)

$$\varrho_{\mathrm{class}}(\nu, T) := \lim_{V\to\infty}\left(\frac{\mathcal{D}_{\mathrm{em}}(\nu)}{V}\right)k_{\mathrm{B}}T\ , \tag{1.24}$$

where $\mathcal{D}_{\mathrm{em}}(\nu)$ denotes the density of states of the electromagnetic vibrations in a cavity of volume V. Furthermore, we have taken care of the fact that the predictions of thermodynamics involve the so-called *thermodynamic limit* $V \to \infty$. Here

$$\mathcal{D}_{\mathrm{em}}(\nu) := \frac{\mathrm{d}\overline{\mathcal{N}}_{\mathrm{em}}(\nu)}{\mathrm{d}\nu} \text{ with } \overline{\mathcal{N}}_{\mathrm{em}}(\nu) = 2\overline{\mathcal{N}}(\nu) = 2\overline{N}\left((2\pi/c\nu)^2\right)\ ,$$

where $\overline{N}(\lambda)$ denotes the two asymptotic terms of the counting function (1.7) for the three-dimensional case, and the factor 2 comes from the two polarizations of the photon. We then obtain

$$\overline{\mathcal{N}}_{\mathrm{em}}(\nu) = V\frac{8\pi}{3c^3}\nu^3 + O_{|\partial\Omega|}(\nu^2) \tag{1.25}$$

which leads to

$$\lim_{V\to\infty}\left(\frac{\mathcal{D}_{\mathrm{em}}(\nu)}{V}\right) = \frac{8\pi}{c^3}\nu^2 \tag{1.26}$$

since $\lim\limits_{V\to\infty}(|\partial\Omega|/V) = 0$, where $|\partial\Omega|$ denotes the surface area of the cavity.

In his famous book, originally published in 1928 in German under the title "Gruppentheorie und Quantenmechanik" [31, p. 103–104 and p. 402] Weyl treats the black-body radiation and proves that it "is mathematically equivalent to a system of infinitely many oscillators." He then states, without proof: "For high frequencies ν there are approximately $V\left(8\pi\nu^2\,\mathrm{d}\nu/c^3\right)$ modes of oscillation in the frequency interval $\nu, \nu + \mathrm{d}\nu$. We are interested above all in the limiting case of an infinitely large cavity; the spectrum then becomes continuous and our formula for the density of frequencies becomes exact." In a footnote he adds: "This result is easily obtained by elementary methods for a rectangular parallelepiped. For the general proof see H. Weyl [4, 5, 8]." It is clear that the limit $V \to \infty$ is an idealization which can never be realized in a physical experiment. Rather the "assumption must always hold that the linear dimensions of all cavities considered and also the curvature of the radii of all surfaces considered must be large compared with the wavelengths of the radiation. Then we are allowed, without making a noticeable error, to neglect the influences of the form of the boundaries caused by diffraction." [32, p. 2]

Inserting (1.26) into (1.24), one obtains

$$\varrho_{\text{class}}(\nu, T) = \frac{8\pi k_{\text{B}}}{c^3}\nu^2 T \tag{1.27}$$

which is precisely the Rayleigh–Einstein–Jeans law (1.19) with the correct power behavior in ν and the same coefficient (1.23) as obtained from the exact Planck formula. It is thus seen that heat radiation (in the classical limit) is indeed independent of the shape of the cavity due to Weyl's law and the performance of the thermodynamical limit.

As shown above, Planck's radiation law (1.21) from October 1900 can be considered as a simple interpolation formula which smoothly interpolates between the Rayleigh–Einstein–Jeans law (1.27) and Wien's law (1.18). In fact, it differs from Wien's law only by the –1 in the denominator. It has rightly been said [15], that even if Planck had stopped after October 19, he would forever be remembered as the discoverer of the radiation law. It is a true measure of his greatness that he went further. He wanted to interpret (1.21). That made him to discover the quantum theory. Already on December 14, 1900, Planck presented a theoretical derivation of his formula to the German Physical Society in Berlin [33] and shortly afterwards (7 January 1901) submitted his famous paper [34]. More and more precise measurements carried out during the following years established Planck's formula as the correct phenomenological law of black body radiation. It is thus quite astonishing to learn that several excellent theoretical physicists, in particular Lorentz, Lord Rayleigh, and Jeans, worked on alternative theories leading to formulae different from Planck's. Ironically, since Planck's derivation does not rely on the density of states, the origin of Weyl's law lies just in these alternative approaches. Therefore, a history of Weyl's law without a discussion of these differing theories would be incomplete.

1.2.7
Black Body Radiation: The Rayleigh–Einstein–Jeans Law

First of all, one should understand why some theorists were seeking for different theories of black body radiation despite the great empirical success of Planck's formula. The explanation is quite obvious: they realized that Planck's radiation theory was not satisfactory from a theoretical point of view; in fact, it was inconsistent! As the above quotation from Einstein [23] shows, the problem still existed in 1913; the ultimate derivation of Planck's formula was only provided in 1924 using the correct Bose–Einstein quantum statistics.

In 1903, Lorentz [35] derived (1.19) in the low-frequency limit together with the value $c_1 = \left(16\pi\alpha/3c^3\right)$ for the coefficient c_1 where α is a constant such that αT represents the mean kinetic energy of a molecule of a gas. Comparing (1.19) with the low-frequency limit of Planck's formula (1.21), he obtained $\alpha = (3/2)k_\mathrm{B}$ (see also (1.23)) and states: "Now the mean kinetic energy of a molecule of a gas would be $(3/2)kT$ according to Planck ... there appears therefore to be full agreement between the two theories in the case of long waves, certainly a remarkable conclusion, as the fundamental assumptions are widely different."

The year 1905 is one of the most amazing ones in the history of science: it marks, first of all, Einstein's annus mirabilis with his five seminal papers, where only the first one on the famous light quantum hypothesis [36] concerns us here, since it deals with the radiation problem, and, secondly, the series of papers published by Rayleigh [37, 38] and Jeans [39–42] on the radiation problem using the Weyl asymptotics.

From reading these papers it becomes clear that Einstein is the only one who takes Planck's formula serious since it "agrees with all experiments to date" [36]. But in Section 1.1 of this paper entitled "On a difficulty concerning the theory of the « black radiation »" [36] he implicitly expresses his doubts on Planck's derivation by showing that Planck should have obtained (1.27) instead of his formula (1.21)! The argument is very simple. Planck's starting point in his derivation is the formula

$$\varrho(\nu, T) = \frac{8\pi\nu^2}{c^3}\langle E\rangle(\nu, T) \ , \tag{1.28}$$

where $\langle E\rangle(\nu, T)$ is the average energy of a Planck resonator of frequency ν at the joint equilibrium of matter and radiation at temperature T. Furthermore, the equilibrium energy of a one-dimensional resonator is according to the equipartition theorem given by $\langle E\rangle(\nu, T) = k_\mathrm{B}T$, and inserting this into (1.28), Einstein obtains (1.27). We thus see that the radiation law (1.27), commonly known as the Rayleigh–Jeans law, ought to be called the Rayleigh–Einstein–Jeans law [15]. Many years later Einstein said: "If Planck had drawn this conclusion, he probably would not have made his great discovery, because the foundation would have been withdrawn from his deductive reasoning." [43]

Years later Planck himself presented two derivations of (1.27) in his famous book "Theorie der Wärmestrahlung" [32] and concluded: "It is not too much asserted if we say in generalizing: The classical theory leads of necessity to Rayleigh's radiation law."

Einstein's paper [36] was submitted on 17 March 1905, and thus is the earliest among the above mentioned papers by Rayleigh and Jeans. (Rayleigh's first paper [37] was submitted on 6 May 1905; Jeans' first paper on radiation [39] on 20 May 1905.)

As discussed in Section 1.2.5, Rayleigh was the first [26] to have already counted in 1900 "the number of modes corresponding to any finite space occupied by radiation" [37] and to obtain the law (1.19), however, without determining the coefficient c_1. "Elicited by the very clear statement of his view which Mr. Jeans gives in NATURE of April 27 (1900) [44]", he repeats the arguments of his former paper [26] "with an extension designed to determine the coefficient as well as the law of radiation" [37]. By counting the modes within a cube of length l (Weyl's law), he obtains again (1.19) "as probably representing the truth when ν is small." He remarks that this formula agrees with Planck's in the limit when ν is small apart from the fact that his value for c_1 "is eight times as large as that found by Planck." Rayleigh adds: "A critical comparison of the two processes would be of interest, but not having succeeded in following Planck's reasoning I am unable to undertake it. As applying to all wavelengths, his formula would have the greater value if satisfactorily established. On the other hand, the reasoning leading to (1.19) is very simple, and this formula appears to me a necessary consequence of the law of equipartition as laid down by Boltzmann and Maxwell. My difficulty is to understand how another process, also based on Boltzmann's ideas, can lead to a different result." [37]

Two days after Rayleigh's letter [37] Jeans submitted a short letter [39] in reply to Rayleigh. His main point was "the general question of the applicability of the theorem of equipartition to the energy of the ether" as opened up by Rayleigh. He takes up "Lord Rayleigh's example of a stretched string, say a piano wire" and then discusses the "vibrations of the ether in a finite enclosure". He writes: "It is easily seen that the number of slow vibrations is approximately proportional to the volume of the enclosure, so that roughly the energy of ether must be measured per unit volume in order to be independent of the size of the enclosure." He then arrives at (1.19), but without determining the value for c_1. On June 7, Jeans adds a "postscript" to his paper [40] and calculates again "the number of degrees of freedom of the æther" by referring to Rayleigh's book [27](!). From this he obtains the radiation law (1.19) together with the correct value (1.23) for the coefficient c_1. "This is one-eighth of the amount found by Lord Rayleigh, but agrees exactly with that given by Planck for large values of λ. It seems to me that Lord Rayleigh has introduced an unnecessary factor 8 by counting negative as well as positive values of his integers p, q, r." (See the discussion before equation (1.19).) A month later, Rayleigh replies to Jeans [38]: "In NATURE, May 18, I gave a calculation of the coefficient of complete radiation at a given absolute temperature for waves of great length on principles laid down in 1900, and it appeared that the result was eight times as great as that deduced from Planck's formula for this case. In connection with similar work of his own, Mr. Jeans has just pointed out that I have introduced a redundant factor 8 by counting negative as well as positive values of my integers p, q, r – I hasten to admit the justice of this correction. But while the precise agreement of results in the case of very long waves is satisfactory so far as it goes, it does not satisfy the wish expressed in my former letter for a comparison of processes. In the application to waves that are not long, there must be some limitation on the principle of equipartition. Is there any affinity in this respect between the ideas of Prof. Planck and those of Mr. Jeans?"

On July 27, Jeans published another letter [41]: "On two occasions (NATURE, May 18 and July 13) Lord Rayleigh has asked for a critical comparison of two theories of radiation, the one developed by Prof. Planck and the other by myself, following the dynamical principles laid down by Maxwell and Lord Rayleigh. It is with the greatest hesitation that I venture to express my disagreement with some points in the work of so distinguished a physicist as Prof. Planck, but Lord Rayleigh's second demand for a comparison of the two methods leads me to offer the following remarks, which would not otherwise have been published, on the theory of Prof. Planck." Jeans then criticises Planck's concept of the "entropy of a single resonator" given by the formula $S = k_\mathrm{B} \log W + \text{constant}$ by saying: "The function W, as at present defined, seems to me to have no meaning. Planck (in common, I know, with many other physicists) speaks of the 'probability' of an event, without specifying the basis according to which the probability is measured. This conception of probability seems to me an inexact conception, and as such to have no place in mathematical analysis." [41]

Jeans' critique of Planck's derivation is fully justified as one can infer from Einstein's "laudatio" for Planck written in 1913: "This [that is Planck's] calculation which, due to the not sufficiently sharp definition of W, could not be performed without arbitrariness, led to the radiation formula (1.21) . . . " [23].

Jeans then continues [41] by criticising Planck's introduction of his famous constant h via the fundamental relation $\varepsilon = h\nu$. "Here ε is a small quantity, a sort of indivisible atom of energy, introduced to simplify the calculations. We may legitimately remove this artificial quantity by passing to the limit in which $\varepsilon = 0$ . . . The relation $\varepsilon = h\nu$ is assumed by Planck in order that the law ultimately obtained may satisfy Wien's 'displacement law' i.e. may be of the form (1.17). This law is obtained by Wien from thermodynamical considerations on the supposition that the energy of the ether is in statistical equilibrium with that of matter at a uniform temperature. The method of statistical mechanics, however, enables us to go further and determine the form of the function $f(\nu/T)$; it is found to be $8\pi k_\mathrm{B}(T/\nu)$, so that Wien's law (1.17) reduces to the law given by expression (1.27). In other words, Wien's law directs us to take $\varepsilon = h\nu$, but leaves h indeterminate, whereas statistical mechanics gives us the further information that the true value of h is $h = 0$. Indeed, this is sufficiently obvious from general principles. The only way of eliminating the arbitrary quantity ε is by taking $\varepsilon = 0$, and this is the same as $h = 0$. – Thus it comes about that in Planck's final law (1.21) the value of h is left indeterminate; on putting $h = 0$, the value assigned to it by statistical mechanics, we arrive at once at the law (1.27). . . . I carry the method further than Planck, since Planck stops short of the step of putting $h = 0$. I venture to express the opinion that it is not legitimate to stop short at this point, as the hypotheses upon which Planck has worked lead to the relation $h = 0$ as a necessary consequence. Of course, I am aware that Planck's law is in good agreement with experiment if h is given a value different from zero, while my own law, obtained by putting $h = 0$, cannot possibly agree with experiment. This does not alter my belief that the value $h = 0$ is the only value which it is possible to take." [41]

Although Jeans' conclusion [41] that Planck should have arrived at the radiation law (1.27) instead of his formula (1.21) agrees with the conclusions drawn earlier by Einstein [36] and Rayleigh [37, 38]; his belief that the value $h = 0$ is the only value which Planck's constant can possibly take shows that he did not realize the importance of the equation $\varepsilon = h\nu$ (neither did Planck nor Rayleigh!). It was Einstein's revolutionary light-quantum paper [36] (the only contribution he himself called revolutionary) which gave a deep meaning to this equation and thus paved the way towards a quantum theory. Einstein put forward the following "heuristic view" [36]. "Monochromatic radiation of low density (within the domain of validity of Wien's radiation formula) behaves in thermodynamic respect as if it would consist of mutually independent energy quanta of magnitude $R\beta\nu/N_\text{A}$ [$\equiv h\nu$ using $\beta = h/k_\text{B}$]. – If, in regard to the volume dependence of the entropy, monochromatic radiation (of sufficiently low density) behaves as a discontinuous medium, which consists of energy quanta of magnitude [$h\nu$], then this suggests an inquiry as to whether the laws of the generation and conservation of light are also constituted as if light were to consist of energy quanta of this kind."

1.2.8
From Acoustics to Weyl's Law and Kac's Question

In the previous sections we have discussed how the heat radiation problem was at the origin of Weyl's famous work. Furthermore, we have seen that the idea of expressing the spectral energy density $\varrho(\nu, T)$ of the black body radiation in terms of the density of states $D(\nu)$ goes back to Rayleigh [26] who in turn reduced the problem to the "vibrations of a cubical mass of air". Thus Weyl's law actually has its roots in acoustics. In view of the fact that Rayleigh was a leading expert in acoustics and the author of the famous book "The Theory of Sound" [27], first published in 1877, it is not surprising that he realized that the radiation problem can be related to the number of vibrational modes of a black body cavity. All the more reason that it is strange to observe that he had difficulties in obtaining the correct value for the constant c_1 in his radiation law (1.19). The problem was of course a question of the correct boundary conditions in the electromagnetic case. In his book, Rayleigh writes: "Some of the natural notes of the air contained within a room may generally be detected on singing the scale. Probably it is somewhat in this way that blind people are able to estimate the size of rooms." [27] And in a footnote he adds: "A remarkable instance is quoted in Young's *Natural Philosophy*, II. p. 272, from Darwin's *Zoonomia*, II. 487. "The late blind Justice Fielding walked for the first time into my room, when he once visited me, and after speaking a few words said, 'This room is about 22 feet long, 18 wide, and 12 high'; all which he guessed by the ear with great accuracy." And then Rayleigh continues: "In long and narrow passages the vibrations parallel to the length are too slow to affect the ear, but notes due to transverse vibrations may often be heard. The relative proportions of the various overtones depend upon the place at which the disturbance is created. In some cases of this kind the pitch of the vibrations, whose direction is principally transverse, is influenced by the occurrence of longitudinal motion. . . . "

These remarks on acoustics lead us directly to Kac's famous question: "Can one hear the shape of a drum?" [9], which will be discussed in Sections 1.3.6 and 1.3.7, and the more general question: "Can one hear the periodic orbits of a drum?" to be discussed in Section 1.3.7.

1.3 Weyl's Law with Remainder Term. I

1.3.1 The Laplacian on the Flat Torus $\mathbb{T}^2$

In special cases it is possible to derive exact formulae for the counting function $N(\lambda)$ which contain in addition to the Weyl term (and possible higher order terms) an explicit expression for a remainder function. The most elegant way to derive these formulae is based on *trace formulae*; a famous example is the Selberg trace formula [45–47] to be discussed in Section 1.4. To illustrate the method in a simple case, we consider the eigenvalue problem $-\Delta_{\mathbb{T}^2}u = \lambda u$, where $\Delta_{\mathbb{T}^2}$ denotes the *Laplacian on a flat torus* $\mathbb{T}^2 := S^1_L \times S^1_L = \mathbb{R}^2/(L\mathbb{Z} \times L\mathbb{Z})$ characterized by a length scale $L > 0$. $\mathbb{T}^2$ can be represented by the fundamental domain $\Omega = [0, L] \times [0, L] \in \mathbb{R}^2$ i.e. by a square with side L, where opposite sides are glued together. Obviously, all of $\mathbb{R}^2$ is covered by the Γ-translates of Ω where Γ is the translation group $(L\mathbb{Z})^2$. This produces a tessellation of $\mathbb{R}^2$ and leads to the periodic boundary conditions

$$u(x_1 + \mu_1 L, x_2 + \mu_2 L) = u(x_1, x_2), \quad (x_1, x_2) \in \Omega, (\mu_1, \mu_2) \in \mathbb{Z}^2 \ .$$

Note that $\mathbb{T}^2$ is a smooth, compact manifold with area $|\Omega| = L^2$ (but with no boundary). It is easy to see that $(e_m)_{m\in\mathbb{Z}^2} = \left(e^{2\pi i(m\cdot x)/L}\right)_{m\in\mathbb{Z}^2}$ is an orthonormal basis of $L^2(\Omega)$ consisting of eigenvectors of $-\Delta_{\mathbb{T}^2}$ with discrete eigenvalues $(\lambda_m)_{m\in\mathbb{Z}^2} = \left((4\pi^2)/L^2\left(m_1^2 + m_2^2\right)\right)_{(m_1,m_2)\in\mathbb{Z}^2}$.

Let $r(n) = \#\left\{(m_1, m_2) \in \mathbb{Z}^2, n = m_1^2 + m_2^2\right\}$, $n \in \mathbb{N}_0$, with $r(0) = 1$, i.e. $r(n)$ denotes the number of representations of $n \in \mathbb{N}_0$ as a sum of two squares of integers. Obviously, the distinct eigenvalues of $-\Delta_{\mathbb{T}^2}$,

$$(\bar{\lambda}_n)_{n\in\mathbb{N}_0} = \left(\frac{4\pi^2}{|\Omega|}n\right)_{n\in\mathbb{N}_0} \ ,$$

occur with multiplicity $r(n)$. Then the counting function on the torus reads

$$N(\lambda) = \sum_{\bar{\lambda}_n \le \lambda} r(n) = \sum_{0\le n\le(|\Omega|/4\pi^2)\lambda} r(n) \ . \tag{1.29}$$

The very irregular ("valde irregulariter" [48]) number theoretical function $r(n)$ had already been studied by Gauss [48] who derived the formula $r(n) = 4(d_1(n) - d_3(n))$, $n \ge 1$, where $d_1(n)$ and $d_3(n)$ are the number of divisors of n of the form $4m + 1$ and $4m + 3$, $m \in \mathbb{N}_0$, respectively. The first values are $r(0) = 1$, $r(1) = 4$, $r(2) = 4$, $r(3) = 0$, $r(4) = 4$, $r(5) = 8$. If $n \equiv 3(\bmod 4)$ then $r(n) = 0$. For large n one has $r(n) = O(n^\varepsilon)$ for every $\varepsilon > 0$; $r(n) = O\left((\log n)^\delta\right)$ is false for every δ. The average order of $r(n)$ is

$$\bar{r} := \lim_{x\to\infty} \frac{1}{x} \sum_{0\le n\le x} r(n) = \pi$$

(Gauss resp. the Weyl law, see (1.31)). For further information about $r(n)$, see [49, pp. 241].

1.3.2 The Classical Circle Problem of Gauss

Let

$$\nu(x) := \sum_{0 \le n \le x} r(n) = \sum_{\substack{m_1^2+m_2^2 \le x \\ (m_1,m_2)\in\mathbb{Z}\times\mathbb{Z}}} 1 \;, \tag{1.30}$$

then

$$N(\lambda) = \nu\left(\frac{|\Omega|}{4\pi^2}\lambda\right)$$

and the derivation of Weyl's law is reduced to a *lattice point problem*, since $\nu(x)$ has a simple geometric interpretation as the number of lattice points in the interior and on the boundary of a circle with center $(0,0)$ and of radius $\sqrt{x}$. The problem of calculating the leading asymptotic behavior of $\nu(x)$ for $x \to \infty$ was already considered by Gauss in 1834 [48] (see also [50, pp. 32–39]). He realized that $\nu(x)$ is approximately given by the sum of the areas of all squares of unit side length which are inscribed in the circle of radius $\sqrt{x}$, and thus $\nu(x)$ is in first approximation equal to the area of the circle $\pi\left(\sqrt{x}\right)^2 = \pi x$. Actually, Gauss proved

$$\lim_{x\to\infty} \frac{\nu(x)}{x} = \pi \;, \tag{1.31}$$

which implies *Weyl's law*

$$\lim_{\lambda\to\infty} \frac{N(\lambda)}{\lambda} = \frac{|\Omega|}{4\pi} \tag{1.32}$$

for the counting function (1.29). Based on his result (1.31), Gauss considered $\nu(x)/x$ as an approximation method to calculate π. To this purpose, he thought about the error one makes at finite x. Again, by geometrical intuition, one sees that the error should not be larger than the combined area of those squares that are cut by the boundary of the circle i.e. those contained in an annulus of width $2\sqrt{2}$, and thus is approximately given by $2\sqrt{2}$ times the perimeter of the circle $2\pi\sqrt{x}$, and, indeed, Gauss was able to prove

$$\nu(x) = \pi x + O\left(\sqrt{x}\right) \quad (x \to \infty) \;,$$

which implies

$$N(\lambda) = \frac{|\Omega|}{4\pi}\lambda + O\left(\sqrt{\lambda}\right) \quad (\lambda \to \infty) \;.$$

Defining a remainder term $P(x)$,

$$\nu(x) = \pi x + P(x) \;,$$

we are led to the *classical circle problem*, a famous problem in analytic number theory [51, pp. 181–308]: estimate the remainder function $P(x)$ as accurately as possible. In particular, determine $\alpha_0 = \inf \alpha$ in the estimate

$$P(x) = O(x^{\alpha}) \quad (x \to \infty) \ .$$

In Figures 1.1 and 1.2 we show plots of $\nu(x)$ and $P(x)$, respectively, from which it becomes clear that $P(x)$ – due to the erratic behavior of $r(n)$ – is a very irregular function wildly fluctuating about zero. It is therefore no big surprise that to determine the actual size of $P(x)$, and thus the remainder to Weyl's law, is a difficult problem. Considering the difference $\nu(n + 1/2) - \nu(n)$, $n \in \mathbb{N}$, it is easy to see that $P(x) = o(1)$ is false, and thus $0 \leq \alpha_0 \leq 1/2$. An important result showing that $P(x)$ is much smaller than the classical result $\alpha_0 \leq 1/2$ is due to Sierpiński who proved $\alpha_0 \leq 1/3$ in 1906 [52, pp. 73–108]. A famous conjecture by Hardy from 1915 states that α_0 should be 1/4, i.e. $P(x) = O\left(x^{1/4+\varepsilon}\right)$ for every $\varepsilon > 0$ [53, 54]. Actually, Hardy proved $\alpha_0 \geq 1/4$.

During the last 100 years, the values for α_0 decreased only by a tiny amount: $\alpha_0 \leq 37/112 = 0.330\ldots$ (van der Corput 1923 [55]), $\alpha_0 \leq 12/37 = 0.324\ldots$ (Wen-Lin Yin 1962 [56]), $\alpha_0 \leq 7/22 = 0.318\ldots$ (Iwaniec and Mozzochi 1988 [57]). The best bound known today is due to Huxley who proved in 1992 that

$$P(x) = O\Big(x^{23/73} \left(\log x\right)^{315/146}\Big) \ ;$$

note that $23/73 = 0.315\ldots$ is still far away from 1/4! (For a review, see [58].) Since $P(x)$ is a wildly fluctuating function, it might be that some very rare spikes exceeding the conjectured $x^{1/4}$-behavior make it extremely difficult to improve the best existing bound. In order to "tame" these spikes, one can consider moments of $P(x)$ and hope that the spikes are being washed out. We shall come back to this idea in Section 1.3.9 making use of the trace formula for $\nu(x)$ which we shall now derive.

Note added in proof: in a recent unpublished paper [59] it is claimed to present a proof of Hardy's conjecture.

1.3.3 The Formula of Hardy–Landau–Voronoï

The counting function $\nu(x)$ can be rewritten as

$$\nu(x) = \sum_{m \in \mathbb{Z}^2} \theta\left(x - m^2\right) ,$$

where $\theta(x)$ denotes the Heaviside step function. Instead of $\theta(x - m^2)$, let us consider a function $g(m)$ with

- $g : \mathbb{R}^2 \to \mathbb{C}$, continuous
- $g(x) = O\Big(1/(\|x\|^{2+\varepsilon})\Big)$ for $\|x\|^2 = x_1^2 + x_2^2 \to \infty$, $\varepsilon > 0$,

and let us study the sum $\sum\limits_{m \in \mathbb{Z}^2} g(m)$. Using the *Poisson summation formula*, we obtain

$$\sum_{m \in \mathbb{Z}^2} g(m) = \sum_{l \in \mathbb{Z}^2} \tilde{g}(l) \ , \tag{1.33}$$

where $\tilde{g}$ denotes the Fourier transform of g:

$$\tilde{g}(l) = \int\limits_{\mathbb{R}^2} g(x) e^{-2\pi i (l\cdot x)} \, d^2 x \, . \tag{1.34}$$

To apply this to the circle problem, we make the further assumption that $g(x)$ is a radial function which depends only on $\varrho = \|x\|$, i.e.

$$g(x) = g(x_1, x_2) = \phi\left(x_1^2 + x_2^2\right) = \phi\left(\varrho^2\right) .$$

Thus

$$\begin{aligned}\tilde{g}(l) = \tilde{g}(l_1, l_2) &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} \phi\left(x_1^2 + x_2^2\right) e^{-2\pi i (l_1 x_1 + l_2 x_2)} \, dx_1 \, dx_2 \\ &= \int\limits_0^{\infty} \varrho \phi\left(\varrho^2\right) \int\limits_0^{2\pi} e^{-2\pi i \|l\| \varrho \cos\varphi} \, d\varphi \, d\varrho = 2\pi \int\limits_0^{\infty} \varrho \phi\left(\varrho^2\right) J_0(2\pi \|l\| \varrho) \, d\varrho .\end{aligned}$$

Here we have introduced polar coordinates in $\mathbb{R}^2$, $x_1 = \varrho \cos\varphi$, $x_2 = \varrho \sin\varphi$, $0 \le \varphi \le 2\pi$, and have used the integral representation

$$J_0(z) = \frac{1}{2\pi} \int\limits_0^{2\pi} e^{-iz\cos\varphi} \, d\varphi$$

for the Bessel function $J_0(z)$. Now the Poisson summation formula (1.33) reads

$$\sum_{m\in\mathbb{Z}^2} \phi\left(m^2\right) = 2\pi \sum_{l\in\mathbb{Z}^2} \int\limits_0^{\infty} \varrho \phi\left(\varrho^2\right) J_0(2\pi \|l\| \varrho) \, d\varrho \ ,$$

or, by introducing the multiplicity $r(n)$ and $\varrho = \sqrt{x}$, $x \ge 0$:

$$\sum_{n=0}^{\infty} r(n) \phi(n) = \pi \sum_{n=0}^{\infty} r(n) \int\limits_0^{\infty} \phi(x) J_0\left(2\pi \sqrt{nx}\right) dx \, . \tag{1.35}$$

This is the *theorem due to Hardy* [54, 60, 61], *Landau* [51, pp. 189] and *Voronoï* [62].

1.3.4
The Trace Formula on the Torus $\mathbb{T}^2$ and the Leading Weyl Term

We recall that the distinct eigenvalues on the torus $\mathbb{T}^2$ are given by $\bar{\lambda}_n = (2\pi/L)^2 \, n = p_n^2$ with $p_n := (2\pi/L)\sqrt{n}$, $n \in \mathbb{N}_0$, and multiplicities $r(n)$. Introducing in the theorem (1.35) the *spectral function* $h((2\pi/L)\,\varrho) := \phi\left(\varrho^2\right)$ with

- $h : \mathbb{R} \to \mathbb{C}$, continuous
- h even i.e. $h(-p) = h(p)$ (1.36)
- $h(p) = O\left(\dfrac{1}{|p|^{2+\varepsilon}}\right)$, $|p| \to \infty$, $\varepsilon > 0$,

we arrive at the *trace formula on the torus* $\mathbb{T}^2$

$$\sum_{n=0}^{\infty} r(n)h(p_n) = \frac{|\Omega|}{2\pi}\int_0^{\infty} ph(p)\,\mathrm{d}p + |\Omega| \sum_{n=1}^{\infty} r(n)\hat{h}\left(L\sqrt{n}\right) , \tag{1.37}$$

where $\hat{h}(x)$ denotes the Fourier–Bessel (or Hankel) transform of $h(p)$:

$$\hat{h}(x) := \frac{1}{2\pi}\int_0^{\infty} ph(p)J_0(px)\,\mathrm{d}p .$$

(In deriving the first term on the right-hand side of (1.37), we have used $r(0) = 1 = J_0(0)$ and $L^2 = |\Omega|$.) Note that the left-hand side of (1.37) can be written as the trace of the trace class operator

$$h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right) : L^2(\Omega) \to L^2(\Omega)$$

with

$$h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right)f = \sum_{m\in\mathbb{Z}^2} h\left(\sqrt{\lambda_m}\right)(e_m \mid f)\, e_m , \quad \textit{for} \quad f \in L^2(\Omega) ,$$

i.e.

$$\sum_{n=0}^{\infty} r(n)h(p_n) = \mathrm{Tr}\, h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right) ,$$

which explains why (1.37) is called a trace formula.

Due to the conditions (1.36) on the spectral function $h(p)$, the operator $h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right)$ is actually a Hilbert–Schmidt operator with kernel $G_h(x, y) = \bar{G}_h(y, x) \in L^2(\Omega \times \Omega)$ satisfying, for $f \in L^2(\Omega)$,

$$\left(h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right)f\right)(x) = \int_{\Omega} G_h(x, y) f(y)\,\mathrm{d}^2 y . \tag{1.38}$$

Furthermore, $G_h(x, y)$ has the uniformly convergent expression in terms of the orthonormal eigenfunctions $e_m \in L^2(\Omega)$ (Mercer's theorem)

$$G_h(x, y) = \sum_{m\in\mathbb{Z}^2} h\left(\sqrt{\lambda_m}\right) e_m(x)\bar{e}_m(y) , \tag{1.39}$$

which expresses the fact that e_m is an eigenfunction of the operator $h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right)$ with eigenvalue $h\left(\sqrt{\lambda_m}\right)$. From this one immediately derives the *pre-trace formula*

$$\mathrm{Tr}\, h\left((-\Delta_{\mathbb{T}^2})^{1/2}\right) = \int_{\Omega} G_h(x, x)\,\mathrm{d}^2 x . \tag{1.40}$$

Pre-trace formulae of this type are the starting point for the derivation of trace formulae in the general case, for example in quantum mechanics, where the right-hand side of (1.40) is expressed by the volume of the classical phase space and the classical actions evaluated along the periodic orbits of the corresponding classical system [63].

An alternative way to write the left-hand side of (1.37) is

$$\sum_{n=0}^{\infty} r(n)h(p_n) = \int_0^{\infty} h\left(\sqrt{\lambda}\right) \mathrm{d}N(\lambda) \ ,$$

where $N(\lambda)$ is the counting function, and the integral is understood as a Stieltjes integral. Rewriting in a similar way the first term on the right-hand side of (1.37)

$$\frac{|\Omega|}{2\pi} \int_0^{\infty} p h(p) \, \mathrm{d}p =: \int_0^{\infty} h\left(\sqrt{\lambda}\right) \mathrm{d}\overline{N}(\lambda) \ ,$$

one obtains $\mathrm{d}\overline{N}(\lambda) = |\Omega| / (4\pi) \, \mathrm{d}\lambda$ and thus, immediately, the smooth term

$$\overline{N}(\lambda) = \frac{|\Omega|}{4\pi} \lambda \ , \tag{1.41}$$

which turns out to be exactly the leading *Weyl term* of $N(\lambda)$, see (1.4) and Section 1.3.6.

1.3.5 Spectral Geometry: Interpretation of the Trace Formula on the Torus $\mathbb{T}^2$ in Terms of Periodic Orbits

While the left-hand side of the trace formula (1.37) has a simple *spectral* interpretation (being just the sum over the "frequencies" $p_n = \sqrt{\lambda_n}$ of the eigenvibrations on $\mathbb{T}^2$, evaluated on a large class of spectral functions $h(p)$, see Equation (1.36)), the infinite series on the right-hand side has a simple *geometrical* interpretation as can be seen by rewriting (1.37) as follows

$$\sum_{n=0}^{\infty} r(n)h(p_n) = |\Omega| \, \hat{h}(0) + |\Omega| \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} r\left(k^2 l_n^2\right) \hat{h}\,(k l_n) \ . \tag{1.42}$$

Here $\{l_n\}_{n \in \mathbb{N}}$ denotes the *primitive length spectrum* on $\mathbb{T}^2$ with

$$l_n = L \sqrt{m_1^2 + m_2^2} = L \sqrt{n} \ ,$$

where n is a square-free integer with $r(n) \neq 0$. l_n is the geometrical length of a primitive periodic orbit (closed geodesic) of the classical geodesic flow on $\mathbb{T}^2$. The non-primitive periodic orbits have lengths $k l_n$, $k \geq 2$, where k counts the k^{th} traversal of the corresponding primitive periodic orbit with length l_n. The trace formula (1.42) displays a beautiful relation in *spectral geometry* relating the *spectrum of the Laplacian* to the *length spectrum of the geodesic flow*. The torus $\mathbb{T}^2$ is a compact Riemann surface of genus 1 and Gaussian curvature $K = 0$. A generalization to surfaces of higher genus is given by the famous Selberg trace formula [45, 46] which has been much studied in the field of quantum chaos (see for example [47, 64–66]) and string theory (see for example [67, 68]) and will be discussed in Section 1.4.4.

1.3.6
The Trace of the Heat Kernel on *d*-Dimensional Tori and Weyl's Law

The trace formula (1.37), respectively (1.42), has the typical structure of a trace formula and is in some sense a "meta formula" since it allows one to derive an infinite number of relations depending on the special choice of the spectral function $h(p)$ satisfying the conditions (1.36). As a first example, let us calculate the *trace of the heat kernel*, which is obtained for the choice $h(p) = e^{-p^2 t}$, $t > 0$. With

$$\hat{h}(x) = \frac{1}{2\pi}\int_0^\infty p e^{-p^2 t} J_0(px)\,\mathrm{d}p = \frac{1}{4\pi t} e^{-x^2/4t}$$

we get ($t > 0$)

$$\begin{aligned}\Theta_{\mathbb{T}^2}(t) := \operatorname{Tr} e^{t\Delta_{\mathbb{T}^2}} &= \sum_{n=0}^\infty r(n) e^{-\left(4\pi^2/|\Omega|\right)nt} = \frac{|\Omega|}{4\pi t} + \frac{|\Omega|}{4\pi t}\sum_{n=1}^\infty r(n) e^{-(|\Omega|/4)tn} \\ &= \frac{|\Omega|}{4\pi t} + \frac{|\Omega|}{4\pi t}\sum_{n=1}^\infty\sum_{k=1}^\infty r\left(k^2 l_n^2\right) e^{-k^2 l_n^2/4t} .\end{aligned} \tag{1.43}$$

For $t \to 0^+$ one thus obtains the correct $|\Omega|/(4\pi t)$-term (and no higher order terms of the type $\sum\limits_{n=-1}^{\infty} a_n t^{n/2}$ as occurring in the general case), which yields the correct Weyl term, and an exponentially small remainder term behaving as $O\left(t^{-1}e^{-L^2/4t}\right)$. It is thus seen that the Weyl term corresponds to the "zero-length contribution" in the periodic orbit sum i.e. to the term obtained for $l_0 := 0$, while the exponential remainder term is determined by the shortest primitive periodic orbit on $\mathbb{T}^2$ having the length $l_1 = L$. As to physical applications, let us point out that the function $\Theta_{\mathbb{T}^2}(t)$ is for $t \sim 1/T$, where T denotes absolute temperature, identical to the *partition function in statistical mechanics*, and thus the Weyl term determines the high-temperature limit of the corresponding thermodynamical system.

Note that the trace of the heat kernel rewritten as $f(q) := \sum\limits_{n=0}^{\infty} r(n) q^n$ with $q = e^{i\pi\tau}$, $\tau = i\,(4\pi/|\Omega|)\,t$, plays the role of a generating function of the arithmetic function $r(n)$. $f(q)$ was already introduced by Jacobi in 1829 who derived

$$f(q) = \left(\sum_{m\in\mathbb{Z}} q^{m^2}\right)^2 = (\theta_3(0|\tau))^2$$

for $\operatorname{Im}\tau > 0$ in terms of the elliptic theta function θ_3. Using the transformation formula $\theta_3(0 \mid \tau) = (-i\tau)^{-1/2}\theta_3(0 \mid -1/\tau)$ derived by Poisson in 1823, one obtains again relation (1.43).

It is not difficult to generalize the result (1.43) to d-dimensional flat tori $\mathbb{T}^d := \mathbb{R}^d/\Gamma$ with $\Gamma = (L\mathbb{Z})^d$. The translation group Γ has a natural embedding as a lattice in $\mathbb{R}^d$. To Γ there is associated a uniquely defined dual lattice Γ^* (called a reciprocal lattice in physics): $\Gamma^* = \{\gamma^* \in \mathbb{R}^n : \gamma \cdot \gamma^* \in \mathbb{Z} \text{ for all } \gamma \in \Gamma\}$. With $\gamma = Ln$, $n \in \mathbb{Z}^d$, $\gamma^* = 1/Lm$, $m \in \mathbb{Z}^d$, the eigenvalues of $-\Delta_{\mathbb{T}^d}$ are given by $\left(\lambda_{\gamma^*}\right)_{\gamma^*\in\Gamma^*} = 4\pi^2 \left\|\gamma^*\right\|^2$ with eigenvectors $\left(e_{\gamma^*}\right)_{\gamma^*\in\Gamma^*} = \left(e^{2\pi i(\gamma^*\cdot x)}\right)$. Furthermore, the length spectrum of the classical periodic orbits on $\mathbb{T}^d$ is given by $\left(\|\gamma\|\right)_{\gamma\in\Gamma}$. Using the Poisson summation formula as in the case $d = 2$, it is straightforward to derive a trace formula on $\mathbb{T}^d$ from which one obtains, for example, for the trace of the heat kernel ($t > 0$)

$$\begin{aligned}\Theta_{\mathbb{T}^d}(t) &:= \operatorname{Tr} e^{t\Delta_{\mathbb{T}^d}} = \sum_{\gamma^*\in\Gamma^*} e^{-4\pi^2\|\gamma^*\|^2 t} = \frac{|\Omega|}{(4\pi t)^{d/2}} \sum_{\gamma\in\Gamma} e^{-\|\gamma\|^2/4t} \\ &= \frac{|\Omega|}{(4\pi t)^{d/2}} + O\left(t^{-d/2}e^{-L^2/4t}\right) \qquad (t \to 0^+). \end{aligned} \tag{1.44}$$

Here the first term on the right-hand side corresponding to the identity element $I \in \Gamma$ with $\|I\| = 0$ yields via the Tauberian theorem of Karamata (see Theorem 1.1 in Section 1.6) *Weyl's law for* $\mathbb{T}^d$ ($\lambda \to \infty$)

$$N(\lambda) = \frac{|\Omega|}{(4\pi)^{d/2}\Gamma(1 + d/2)}\lambda^{d/2} + O\left(\lambda^{d/2}\right), \tag{1.45}$$

but the trace formula yields, in addition, an exact expression for the remainder term in the same way as discussed in detail for $\mathbb{T}^2$ in Section 1.3.9 below.

The case $d = 3$ has important applications in several fields. For example, in solid state physics, chemistry and crystallography, one identifies the lattice Γ with the atomic structure of crystals. Furthermore, the reciprocal lattice Γ^* is very useful in analyzing diffraction phenomena in light and neutron scattering off crystals. In cosmology it has been proposed that the spatial section of our Universe is given by a 3-torus whose fundamental domain is a cube with side length $L \simeq 5 \times 10^{26}$ m $\simeq 5.6 \times 10^{10}$ light years [69].

Finally we would like to mention that the case $d = 16$, i.e. the tori $\mathbb{R}^{16}/\mathbb{Z}^{16}$ have played an important role in the attempts to answer Kac's question [9], since it had already been noticed by John Milnor in 1964 that the tori $\mathbb{T}^{16}$ give examples of nonisometric compact manifolds with the same spectrum of the Laplacian [70]. The construction of these lattices for $d = 16$ had already been found by Witt in 1941 [71].

1.3.7 Going Beyond Weyl's Law: One can Hear the Periodic Orbits of the Geodesic Flow on the Torus $\mathbb{T}^2$

Let us consider another admissible spectral function $h(p)$ in the trace formula (1.42) which is slightly more general than the one used in the previous section for the heat kernel:

$$h(p) := J_0(ps)e^{-p^2 t}\,, \quad s \in \mathbb{R}, t > 0\,.$$

With

$$\hat{h}(x) = \frac{1}{2\pi}\int_0^\infty pJ_0(ps)e^{-p^2 t}J_0(px)\,\mathrm{d}p = \frac{1}{4\pi t}e^{-(s^2+x^2)/4t}I_0\left(\frac{sx}{2t}\right) \tag{1.46}$$

($I_0(z)$ is the modified Bessel function) we arrive at the trace formula ($s \in \mathbb{R}, t > 0$)

$$\begin{aligned} G(s,t) &:= \mathrm{Tr}\left(J_0\left(s(-\Delta_{\mathbb{T}^2})^{1/2}\right)e^{t\Delta_{\mathbb{T}^2}}\right) = \sum_{n=0}^\infty r(n)J_0\left(s\sqrt{\bar{\lambda}_n}\right)e^{-\bar{\lambda}_n t} \\ &= \frac{|\Omega|}{4\pi t}e^{-s^2/4t} + \frac{|\Omega|}{4\pi t}\sum_{n=1}^\infty\sum_{k=1}^\infty r\left(k^2 l_n^2\right)e^{-(s^2+k^2l_n^2)/4t}I_0\left(\frac{skl_n}{2t}\right). \end{aligned} \tag{1.47}$$

Since $I_0(0) = 1$, it follows that (1.47) coincides in the limit $s \to 0$ with the trace of the heat kernel (1.43), $G(0,t) = \Theta_{\mathbb{T}^2}(t)$. Performing on the other hand for fixed $s > 0$ the limit $t \to 0^+$ i.e. eliminating the "regulator" t, one obtains the remarkable relation ($s > 0$)

$$G(s,0) = \sum_{n=0}^\infty r(n)J_0\left(s\sqrt{\bar{\lambda}_n}\right) = \frac{|\Omega|}{2\pi}\sum_{n=1}^\infty\sum_{k=1}^\infty \frac{r\left(k^2l_n^2\right)}{kl_n}\delta(s-kl_n)\,, \tag{1.48}$$

which is to be understood as an identity in the sense of distributions. Here we have used the asymptotic expansion (valid for $z \to +\infty$)

$$I_0(z) = \frac{1}{\sqrt{2\pi z}}e^z\left(1 + O\left(\frac{1}{z}\right)\right)$$

and the delta-sequence

$$\frac{1}{2\sqrt{\pi t}}e^{-x^2/4t} \to \delta(x) \quad (t \to 0^+)\,.$$

Relation (1.48) tells us that the formal trace $G(s,0) = \mathrm{Tr}\,J_0\left(s(-\Delta_{\mathbb{T}_2})^{1/2}\right)$ yields a well-defined distribution whose singular support is given for $s > 0$ by

$$\text{singsupp}\,G(s,0) = \{kl_n\}\,, k \in \mathbb{N}\,,$$

i.e. by the primitive length spectrum $\{l_n\}$ of the geodesic flow on the torus and the nonprimitive length spectrum $\{kl_n\}\,, k \geq 2$. Thus the eigenvalues $\left\{\bar{\lambda}_n\right\}$ of the

Laplacian on $\mathbb{T}^2$ together with their multiplicities $\{r(n)\}$ "know" the length spectrum of the closed geodesics of the classical motion on $\mathbb{T}^2$, i.e. one can hear the periodic orbits of the torus! Since the torus is uniquely given by its area $|\Omega|$ and its length spectrum $\{l_n\}$, we can conclude that the complete shape of the torus is audible.

A slightly different operator has been studied by Chazarain [72], Colin de Verdière [73, 74], and Duistermaat and Guillemin [75, 76], where the Bessel function J_0 is replaced by $\cos\left(s(-\Delta)^{1/2}\right)$ respectively $\exp\left(is(-\Delta)^{1/2}\right)$.

1.3.8
The Spectral Zeta Function on the Torus $\mathbb{T}^2$

Define for $s \in \mathbb{C}$, $\operatorname{Re} s > 1$, the *spectral zeta function* on $\mathbb{T}^2$:

$$\zeta_{\mathbb{T}^2}(s) := \operatorname{Tr}'\left(-\Delta_{\mathbb{T}^2}\right)^{-s} = \sum_{n=1}^{\infty} \frac{r(n)}{\bar{\lambda}_n^s} = \frac{|\Omega|^s}{(2\pi)^{2s}} \sum_{n=1}^{\infty} \frac{r(n)}{n^s} , \tag{1.49}$$

where the prime at the trace denotes that the eigenvalue $\bar{\lambda}_0 = 0$ has been omitted. (Zeta functions of this type for general Laplace–Beltrami operators were introduced in [77, 78] following a suggestion of Weyl. See also [14].) With the help of

$$\frac{1}{n^s} = \frac{1}{\Gamma(s)} \int_0^{\infty} \tau^{s-1} e^{-n\tau}\, d\tau, \quad n > 0, \operatorname{Re} s > 0 ,$$

we obtain for $\operatorname{Re} s > 1$

$$\Gamma(s)\zeta_{\mathbb{T}^2}(s) = \int_0^1 t^{s-1}\left[\Theta_{\mathbb{T}^2}(t) - 1\right] dt + \int_1^{\infty} t^{s-1}\left[\Theta_{\mathbb{T}^2}(t) - 1\right] dt . \tag{1.50}$$

Hence $\zeta_{\mathbb{T}^2}(s)$ is the Mellin transform of $\Theta_{\mathbb{T}^2}(t)$ with the eigenvalue zero omitted. Since

$$\Theta_{\mathbb{T}^2}(t) = 1 + O\left(e^{-\left(4\pi^2/|\Omega|\right)t}\right) \quad \text{for} \quad t \to \infty ,$$

the second integral has an analytic continuation to the whole complex s-plane as an entire function. Inserting in the first integral for $\Theta_{\mathbb{T}^2}(t)$ the expansion (1.43), we obtain for $\operatorname{Re} s > 1$

$$\zeta_{\mathbb{T}^2}(s) = \frac{|\Omega|/(4\pi)}{s-1} + F(s) , \tag{1.51}$$

where $F(s)$ is an entire function. Thus we can extend the Dirichlet series (1.49) meromorphically to all $s \in \mathbb{C}$ having only one simple pole at $s = 1$ with residue $|\Omega|/(4\pi)$, which is given by the area of the torus. This pole is a direct consequence of the leading Weyl term in the expansion (1.43). It thus follows that the Dirichlet series $\sum\limits_{n=1}^{\infty} r(n)/n^s$ diverges for $\operatorname{Re} s \leq 1$, but is convergent for $\operatorname{Re} s > 1$, which will be important in the explicit formula for the remainder term in Weyl's law. Note that there exists the following closed expression, which can be considered as another generating function of $r(n)$ (see for example [79, pp. 265])

$$\sum_{n=1}^{\infty} \frac{r(n)}{n^s} = 4\zeta(s)L(s)$$

in terms of the Riemann zeta function $\zeta(s)$ and the Dirichlet L-series $L(s) := 1 - 1/3^s + 1/5^s - \ldots$ with $L(1) = \pi/4$, which has an entire extension.

The result (1.51) holds in general for a large class of eigenvalue problems; see for example reference [47] for the Laplace–Beltrami operator on compact Riemannian surfaces of genus $g \geq 2$. In the case of the Dirichlet Laplacian acting on a smooth bounded open set $\Omega \subset \mathbb{R}^d$ one can show [80] that $\zeta_\Omega(s) := \mathrm{Tr}\left(-\Delta_\Omega^D\right)^{-s}$ possesses a meromorphic analytic continuation into the complex s-place with a leading simple pole at $s = d/2$ and residue $|\Omega| / \left((4\pi)^{d/2}\,\Gamma(d/2)\right)$. In particular, $s = 0$ turns out to be a regular point such that the first derivative at $s = 0$, $\zeta'_\Omega(0)$, is well defined. This fact is then used to regularize the *functional determinant* of $-\Delta_\Omega^D$ by [80]

$$\det\left(-\Delta_\Omega^D\right) := \exp\left(-\zeta'_\Omega(0)\right) .$$

This method was introduced into physics by Stephen Hawking [81] as a convenient way to compute the determinants arising in the Feynman path integral approach to quantum field theory and quantum gravity. For applications of this method, see for example [82, pp. 37–43] in the case of quantum mechanics, and [67] in the case of string theory.

1.3.9 An Explicit Formula for the Remainder Term in Weyl's Law on the Torus $\mathbb{T}^2$ and for the Circle Problem

To derive $N(\lambda)$ from the trace formula (1.37), we choose the function $h(p) = \theta\left(\lambda - p^2\right)$, $\lambda > 0$. We then obtain with

$$\frac{1}{2\pi}\int_0^\infty p h(p)\,\mathrm{d}p = \frac{1}{2\pi}\int_0^{\sqrt{\lambda}} p\,\mathrm{d}p = \frac{\lambda}{4\pi}$$

and

$$\hat{h}(x) = \frac{1}{2\pi}\int_0^{\sqrt{\lambda}} p J_0(px)\,\mathrm{d}p = \frac{\sqrt{\lambda}}{2\pi x} J_1\left(\sqrt{\lambda}x\right)$$

the relation

$$N(\lambda) = \frac{|\Omega|}{4\pi}\lambda + \frac{L}{2\pi}\sqrt{\lambda}\sum_{n=1}^{\infty}\frac{r(n)}{\sqrt{n}} J_1\left(L\sqrt{n\lambda}\right) . \tag{1.52}$$

This equation was found for the first time in Hardy's paper [54] who writes in a footnote: "The form of this equation was suggested to me by Mr. S. Ramanujan, ...". (As we shall see below, the sum in (1.52) is not absolutely convergent since the function $h(p)$ used in the derivation is not continuous. Relation (1.52) can be derived, however, by using an appropriate smoothing [65, 83].)

In order to study the asymptotic behavior of the remainder term, we employ the asymptotic formula

$$J_1(x) = \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{3\pi}{4}\right) + O\left(\frac{1}{x^{3/2}}\right) \quad (x \to \infty) ,$$

and obtain in the limit $\lambda \to \infty$

$$N_{\text{fl}}(\lambda) = \lambda^{1/4} \frac{1}{\pi} \sqrt{\frac{L}{2\pi}} \sum_{n=1}^{\infty} \frac{r(n)}{n^{3/4}} \cos\left(L\sqrt{\lambda n} - \frac{3\pi}{4}\right) + O\left(\frac{1}{\lambda^{1/4}} \sum_{n=1}^{\infty} \frac{r(n)}{n^{5/4}}\right) , \tag{1.53}$$

where we have defined the "*fluctuating part*" of the counting function by $N_{\text{fl}}(\lambda) := N(\lambda) - (|\Omega| / (4\pi))\,\lambda$. $N_{\text{fl}}(\lambda)$ describes the fluctuations of $N(\lambda)$ about the mean behavior $\overline{N}(\lambda) := (|\Omega| / (4\pi))\,\lambda$ given by Weyl's law, see (1.41). In Figure 1.1 we show a plot of $N(\lambda)$ for $L = 2\pi$ (which implies $N(\lambda) = \nu(\lambda)$ and $P(\lambda) = N_{\text{fl}}(\lambda)$ for the remainder term in Gauss' circle problem) for small values of $\lambda \equiv x$ $(0 \le x \le 50)$. Weyl's law is indicated as a straight line. One observes that the Weyl term does indeed describe the mean behavior of the staircase function $\nu(x)$ very well, even at small values of x. The fluctuating part $P(x)$ is shown in Figure 1.2 for small values $(0 \le x \le 50)$ and for large values $(10^{11} \le x \le 10^{11} + 10^7)$ of x and shows a very erratic behavior fluctuating about zero. In order to understand this behavior, one has to study the series in (1.53), which is a trigonometric series and therefore more difficult to control than a Fourier series. (Since $\sum_1^{\infty} r(n)/n^{5/4} < \infty$, see Section 1.3.8, the second term in (1.53) is bounded by $\lambda^{-1/4}$, and thus can be neglected.) Due to the divergence of the sum $\sum_1^{\infty} r(n)/n^{3/4}$, the trigonometric sum is only conditionally convergent, explaining the difficulty in proving Hardy's conjecture which amounts to the bound $O(\lambda^{\varepsilon})$ for every $\varepsilon > 0$ for this sum. (It is possible, however, to replace the sharp counting function $N(\lambda)$ by a smooth counting function depending on a smoothness parameter which leads to better convergence properties, see [65] and [83].)

In order to quantify the numerical observation that $N_{\text{fl}}(\lambda)$ oscillates about zero, let us calculate the mean value of $P(x)$ $(= N_{\text{fl}}(x)$ for $L = 2\pi)$:

$$\bar{P}(x) := \frac{1}{x} \int_0^x P(y)\, \mathrm{d}y .$$

We then obtain from (1.52) using

$$\int_0^x \sqrt{y} J_1\left(2\pi\sqrt{ny}\right) \mathrm{d}y = \frac{x}{\pi\sqrt{n}} J_2\left(2\pi\sqrt{nx}\right)$$

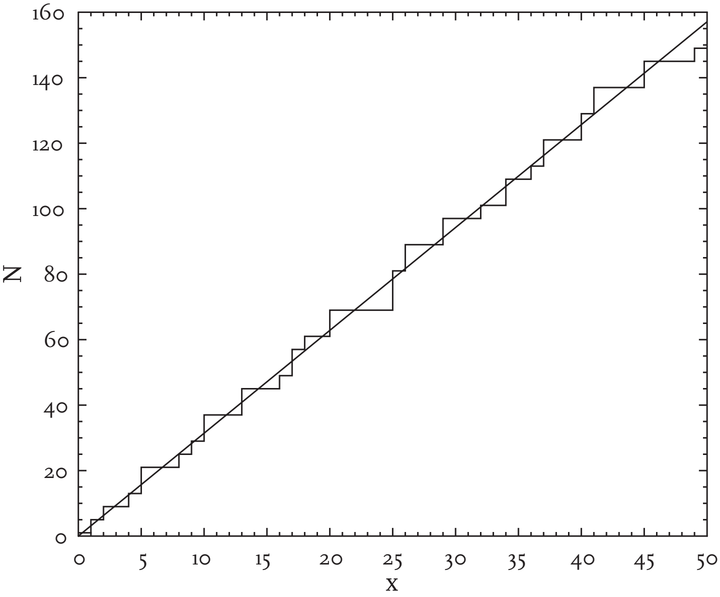


Figure 1.1 The counting function $\nu(x)$ for the Gaussian circle problem (respectively for a torus with $L = 2\pi$). The straight line shows the leading term πx (Weyl's law).

and the asymptotics of the Bessel function ($x \to \infty$)

$$\begin{aligned}\bar{P}(x) &= \frac{1}{\pi}\sum_{n=1}^{\infty}\frac{r(n)}{n}J_2\left(2\pi\sqrt{nx}\right)\\ &= \frac{x^{-1/4}}{\pi}\sum_{n=1}^{\infty}\frac{r(n)}{n^{5/4}}\cos\left(2\pi\sqrt{nx}-\frac{5\pi}{4}\right)+O\left(x^{-3/4}\right)\ ,\end{aligned} \tag{1.54}$$

which implies, since the sums in (1.54) are now absolutely convergent, $\lim\limits_{x\to\infty}\left|\bar{P}(x)\right| = 0$ [51, pp. 206]. A method to smooth possible spikes in $P(x)$, which originates in a paper by Cramér in 1922 [84], is to study higher moments of $P(x)$

$$M_k(x) := \frac{1}{x}\int_0^x \left|P(y)\right|^k \,\mathrm{d}y \tag{1.55}$$

for $k > 0$ and

$$m_k(x) := \frac{1}{x}\int_0^x (P(y))^k \,\mathrm{d}y \tag{1.56}$$

for $k = 1, 3, 5, \ldots$. (Note that $m_1(x) = \bar{P}(x)$). The following results are known [85]

$$\begin{aligned} M_k(x) &\to C_k x^{k/4}, \quad k \in [0, 9] \\ m_k(x) &\to c_k x^{k/4}, \quad k = 3, 5, 7, 9. \end{aligned} \qquad (x \to \infty) \tag{1.57}$$

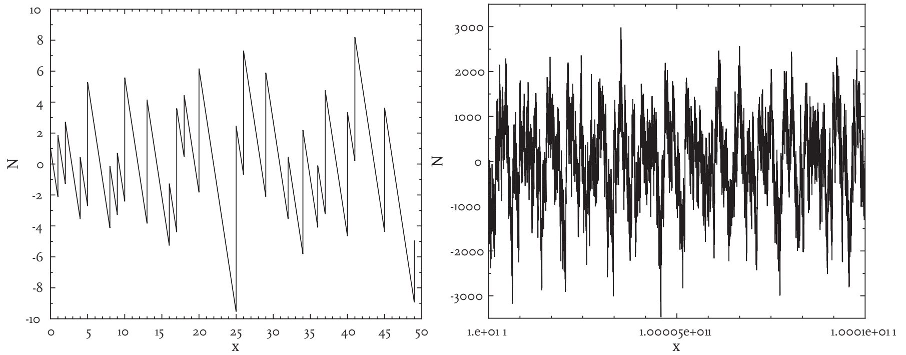


Figure 1.2 The remainder term $P(x)$ of the Gaussian circle problem (respectively the fluctuating part of the torus problem with $L = 2\pi$) is shown in different intervals.

($C_2 = 1/(3\pi^2) \sum_{n=1}^{\infty} r(n)^2/n^{3/2}$ [84]). It follows that the moments (1.57) are consistent with Hardy's conjecture, $P(x) = O\left(x^{1/4+\varepsilon}\right)$, since this implies $(m_k(x))^{1/k} = O\left(x^{1/4}\right)$ resp. $(M_k(x))^{1/k} = O\left(x^{1/4}\right)$, but of course they do not prove it. Nevertheless it seems justified to say that the "mean" behavior of $P(x)$ is proportional to $x^{1/4}$ for $x \to \infty$.

1.3.10 The Value Distribution of the Remainder Term in the Circle Problem

In the preceding section we saw that the remainder term $P(x)$ in the circle problem (respectively the fluctuating part $N_{\text{fl}}(\lambda)$ in Weyl's law for the torus problem) is a very irregular function fluctuating about zero (see Figures 1.1 and 1.2). It thus appears natural to consider $P(x)$ as a random function of x and to study its statistical properties in the limit $x \to \infty$, like its moments as in Equations (1.55) and (1.56), its limit distribution (if it exists), correlations etc., rather than to estimate its magnitude, i.e. trying to prove Hardy's conjecture. Since the moment $M_2(x)$, see (1.55), is the variance of $P(x)$, an obvious quantity to study is the *normalized remainder term*

$$W(x) := \frac{P(x)}{\sqrt{M_2(x)}} .$$

Since $M_2(x) \to C_2 \sqrt{x}$ for $x \to \infty$, it turns out to be convenient to consider the function

$$F(p) := \frac{P(p^2)}{\sqrt{p}} = \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{r(n)}{n^{3/4}} \cos\left(2\pi \sqrt{n} p - \frac{3\pi}{4}\right) + O\left(\frac{1}{p}\right) \quad (p \to \infty) \tag{1.58}$$

as a function of $p := \sqrt{x} > 1$ and $F(p) = 0$ for $p < 1$. Obviously, $F(p)$ fluctuates about zero and its mean value vanishes asymptotically for $p \to \infty$, whereas Cramér's result [84] implies that the second moment of $F(p)$ exists. There now arise the following questions. i) Does $F(p)$, where p is randomly chosen from the interval

$[1, p_m]$, have for $p_m \to \infty$ a limit distribution $f(\alpha)\,\mathrm{d}\alpha$ with probability density $f(\alpha)$? ii) Assuming that $f(\alpha)$ exists, what is its form? In view of the erratic behavior of $P(p^2)$ and thus of $F(p)$, one may guess that the central limit theorem can be applied to $F(p)$ and thus $f(\alpha)$ should be a Gaussian.

The study of the distribution of $F(p)$ was initiated by Heath-Brown [85] who proved that $F(p)$ has indeed a distribution function $f(\alpha)$ in the sense that, for any interval $[a, b] \subset \mathbb{C}$ we have

$$\lim_{p_m\to\infty} \frac{1}{p_m}\mu\{p \in [0, p_m] : F(p) \in [a, b]\} = \int_a^b f(\alpha)\,\mathrm{d}\alpha \tag{1.59}$$

(here μ denotes the Lebesgue measure.) Moreover, he proved that $f(\alpha)$ can be extended to an entire function on $\mathbb{C}$ and decreases faster than polynomially on the real line as $|\alpha| \to \infty$.

The results of Heath-Brown were developed further by Bleher, Cheng, Dyson and Lebowitz [86] who proved

$$\lim_{p_m\to\infty} \frac{1}{p_m}\int_0^{p_m} g(F(p))\varrho\left(\frac{p}{p_m}\right)\mathrm{d}p = \int_{-\infty}^{\infty} g(\alpha)f(\alpha)\,\mathrm{d}\alpha \tag{1.60}$$

for every piecewise continuous bounded function $g(x)$ on $\mathbb{R}$ and for an arbitrary probability density $\varrho(x) \geq 0$ on $[0, 1]$. In addition, they showed that for every $\varepsilon > 0$ there exists $\alpha_0 = \alpha_0(\varepsilon) > 0$ such that, on the real line $\alpha \in \mathbb{R}$, we have the upper bound

$$0 \leq f(\alpha) < e^{-|\alpha|^{4-\varepsilon}} \tag{1.61}$$

when $|\alpha| > \alpha_0$, and that the cumulative distribution $C(\alpha) := \int_{-\infty}^{\alpha} f(\alpha')\,\mathrm{d}\alpha'$ satisfies for every $\alpha > \alpha_0$ the lower bound

$$C(-\alpha), 1 - C(\alpha) > e^{-\alpha^{4+\varepsilon}} . \tag{1.62}$$

These results [85, 86] came as a great surprise since they imply that $f(\alpha)$ decreases for $|\alpha| \to \infty$ roughly as $e^{-\alpha^4}$ and thus faster than a Gaussian density! A numerical computation of $f(\alpha)$ is shown in Figure 1.3 and compared with a normal Gaussian distribution. The deviation from a Gaussian distribution is clearly visible; moreover, one observes that $f(\alpha)$ is skewed towards positive values of α.

In the next section we shall formulate a conjecture which states that the non-Gaussian behavior of $f(\alpha)$ has its origin in the fact that the circle problem can be related to the remainder term of Weyl's law for a quantum mechanical system whose corresponding classical system (i.e. the geodesic flow on a torus with $L = 2\pi$) is integrable and thus regular.

The proof of the properties (1.60)–(1.62) is based on the fact that $F(p)$ is an almost periodic function of Besicovitch class B^2 [86, 87], which means

$$\lim_{N\to\infty} \lim_{p_m\to\infty} \frac{1}{p_m}\int_0^{p_m} \left|F(p) - \frac{1}{\pi}\sum_{n=1}^{N} \frac{r(n)}{n^{3/4}} \cos\left(2\pi\sqrt{n}p - \frac{3\pi}{4}\right)\right|^2 \mathrm{d}p = 0 . \tag{1.63}$$

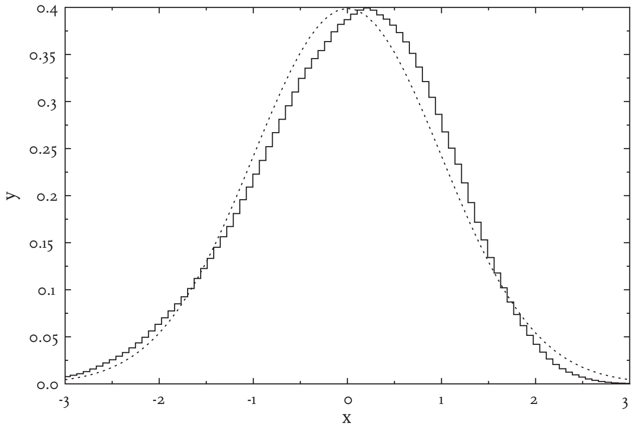


Figure 1.3 The distribution function $f(\alpha)$ is shown for the circle problem in comparison with a Gaussian normal distribution (dashed curve).

1.3.11
A Conjecture on the Value Distribution of the Remainder Term in Weyl's Law for Integrable and Chaotic Systems

In this section we would like to mention an application of Weyl's law in quantum mechanics. Let us consider a bound quantum system i.e. a system whose quantum Hamiltonian has a purely discrete energy spectrum $\{\lambda_n\}_{n\in\mathbb{N}}$. To have a specific example in mind, think of two-dimensional quantum billiards on a bounded domain Ω with area $|\Omega|$, for which the time-independent Schrödinger equation reads (in appropriate units) $-\Delta_\Omega u_n(x) = \lambda_n u_n(x)$ imposing (for example) Dirichlet or Neumann boundary conditions on $\partial\Omega$ (see (1.1) and (1.2)). Moreover, let us assume that Weyl's law holds in the form

$$N(\lambda) = \overline{N}(\lambda) + N_{\mathrm{fl}}(\lambda)\ , \tag{1.64}$$

where the smooth part $\overline{N}(\lambda)$ describes asymptotically the mean behavior of the counting function $N(\lambda)$, i.e. the fluctuating remainder term $N_{\mathrm{fl}}(\lambda)$ satisfies

$$\frac{1}{\lambda}\int\limits_{\lambda_1}^{\lambda} N_{\mathrm{fl}}(\lambda')\,\mathrm{d}\lambda' \to 0 \quad \text{for} \quad \lambda \to \infty\ . \tag{1.65}$$

For "generic" two-dimensional billiards, there exists a three-term formula for $\overline{N}(\lambda)$

$$\overline{N}(\lambda) = \frac{|\Omega|}{4\pi}\lambda \mp \frac{|\partial\Omega|}{4\pi}\sqrt{\lambda} + C\ , \tag{1.66}$$

where the first two terms correspond to Weyl's conjecture (see (1.6) and the remarks after (1.7)), and the constant C takes the curvature of $\partial\Omega$ and corner corrections into account (see (1.67)).

The definition of what is meant by "generic" is a very subtle subject, the discussion of which is beyond the scope of this contribution. Examples of generic and nongeneric systems are discussed in [88]. A rigorous definition requires the introduction of geometrical concepts like "nonperiodicity" and "nonblocking"; see for example [89]. To derive the smoothed counting function $\overline{N}(\lambda)$, several averaging procedures have been invented, in particular by Brownell [90], which are described in [91]. For a simply connected domain Ω possessing piecewise smooth arcs of length γ_i and corners of angle $\varphi_j \in (0, 2\pi]$ one obtains [91, p. 62] (1.66) with

$$C = \frac{1}{12\pi} \sum_i \int\limits_{\gamma_i} \kappa(l)\,\mathrm{d}l + \frac{1}{24} \sum_j \left(\frac{\pi}{\varphi_j} - \frac{\varphi_j}{\pi} \right), \tag{1.67}$$

where $\kappa(l)$ ($l \in \operatorname{arc} \gamma_i \subset \partial\Omega$) denotes the curvature of the arc γ_i. It should be noted, however, that the three-term formula (1.66) does not imply $N_{\mathrm{fl}}(\lambda) = O(1)$. On the contrary, the problem of determining $\alpha_0 = \inf \alpha$ in the estimate $N_{\mathrm{fl}}(\lambda) = O(\lambda^{\alpha})$ is a very difficult one; the circle problem discussed in Section 1.3.2 being an illustrative example.

To compare the quantal spectra of different systems, one has to get rid of the system-dependent constants in $\overline{N}(\lambda)$, which is achieved by "unfolding" the spectra by $x_n := \overline{N}(\lambda_n)$. The unfolded spectrum $\{x_n\}_{n\in\mathbb{N}}$ has by construction a unit mean level spacing, and thus the corresponding counting function $\hat{N}(x) := \#\{x_n \le x\}$ reads $\hat{N}(x) = x + \hat{N}_{\mathrm{fl}}(x)$. Obviously,

$$\frac{1}{x - x_1} \int\limits_{x_1}^{x} \hat{N}_{\mathrm{fl}}(y)\,\mathrm{d}y \to 0 \quad \text{for } x \to \infty. \tag{1.68}$$

In analogy to the approach discussed in Section 1.3.10 for the circle problem, we are interested in the statistical properties of the *normalized remainder term*

$$W(x) := \frac{\hat{N}_{\mathrm{fl}}(x)}{\sqrt{D(x)}}, \tag{1.69}$$

where $D(x)$ denotes the variance (ξ is a constant to be given below)

$$D(x) := \frac{\xi}{x - x_1} \int\limits_{x_1}^{x} \left(\hat{N}_{\mathrm{fl}}(y) \right)^2 \mathrm{d}y\,. \tag{1.70}$$

We now consider $W(x)$ as a random variable, where x is chosen randomly from the interval $[x_1, x_m]$ and ask whether $W(x)$ possesses in the limit $x_m \to \infty$ a limit distribution. If a limit distribution exists, it has by construction a second moment of one (if the second moment exists) and a first vanishing moment.

We are now in a position to formulate the following

Conjecture 1.1 ([92, 93]) *For bound conservative and scaling quantum systems the quantity $W(x)$, Equation (1.69), possesses for $x \to \infty$ a limit distribution with zero mean and unit variance. This distribution is absolutely continuous with respect to Lebesgue measure on the real line with a density $f(\alpha)$ i.e.*

$$\lim_{x_m \to \infty} \frac{1}{x_m} \int_{x_1}^{x_m} \mathrm{g}(W(x)) \varrho\left(\frac{x}{x_m}\right) \mathrm{d}x = \int_{-\infty}^{\infty} \mathrm{g}(\alpha) f(\alpha) \, \mathrm{d}\alpha \,, \tag{1.71}$$

where $\mathrm{g}(x)$ is a bounded continuous function on $\mathbb{R}$, and $\varrho(x) \geq 0$ a probability density on $[0, 1]$. Furthermore,

$$\int_{-\infty}^{\infty} \alpha f(\alpha) \, \mathrm{d}\alpha = 0, \qquad \int_{-\infty}^{\infty} \alpha^2 f(\alpha) \, \mathrm{d}\alpha = 1 \,. \tag{1.72}$$

If the corresponding classical system is strongly chaotic, having only isolated and unstable periodic orbits, then $f(\alpha)$ is universally a Gaussian,

$$f(\alpha) = \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2} \,. \tag{1.73}$$

In contrast, a classically integrable system leads to a system-dependent non-Gaussian density $f(\alpha)$.

Here a few remarks are in order. i) The normalization used in the definition (1.69) is crucial in order for a limit distribution to exist since in all interesting cases $D(x)$ diverges for $x \to \infty$. From Berry's [94] semiclassical analysis one obtains for generic *integrable billiards*

$$D(x) \to c\sqrt{x}, \quad x \to \infty \,, \tag{1.74}$$

where c is some nonuniversal constant. (For rigorous results, see the discussion of the torus billiard in Section 1.3.9 and [95]). In contrast, for generic *classically chaotic systems* one expects

$$D(x) \to \frac{1}{2\pi^2 \beta} \ln x, \quad x \to \infty \,, \tag{1.75}$$

with $\beta = 1$ for systems with anti-unitary symmetry (for example time-reversal symmetry) and $\beta = 2$ for systems without such a symmetry. ii) The constant ξ in (1.70) takes the value $\xi = 2/3$, if $D(x)$ obeys (1.74), and $\xi = 1$ in the case of (1.75). iii) The conjecture is proven for some integrable systems like the torus (Gauss circle) problem, see [96] for a review. iv) The conjecture has been checked numerically for several integrable (like the isospectral billiard shown in Figure 1.5) and chaotic systems [88, 93] and has been found to hold with high statistical significance. v) In Figure 1.4 we show the numerical evaluation of $f(\alpha)$ for the strongly chaotic Hadamard–Gutzwiller model [64] which is the quantum version of the geodesic flow on a compact Riemann surface of genus two (for details, see Section 1.4). For this system there exists the rigorous Selberg trace formula [45] (see Equation (1.95) below) which yields for the remainder term $\hat{N}_{\mathrm{fl}}(x)$ the explicit expression (see (1.107) below)

$$\hat{N}_{\mathrm{fl}}(x) = \frac{1}{\pi} \arg Z\left(\frac{1}{2} + ix\right) \tag{1.76}$$

in terms of the Selberg zeta function $Z(s)$ evaluated on the critical line $s = 1/2 + ix$. For the numerical computation in Figure 1.4 we used the first 6000 eigenvalues with positive parity (computed by the boundary-element method [97]) of a generic (nonarithmetic) Riemann surface whose fundamental domain in the Poincaré-disk model for hyperbolic geometry is described in [97]. We conclude that the computed histogram is in nice agreement with the conjecture (1.73). vi) In many respects the nontrivial zeros of the Riemann zeta function $\zeta(s)$ behave like the scaled eigenvalues of a hypothetical classically chaotic system without anti-unitary symmetry, see Sections 1.4.8 and 1.4.9. The analogue of (1.76) reads

$$\hat{N}_{\mathrm{fl}}(x) = \frac{1}{\pi} \arg \zeta\left(\frac{1}{2} + ix\right)$$

(see (1.108) below) counting only the zeros $\{x_n\}_{n\in\mathbb{N}}$, $\zeta(1/2 + ix_n) = 0$, with $\operatorname{Re} x_n > 0$ and $-1/2 < \operatorname{Im} x_n < 1/2$. It has been shown by Selberg's moment method [98–100] that the corresponding quantity $W(x)$, with $D(x) \sim 1/2\pi^2 \ln\ln x$, has a Gaussian limit distribution in accordance with the conjecture. For a numerical calculation of $f(\alpha)$ using the first 50 000 zeros and the 50 000 zeros starting from the $10^{20} +$ 143 780 420 th zero, respectively, see Figure 8 in [101], which shows that the convergence of the probability distribution to the proven Gaussian limit distribution is very slow.

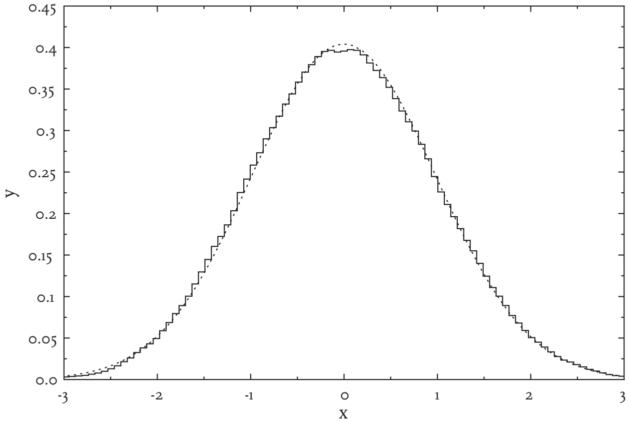


Figure 1.4 The distribution function $f(\alpha)$ is shown for the strongly chaotic Hadamard-Gutzwiller model in comparison with the conjectured Gaussian normal distribution (dashed curve).

1.4 Weyl's Law with Remainder Term. II

1.4.1 The Laplace–Beltrami Operator on *d*-Dimensional Compact Riemann Manifolds $\mathcal{M}^d$ and the Pre-Trace Formula

In many physical applications (ergodic theory, quantum mechanics, nonlinear optics, general relativity, string theory, and cosmology) one has to deal with the wave equation (or heat or Schrödinger equation) on non-Euclidean spaces. Important examples are d-dimensional manifolds or orbifolds $\mathcal{M}^d$ endowed with a Riemannian metric for which the Euclidean Laplacian has to be replaced by the corresponding Laplace–Beltrami operator. For simplicity, we discuss only manifolds with constant Gaussian curvature K.

Let us first consider smooth compact Riemannian manifolds $\mathcal{M}^d$ without boundary which are well studied and for which one can derive exact trace formulae and therefore can obtain full information on Weyl's law and even on Carleman's law [102, 103] involving the eigenfunctions. The simplest case of zero curvature $K = 0$ i.e. flat tori $\mathcal{M}^d_\Gamma = \mathbb{R}^d/\Gamma$, where Γ is a group of motions isomorphic to $\mathbb{Z}^d$, which are compact Riemannian manifolds, has already been discussed in Section 1.3.

The case of homogeneous manifolds with constant positive curvature $K = +1$ is also well understood but will not be treated here.

The case of compact manifolds with constant negative curvature $K = -1$ and dimension $d \geq 2$ is highly nontrivial since the eigenvalues and eigenfunctions of the Laplace–Beltrami operator corresponding to the non-Euclidean (hyperbolic) metric are not known analytically. The geodesic flow i.e. the free motion of a point particle on these hyperbolic manifolds was already studied by Jacques Hadamard in 1898 [104, 105] and has played an important role in the development of ergodic theory ever since. Hadamard proved that all trajectories in this system are unstable and that neighboring trajectories diverge in time at an exponential rate, the most striking property of *deterministic chaos*. In 1980, Martin Gutzwiller drew attention to this system as a prototype example for *quantum chaos* [106]. Today the quantum system governed by the free Schrödinger equation i.e. the eigenvalue problem of the Laplace–Beltrami operator on these hyperbolic manifolds (or orbifolds), is known as the Hadamard–Gutzwiller model [64, 65, 107]. In dimension $d = 3$, hyperbolic manifolds are possible candidates for the spatial section of the Universe and are investigated in cosmology [108].

In order to define a hyperbolic manifold, one considers $\mathrm{Iso}\,\mathbb{H}^d$, the group of isometries on $\mathbb{H}^d$ (i.e. the distance-preserving bijections on $\mathbb{H}^d$), where $\mathbb{H}^d$ is the d-dimensional hyperbolic space. The action of an isometry γ of $\mathbb{H}^d$ is denoted by $\gamma(z)$ with $z \in \mathbb{H}^d$. Take a discrete subgroup Γ of $\mathrm{Iso}\,\mathbb{H}^d$ and identify all points of $\mathbb{H}^d$ which can be transformed into each other by an element of Γ. Those points are called Γ-equivalent, and we put them into an equivalence class $\Gamma(z) = \{\gamma(z) : \gamma \in \Gamma\}$ with $z \in \mathbb{H}^d$. The set of those classes defines the hyperbolic d-manifold represented by the quotient space $\mathcal{M}^d := \mathbb{H}^d/\Gamma = \left\{\Gamma(z) : z \in \mathbb{H}^d\right\}$. To visualize a given manifold, we have to take one representative from each class such that the set of all representatives yields a simply connected set in $\mathbb{H}^d$, called the fundamental domain Ω_Γ. Here we discuss only compact manifolds whose fundamental domain is of finite volume, $|\Omega_\Gamma| < \infty$. One can cover all of $\mathbb{H}^d$ with Γ-translates of Ω_Γ. This produces a tessellation of $\mathbb{H}^d$ in analogy to the case discussed in Section 1.3 for flat tori. The group Γ is then called a hyperbolic crystallographic group or simply a hyperbolic lattice. The task is then to study the eigenvalue problem of the hyperbolic Laplacian $-\Delta u(z) = \lambda u(z)$, $z \in \mathbb{H}^d$, $u \in L^2\left(\mathbb{H}^d/\Gamma, \chi\right)$, where u is *automorphic* i.e. satisfies $u(\gamma(z)) = \bar{\chi}(\gamma)u(z)$ for all $\gamma \in \Gamma$ and $z \in \mathbb{H}^d$. Here χ is any one-dimensional unitary representation of Γ, also called a character which satisfies $\left|\chi(\gamma)\right|^2 = 1$ for all $\gamma \in \Gamma$. Due to the compactness of $\mathcal{M}^d$, the spectrum of $-\Delta$ is discrete with $0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \ldots$. (whether $\lambda_0 = 0$ exists depends on $\mathcal{M}^d$).

Let us consider the *resolvent kernel* $G_\Gamma(z, z'; \lambda)$ on $\mathbb{H}^d/\Gamma$ for $f \in L^2\left(\mathbb{H}^d/\Gamma, \chi\right)$

$$\left[(-\Delta - \lambda)^{-1} f\right](z) = \int_{\Omega_\Gamma} G_\Gamma(z, z'; \lambda) f(z')\, \mathrm{d}\mu(z') \ , \tag{1.77}$$

where $\lambda \in \mathbb{C}\setminus[0, \infty)$. We then obtain the correlation function [107]

$$C_F(z, z') := \sum_n F(\lambda_n) e_n(z) \bar{e}_n(z') = \frac{1}{\pi} \int_0^\infty F(\lambda')\, \mathrm{disc}\, G_\Gamma(z, z'; \lambda')\, \mathrm{d}\lambda' \ , \tag{1.78}$$

where the spectral function $F(\lambda)$ is assumed to obey the following sufficient conditions:

- $F(\lambda)$ is holomorphic in a strip enclosing the positive real axis,
- $F(\lambda)$ drops faster than $\lambda^{-d/2}$ for $\lambda \to \infty$.

The last condition is imposed to ensure convergence of the above expression for all $z, z' \in \mathbb{H}^d$ including the diagonal $z = z'$. For $z \neq z'$ weaker conditions are sufficient. Furthermore, we have introduced the discontinuity of G_Γ across the cut in the λ-plane

$$\mathrm{disc}\, G_\Gamma(z, z'; \lambda) := \lim_{\varepsilon \to 0^+} \frac{1}{2i} \left[G_\Gamma(z, z'; \lambda + i\varepsilon) - G_\Gamma(z, z'; \lambda - i\varepsilon)\right] \ .$$

Since $C_F(z, z')$ is identical to the automorphic kernel of the operator $F(-\Delta)$, we obtain the *pre-trace formula*

$$\sum_n F(\lambda_n) = \mathrm{Tr}\, F(-\Delta) = \int_{\Omega_\Gamma} C_F(z, z)\, \mathrm{d}\mu(z) \ .$$

1.4.2 The Sum Rule for the Automorphic Eigenfunctions on $\mathcal{M}^d$

In the next step, we make use of the alternative representation of the resolvent kernel which expresses the Γ-invariant kernel G_Γ as a sum ("method of images") over the free resolvent kernel $G_0^{(d)}(z, z'; \lambda)$ on $\mathbb{H}^d$

$$G_\Gamma(z, z'; \lambda) = \sum_{\gamma \in \Gamma} \chi(\gamma) G_0^{(d)}(z, \gamma(z'); \lambda) \; .$$

The crucial point now is that $G_0^{(d)}$ is explicitly known for all $d \geq 2$, see [109]. Introduce the wave numbers p_n via $p_0 := ((d-1)/2)\, i$ and $p_n := \sqrt{\lambda_n - (d-1)^2/4} \geq 0$ for $n \geq 1$. Here p_0 belongs to $\lambda_0 = 0$ (if it exists), and $p_n, n \geq 1$, to the eigenvalues $\lambda_n \geq (d-1)^2/4$, where we have assumed that there are no so-called "small eigenvalues" with $0 < \lambda_n < (d-1)^2/4$. It is now convenient to replace the spectral function $F(\lambda)$ by a new *spectral function*

$$h(p) := F\left(p^2 + \frac{(d-1)^2}{4}\right) = F(\lambda) : \mathbb{C} \to \mathbb{C} \; ,$$

which has to fulfil the following sufficient conditions

$$\begin{aligned} &\bullet \; h(-p) = h(p) \\ &\bullet \; h(p) \text{ is holomorphic in the strip } |\mathrm{Im}\, p| \leq \frac{d-1}{2} + \varepsilon, \; \varepsilon > 0 \\ &\bullet \; h(p) = O\left(p^{-d-\delta}\right), \delta > 0 \text{ for } |p| \to \infty. \end{aligned} \tag{1.79}$$

Then the correlation function takes the final form of *a "sum rule" for the automorphic eigenfunctions* e_n $(d \geq 2)$ [107]

$$\sum_{n=0}^{\infty} h(p_n) e_n(z) \bar{e}_n(z') = \frac{2}{\pi} \sum_{\gamma \in \Gamma} \chi(\gamma) \int_0^\infty p h(p) \hat{\Phi}^{(d)}(\cosh d(z, \gamma(z')); p)\, \mathrm{d}p \; , \tag{1.80}$$

where $d(z, z')$ denotes the hyperbolic distance between arbitrary points $z, z' \in \mathbb{H}^d$. $d(z, z')$ is a point-pair invariant, i.e. $d(\gamma(z), \gamma(z')) = d(z, z')$ for all $\gamma \in \Gamma$ and $z, z' \in \mathbb{H}^d$. For $z = z'$ the distance $\tau_\gamma := d(z, \gamma(z))$ is the length of a closed orbit, but which is in general not a periodic one. The function $\hat{\Phi}^{(d)}(y; p)$ is explicitly given by $(y \geq 1)$

$$\hat{\Phi}^{(d)}(y; p) = \frac{\pi}{(2\pi)^{d/2}} \frac{(y^2-1)^{(2-d)/4}}{2p} \left| \frac{\Gamma(ip + (d-1)/2)}{\Gamma(ip)} \right|^2 P^{(2-d)/2}_{-1/2+ip}(y) \; , \tag{1.81}$$

where $P_\nu^\mu(y)$ is the associated Legendre function of the first kind.

1.4.3 Weyl's Law on $\mathcal{M}^d$ and its Generalization by Carleman

At this point let us introduce the *generalized counting function*

$$N_\Gamma^{(d)}(\lambda; z, z') := \sum_{\lambda_n \leq \lambda} e_n(z) \bar{e}_n(z'), \tag{1.82}$$

which for $z = z'$ gives *Carleman's function* $\sum\limits_{\lambda_n \leq \lambda} \left|e_n(z)\right|^2$ [102,103] and after integrating over Ω_Γ the usual counting function

$$N_\Gamma^{(d)}(\lambda) = \int\limits_{\Omega_\Gamma} N_\Gamma^{(d)}(\lambda; z, z)\,\mathrm{d}\mu(z) = \sum_{\lambda_n \leq \lambda} 1 \tag{1.83}$$

(since $\int\limits_{\Omega_\Gamma} e_m(z)\bar{e}_n(z)\,\mathrm{d}\mu(z) = \delta_{mn}$).

We then obtain from the sum rule (1.80) the explicit formula

$$\mathrm{d}N_\Gamma^{(d)}(\lambda; z, z') = \mathrm{d}\overline{N}_\Gamma^{(d)}(\lambda; z, z') + \mathrm{d}N_{\Gamma,\text{fl}}^{(d)}(\lambda; z, z') \tag{1.84}$$

with

$$\mathrm{d}\overline{N}_\Gamma^{(d)}(\lambda; z, z') := \frac{1}{\pi}\hat{\Phi}^{(d)}\left(\cosh d(z, z');\ \sqrt{\lambda - \left(\frac{d-1}{2}\right)^2}\right)\mathrm{d}\lambda \tag{1.85}$$

and

$$\mathrm{d}N_{\Gamma,\text{fl}}^{(d)}(\lambda; z, z') := \frac{1}{\pi}\sum_{\gamma \in \Gamma'} \chi(\gamma)\hat{\Phi}^{(d)}\left(\cosh d(z, \gamma(z'));\ \sqrt{\lambda - \left(\frac{d-1}{2}\right)^2}\right)\mathrm{d}\lambda\ ,$$

where $\Gamma' := \Gamma \setminus \{I\}$ (I denotes the identity) and $\chi(I) = 1$ was used. From our discussion of the trace formula for the tori $\mathbb{T}^d$ we expect that (1.85) gives the asymptotically leading smooth contribution to the generalized counting function (1.82). With $d(z, z) = 0$ we obtain from (1.85) for $z = z'$ $\left(p := \sqrt{\lambda - ((d-1)/2)^2}\right)$

$$\overline{N}_\Gamma^{(d)}(\lambda; z, z) := \int\limits_{((d-1)/2)^2}^{\lambda} \mathrm{d}\overline{N}_\Gamma^{(d)}(\lambda; z, z) = \frac{2}{\pi}\int\limits_0^p \hat{\Phi}^{(d)}(1; p')p'\,\mathrm{d}p'\ ,$$

which no longer depends on z! Here $\hat{\Phi}^{(d)}(1; p)$ follows from (1.81)

$$\begin{aligned}\hat{\Phi}^{(d)}(1; p) &= \frac{\pi}{(2\pi)^{d/2}}\frac{1}{2p}\left|\frac{\Gamma(ip + (d-1)/2)}{\Gamma(ip)}\right|^2 \lim_{y \to 1^+}\frac{P_{-1/2+ip}^{(2-d)/2}(y)}{(y^2 - 1)^{(d-2)/4}} \\ &= \frac{d}{(4\pi)^{d/2}\Gamma(1 + d/2)} \cdot \frac{\pi}{2p} \cdot \left|\frac{\Gamma(ip + (d-1)/2)}{\Gamma(ip)}\right|^2\end{aligned} \tag{1.86}$$

and thus

$$\overline{N}_\Gamma^{(d)}(\lambda; z, z) = \frac{d}{(4\pi)^{d/2}\Gamma(1 + d/2)}\int\limits_0^p \left|\frac{\Gamma(ip' + (d-1)/2)}{\Gamma(ip')}\right|^2 \mathrm{d}p'\ . \tag{1.87}$$

Using the asymptotic expansion

$$\left|\frac{\Gamma(ip+(d-1)/2)}{\Gamma(ip)}\right|^2 = p^{d-1}\left(1+O\left(\frac{1}{p^2}\right)\right) \qquad (p\to\infty)\ ,$$

we immediately obtain

$$\overline{N}_\Gamma^{(d)}(\lambda;z,z) = \frac{1}{(4\pi)^{d/2}\Gamma(1+d/2)}\lambda^{d/2} + O\left(\lambda^{d/2-1}\right) \qquad (\lambda\to\infty) \tag{1.88}$$

and after integration over Ω_Γ the non-Euclidean analog of *Weyl's law* ($d \geq 2$)

$$\overline{N}_\Gamma^{(d)}(\lambda) = \frac{|\Omega_\Gamma|}{(4\pi)^{d/2}\Gamma(1+d/2)}\lambda^{d/2} + O\left(\lambda^{d/2-1}\right)\ . \tag{1.89}$$

Since one can show that the remainder term satisfies $N_{\Gamma,\text{fl}}^{(d)}(\lambda;z,z) = O\left(\lambda^{d/2}\right)$, we obtain *Carleman's law*

$$N_\Gamma^{(d)}(\lambda;z,z) = \sum_{\lambda_n\leq\lambda}\left|e_n(z)\right|^2 = \frac{\lambda^{d/2}}{(4\pi)^{d/2}\Gamma\left(1+d/2\right)} + O_z\left(\lambda^{d/2}\right) \quad (\lambda\to\infty)\ , \tag{1.90}$$

which is a generalization of Weyl's law since it is not only a statement about the eigenvalues but also about the eigenfunctions. Note, however, that the sum rule (1.80) – being an exact explicit expression – contains much more information. To see this, let us consider the simplest case $d = 2$ in more detail.

1.4.4
The Selberg Trace Formula and Weyl's Law

In the case $d = 2$ we consider compact Riemann surfaces $\mathcal{M}^2 = \mathbb{H}^2/\Gamma$ of genus $g \geq 2$ with Γ a strictly hyperbolic Fuchsian group of the first kind, $\Gamma \in \text{PSL}(2,\mathbb{R})$. Such groups are characterized by the fact that all their group elements γ (except the unity I) are hyperbolic. Here we choose for $\mathbb{H}^2$ the Poincaré!upper half plane $\mathbb{H}^2 = \{z = x + iy : x, y \in \mathbb{R}, y > 0\}$ with the hyperbolic metric

$$\mathrm{d}s^2 = \frac{\mathrm{d}x^2+\mathrm{d}y^2}{y^2}\ ,$$

which is invariant under fractional linear transformations:

$$z \to \gamma(z) := \frac{az+b}{cz+d}\ ,$$

where $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. Then the Laplace–Beltrami operator is $\Delta = y^2\left(\partial^2/\partial x^2 + \partial^2/\partial y^2\right)$. It is also invariant under the group actions $\gamma \in \Gamma$. We then obtain from (1.81)

$$\hat{\Phi}^{(2)}(y;p) = \frac{1}{4}\tanh(\pi p)P_{-1/2+ip}(y)\ ,$$

where $P_\nu(y)$ denotes the Legendre function of the first kind.

Then the sum rule (1.80) takes the simple form ($\chi(\gamma) = 1 \ \forall \gamma \in \Gamma$, $p_0 = i/2$, $p_n = \sqrt{\lambda_n - 1/4} \geq 0$, $n \geq 1$) [110]

$$\sum_{n=0}^{\infty} h(p_n) e_n(z) \bar{e}_n(z') = \frac{1}{2\pi} \sum_{\gamma \in \Gamma} \hat{h}\left(\cosh d\left(z, \gamma(z')\right)\right) , \tag{1.91}$$

where the hyperbolic distance on $\mathbb{H}^2$ is given by

$$\cosh d(z, z') = 1 + \frac{(x - x')^2 + y^2 + y'^2}{2yy'} .$$

Here $\hat{h}$ denotes the Mehler transform of the spectral function h which is defined by the relations

$$h(p) = \int_1^\infty \hat{h}(y) P_{-1/2+ip}(y) \, \mathrm{d}y \tag{1.92}$$

$$\hat{h}(y) = \int_0^\infty p \tanh(\pi p) h(p) P_{-1/2+ip}(y) \, \mathrm{d}p. \tag{1.93}$$

In [107, 110] it was shown that the sum rule (1.91) can be used to compute numerically the eigenfunctions $e_n(z)$ called nonholomorphic (or automorphic) forms or *Maass waveforms,* at least if the eigenvalues λ_n are not too large. Taking the trace of the sum rule (1.91) one gets with (1.93) and $P_{-1/2+ip}(1) = 1$ (the $\mathrm{SL}(2, \mathbb{R})$-invariant area element on $\mathbb{H}^2$ is $\mathrm{d}\mu(z) = \mathrm{d}x \, \mathrm{d}y / y^2$)

$$\sum_{n=0}^{\infty} h(p_n) = \frac{|\Omega_\Gamma|}{2\pi} \int_0^\infty p \tanh(\pi p) h(p) \, \mathrm{d}p + \frac{1}{2\pi} \sum_{\gamma \in \Gamma'} \int_{\Omega_\Gamma} \hat{h}\left(\cosh d\left(z, \gamma(z)\right)\right) \mathrm{d}\mu(z) . \tag{1.94}$$

To evaluate the sum over $\gamma \in \Gamma'$ involving the integral over $\hat{h}$ is a nontrivial task and was first achieved by Atle Selberg [45, 46] leading to the famous *Selberg trace formula*

$$\sum_{n=0}^{\infty} h(p_n) = \frac{|\Omega_\Gamma|}{2\pi} \int_0^\infty p \tanh(\pi p) h(p) \, \mathrm{d}p + \sum_{\{\gamma\}_p}^{\infty} \sum_{n=1}^{\infty} \frac{l(\gamma)}{2 \sinh\left(n l(\gamma)/2\right)} \tilde{h}\left(n l(\gamma)\right) , \tag{1.95}$$

where $\tilde{h}(x)$ denotes the Fourier transform of $h(p)$

$$\tilde{h}(x) := \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{ipx} h(p) \, \mathrm{d}p .$$

The sum on the right-hand side of (1.95) runs over the *length spectrum* $\{l(\gamma)\}_p$ of the primitive periodic orbits of the geodesic flow on the surface $\mathcal{M}^2 = \mathbb{H}^2/\Gamma$. Notice that the length spectrum is uniquely given by the conjugacy classes of the hyperbolic elements in Γ as can be seen as follows. The elements $\gamma \in \Gamma$ of the discrete subgroups of $\mathrm{PSL}(2,\mathbb{R})$ can be represented as 2×2 matrices $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ with real entries and $\det\gamma = ad - bc = 1$. For a strictly hyperbolic group one has, for all $\gamma \neq \pm I$: $\left|\mathrm{Tr}\,\gamma\right| = \left|a + d\right| > 2$. The Jordan form of these matrices takes the form $\left(\begin{smallmatrix} a & 0 \\ 0 & 1/a \end{smallmatrix}\right)$ with $|a| > 1$, and the action of γ gives $z \to \gamma(z) = a^2 z$, where $N(\gamma) := a^2$ is called the norm of the element γ. Since there exists a unique relationship between the conjugacy classes in Γ and the homotopy classes of closed paths on $\mathbb{H}^2$, one can define in each class a length $l(\gamma)$ by the length of the shortest path, and then obtains $N(\gamma) = e^{l(\gamma)}$, $l(\gamma) > 0$. The length $l(\gamma)$ is then given by $\cosh(l(\gamma)/2) = \left|\mathrm{Tr}\,\gamma\right|/2$.

The sums and integrals in the Selberg trace formula are all absolutely convergent if the spectral function $h(p)$ satisfies conditions (1.79) for $d = 2$. The Selberg trace formula (1.95) can be considered as a generalization and noncommutative analogue of the classical Poisson summation formula (1.33), respectively of the trace formulae (1.37) and (1.42–1.44) for flat tori.

From the Selberg trace formula (1.95) we can immediately read off the complete Weyl term of the counting function (see the discussion above for general $d \geq 2$)

$$\overline{N}_\Gamma^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) := \frac{|\Omega_\Gamma|}{2\pi} \int_0^p p' \tanh(\pi p')\,\mathrm{d}p' \;, \tag{1.96}$$

which behaves as

$$\overline{N}_\Gamma^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) = \frac{|\Omega_\Gamma|}{6} p^3 + O\left(p^5\right) \quad \text{for } p \to 0 \;,$$

and hence we obtain *Weyl's law on compact Riemann surfaces of genus g* ≥ 2

$$\overline{N}_\Gamma^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) = \frac{|\Omega_\Gamma|}{4\pi}\left(p^2 - \frac{1}{12}\right) + O\left(p e^{-2\pi p}\right) \text{ for } p \to \infty \;. \tag{1.97}$$

This asymptotic formula contains the standard Weyl term proportional to λ and the volume $|\Omega_\Gamma|$, no term proportional to $\sqrt{\lambda}$, since $\mathcal{M}^2$ has no boundary, it has a constant term and then an exponentially small correction. Below we shall also derive the fluctuating remainder term of the counting function.

1.4.5 The Trace of the Heat Kernel on $\mathcal{M}^2$

Choosing the spectral function $h(p) = e^{-\left(p^2+1/4\right)t}$, $t > 0$, we obtain for the *trace of the heat kernel on a compact Riemann surface* $\mathcal{M}^2$ *of genus g* ≥ 2 possessing the area

$|\Omega_\Gamma| = 4\pi\,(g-1)$ (Gauss–Bonnet) the explicit formula ($t > 0$) [47]

$$\begin{aligned}
\Theta^{\mathcal{M}^2}(t) &:= \sum_{n=0}^{\infty} e^{-\lambda_n t} = \sum_{n=0}^{\infty} e^{-\left(p_n^2+1/4\right)t} = \Theta_1^{\mathcal{M}^2}(t) + \Theta_2^{\mathcal{M}^2}(t), \\
\Theta_1^{\mathcal{M}^2}(t) &:= |\Omega_\Gamma|\,\frac{e^{-t/4}}{(4\pi t)^{3/2}} \int_0^{\infty} \frac{x}{\sinh(x/2)}\, e^{-x^2/4t}\,\mathrm{d}x \\
&= \frac{|\Omega_\Gamma|}{4\pi t} \sum_{n=0}^{N} b_n t^n + O\!\left(t^N\right), \qquad t \to 0^+ \,, \qquad (1.98) \\
b_0 &= 1,\, b_n = \frac{(-1)^n}{2^{2n} n!} \left[1 + 2 \sum_{k=1}^{n} \binom{n}{k} \left(2^{2k-1} - 1\right) |B_{2k}| \right] , \qquad n \in \mathbb{N}, \\
\Theta_2^{\mathcal{M}^2}(t) &:= \frac{e^{-t/4}}{4\sqrt{\pi t}} \sum_{\{\gamma\}_p} \sum_{n=1}^{\infty} \frac{l(\gamma)}{\sinh\left((nl(\gamma))/2\right)}\, e^{-n^2 l^2(\gamma)/4t} \,,
\end{aligned}$$

where B_{2k} are the Bernoulli numbers ($b_1 = -1/3$, $b_2 = 1/15$). This formula is the generalization of Poisson's transformation formula for the elliptic theta function θ_3 discussed in Section 1.3.6 to Riemann surfaces of genus $g \geq 2$. Thus $\Theta^{\mathcal{M}^2}(t)$ can be called the non-Euclidean theta function. The formula (1.98) is quite interesting since it shows that for compact Riemann surfaces of genus $g \geq 2$ the complete small-t asymptotics is explicitly known, see the term $\Theta_1^{\mathcal{M}^2}$, and not just the leading Weyl term $|\Omega_\Gamma|/4\pi t$. Furthermore, there even exists a closed expression for this contribution as an integral which is valid for all $t > 0$ and is not just an asymptotic result in the limit $t \to 0^+$. Moreover, the remainder term $\Theta_2^{\mathcal{M}^2}$ also has an explicit representation as a sum over the length spectrum of periodic orbits. This term is exponentially small in the limit $t \to 0^+$ and is determined by the shortest periodic orbit with primitive length $l(\gamma_1) > 0$, i.e. $\Theta_2^{\mathcal{M}^2}(t) = O\!\left(t^{-1/2} e^{-l^2(\gamma_1)/4t}\right)$ in close analogy with the behavior on the torus $\mathbb{T}^2$.

1.4.6 The Trace of the Resolvent on $\mathcal{M}^2$ and Selberg's Zeta Function

In order to calculate the trace of the resolvent of $-\Delta$ on $\mathcal{M}^2$, one is led to substitute $h(p) = \left(1/4 + p^2 - \lambda\right)^{-1}$ in the trace formula. This function violates, however, the asymptotic condition in Equation (1.79) for $|p| \to \infty$, i.e. the resolvent is not of trace class as a consequence of Weyl's law which tells us that the eigenvalues behave as $\lambda_n = 1/4 + p_n^2 \sim (4\pi/\Omega_\Gamma)\, n$ for $n \to \infty$. Thus the resolvent has to be regularized properly. A very convenient regularization is given by the following choice. ($\operatorname{Re} s$, $\operatorname{Re} \sigma > 1$)

$$h(p) = \frac{1}{p^2 + (s - 1/2)^2} - \frac{1}{p^2 + (\sigma - 1/2)^2} \,,$$

which fulfills all the conditions (1.79) in the trace formula. For the integral (Weyl) term in the trace formula (1.95) one then obtains

$$\frac{|\Omega_\Gamma|}{2\pi}\int_0^\infty p\tanh(\pi p)h(p)\,\mathrm{d}p = -\frac{|\Omega_\Gamma|}{2\pi}\left(\psi(s)-\psi(\sigma)\right) ,$$

where $\psi(s) := \Gamma'(s)/\Gamma(s)$ is the digamma function. Using the Fourier transform ($\mathrm{Re}\, s > 1/2, x \geq 0$)

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{ipx}}{p^2+(s-1/2)^2}\,\mathrm{d}p = \frac{1}{2s-1}e^{-(s-1/2)x},$$

the *Selberg trace formula for the trace of the regularized resolvent* reads ($\mathrm{Re}\, s, \mathrm{Re}\, \sigma > 1$)

$$\sum_{n=0}^{\infty}\left(\frac{1}{\lambda_n+s(s-1)}-\frac{1}{\lambda_n+\sigma(\sigma-1)}\right) = -\frac{|\Omega_\Gamma|}{2\pi}\left(\psi(s)-\psi(\sigma)\right) + \frac{1}{2s-1}A(s)-\frac{1}{2\sigma-1}A(\sigma) , \tag{1.99}$$

where the function $A(s)$ is for $\mathrm{Re}\, s > 1$ given by the absolutely convergent double sum

$$A(s) := \sum_{\{\gamma\}_p}\sum_{n=1}^{\infty}\frac{l(\gamma)e^{-(s-1/2)nl(\gamma)}}{2\sinh\left(nl(\gamma)/2\right)} .$$

It was one of Selberg's deep insights to realize that $A(s)$ can be rewritten for $\mathrm{Re}\, s > 1$ as the logarithmic derivative of a kind of zeta function $Z(s)$:

$$\begin{aligned} A(s) &= \sum_{\{\gamma\}_p}\sum_{n=1}^{\infty}\frac{l(\gamma)e^{-(s-1/2)nl(\gamma)}}{e^{nl(\gamma)/2}-e^{-nl(\gamma)/2}} = \sum_{\{\gamma\}_p}\sum_{n=1}^{\infty}\frac{l(\gamma)e^{-snl(\gamma)}}{1-e^{-nl(\gamma)}} \\ &= \sum_{\{\gamma\}_p}\sum_{n=1}^{\infty}l(\gamma)e^{-snl(\gamma)}\sum_{k=0}^{\infty}e^{-knl(\gamma)} = \sum_{\{\gamma\}_p}\sum_{k=0}^{\infty}l(\gamma)\sum_{n=1}^{\infty}e^{-(s+k)nl(\gamma)} \\ &= \sum_{\{\gamma\}_p}\sum_{k=0}^{\infty}l(\gamma)\frac{e^{-(s+k)l(\gamma)}}{1-e^{-(s+k)l(\gamma)}} = \sum_{\{\gamma\}_p}\sum_{k=0}^{\infty}\frac{\mathrm{d}}{\mathrm{d}s}\ln\left(1-e^{-(s+k)l(\gamma)}\right) \\ &= \frac{\mathrm{d}}{\mathrm{d}s}\ln\left[\prod_{\{\gamma\}_p}\prod_{k=0}^{\infty}\left(1-e^{-(s+k)l(\gamma)}\right)\right] =: \frac{Z'(s)}{Z(s)} . \end{aligned}$$

Here we have defined the *Selberg zeta function* ($\mathrm{Re}\, s > 1$) [45]

$$Z(s) := \prod_{\{\gamma\}_p}\prod_{k=0}^{\infty}\left(1-e^{-(s+k)l(\gamma)}\right) , \tag{1.100}$$

which is given as a generalized Euler product over the lengths of the primitive periodic orbits. It follows from Selberg's trace formula that the infinite products in (1.100) are absolutely convergent for $\operatorname{Re} s > 1$. Replacing $A(s)$ and $A(\sigma)$ in (1.99) by Selberg's zeta function, we obtain an exact relation [47] which expresses the trace of the regularized resolvent of $-\Delta$ on an arbitrary compact Riemann surface of genus $g \geq 2$ in terms of the well-known ψ-function and Selberg's zeta function. On the other hand, this relation allows us to prove that $Z(s)$ can be continued to the left of $\operatorname{Re} s = 1$. This can be seen by rewriting (1.99) as follows [47]

$$\frac{1}{2s-1}\frac{Z'(s)}{Z(s)} = -2\,(g-1)\,\psi(\sigma) + \left(\frac{1}{2\sigma-1}\frac{Z'(\sigma)}{Z(\sigma)} - \frac{1}{\sigma(\sigma-1)}\right) \tag{1.101}$$

$$+\,2\,(g-1)\,\psi(s) + \frac{1}{s(s-1)} + \sum_{n=1}^{\infty}\left(\frac{1}{\lambda_n + s(s-1)} - \frac{1}{\lambda_n + \sigma(\sigma-1)}\right).$$

Note that the sum over the eigenvalues no longer contains the zero mode $\lambda_0 = 0$. Keeping the regulator σ fixed with $\operatorname{Re}\sigma > 1$, we see that the right-hand side of (1.101), derived for $\operatorname{Re} s > 1$, is actually meromorphic for all $s \in \mathbb{C}$. Thus the left-hand side of (1.101) is also meromorphic, and so we obtain the analytic continuation of $Z(s)$ on $\mathbb{C}$. In fact, further inspection shows that the Selberg zeta function is an entire function of s of order 2 whose "trivial" zeros are at $s = -k$, $k \in \mathbb{N}$, with multiplicity $2(g-1)(2k+1)$. Furthermore, $s = 1$ is a simple zero, and $s = 0$ is a zero of multiplicity $2g-1$. In addition $Z(s)$ can have a finite number of zeros on the real axis between 0 and 1 located at $s = 1/2 \pm \sqrt{1/4 - \lambda_n}$ corresponding to the so-called "small" eigenvalues $0 < \lambda_n < 1/4$. For surfaces of genus $g > 2$, one has at most $4g-3$ small eigenvalues [111, 112], while in the case of $g = 2$ there is at most one small eigenvalue [113].

More importantly, $Z(s)$ has an infinite number of *"nontrivial" zeros* located at $s = 1/2 \pm ip_n$, $p_n \geq 0$, i.e. lying on the *critical line* $\operatorname{Re} s = 1/2$, and thus one can say that the *Riemann hypothesis* is valid for $Z(s)$, a very remarkable result! One therefore has the exact *quantization condition* ($p_n \in \mathbb{R}$)

$$Z\left(\frac{1}{2} + ip_n\right) = 0 \tag{1.102}$$

for the quantal eigenvalues $\lambda_n = p_n^2 + 1/4 \geq 1/4$ of the Schrödinger equation, which are completely determined by the lengths of the classical periodic orbits of the corresponding classical Hamiltonian system.

The reason behind the validity of the Riemann hypothesis in this case is obviously that $s(s-1)$ is an eigenvalue of a self-adjoint operator, and hence is real, whenever s is a zero of $Z(s)$ within the critical strip. The question of whether something similar holds for the nontrivial zeros of the Riemann zeta function, will be discussed below.

The information on the zeros of $Z(s)$ enables us now to eliminate the regulator σ in (1.101) by taking the limit $\sigma \to 1$. With $\psi(1) = -\gamma$, where γ is Euler's constant, we define the *generalized Euler constant* γ_Δ

$$\gamma_\Delta := 2\,(g-1)\,\gamma + B$$

with

$$B := \lim_{\sigma\to 1}\left(\frac{1}{2\sigma-1}\frac{Z'(\sigma)}{Z(\sigma)} - \frac{1}{\sigma(\sigma-1)}\right) = \frac{1}{2}\frac{Z''(1)}{Z'(1)} - 1 \ .$$

Since $Z(s)$ possesses a simple zero at $s = 1$, one has $Z'(1) \neq 0$ (actually $Z'(1) > 0$ holds) and thus the constant B is well defined. We then obtain for the *trace of the regularized resolvent of* $-\Delta$ *on* $\mathcal{M}^2 = \mathbb{H}^2/\Gamma$ the final result [47]

$$\frac{1}{s(s-1)} + \sum_{n=1}^{\infty}\left(\frac{1}{\lambda_n + s(s-1)} - \frac{1}{\lambda_n}\right) = \frac{1}{2s-1}\frac{Z'(s)}{Z(s)} - \gamma_\Delta - 2(g-1)\psi(s) \ . \qquad (1.103)$$

1.4.7 The Functional Equation for Selberg's Zeta Function $Z(s)$

To derive the functional equation for $Z(s)$, we notice that $s(s-1)$ is invariant under $s \to 1-s$ and $2s-1$ changes sign. If we then subtract (1.103) evaluated at $1-s$ from the same expression evaluated at s, we obtain

$$\frac{1}{2s-1}\frac{\mathrm{d}}{\mathrm{d}s}\ln\frac{Z(s)}{Z(1-s)} = 2(g-1)\left(\psi(s) - \psi(1-s)\right) \ .$$

Using the functional equation

$$\psi\left(\frac{1}{2}+z\right) - \psi\left(\frac{1}{2}-z\right) = \pi\tan(\pi z)$$

for the digamma function this then leads, with $z = s-1/2$, to the *functional equation for* $Z(s)$

$$Z(s) = \exp\left(|\Omega_\Gamma|\int_0^{s-1/2} x\tan(\pi x)\,\mathrm{d}x\right)Z(1-s) \ . \qquad (1.104)$$

Evaluating the functional equation on the critical line i.e. choosing $s = 1/2 + ip$, $p \in \mathbb{R}$, we get

$$Z\left(\frac{1}{2}+ip\right) = e^{-2\pi i \overline{N}_\Gamma^{\mathcal{M}^2}(p^2+1/4)} Z\left(\frac{1}{2}-ip\right) , \qquad (1.105)$$

where the smooth term $\overline{N}_\Gamma^{\mathcal{M}^2}$ of the counting function given in (1.96) enters as a phase. It follows that the function

$$\xi(p) := Z\left(\frac{1}{2}+ip\right)e^{i\pi\overline{N}_\Gamma^{\mathcal{M}^2}(p^2+1/4)}$$

satisfies the simple functional equation $\xi(p) = \xi(-p)$, and furthermore that $\xi(p)$ is real if $p \in \mathbb{R}$, i.e. on the critical line.

1.4.8 An Explicit Formula for the Remainder Term in Weyl's Law on $\mathcal{M}^2$ and the Hilbert–Polya Conjecture on the Riemann Zeros

Using the argument principle, one can derive the exact *Weyl formula* ($p \geq 0$, $p \neq p_n$) for the counting function

$$N_\Gamma^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) = \overline{N}_\Gamma^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) + \frac{1}{\pi} \arg Z\left(\frac{1}{2} + ip\right), \tag{1.106}$$

which proves that the fluctuating term (remainder term) of the counting function is determined by the Selberg zeta function on the critical line

$$N_{\Gamma,\mathrm{fl}}^{\mathcal{M}^2}\left(p^2 + \frac{1}{4}\right) = \frac{1}{\pi} \arg Z\left(\frac{1}{2} + ip\right). \tag{1.107}$$

The derivation of (1.106) is completely analogous to the well-known calculation leading to the counting function $N_\mathrm{R}(t)$ for the *nontrivial Riemann zeros*

$$N_\mathrm{R}(t) = \overline{N}_\mathrm{R}(t) + \frac{1}{\pi} \arg \zeta\left(\frac{1}{2} + it\right), \tag{1.108}$$

which counts the number of zeros of the Riemann zeta function $\zeta(s)$ in the region $0 < \operatorname{Re} s < 1$, $0 < \operatorname{Im} s \leq t$. Here the smooth term $\overline{N}_\mathrm{R}(t)$ is given by the famous *Riemann–von Mangoldt formula* [114]

$$\overline{N}_\mathrm{R}(t) = \frac{t}{2\pi} \ln t - \frac{1 + \ln 2\pi}{2\pi} t + \frac{7}{8} + O\left(\frac{1}{t}\right) \qquad (t \to \infty) . \tag{1.109}$$

Note that Selberg introduced his zeta function $Z(s)$ around 1950 in analogy with the Riemann zeta function $\zeta(s)$ with the intention to shed some light on the properties of the nontrivial Riemann zeros and the *Riemann hypothesis*. He noticed the striking similarities between his trace formula (1.95) and the so-called explicit formulae in the theory of prime numbers [115], whose most general form is André Weil's explicit formula [116].

Weil's explicit formula establishes a deep relation between the nontrivial zeros $\varrho_n = 1/2 + i\tau_n$, $\tau_n \in \mathbb{C}$, of $\zeta(s)$ and the prime numbers p:

$$\sum_{n=1}^{\infty} h(\tau_n) = \frac{1}{4\pi} \int_{-\infty}^{\infty} \psi\left(\frac{1}{4} + i\frac{\tau}{2}\right) h(\tau)\, \mathrm{d}\tau + h\left(\frac{i}{2}\right) - \tilde{h}(0) \frac{\ln \pi}{2} - \sum_p \sum_{n=1}^{\infty} \frac{\ln p}{p^{n/2}} \tilde{h}(n \ln p) , \tag{1.110}$$

where the "test function" $h(\tau)$ satisfies the same conditions (1.79) as the spectral function in the Selberg trace formula for $d = 2$, and $\tilde{h}(x)$ is again its Fourier transform. Here the sum on the right-hand side runs over all primes p. Comparing Weil's formula (1.110) with Selberg's trace formula (1.95), one is tempted to interpret the nontrivial zeros of $\zeta(s)$ as eigenvalues of a hypothetical "Riemann operator" and the logarithm of the prime numbers as the "lengths" $l(p) := \ln p$ of the primitive "periodic orbits" of the corresponding hypothetical geodesic flow. The term on the right-hand side of (1.110) involving the summation over the primes then reads

$$-\sum_{p}\sum_{n=1}^{\infty}\frac{l(p)}{e^{nl(p)/2}}\tilde{h}(nl(p))\ , \tag{1.111}$$

which is strikingly similar to the corresponding term in the Selberg trace formula (1.95) involving the summation over periodic orbits. Note, however, the difference between the denominator

$$+2\sinh\left(\frac{nl(\gamma)}{2}\right) = e^{nl(\gamma)/2} - e^{-nl(\gamma)/2}$$

in (1.95) which has a dynamical interpretation in terms of the linearized Poincaré recurrence map for unstable hyperbolic periodic orbits, see for example [82,92], and the corresponding denominator $-e^{nl(p)/2}$ in (1.111), for which no dynamical interpretation has been found until now; see, however, the paper by Alain Connes [117] who has devised a hermitian operator whose eigenvalues are the nontrivial Riemann zeros. His operator is the Perron–Frobenius operator (called the transfer operator in physics) of a classical dynamical system. In his framework he has found an explanation for the minus sign in (1.111).

At first sight it seems that there is another obstruction to the interpretation of the Riemann zeros as the eigenvalues of a dynamical system since the smooth counting function $\overline{N}_{\mathrm{R}}(t)$ (1.109) goes asymptotically as $\lambda/(2\pi)\ln\lambda$, if we put $t = \lambda$, which differs from the standard behavior according to Weyl's law in dimension 2. It will be seen, however, in Section 1.5.2 that such logarithmic modifications to Weyl's law can occur, for example in membrane problems, for which the domain Ω is unbounded.

Mathematical wisdom has usually attributed the formulation of the idea of a hypothetical Riemann operator to Hilbert and Polya, independently, some time in the 1910s. (See Odlyzko's correspondence with Polya [118].)

There is another difference between the Riemann and the Selberg case. In the definition of $Z(s)$ in (1.100) one has a double product, whereas $\zeta(s)$ involves only a single one. Furthermore, the "Euler factor" occurs in $Z(s)$ with the (+1) in the exponent, and in the case of $\zeta(s)$ with a (–1). It turns out that, when one generalizes the Selberg zeta function to spaces of higher rank, the natural exponents are certain Euler characteristics which can take positive or negative values [119]. To get rid of the second product in (1.100), one simply considers the ratio

$$R(s) := \frac{Z(s)}{Z(s+1)} = \prod_{\{\gamma\}_p}\left(1 - e^{-sl(\gamma)}\right)\ , \tag{1.112}$$

and ends up with the *Ruelle zeta function* $R(s)$ [120], which is now a meromorphic function. $R(s)$ or rather $1/R(s)$ has been discussed in terms of Beurling's generalized prime numbers and in connection with a generalized prime number theorem [121].

1.4.9 The Prime Number Theorem vs. the Prime Geodesic Theorem on $\mathcal{M}^2$

The famous prime number theorem states that the number of primes up to x, $\pi(x) := \#\{p : p \leq x\}$, is asymptotically equal to the logarithmic integral, given for $x > 1$ by ($\fint$ means the Cauchy principal value of the integral)

$$\mathrm{li}(x) := \fint_0^x \frac{\mathrm{d}t}{\ln t} = \frac{x}{\ln x} + \frac{x}{(\ln x)^2} + \ldots \qquad (x \to \infty) .$$

The fact that the density of primes near x is about $1/\ln x$ was already conjectured by Gauss in 1792 at the age of 15. To derive a formula for $\pi(x)$ was Riemann's main goal in his famous paper from 1859, and it was for this purpose that he studied $\zeta(s)$ which had been introduced for integer argument already in 1735 by Euler who discovered among several other relations the formula $\zeta(2) = \pi^2/6$ and in 1737 established the Euler product for $\zeta(m)$, $m \geq 2$. The prime number theorem was proved in 1896 independently by Hadamard and de la Vallée Poussin by using the Riemann zeta function. It is worthwhile noticing that the first "elementary" proof was found by Selberg in 1949, see for example [46].

If one associates the prime numbers with the "lengths" $l(p) := \ln p$, the counting function $\mathcal{N}(l) := \#\{p : l(p) \leq l\}$ counts the number of hypothetical "periodic orbits" with length up to l. The prime number theorem is then converted into

$$\mathcal{N}(l) \equiv \pi\left(e^l\right) \sim \mathrm{li}\left(e^l\right) \sim \frac{e^l}{l}, \qquad (l \to \infty) . \tag{1.113}$$

It is this result which gives perhaps the strongest support to the Hilbert–Polya conjecture, since it turns out that the counting function $\mathcal{N}_\Gamma^{\mathcal{M}^2}(l) := \#\{\gamma \in \Gamma : l(\gamma) \leq l\}$ of the genuine periodic orbits of the geodesic flow on $\mathcal{M}^2$ obeys *Huber's law* [122]

$$\mathcal{N}_\Gamma^{\mathcal{M}^2}(l) = \mathrm{li}\left(e^l\right) + O\left(\frac{e^{(3/4)l}}{l}\right), \qquad (l \to \infty) . \tag{1.114}$$

This is a special case of the general *prime geodesic theorem* valid for the counting function of the lengths of the unstable periodic orbits of chaotic systems with a *topological entropy* $\tau > 0$. In the general case, one has as leading term $e^{\tau l}/\tau l$. Thus Huber's law is consistent with the well-known fact that the geodesic flow on $\mathcal{M}^2$ is strongly chaotic, i.e. ergodic, mixing, possesses the Bernoulli property, and has topological entropy $\tau = 1$. (Actually, all periodic orbits on $\mathcal{M}^2$ are unstable and possess the same Lyapunov exponent $\lambda(\gamma) = 1$.)

Comparing (1.113) with (1.114), one concludes that the hypothetical dynamical system associated with the Riemann zeros should be chaotic, should have topological entropy $\tau = 1$, and should possess a length spectrum of primitive periodic orbits exactly given by the logarithm of the primes, $l(p) = \ln p$!

The validity of Huber's law (1.114) can be seen as follows. Due to the existence of the zero mode $\lambda_0 = 0$ with multiplicity one, $\Theta^{\mathcal{M}^2}(t) = 1 + O\left(e^{-\lambda_1 t}\right)$, $t \to \infty$, holds for the trace of the heat kernel on $\mathcal{M}^2$. Furthermore, one infers from (1.98) that the complete Weyl term $\Theta_1^{\mathcal{M}^2}(t)$ satisfies $\lim\limits_{t\to\infty}\Theta_1^{\mathcal{M}^2}(t) = 0$, and thus the remainder term $\Theta_2^{\mathcal{M}^2}(t)$ in (1.98) must satisfy $\lim\limits_{t\to\infty}\Theta_2^{\mathcal{M}^2}(t) = 1$. One therefore obtains the condition

$$\lim_{t\to\infty}\frac{e^{-t/4}}{2\sqrt{\pi t}}\int_{l_1}^{\infty} le^{-l^2/4t-l/2}\,\mathrm{d}\mathcal{N}_\Gamma^{\mathcal{M}^2}(l) = 1\ ,$$

which yields $\mathrm{d}\mathcal{N}_\Gamma^{\mathcal{M}^2}(l) = e^l/l\,\mathrm{d}l + \ldots$ for $l\to\infty$ in complete agreement with Huber's law (1.114).

In [123] an explicit formula for $\mathrm{d}\mathcal{N}_\Gamma^{\mathcal{M}^2}(l)$ was derived including an oscillating remainder term. The derivation starts from Selberg's trace formula (1.95) and uses the Möbius inversion formula in complete analogy with Riemann's explicit formula for $\pi(x)$. The formula was used to compute the lowest part of the length spectrum for the most symmetric compact Riemann surface of genus $g = 2$ using the first 200 eigenvalues, see Figure 1 in [123].

1.5 Generalizations of Weyl's Law

1.5.1 Weyl's Law for Robin Boundary Conditions

In Equations (1.66) and (1.67) we have given the three-term formula for the smooth term $\overline{N}(\lambda)$ for simply connected and bounded two-dimensional domains Ω with smooth boundary for Dirichlet as well as for Neumann boundary conditions. A generalization encountered in a nuclear physics context [124–126] are mixed or so-called *Robin boundary conditions*

$$\alpha(x)u(x) + \partial_n u(x) = 0 \quad (x\in\partial\Omega)\ , \tag{1.115}$$

which leaves the problem self-adjoint when α is real. The Dirichlet and Neumann boundary conditions are recovered in the limit $\alpha\to\infty$ and $\alpha\to 0$, respectively. For constant $\alpha \geq 0$ and excluding corners, Sieber *et al.* [127] derived the *three-term Weyl formula*

$$\begin{aligned}\overline{N}(\lambda) = {}&\frac{|\Omega|}{4\pi}\lambda - \frac{|\partial\Omega|}{4\pi}\left[1 - 2\left(\sqrt{1+\frac{\alpha^2}{\lambda}} - \frac{\alpha}{\sqrt{\lambda}}\right)\right]\sqrt{\lambda}\\ &+\left[1 - 3\frac{\sqrt{\lambda}}{\alpha}\frac{\sqrt{1+\alpha^2/\lambda}-1}{\sqrt{1+\alpha^2/\lambda}}\right]\frac{1}{12\pi}\int_{\partial\Omega}\kappa\,\mathrm{d}l\ .\end{aligned} \tag{1.116}$$

Since $\partial_n u = O\left(\sqrt{\lambda}\right)$ in the limit $\lambda\to\infty$, the term $\partial_n u$ is asymptotically dominant in the boundary condition (1.115), and hence the mean spectrum will for fixed α always tend to the Neumann case. Therefore in the derivation of (1.116), λ and $\alpha/\sqrt{\lambda}$

have been considered as independent parameters. One observes that the generalized Weyl law (1.116) interpolates neatly between the law (1.66), (1.67) for Dirichlet and Neumann boundary conditions. Formula (1.116) has been checked [127] in the case of the circle billiard, where $1/(12\pi) \int_{\partial\Omega} \kappa \, \mathrm{d}l = 1/6$, for which the exact resolvent kernel is known in closed form.

Apart from applications in nuclear physics, it was shown in [127] that the parametric dependence of the spectrum on the boundary condition is a very useful diagnostic tool in the analysis of spectra.

1.5.2
Weyl's Law for Unbounded Quantum Billiards

In Section 1.4.8 we have observed that the smooth term $\overline{N}_{\mathrm{R}}(\lambda)$ of the counting function of the nontrivial zeros of the Riemann zeta function grows asymptotically as $\lambda \ln \lambda$ which contradicts the classical eigenvalue asymptotics given by Weyl's law. Thus it appears that the interpretation of the nontrivial Riemann zeros as eigenvalues of the Laplacian is ruled out. It was pointed out, however, by Barry Simon [128, 129] that an asymptotic behavior of the form $\lambda \ln \lambda$ can occur for the eigenvalues of the two-dimensional Dirichlet Laplacian for certain unbounded regions which have a purely discrete spectrum. Since this nonclassical Weyl asymptotics again opens the possibility of identifying the nontrivial Riemann zeros with the eigenvalues of a hypothetical Riemann operator, it is important to determine also the nonleading terms of the counting function for such unbounded systems. As a representative example we here quote only the result for the so-called *hyperbola billiard* which is defined by the two-dimensional Euclidean Dirichlet Laplacian in the "horn-shaped" region

$$\Omega = \left\{(x, y) \in \mathbb{R}_+^2 : 0 \le x \cdot y \le 1\right\} .$$

It was shown by Simon [128] that this quantum system possesses a purely discrete spectrum although the corresponding classical billiard has a continuous spectrum. In [130] the following asymptotic expansion for the *trace of the heat kernel of the hyperbola billiard* was derived ($t \to 0^+$)

$$\Theta(t) := \operatorname{Tr} e^{t\Delta} = -\frac{\ln t}{4\pi t} - \frac{a'}{4\pi t} + \frac{b}{8\sqrt{\pi t}} + O\left(t^{-1/4}\right) , \tag{1.117}$$

where $a' = 2\ln(2\pi) - 1 - \gamma = 2.0985\ldots$, $b = 4\pi^{3/2}/\Gamma^2(1/4) = 1.6944\ldots$. Using the Karamata–Tauberian theorem in the form [129]: $\lim\limits_{t\to 0^+} \left[-(t^r/\ln t) \operatorname{Tr} e^{t\Delta}\right] = c$ if and only if $\lim\limits_{\lambda\to\infty} [(\lambda^{-r}/\ln\lambda)\, N(\lambda)] = c/\Gamma(r+1)$, one derives from (1.117) the leading term for the counting function

$$N(\lambda) = \frac{1}{4\pi} \lambda \ln \lambda + \ldots (\lambda \to \infty).$$

To obtain the next terms one uses a theorem by Brownell [90] which allows to obtain a smoothed counting function $\overline{N}(\lambda)$. Form (1.117) one then obtains the mean asymptotic growth of the number of eigenvalues of the hyperbola billiard [130]

$$\overline{N}(\lambda) = \frac{1}{4\pi}\lambda \ln \lambda - \frac{a}{4\pi}\lambda + \frac{b}{4\pi}\sqrt{\lambda} + O\left(\lambda^{1/4}\ln \lambda\right) \quad (\lambda \to \infty) \,, \tag{1.118}$$

where $a = 2\,(\ln(2\pi) - \gamma) = 2,5213\ldots$. While the leading term in the last expression coincides with the first term of $\overline{N}_{\mathrm{R}}(\lambda)$, Equation (1.109), the second and third terms are different.

The hyperbola billiard has been extensively investigated in classical and quantum mechanics as a model for quantum chaos [131–133]. It turns out that the classical periodic orbits can be effectively enumerated using symbolic dynamics with a ternary code, and thus the length spectrum together with the Lyapunov exponents can be calculated with high precision. The topological entropy of this system is $\tau \approx 0.6$. Using the boundary-element method, a large number of eigenvalues could be calculated. The statistics of the eigenvalues is found to be consistent with the predictions of random matrix theory for the Gaussian orthogonal ensemble. Using the semiclassical Gutzwiller trace formula, one can define a dynamical zeta function defined by an Euler product over the classical periodic orbits in analogy with the Selberg zeta function (1.100). This zeta function satisfies an approximate functional equation and thus can be effectively used as a semiclassical quantization condition in analogy to the exact quantization condition (1.102).

1.6 A Proof of Weyl's Formula

Only for very special geometries of Ω is it possible to give an explicit formula for the eigenvalues of the Dirichlet Laplacian. Such a situation had been considered in the previous sections, another is given by rectangles and cubes. Weyl's original proof for Jordan measurable domains consisted in exhausting the domain by rectangles. This proof needs technical computations which we do not want to cover here. There is another more structured proof which uses properties of the heat equation and reveals an interesting connection between the heat kernel and the eigenvalues.

Let $\Omega \subset \mathbb{R}^N$ be open and bounded with boundary $\partial\Omega$. We want to impose a mild regularity condition on Ω, namely we assume that for each $\varphi \in \mathrm{C}(\partial\Omega)$ the Dirichlet problem

$$\begin{cases} u \in \mathrm{C}(\overline{\Omega}) \cap \mathrm{C}^2(\Omega) \\ \Delta u = 0 \\ u|_{\partial\Omega} = \varphi \end{cases} \tag{$D(\varphi)$}$$

has a unique solution; i.e. we assume that Ω is *Dirichlet regular*. The Dirichlet problem is a classical subject of Potential Theory with physical interpretation in electrostatics.

There is a beautiful mathematical theory on the Dirichlet problem, and precise conditions on the boundary are known which imply Dirichlet regularity. It is a mild regularity condition on the boundary. If Ω has C^1-boundary or if Ω is a polygon, then Ω is Dirichlet regular. More generally, Lipschitz continuity of the boundary suffices. In dimension 2 each simply connected domain (i.e. each open set without holes) is Dirichet regular.

Dirichlet regularity implies that all eigenfunctions of the Dirichlet Laplacian are continuous up to the boundary i.e. they lie in the space

$$C_0(\Omega) := \left\{u \in C(\overline{\Omega}) : u|_{\partial\Omega} = 0\right\} .$$

Thus we may describe the Dirichlet Laplacian very simply by its spectral decomposition. We consider the Hilbert space $L^2(\Omega)$ with respect to the Lebesgue measure. Then there exists an orthonormal basis $\{e_n : n \in \mathbb{N}\}$ of $L^2(\Omega)$ such that

$$\begin{aligned} e_n &\in C^\infty(\Omega) \cap C_0(\Omega) , \\ -\Delta e_n &= \lambda_n e_n , \end{aligned}$$

where $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n \to \infty$. We call λ_n the n^{th} *eigenvalue of the Dirichlet Laplacian*. Now Weyl's law says that

$$\lim_{\lambda\to\infty} \frac{N(\lambda)}{\lambda^{N/2}} = \frac{\omega_N}{(4\pi)^{N/2}} |\Omega| \tag{1.119}$$

where $|\Omega|$ is the volume of Ω and $\omega_N = \pi^{N/2}\Gamma(1 + N/2)$ is the volume of the unit ball in $\mathbb{R}^N$. By $N(\lambda) = \#\{n : \lambda_n \le \lambda\}$ we denote the counting function.

For $f \in L^2(\Omega)$ we let

$$e^{t\Delta_\Omega^D} f = \sum_{n=1}^{\infty} e^{-\lambda_n t} (f|\, e_n)\, e_n , \tag{1.120}$$

where $(f|\,g) = \int_\Omega f g \,dx$ denotes the scalar product in $L^2(\Omega)$. Then $e^{t\Delta_\Omega^D}$ is a compact, self-adjoint operator on $L^2(\Omega)$. We call the family of operators $\left(e^{t\Delta_\Omega^D}\right)_{t\ge 0}$ the *semigroup generated by the Dirichlet Laplacian*. This semigroup is positive and dominated by the Gaussian semigroup $(G(t))_{t\ge0}$, i.e. for $0 \le f \in L^2(\Omega)$ we have

$$0 \le e^{t\Delta_\Omega^D} f \le G(t)f , \qquad (t > 0) \tag{1.121}$$

where

$$\begin{aligned} (G(t)f)(x) &:= \int_\Omega k_t^0(x,y) f(y)\, dy , \\ k_t^0(x,y) &:= (4\pi t)^{-N/2}\, e^{-|x-y|^2/4t} , \\ |x-y|^2 &:= \sum_{j=1}^{N} \left(x_j - y_j\right)^2 , \quad x, y \in \mathbb{R}^N . \end{aligned}$$

The domination property (1.121) implies also that $e^{t\Delta_\Omega^D}$ is defined by a measurable kernel $\tilde{k}_t(x,y)$ such that

$$0 \le \tilde{k}_t(x,y) \le k_t^0(x,y) \qquad \text{for all } x, y \in \Omega . \tag{1.122}$$

We will express the kernel $\tilde{k}_t$ in terms of the eigenfunctions in (1.124). But here we recall that those operators S on $L^2(\Omega)$ given by

$$(Sf)(x) = \int_\Omega q(x,y)f(y)\,\mathrm{d}y$$

for some $q \in L^2(\Omega \times \Omega)$ are called *Hilbert Schmidt operators*. Such a Hilbert Schmidt operator S is always compact. And if S is self-adjoint, then its eigenvalues $(\mu_n)_{n\in\mathbb{N}}$ satisfy $\sum_{n=1}^\infty \mu_n^2 < \infty$. Hence in our case

$$\sum_{n=1}^\infty \mathrm{e}^{-2t\lambda_n} < \infty \qquad \text{for all } t > 0 \,.$$

Replacing t by $t/4$ we deduce that

$$\sum_{n=1}^\infty \mathrm{e}^{-t\lambda_n/2} < \infty \qquad \text{for all } t > 0 \,. \tag{1.123}$$

Note that (1.122) implies that

$$\left|\mathrm{e}^{-\lambda_n t} e_n\right| = \left|\mathrm{e}^{t\Delta_\Omega^{\mathrm{D}}} e_n\right| \le G(t)\,|e_n| \;.$$

Since $\|e_n\|_{L^2} = 1$, it follows from the Cauchy Schwarz inequality that

$$(G(t)\,|e_n|)\,(x) \le ct^{-N/4} \;, \text{ where } c = \pi^{-N/4}2^{-(3/4)N} \,.$$

Thus

$$\left|e_n(x)\right| \le ct^{-N/4}\mathrm{e}^{\lambda_n t} \,.$$

Letting $t = 1/\lambda_n$ we obtain

$$\left|e_n(x)\right| \le \tilde{c}\lambda_n^{N/4} \qquad (x \in \Omega, n \in \mathbb{N})$$

where $\tilde{c} = c \cdot e$.

In view of (1.123), this estimate asserts that for each $t > 0$, the series

$$k_t(x,y) := \sum_{n=1}^\infty \mathrm{e}^{-\lambda_n t} e_n(x) e_n(y) \tag{1.124}$$

converges uniformly on the set $\overline{\Omega} \times \overline{\Omega}$ and defines a continuous, bounded function $k_t \colon \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$ such that $k_t(x,y) = 0$ whenever $x \in \partial\Omega$ or $y \in \partial\Omega$.

Note that

$$\left(\mathrm{e}^{t\Delta_\Omega^{\mathrm{D}}} f\right)(x) = \int_\Omega k_t(x,y)f(y)\,\mathrm{d}y \,, \tag{1.125}$$

whenever $f \in \{e_n : n \in \mathbb{N}\}$. Since the e_n form an orthonormal basis of $L^2(\Omega)$ it follows that (1.125) remains true for all $f \in L^2(\Omega)$. We have shown that the function k_t is the kernel of the operator $\mathrm{e}^{t\Delta_\Omega^{\mathrm{D}}}$ i.e. $\tilde{k}_t = k_t$.

For our purposes the following immediate consequence is crucial.

$$\int_\Omega k_t(x,x)\,\mathrm{d}x = \sum_{n=1}^{\infty} \mathrm{e}^{-\lambda_n t} \tag{1.126}$$

This formula allows us to estimate the counting function $N(\lambda) = \#\{n : \lambda_n < \lambda\}$ with the help of the kernel k_t. For this we will make use of the following Tauberian theorem due to Karamata [134].

Theorem 1.1 *Let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers such that the series $\sum_{n\in\mathbb{N}} \mathrm{e}^{-\lambda_n t}$ converges for every $t > 0$. Then for $r > 0$ and $a \in \mathbb{R}$ the following are equivalent.*

(a) $\lim\limits_{t\to 0} t^r \sum\limits_{n\in\mathbb{N}} \mathrm{e}^{-\lambda_n t} = a$

(b) $\lim\limits_{\lambda\to\infty} \lambda^{-r} N(\lambda) = \dfrac{a}{\Gamma(r+1)}$

Here N denotes the counting function $N(\lambda) = \#\{\lambda_n \le \lambda\}$, and $\Gamma(r) = \int_0^\infty x^{r-1}\mathrm{e}^{-x}\,\mathrm{d}x$ is the usual Gamma function.

Combining formula (1.126) and Theorem 1.1 we see that Weyl's law (1.119) is equivalent to the kernel estimate

$$\lim_{t\to 0} t^{N/2} \int_\Omega k_t(x,x)\,\mathrm{d}x = \frac{|\Omega|}{(4\pi)^{N/2}} \,. \tag{1.127}$$

It is easily seen that the left-hand side of (1.127) is not greater than the right-hand side as the kernel k_t is bounded by the *Gaussian kernel* i. e. $k_t(x,y) \le k_t^0(x,y)$ for $x, y \in \Omega$, $t > 0$.

The lower estimate is more delicate. For this we will consider the heat equation on the infinite cylinder $\mathbb{R}_+ \times \overline{\Omega}$ whose boundary we denote by $\Gamma = (\{0\} \times \overline{\Omega}) \cup ((0,\infty) \times \partial\Omega)$. It is a remarkable fact that Dirichlet regularity of Ω also implies that the following boundary value problem for the heat equation is well-posed.

Theorem 1.2 ([135, Theorem 6.2.8], [136]) *Let $\psi \in C(\Gamma)$. Then there exists a unique solution of*

$$\begin{cases} u \in C\left(\mathbb{R}_+ \times \overline{\Omega}\right) \cap C^\infty\left((0,\infty) \times \Omega\right), \\ \dfrac{\partial}{\partial t} u(t,x) = \Delta u(t,x)\,, \qquad (t > 0, x \in \Omega) \\ u|_\Gamma = \psi\,. \end{cases} \tag{1.128}$$

This solution satisfies the parabolic maximum principle*, which says that for all $t > 0$ and all $0 \le s \le t$, $x \in \overline{\Omega}$,*

$$u(s,x) \le \max_{\Gamma_t} u$$

where $\Gamma_t := \Gamma \cap \left([0,t] \times \overline{\Omega}\right)$.

Example 1.1 *Let $f \in C_0(\Omega)$ and define $\psi \in C(\Gamma)$ by $\psi(0, x) = f(x)$ for $x \in \overline{\Omega}$, and $\psi(t, z) = 0$ for $t > 0$, $z \in \partial\Omega$. Then the solution of* (1.128) *is given by $u(t, x) = \left(e^{t\Delta_\Omega^D} f\right)(x)$. Thus, the semigroup $\left(e^{t\Delta_\Omega^D}\right)_{t\geq 0}$ governs the homogeneous boundary value problem* (1.128). *Its solution can be expressed by the kernel k_t, namely,*

$$u(t, x) = \int_\Omega k_t(x, y) f(y)\,dy .$$

For this reason we call k_t the heat kernel associated with the Dirichlet Laplacian.

To obtain a lower bound for the kernel we formalize the idea that at some distance away from the boundary, k_t behaves just like the Gaussian kernel.

Lemma 1.1 *Let $x \in \Omega$ be arbitrary, and for $y \in \Omega$ let $t_0(y) := \operatorname{dist}(y, \partial\Omega)^2/2N$ denote the scaled squared distance of y to the boundary of Ω. Then*

$$k_t^0(x, y) - k_t(x, y) \leq \begin{cases} (4\pi t)^{-N/2}\, e^{-\operatorname{dist}(y,\partial\Omega)^2/4t} , & t \leq t_0(y) , \\ (4\pi t_0(y))^{-N/2}\, e^{-N/2} , & t > t_0(y) . \end{cases}$$

Proof Fix $y \in \Omega$. Then by Theorem 1.2 there exists a unique function $p(\cdot, \cdot, y)$ solving the parabolic boundary value problem

$$\begin{cases} p(\cdot, \cdot, y) \in C\left(\mathbb{R}_+ \times \overline{\Omega}\right) \cap C^\infty\left((0, \infty) \times \Omega\right) , & \\ \dfrac{\partial}{\partial t} p(t, x, y) = \Delta_x p(t, x, y) , & (t > 0, x \in \Omega) \\ p(t, x, y) = 0 , & (x \in \overline{\Omega}) \\ p(t, x, y) = (4\pi t)^{-N/2}\, e^{-|x-y|^2/4t} . & (t > 0, x \in \partial\Omega) \end{cases}$$

Then $p(t, x, y) = k_t^0(x, y) - k_t(x, y)$. In fact, let $f \in C_0(\Omega)$ be arbitrary, and let

$$u(t, x) := \int_\Omega \left(k_t^0(x, y) f(y) - p(t, x, y) f(y)\right) dy .$$

The properties $u \in C^\infty\left((0, \infty) \times \Omega\right)$, $u_t = \Delta u$ on $(0, \infty) \times \Omega$ and $u(t, x) = 0$ if $x \in \partial\Omega$, $t > 0$ are obvious. Moreover, it is easy to prove that u can be continuously extended to $t = 0$ and $u(0, x) = f(x)$ for all $x \in \overline{\Omega}$. Thus $u(t, \cdot) = e^{t\Delta_\Omega^D} f$ according to Example 1.1.

Since p solves a parabolic problem, we can use the parabolic maximum principle to deduce that p attains its maximum on the boundary i. e.

$$p(t, x) \leq \sup_{\substack{0\leq s\leq t \\ x\in\partial\Omega}} (4\pi s)^{-N/2}\, e^{-|x-y|^2/4s} \leq \sup_{0\leq s\leq t} (4\pi s)^{-N/2}\, e^{-\operatorname{dist}(y,\partial\Omega)^2/4s} . \tag{1.129}$$

Calculating the derivative of $(4\pi t)^{-N/2}\, e^{-\operatorname{dist}(y,\partial\Omega)^2/4t}$ as a function in the variable t one sees that the maximum is attained at time $t = t_0(y)$. We can thus simplify (1.129) accordingly which completes the proof. □

We are interested in the error $\int_\Omega \left(k_t^0(x,x) - k(x,x)\right) \mathrm{d}x$ of the approximation of k_t^0 by k_t as t tends to 0. Since the lemma essentially says that problems may only arise near the boundary, it is natural to decompose Ω into a good part $\Omega_1(t) := \left\{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \geq t^{1/4}\right\}$ and a bad part $\Omega_2(t) := \Omega \setminus \Omega_1(t)$. Note that $\left|\Omega_2(t)\right| \to 0$ as $t \to 0$. If $t \leq 1/4N^2$, then for every $x \in \Omega_1(t)$ we have $t_0(x) \geq \sqrt{t}/2N \geq t$. Hence we can apply the lemma to obtain

$$t^{N/2} \int_{\Omega_1(t)} \left(k_t^0(x,x) - k_t(x,x)\right) \mathrm{d}x \leq |\Omega| \, (4\pi)^{-N/2} \, \mathrm{e}^{-\sqrt{t}/4t} \to 0 \quad (t \to 0) \, .$$

On the other hand, using the trivial estimate $k_t \geq 0$ we see

$$t^{N/2} \int_{\Omega_2(t)} \left(k_t^0(x,x) - k_t(x,x)\right) \mathrm{d}x \leq \left|\Omega_2(t)\right| (4\pi)^{-N/2} \to 0 \quad (t \to 0) \, .$$

Combining these two estimates, we have proved

$$\liminf_{t \to 0} \, t^{N/2} \int_\Omega k_t(x,x) \, \mathrm{d}x \geq \liminf_{t \to 0} \, t^{N/2} \int_\Omega k_t^0(x,x) \, \mathrm{d}x = \frac{|\Omega|}{(4\pi)^{N/2}}$$

This was the missing inequality required to prove (1.127). Since (1.127) has been shown to be equivalent to Weyl's law, we have completed the proof.

Weyl's law also holds for arbitrary bounded open sets, [137, Theorem 1.11]. A simple proof by approximating an arbitrary open set by regular sets from the interior is given in [138, Section 6.5.2]. For further results on domain approximation we refer to the survey article [139] by Daners.

The proof given here is essentially the one given by Kac [9] who found formula (1.126) and used Karamata's Tauberian theorem. We were also inspired by Dodzink [140] and the Diploma thesis by E. Michel [141]. However, the use of Dirichlet regularity and Theorem 1.2 in particular comes from [138, Chapter 5] where more details can be found. Concerning the Dirichlet problem we refer to [142, 143] and the literature mentioned there.

1.7
Can One Hear the Shape of a Drum?

Weyl's law shows us in particular the following. Assume that $\Omega \subset \mathbb{R}^N$ is a bounded open set and we know all the eigenvalues of the Dirichlet Laplacian. Then we also know the volume of Ω. Thus the spectrum of the Dirichlet Laplacian determines the volume. It is natural to ask whether there are other properties or qualities which we may deduce from the spectrum. Those types of questions are called *inverse (spectral) problems*. Let us say that two open bounded sets Ω_1 and Ω_2 in $\mathbb{R}^N$ are *isospectral* if the corresponding Dirichlet Laplacians $\Delta_{\Omega_1}^{\mathrm{D}}$ and $\Delta_{\Omega_2}^{\mathrm{D}}$ have the same sequence of eigenvalues. We already know that isospectral sets have the same volume. There is another result of this kind.

Theorem 1.3 *Let $\Omega_1, \Omega_2 \subset \mathbb{R}^N$ be open bounded sets with Lipschitz boundary. If Ω_1 and Ω_2 are isospectral, then they have the same surface area.*

Here we use the natural measure σ on the boundary $\partial\Omega_i$ of Ω_i i. e. the surface measure or (which is the same) the $(N-1)$-dimensional Hausdorff measure. The surface area of Ω_i is by definition $\sigma(\partial\Omega_i)$. For a proof, we refer to [144].

The most radical inverse spectral problem is whether the spectrum determines the domain completely. This question became famous by Marc Kac's article [9] from 1966. We want to formulate it more precisely. Two open sets $\Omega_1, \Omega_2 \subset \mathbb{R}^N$ are called *congruent* if there exists an orthogonal matrix B and a vector b in $\mathbb{R}^N$ such that $\Omega_2 = \{Bx + b : x \in \Omega_1\}$. This is just congruence in the Euclidean sense. It is obvious that congruent open sets are isospectral.

Question 1.1 (Kac's Question) *Let $\Omega_1, \Omega_2 \subset \mathbb{R}^2$ be two bounded smooth domains which are isospectral. Are they necessarily congruent?*

By a *domain* we mean an open connected set. An open bounded set is called *smooth* if the boundary is of class C^∞.

Kac's question became so popular because it has a fascinating physical interpretation. We consider a bounded smooth domain $\Omega \subset \mathbb{R}^2$ as a membrane which is fixed at the boundary Γ of Ω. If it is set into motion, then the vertical displacement $u(t, x)$ at time $t > 0$ at the point $x \in \Omega$ satisfies the wave equation

$$u_{tt} = c\Delta u(t, x) \qquad (t > 0, x \in \Omega) \ .$$

We normalize physical units in such a way that $c = 1$.

Of particular interest are solutions of the form $u(t, x) = v(x)e^{i\omega t}$ which are called the *pure tones* of the membrane. In order that such u be a solution of the wave equation it is necessary and sufficient that

$$-\Delta v = \omega^2 v \ .$$

Thus u is a solution if and only if v is an eigenfunction of the Dirichlet Laplacian for the eigenvalue ω^2, where ω is the frequency of the displacement u. Now we see that the eigenvalues of the Dirichlet Laplacian correspond exactly to the pure tones of the membrane which we can hear. This lead Kac to reformulate his question by asking "Can one hear the shape of a drum?". Following Kac, people like to formulate inverse spectral problems by asking which properties of Ω one can hear. For example, we already know that we can hear the volume and the surface area of a Lipschitz domain.

Kac himself said in [9]: "I believe that one cannot hear the shape of a tambourine but I may be wrong and I am not prepared to bet large sums either way."

Today the question raised by Kac is still open. But much more is known about it. In fact, we may ask more generally if two bounded isospectral domains in $\mathbb{R}^N$ are congruent. That is, we consider arbitrary dimensions now and give up the very restrictive smoothness hypothesis. Let us note though that some hypothesis on the boundary is needed to avoid trivialities. For instance, if we consider the disc $\Omega_1 = \left\{x \in \mathbb{R}^2 : |x| < 1\right\}$ and the punctured disc $\Omega_2 = \Omega_1 \setminus \{0\}$, then they are isospectral but not congruent. In fact, $L^2(\Omega_1) = L^2(\Omega_2)$ and also the Dirichlet Laplacians with respect to these two open sets are identical. We will describe below precisely which regularity of the boundary is needed to avoid such simple counterexamples. *Here we want to impose throughout that all bounded domains have a Lipschitz boundary*, and we call them *Lipschitz domains* for short. They include all polygons in particular.

Before we describe some of the results concerning Kac's question we mention that the analogous question for compact manifolds has a negative answer as John Milnor [70] had already shown in 1964. So the challenge concerns the Euclidean case. A first counterexample was given by Urakawa [145] in 1982 who constructed two isospectral Lipschitz domains in $\mathbb{R}^4$ which are not congruent. Ten years later, Gordon, Webb and Wolpert [146] found a two-dimensional example. By putting together seven triangles they obtained two polygons in $\mathbb{R}^2$ which are isospectral but not congruent, see Figure 1.5. These two polygons are not convex, though. It is an open question whether convex isospectral polygons in $\mathbb{R}^2$ are congruent. However, in four dimensions convexity alone does not help. There are convex isospectal sets which are not congruent. In fact, by modifying Urakawa's example, Gordon and Webb [147] obtained two truncated convex cones in $\mathbb{R}^4$ which are isospectral but not congruent. These cones are induced by some vector space bases in $\mathbb{R}^4$. Here is an explicit formulation.

Example 1.2 (Gordon, Webb) *Let*

$$u_1 := \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, u_2 := \begin{pmatrix}1\\-1\\0\\0\end{pmatrix}, u_3 := \begin{pmatrix}1\\1\\1\\0\end{pmatrix}, u_4 := \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

be the first basis of $\mathbb{R}^4$ *and*

$$v_1 := \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, v_2 := \begin{pmatrix}1\\\sqrt{3}\\0\\0\end{pmatrix}, v_3 := \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, v_4 := \begin{pmatrix}0\\0\\1\\1\end{pmatrix}$$

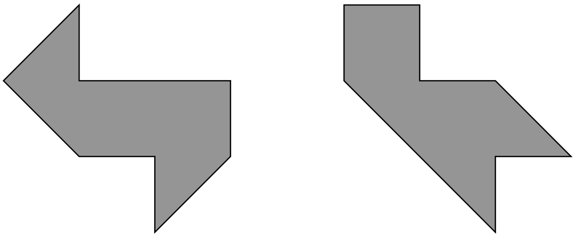


Figure 1.5 Isospectral polygons in two dimensions.

the second. Consider the corresponding positive cones

$$C_1 := \left\{ \sum_{i=1}^{4} a_i u_i : a_i \geq 0, i = 1, \ldots, 4 \right\}, \quad C_2 := \left\{ \sum_{i=1}^{4} a_i v_i : a_i \geq 0, i = 1, \ldots, 4 \right\}.$$

Let $B_0 := \left\{ x \in \mathbb{R}^4 : 0 < |x| < 1 \right\}$ *be the punctured unit ball in* $\mathbb{R}^4$ *with respect to the Euclidean norm* $|x| = \sqrt{\sum_{j=1}^4 x_j^2}$. *Then* $\Omega_1 := B_0 \cap C_1$ *and* $\Omega_2 := B_0 \cap C_2$ *are isospectral but not congruent.*

So far no smooth counterexample is known in any dimension. But in a very recent work Zelditch [148] showed that isospectral domains with an analytic boundary, having some symmetry, are congruent. A simple class of domains having such a symmetry are ellipses and stadiums. Thus he shows in particular that those domains can be distinguished by their spectra.

Now we describe further positive results. We mention that two isospectral triangles are congruent, see [149] and references therein. Moreover, one can hear whether a Lipschitz domain in $\mathbb{R}^N$ is a ball.

Theorem 1.4 *Let* $\Omega_1 \subset \mathbb{R}^N$ *be a ball and* $\Omega_2 \subset \mathbb{R}^N$ *a Lipschitz domain. If* Ω_1 *and* Ω_2 *are isospectral, then they are congruent.*

Proof If Ω is a Lipschitz domain, then one can hear its volume $|\Omega|$ according to Weyl's law. The Faber–Krahn inequality

$$\lambda_1^{\Omega} \geq c_N |\Omega|^{-2/N} \tag{1.130}$$

holds for all such domains, where λ_1^{Ω} denotes the first eigenvalue of the Dirichlet Laplacian on Ω and c_N is an optimal constant which depends only on the dimension N [150, Theorem 3.1]. Moreover, (1.130) is an equality if and only if Ω is a ball, see [151, Theorem 1.2]. □

The above theorem can be found in Kac's paper [9]. However, Kac uses the isoperimetric inequality together with Theorem 1.3 instead of (1.130). For this argument one has to be able to define the surface area of the domain. The above proof on the other hand works in much more generality. The result can even be made optimal in a sense that we will describe now. For this, we need the notion of *capacity* which is used to describe the size of sets in $\mathbb{R}^N$ in terms of Sobolev norms. For a systematic introduction we refer to [152]. The capacity cap(A) of a set $A \subset \mathbb{R}^N$ may be any number in $[0, \infty]$, but here we only need to know whether a set has capacity 0. Sets of capacity 0 are also called *polar sets.* Although it is not trivial to characterize all polar sets, thinking of them as subsets of $\mathbb{R}^N$ of dimension at most $N - 2$ gives a good impression of how they look. For example, single points in $\mathbb{R}^2$ and smooth curves in $\mathbb{R}^3$ are polar, but curves in $\mathbb{R}^2$ and surfaces in $\mathbb{R}^3$ are not polar. Moreover, subsets of polar sets and countable unions of polar sets are also polar.

What makes the notion of capacity paticularly interesting is the fact that the Dirichlet Laplacian "does not see" polar sets. More precisely, if Ω_1 and Ω_2 are open subsets of $\mathbb{R}^N$ that only differ by a polar set i.e. $\Omega_2 \setminus \Omega_1$ and $\Omega_1 \setminus \Omega_2$ are both polar, then the sets differ only by a set of Lebesgue measure zero, hence $L^2(\Omega_1) = L^2(\Omega_2)$ as subspaces of $L^2(\mathbb{R}^N)$. But in fact, $\Delta_{\mathrm{D}}^{\Omega_1} = \Delta_{\mathrm{D}}^{\Omega_2}$ as operators on this space, thus they have the same spectrum. This shows that inverse spectral problems for the Dirichlet Laplacian are meaningful only up to polar sets. Thus we are lead to introduce a notion of regularity which asserts that there are no artificial "polar holes" in the set. More precisely, call an open set Ω in $\mathbb{R}^N$ *regular in capacity* if $\operatorname{cap}(B(z,r) \setminus \Omega) > 0$ for all $z \in \partial\Omega$ and all $r > 0$, where $B(z,r)$ denotes the ball of radius r centered in z. We refer to [153] where this regularity assumption is introduced and discussed. Here we only mention that all Dirichlet regular sets, and hence all Lipschitz domains, are regular in capacity.

Given any open set $\Omega \subset \mathbb{R}^N$, there exists a unique open set Ω' which is regular in capacity such that $\Omega \subset \Omega'$ and $\operatorname{cap}(\Omega' \setminus \Omega) = 0$. Since the Laplacian does not see polar sets it is natural to consider merely open sets which are regular in capacity. An inspection of Daners' proof [151] shows that for a bounded open set Ω which is regular in capacity the Faber–Krahn inequality becomes an identity if and only if Ω is a ball. Thus Theorem 1.4 remains true if we assume that Ω_2 is regular in capacity instead of being a Lipschitz domain. In other words, if Ω_2 is an arbitrary open set which is isospectral to a ball Ω_1, then the regular version of Ω_2 is a ball of the same radius, or, what is the same, there exists a ball $B \subset \mathbb{R}^N$ which is a translation of Ω_1 such that $\Omega_2 \subset B$ and $\operatorname{cap}(B \setminus \Omega_2) = 0$.

1.8 Does Diffusion Determine the Domain?

In this short section we follow a paradigm which is slightly different from that in the last section. Instead of the wave equation let us consider the diffusion equation

$$\left\{\begin{array}{rcll} u_t(t,x) & = & \Delta u(t,x) & (t > 0, x \in \Omega)\,, \\ u(0,x) & = & u_0(x) & (x \in \Omega)\,, \\ u(t,z) & = & 0 & (z \in \partial\Omega)\ . \end{array}\right. \tag{D}$$

Here again Ω is a Lipschitz domain with boundary Γ. The solution u of (D) has the following interpretation. Assume that Ω is a body containing water and some dissolving liquid, for instance ink. Then u_0 is the initial concentration of the ink i. e. for $\omega \subset \Omega$ the amount of ink in ω is given by $\int_\omega u_0(x)\,\mathrm{d}x$. The solution $u(t,x)$ gives the concentration at time $t > 0$ i. e. for $\omega \subset \Omega$, $\int_\omega u(t,x)\,\mathrm{d}x$ is the amount of ink in ω at time t.

Given $u_0 \in L^2(\Omega)$, Equation (D) has a unique solution $u : \mathbb{R}_+ \to L^2(\Omega)$, where we let $u(t,x) = u(t)(x)$, given by

$$u(t) = \mathrm{e}^{t\Delta_\Omega^{\mathrm{D}}} u_0 = \sum_{n\in\mathbb{N}} \mathrm{e}^{-\lambda_n t} (u_0 \mid e_n)\, e_n\ ,$$

(compare (1.120)). In fact, since $(\mathrm{d}/\mathrm{d}t)\mathrm{e}^{-\lambda_n t}e_n = \mathrm{e}^{-\lambda_n t}(-\lambda_n e_n) = \mathrm{e}^{-\lambda_n t}\Delta e_n$, u is a solution of (D). Its uniqueness follows from Theorem 1.128, the parabolic maximum principle. Thus the semigroup generated by $\Delta_\Omega^{\mathrm{D}}$, $\mathrm{e}^{-t\Delta_\Omega^{\mathrm{D}}}$, is frequently called the *diffusion semigroup*.

Now let Ω_1 and Ω_2 be two Lipschitz domains. If Ω_1 and Ω_2 are isospectral, then we find orthonormal bases $(e_n)_{n\in\mathbb{N}}$ of $L^2(\Omega_1)$ and $(f_n)_{n\in\mathbb{N}}$ of $L^2(\Omega_2)$ such that

$$-\Delta_{\Omega_1}^{\mathrm{D}} e_n = \lambda_n e_n \qquad \text{and} \qquad -\Delta_{\Omega_2}^{\mathrm{D}} f_n = \lambda_n f_n$$

for all $n \in \mathbb{N}$. Consider the unitary operator $U : L^2(\Omega_1) \to L^2(\Omega_2)$ satisfying $Ue_n = f_n$. Then

$$U\mathrm{e}^{t\Delta_{\Omega_1}^{\mathrm{D}}} = \mathrm{e}^{t\Delta_{\Omega_2}^{\mathrm{D}}} U \qquad (t > 0)\ , \tag{1.131}$$

i. e. *U intertwines* the two diffusion semigroups. In other words, U maps solutions of the first diffusion equation to solutions of the other diffusion equation. Conversely, if we find an intertwining invertible operator $U : L^2(\Omega_1) \to L^2(\Omega_2)$, then Ω_1 and Ω_2 are isospectral. Now we remember that for the physical interpretation only positive concentrations $0 \le u_0 \in L^2(\Omega_1)$ are meaningful. If $u_0(x) \ge 0$ for all $x \in \Omega_1$, then $u(t, x) \ge 0$ for all $x \in \Omega_1$ and all $t > 0$. This is the positivity property of the diffusion equation. The physical interpretation motivates us to consider, instead of unitary operators, operators U which preserve positivity. A linear bijective mapping $U : L^2(\Omega_1) \to L^2(\Omega_2)$ is called an *order isomorphism* if for all $f \in L^2(\Omega_1)$, $f \ge 0$ if and only if $Uf \ge 0$. If in (1.131) instead of unitary we assume that U is an order isomorphism, then we obtain a positive result.

Theorem 1.5 *Let Ω_1 and Ω_2 be two Lipschitz domains in $\mathbb{R}^N$. Assume that there exists an order isomorphism $U : L^2(\Omega_1) \to L^2(\Omega_2)$ such that* (1.131) *holds. Then Ω_1 and Ω_2 are congruent.*

For a proof, we refer to [153, Corollary 3.17]. We remark that this result also remains true if we only assume the domains to be regular in capacity.

This theorem is no longer a purely spectral problem, but it is an inverse problem. To say that U is an intertwining order isomorphism is the same as saying that U maps positive solutions to positive solutions. Thus we may rephrase the result by saying that "Diffusion determines the domain".

References

1 Weyl, H. (**1911**) Über die asymptotische Verteilung der Eigenwerte. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen. Mathem.-physikal. Klasse*, 110–117.

2 Weyl, H. (**1912**) Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Mathematische Annalen*, **71**(4), 441–479.

3 Weyl, H. (**1912**) Über die Abhängigkeit der Eigenschwingungen einer Membran von deren Begrenzung. *J. Reine Angew. Math.*, **141**, 1–11.

4 Weyl, H. (**1912**) Über das Spektrum der Hohlraumstrahlung. *J. Reine Angew. Math.*, **141**, 163–181.

5 Weyl, H. (**1913**) Über die Randwertaufgabe der Strahlungstheorie und asymptotische Spektralgeometrie. *J. Reine Angew. Math.*, **143**, 177–202.

6 Ivrii, V.Y. (**1980**) Second term of the spectral asymptotic expansion of the Laplace–Beltrami operator on manifolds with boundary. *Functional Analysis and Its Applications*, **14**(2), 98–106.

7 Melrose, R. (**1980**) Weyl's conjecture for manifolds with concave boundary. *Proc. Sympos. Pure Math.*, **36**, 257–274.

8 Weyl, H. (**1915**) Das asymptotische Verteilungsgesetz der Eigenschwingungen eines beliebig gestalteten elastischen Körpers. *Rend. Circ. Mat. Palermo*, **39**, 1–50.

9 Kac, M. (**1966**) Can One Hear the Shape of a Drum? *The American Mathematical Monthly*, **73**(4), 1–23.

10 Sommerfeld, A. (**1910**) Die Greensche Funktion der Schwingungsgleichung für ein beliebiges Gebiet. *Physikal. Zeitschr.*, **11**, 1057–1066.

11 Lorentz, H.A. (**1910**) Alte und neue Fragen der Physik. *Physikal. Zeitschr.*, **11**, 1234–1257.

12 Born, M., Einstein, A. (**1972**) *Briefwechsel 1916–1955.* Rowohlt Verlag, 53–55.

13 Reudler, J. (**1912**) PhD thesis, Leiden.

14 Weyl, H. (**1950**) Ramifications, old and new, of the eigenvalue problem. *Bull. Amer. Math. Soc.*, **56**(2), 115–139.

15 Pais, A. (**1982**) *"Subtle is the Lord ...", The Science and the Life of Albert Einstein.* Oxford University Press.

16 Kirchhoff, G. (**1860**) Ueber das Verhältniss zwischen dem Emissionsvermögen und dem Absorptionsvermögen der Körper für Wärme und Licht. *Annalen der Physik und Chemie*, **19**, 275–301.

17 Hilbert, D. (**1912**) Begündung der elementaren Strahlungstheorie. *Physik. Zeitschrift*, 13:1056–1064.

18 Hilbert, D. (**1913**) Bemerkung zur Begründung der elementaren Strahlungstheorie. *Physik. Zeitschrift*, **14**, 592–595.

19 Hilbert, D. (**1914**) Zur Begründung der elementaren Strahlungstheorie. Dritte Mitteilung. *Physik. Zeitschrift*, **15**, 878–889.

20 Boltzmann, L. (**1884**) Ableitung des Stefan'schen Gesetzes, betreffend die Abhängigkeit der Wärmestrahlung von der Temperatur aus der electromagnetischen Lichttheorie. *Annalen der Physik und Chemie*, **22**, 291–294.

21 Wien, W. (**1893**) Eine neue Beziehung der Strahlung schwarzer Körper zum zweiten Hauptsatz der Wärmetheorie. *Sitzungsberichte der Königlichen Preußischen Akademie der Wissenschaften zu Berlin*, 55–62.

22 Paschen, F. (**1896**) Ueber Gesetzmässigkeiten in den Spectren fester Körper. (Erste Mittheilung). *Annalen der Physik und Chemie*, **58**, 455–492.

23 Einstein, A. (**1913**) Max Planck als Forscher. *Naturwissenschaften*, **1**(45), 1077–1079.

24 Wien, W. (**1896**) Ueber die Energievertheilung im Emissionsspectrum eines schwarzen Körpers. *Annalen der Physik und Chemie*, **58**, 662–669.

25 Paschen, F. (**1897**) Ueber Gesetzmässigkeiten in den Spectren fester Körper. (Zweite Mittheilung). *Annalen der Physik und Chemie*, **60**, 662–723.

26 Rayleigh, L. (**1900**) Remarks upon the law of complete radiation. *Phil. Mag.*, **49**, 539–540.

27 Rayleigh, L. (**1896**) *The Theory of Sound*, volume 2. Reprinted by Dover Publications, New York, 1945. First edition printed 1877; second revised and enlarged edition.

28 Lummer, O. and Pringsheim, E. (**1900**) Über die Strahlung des schwarzen Körpers für lange Wellen. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, **2**, 163–180.

29 Rubens, H. and Kurlbaum, F. (**1900**) Über die Emission langwelliger Wärmestrahlen durch den schwarzen Körper bei verschiedenen Temperaturen. *Sitzungsberichte der Preußischen Akademie der Wissenschaften*, 929–941.

30 Planck, M. (**1900**) Über eine Verbesserung der Wienschen Spektralgleichung. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, **2**, 202–204.

31 Weyl, H. (**1931**) *The theory of groups and quantum mechanics*. Dover Publications, New York. Translation of the second (revised) German Edition.

32 Planck, M. (**1966**) *Theorie der Wärmestrahlung*. Johann Ambrosius Barth, Leipzig, 6th edition.

33 Planck, M. (**1900**) Zur Theorie des Gesetzes der Energieverteilung im Normalspectrum. *Verhandlungen der Deutschen Physikalischen Gesellschaft*, **2**, 237–245.

34 Planck, M. (**1901**) Ueber das Gesetz der Energieverteilung im Normalspectrum. *Annalen der Physik*, **4**, 553–563.

35 Lorentz, H.A. (**1903**) On the emission and absorbtion by metals of rays of heat of great wave-lengths. *Proc. Acad. Amsterdam*, **5**, 666–685.

36 Einstein, A. (**1905**) Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt. *Annalen der Physik*, **17**, 132–148.

37 Rayleigh, L. (**1905**) The dynamical theory of gases and of radiation. *Nature*, **72**, 54–55.

38 Rayleigh, L. (**1905**) The constant of radiation as calculated from molecular data. *Nature*, **72**, 243–244.

39 Jeans, J.H. (**1905**) The dynamical theory of gases and of radiation. *Nature*, **72**, 101–102.

40 Jeans, J.H. (**1905**) On the partition of energy between matter and aether. *Phil. Mag.*, **10**, 91–98.

41 Jeans, J.H. (**1905**) A comparison between two theories of radiation. *Nature*, **72**, 293–294.

42 Jeans, J.H. (**1905**) On the laws of radiation. *Proc. R. Soc. London A*, **76**(513), 545–552.

43 Einstein, A. (**1949**) *Albert Einstein: Philosopher-Scientist*, (ed P.A. Schilpp), Cambridge University Press, London, 43.

44 Jeans, J.H. (**1905**) The dynamical theory of gases. *Nature*, 71:607.

45 Selberg, A. (**1956**) Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc.*, **20**, 47–87.

46 Selberg, A. (**1962**) Discontinuous groups and harmonic analysis, in *Proc. Int. Math. Congr. Stockholm*, 177–189.

47 Steiner, F. (**1987**) On Selberg's zeta function for compact Riemann surfaces. *Phys. Lett. B*, **188**, 447–454.

48 Gauss, C.F. (**1981**) De nexu inter multitudinem classium, in quas formae binariae secundi gradus distribuntur, earumque determinantem, in *Werke*, **2**, 269–280. Georg Olms Verlag, Hildesheim, New York.

49 Hardy, G.H. and Wright, E.M. (**2002**) *An Introduction to the Theory of Numbers*, Clarendon Press, Oxford.

50 Hilbert, D. and Cohn-Vossen, S. (**1952**) *Geometry and the Imagination (Anschauliche Geometrie)*, Chelsea Publishing Company.

51 Landau, E. (**1927**) *Vorlesungen über Zahlentheorie*, volume 2, Hirzel Verlag, Leipzig.

52 Sierpiński, W.M. (**1974**) *Oevres Choisies*, volume 1, Polish Scientific Publishers PWN, Warsaw.

53 Hardy, G.H. (**1915**) The average order of the arithmetical functions $P(x)$ and $\Delta(x)$. *Proc. London Math. Soc.*, **15**, 192–213.

54 Hardy, G.H.(**1915**) On the expression of a number as the sum of two squares. *Quart. J. of Math.*, **46**, 263–283.

55 van der Corput, J.G. (**1923**) Neue zahlentheoretische Abschätzungen, erste Mitteilung. *Math. Annalen*, **89**, 215–254.

56 Yin, W.-L. (**1962**) The lattice points in a circle. *Sci. Sinica.*, **11**, 10–15.

57 Iwaniec, H. and Mozzochi, C.J. (**1988**) On the divisor and circle problems. *J. Number Theory*, **29**, 60–93.

58 Huxley, M.N. (**1993**) Exponetial sums and lattice points II. *Proc. London Math. Soc.*, **66**, 279–301.

59 Cappell, S.E. and Shaneson, J.L. (**2007**) Some problems in number theory I: The circle problem. arXiv:math/0702613v3 [math.NT].

60 Hardy, G.H. and Landau, E. (**1925**) The lattice points of a circle. *Proc. Roy. Soc London A*, **105**, 244–258.

61 Hardy, G.H. (**1925**) The lattice points of a circle. *Proc. Roy. Soc London A*, **107**, 623–635.

62 Voronoï, G. (**1904**) Sur le développement, à l'aide des fonctions cylindriques, des sommes doubles $\sum f(pm^2 + 2qmn + rn^2)$. *Verh. Math. Kongr. Heidelberg*, 241–245.

63 Steiner, F. (**2003**) Space–time approach to quantum chaos. *physica status solidi (b)*, **237**, 133–145.

64 Aurich, R., Sieber, M. and Steiner, F. (**1988**) Quantum chaos of the Hadamard–Gutzwiller model. *Phys. Rev. Lett.*, **61**(5), 483–487.

65 Aurich, R. and Steiner, F. (**1989**) Periodic-orbit sum rules for the Hadamard-Gutzwiller model. *Physica D*, **39**, 169–193.

66 Aurich, R. and Steiner, F. (**1992**) From classical periodic orbits to the quantization of chaos. *Proc. R. Soc. Lond. A*, **437**, 693–714.

67 Bolte, J. and Steiner, F. (**1990**) Determinants of Laplace-like operators on Riemann surfaces. *Commun. Math. Phys.*, **130**, 581–597.

68 Bolte, J. and Steiner, F. (**1993**) The Selberg trace formula for bordered Riemann surfaces. *Commun. Math. Phys.*, **156**, 1–16.

69 Aurich, R., Janzer, H.S., Lustig, S. and Steiner, F. (**2008**) Do we live in a "small Universe"? Classical and Quantum Gravity, 25, 125006, 1–12.

70 Milnor, J. (**1964**) Eigenvalues of the Laplace Operator on Certain Manifolds. *Proc. Nat. Acad. Sci. USA*, **51**(4), 542ff.

71 Witt, E. (**1941**) Eine Identität zwischen Modulformen zweiten Grades. *Abh. Math. Sem. Hansischen Universität*, **14**, 323–337.

72 Chazarain, J. (**1974**) Formule de poisson pour les variétés riemanniennes. *Invent. Math.*, **24**, 65–82.

73 Colin de Verdière, Y. (**1973**) Spectre du laplacien et longueurs des géodésiques périodiques. i. *Compos. Math.*, **27**(1), 83–106.

74 Colin de Verdière, Y. (**1973**) Spectre du laplacien et longueurs des géodésiques périodiques. ii. *Compos. Math.*, **27**(2), 159–184.

75 Duistermaat, J.J. and Guillemin, V.W. (**1975**) The spectrum of positive elliptic operators and periodic bicharacteristics. *Invent. Math.*, **29**, 39–79.

76 Guillemin, V.W. (**1977**) Lectures on spectral theory of elliptic operators. *Duke Math. J.*, **44**, 485–517.

77 Minakshisundaram, S. and Pleijel, Å. (**1949**) Some properties of the eigenfunctions of the Laplace-operator on Riemannian manifolds. *Can. J. Math.*, **1**, 242–256.

78 Minakshisundaram, S. (**1949**) A generalization of Epstein zeta function. *Can. J. Math.*, **1**, 320–329.

79 Titchmarsh, E.C. (**1948**) *Introduction to the theory of Fourier integrals.* Oxford University Press, 2nd edition.

80 Ray, D. and Singer, I.M. (**1973**) Analytic torsion for complex manifolds. *Ann. Math.*, **98**, 154–177.

81 Hawking, S. (**1977**) Zeta function regularization of path integrals in curved space time. *Commun. Math. Phys.*, **55**, 133–148.

82 Grosche, C. and Steiner, F. (**1998**) *Handbook of Feynman path integrals*, volume 145 of *Springer Tracts in Modern Physics*. Springer.

83 Hughes, C.P. and Rudnick, Z. (**2004**) On the distribution of lattice points in thin annuli. *Int. Math. Research Notices*, **13**, 637–658.

84 Cramér, H. (**1922**) Über zwei Sätze von Herrn G.H. Hardy. *Math. Z.*, **15**, 201–210.

85 Heath-Brown, D.R. (**1992**) The distribution and moments of the error term in the Dirichlet divisor problem. *Acta Arithmetica*, LX(4):389–415.

86 Bleher, P.M., Cheng, Z., Dyson, F.J. and Lebowitz, J.L. (**1993**) Distribution of the error term for the number of lattice points inside a shifted circle. *Communications in Mathematical Physics*, **154**(3), 433–469.

87 Levitan, B.M. and Zhikov, V.V. (**1968**) *Almost Periodic Functions and Differential Equations*. Cambridge University Press.

88 Aurich, R., Bäcker, A. and Steiner, F. (**1997**) Mode fluctuations as fingerprints of chaotic and non-chaotic systems. *Int. J. Modern Phys. B*, **11**, 805–849.

89 Safarov, Y. and Vassiliev, D. (**1996**) *The Asymptotic Distribution of Eigenvalues of Partial Differential Operators*, volume 155 of *Translations of Mathematical Monographs*. AMS.

90 Brownell, F.H. (**1957**) Extended asymptotic eigenvalue distributions for bounded domains in *n*-space. *J. Math. Mech.*, **6**, 119–166.

91 Baltes, H.P. and Hilf, E.R. (**1976**) *Spectra of Finite Systems*, Bibliographisches Institut Mannheim, Zürich.

92 Steiner, F. (**1994**) Quantum chaos. In R. Ansorge, editor, *Schlaglichter der Forschung. Zum 75. Jahrestag der Universität Hamburg 1994*, Dietrich Reimer Verlag, Berlin, 543–564.

93 Aurich, R., Bolte, J. and Steiner, F. (**1994**) Universal signatures of quantum chaos. *Phys. Rev. Lett.*, **73**, 1356–1359.

94 Berry, M.V. (**1985**) Semiclassical theory of spectral rigidity. *Proc. R. Soc. London, A* **400**, 229–251.

95 Bleher, P.M. and Lebowitz, J.L. (**1994**) Energy-level statistics of model quantum systems: Universality and scaling in a lattice-point problem. *J. Statistical Physics*, **74**, 167–217.

96 Bleher, P.M. (**1999**) Trace formula for quantum integrable systems, lattice-point problem, and small divisors, in (eds D.A. Hejhal, J. Friedman, M.C. Gutzwiller and A.M. Odlyzko), *Emerging Applications of Number theory*, Springer, New York, 1–38.

97 Aurich, R. and Steiner, F. (**1993**) Statistical properties of highly excited quantum eigenstates of a strongly chaotic system. *Physica D*, **64**, 185–214.

98 Selberg, A. (**1946**) Contributions to the theory of the Riemann zeta-function. *Arch. Math. Naturvid.*, **48**(5), 89–155.

99 Ghosh, A. (**1983**) On the Riemann zeta-function – mean value theorems and the distribution of S(T). *J. Number Theo.*, **17**(1), 93–102.

100 Montgomery, H.L. (**1989**) Selberg's work on the zeta-function, in (eds K.E. Aubert, E. Bombieri and D. Goldfeld), *Number Theory, Trace Formulas and Discrete Groups*, Academic, New York, 157–168.

101 Bäcker, A. and Steiner, F. (**2001**) Quantum chaos and quantum ergodicity, in (ed B. Fiedler), *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems*, Springer, Berlin, 717–751.

102 Carleman, T. (**1934**) Propriétés asymptotiques des fonctions fondamentales des membranes vibrantes. *Comptes Rendus des Mathematiciens Scandinavesa Stockholm*, 14–18.

103 Carleman, T. (**1936**) Über die asymptotische Verteilung der Eigenwerte partieller Differentialgleichungen. *Ber. der Sächs. Akad. d. Wiss. Leipzig*, **88**, 119–132.

104 Hadamard, J. (**1898**) Sur le billard non-Euclidean. *Soc. Sci. Bordeaux, Proc. Verbaux*, p. 147.

105 Hadamard, J. (**1898**) Les surfaces à courbures opposées et leurs lignes géodésiques. *J. Math. Pure Appl.*, **4**, 27–73.

106 Gutzwiller, M.C. (**1980**) Classical quantization of a Hamiltonian with ergodic behavior. *Phys. Rev. Lett.*, **45**, 150–153.

107 Aurich, R. and Steiner, F. (**2001**) Orbit sum rules for the quantum wave functions of the strongly chaotic Hadamard billiard in arbitrary dimensions. *Foundations of Physics*, **31**, 423–444.

108 Aurich, R. and Steiner, F. (**2001**) The cosmic microwave background for a nearly flat compact hyperbolic universe. *Monthly Notices of the Royal Astronomical Society*, **323**, 1016–1024.

109 Grosche, C. and Steiner, F. (**1988**) The path integral on the pseudosphere. *Annals of Physics (NY)*, **182**, 120–156.

110 Aurich, R. and Steiner, F. (**1991**) Exact theory for the quantum eigenstates of a strongly chaotic system. *Physica D*, **48**, 445–470.

111 Buser, P. (**1977**) Riemannsche Flächen mit Eigenwerten in (0,1/4). *Comment. Math. Helvet.*, **52**(1), 25–34.

112 Schoen, R., Wolpert, S. and Yau, S.T. (**1980**) Geometric bounds on the low eigenvalues of a compact surface, in *Geometry of the Laplace operator*, volume 36 of *Proc. Sympos. Pure Math.*, American Mathematical Society, Providence, 279–285.

113 Schmutz, P. (**1991**) Small eigenvalues on Riemann surfaces of genus 2. *Invent. Math.*, **106**(1), 121–138.

114 Titchmarsh, E.C. (**1986**) *The theory of the Riemann zeta-function*. Oxford University Press, Second edition revised by D.R. Heath-Brown.

115 Ingham, A.E. (**1971**) *The distribution of prime numbers*, Hafner, New York.

116 Weil, A. (**1952**) Sur les 'formules explicites' de la théorie des nombres premiers. *Comm. Sém. Math. Univ. Lund (Medd. Lunds Univ. Mat. Sem) Tome Supplémentaire*, 252–265.

117 Connes, A. (**1996**) Formule de trace en géométrie non-commutative et hypothèse de Riemann. *C.R. Acad. Sci. Paris*, 323:(**1231**)–(**1236**).

118 http://www.dtc.umn.edu/~odlyzko/polya/index.html.

119 Deitmar, A. (**2000**) Geometric zeta-functions of locally symmetric spaces. *Am. J. Math.*, **122**(5), 887–926.

120 Ruelle, D. (**1976**) Zeta-functions for expanding maps and Anosov flows. *Invent. Math.*, **34**(3), 231–242.

121 Aurich, R. and Steiner, F. (**1992**) Asymptotic distribution of the pseudo-orbits and the generalized Euler constant γ_Δ for a family of strongly chaotic systems. *Phys. Rev. A*, **46**, 771–781.

122 Huber, H. (**1959**) Zur analytischen Theorie hyperbolischer Raumformen und Bewegungsgruppen. *Math. Ann.*, **138**(1), 1–26.

123 Aurich, R. and Steiner, F. (**1992**) Staircase functions, spectral rigidity, and a rule for quantizing chaos. *Phys. Rev. A*, **45**(2), 583–592.

124 Balian, R. and Bloch, C. (**1970**) Distribution of eigenfrequencies for the wave equation in a finite domain. I. Three-dimensional problem with smooth boundary surface. *Ann. Phys.*, **60**(43), 401–447.

125 Balian, R. and Bloch, C. (**1971**) Distribution of Eigenfrequencies for the Wave Equation in a Finite Domain. II. Electromagnetic Field. Riemannian Spaces. *Ann. Phys.*, **64**(43), 271–307.

126 Balian, R. and Bloch, C. (**1974**) Errata. *Ann. Phys.*, **84**, 559–563.

127 Sieber, M., Primack, H., Smilansky, U., Ussishkin, I. and Schanz, H. (**1995**) Semiclassical quantization of billiards with mixed boundary conditions. *J. Phys. A: Math. Gen.*, **28**, 5041–5078.

128 Simon, B. (**1983**) Some quantum operators with discrete spectrum but classically continuous spectrum. *Ann. Phys. (NY)*, **146**(1), 209–220.

129 Simon, B. (**1983**) Nonclassical eigenvalue asymptotics. *J. Functional Analysis*, **53**(4), 84–98.

130 Steiner, F. and Trillenberg, P. (**1990**) Refined asymptotic expansion for the partition function of unbounded quantum billiards. *J. Math. Phys.*, **31**, 1670–1676.

131 Sieber, M. and Steiner, F. (**1990**) Classical and quantum mechanics of a strongly chaotic billiard system. *Physica D*, **44**, 248–266.

132 Sieber, M. and Steiner, F. (**1990**) Quantum chaos in the hyperbola billiard. *Phys. Lett. A*, **148**(8-9), 415–420.

133 Sieber, M. and Steiner, F. (**1991**) Quantization of chaos. *Phys. Rev. Lett.*, **67**(15), 1941–1944.

134 Karamata, J. (**1931**) Neuer Beweis und Verallgemeinerung der Tauberschen Sätze, welche die Laplace'sche und Stieljes Transformation betreffen. *J. Reine Angew. Math.*, **164**, 27–39.

135 Arendt, W., Batty, C., Hieber, M. and Neubrander, F. (**2001**) *Vector-Valued Laplace Transforms and Cauchy Problems*, Birkhäuser.

136 Arendt, W. (**2000**) Resolvent positive operators and inhomogeneous boundary conditions. *Scuola Normale Superiore Pisa*, **29**, 639–670.

137 Birman, M.S. and Solomjak, M.Z. (**1980**) *Quantitative Analysis in Sobolev Imbedding Theorems and Applications to Spectral Theory*. American Mathematical Society.

138 Arendt, W. (**2005/06**) Heat Kernels (Internet Seminar 2005/2006). https://tulka.mathematik.uni-ulm.de/2005/lectures/internetseminar.pdf.

139 Daners, D. (**2008**) *Handbook of Differential Equations: Stationary Partial Differential Equations*, **6**, 1–81, Elesevier.

140 Dodziuk, J. (**1981**) Eigenvalues of the Laplacian and the heat equation. *The American Mathematical Monthly*, **88**(9), 686–695.

141 Michels, E. (**2001**) Zur Spektraltheorie elliptischer Differentialoperatoren: Beschreibung und Informationsgehalt diskreter Eigenwerte. Master's thesis, Eberhard-Karls-Universität Tübingen.

142 Arendt, W. and Daners, D. (**2008**) The Dirichlet problem by variational methods. *Bulletin of the London Mathematical Society*, **40**, 51–56.

143 Arendt, W. and Daners, D. (**2008**) Varying Domains: Stability of the Dirichlet and Poisson Problem. *Discrete Continuous Dynamical Systems – Series A*, **21**, 21–39.

144 Brown, R.M. (**1993**) The trace of the heat kernel in Lipschitz domains. *Transactions of the American Mathematical Society*, **339**(2), 889–900.

145 Urakawa, H. (**1982**) Bounded domains which are isospectral but not congruent. *Annales Scientifiques de l'École Normale Supérieure Sér. 4*, **15**(3), 441–456.

146 Gordon, C., Webb, D. and Wolpert, S. (**1992**) Isospectral plane domains and surfaces via Riemannian orbifolds. *Inventiones Mathematicae*, **110**(1), 1–22.

147 Gordon, C. and Webb, D. (**1994**) Isospectral convex domains in euclidean space. *Math. Res. Lett*, **1**, 539–45.

148 Zelditch, S. (**2007**) Inverse Spectral Problem for Analytic Domains II: $\mathbb{Z}_2$-Symmetric Domains, Annals of Mathematics, to appear.

149 Chang, P.K. and Deturck, D. (**1989**) On hearing the shape of a triangle. *Proceedings of the American Mathematical Society*, **105**(4), 1033–1038.

150 Henrot, A. (**2003**) Minimization problems for eigenvalues of the Laplacian. *Journal of Evolution Equations*, **3**, 443–461.

151 Daners, D. and Kennedy, J. (**2007**) Uniqueness in the Faber–Krahn inequality for Robin problems. *SIAM J. Math. Anal.*, **39**(4), 1191–1207.

152 Ziemer, W.P. and Malỳ, J. (**1997**) Fine Regularity of Solutions of Elliptic Partial Differential Equations. *Am. Math. Soc.*

153 Arendt, W. (**2002**) Does diffusion determine the body? *J. Reine. Angew. Math*, **550**, 97–123.

2
Solutions of Systems of Linear Ordinary Differential Equations

Werner Balser, Claudia Röscheisen, Frank Steiner, Eric Sträng[1]

2.1
Introduction

Systems of ordinary differential equations (ODE) are of great interest, both in mathematics and physics. If their coefficient matrix depends on the variable t, then their solutions can only occasionally be computed in explicit form in terms of *known functions* such as the exponential function (of a matrix), or other so-called *higher transcendental functions* including *Bessel's* or the *hypergeometric function*. In this article we collect and describe methods that are popular among mathematicians and/or physicists for computing the solutions of such a system. In some exceptional cases these methods may lead to explicit solution formulas, while in general they end with representations of solutions as infinite series that may or may not converge, but still give useful insight into the behavior of the solutions. As illustrating examples we shall frequently refer to two simple but nonetheless nontrivial systems of the following very special form:

1. For $d \times d$ constant matrices Λ and A, with Λ being diagonalizable, we shall follow *K. Okubo* [1] and refer to[2]

   $$(\Lambda - t)x'(t) = Ax(t) \tag{2.1}$$

   as *the hypergeometric system* in dimension $d \geq 2$. The name for this system refers to the fact that, for $d = 2$, a fundamental solution can be computed in terms of the hypergeometric function (and other elementary ones); for this, refer to a book of *W. Balser* [2]. For $d \geq 3$, however, it is believed, although perhaps not rigorously proven, say by differential Galois theory, that its solutions only occasionally are *known functions*. This system may not have any direct application in physics or other areas but has, partially in more general form,

[1] Corresponding author.

[2] We shall adopt the convention of writing $\Lambda - t$ instead of $\Lambda - t\,I$, with I being the identity matrix of appropriate dimension.

been frequently investigated by mathematicians. The reason for its popularity with the latter group is that it is complicated enough to make its solutions *new higher transcendental functions*, while on the other hand it is simple in the following sense: The eigenvalues of Λ, defined as the points where $\Lambda - t$ fails to be invertible, as well as the point $t = \infty$, are *regular singularities* of (2.1), hence it is what is called *a Fuchsian system*.

2. For Λ and A as above, we shall call

$$x'(t) = (\Lambda + t^{-1}A)x(t) \tag{2.2}$$

the confluent hypergeometric system in dimension $d \geq 2$. The confluent hypergeometric system is related to (2.1) by means of Laplace transformation, but also by a confluence of all but one singularity of the hypergeometric system, as was shown by *R. Schäfke* [3]. Having an irregular singularity at $t = \infty$, and a regular one at the origin, (2.2) may appear more complicated than the previous one, but owing to their close relation, it is fair to say that they are of the same degree of transcendency in the sense that if we can solve either one of them, then we can also solve the other one. For $d = 2$, a fundamental solution of the system (2.2) can be computed in terms of the *confluent hypergeometric function*.

Mathematicians who have investigated either one of the two systems, analyzing the behavior of their solutions and/or evaluating their Stokes constants, include G.D. Birkhoff [4], H.W. Knobloch [5], K. Okubo [6, 7], M. Kohno [8–10], R. Schäfke [11, 12], Balser, Jurkat, and Lutz [13, 14], Kohno and Yokoyama [15], T. Yokoyama [16,17], and M. Hukuhara [18].

Aside from these two examples, we shall consider a general linear system of ODE, denoted as

$$x'(t) = H(t)x(t) \tag{2.3}$$

with a matrix $H(t)$ whose entries are functions defined in some domain D that is either a real (open) interval or an open and connected subset of the complex numbers. If necessary, we shall require stronger assumptions on $H(t)$, such as analyticity, but for the time being we shall make do with continuity. Note that some of the results to be presented here carry over to, or even have been developed for, the case when $H(t)$ is not a matrix but a more general, perhaps even unbounded, operator in a Banach space. While we shall not attempt to treat such a general and considerably more difficult situation here, we mention as a simple example the situation when $\mathbb{X}$ is a Banach or Hilbert space of functions $f(x)$, with functions that are arbitrarily often differentiable being dense in $\mathbb{X}$, and instead of a matrix $H(t)$ we consider the operator ∂_x^2, assigning to f its second derivative. In this case, instead of a system (2.3) of ODE we deal with the one-dimensional heat or diffusion equation $\partial_t u(t,x) = \alpha\partial_x^2 u(t,x)$, where $\alpha > 0$ is the diffusion constant. Given an initial condition $u(0,x) = \phi(x) \in X$ which is arbitrarily often differentiable, one can formally obtain a solution as

$$u(t,x) = \sum_{k=0}^{\infty} \frac{(\alpha t)^k}{k!} \partial_x^{2k}\phi(x) = \mathrm{e}^{\alpha t\partial_x^2}\phi(x) \ . \tag{2.4}$$

Here, the operator under consideration is independent of time, which is an easy situation when dealing with linear systems of ODE, since for any constant matrix H the series $\mathrm{e}^{tH} = \sum_k (t^k/k!) H^k$ converges. However, owing to the fact that $\partial_x^{2k}\phi(x)$ in general is of magnitude $(2k)!$, the series (2.4) may diverge for every $t \neq 0$. Very recently, it has been shown by Lutz, Miyake, and Schäfke [19] that for many, but not all, functions $\phi(x)$ the series is summable in a sense to be discussed later; for this and other results in this direction, the reader may also refer to a paper by W. Balser [20], as well as to the literature listed there. If the diffusion constant α is allowed to be purely imaginary, say $\alpha = \mathrm{i}\tilde{\alpha}$, then instead of the diffusion equation one obtains the one-dimensional Schrödinger equation for a free particle. The results from the papers quoted above easily carry over to this situation as well. For a somewhat more detailed description of this situation, see Section 2.5.6.

According to the general theory, we may regard (2.3) as solved, if we can find d linearly independent solution vectors, since it is well known that the set of all solutions is a linear space of dimension d. Such solutions can be arranged into a $d \times d$ *fundamental matrix* $X(t)$, and if $X(t)$ is any matrix whose columns solve (2.3), then it is fundamental if, and only if, its determinant is non-zero *at least at one point* t_0, implying $\det X(t) \neq 0$ at all $t \in D$.

The methods that we are going to describe in the following three sections are quite different at first glance, but are all based upon the following simple observation:

Suppose that the integral

$$Q(t) = \int_{t_0}^{t} H(\tau)\,\mathrm{d}\tau \tag{2.5}$$

(which exists for any $t_0 \in D$, owing to continuity of $H(\tau)$) gives rise to a matrix function that commutes with $H(t)$. Then one can verify that the matrix

$$X(t) = \mathrm{e}^{Q(t)} := \sum_{n=0}^{\infty} \frac{1}{n!} Q(t)^n$$

with the series being absolutely convergent for all $t \in D$, gives a fundamental solution of (2.3), which is normalized by the fact that $X(t_0) = I$.

This assumption is certainly satisfied whenever $H(t)$ is a constant matrix H, in which case the fundamental solution is $X(t) = \mathrm{e}^{(t-t_0)H}$. So the difficulty in computing a fundamental solution for (2.3) is caused by the fact that, in general, the commutator $[Q(t), H(t)] := Q(t)\,H(t) - H(t)\,Q(t)$ is not going to vanish. For example, in the case of (2.2) we have

$$[Q(t), H(t)] = \big(1 - t_0/t - \log|t/t_0|\big)[\Lambda, A]$$

which vanishes if, and only if, Λ and A commute, and this is a relatively rare situation. If Λ is a diagonal matrix with all distinct diagonal entries, then it commutes with A if, and only if, A is also diagonal. However, if they do commute, then the matrix $X(t) = |t/t_0|^A \, e^{(t-t_0)\Lambda}$ is a fundamental solution. In the case of complex t, an even simpler one is given by

$$X(t) = t^A \, e^{t\Lambda} \tag{2.6}$$

for any choice of the branch of $t^A = e^{(\log t)\, A}$.

We also wish to mention that the commutator $[Q(t), H(t)]$ certainly vanishes if for arbitrary values $t_1, t_2 \in D$ we have $[H(t_1), H(t_2)] = 0$, and if this is so, we say that $H(t)$ *satisfies the commutator condition.*

Roughly speaking, the methods to be discussed treat a general system (2.3) as a perturbation of a second one for which the commutator condition is satisfied or its exact solution is known. In the first two approaches, the transition between the two systems is by introducing a perturbation parameter λ (which in physical applications often plays the role of a coupling strength) and analyzing the dependence of a fundamental solution upon λ, while the third method is best understood as finding linear transformations linking the solution spaces of the two systems. In all approaches, power series either in λ or t are used. While one at first may proceed in a formal manner, one eventually is forced to ensure convergence of these series. We shall indeed see in the last section that, in some situations, power series occur which do not converge, but it will be indicated briefly that even then one can use a technique of *summation* to still make good use of these divergent series.

2.2 The Exponential Ansatz of Magnus

Since a fundamental solution $X(t)$ of (2.3) has a non-zero determinant, we may define $Q(t) = \log X(t)$, or equivalently write $X(t) = e^{Q(t)}$, with whatever determination of the multi-valued logarithm of a matrix. This, however, leaves the question of whether one may compute $Q(t)$ without presuming $X(t)$ to be known. That this can be done, even in situations more general than (2.3), has been shown in an article by Magnus [21]. However, observe that the suggestive idea of saying that

$$\frac{d}{dt} \log X(t) = X'(t)X(t)^{-1} = H(t)$$

implying $\log X(t) = \int_{t_0}^{t} H(\tau)\, d\tau$, may not hold except when $H(t)$ satisfies the commutator condition, which brings us back to what was discussed above! Hence we need a more sophisticated approach, and to facilitate computation, it is best to slightly generalize (2.3), introducing a (complex) parameter λ and write

$$x' = \lambda H(t) x \, . \tag{2.7}$$

A fundamental solution $X(t;\lambda)$ then depends upon t as well as λ, and we wish to represent it as

$$X(t;\lambda) = e^{\lambda Q(t;\lambda)} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} Q(t;\lambda)^k \, , \qquad Q(t;\lambda) = \sum_{j=0}^{\infty} \lambda^j Q_j(t) \tag{2.8}$$

with coefficient matrices $Q_j(t)$ to be determined, and convergence of the second series to be investigated later. While the computation to follow can be facilitated by using some well-known identities, such as those for the computation of the derivative of an exponential matrix, we shall follow a more direct approach, leading to an identity from which one can recursively compute the matrices $Q_j(t)$. For every natural number $k \geq 2$, we set

$$Q(t;\lambda)^k = \sum_{j=0}^{\infty} \lambda^j Q_{jk}(t) , \quad Q_{jk}(t) = \sum_{\nu=0}^{j} Q_{j-\nu}(t) Q_{\nu,k-1}(t) .$$

Setting $Q_{j1}(t) = Q_j(t)$ and interchanging the order of summation, we conclude

$$X(t;\lambda) = I + \sum_{\mu=1}^{\infty} \lambda^\mu \sum_{j=0}^{\mu-1} \frac{1}{(\mu-j)!} Q_{j,\mu-j}(t)$$

and in order that this expression is a solution of (2.7), with $X(t_0;\lambda) = I$, we need to have $Q_0(t) = \int_{t_0}^{t} H(\tau)\,\mathrm{d}\tau$, and for $\mu \geq 1$

$$\sum_{j=0}^{\mu} \frac{1}{(\mu+1-j)!} Q_{j,\mu+1-j}(t) = \int_{t_0}^{t} H(\tau) \sum_{j=0}^{\mu-1} \frac{1}{(\mu-j)!} Q_{j,\mu-j}(\tau)\,\mathrm{d}\tau . \tag{2.9}$$

Suppose that for some $\mu \geq 1$ we would already know $Q_0(t), \ldots, Q_{\mu-1}(t)$ and this is certainly correct for $\mu = 1$. Then we also know $Q_{jk}(t)$ for all $j = 0, \ldots, \mu - 1$ and all $k \geq 1$. Hence we may use (2.9) to explicitly find the next matrix $Q_\mu(t) = Q_{\mu 1}(t)$. We leave it to the reader to verify that

$$\begin{aligned} Q_1(t) &= \frac{1}{2} \int_{t_0}^{t} \int_{t_0}^{t_1} [H(t_1), H(t_2)]\,\mathrm{d}t_2\,\mathrm{d}t_1 \\ Q_2(t) &= \frac{1}{6} \int_{t_0}^{t} \int_{t_0}^{t_1} \int_{t_0}^{t_2} \Big(\big[[H(t_1), H(t_2)], H(t_3)\big] \\ &\quad + \big[[H(t_3), H(t_2)], H(t_1)\big] \Big) \mathrm{d}t_3\,\mathrm{d}t_2\,\mathrm{d}t_1 . \end{aligned}$$

Similarly, the other coefficients can be computed in terms of higher order commutators. For details, and a different proof of these identities, refer to articles [22] and [23].

Note that all $Q_k(t)$ with $k \geq 1$ vanish whenever $H(t)$ satisfies the commutator condition, and there are other situations possible when Magnus' series for $Q(t;\lambda)$ may terminate. In general, however, we have to deal with investigating the convergence of the power series $Q(t;\lambda)$, in particular for the value of $\lambda = 1$. While we shall postpone the discussion of the general case until later, we conclude this section with the following easy but instructive example, showing that we cannot always expect convergence at $\lambda = 1$:

Suppose that

$$H(t) = \begin{bmatrix} a & 0 \\ t & 0 \end{bmatrix} , \qquad a \neq 0 .$$

In this case, the fundamental solution $X(t;\lambda)$, with $X(0;\lambda) = I$, of (2.7) can be verified to be

$$X(t;\lambda) = \begin{bmatrix} e^{\lambda at} & 0 \\ \dfrac{1 + (\lambda at - 1)\, e^{\lambda at}}{\lambda a^2} & 1 \end{bmatrix} .$$

This matrix has a removable singularity at $\lambda = 0$. Using the theory of logarithms of a matrix, one finds that $X(t;\lambda) = e^{Q(t;\lambda)}$ with

$$Q(t;\lambda) = \log X(t;\lambda) = \begin{bmatrix} \lambda at & 0 \\ \dfrac{t\left(1 + (\lambda at - 1)\, e^{\lambda at}\right)}{a\left(e^{\lambda at} - 1\right)} & 0 \end{bmatrix} .$$

Again, the singularity of $Q(t;\lambda)$ at $\lambda = 0$ is removable, and hence an expansion as in (2.8) holds for sufficiently small values of $|\lambda|$. For $t \neq 0$, however, $Q(t;\lambda)$ has a first-order pole at $\lambda = 2\pi i/(at)$, so that the radius of convergence is smaller than 1 whenever $|t| > 2\pi/|a|$. We shall analyze this effect in more detail in the following section.

2.3
The Feynman–Dyson Series, and More General Perturbation Techniques

Here we briefly mention that the fundamental solution of (2.7) can also be represented by a convergent power series

$$X(t;\lambda) = \sum_{k=0}^{\infty} \lambda^k X_k(t) \tag{2.10}$$

with $X_0(t) = I$ and

$$X_k(t) = \int_{t_0}^{t} H(\tau) X_{k-1}(\tau)\, d\tau\,, \quad k \geq 1\,.$$

By repeated insertion of this recursion relation into itself, one can also write $X_k(t)$ as an n-fold integral, and after some manipulation one can obtain a form that is referred to as the *Feynman–Dyson series* [24] which contains so-called time-ordered products of the matrix $H(t)$. This shall not be discussed here, but we should like to say that the series in its original form is intimately related to the Liouville–Neumann method, see for example [25], which is also used in the proof of Picard–Lindelöf's Theorem on existence and uniqueness of solutions to initial value problems. From estimates given there one can show that in our situation the series converges for every λ, hence $X(t;\lambda)$ is an entire function of λ. Knowing this, one can conclude that the matrix $Q(t;\lambda) = \log X(t;\lambda)$, studied in the previous section, is holomorphic at least in a sufficiently small disc about the origin, so that Magnus' series in (2.8) has indeed a positive radius of convergence. From the theory of logarithms of a matrix one knows that $Q(t;\lambda) = \log X(t;\lambda)$, regarded as a function of λ, may become singular once two eigenvalues of $X(t;\lambda)$ differ by a nonzero multiple of $2\pi \mathrm{i}$, and this may or may not happen for values of λ in the unit disc, as is seen in the example given at the end of the previous section. Therefore, the radius of convergence of Magnus' series may, for any fixed $t \neq t_0$ be smaller than 1, in which case the series fails to converge at $\lambda = 1$. As a way out of this dilemma, one may use explicit summation methods providing continuation of holomorphic functions to compute $Q(t;\lambda)$ outside of the circle of convergence of Magnus' series, but we will not discuss this here in detail.

A similar approach to that above works when investigating a system of the form

$$x' = \big(H_0(t) + \lambda H(t)\big)x \tag{2.11}$$

where $H_0(t)$ satisfies the commutator condition, so that for $\lambda = 0$ (the "unperturbed" system) a fundamental solution $X_0(t)$ (which in the case of $H_0(t) \equiv 0$ may be taken as the identity matrix) of (2.11) is known. The matrix $\lambda H(t)$ is considered to be a "small perturbation" if $|\lambda|$ is considered to be small enough. For example, in the case of confluent hypergeometric system (2.2), we may choose $H_0(t) = \Lambda + t^{-1}D$, with D being a diagonal matrix consisting of the diagonal elements of A. In this case D and Λ commute, so that $X_0(t) = t^D \mathrm{e}^{t\Lambda}$. The series (2.10) then is a solution of (2.11) if, and only if,

$$X'_k(t) = H_0(t)X_k(t) + H(t)X_{k-1}(t)\,, \qquad \forall k \geq 1\,.$$

With the standard technique of variation of constants one obtains the recursion

$$X_k(t) = X_0(t)\left[C_k + \int_{t_0}^{t} X_0^{-1}(\tau)H(\tau)X_{k-1}(\tau)\,\mathrm{d}\tau\right], \qquad \forall k \geq 1$$

with constant matrices that can be chosen arbitrarily. To have $X(t_0;\lambda) = I$, one should pick $C_k = 0$. However, if one wishes to obtain a fundamental solution with a prescribed behavior as $t \to \infty$, say, then other choices for C_k are more appropriate. In the *Diplomarbeit* of Röscheisen [26], this technique has been used for the system (2.2) to obtain fundamental solutions that, in sectors in the complex plane, have a certain asymptotic behavior.

The approach discussed so far is referred to as a regular perturbation of linear systems, since (2.11), for $\lambda = 0$, is still a linear system of ODE. Other cases arise when the parameter λ also occurs in front of the derivative x', in which case one refers to a singular perturbation. Such cases have been analyzed, see for example [27], and it has been shown there that one meets power series that are divergent for every $\lambda \neq 0$, but can be summed using the techniques to be discussed later.

Singular perturbation theory has important applications in quantum mechanics. The evolution of a quantum mechanical system is governed by a PDE, the Schrödinger equation,

$$\mathrm{i}\hbar\partial_t\psi(x,\,t) = -\frac{\hbar^2}{2}\Delta_x\psi(x,t) + V(x)\psi(x,t) \tag{2.12}$$

where Δ_x is an appropriately defined Laplacian on a Hilbert space and $\hbar$ is Planck's constant. The study of the *semiclassical limit* $\hbar \to 0$ leads to a singular perturbation problem identifying $\lambda = \mathrm{i}\hbar$. Of particular interest are the simultaneous limits $\hbar \to 0$, $t \to \infty$ of the unitary evolution generated by (2.12) which do not, in general, commute. This is the subject of Quantum Chaos, an introduction to which can be found in the book of Grosche and Steiner (fifth chapter) [28] and the article by Bäcker and Steiner [29]. Quantum chaos deals with, amongst others, the properties of the spectrum of operators or the ergodic properties of the evolution generated by (2.12) using such techniques as microlocal analysis and the theory of pseudo-differential operators (see [30] and references therein).

2.4
Power Series Methods

The methods discussed in the previous sections have the advantage of being applicable to systems where the coefficient matrix $H(t)$ is fairly general. What we shall do here is restricted to cases when $H(t)$ is a meromorphic function for $t \in D$, meaning that D is an open and connected subset of the complex numbers $\mathbb{C}$, and $H(t)$ is either holomorphic or has a pole at any point $t_0 \in D \subset \mathbb{C}$. As we shall see, it is natural in this context to distinguish three different cases:

2.4.1
Regular Points

If $H(t)$ is holomorphic at a point $t_0 \in D$, then t_0 is referred to as a *regular point* of (2.3). In this case, we can expand $H(t)$ into its power series about t_0, and hence for some $\varrho > 0$ we have

$$H(t) = \sum_{k=0}^{\infty} (t - t_0)^k H_k \,, \qquad |t - t_0| < \varrho \tag{2.13}$$

with coefficient matrices H_k that we assume known. *Assuming* that the fundamental solution $X(t)$ can also be represented by a power series, we write analogously

$$X(t) = \sum_{k=0}^{\infty} (t - t_0)^k X_k$$

and inserting into (2.3) and comparing coefficients, we obtain

$$(k+1) X_{k+1} = \sum_{j=0}^{k} H_{k-j} X_j \,, \qquad \forall k \geq 0 \,.$$

Selecting $X_0 = I$, the remaining coefficients X_k are determined by this identity, and a direct estimate shows that the power series so obtained converges for $|t - t_0| < \varrho$, and its sum indeed is the fundamental solution of (2.3) normalized by $X(t_0) = I$. This argument shows that theoretically we can compute a fundamental solution of (2.3) by a power series ansatz, provided that the coefficient matrix $H(t)$ is holomorphic in the disk $|t - t_0| < \varrho$, and moreover, we obtain holomorphy of $X(t)$ in the same disc! We can even do better than this. If we choose any curve from t_0 to any other point $t \in D$, we can cover the curve with discs that remain in D, and by successive re-expansion of $X(t)$ compute its continuation to the point t. However, note that examples show that continuation along a closed curve may not end with the same fundamental solution with which we started!

2.4.2
Singularities of the First Kind

An important issue in the theory of ODE is to analyze how solutions behave when the variable t tends to a singularity t_0 of the coefficient matrix $H(t)$. Even if we succeed in calculating a fundamental solution in closed form, or by means of a convergent power series about a regular point, this may still be a difficult problem. An explicit formula for $X(t)$ may be so complicated that we cannot find out whether or not $X(t)$ grows, or stays bounded, or even goes to 0 as $t \to t_0$; the power series, even when t_0 is a point on the boundary of its circle of convergence, will not immediately say much about the behavior of $X(t)$ at t_0 anyway. So this is why other ways of representing $X(t)$ are still to be desired. This can be done relatively easily at a singularity of the first kind, meaning any point t_0 where $H(t)$ has, at most, a first-order pole: Suppose that

$$H(t) = (t - t_0)^{-1-r} \sum_{k=0}^{\infty} (t - t_0)^k H_k \,, \qquad |t - t_0| < \varrho \tag{2.14}$$

then one refers to r as the *Poincaré rank* of (2.3) at t_0, and a singularity of the first kind is characterized by $r = 0$. In addition, we assume for simplicity that the matrix H_0 satisfies the following eigenvalue condition:

(E) If λ and μ are two distinct eigenvalues of H_0, then $\lambda - \mu$ is not an integer.

In this situation, a fundamental solution $X(t)$ exists that has a representation of the form

$$X(t) = \left(\sum_{k=0}^{\infty} (t - t_0)^k X_k \right) (t - t_0)^{H_0} \,. \tag{2.15}$$

Choosing $X_0 = I$, the remaining coefficients are uniquely determined by the identity

$$X_k(H_0 + k) - H_0 X_k = \sum_{j=0}^{k-1} H_{k-j} X_j \,, \qquad \forall\, k \geq 1 \tag{2.16}$$

since the eigenvalue assumption made above ensures that the left-hand side, which is simply a system of linear equations in the entries of X_k, has a unique solution.

Again, estimating coefficients implies that the power series in (2.15) converges for $|t - t_0| < \varrho$, and this representation immediately explains how $X(t)$ behaves as $t \to t_0$, since we see that $X(t)(t - t_0)^{-H_0}$ is even holomorphic at t_0.

Another way of looking at the above result is as follows. The convergent power series

$$T(t) = \sum_{k=0}^{\infty} (t - t_0)^k X_k \ ,$$

when used as a transformation $x = T(t)y$, changes (2.3) to the system $y' = (t - t_0)^{-1} H_0 y$, whose fundamental solution is $Y(t) = (t - t_0)^{H_0}$. In general, if $T(t)$ is any invertible matrix, then the linear transformation $x = T(t)\,y$ takes (2.3) to the new system

$$y' = \tilde{H}(t) y \ , \qquad \tilde{H}(t) = T^{-1}(t)\big(H(t)T(t) - T'(t)\big)$$

and one may hope that the system so obtained can be solved more easily than the original one, perhaps since the commutator relation discussed in the introduction is satisfied. The same approach works with a singularity of the first kind when the eigenvalue assumption (E) is violated, leading to an analogous result. For more details on this, refer to [31] or [2]. As we shall see in the following subsection, this idea can also be used for systems with singularity of higher Poincaré rank.

Applying this result to the hypergeometric system (2.1), which for diagonal Λ has singularities of the first kind at all diagonal elements of Λ, plus an additional one at $t = \infty$, we see that in principle we may compute fundamental solutions at each singularity, and then by successive re-expansion even find out how these matrices are connected with one another. These *connection formulas* have important applications and have therefore been much studied in the literature.

A related method is commonly used to solve the Schrödinger equation

$$\mathrm{i}\hbar\partial_t U(t) = H(t)\, U(t)$$

in quantum mechanics and quantum field theory, where the Hamiltonian $H(t) = H_0(t) + \lambda H_1(t)$ can be split into a free part $H_0(t)$ and an interacting part $H_1(t)$. (Here $H(t)$, $H_0(t)$, $H_1(t)$ denote hermitian matrices or self-adjoint operators.) With $U(t) = U_0(t)\, U_1(t)$, $\mathrm{i}\hbar\partial_t U_0(t) = H_0(t)\, U_0(t)$, one obtains the Schrödinger equation in the Dirac interaction picture [32, 33]

$$\mathrm{i}\hbar\partial_t U_1(t) = \lambda \tilde{H}_1(t)\, U_1(t) \ , \qquad \tilde{H}_1(t) := U_0^{-1}(t) H_1(t)\, U_0(t) \ .$$

2.4.3 Singularities of Second Kind

The confluent hypergeometric system (2.2) has a singularity of the first kind at the origin, hence we may compute a fundamental solution of the form (2.15), with $t_0 = 0$. Owing to the absence of other finite singularities, the power series in this representation converges for every $t \in \mathbb{C}$. However, it is not obvious how the solutions so obtained behave as $t \to \infty$. By means of a change of variable $t = 1/\tau$, system (2.2) becomes equal to $y' = -(\tau^{-2}\Lambda + \tau^{-1}A)y$, with y' denoting the derivative of y with respect to τ. This new system has a singularity of Poincaré rank $r = 1$ at the origin, and this is why we say that (2.2) has the same rank at infinity. Hence, the methods of the previous subsection do not apply. Nonetheless, it is natural to look for a transformation $x = T(t)y$ which changes (2.2) into a new system that may be solved directly, and since we want the fundamental solution $Y(t)$ of the transformed system to have the same behavior at ∞ as that of the original equation, we wish to represent $T(t)$ as a power series in t^{-1}, denoted by

$$T(t) = \sum_{k=0}^{\infty} t^{-k} T_k \,. \tag{2.17}$$

Since $T^{-1}(t)$ should also be such a power series, we require in addition that the matrix T_0 be invertible. For simplicity, we require that Λ is not only diagonalizable, but is indeed a diagonal matrix, whose diagonal entries are all distinct. Then we may even restrict to $T_0 = I$, and it can be shown that a transformation as above exists, for which the transformed system has the form

$$y' = (\Lambda + t^{-1}D)y$$

with a diagonal matrix D that is equal to the diagonal entries of the original matrix A in (2.2). So in a sense the matrix $T(t)$ is a *diagonalizing transformation* for the confluent hypergeometric system.

In [26], confluent hypergeometric systems have been considered, without the restriction that all diagonal entries of Λ are distinct (the so-called case of multiple eigenvalues), but that identical entries appear consecutively. With a formal diagonalizing transformation, recursion formulas for a formal fundamental solution of triangular systems has been computed and it has been shown that this formal transformation in triangular form is unique under the assumption that $T_0 = I$.

Even for a general system of arbitrary Poincaré rank, say, at the point ∞ it is well known that a transformation (2.17) exists for which the transformed system satisfies the commutator condition needed to compute its fundamental solution. However, in all but some exceptional situations, the series in (2.17) fails to converge for every t. But there is a relatively recent theory of multi-summability that allows one still to make use of this series and compute a fundamental solution with the help of finitely many integral transformations. We shall briefly describe these results in the following section, and refer to the books by Balser [2, 34] for details. In the first one, one can even find a proof for the fact that all power series that arise as formal solutions even for nonlinear equations are multi-summable. Unfortunately, this is no longer the case with series that solve even very simple partial differential equations; for example, the series (2.4) solving the heat equation fails to be multi-summable for certain initial conditions.

2.5 Multi-Summability of Formal Power Series

Here we attempt to briefly describe the theory of Gevrey asymptotics and (multi-) summability of formal power series. This theory was proven to be very appropriate when dealing with formal solutions of ordinary differential equations, and is now investigated for applications to partial differential equations, or other functional equations. In this case one is concerned with power series whose coefficients, instead of being complex numbers, are functions of one or several variables, and therefore this summation method has been shown in [2] to generalize to power series whose coefficients are in an abstract Banach space $\mathbb{E}$ over the field $\mathbb{C}$ of complex numbers. However, for simplicity of presentation we shall here restrict ourselves to the special situation where $\mathbb{E}$ is one-dimensional, in other words where it is in fact equal to $\mathbb{C}$.

2.5.1 Asymptotic Power Series Expansions

It is standard to say that a function $f(z)$, which is assumed here to be analytic in a sector $S = \{z : 0 < |z| < r, \alpha < \arg z < \beta\}$ of the complex plane, is *asymptotic to a formal power series* $\hat{f}(z) = \sum_n z^n f_n$ *as* $z \to 0$ *in* S, provided that for every (closed) subsector $S_\delta = \{z : 0 < |z| \le r - \delta, \alpha + \delta \le \arg z \le \beta - \delta\}$, with $\delta > 0$, and every $N \ge 0$, there exist constants $c_{\delta,N} > 0$ such that

$$\left| f(z) - \sum_{n=0}^{N-1} z^n f_n \right| \le c_{\delta,N} |z|^N , \qquad \forall\, z \in S_\delta . \tag{2.18}$$

If this is so, we also say that the power series $\hat{f}(z)$ is *the asymptotic expansion* of the function $f(z)$ in the sector S. This statement can be seen to be equivalent to the fact that for every $n \ge 0$ the nth derivative $f^{(n)}(z)$ of $f(z)$ tends to $n!\, f_n$ as $z \to 0$ in S_δ. Yet another way of looking at this is by saying that there exists a unique value $f(0)$ for which $f(z)$ not only becomes continuous, but in fact arbitrarily often differentiable at the origin (however, only when restricting the corresponding limits for $z \to 0$ to values $z \in S_\delta$), and then the formal series $\hat{f}(z)$ is simply the Taylor series of f at the origin.

The observation that an asymptotic expansion of a function $f(z)$ is equal to its Taylor series may suggest that this series converges to $f(z)$ for sufficiently small $|z|$, but this is not the case in general. On the contrary, the radius of convergence of $\hat{f}(z)$ may be equal to zero, and if not, the series may converge to a function different from $f(z)$. Moreover, while a given function $f(z)$ can have at most one asymptotic expansion, it is so that infinitely many functions exist which are asymptotic to a given power series $\hat{f}(z)$ in a given sector S (of finite opening). The existence of such a function in fact follows from a theorem named after J.F. Ritt [35], while its nonuniqueness is due to the fact that $\exp[az^{-c}]$, for sufficiently small $c > 0$ and suitable complex a, is asymptotic to the zero power series in a given sector S.

2.5.2 Gevrey Asymptotics

The fact that there are always infinitely many functions which are asymptotic to a given formal power series, even if this series converges, is very annoying in certain applications, such as using formal power series solutions for differential equations to construct proper solutions. Therefore, it is natural to ask whether there is one particular function $f(z)$, asymptotic to a given series $\hat{f}(z)$, which in some sense is the most natural choice to make, and which therefore might be considered as a generalized sum for $\hat{f}(z)$. In order to discuss this, we introduce the notion of Gevrey asymptotics of order $s > 0$ by requiring that (2.18) holds for constants $c_{\delta,N}$ satisfying

$$c_{\delta,N} \le C_\delta K_\delta^N \Gamma(1+sN)\,, \qquad \forall\ N \ge 0$$

with suitable constants C_δ, K_δ that may depend upon δ but not upon N. In this situation, if we want a function $f(z)$ to have a given series $\hat{f}(z)$ as its Gevrey asymptotic of order s, then $|f_n| \le CK^n\Gamma(1+sn)$ for all $n \ge 0$ follows, with suitable $C, K > 0$. This fact is expressed verbally by saying that $\hat{f}(z)$ is a formal power series of Gevrey order s. Given such a power series, the following can be shown:

1. If S is a sector of opening not larger than $s\pi$, then there always exists a function $f(z)$ that, in this sector, has $\hat{f}(z)$ as its Gevrey asymptotic of order s, but this function is still not uniquely determined. This result is usually referred to as the *Gevrey–Ritt theorem*.

2. If the opening of S is larger than $s\pi$, then a function as in (1) may not exist, but if it does, then it is uniquely determined. To show uniqueness of $f(z)$, one uses the so-called *Phragmén–Lindelöf theorem*, which is a variant of the maximum principle for analytic functions.

Hence we see that for sectors of large opening we have, at most, one function having a given series as its Gevrey asymptotic, and therefore this function, if it exists, is awarded the title (generalized) *sum* of the divergent series $\hat{f}(z)$. In fact, it turns out that this sum coincides with the one obtained by the process of k-summability, for $k = 1/s$, that we shall discuss later. Note, however, that this sum in general depends upon the sector S, and there may be some sectors where no such sum exists, even if we are willing to make the opening of the sector smaller (but still larger than $s\pi$), keeping its bisecting ray fixed.

2.5.3 Asymptotic Existence Theorems

A nonlinear system of ordinary differential equations may have one, or even several, solution vectors whose coordinates are formal power series in z. If so, Ritt's theorem implies the existence of functions that are asymptotically equal to these series, but it is not clear whether any of them may be used to build a solution of the given system of ODE. However, results exist showing that this indeed can be done under fairly weak assumptions about the form of the system.

Results of this kind, usually called *asymptotic existence theorems*, were proven, first for linear and later for nonlinear systems, as follows. First, one uses the fact that a formal power series solution exists, in order to rewrite the system in the form of an integral equation that, aside from the other data, depends upon finitely many arbitrary parameters. Then one shows that this integral equation can be solved by the usual iteration procedure for contractive operators in a suitable Banach space.[3] Last, the solution obtained by this technique is shown to have the desired asymptotic behavior. In fact, one can show that the asymptotic expansion is of some Gevrey order, but the solution so obtained is not unique, since the sector in which this expansion holds, generally has too small an opening. Many results of this type have been obtained in the first half of the last century, without using the terminology of Gevrey asymptotics, or even that of multi-summability that we shall discuss later on. The fact that for one and the same formal power series solution one has different solutions having the formal one as their asymptotic expansion in different sectors, or equivalently, that the same solution in general has a different asymptotic behavior in different sectors, is usually called *Stokes' phenomenon*, since Stokes [36] first analyzed this behavior in some special examples.

2.5.4 **k-Summability**

Formal power series solutions of linear, and to a lesser degree even nonlinear systems of ordinary differential equations are relatively easy to find – if we are satisfied with finding a recursion equation for the coefficients of the power series. Given a sector of sufficiently small opening, an asymptotic existence theorem shows that at least one solution of the system has the formal one as its asymptotic expansion in this sector. However, one would certainly prefer to have a way of computing such a solution directly in terms of the formal power series that one has computed before. To do this, a summation method is needed that can sum power series with rapidly growing coefficients. Abstractly speaking, such a summation method is an operator $\mathcal{S}$, mapping a set $\mathbb{D}$ of formal power series into a set of holomorphic functions, and in order to be suitable for summation of formal solutions of ordinary differential equations, it should have the following properties:

- The domain $\mathbb{D}$ of $\mathcal{S}$ should be a differential algebra, that is, a vector space over $\mathbb{C}$ that is closed with respect to multiplication and termwise differentiation.
- The operator $\mathcal{S}$ should be a homomorphism, meaning that it should map a sum, respectively, a product of series in $\mathbb{D}$ to the sum, respectively, product

[3] Observe that the integral equation has a unique solution. However, due to the existence of free parameters, one has in general more than one solution of the original system of ODE.

of the corresponding functions, and a derivative of a series should be mapped to the derivative.

- The domain $\mathbb{D}$ should include all convergent power series, and $\mathcal{S}$ should map every one of those to its natural sum.

Every summation method satisfying these requirements has the property that, given any ordinary differential equation satisfying some reasonable assumptions, any formal power series solution that belongs to the domain $\mathbb{D}$ is mapped by $\mathcal{S}$ to a function that is again a solution of the equation. Many of the classical summation methods such as Cesaro's, or Abel's method, fail to have one or more of these properties. However, Ramis [37] presented a family of methods which all are variants of the Borel summation and can be seen to obey all the requirements listed above:

Let $k > 0$ and d be given.[4] Then a formal power series $\hat{f}(z) = \sum f_j z^j$ is said to be k-summable in the direction d, provided that the following two conditions hold.

1. The series $g(t) = \sum f_j t^j / \Gamma(1 + j/k)$ has positive radius of convergence.
2. There exists a $\delta > 0$ so that the function $g(t)$ can be continued into a sector $S(d, \delta) = \{t : |d - \arg t| \le \delta\}$, and for suitably large C, K we have
$$|g(t)| \le C e^{K|t|^k} , \quad t \in S(d, \delta) .$$

If so, then the function f defined by the integral[5]

$$f(z) = z^{-k} \int_0^{\infty(d)} g(t)\, e^{-(t/z)^k}\, dt^k , \quad z \in G_d \tag{2.19}$$

is the k-sum of $\hat{f}(z)$ in the direction d, or for short: the sum of $\hat{f}(z)$. We denote by G_d the set of values z where the integral converges. Observe that the integral in (2.19) is intimately related to Laplace transformation, and if we would replace $g(t)$ by its power series representation and interchange sum and integral, then instead of $f(z)$ we would re-obtain the series $\hat{f}(z)$. This, in fact, proves that if the series $\hat{f}(z)$ has positive radius of convergence, then it is k-summable in the direction d, for every value of $k > 0$ and d, and $f(z)$ in this case is its natural sum. Moreover, one can show that if a series is k-summable in every direction d, then it is necessarily convergent. On the other hand, there are simple examples of series that are divergent but k-summable in all but finitely many directions. Consider the series $\hat{f}(z) = \sum \Gamma(1 + n/k) z^n$; its radius of convergence is zero, owing to the rapid growth of the Gamma function. However, the corresponding function $g(t)$ is just $(1 - t)^{-1}$, and therefore we conclude from the definition that this series is k-summable in all directions d except for the positive real axis, and its sum is given by

$$f(z) = z^{-k} \int_0^{\infty(d)} \frac{e^{-(t/z)^k}}{1 - t}\, dt^k .$$

[4] Here d may be restricted to a half-open interval of length 2π, say $d \in (-\pi, \pi]$ or, more conveniently, values of d differing by integer multiples of 2π should be considered as equivalent.

[5] For convenience we write dt^k instead of $kt^{k-1}\, dt$.

Here one can see that $f(z)$ is a holomorphic (but not single-valued) function in the sector $-\pi/2 < \arg z < 5\pi/2$ and has the formal series $\hat{f}(z) = \sum \Gamma(1 + n/k)z^n$ as its asymptotic expansion of Gevrey order $s = 1/k$, as $z \to 0$ in this sector. Even in general, the set G_d where the integral (2.19) is convergent is a sectorial region of opening larger than π/k, and the function so defined is asymptotic to $\hat{f}(z)$ of Gevrey order $s = 1/k$.

In [26], Röscheisen used the method of 1-summability to get proper fundamental solutions of a triangular confluent hypergeometric system of differential equations. After computing recursion formulas for a formal power series $\hat{F}(z) = \sum_{j=0}^{\infty} z^{-j} F_j$, it is possible to define a function $G(u)$ by

$$G(u) := \sum_{j=1}^{\infty} \frac{u^{j-1}}{(j-1)!} F_j \ ,$$

which is a slight modification of the definition above. So, for certain directions $d_\nu, \nu \in \mathbb{Z}$, depending upon the eigenvalues of the system, and $d \in (d_{\nu+1}, d_\nu)$ the 1-sum of $\hat{F}(z)$ in direction d has been given by

$$F_\nu(z) := \int_0^{\infty(d)} e^{-zu} G(u)\,\mathrm{d}u \ .$$

Observing the intimate relation of (2.19) to a Laplace transformation, one can verify that the method of k-summability has all the properties listed above as desirable for dealing with formal solutions of differential equations. Unfortunately, examples show that some equations have formal solutions that are not k-summable in any direction d, for whatever value of $k > 0$. On the other hand, a result by Ramis [37] that was also shown independently in [38, 39], shows that such solutions can be decomposed into sums of products of (finitely many) series that individually are summable, but for values of k depending upon the term in the decomposition. Since this decomposition cannot be found effectively, this result cannot be used to compute the sum of formal solutions for general systems of ODE. To do this, one needs a more powerful method, called multi-summation, that was originally introduced by Ecalle [40] and will be presented in the next subsection in a form which in [41] was shown to be equivalent to that of Ecalle.

It may be surprising to note that the method of k-summation does not improve when k is varied in either direction. The first condition in the definition may fail to hold when k becomes larger, that is, when $s = 1/k$ gets smaller. On the other hand, if $k > \tilde{k} > 0$, then there exist series that are k-summable in a direction d, but for which the second condition in the definition does not hold when k is replaced by $\tilde{k}$. In addition, one can show that a series which is k-summable in all but finitely many directions d, and which at the same time is of Gevrey order $s < 1/k$, is necessarily convergent.

2.5.5 Multi-Summability

As noted above, the methods of k-summation are still not sufficiently powerful to handle all formal power series solutions of ODE. In order to produce some stronger method, we again look at the two conditions in the definition of k-summability. The first one is important to provide a function $g(t)$ which then, according to the second condition, can be inserted into the integral (2.19) in order to obtain the sum of the formal series $\hat{f}(z)$. So to proceed from k-summability to multi-summability, we say that in order to have a function $g(t)$ at our disposal, it is not important to require convergence of the series $\sum f_j t^j / \Gamma(1+j/k)$, but instead we may allow that this series is just summable in some sense, for example, κ-summable in a direction δ. This idea of iteration of summation methods is already present in the work of Hardy and his student and collaborator Good, but its potential for summation of formal solutions of ODE was not observed until the fundamental contributions from Ecalle.

In detail, the idea described above leads to the following definition. Let $\kappa_j > 0$ and d_j, $1 \le j \le q$, be so that $|d_j - d_{j-1}| \le \pi/\kappa_j$, $2 \le j \le q$. Then $\hat{f}(z) = \sum f_j t^j$ is said to be $(\kappa_1, \dots, \kappa_q)$-summable in the multidirection $(d_1, \dots, d_q)$, provided that the following two conditions hold.

1. The series $\hat{f}_1(t) = \sum f_j t^j / \Gamma(1 + j/\kappa_1)$ is $(\kappa_2, \dots, \kappa_q)$-summable in the multidirection $(d_2, \dots, d_q)$; its sum will be called $f_1(t)$.
2. For some $\delta > 0$, the function $f_1(t)$ can be continued into the sector $S(d_1, \delta) = \{t : |d_1 - \arg t| \le \delta\}$, and for some $C, K > 0$ we have
$$|f_1(t)| \le C e^{K|t|^{\kappa_1}}, \quad t \in S(d_1, \delta) .$$

If so, then

$$f(z) = z^{-\kappa_1} \int_0^{\infty(d_1)} f_1(t) \, e^{-(t/z)^{\kappa_1}} \, dt^{\kappa_1} , \quad z \in G_{d_1}$$

is the $(\kappa_1, \dots, \kappa_p)$-sum of $\hat{f}(z)$ in the multi-direction $(d_1, \dots, d_p)$.

Observe that here in the case of $q = 2$, one should interpret $(\kappa_2, \dots, \kappa_q)$-summability in the sense of Ramis. Moreover, the sum $f(z)$ is in fact given by an iterated integral, in which the order of integration, in general, cannot be interchanged. One usually refers to the vector $(\kappa_1, \dots, \kappa_q)$ as *the type* of multi-summability, and notes that a formal power series is multi-summable, provided that a type $(\kappa_1, \dots, \kappa_q)$ exists for which it is $(\kappa_1, \dots, \kappa_q)$-summable in all but finitely many multi-directions $(d_1, \dots, d_q)$. If we define numbers $k_1 > \dots > k_q > 0$ by

$$1/k_j = 1/\kappa_1 + \dots + 1/\kappa_j , \qquad 1 \le j \le q$$

then one can show that every series $\hat{f}_j(z)$ that is k_j-summable in the direction d_j is also $(\kappa_1, \dots, \kappa_q)$-summable in the multi-direction $(d_1, \dots, d_q)$, and the two sums agree. Moreover, since multi-summability can also be shown to have all the properties listed in Section 2.5.4, the sum $\hat{f}(z) = \hat{f}_1(z) + \ldots + \hat{f}_q(z)$ is also summable in this sense. In fact, it has been shown in [42, 43] that in the case of $\kappa_j > 1/2$ the converse of this statement holds. If $\hat{f}(z)$ is $(\kappa_1, \ldots, \kappa_q)$-summable in the multi-direction $(d_1, \ldots, d_q)$, then $\hat{f}(z) = \hat{f}_1(z) + \ldots + \hat{f}_q(z)$ with $\hat{f}_j(z)$ being k_j-summable in the direction d_j; the same statement, however, fails to hold if one, or several, $\kappa_j \leq 1/2$.

When Ecalle gave his definition of multi-summability, it followed from results mentioned above that all power series arising in formal solutions of *linear* ODE are multi-summable. Meanwhile, there are three independent proofs by Braaksma [44], Ramis and Sibuya [45], and in [34], showing the same even for nonlinear systems of ODE. However, while many of the best-known ODE have their roots in physics, and some of them have divergent power series solutions, there is no such equation for which a formal power series solution cannot be summed by the method of k-summability, but instead requires the technique of multi-summation. Nonetheless, it is fair to say that multi-summability is the perfect tool for handling divergent solutions of ODE. The situation is slightly different for partial differential equations, as we shall briefly describe in the next subsection.

2.5.6
Applications to PDE

The technique of multi-summability can be applied to formal solutions of partial differential equations as well but, as one may expect, such an attempt meets with new difficulties. In order to understand this, it suffices to look at the very simple Cauchy problem

$$\partial_t u = \alpha \partial_x^2 u \,, \qquad u(0, x) = \phi(x)$$

where $\alpha \neq 0$ is an arbitrary complex constant, while the initial condition ϕ is assumed to be holomorphic in a disc about the origin. As was made clear in the Introduction, this problem includes both the heat or diffusion equation, where α would be a positive real number, as well as the Schrödinger equation, in which case α should be purely imaginary. For applications, one should interpret the variables t and x as real (and perhaps restrict t to positive values), but presently it is more natural to allow the two variables to be complex. In any case the above problem has the unique formal power series solution (2.4), and since the $2k$th derivative of a function, in general, is of magnitude $(2k)!$, one can see that this series diverges and is, for fixed x, of Gevrey order at most 1. Therefore, it is natural to investigate its 1-summability, for fixed x, and to do so, we form the series

$$v(t, x) = \sum_{k=0}^{\infty} \frac{(\alpha t)^k}{(k!)^2} \partial_x^{2k} \phi(x)$$

which has a positive radius of convergence. For $u(t, x)$ to be 1-summable in a direction d, it is necessary and sufficient to show that $v(t, x)$ can be continued with respect to t into a small sector bisected by the ray $\arg t = d$, and can be estimated by $C\exp(K|t|)$, for some constants $C, K > 0$. To do this, it is convenient to recall from [2] that this is possible for the function $v(t, x)$ if, and only if, it can be done for the slightly different function $w(t, x)$ given by

$$w(t,x) = \sum_{k=0}^{\infty} \frac{(\alpha t)^k}{(2k)!} \partial_x^{2k} \phi(x) = \frac{1}{2}\Big(\phi(x + (\alpha t)^{1/2}) + \phi(x - (\alpha t)^{1/2})\Big) .$$

From this formula, we read off that for 1-summability of the series (2.4) in a direction d, it suffices to assume that the initial condition $\phi(x)$, which by assumption is holomorphic in a disc about the origin, can be continued into a sector with bisecting direction $\arg x = (d + \arg\alpha)/2$, and can be estimated by $C\exp(K|x|^2)$. It is not obvious, but can be derived from general properties of k-summability that this condition upon $\phi(x)$ is also necessary. It is clear that every rational function $\phi(x)$ has this property for all but finitely many directions d, with the exceptional ones related to the location of its poles. On the other hand, there exist functions that are holomorphic in the unit disc, say, but cannot be continued beyond its boundary, and for such an initial condition ϕ, we conclude that the formal solution (2.4) is not 1-summable in whatever direction d.

The result described above has, in the case of $\alpha = 1$, been shown in [19]. For an application, it is most natural to choose the direction $d = 0$, since then, under the corresponding condition on ϕ, the sum of (2.4) is defined for all positive values of t.

2.5.7
Perturbed Ordinary Differential Equations

This subsection is about singularly perturbed systems of ordinary differential equations of the form

$$t^{r+1}\varepsilon^{\sigma}x' = g(t, x, \varepsilon)$$

where σ and the Poincaré rank r are natural numbers, and $g(t, x, \varepsilon)$ is holomorphic in a neighborhood of the origin of $\mathbb{C} \times \mathbb{C}^{\nu} \times \mathbb{C}$. Analysis of the dependence of solutions on the perturbation parameter ε is referred to as a singular perturbation problem. Under suitable assumptions on the right-hand side, such a system will have a formal solution

$$\hat{x}(t,\varepsilon) = \sum_{n=0}^{\infty} x_n(t)\varepsilon^n = \sum_{k=0}^{\infty} x_k(\varepsilon)t^k = \sum_{k,n=0}^{\infty} x_{kn}t^k\varepsilon^n$$

with coefficients $x_n(t), x_k(\varepsilon), x_{kn}$ given by differential recursion relations. In general, the series are divergent, but they are still useful if one applies summability in several variables.

In [27], Balser and Mozo-Fernández investigated the case of linear perturbation systems of the form

$$t^{r+1}\varepsilon x' = A(t,\varepsilon)x - f(t,\varepsilon) \qquad (2.20)$$

with a coefficient matrix $A(t,\varepsilon) = \sum_{k,n} A_{kn} t^k \varepsilon^n$, whose leading term is invertible (that is $A(0,0)$ is invertible) and an inhomogeneity vector $f(t,\varepsilon)$ whose entries are holomorphic near the origin.

A typical example in dimension $\nu = 1$ is

$$t^{r+1}\varepsilon x' = ax - f(t) \tag{2.21}$$

where $a \in \mathbb{C}\backslash\{0\}$ and f is holomorphic in a neighborhood of the origin. In this example, let r be not just a natural number, but $r \geq -1$. The problem has a unique formal solution of the form $\hat{x}(t,\varepsilon) = \sum_{n=0}^{\infty} x_n(t)\varepsilon^n$. The coefficients $x_n(t)$ can be recursively computed by

$$ax_0(t) = f(t)\ , \quad ax_{m+1}(t) = t^{r+1}x'_m(t)\ , \quad m \geq 0$$

If $r = -1$, this formal series is 1-summable in direction d for t in a sufficiently small disc around the origin if and only if the function f is holomorphic and of exponential growth at most 1 in some sector with bisecting direction $d - \arg a$. Furthermore, the coefficients $x_n(t)$ are holomorphic functions in t and if one considers $\hat{x}(t,\varepsilon)$ as a power series in t, it is even convergent.

The most important case, however, is $r \geq 1$. In this case, the formal series is 1-summable in a direction d for t in r disjoint sectorial regions of opening π/r, whose bisecting directions depend upon d. So, although the higher value of the Poincaré rank r causes a more complicated singular behavior of the solutions of (2.21) with respect to the variable t, it improves the situation when studying the dependence upon the perturbation parameter ε.

The case $r = 0$ is a special case, but not very interesting and therefore it will not be discussed here.

2.6 Periodic ODE

Other ODE of importance are those where $H(t)$ is *periodic* in the variable t. As examples, these appear when seeking periodic solutions of ODE [46] or in the Fredholm theory of periodic operators on Banach spaces and perturbations thereof. Historically, periodic ODE have played an important role in the study of planetary and lunar motion [47–49] and later in the quantum mechanical description of solids [50]. We will here consider only linear ODE of real variables.

2.6.1 Floquet–Lyapunov Theorem and Floquet Theory

For linear ODE of the form (2.3) with periodic coefficients $H(t)$ of period T and Lebesgue measurable entries, the Floquet–Lyapunov theorem [49, 51] implies the existence of fundamental matrices $X(t)$ of the form

$$X(t) = U(t)\, e^{\Lambda t} \tag{2.22}$$

where Λ is a constant matrix and $U(t)$ is a periodic function of period T [52].

2.6.2 The Mathieu Equation

The inherent problem of Floquet theory is to find just the matrix Λ. In important cases, Λ is diagonalizable whereupon its eigenvalues λ_i are called Floquet exponents. In many examples one is particularly interested in the dependence of the Floquet exponents upon some parameter of $H(t, \varepsilon)$ where H depends analytically on ε, in other words, one wishes to find the functions $\lambda_i(\varepsilon)$. We will here describe the determinantal approach to finding these exponents.

For the sake of simplicity, we will consider the 1-dimensional Mathieu equation [46]

$$x'' + (\varepsilon - 2q\cos(2t))x = 0 \tag{2.23}$$

where ε and q are real parameters. The Mathieu equation is the simplest nontrivial case of the Hill equation [47] where the periodic function is not restricted to one Fourier mode. The Mathieu equation has come to play an important role in the theory of ODE, in particular, its stability regions, defined by the values of ε for which Re $(\lambda_i) > 0$, have been studied extensively. A similar treatment for systems of linear ODE is also possible [53].

Inserting the Floquet representation

$$x(t) = \mathrm{e}^{\mathrm{i}\lambda t} \sum_{\kappa\in\mathbb{Z}} c_{2\kappa}(\lambda) e^{2\mathrm{i}\kappa t} \tag{2.24}$$

in (2.23), one is led to investigate the Fourier coefficients of the periodic function $U(t)$, $c_{2\kappa}$, which satisfy the recursion

$$c_{2\kappa} + q\frac{c_{2(\kappa-1)} + c_{2(\kappa+1)}}{(2\kappa-\lambda)^2 - \varepsilon} = 0\,. \tag{2.25}$$

In the case of a Hill equation with n Fourier coefficients, the corresponding recursion relation would be of order $2n + 1$.

2.6.3 The Whittaker–Hill Formula

Finding the non-trivial solutions of (2.25) is equivalent to finding λ such that

$$\det(A(\lambda; \varepsilon, q)) = 0 \tag{2.26}$$

where $A(\lambda; \varepsilon, q)$ is a linear operator on $\ell_2(\mathbb{Z})$ giving the system of recursions as $A(\lambda;\ \varepsilon,\ q)\bar{c} = 0$.

Such determinants have contributed to the creation of the theory of determinants of infinite matrices [47,54]. The latter determinant is to be understood in the sense of Poincaré [54], who defined it as the limit of a sequence of determinants of finite dimensional matrices. In this particular case, this determinant can be shown to converge absolutely for $2\kappa - \lambda \neq \pm\sqrt{\varepsilon}$ since $A - \mathbb{1}_{\ell_2(\mathbb{Z})} : \ell_2(\mathbb{Z}) \to \ell_2(\mathbb{Z})$ is *trace class*. A trace class operator T defined on a separable Hilbert space with eigenvalues μ_i satisfies $\sum_i |\mu_i| < \infty$. The determinant of $T + \mathbb{1}$ is then correctly defined by $\prod_i(1 + |\mu_i|)$ [55].

The determinant is an analytic function of the variable λ except at its poles, for which $2\kappa - \lambda = \pm\sqrt{\varepsilon}$. It is easily shown to be 1 periodic in λ with simple poles. This is also valid for the function

$$\mathcal{D}(\varepsilon, \lambda) = \frac{1}{\cos(\pi\lambda) - \cos\left(\pi\sqrt{\varepsilon}\right)} . \tag{2.27}$$

Hence, applying Liouville's theorem, there exists $C(\lambda)$ such that

$$\det(A(\lambda; \varepsilon, q)) - C(\lambda)\mathcal{D}(\varepsilon, \lambda) = 0 \tag{2.28}$$

Further asymptotic considerations in λ permit one to determine C, leading to the so-called Whittaker–Hill formula [25]

$$\sin^2\left(\frac{\pi\lambda}{2}\right) = \det(A(0; \varepsilon, q))\sin^2\left(\frac{\pi\sqrt{\varepsilon}}{2}\right) . \tag{2.29}$$

Note that this result still holds for the full Hill problem [25].

2.6.4 Calculating the Determinant

There are many schemes that attempt to calculate the Floquet exponent using the Whittaker–Hill formula [25,56–59]. These schemes rely on the analytic or algebraic properties of the sequence of matrices with index i of size $(2i + 1) \times (2i + 1)$ needed to determine $\det(A(0; \varepsilon, q))$ in the Poincaré sense.

In [59], the determinants, Δ_i of adjacent matrices in the sequence are put in relation to each other. This results in a third-order recursion for these determinants

$$\Delta_i = \beta_i\Delta_{i-1} - \alpha_i\beta_i\Delta_{i-2} + \alpha_i\alpha_{i-1}^2\Delta_{i-3} \tag{2.30}$$

where α_i, β_i are known elementary functions of the elements of the i-th matrices in the sequence. Note that this technique can easily be generalized to potentials with a greater number of Fourier coefficients in (2.23). The order of the recursion (2.30) will, however, also be higher. Such recursions can be brought to explicit closed forms [59]. The sought-for determinant is then $\det(A(0; \varepsilon, q)) = \lim_{i\to\infty}\Delta_i$.

Although such considerations do not explicitly enable one to compute λ using the Whittaker–Hill formula, they provide iterative methods to numerically approximate λ.

2.6.5 Applications to PDE

The Hill equations appear intrinsically in the eigenvalue problems of so-called Hill operators. They are differential operators of the form

$$\widehat{H} = -\hbar^2\partial_x^2 + V_\Gamma(x) \tag{2.31}$$

where $V_\Gamma(x) \in L_1\left(\mathbb{R}^d\right)$ is periodic with respect to some d-dimensional lattice Γ. The spectral properties of such operators have been extensively studied (for a short review see [60]). Lyapunov [49] proved that there are, in general, infinitely many stability intervals, which implies for periodic Schrödinger operators that the spectrum of $\widehat{H}$ cannot lie in certain bands. Moreover, such operators have an absolutely continuous spectrum. It is, in this context, customary to denote the eigenvalue band n as a function $E_n(\lambda)$ of the characteristic exponent λ. These functions play an important role in the quantum theory of solids. Considering the Schrödinger equation

$$\mathrm{i}\hbar\partial_t\psi = \left(\widehat{H} + \varphi(\varepsilon x)\right)\psi\ , \quad \psi(0,\ x) = \phi(x) \in L_2\left(\mathbb{R}^d\right) \tag{2.32}$$

the evolution on the subspace of the n-th energy band will, for small enough ε, be well approximated by Peierls' substitution. This consists in replacing the periodic part of the Schrödinger operator (2.32) $\widehat{H}$ by $E_n(-i\hbar\partial_x)$ [61]. Such approximate adiabatic evolutions are used exhaustively in atom and solid state physics.

References

1 Okubo, K. (**1971**) Connection problems for systems of linear differential equations, in *Japan-United States Seminar on Ordinary Differential and Functional Equations (Kyoto, 1971)*. Lecture Notes in Math., Springer, Berlin, **243**, pp. 238–248.

2 Balser, W. (**2000**) *Formal power series and linear systems of meromorphic ordinary differential equations*. Lecture Notes in Math. Springer Verlag, New York.

3 Schäfke, R. (**1998**) Confluence of several regular singular points into an irregular singular one. *Journal of Dynamical and Control Systems*, **4**(3), 401–424.

4 Birkhoff, G.D. (**1968**) *Collected Mathematical Papers Vol. 1*. Dover Publications, New York.

5 Knobloch, H.W. (**1958**) Zusammenhänge zwischen konvergenten und asymptotischen Entwicklungen bei Lösungen linearer Differentialgleichungs-Systeme vom Range 1. *Tokyo Journal of Mathematics*, **134**, 260–288.

6 Okubo, K. (**1963**) A global representation of a fundamental set of solutions and a Stokes phenomenon for a system of linear ordinary differential equations. *Journal of the Mathematical Society of Japan*, **15**, 268–288.

7 Okubo, K. (**1965**) A connection problem involving a logarithmic function. *Publications of the Research Institute for Mathematical Sciences Kyoto University series A*, **1**, 99–128.

8 Kohno, M. (**1966**) A two points connection problem involving logarithmic polynomials. *Publications of the Research Institute for mathematical Sciences Kyoto University series A*, **2**, 269–305.

9 Kohno, M. (**1968/1969**) On the calculation of the approximate values of Stokes' multipliers. *Publications of the Research Institute for Mathematical Sciences Kyoto University series A*, **4**, 277–297.

10 Kohno, M. (**1968**) The convergence condition of a series appearing in connection problems and the determination of

Stokes' multipliers. *Publications of the Research Institute for Mathematical Sciences Kyoto University series A*, **3**, 337–350.

11 Schäfke, R. (**1979**) *Über das globale analytische Verhalten der Lösungen der über die Laplace–Transformation zusammenhängenden Differentialgleichungen $tx'(t) = (A + tB)x$ und $(s - B)v' = (\varrho - A)v$.* PhD thesis, Essen.

12 Schäfke, R. (**1985**) Über das globale Verhalten der Normallösungen von $x'(t) = (B+t^{-1}A)x(t)$ und zweier Arten von assoziierten Funktionen. *Mathematische Nachrichten*, **121**, 123–145.

13 Balser, W., Jurkat, W.B., and Lutz, D.A. (**1981**) On the reduction of connection problems for differential equations with an irregular singular point to ones with only regular singularities; Part I. *Society for Industrial and Applied Mathematics Journal of Mathematical Analysis*, **12**.

14 Balser, W, Jurkat, W.B., and Lutz, D.A. (**1988**) On the reduction of connection problems for differential equations with an irregular singular point to ones with only regular singularities; Part II. *Society for Industrial and Applied Mathematics Journal of Mathematical Analysis*, **12**, 398–443.

15 Kohno, M. and Yokoyama, T. (**1984**) A central connection problem for a normal system of linear differential equations. *Hiroshima Mathematical Journal*, **14**, 257–263.

16 Yokoyama, T. (**1987**) Characterization of connection coefficients for hypergeometric systems. *Hiroshima Mathematical Journal*, **17**, 219–233.

17 Yokoyama, T. (**1988**) On the structure of connection coefficients for hypergeometric systems. *Hiroshima Mathematical Journal*, **18**, 309–339.

18 Hukuhara, M. (**1982**) Développements asymptotiques des solutions principales d'un système différentiel linéaire du type hypergéométrique. *Tokyo Journal of Mathematics*, **5**, 491–499.

19 Lutz, D.A., Miyake, M., and Schäfke, R. (**1999**) On the Borel summability of divergent solutions of the heat equation. *Nagoya Mathematical Journal*, **154**, 1–29.

20 Balser, W. (**2004**) Multisummability of formal power series solutions of partial differential equations with constant coefficients. *Journal of Differential Equations*, **201**(1), 63–74.

21 Magnus, W. (**1954**) On the exponential solution of differential equations for a linear operator. *Communications On Pure & Applied Mathematics*, **7**, 649–673.

22 Dahmen, H.D. and Steiner, F. (**1981**) Asymptotic dynamics of QCD, coherent states, and the quark form factor. *Zeitschrift für Physik C*, **11**, 247–249.

23 Dahmen, H.D., Scholz, B. and Steiner, F. (**1982**) First results of a calculation of the long range quark-antiquark potential from asymptotic QCD dynamics. *Zeitschrift für Physik C*, **12**, 229–234.

24 Dyson, F.J. (**1949**) The radiation theories of Tomonaga, Schwinger, and Feynman. *Physical Review (2)*, **75**, 486–502.

25 Whittaker, E.T. and Watson, G.N. (**1973**) *A Course of Modern Analysis. An introduction to the general theory of infinite processes and of analytic functions: with an account of the principal transcendental functions.* Fourth edition. Reprinted. Cambridge University Press, New York.

26 Röscheisen, C. (**2005**) Konfluente hypergeometrische Differentialgleichungssysteme mit mehrfachen Eigenwerten. Diplomarbeit.

27 Balser, W. and Mozo-Fernández, J. (**2002**) Multisummability of formal solutions of singular perturbation problems. *Journal of Differential Equations*, **183**(2), 526–545.

28 Grosche, C. and Steiner, F. (**1998**) *Handbook of Feynman Path Integrals*, volume 145 of Springer Tracts in Modern Physics. Springer-Verlag, Berlin.

29 Bäcker, A. and Steiner, F. (**2001**) Quantum chaos and quantum ergodicity, in *Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems* (ed Bernold Fiedler). Springer-Verlag, Berlin, pp. 717–751.

30 Steiner, F. (**2003**) Space–time aproach to quantum chaos. *Physica Status Solidi (b)*, **237**, 133–145.

31 Gantmacher, F.R. (**1959**) *Theory of Matrices, Vol. I & II*. Chelsea, New York.

32 Dirac, P.A.M. (**1926**) On the theory of quantum mechanics. *Proceedings of the Royal Society of London A*, **112**, 661–677.

33 Dirac, P.A.M. (**1927**) The quantum theory of emission and absorption. *Proceedings of the Royal Society of London A*, **114**, 243–265.

34 Balser, W. (**1994**) *From Divergent Power Series to Analytic Functions*, volume 1582 of Lecture Notes in Math. Springer Verlag, New York.

35 Ritt, J.F. (**1916**) On the derivatives of a function at a point. *Annals of Mathematics*, **18**, 18–23.

36 Stokes, G.G. (**1857**) On the discontinuity of arbitrary constants which appear in divergent developments. *Transactions of the Cambridge Philosophical Society*, **10**, 106–128.

37 Ramis, J.-P. (**1980**) Les séries *k*-sommables et leurs applications, in *Complex Analysis, Microlocal Calculus and Relativistic Quantum Theory*, (ed D. Iagolnitzer), Springer Verlag, New York, **126**, 178–199.

38 Balser, W. (**1978**) Einige Beiträge zur Invariantentheorie meromorpher Differentialgleichungen. *Habilitationsschrift*, Universität Ulm.

39 Balser, W. (**1982**) Solutions of first level of meromorphic differential equations. *Proceedings of the Edinburgh Mathematical Society*, **25**, 183–207.

40 Ecalle, J. (**1981**) Les fonctions résurgentes I–II. *Publ. Math. d'Orsay, Université Paris Sud*.

41 Balser, W. (**1992**) Summation of formal power series through iterated Laplace integrals. *Mathematica Scandinavica*, **70**(2), 161–171.

42 Balser, W. (**1992**) A different characterization of multisummable power series. *Analysis*, **12**, 57–65.

43 Balser, W. (**1993**) Addendum: "A different characterization of multi-summable power series" [Analysis **12** (**1992**), no. 1–2, 57–65. *Analysis. International Mathematical Journal of Analysis and its Applications*, **13**(3), 317–319.

44 Braaksma, B.L.J. (**1991**) Multisummability and Stokes multipliers of linear meromorphic differential equations. *Journal of Differential Equations*, **92**, 45–75.

45 Ramis, J.-P. and Sibuya, Y. (**1994**) A new proof of multisummability of formal solutions of nonlinear meromorphic differential equations. *Annales de L'Institut Fourier (Grenoble)*, 44:811–848.

46 Émile Mathieu. Mémoire sur le mouvement vibratoire d'une membrane de forme elliptique. *Journal de mathématiques pures et appliquées*, XIII (2ème série)(7-12):137–203, **1868**.

47 Hill, G.W. (**1886**) On the part of the motion of the lunar perigee which is a function of the mean motions of the sun and the moon, (1877), reprinted. *Acta Mathematica*, **8**(7-12), 1–36.

48 Poincaré, H. (**1987**) *Les Méthodes Nouvelles de la Mécanique Céleste*. A. Blanchard.

49 Liapounoff, A. (**1907**) Problème général de la stabilité du mouvement, (1892), reprinted from the mathematical society of Karkhow. *Annales de la Faculté des Sciences de Toulouse pour les Sciences Mathématiques et les Sciences Physiques. Série 2*, **9**, 203–474.

50 Bloch, F. (**1928**) Über die Quantenmechanik der Elektronen in Kristallgittern. *Zeitschrift für Physik*, **52**, 555–600.

51 Floquet, G. (**1883**) Sur les équations différentielles linéaires à coefficients périodiques. *Annales Scientifiques de l'École Normale Supérieure. Deuxiéme Série*, **12**, 47–88.

52 Coddington, E.A. and Levinson, N. (**1955**) *Theory of Ordinary Differential Equations*. McGraw-Hill Book Company, Inc., New York-Toronto-London.

53 Denk, R. (**1995**) Hill's equation system and infinite determinants. *Mathematische Nachrichten*, **175**, 47–60.

54 Poincaré, H. (**1886**) Sur les déterminants d'ordre infini. *Bulletin de la Société Mathématique de France*, **14**, 77–90.

55 Simon, B. (**1979**) *Trace ideals and their applications*, volume 35 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge.

56 Meixner, J. and Schäfke, F.W. (**1954**) *Mathieusche Funktionen und Sphäroidfunktionen mit Anwendungen auf physikalische und technische Probleme.* Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete, Band LXXI. Springer-Verlag, Berlin.

57 McLachlan, N.W. (**1964**) *Theory and Application of Mathieu functions*. Dover Publications Inc., New York.

58 Magnus, W. and Winkler, S. (**1966**) *Hill's equation.* Interscience Tracts in Pure and Applied Mathematics, No. 20. Interscience publishers John Wiley & Sons, New York, London, Sydney.

59 Sträng, J.E. (**2005**) On the characteristic exponents of Floquet solutions to the Mathieu equation. *Académie Royale de Belgique. Bulletin de la Classe des Sciences. 6ème Série*, **16**(7–12), 269–287.

60 Reed, M. and Simon, B. (**1978**) *Methods of Modern Mathematical Physics. IV. Analysis of Operators.* Academic Press (Harcourt Brace Jovanovich publishers), New York.

61 Panati, G., Spohn, H., and Teufel, S. (**2003**) Effective dynamics for Bloch electrons: Peierls', substitution and beyond. *Communications in Mathematical Physics*, **242**(3), 547–578.

3
A Scalar–Tensor Theory of Gravity with a Higgs Potential

Nils Manuel Bezares-Roder[1], *Frank Steiner*[1]

3.1
Introduction

3.1.1
General Relativity and the Standard Model of Particle Physics

One might imagine science as a huge building that has been growing through time. At first sight, this building seems to consist of independent blocks. However, in its foundations general topics are closely related with each other, but also still far apart in the lack of unifying concepts and theories, given the different interpretations that may exist. In theoretical physics, for instance, there are the theories of elementary particle physics for quantum phenomena. Together with their interpretation, they are now extremely well-grounded experimentally. On the other hand, there is General Relativity (GR) as the theory of gravitation. This represents another conceptual revolution fulfilled and grounded in the 20th century. It is formally comparable to "theories of quanta". However, both cannot even now be unified in a unique and consistent theory.

The construction plan and design of science is not always clear, but it develops, expands and interrelates its different subjects, implementing new ideas and results which come over time. Science has evolved progressively, so that new concepts always work as a seed for future interrelations, to produce a better understanding and new and wider horizons. It is exactly the same with learning. There, a neural network is founded as a building and new synapses are created for better interconnections through the axons between individual somata of the neurons which form the net. Such a unification may be fulfilled only in a progressive way. When complete, however, stimuli may be better transmitted and received. At the same time, these stimuli lead to new and better interconnections which need continuous refreshment. Between neurons, myelin is then needed for the isolation of axons, and the ability to create and modulate neurotransmitters to transmit the information is required.

[1] Corresponding authors.

In science, stimuli come from both theory and experiment and it cannot be ensured how they (as new concepts, for instance) may appear or mix, nor which of the channels in the existing net are best for them. Their correct reception, transmission and modulation, however, is part of the architecture of science.[2)]

A first grounding for transmission and reception of stimuli in scientific models and theories lies in empiricy. Hence, there are degrees of freedom in their construction and also in their final form. New concepts may come from new experiments and the hypotheses that follow them. Thus, interrelation between the established models and the feedback from evolving theoretical and experimental techniques is crucial. Self-doubt acquires a fundamental character to strengthen scientific achievements, too, with the viewpoint that scientific ideas may change and that science, in the end, does not attempt to dictate reality, but only to describe it.

In this context, it may be concluded that scientific theories and interpretations in natural science are related to epistemological ideas that are based on an empiricist-transcendental principle which implies that the validity of the theories depends on the knowledge obtained so far (see, for instance, the discussion by Einstein in [2] in relation to his opinion about metaphysics in interpretations). This knowledge, however, together with its interpretation, may change in time. Thus, a scientific theory cannot expect to be the final truth. However, within more general theories, the descriptions of nature can seek for a wider application area and for new predictions to be contradicted by experiments.[3)] Fundamental maxims, further, are unification, consistency and simplicity. These are followed, for instance, when extending gravitational models from Newtonian mechanics to Einstein's theory of General Relativity (GR). Not only does the latter entail Newtonian gravitation in the nonrelativistic limit of weak gravitational fields. It describes and accurately predicts effects such as perihelion advance (also known as perihelion shift; most notably in the case of Mercury's orbit), light deviation (leading to "gravitational lenses") and the Thirring–Lense effect (an effect caused by rotating massive bodies), which all lie beyond Newtonian dynamics. At the same time, GR unifies the concepts of space and time (therefore, we use Greek letters here as indices counting from zero to three) introducing curved Lorentzian manifolds and tensors as functions on the manifold itself to describe its properties (like energy and momentum, which now discovers a relationship with geometry!).

[2)] As in science, for progressive development, the reader should have the right to pick up some concepts as stimuli and to compare them with related ones in theories he might know or might be interested to know in the future. Paraphrasing Sartre, and in relation to positivistic thoughts of modern science, "the fundament of truth is freedom. Recognition is to bring what is ('das Sein' or ' l'être ') to the light, to act and search on the margin of mistakes, rejecting in this way ignorance and lies related with predisposition. And truth is this progressive disclosure, even though truth itself may be relative to the epoch in which it is achieved" (cf. [1]).

[3)] In the philosophy of science, [3] this is called *fecundity* and *independent testability*, which, together with *unity* i.e. a unique problem-solving strategy, should characterize a "good" scientific theory. In the same context, Karl R. Popper defines scientific theories and human knowledge in general as dependent on historico-cultural settings. Further, while absolute verifications cannot exist, nonrefutability of a theory is stated as nonscientific [4].

Within GR, the field equations (which give the dynamics) are usually known as Einstein's field equations. They were first derived in 1915 [5] and give the dynamics of gravitation. According to them, gravitation is understood geometrically as the curvature of the space–time manifold (that is, a generalization of space carrying time as a further dimension) and no longer as a force. This curvature is given by the Einstein tensor $G_{\mu\nu} = R_{\mu\nu} - 1/2Rg_{\mu\nu}$.[4] According to Einstein's field equations, $G_{\mu\nu}$ is related to the energy-momentum tensor $T_{\mu\nu}$ of matter:

$$G_{\mu\nu} = -\kappa T_{\mu\nu} \ , \tag{3.1}$$

whereas the latter is derived from the so-called Hilbert–Einstein action

$$S_{\mathrm{HE}} = \int \left[\frac{1}{\kappa} R + \mathcal{L}_{\mathrm{M}} \right] \sqrt{-g} \mathrm{d}^4 x \ . \tag{3.2}$$

For an action, the variation is postulated as vanishing, according to the Hamilton Principle of Least Action. This principle leads correctly to Newtonian or quantum mechanics; electromagnetism or gravitation, for instance. This depends on the Lagrangian (or Lagrange density) chosen under the integral of the action S. Thus, (3.2) is the fundament of normal GR and defines all gravitational interactions.

Here, g denotes the determinant of the metrical tensor $g_{\mu\nu}$, and $\kappa = 8\pi G/c^4$ is related to Newton's gravitational constant G and the speed of light c, which is postulated as constant. Furthermore, the energy-momentum tensor $T_{\mu\nu}$ is related to the Lagrangian of matter $\mathcal{L}_{\mathrm{M}}$, which is the fundament of a theory, where (especially fermionic) matter is defined in terms of the wave function given by the state ψ, in accordance with quantum mechanics. And it is quantum mechanics which indeed leads to the idea of field theories instead of only theories for the dynamics of particle systems. Within quantum theories, trajectories are no longer defined. But their analog can be found in the quantum mechanical state, as the system "blurred" in space. As the eigenvector of an observable, the state gives the probability of qualities to be measured.

Within quantum mechanics, a measure of a property given by an operator $\hat{A}$ with real-valued eigenvalues (an observable) is related to the so-called *collapse* of the wave function. The observer is understood as part of the whole quantum mechanical system, and as such there is an intrinsic interaction between him and

[4] In the notation, the Einstein convention for the addition states that for a twice-appearing index; once as an upper (in physics usually called covariant) and one as a lower (contravariant) one, is to be summed over it. Further, here $R_{\mu\nu}$ denotes the so-called Ricci tensor, while R is the Riemann (or Ricci) scalar of the metric. Since in curved space inversion of the first and second derivatives of a vector does not lead to the same mathematical object, the so-called Riemann tensor $R^{\lambda}_{\mu\sigma\nu}$ may be defined as a measure of this permutation loss. Its trace $R^{\lambda}_{\mu\lambda\nu} = R_{\mu\nu}$ is the Ricci tensor, and the trace of it then R. All of these can be written as functions of derivatives of the metric $g_{\mu\nu}$, which gives the properties of the space–time and mathematically defines the scalar product. In a flat space–time like the pseudo-Euclidean, Minkowski one, of Special Relativity (which can be found in normal quantum mechanics where gravitation is neglected), the curvature tensors vanish.

the analyzed subsystem. This interaction leads to loss of information in form of the "collapse" (to the basis of the observed property with eigenvalues a, which are the statistical mean value of measurements). If the eigenvalues of $\hat{A}$ and of a further observable $\hat{B}$ possess different eigenvectors, the measurement of both is not commutative ($\hat{A}\hat{B} \neq \hat{B}\hat{A}$), and a change in the order of measurement leads to different mean results. The quantum mechanical state, which gives the properties of the analyzed system, changes after the first measurement. A second measurement represents an interaction with a different system where information was lost (this is closely related with Heisenberg's *uncertainty relation*, often given for the measurement of a momentum $\hat{p}_x$ and a coordinate $\hat{x}$ which a particle may possess as properties at a specific moment. As canonical conjugate operators, $\hat{p}_x$ and $\hat{x}$ are orthogonal to each other and thus possess a different basis. Consequence: it is not possible to know both the exact momentum (velocity) and the place of say an electron at the same time).

In the same way, since a realistic quantum system is never isolated, the interaction of the state with its environment is important. There are quantum correlations between them, and this interactions may be understood as a sort of measurement, again related with a collapse, but especially with the so-called *decoherence* [24]. Through this, superpositions of the wave functions, a fundamental property of quantum physics (mathematically founded in the linearity of the Hilbert space, in which quantum mechanical states exist), vanish.[5] The idea is that classical mechanics should be recovered from quantum mechanics by means of the quantum properties themselves, especially for large sizes and masses of the observed system. Quantum properties cancel out, leading to the classical world, it is said. The dynamics that will explain the collapse and define a complete theory of measurement has not yet been completely explained, though. It is related to what is called "the problem of definite outcomes" and that of the "preferred basis". Together these form *the measurement problem*, and their further research relies on quantum information theory.

Research on quantum information, related for instance to quantum and nonlinear optics, leads to many new and classically unexplicable effects such as entanglement (from the so-called EPR paradox of Einstein, Podolsky and Rosen [6]) and quantum teleportation [7] (an experimental fact since Zeilinger's experiment in 1997 [8]!), which is fundamental to the concept of quantum computers [9] and which should further be explained in relativistic contexts (where *a priori* information of the state to be teleported is necessary to label identical particles in order to make them effectively distinguishable [10]). However, neither usual (Schrödinger's nonrelativistic, or even Dirac's special-relativistic for particles with spin) quantum mechanics alone, nor General Relativity, can describe the nature of matter itself. This is rather fulfilled within the context of (special- relativistic as well as quan-

[5] A well-known example of superpositions in quantum mechanics is *Schrödinger's cat*. It is described as alive and dead at the same time until measurement (observation) "decides" its classical status.

tum theoretic) elementary particle and high-energy physics. The latter evolved out of nuclear physics with the desire to discover the foundations of matter and its fundamental dynamics.

Modern particle physics theories have their beginning and interpretation basis in the early 20th century. Back then, H. Yukawa proposed that nuclear particles were held together (against electromagnetic repulsion) by the mediation of particles that he proposed (the pi-mesons or pions) [11]. These particles are massive (with about 140 MeV/$c^2 \approx 2.5 \times 10^{-28}$ kg mass) and do indeed mediate short-range interactions within the nucleus.[6] However, pions decay in leptons, and are thus not fundamental. They mediate only residual interactions. Further, the nucleus-conforming particles (called hadrons in the general case) possess a finite diameter of about 10^{-15} m and also an inner structure [13]. In this context, Gell-Mann [15] and Zweig [16], independently of each other, interpreted a nonelementarity and introduced constituent particles of hadrons back in 1964. The experimental evidence for these [14], finally, was acknowledged with the Nobel prize in 1990.

Elementary particle physics describes dynamics on the basis of quantum field theories and hence of quantum mechanical states as property carriers for measurements. States are there related to constituent particles. Furthermore, modern theories are based on Yang's and Mills' field theory of 1954 [17], utilizing *N*-dimensional wave functions as states.[7] The latter, however, is necessarily combined with the Higgs (or Higgs–Kibble) mechanism of symmetry breakdown [18] (cf. later in this subsection), through which the theory is renormalizable [19] ('t Hooft and Veltman were awarded the Nobel prize for this proof in 1999). It further neglects gravitation since gravitation is by far the weakest of the four known fundamental interactions (gravitational, electromagnetic, weak and strong).

The main example within elementary particle theories is the unifying Glashow–Salam–Weinberg Model of the electroweak interactions (Nobel prize awarded in 1971), which, together with so-called Quantum Chromodynamics (QCD) of the strong interactions (named in analogy to electrodynamics but with "color" – $\chi\varrho\acute{\omega}\mu\alpha$ – charge instead of only an electric one) of Gell-Mann and others, leads to the Standard Model (SM) of elementary particles under the symmetry group $\mathrm{SU}(3)_{\mathrm{C}} \times \mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)_{\mathrm{Y}}$. Here, the constituents are part of a multiplet or isovector which is conformed by those particles which are indistinguishable within a specific interaction. The group dimension (and subscript), hence, is given by the particles represented in each group: three differently "colored" (C) quarks for the strong interactions, electrons

[6] For his prediction in 1935 and for the development of experimental techniques which resulted in the discovery of pions [12], Yukawa and Powell were awarded with the Nobel prize in 1949 and 1950, respectively.

[7] The Yang–Mills theory is a non-Abelian (non-commutative) theory with SU(N) transformations (i.e. unitary matrix-valued transformations for *N* dimensions and determinant +1 for the transformation operator or matrix) and thus with self-interactions that generalize the Maxwell equations of (Abelian U(1)-) electrodynamics (where photons as gauge bosons – mediators – do not self-interact) to the so-called and analog Yang–Mills equations. The dimension *N* of the symmetry group gives the number of so-called gauge currents that mediate the interactions, according to the usual interpretation of these theories.

and neutrinos (leptons L) for the weak interactions, and electrons for electromagnetism. Y stays for the "hypercharge", which is related to the usual electromagnetic charge of electrons by the so-called Gell-Mann–Nishijima formula.

The SM, as a quantum field theory of interacting fundamental fields, is based on the so-called gauge principle or gauge invariance, which leads to the covariant derivatives. These make it possible for derivatives to maintain their tensorial character, and can be introduced in terms of parallel transport (and holonomy) in curved space (a sphere, for instance). There, a usual derivative leads outside of the manifold. An additional term is needed for correction, that is to move parallel to the surface of the sphere during the derivation. This additional term is related to so-called connections, such as Christoffel symbols $\Gamma^{\mu}_{\nu\lambda}$ in GR or gauge fields (or potentials) A^{μ} in usual gauge theories. Furthermore, the SM has built in Gell-Mann's [15] and Zweig's [16] idea of quarks as fundamental constituents of hadrons (like protons, neutrons and more massive particles). Interactions between quarks, then, are understood to be mediated through the gauge fields, with the so-called gauge bosons as the field quanta of the interactions. Gauge bosons thus possess an analogous property to the mesons in Yukawa's earlier theory, but they do have a truly fundamental character, related to potentials (photons are related to the electromagnetic potentials, for instance).

The above ideas suggest the composition of the universe through different eras after the Big Bang, after which more complex types of matter evolved. The composition of the early universe includes different types of composite and noncomposite matter. The composite one is then divisible as the superposition of the main constituents bound together by gluons as gauge bosons of the strong interactions. Gluons, as gauge bosons, are the mediators of the strong interactions, which, in the end, hold atoms together through short-ranged residual interactions.

In analogy to quantum electrodynamics (QED), in quantum chromodynamics (QCD) of strong interactions, a (strong) color charge is defined. However, this charge exists in three types (named blue, red and green, like the three primary colors for additive combinations in color theory) and three "anti-types" (named anti-blue, anti-red and anti-green). Further, although both photons and gluons, gauge bosons, are massless and thus QED- and QCD- interactions are long-ranged, other than photons, gluons self-interact and carry and mediate a charge (actually both a color- and an anti-color-charge). And while quarks are "colored" i.e. possess a color charge, the resulting superposition of all free particles in nature is assumed to be "colorless" ("white"-charged, following the analogy with color theory). This is related with the problem of *confinement* and *asymptotic freedom*: quarks cannot be detected as free particles, since strong-interaction (color) forces should augment with distance. Within a hadron (femtometer scale), however, quarks would move freely.[8]

[8] The predicted interaction between the color-mediating gluons and quarks in hadrons, first discovered in the early 1970s, achieved the Nobel Prize for Gross, Wilczek and Politzer in 2004.

Unlike hadrons, quarks do not seem to possess an inner structure and in particular, the SM assumes none. What makes them different from each other is simply called *flavor* (there are six of them, and six anti-flavors). Otherwise, quarks are only divided into three families or generations. In each of them, an (iso-)pair (anti-pair) of quarks is found. With a member with positive fractional electric charge $Q = +2/3e$ and another one with negative fractional electric charge $Q = -1/3e$ (up and down, charm and strange, top and bottom, and antiquark analogs). The first generation consists of the less massive quarks with $m_{\mathrm{up}}/m_{\mathrm{down}} \approx 0.56$ and m_{up} of about 2 MeV/c^2. Top quarks (proven experimentally only in 1995 at Fermilab [20]), of the third and most massive family, on the other hand, possess a mass of about 171 GeV/c^2 (an electron Volt eV is about 1.6×10^{-19} Joules. It is the amount of energy equivalent to that gained by a single unbound electron when it is accelerated *in vacuo* through an electrostatic potential difference of one volt): about 1000 times more massive than (composite) pions and with almost twice the mass of weakons!

The matter class composed especially by quarks is called hadronic. The relation of this composite matter with its main constituents as well as its subclassification are listed in a simplified way below:

Class of composite **HADRONS (H)** **H1) Baryons**	**Constituents** **QUARKS and GLUONS** **3 quarks OR 3 antiquarks**	**Examples**
H1.1) *Nucleons*	up- and down- quarks (antiquarks) and gluons	proton neutron antiproton antineutron
H1.2) *Hyperons*	with top-, bottom-, charm- or strange- quarks (antiquarks) also	Ω^- (3 strange-quarks) Λ_C^+ (up, down, charm)
H2) **Mesons**	**ONE quark and ONE antiquark**	
H2.1) *Flavorless*	up-, down- quarks and antiquarks and gluons	pion π^+ (up, anti-down)
H2.2) *Flavored*	up-, down, top-, bottom-, strange- quarks and antiquarks and gluons	kaon K$^+$ (up, anti-strange)

Nuclear matter consists only of quarks of the first generation. Thus, given that such combinations are generally preferred energetically, the main type of baryonic matter is nucleonic. However neither baryonic, nor hadronic matter in general, are the only types. Not even including composite matter. Other classes that are known experimentally are those consisting of bosons or leptons. Leptons are sometimes bound in composites as leptonia (like electron–antielectron–pairs, for instance). At the same time, gluons, for example, bound in so-called glue-balls which acquire dynamic mass and may explain in the SM the short range of (effective) nuclear forces (about 2.5 $\times 10^{-15}$ m, in contrast to pure strong interactions, which are long-ranged, since gluons do not possess mass).

The elementary nonhadronic (and thus nonbaryonic) matter is listed below:

Class of matter	Constituents	Frequent symbol	Some properties
Bosonic matter	photons	A or γ	Mediate electromagnetism. Uncharged. No mass
	gluons	G_i (nine types)	Mediate strong interactions. Possess color-charge and -anticharge
	weakons	W^+, W^-, Z^0	Mediate weak interactions. Lead to β-decay. Massive
Leptonic matter	electron- (positron-) analogs	$e^\pm$, $\mu^\pm$, $\tau^\pm$	Massive. Three leptonic generations with $m_e < m_\mu < m_\tau$
	neutrinos (antineutrinos)	ν_i and $\bar{\nu}_i$ ($i = e, \mu, \tau$)	Only (gravitationally and) weakly interacting. Nonvanishing small mass

The latter category (of neutrinos and their antimatter counterpart) is especially relevant in an astrophysical description of matter. In astrophysics, there is a crucial discussion about which kind of matter may be perceived (almost) only gravitationally. And this seems necessary to explain some empirical data which may contradict the usual models of astrophysical dynamics if this type of matter is not taken into account. Neutrinos do not couple electromagnetically and are thus very difficult to detect directly. In an astrophysical context they are therefore called hot dark matter (HDM). Dark because they lack electromagnetic coupling (which makes them very difficult to detect, so that 25 years passed from their prediction by Pauli in the late 1930s [21], even before neutrons were discovered, until their 1995 Nobel-prize winning discovery by Reines and Cowan in 1956 [22]. Hot, because of the high velocity of neutrinos related to their almost, but according to neutrino oscillations not vanishing, mass of, at most a few eV/c^2).

Dark matter (DM), though, can be baryonic (like a brown dwarf, for instance) or nonbaryonic. Nonbaryonic DM can be like the HDM. But it may be an exotic example as well, generally called "cold"; some kind of as – yet undiscovered matter, a particularly important likely candidate of which being light supersymmetric particles like neutralinos or gravitinos. These are introduced under the assumption that there is a special symmetry between bosons (with integer spin as "quantum number") and fermions (with half-integer spin) in nature, so that for each boson (fermion) there exists a partner fermion (boson) as a so-called superpartner. Here, gravitinos are superpartners of gravitons in a quantum theory of gravitation, and neutralinos are quantum theoretical superpositions of the superpartners of the Z-bosons, of photons (neutral gauginos) and of neutral Higgs particles of supersymmetric theories (Higgsinos). The latter are thought to mix due to the effects of electroweak symmetry breaking (when both electromagnetic and weak become independent interactions, leading to *massive* weakons characterizing the broken symmetry). As heavy, stable particles, neutralinos, in particular, seem to be good candidates for cold dark matter (CDM) as very weakly interacting massive particles (WIMPs). They are assumed to decay finally especially in τ-leptons, although decay channels including supersymmetric particles such as neutral Higgsinos, for instance, are also expected [23]. The neutralino mass is thought to be over 10 GeV/c^2, and evidence of annhihilation of such particles in regions which are expected to be highly "dark-densed" is hoped will be found in γ-ray and neutrino telescopes.

From a theoretical point of view though, it is not only possibly still unobserved supersymmetric particles which should be taken into account. There are also cosmological relics from symmetry-breaking processes which are predicted by high-energy physics to be included in a list of the universe's components [25]. All these particles and fields, as far as they really exist in the physical world, should have played a role in structure formation. They therefore imply the existence of an exotic part of the dark components of the density of the universe (that is, of the components such as those of dark matter which we do not directly see, or the nature of which is still unclear).

According to GR dynamics, nowadays it is generally believed that dark components are by far the major part of the total matter-like density that exists (cf. [26]), but not only of the matter-like components, though! However, at this point the nature of these dark density components is unclear and also whether this complexity in interpretation might indicate a deeper, new physics, better described by more general models. Here, for instance, we intend to ground a model for which the assumption of scale-independent dynamics is no longer valid. This means that small or large ranges show different dynamics to those we usually perceive. We intend to change the fundamental properties of the models in order to obtain a more general one which might be used in a simplified and unified way to explain empirical complexity by means of changed dynamics, in comparison with usual GR. The methods and ideas in the subsequent sections are, however, not new. They may be traced back to the beginning of the 20th century but might allow the reader to view some concepts and building properties of modern theoretical physics and its structure. The general, covariant form of such theories (usual in Relativity) as in Section 3.2.2 is very compact and helpful. It is, however, rather difficult to understand for readers not used to it, though. Thus, the outlook in Section 3.2.4 will help with two special cases which are then simplified in terms of dark matter and cosmology.

The dark components of density are usually divided into those of matter, the main component of which is called cold dark matter (CDM) (as opposed to the HDM components), and those of dark energy (which acts anti-gravitationally i.e. against gravity, or, equally, as having negative pressure). The latter may be identical with Einstein's cosmological constant Λ_0, introduced by Einstein in 1917 [27] (however, with the idea of a closed and static universe) by replacing $G_{\mu\nu}$ in his Equations (3.1) by $G_{\mu\nu} + \Lambda_0 g_{\mu\nu}$. Λ_0 is interpretable as the energy density of a vacuum. A particular candidate for this is the scalar field commonly known as quintessence or "cosmon" field [28, 29] as a theoretical carrier (generally coupled minimally to gravitation in modern standard theories or with a scalar field coupling to R which stays almost constant). The cosmological constant represents a special case of dark energy that does not change with time (cf. [30]) but which should also be explained within a quantum theory of gravitation.

Dark energy is related to the phenomenon of cosmic acceleration (see Section 3.2.4) since anti-gravitation interactions would lead to a repulsion of matter after the Big Bang. However, for a long time, it was strongly believed to be vanishing. There were only few empirical data which would point to anti-gravitation. And it was only in the last decade of the 20th century that this assumption of a vanishing Λ began to fall apart. A nonvanishing value for dark energy was measured within the context of GR for Super Novae of type Ia (SNeIa) as extragalactic distance indicators [31–33].[9] And in the years that followed, the results were corroborated. Thus, cosmic expansion seems to be accelerated. However, by now it is unclear if the value of this dark energy (as the anti-gravitation component) stays constant in time, as a true cosmological constant Λ_0, or wether today's dark energy component is the remainder of some cosmological function. This function should contribute as $\Omega_\Lambda \approx 0.7$ today to the total density parameter Ω_T of the current universe. Here, Ω_i is defined as a dimensionless parameter for a given energy (mass) density $\varepsilon_i = \varrho_i c^2$, given by $\Omega_i = \varrho_i/\varrho_\mathrm{c} = \varepsilon_i/\varepsilon_\mathrm{c}$, where ϱ_c is a critical density defined in terms of G, c and the Hubble constant H_0, which, on the other hand, is a measure of the cosmic expansion, see Section 3.2.4. Earlier experimental works like [34] and [35] have already proposed a nonvanishing, but over-abundant cosmological constant (for a slightly closed ($K = 1$, $\Omega_\mathrm{T} \approx 1$) baryonic-matter dominated universe). Nowadays standard measured values of the models, however, are $\Omega_\mathrm{m} = 0.127\ \mathrm{h}^{-2}$ for matter, including $\Omega_\mathrm{B} = 0.0223\ \mathrm{h}^{-2}$ for baryons and $\Omega_\mathrm{DM} = 0.105\ \mathrm{h}^{-2}$ for dark matter ($h = 0.73$ gives the normalized modern Hubble expansion rate). For neutrinos, the constraint lies at $\Omega_\nu < 0.007h^{-2}$, and the cosmological constant density yields $\Omega_\Lambda = 0.76$. According to the three-year results of WMAP, the total energy density parameter lies around $\Omega_\mathrm{T} = 1.003^{+0.013}_{-0.017}$ [36]. A value of 1 means a curvature $K = 0$ of a flat universe, while higher values mean a closed universe with $K = 1$, and lower ones indicate an hyperbolic universe with $K = -1$.

If dark energy components should change in time, the scalar field of quintessence might be one that acts on local planetary [37] or galactic scales [38]. Moreover, if coupled nonminimally to gravity, massive fields of that kind might even account for the phenomenology of dark matter components of galaxies and thus contribute to the understanding of their flat rotation curves [39, 41]. According to Newtonian dynamics, the tangential velocity of spiral galaxies should decrease

[9] SNe Ia are variable stars which (simplified) result from a violent explosion of a white dwarf star which has completed its normal stellar life and where fusion has ceased. After having ignited carbon fusion, the released energy and subsequent collapse has unbound the star in the supernova explosion. For the type Ia especially, the spectrum shows a lack of hydrogen lines but indicates singly-ionized silicon.

inverse-proportionally to the distance from the galactic nucleus. Empirically, this is not the case and the velocity stays nearly constant (flat), as if there were a halo of invisible massive matter (dark matter), the dynamics of which is dependent of the visible matter of the galactic disc. A scalar field, though, may lead to the same dynamics and flattened curves of the tangential velocity.

The question of whether scalar fields exist at all is still open. However, the Higgs field for spontaneous symmetry breaking (SSB) [18] is of special relevance as a basis for the SM (but not the only one!). The underlying mechanism, related to a scalar field (which can be related to an order parameter within the context of solid state physics, and specifically with the complex one of the Ginzburg–Landau theory, or with Cooper pairs – bosons – in the BCS theory, both of superconductivity), predicts the existence of new particles, especially the massive ones known as Higgs particles in the case of the Higgs' mechanism. In the case of Goldstone's mechanism, on the other hand, there are also massless Goldstone particles. Through a unitary transformation they are absorbed by the gauge fields and the Higgs mode is achieved, with the appearance of only massive particles (cf. [45]).

The appearance of these particles is related to a breaking of the symmetry when energy scales are low enough and the ordered state becomes unstable. The scalar field as the order parameter and as the most likely state becomes nonvanishing (order, symmetry, is broken). And the latter is related to the scalar field potential and its degeneracy after the universe cools following the Big Bang. Then, the ground-state value (or vacuum expectation value (VEV)) of the field becomes nonvanishing and it is energetically more favorable to create a particle (the Higgs particle) than to have it disappear. This particle is related to the nonvanishing field. If, on the other hand, there were a real invariant vacuum, the symmetry-breaking mode becames a Wigner mode. Modes of this kind are found in relation to degeneracies like the ones within the Zeeman effect (see discussion in [45]).

The Higgs mode within Abelian Higgs models (coupled with electrodynamics) gives a more elementary explanation of the Meissner effect and superconductivity from the appearance of the (dynamical) mass of photons, and within QCD it gives a method for the explanation of the confinement of quarks in color singlets within hadrons through a dual form of superconductivity i.e. with a color-magnetic superconductor (see [46]).

Higgs particles are expected to be found in high-energy experiments such as in the LHC, the particle accelerator in Geneva, Switzerland (see [45]). They represent the one still unverified prediction of the SM, which has proven very successful. Still, the SM postulates Higgs fields in order to be renormalizable [42] (i.e. especially avoiding divergences in perturbation theory) and thus so to obtain a realistic physical description.

The SM of elementary particle physics has been remarkably successful in providing the astonishing synthesis of the electromagnetic, weak and strong interactions of fundamental particles in nature [43,44]. According to this, inertial as well as passive gravitational mass[10] are introduced as generated simultaneously with respect to gauge-invariance by the interaction with a scalar Higgs field through the SSB. Then, considering the Higgs field for small enough energy scales, the Higgs field couples to matter. By means of this interaction, matter no longer moves as fast as the speed of light. It spontaneously possesses mass. However, the latter is generated or explained in the theory by an interaction between particles, but only within elementary particle physics and not within GR. The Higgs mechanism of SSB [18] provides a way for the acquisition of mass by the gauge bosons and fermions in nature, but the mass is reduced to the parameters of the Higgs potential, the physical meaning of which is still not completely understood.

On the other hand, following Einstein's idea of the principle of relativity of inertia, mass should be produced by the interaction with the gravitational field [47]. Einstein had argued that the inertial mass is only a measure for the resistance of a particle against a relative acceleration with respect to other particles; therefore, within the theory of relativity, the mass of a particle should originate from an interaction with all other particles of the universe. This interaction should be the gravitational one, which couples to all particles i.e. to their masses or energies. Furthermore, Einstein postulated that the value of the mass should go to zero if one put the particle at an infinite distance from all the others. GR, however, was not able to realize this principle, commonly known as (Einstein's version of the) Mach's principle. Nevertheless, these ideas led Carl H. Brans and Robert H. Dicke to develop their scalar–tensor theory (STT) [48] (see Section 3.1.2), although an explanation of mass did not follow from it either. However, the cosmology of scalar–tensor theories leads naturally to cosmic acceleration [49]. This makes the scalar fields of such theories the natural candidates to be quintessential-like fields [50–52].

Interestingly, the Higgs mechanism lies in the direction of Einstein's idea of mass production. As a result of the fact that the Higgs field itself becomes massive after symmetry breaking, it mediates a scalar gravitational interaction between massive particles. This interaction is, however, of Yukawa type (short-ranged): there is a type of gravitational interaction of the Higgs field between massive fermions [53] as well as bosons within the Glashow–Salam–Weinberg model of electroweak interactions based on the localized group $SU(2) \times U(1)$ [54] but not restricted to it; further, this is valid in all cases of mass production by symmetry breakdown via the Higgs mechanism [55], so that it is natural to couple scalar fields, and especially Higgs-like fields, to the curvature of GR within scalar–tensor theories (STTs) (see also the analysis in [56, 57]). Actually, both the 4-force of the gauge field and the Higgs field act on the matter field. The gauge-field strength couples to the gauge-currents, that is to the gauge-coupling

[10] Inertial mass is defined as a measure of an object's resistance to the change of its position due to an applied force. Passive gravitational mass is also a measure of the strength of the gravitational field due to a particular object (see [45], especially in relation to symmetry breaking modes and the Higgs mechanism). Although conceptually different, Einstein's principle of equivalence asserts that they are equal for a given body and, this has now been well-grounded experimentally.

constant g, whereas the Higgs field strength (gradient of the Higgs field) couples to the fermionic mass-parameter k. This points to a gravitational action of the scalar Higgs field. In the case of a scalar gravity only massive particles should interact and the only possible source of a scalar gravity is the trace of the energy-momentum tensor [55]. Further, even in the field equations of the SM, the scalar field plays the role of an attractive scalar gravitational potential between massive particles.

3.1.2 Alternative Theories of Gravity and Historical Overview

In modern quantum theories, interactions between equally charged particles mediated by bosons with odd spin are repulsive (as in quantum electrodynamics for equally charged particles, since photons possess spin 1, while QCD-confinement comes from an attractive force, given the differently color-charged quarks of hadrons). Those by bosons with even spin are attractive. Further, from Einstein's GR (in analogy to quantum theories) it follows that the gravitational interaction is, in its quantum-mechanical nature, mediated by massless spin-2 excitations only [58]. This is expected to be related to the still-hypothetical gravitons as intermediate particles of a quantum theory of gravity. Classically, to describe this interaction, the gravitational Lagrangian of the theory (which follows the Euler–Lagrange equations for a field) describes the propagation and self-interaction of the gravitational field only through the Ricci scalar R (see (3.2)).

Scalar–tensor theories (STTs), on the other hand, postulate in this context the existence of more complex dynamics from further mediating particles, named in this case *graviscalars*, within the context of quantum theories. This means that STTs modify classical GR by the addition of scalar fields to the tensor field of GR. They further demand that the "physical metric" $g_{\mu\nu}$ (coupled to ordinary matter) be a composite object of the form $g_{\mu\nu} = A^2(\hat{\phi})g^*_{\mu\nu}$, with a coupling function $A(\hat{\phi})$ of the scalar field $\hat{\phi}$ [59].

The first attempts at a scalar–tensor theory were started independently by M. Fierz in 1956 [60] and by Pascual Jordan in 1949 [61]. The latter noticed through his isomorphy theorem that projective spaces such as Kaluza–Klein's (five-dimensional) can be reduced to usual Riemannian four-dimensional spaces and that a scalar field as a fifth component of such a projective metric can play the role of a variable effective gravitational "constant" G_{eff}, which is typical for STTs and by which it is possible to vary the strength of gravitation [62] (thus, obviously violating to some degree the strong equivalence principle). Furthermore, this kind of general-relativistic model with a scalar field is equivalent to a multi-dimensional general-relativistic model [63]. Many theories do involve this physics (e.g. string theories or brane theories), but scalar-tensor theories are typically found to represent classical descriptions of them [64].

In his theory, Jordan introduced two coupling parameters of the scalar field. One parameter produced a variation in the gravitational constant. The other one broke the energy conservation by a nonvanishing divergence of the energy–momentum tensor to increase the mass in time, in accordance with the ideas of Jordan and Dirac [65]. However, the cosmic microwave background radiation (CMB) as a real black-body radiation discovered in 1965 [66][11] forces one to accept general energy conservation as experimental fact [67].

Jordan's theory was worked out independently by Brans and Dicke in 1961 [48] without breaking energy conservation, but again introducing a scalar field with an infinite length scale and playing the role of a variable gravitational coupling. The generalization to GR's action (3.2) was then proposed as

$$S_{\mathrm{JBD}} = \int \left[\hat{\phi} R + (16\pi/c^4)\, \mathcal{L}_{\mathrm{M}} - \frac{\omega}{\hat{\phi}} \left(\frac{\partial \hat{\phi}}{\partial x_\mu} \frac{\partial \hat{\phi}}{\partial x^\mu} \right) \right] \sqrt{-g}\, \mathrm{d}^4 x \,. \tag{3.3}$$

Here, we have the determinant g of the metric tensor, the (Ricci) curvature scalar R, the matter Lagrangian $\mathcal{L}_{\mathrm{M}}$ and a scalar field $\hat{\phi}$ which plays the role of the reciprocal Newtonian constant G^{-1}. The first term of (3.3) couples the scalar field and gravitation given by R, while the third term represents the kinetic energy of $\hat{\phi}$, since the Lagrange density $\mathcal{L}$ (conceptually derived from the Lagrange function of mechanics) is usually defined in terms of the substraction of the potential from the kinetic energy of the analyzed system.

Other than in the original theory of Jordan, the Brans and Dicke theory in (3.3) does not contain a mass-creation principle. The wave equation of $\hat{\phi}$ can be transformed so as to make the source term appear as the contracted energy–momentum tensor of matter alone. In other words, the inhomogeneous part of the wave equation is only dependent on the trace T of the tensor $T_{\mu\nu}$, and this is in accordance with the requirements of Mach's principle: $\hat{\phi}$ is given by the matter distribution in space.

In 1968, P. Bergmann [68], and 1970 R. Wagoner [69], discussed a more general scalar–tensor theory which possesses an additive cosmological function term $\Lambda(\hat{\phi})$ in the Lagrangian. Furthermore, the latter may now possess a functional parameter $\omega = \omega(\hat{\phi})$ for a scalar field $\hat{\phi}$. This general kind of theory, now often called *Bergmann–Wagoner* (BW) class of STTs, possesses the Jordan–Brans–Dicke (JBD) class as a special case for $\omega = const$ and $\Lambda(\hat{\phi}) = 0$.

The Bergmann–Wagoner-formed models are not canonical, and in physics a theory is said to be in a canonical form if it is written in the paradigmatic form taken from the classical one (an ideal which is, in principle, freely eligible and a matter of definition).[12] The Equation (3.3) (considered to be in the *Jordan frame*) is not in this form. However, STTs can be transformed conformally into a canonical form (*Einstein frame*) in which a cosmological function still appears, but $\hat{\phi}$ is minimally coupled. This is achieved by changing from the so-called Jordan frame (with mixed

[11] Work which resulted in 1984 in the Nobel prize for physics for A.A. Penzias and R.W. Wilson. Further, the exact analysis and corroboration of the qualities of CMB, together with the small anisotropy present, led to the Nobel prize award to John C. Mather and George F. Smoot in 2006.

[12] Cf. an interesting analysis about the historical origin and meaning of the concept of "canon" by J. Assmann in [70], which he further relates to Halbwachs' "mémoire volontaire" of a society.

degrees of freedom of metric and scalar field) to the Einsteinian one (with unmixed degrees of freedom). In the four-dimensional case, this is fulfilled through $g_{\mu\nu} \rightarrow \hat{\phi}^{-1} g_{\mu\nu}$ and a redefinition of the scalar field and cosmological function. It is still the subject of discussion which frame is best. The Jordan one, however, is usually called the physical frame [71].

The scalar field in the Jordan–Brans–Dicke theory is massless. However, a generally covariant theory of gravitation can accommodate a massive scalar field in addition to the massless tensor field [72, 73]. Thus, a version of the JBD or BW theory with massive scalar fields may be postulated [74]; indeed, A. Zee incorporated the first concept of SSB to gravity within a STT [75], suggesting that the same symmetry-breaking mechanism was responsible for breaking a unified gauge theory into strong, weak and electromagnetic interactions (mediated by their corresponding gauge bosons). Spontaneous symmetry breaking (SSB) causes some scalar field to have a vacuum expectation value v, thus generating the mass of the intermediate bosons and of fermions, relating them to the ground state of the scalar field after the breakdown of symmetry. Zee attributed the smallness of Newton's gravitational constant G_{N} (of order of magnitude of about $10^{-11}\ \mathrm{m}^3\,\mathrm{kg}^{-1}\,\mathrm{s}^{-2}$) to the massiveness of some particle (this may be compared with the result of [53]) with

$$G_{\mathrm{N}} \sim 1/v^2 \ , \tag{3.4}$$

where $v^2 = \sqrt{2}/(8\pi G_{\mathrm{F}}) \approx 6.07 \times 10^4\ (\mathrm{GeV})^2$. Thus, SSB generates the mass of the intermediate boson such that for the Fermi constant ($G_{\mathrm{F}} \approx 1/(2\pi(294\ \mathrm{GeV})^2)$) (with weakon mass M_{W} ($\sim 80\ \mathrm{GeV/c^2}$) and elementary charge e ($\sim 10^{-19}$ C)) we have

$$G_{\mathrm{F}} \sim e^2/M_{\mathrm{W}}^2 \sim 1/v^2 \ , \tag{3.5}$$

which may be compared with (3.4). Since SSB has proven extraordinarily faithful in many areas of physics, Zee considered it worthwile to incorporate this mechanism into gravitation [75] and explain the smallness of Newtons's constant through the mass of Higgs particles.[13] This mechanism includes a self-interaction of the scalar field and, thus, a potential V as part of the cosmological function of the BW class of STTs, but which is missing from usual JBD theories.

As a result of the missing potential, Brans–Dicke's theory is inconsistent with observation unless a certain parameter is very large [76]. In fact, from measurements of radio-signal current time delay with Viking probes from Mars, the coupling parameter ω of the usual JBD theory in (3.3) (a measure of the strength of the scalar field coupling to matter) is required to being greater than about 500 [77].

In any sensitive theory, as Brans and Dicke proved in their original work [48], the dimensionless constant ω must be of the general order of unity. For $\omega \rightarrow \infty$, however, GR is obtained, which entails that the JBD theory leads nearly to the same

[13] The smallness of G can also, as will be seen, be explained through a high expectation value of the scalar field as well as through a strong coupling of the scalar field to gravitation, analogously to [54].

results as GR. In contrast, however, in a STT with the scalar field anchored by the SSB potential, this strength of the scalar field coupling may naturally be smaller. Thus, new physics of a higher order is possible. Given the gravitational properties of Higgs-like fields (see Section 3.1.1), for instance, it seems natural to couple them to gravitation and analyze new properties of the model. This kind of Higgs-like field (because of the coupling to SSB and the possession of a nontrivial vacuum state) may then be relevant in view of a gravitational theory which might entail long-range changes in the dynamics to explain dark components, anchored or not with elementary particle physics. Indeed, the simplest "Higgs-field model" beyond the standard model consists in the addition of a singlet particle that only interact with the Higgs sector of the SM, in which the sector does not couple directly to vector bosons. With a fundamental gauge-invariant construction block $\phi^\dagger\phi$, the simplest coupling of a particle to a Higgs or Higgs-like field is [78]

$$\text{Lagrangian term of Higgs sector} = \tilde{\lambda} X \phi^\dagger \phi \ , \tag{3.6}$$

where X is a scalar field and † represents the Hermitean conjugation, the transposition of a tensor for real-valued components, and complex conjugation for purely scalar quantities.

The Higgs-like field develops a vacuum expectation value and, after shifting it, the vertex (3.6) leads to a mixing between the scalar field and the Higgs-like field. Thus, it may give rise to new effects that do not involve the scalar explicitly [78]. Furthermore, the X-field may be considered as not fundamental, but an effective description of an underlying dynamical mechanism, and a relation between gravity and the generalized Higgs sector may be assumed. Both gravity and a Higgs particle possess some universal characteristics; gravity couples universally to the energy–momentum tensor and the Higgs particle to mass, which corresponds to the trace of the energy–momentum tensor. This suggest a relation between the generalized Higgs sector and gravity be conjectured, which is indeed given by Higgs gravity in [54]. Furthermore, there is a similarity between X and the hypothetical graviton, since both are singlets under the gauge group [57].

Because they have no coupling to ordinary matter, singlet fields are not well constrained by experiments. Typically, one can argue that they are absent from the theory because they can have a bare mass term which can be made to be of the order of the Planck mass M_P, making these fields invisible. However, one can take the attitude that the Planck length be not a fundamental constant but rather a property of today's state of the world, which evolve in time and be typically given by a vacuum expectation value of some scalar field [28]. With a Higgs coupling to gravity, then, all masses, including the Planck mass, should be given by SSB. In this case there is a hierarchy of mass scales $M_\text{P} \gg v$. Given these similarities, X can be considered to be essentially the graviton and be identified as constant $\cdot$ R, with the curvature scalar R [57]. Moreover, this possibility may be used to explain the naturalness problem, especially since other candidates such as top-quark condensation or technicolor (in which quarks are no longer primordial) have not functioned so far and supersymmetry doubles the spectrum of elementary particles, replacing

Bose (Fermi) degrees of freedom with Fermi (Bose) degrees of freedom, whereas all supersymmetric particles are by now beyond physical reality.

Making a low-energy expansion ([57]) and ignoring higher derivative terms, a spontaneous symmetry breaking theory of gravity with a Higgs field as the origin of the Planck mass may be derived [57, 79]. Moreover, this is the theory which was first derived in [80] and [81]. The remnant of originally very strong interactions is the parameter $\breve{\alpha}$, which in Section 3.2.1 will be introduced as the coupling strength of the Higgs field to gravitation. It will essentially give Newton's gravitational constant, and its high value will enable the model to be distinguishable to gravity at low energy scales, other than the case within usual JBD-theories.

The class of STTs with massive scalar fields is given within the Bergmann–Wagoner (BW) class with the Lagrangian[14]

$$\mathcal{L}_{\mathrm{BW}} = \frac{1}{16\pi}\left\{\hat{\phi}R + \frac{\omega(\hat{\phi})}{\hat{\phi}}\hat{\phi}_{,\lambda}\hat{\phi}^{,\lambda} - 2\hat{\phi}\,U(\hat{\phi}) + \mathcal{L}_{\mathrm{M}}\right\}\sqrt{-g}\,, \tag{3.7}$$

whereas $S = \int \mathcal{L}d^4x$ and $\delta S \equiv 0$ are valid. Further, $\hat{\phi}\,U(\hat{\phi}) = \tilde{\Lambda}(\hat{\phi})$ gives a cosmological function and $_{,\lambda}$ the derivative in respect to the λ-coordinate.

Within the Bergmann–Wagoner class, there is a wide account of analyses, although most of them focus on $U(\hat{\phi}) = 0$ as special case. However, analyses within the general BW class such as on the existence of black holes as well as global properties of vacuum, static, spherically symmetric configurations can be found, for instance, in [82–84], and in deSitter and warm inflation models in the framework of STTs in [85], and with the Higgs potential in [86, 87]. Friedmann–Lemaître–Robertson–Walker (FLRW or simply RW) models for Friedmann–Lamaître universes for cosmology, further, are analyzed in [88], obtaining a class of separable Wheeler–deWitt equations after a quantization of the models. That is, equations which a wave function of the universe should satisfy in a theory of quantum gravity.

3.2 Scalar-Tensor Theory with Higgs Potential

3.2.1 Lagrange Density and Models

Let us take a closer look at a Bergmann–Wagoner (BW) model with a generally nonvanishing cosmological function. Then let the scalar field be defined through a U(N) isovector which is a scalar field also, with

$$\hat{\phi} = \breve{\alpha}\phi^\dagger\phi \quad \text{and the definition} \quad \omega = \frac{2\pi}{\breve{\alpha}} = \text{constant}\,, \tag{3.8}$$

[14] For purposes of completeness, the BW class can be given in an even more general form for D dimensions and with a nonminimal coupling $f(\phi)R$ (see [82]).

with the gravitational strength $\breve{\alpha}$ (as remnant of strong interactions [56]), and the cosmological function of the BW class given by

$$U(\hat{\phi}) = U(\phi^\dagger\phi) = \frac{1}{\breve{\alpha}\phi^\dagger\phi}\left[8\pi V^*(\phi^\dagger\phi)\right] , \tag{3.9}$$

whereas $V^*(\phi) \equiv V^*(\phi^\dagger\phi)$ be the potential of the scalar field.

As can be easily seen, such a model does not possess a dimension-loaded coupling constant like G, which is the main problem for renormalizing Einstein's theory. Through (3.19), G will be replaced with the reciprocate dimensionless constant $\breve{\alpha}$ multiplied by $\phi^\dagger\phi$. Thus, the dimension problem for renormalization disappears.

The scalar field will couple non-minimally with the Ricci curvature scalar R with the gravitational strength $\breve{\alpha}$. In this way, we can give the Lagrangian of a scalar–tensor theory in Jordan frame of the form

$$\mathcal{L} = \left[\frac{1}{16\pi}\breve{\alpha}\phi^\dagger\phi R + \frac{1}{2}\phi^\dagger_{;\mu}\phi^{;\mu} - V^*(\phi) - \mathcal{L}_\mathrm{M}\right]\sqrt{-g} , \tag{3.10}$$

whereas $\hbar = 1$ and $c = 1$ are set, and $_{;\mu}$ means the covariant derivative with respect to all gauge groups.[15] The subscript $_{,\mu}$ represents the usual derivative (see discussion in relation with the Lagrangian (3.7)). The Lagrangian (3.10) postulates possible gravitational interactions not only mediated by massless spin-2 excitations as is postulated on one hand in usual GR, but also takes into account gravitational interactions of massive scalar fields. Further, let the potential $V^*(\phi)$ of the scalar field be of the form of that of the Higgs field of elementary particle physics, that is a ϕ^4-potential with

$$V(\phi) = \frac{\lambda}{24}\left(\phi^\dagger\phi + 6\frac{\mu^2}{\lambda}\right)^2 = \frac{\mu^2}{2}\phi^\dagger\phi + \frac{\lambda}{24}\left(\phi^\dagger\phi\right)^2 + \frac{3}{2}\frac{\mu^4}{\lambda} . \tag{3.11}$$

The potential in (3.11) possesses an additive factor $3/2\mu^4/\lambda$ which does not usually appear in the standard theory. This factor lowers the minimum so that the energy density for vanishing scalar field excitations is defined as zero with $V(\phi_0^\dagger\phi_0) = 0$ for the ground state ϕ_0 of the scalar field. The additive term is thus related to the election of a vanishing formal cosmological constant which, however, can be inserted in the theory by adding a constant term

$$V_0 = -\frac{3\breve{\alpha}\mu^2}{4\pi\lambda}\Lambda_0 \tag{3.12}$$

with Λ_0 as a true cosmological constant and with a total potential of the form

$$V^*(\phi) = V(\phi) + V_0 \tag{3.13}$$

with a cosmological function $\Lambda(\phi)$ dependent on this generalized Higgs potential, as will be seen in Section 3.2.2. The cosmological constant Λ_0 is often expected to

[15] As explained before, the connection coefficients $\Gamma^\mu_{\nu\lambda}$ of the affine connection are now introduced. This connection is a rule which describes how to legitimately move a vector along a curve on the manifold without changing its direction. $\Gamma^\mu_{\nu\alpha}$ are the so-called Christoffel symbols, which give the Riemann tensor $R^\mu_{\nu\lambda\sigma}$ as introduced in footnote 4.

be vanishing for physical economy. However, together with quintessence in general, it is related to our understanding of the nature of gravity. It might indeed be a low-energy appearance coming from primary gravitation in the early universe, as proposed in [89], but also related to dynamical quintessential fields. Nevertheless, the constant part of the cosmological function coming from the Higgs potential (3.13) (i.e. the true cosmological constant) will further be taken as vanishing and, if written, then only for purposes of completeness.

In (3.10), $\mathcal{L}_\mathrm{M}$ is the Lagrange density of the fermionic and massless bosonic fields,

$$\mathcal{L}_\mathrm{M} = \frac{i}{2}\bar{\psi}\gamma^{\mu}_{\mathrm{L,R}}\psi_{;\mu} + h.c. - \frac{1}{16\pi}F^{a}_{\mu\nu}F^{\mu\nu}_{a} - (1-\hat{q})k\bar{\psi}_{\mathrm{R}}\breve{\phi}^{\dagger}\hat{x}\psi_{\mathrm{L}} + h.c.\,, \tag{3.14}$$

while ψ in (3.14) are the fermionic fields, and

$$\begin{aligned} F^{a}_{\mu\nu} = \frac{1}{ig}\left[D^{a}_{\mu}, D^{a}_{\nu}\right] &= A^{a}_{\nu,\mu} - A^{a}_{\mu,\nu} + ig\left[A^{a}_{\mu}, A^{a}_{\nu}\right] \\ &= A^{a}_{[\nu,\mu]} + ig\left[A^{a}_{\mu}, A^{a}_{\mu}\right] \end{aligned}$$

is the field-strength tensor for the gauge potentials A^a_μ. It is defined by the commutator of the covariant derivative D^a_μ, analogous to electrodynamics for the electric and magnetic strengths $\boldsymbol{E}$ and $\boldsymbol{B}$. The exact form of covariant derivatives, that is of the potentials, however, depends on the chirality and form of the actual fermionic field. Chirality is important since parity violation, and thus a different coupling dependent on chirality, is an experimental fact, characterizing the weak interactions and β-decay [90, 91]. For the electroweak interactions, left-handed wave functions are thus described by (iso-)doublets, while right-handed ones are described by (iso-)singlets.

Within electrodynamics, the homogeneous Maxwell equations are derived using Jacobi identities with covariant derivatives (Bianchi identities). The inhomogeneous ones depend on the Lagrangian and thus on the exact system (and thus on the environment, as reflected in the appearance of magnetization $\boldsymbol{M}$ and polarization $\boldsymbol{P}$ in the field equations). The more general equations of Yang–Mills' theories, for the dynamics of $F_{\mu\nu}$ and isovectorial ψ, are derived analogously. However, unlike within QED, the commutator $[A_\mu, A_\nu] \equiv A_\mu A_\nu - A_\nu A_\mu$ is not vanishing. It presents self-interactions of the gauge potentials. Through them, in QCD, for instance, gluons interact with each other, while such interactions vanish within QED given the Abelian (commutative) character of the symmetry group U(1). Photons as gauge bosons in QED, do not self-interact.

In (3.14), $\hat{x}$ give the Yukawa coupling operator, k be a constant factor, and the subscripts R and L refer to the right- and left- handed fermionic states of ψ. The index a be the iso-spin index, which counts the N elements of the multiplet ψ, given by the particles which are indistinguishable for an interaction. Further, let us take $\breve{\phi} = \phi$ in the following; this means the same scalar field coupled with R and matter for the case $\hat{q} \neq 1$.

Equation (3.14), together with (3.10) leads to the field equations as derived first in [57, 80, 81]. The model parting from (3.10) and (3.14) does not possess bare gravitational vertices as an Einsteinian quantum theory would. This lack of only gravitational vertices should further exclude outer gravitational lines (as long as no primordial gravitational constant is assumed) [92]. This theory is renormalizable [80, 92] according to deWitt's power-counting criterion.

The parameter $\hat{q}$ is defined to give the fermionic coupling with the scalar field. It will represent the fact that, in the case of a coupling of the scalar field ϕ to the fermionic fields, the source of this Higgs field is canceled by the fermionic term identically after symmetry breaking (see Section 3.2.3 and especially (3.35) and [45, 81]). If the model is, though, not only for astronomical considerations resulting from changes in the standard dynamics of GR, and a further unification is considered, the scalar field may be coupled to $\mathcal{L}_\mathrm{M}$ and assumed to be the Higgs field coupled to the curvature scalar in (3.10). Then the model is the one first presented in [81] in which $\hat{q}$ is chosen as 0. This model corresponds to a coupling with the SM, with mass production via the Higgs mechanism but with sourceless, only gravitationally coupled stable Higgs particles as a type of self-interacting dark matter in the sense of [93] and [94]. The scalar field would be assumed to be an analog to the standard model Higgs field. Further, for $\hat{q} = 1$ (not assuming another scalar field here) [80], the Higgs-like field does not produce the mass of elementary particles, and may be coupled to a GUT (*Grand Unified Theory*) or quintessence model, with an almost only gravitationally coupled scalar field that, for a very small mass, may contribute to the phenomenon of dark matter as shown in [39, 95]. Such particles are still not detectable through current fifth-force experiments like [96] and lead to a quintessential behavior, canceling the appearance of an horizon for central symmetry and very high scalar field length scales [40].

3.2.2
The Field Equations

Using the Hamilton Principle of Least Action and the Euler–Lagrange equations for relativistic fields,

$$\left(\frac{\partial \mathcal{L}}{\partial \psi_{,\mu}}\right)_{,\mu} - \frac{\partial \mathcal{L}}{\partial \psi} = 0 \quad \text{(and Hermitean conjugation of the same)}\,, \tag{3.15}$$

one achieves generalized Einstein field equations and a Higgs-like field equation with a coupling of the scalar field ϕ to the curvature scalar R and the symmetric metrical energy–momentum tensor $T_{\mu\nu}$:[16)]

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda^*(\phi) g_{\mu\nu} &= -\frac{8\pi}{\breve{\alpha}\phi^\dagger\phi} T_{\mu\nu} - \frac{8\pi}{\breve{\alpha}\phi^\dagger\phi}\left[\phi^\dagger_{(;\mu}\phi_{;\nu)} - \frac{1}{2}\phi^\dagger_{;\lambda}\phi^{;\lambda} g_{\mu\nu}\right] - \\ &- \frac{1}{\phi^\dagger\phi}\left[(\phi^\dagger\phi)_{,\mu;\nu} - \left(\phi^\dagger\phi\right)^{,\beta}{}_{;\beta}\, g_{\mu\nu}\right], \end{aligned} \tag{3.16}$$

$$\phi^{;\mu}_{;\mu} - \frac{\breve{\alpha}}{8\pi}\phi R + \frac{\delta V^*(\phi)}{\delta \phi^\dagger} = 2\frac{\delta \mathcal{L}_\mathrm{M}}{\delta \phi^\dagger}, \quad \text{with} \quad \frac{\delta V(\phi)}{\delta \phi^\dagger} = \mu^2 \phi + \frac{\lambda}{6}(\phi^\dagger\phi)\phi\,. \tag{3.17}$$

[16)] $_{(\ldots)}$ are the antisymmetric Bach parenthesis given by $A_{(i}B_{k)} = 1/2(A_i B_k + A_k B_i)$.

The term on the right-hand side of the Higgs-like field equation (3.17) is the source of the Higgs-like field with

$$2\frac{\delta \mathcal{L}_{\mathrm{M}}}{\delta \phi^{\dagger}} = 2\left(\frac{\delta \mathcal{L}_{\mathrm{M}}}{\delta \phi}\right)^{\dagger} = -2k(1-\hat{q})\bar{\psi}_{R}\hat{x}\psi_{\mathrm{L}} \,. \tag{3.18}$$

Equation (3.18) depends on the fermionic Lagrangian and thus on $\hat{q}$.

We may define in (3.16) a gravitational coupling term

$$G(\phi) = \frac{1}{\breve{\alpha}\phi^{\dagger}\phi} \,, \tag{3.19}$$

in analogy to GR (see (3.1)), whereas $G(\phi)$ here is a field quantity and thus local. It is dependent on the scalar field ϕ and the gravitational strength $\breve{\alpha}$. Analogously, a general cosmological function was defined in (3.16) as

$$\Lambda^{*}(\phi) := \frac{8\pi}{\breve{\alpha}\phi^{\dagger}\phi}V^{*}(\phi) = 8\pi G(\phi)V(\phi) - \frac{6\mu^{2}}{\lambda}\frac{\Lambda_{0}}{\phi^{\dagger}\phi} \,, \tag{3.20}$$

mainly given by the potential of the scalar field and its excitations (as in (3.23) and (3.22)), and related to the cosmological function term $\breve{\alpha}\phi^{\dagger}\phi U(\phi)$ of the BW class of STTs. The field equations for the fermionic fields and the bosonic Yang–Mills fields are neglected.

The Ricci curvature scalar R in the field equations of gravity and of the scalar field is coupled to the scalar field itself. $R = g^{\mu\nu}R_{\mu\nu} \equiv \sum_{\mu,\nu=0}^{3} g^{\mu\nu}R_{\mu\nu}$ can be derived from (3.16), with the form

$$R = \frac{8\pi}{\breve{\alpha}\phi^{\dagger}\phi}\left[T + 4V^{*}(\phi) - \phi^{\dagger}_{;\beta}\phi^{;\beta}\right] - \frac{3}{\phi^{\dagger}\phi}(\phi^{\dagger}\phi)^{;\beta}_{;\beta} \,, \tag{3.21}$$

whereas $V^{*}(\phi) = V(\phi) + V_{0}$ is valid from (3.13).

3.2.3 Field Equations After Symmetry Breakdown

In the spontaneously broken phase of symmetry, developing the scalar field ϕ around its ground state v,

$$\phi_{a} = vN_{a} + \phi'_{a} \,, \tag{3.22}$$

the ground-state value of the scalar field is given by

$$\phi^{\dagger}_{0}\phi_{0} = v^{2} = -\frac{6\mu^{2}}{\lambda} \,, \tag{3.23}$$

with v real-valued and $\mu^{2} < 0$. This can further be resolved in general as $\phi_{0} = vN$ with $N = \mathit{constant}$, satisfying $N^{\dagger}N = 1$, with

$$\phi = \varrho UN = \frac{\varrho}{v}U\phi_{0} \,. \tag{3.24}$$

For the ground state ϕ_0, the potential vanishes with the election of no further additive factor Λ_0 of the cosmological function, following (3.11) through (3.13):

$$V(\phi_0) = u_0 \equiv \frac{1}{8\pi G(\phi_0)} \Lambda_0 \ . \tag{3.25}$$

This is the energy density of the ground state of the scalar field, which is $\check{V}^*(\phi_0) = -3/2(\mu^4/\lambda) + (1/(8\pi G(\phi_0)))\Lambda_0$ if the last factor of (3.11) is not taken. It would lead to a formal cosmological constant added to the cosmological constant itself, which we want to avoid.

After symmetry breaking, two particles appear: a massless particle, called a Goldstone, and a massive particle usually called a Higgs ([45]). The first of these particles can be "gauged away" through the so-called unitary gauge [39, 54]. The scalar field ϕ can be written in terms of the real-valued excited Higgs-like scalar field ξ (a real-valued scalar variable) in the following form:

$$\phi = \varrho N = v\sqrt{1+\xi}N \quad \text{with} \quad \xi = \frac{\phi^\dagger \phi}{v^2} - 1 \ . \tag{3.26}$$

The Higgs-like field equation then yields, using (3.21):

$$\begin{aligned} \xi^{,\mu}_{;\mu} + \frac{4\pi/(9\breve{\alpha})\lambda v^2}{1+4\pi/(3\breve{\alpha})}\xi = {} & \frac{1}{1+4\pi/(3\breve{\alpha})} \cdot \frac{8\pi}{3\breve{\alpha}v^2}\left[\hat{T} - \sqrt{\xi+1}\bar{\psi}\hat{m}\psi\right] \\ & + \frac{4}{3}\Lambda_0\left(1+\frac{4\pi}{3\breve{\alpha}}\right)^{-1} , \end{aligned} \tag{3.27}$$

with the effective energy–momentum tensor $\hat{T}_{\mu\nu}$ (analogous to the SM, [87]) with the trace [81]

$$\hat{T} = \frac{i}{2}\bar{\psi}\gamma^\mu_{\mathrm{L,R}}\psi_{;\mu} + h.c. = \sqrt{1+\xi}\bar{\psi}\hat{m}\psi \ , \tag{3.28}$$

with the fermionic mass matrix (cf. [101])

$$\hat{m} = \frac{1}{2}kv\left(N^\dagger \hat{x} + \hat{x}^\dagger N\right) \ . \tag{3.29}$$

Now, insertion of (3.28) into (3.27) leads to a vanishing of the right-hand side of (3.27) in the case $\hat{q} = 0$ and $\Lambda_0 = 0$. Thus, the coupling (given by $\mathcal{L}_{\mathrm{M}}$ in (3.14)) of these Higgs particles to their source is only weak (this means proportional to G) if this Higgs field does not couple in the fermionic Lagrangian [39, 80], or completely vanishing [45, 81] in the case of a coupling of this Higgs field to SM (with $\Lambda_0 \equiv 0$). According to [56], a wave function renormalization of the scalar field results in the effective coupling of this Higgs field to matter becoming of gravitational strength $O(M/M_{\mathrm{P}})$. Because of this, the Higgs becomes essentially a stable particle, which may have some cosmological consequences. These, however, would depend on the length scale of the scalar field. In particular, the scalar field particles should effectively decouple for large values of λ, that is for a small mass M (see (3.31)). Meanwhile, $\breve{\alpha}$, as the remnant of an original strong interaction, would be the essential cause for the gravitational coupling G being so small (see (3.1.2)). Particularly in the case $\hat{q} = 0$, the scalar field possesses qualities as in [93] as a candidate of self-interacting DM, and in this way might be related to works like [94].

In (3.27), which is a Yukawa equation, a gravitational coupling constant

$$G_0 = \frac{1}{\breve{\alpha} v^2} = -\frac{1}{\breve{\alpha}} \frac{\lambda}{6\mu^2} = G(v) \tag{3.30}$$

may be defined (see (3.42)). Further, the (Compton-)length scale of the scalar field, using (3.30), is given ($\hbar$ and c are inserted only for the definition of the scale factor and the scalar field mass) by

$$l = \left[\frac{1 + 4\pi/(3\breve{\alpha})}{16\pi G_0(\mu^4/\lambda)}\right]^{1/2} = M^{-1}\left(\cdot\frac{\hbar}{c}\right) \tag{3.31}$$

with the scalar field mass M, which in the SM is only given by $\sqrt{|2\mu|^2}$. The latter equation can also be written as

$$\begin{aligned} M^2 &= -\frac{8\pi}{3}\frac{\mu^2}{\breve{\alpha}}\left(1 + \frac{4\pi}{3\breve{\alpha}}\right)^{-1}\left(\cdot\frac{c}{\hbar}\right)^2 \\ &= \frac{4\pi}{9\,\breve{\alpha}}\lambda\, v^2\left(1 + \frac{4\pi}{3\breve{\alpha}}\right)^{-1}\left(\cdot\frac{c}{\hbar}\right)^2 . \end{aligned} \tag{3.32}$$

It is dependent on the reciprocal gravitational coupling strength $\breve{\alpha}^{-1}$. Thus, the squared mass of the scalar field depends essentially on the gravitational coupling strength G_0, which is very weak. Thus, the Compton length given by $l = M^{-1}$ may at this point be expected to be high-valued. This would mean, within the SM, a very small value of $|\mu|$. The constraints of a Higgs-field mass, though, may change here in relation to those in the standard theory. For $\hat{q} = 0$, for instance, the coupling constants g and the ground state (vacuum expectation) value are indirectly known from high-energy experiments. A comparison between current–current coupling within Fermi's theory, low-energetic limits of W^{+} couplings and the weakon mass M_{W}, v can be written dependent on Fermi's constant G_{F} and experimentally determined as $v^2 \approx 6 \times 10^4$ (GeV)^2. Here, though, nonvanishing values of v are possible for small masses M, which may be small-valued without the necessity of small $|\mu|$ values.

The gravitational coupling of (3.30) depends on v^{-2}. Insertion of this and eventually also the length scale (3.31), leads to the Higgs potential in the form

$$\begin{aligned} V^*(\xi) &= \frac{3}{2}\frac{\mu^4}{\lambda}\xi^2 + (8\pi G_0)^{-1}\Lambda_0 = -\frac{1}{4}\mu^2 v^2 \xi^2 + (8\pi G_0)^{-1}\Lambda_0 \\ &= \frac{\lambda\, v^4}{24}\xi^2 + (8\pi G_0)^{-1}\Lambda_0 = \frac{3}{32\, l^2}\frac{\xi^2}{\pi\, G_0}\left(1 + \frac{4\pi}{3\breve{\alpha}}\right) + (8\pi G_0)^{-1}\Lambda_0 . \end{aligned} \tag{3.33}$$

The Higgs-like field equation yields with l,

$$\xi^{,\mu}_{;\mu} + \frac{\xi}{l^2} = \frac{1}{1 + 4\pi/(3\breve{\alpha})} \cdot \frac{8\pi G_0}{3}\left[\hat{T} - \sqrt{\xi + 1}\bar{\psi}\hat{m}\psi\right]\left(+\frac{4}{3}\frac{1}{1 + 4\pi/(3\breve{\alpha})}\Lambda_0\right) . \tag{3.34}$$

In the case $\hat{q} = 0$, the gauge-boson matter terms cancel and the scalar field equation becomes a homogeneous Klein–Gordon equation for $\Lambda_0 \equiv 0$.

Now, let us rewrite (3.34) in the form for $\breve{\alpha} \gg 1$,

$$\xi^{;\mu}_{;\mu} + \frac{1}{l^2}\xi = \frac{8\pi G_0}{3}\hat{q}\hat{T} + \frac{4}{3}\Lambda_0, \quad (\breve{\alpha} \gg 1)\,, \tag{3.35}$$

The vanishing property of its source is easily seen. Further, the trace of the symmetric energy-momentum tensor $T_{\mu\nu}$ belonging to $\mathcal{L}_M\sqrt{-g}$ in the Lagrangian given by (3.10) and (3.14) satisfies the conservation law

$$\hat{T}^{\nu}_{\mu;\nu} = (1-\hat{q})\frac{1}{2}\xi_{,\mu}(1+\xi)^{-1}\,\hat{T}\,. \tag{3.36}$$

In the case when ϕ does not couple to the fermionic state ψ in $\mathcal{L}_M\sqrt{-g}$, then (3.36) does not possess a source and for the SM, the above equation means the production of the fermionic mass through this Higgs field. This leads to a breaking of the conservation law through a new “Higgs force”.

The dimensionless parameter $\breve{\alpha}$ in (3.10) may, further, be defined in terms of the ratio

$$\breve{\alpha} \simeq (M_P/M_B)^2 \gg 1\,, \tag{3.37}$$

where M_P and M_B are the Planck mass and the mass of the gauge boson, respectively. The mass of the gauge boson is given by

$$M_B \simeq \sqrt{\pi}\breve{g}v\,, \tag{3.38}$$

where $\breve{g}$ is the coupling constant of the corresponding gauge group.

Through similitude with the standard theory, an effective gravitational coupling (as screened gravitational strength) may be given by

$$G_{\text{eff}} = G(\xi) = (1+\xi)^{-1}G_0\,. \tag{3.39}$$

The latter reduces to (3.30) in the absence of a Higgs-like scalar field excitation ξ (that is for $\xi = 0$ with the chosen form of Higgs excitations), and becomes singular for a vanishing Higgs-like scalar field with $\xi = -1$ [39, 40].

Also, the generalized Einstein field equations, which reduce to the usual GR ones for vanishing excitations ξ, are now given by (3.16) in the form[17]

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda^*(\xi)g_{\mu\nu} = -8\pi\, G_{\text{eff}}\hat{T}_{\mu\nu} - \frac{\pi}{\breve{\alpha}}\frac{1}{(1+\xi)^2}\Big[2\,\xi_{,\mu}\xi_{,\nu} - \\ - \xi_{,\lambda}\,\xi^{,\lambda}\,g_{\mu\nu}\Big] - \frac{1}{1+\xi}\Big[\xi_{,\mu;\nu} - \xi^{,\lambda}{}_{;\lambda}g_{\mu\nu}\Big]\,, \end{aligned} \tag{3.40}$$

[17] These may be compared with the field equations in [41, 104] within the BW class with a rescaled potential. The Newtonian approximation of it leads to essentially the same equations as here.

whereas the cosmological function is given by

$$\Lambda^*(\xi) = \frac{8\pi G_0}{1+\xi} V(\xi) + \frac{\Lambda_0}{1+\xi} = \frac{12\pi}{\breve{\alpha} v^2}\frac{\mu^4}{\lambda}\frac{\xi^2}{1+\xi} + \frac{\Lambda_0}{1+\xi} . \tag{3.41}$$

It is clear that, for the special case of vanishing scalar field excitations ξ, (3.40) results in the usual Einstein field equations

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda_0 g_{\mu\nu} = -\kappa T_{\mu\nu} . \tag{3.42}$$

As a result of (3.30), the gravitational coupling strength given by $\breve{\alpha}$ is very high, so that the second term on the right-hand side of (3.40), $\pi/\breve{\alpha}$-proportional, can be neglected (see [101]), due to the smallness of the term $4\pi/(3\breve{\alpha})$.

Equation (3.40) can be rewritten for $\breve{\alpha} \gg 1$. This and the insertion of the Higgs-like field equation (3.27) into the Einstein field equations leads to

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \frac{1}{l^2}(1+\xi)^{-1}\xi\left(1+\frac{3}{4}\xi\right) g_{\mu\nu} - \frac{1}{3}(1+\xi)^{-1}\Lambda_0 g_{\mu\nu} \\ = -8\pi G_{\text{eff}}\left(\hat{T}_{\mu\nu} - \frac{\hat{q}}{3}\hat{T} g_{\mu\nu}\right) - (1+\xi)^{-1}\xi_{,\mu;\nu} , \quad (\breve{\alpha} \gg 1) . \end{aligned} \tag{3.43}$$

The cosmological function $\Lambda(\phi)$ after symmetry breaking (3.41) is essentially quadratic in ξ. For $\breve{\alpha} \gg 1$, it yields

$$\Lambda^*(\xi) = \frac{3}{4 l^2}\frac{1}{1+\xi}\xi^2 + \frac{\Lambda_0}{1+\xi} . \tag{3.44}$$

Hence, with

$$\xi = \frac{G(v) - G_{\text{eff}}}{G_{\text{eff}}} \tag{3.45}$$

it can be written in the form

$$\Lambda^*(\xi) = \frac{3}{4 l^2}\left(\frac{G(v)^2 + G_{\text{eff}}^2}{G(v) G_{\text{eff}}} - 2\right) + \frac{G_{\text{eff}}}{G(v)}\Lambda_0 . \tag{3.46}$$

The trace of (3.43) leads to

$$R = \frac{3}{l^2}\xi + 8\pi G_{\text{eff}}(1-\hat{q})\hat{T} = \frac{3}{l^2}\left(\frac{G(v)}{G_{\text{eff}}} - 1\right) + 8\pi G_{\text{eff}}(1-\hat{q})\hat{T} . \tag{3.47}$$

R is independent of Λ_0, since it appears in the Higgs-like field equation (3.35) and in the Einstein field equations (3.40). The trace over Einstein's field equations, using the Higgs-like field equation, leads to a cancelation of the Λ_0-term.

Using (3.47), (3.43) can be rewritten in the form

$$\begin{aligned} R_{\mu\nu} - \frac{1}{2l^2}\left[\frac{1+3/2\xi}{1+\xi}\right]\xi g_{\mu\nu} - \frac{1}{3}(1+\xi)^{-1}\Lambda_0 g_{\mu\nu} = \\ = -8\pi G_{\text{eff}}\left[\hat{T}_{\mu\nu} - \frac{1}{2}\left(1 - \frac{1}{3}\hat{q}\right)\hat{T} g_{\mu\nu}\right] - (1+\xi)^{-1}\xi_{,\mu;\nu} , \quad (\breve{\alpha} \gg 1) , \end{aligned} \tag{3.48}$$

with

$$\frac{1+3/2\xi}{1+\xi} = \frac{1}{2}\left(3 - \frac{G_{\text{eff}}}{G_0}\right) .$$

Obviously, for vacuum and for $\hat{q} = 0$ in general, the Ricci scalar is given by the scalar field only. The matter term of (3.47) leads, however, to a different right-hand side in the square bracket of the Einsteinian field equations where the Ricci curvature has been inserted.

3.2.4 Outlook

To place the above in context, let us take a look at some important and relatively simple symmetries analyzed in physics. An important realization of the field equations is that with spherical (central) symmetry. It is used for many realizations such as galaxies and the solar system, and its line element (which gives the metric through $\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu$) is given by

$$\mathrm{d}s^2 = e^\nu(\mathrm{d}ct)^2 - e^\lambda\,\mathrm{d}r^2 - r^2\,\mathrm{d}\Omega^2 , \tag{3.49}$$

with ν and λ as functions of the r and t coordinates only, and $\mathrm{d}\Omega^2 = (\mathrm{d}\vartheta^2 + \sin^2\vartheta\ \mathrm{d}\varphi^2)$ as the metric of a two-dimensional unit sphere.

Another main realization of symmetry is given by the Friedmann–Lemaître–Robertson–Walker (RW) metric, used for general cosmology and cosmic evolution:

$$\mathrm{d}s^2 = (c\,\mathrm{d}t)^2 - a(t)^2\left[\mathrm{d}\chi^2 + f(\chi)^2\left(\mathrm{d}\vartheta^2 + \sin^2\vartheta\,\mathrm{d}\varphi^2\right)\right] . \tag{3.50}$$

Here, χ is the covariant distance, $a(t)$ is the scale parameter (often found elsewhere, especially as R), $K \in \{1, 0, -1\}$ the curvature constant and $f \in \{\sin\chi, \chi, \sinh\chi\}$ a parameter that depends on K. This last symmetry is based on the long-range well-realized assumption that the cosmos is homogeneous and isotropic.

For both (3.49) and (3.50), the generalized Einstein field equations may be given (and here, c will further be written explicitly, and Λ_0 will be assumed as vanishing).

The line element (3.49), assuming an ideal liquid with energy density distribution $\varepsilon = \varrho c^2$ and pressure p, leads to Einsteinian field equations in the form ($\breve{\alpha} \gg 1$),

$$\begin{aligned}
&e^{\nu-\lambda}\left(\frac{\nu''}{2} + \frac{\nu'^2}{4} - \frac{\nu'\lambda'}{4} + \frac{\nu'}{r}\right) - \frac{1}{c^2}\frac{\ddot{\lambda}}{2} - \frac{1}{c^2}\frac{\dot{\lambda}^2}{4} + \frac{1}{c^2}\frac{\dot{\lambda}\dot{\nu}}{4} + \frac{1}{2l^2}(1+\xi)^{-1}\xi\left(1 + \frac{3}{2}\xi\right)e^\nu \\
&= \frac{8\pi}{(1+\xi)}\frac{G_0}{c^4}\left[\left(e^{-\nu} - \frac{v_1^2}{c^2}e^{-\lambda}\right)^{-1}\left(\varepsilon + \frac{v_1^2}{c^2}p\,e^{\nu-\lambda}\right) - \underline{\frac{1}{2}\left(1 - \frac{1}{3}\hat{q}\right)(\varepsilon - 3p)e^\nu}\right] \\
&\quad + (1+\xi)^{-1}\left[\frac{\ddot{\xi}}{c^2} - \frac{\dot{\nu}}{2c^2}\dot{\xi} - \frac{\nu'}{2}e^{\nu-\lambda}\xi'\right] ,
\end{aligned} \tag{3.51}$$

$$
\begin{aligned}
&e^{\lambda-\nu}\frac{1}{c^2}\left(\frac{\ddot{\lambda}}{2}+\frac{\dot{\lambda}^2}{4}-\frac{\dot{\lambda}\dot{\nu}}{4}\right)-\frac{\nu''}{2}-\frac{\nu'^2}{4}+\frac{\nu'\lambda'}{4}+\frac{\lambda'}{r}-\frac{1}{2l^2}(1+\xi)^{-1}\xi\left(1+\frac{3}{2}\xi\right)e^{\lambda}\\
&=\frac{8\pi}{(1+\xi)}\frac{G_0}{c^4}\left[\left(e^{-\nu}-\frac{v_1^2}{c^2}e^{-\lambda}\right)^{-1}\left(\frac{v_1^2}{c^2}\,\varepsilon+p\,e^{\lambda-\nu}\right)+\underline{\frac{1}{2}\left(1-\frac{1}{3}\hat{q}\right)(\varepsilon-3p)e^{\lambda}}\right]\\
&\quad+(1+\xi)^{-1}\left[\xi''-\frac{\dot{\lambda}}{2c^2}e^{\lambda-\nu}\dot{\xi}-\frac{\lambda'}{2}\xi'\right]\,, \qquad (3.52)
\end{aligned}
$$

$$
\frac{1}{c}\frac{\dot{\lambda}}{r}=-\frac{8\pi}{(1+\xi)}\frac{G_0}{c^4}\left[e^{-\nu}-\frac{v_1^2}{c^2}e^{-\lambda}\right]^{-1}(\varepsilon+p)\,\frac{v_1}{c}-(1+\xi)^{-1}\frac{1}{c}\left[\dot{\xi}'-\frac{\nu'}{2}\dot{\xi}-\frac{\dot{\lambda}}{2}\xi'\right]\,, \qquad (3.53)
$$

$$
\begin{aligned}
&e^{-\lambda}\left(1+\frac{r}{2}(\nu'-\lambda')\right)-1+\frac{r^2}{2l^2}(1+\xi)^{-1}\xi\left(1+\frac{3}{2}\xi\right)=\\
&\quad=-\frac{8\pi}{(1+\xi)}\frac{G_0}{c^4}\left[p\,r^2+\underline{\frac{1}{2}\left(1-\frac{1}{3}\hat{q}\right)(\varepsilon-3p)r^2}\right]-(1+\xi)^{-1}re^{-\lambda}\xi'\,. \qquad (3.54)
\end{aligned}
$$

For $\xi = 0$, the original Einstein field equations for central symmetry are restored. For these, the *Birkhoff theorem* is valid. Thus, for vacuum ($\varepsilon = 0$) all fields are static and $\nu = \nu(r)$ and $\lambda = \lambda(r)$. For nonvanishing excitations ξ, however, this cannot be stated directly.

The Higgs-like equation for the excited Higgs-like field ξ yields

$$
\frac{1}{c^2}\ddot{\xi}e^{-\nu}-\xi''e^{-\lambda}-\frac{1}{c^2}\frac{\dot{\nu}-\dot{\lambda}}{2}e^{-\nu}\dot{\xi}-\frac{\nu'-\lambda'}{2}e^{-\lambda}\xi'-\frac{2}{r}e^{-\lambda}\xi'+\frac{1}{l^2}\xi=+\hat{q}\frac{8\pi}{3}\frac{G_0}{c^4}(\varepsilon-3p)\,. \qquad (3.55)
$$

Within GR, the vacuum solution for central symmetry with a vanishing cosmological constant Λ_0 is the Schwarzschild metric

$$
\mathrm{d}s^2=\left(1-\frac{r_S}{r}\right)c^2\,\mathrm{d}t^2-\frac{\mathrm{d}r^2}{1-r_S/r}-r^2\,\mathrm{d}\vartheta^2-r^2\sin^2\vartheta\,\mathrm{d}\varphi^2\,, \qquad (3.56)
$$

whereas the constant $r_S = 2M_1G_0/c^2 = B$ is the so-called Schwarzschild radius, valid for a constant gravitational coupling G_0 which in GR is $G_0 = G_{\mathrm{N}}$. Here, r_{S} represents the radius which a body of mass M_1 must have so that its rest-mass M_1c^2 is equal to its internal gravitational potential energy $V_{\mathrm{N}} \simeq G_0M_1^2/r_{\mathrm{S}}$ (see [102]).

Within GR, no particle, not even a photon, can escape from a region of radius r_{S} around a body of mass M_1. Hence, the Schwarzschild radius defines the horizon of a black hole, so that for $r = r_{\mathrm{S}}$, there appears a horizon singularity. Then, e^{λ} diverges. However, within this model, for $\hat{q} = 1$, an exact analysis for the static case and in the limit $l \to \infty$ (for which the symmetry stays broken) leads to the disappearance of this horizon singularity. There appears a quintessential-like effect which is coupled to the integration constant A of the scalar field. The scalar field gives the strength of a further term of gravitation added to the usual term from the $1/r$ potential. This Newtonian and classical potential has B as the integration constant. The disappearance of the singularity is usually caused by the new factor with integration constant A, even in the case when this constant is much smaller than the constant B of the Newtonian term [40, 103].[18]

For values as in [39], $1/l^2\xi$ terms are negligible, and the strong equivalence principle is then valid even for supra-solar as well as microscopic distances. However, a linear analysis for $\hat{q} = 1$ leads to the necessity of re-scaling the appearing coupling constant G_0 as $G_0 = 3/4G_{\text{N}}$ in the case $l \gg r$, and $G_0 = G_{\text{N}}$ for $l \ll r$,[19] for G_{N} as the Newtonian coupling constant (see [80] or [105]). Linearization in the ν and λ and not in ξ (which is valid), leads for length scales l of the order of magnitude of some galaxy radii (some kiloparsec, with 1 pc ≈ 3×10^{13} km) to flattened rotation curves (see Section 3.1.1) in a model of galaxies with polytropic density distribution and with polytropic index 2, with or without assuming a very massive core [39]. Further, for the strongest bars in isolated galaxies, a similar value of the length scale, of about 10 kpc, is obtained in [104], within the general BW class and with an arbitrary potential (but analogous field equations, with $p = 0$). This value is beyond the accuracy of the experiments presented in [96], and represents a mass $M = \hbar/(\hbar l) \sim 10^{-26}\,\text{eV/c}^2$. Moreover, with such values, the dynamics of the model seem to be able to give solar-relativistic effects, such as perihelion advance, accurately [105].

Let us now take the linearized scalar field equation

$$\left(\frac{1}{c^2}\frac{\partial}{\partial t} - \Delta\right)\xi + \frac{1}{l^2}\xi = \frac{8\pi G_0}{3\,c^4}\hat{q}\,(\varepsilon - 3p) \tag{3.57}$$

and neglect all time derivatives. Then, the same for the linear form of the 0–0 Einstein field equation (3.51). This leads to a $\hat{q}$-independent Poisson equation for the density distribution $\varrho = \varepsilon/c^2$ after adding the linearized Higgs-like field equation (3.57) to (3.51),[20]

$$\nabla^2\Psi = 4\pi G\left[\varrho + 3\frac{p}{c^2}\right], \quad \text{whereas} \quad \Psi = \Phi + \frac{c^2}{2}\xi\,, \tag{3.58}$$

with the gravitational potential $\Phi = \nu c^2/2$ related to the metric $g_{\mu\nu}$ in the linear case.

[18] $A = -2/3G_0/c^4 \int T\sqrt{-\text{g}}\,\text{d}^3x$ is derived from the Higgs-like field equation $\xi^{,\mu}_{;\mu} = 1/\sqrt{-\text{g}}\left(\sqrt{-\text{g}}\xi^{,\mu}\right)_{,\mu} = (8\pi)/3G_0/c^4\,T$ for $l \to \infty$ (cf. [40]). Note that A depends on the density distribution and on the pressure p.

[19] Actually, the same re-scaling can be found in [104]. However, we have G_{eff} as the actual measured value and the one coupling term when taking both terms of the potential into account. For higher values of ξ, the rescaling, taking all terms, may differ from the one above.

[20] For $p \neq 0$, the measured, effective mass (defined through the asymptotic "gravitational force" with $r \to \infty$) differs from the integral over the density distribution because of terms from the energy of the gravitational field [106, 107] which appear in the Poisson equation (3.58). Further, the Schwarzschild mass differs from the integral over the density exactly by this term (as a usual consequence of $p \neq 0$ within GR). A changed energy of the gravitational field may play an important role within this model.

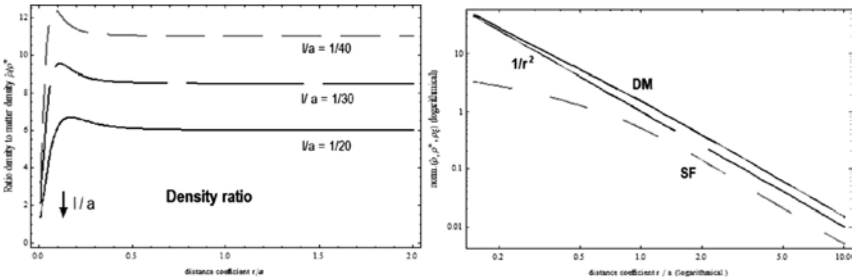


Figure 3.1 Density ratio $\hat{\varrho}/\varrho^*$ for different length scales, and density contributions $\hat{\varrho}(DM)$, $\varrho^*(1/r^2)$ and $\varrho(SF)$ normalized to $v_t^2/(4\pi G_N a^2)$ for $l_a = 1/2$.

If one assumes for $\hat{q} = 1$ that flat rotation curves are obtained through completely linearized fields, then one may derive the form of the density for flat rotation curves by demanding the constancy of the rotation velocity $v_t^2 = r(\mathrm{d}\Phi/\mathrm{d}r)$. It may then be inserted into (3.58) for Φ, which possesses a Yukawa term from ξ. This leads to a density term of Newtonian type and a contribution of the scalar field:

$$\hat{\varrho} = \varrho + \frac{3}{2}\frac{p}{c^2} = \underline{\frac{v_t^2}{4\pi G_\mathrm{N} r^2}} + \frac{\xi c^2}{8\pi G_\mathrm{N} l^2} := \underline{\varrho^*} + \varrho_\xi \ . \tag{3.59}$$

The assumption that only usual density distribution of matter (ϱ^*) acts as source of the Higgs-like field leads to a pressure related to the (energy) density contribution of the scalar field,

$$p = \frac{2}{9}\varrho_\xi c^2 = \frac{2}{9}\varepsilon_\xi \ . \tag{3.60}$$

Analogously to [108], we have

$$\xi = \frac{1}{2r_a}\frac{v_t^2}{c^2}\left[e^{-r_a/l_a}\mathrm{sinhInt}\left(\frac{r_a}{l_a}\right) - \sinh\left(\frac{r_a}{l_a}\right)\mathrm{expIntEi}\left(-\frac{r_a}{l_a}\right)\right] , \tag{3.61}$$

$$\hat{\varrho} = \frac{v_t^2}{4\pi G_N a^2}\left\{\frac{1}{r_a^2} + \frac{1}{4l_a^2 r_a}\left[e^{-r_a/l_a}\mathrm{sinhInt}\left(\frac{r_a}{l_a}\right) - \sinh\left(\frac{r_a}{l_a}\right)\mathrm{expIntEi}\left(-\frac{r_a}{l_a}\right)\right]\right\} , \tag{3.62}$$

with $r_a = r/a$ and $l_a = l/a$, and a as the length scale of the spherical system (e.g. the distance at which galaxies possess flat rotation curves). These are analogous to the solutions for [108]. We use following abbreviations: sinhInt ≡ sinhIntegral, expIntEi ≡ ExpIntegralEi. For (3.61) and (3.62) the size of the halo is assumed larger than a, and l of the order of magnitude of the radius R_0 (the core) of the galaxy. Thus, terms of (3.61) and (3.62) related to the halo radius $\widehat{B}$ through "expIntegralEi($-\widehat{B}/l$)" are neglected. Given this, the density for different l/a and of the different density terms is as given in the figure $\hat{\varrho}$ is what is often known as the DM-profile ϱ_DM. For large l in relation to a, the inverse square contribution of ϱ^* dominates, while for smaller l/a relations the scalar field (SF) contribution becomes more and more dominant for the total density (see figure).

For the Friedmann–Lemaître–Robertson–Walker (RW) metric in (3.50), the continuity condition (3.36) yields for the density ϱ

$$\frac{1}{c}\dot{\varrho} + 3\frac{\dot{a}}{a}\left(\varrho + \frac{p}{c^2}\right) = -(1-\hat{q})\frac{1}{2\,c}\frac{\dot{G}_{\text{eff}}}{G_{\text{eff}}}\left(\varrho - 3\frac{p}{c^2}\right), \tag{3.63}$$

meaning that the scalar field produces no entropy process for $\hat{q} = 1$, other than with $\hat{q} = 0$. Further, the Higgs-like field equation for (3.50) yields

$$\ddot{\xi} + 3\frac{\dot{a}}{a}\dot{\xi} + \frac{c^2}{l^2}\xi = \frac{8\pi G_0}{3}\hat{q}\frac{(\varrho - 3p/c^2)}{(1 + 4\pi/(3\breve{\alpha}))}, \tag{3.64}$$

or with

$$\frac{\ddot{\xi}}{1+\xi} = \frac{1}{G_{\text{eff}}^2}\left(2\,\dot{G}_{\text{eff}}^2 - \ddot{G}_{\text{eff}}\,G_{\text{eff}}\right), \tag{3.65}$$

the Higgs-like field equation with the form

$$\frac{1}{G_{\text{eff}}^2}\left(\ddot{G}_{\text{eff}}\,G_{\text{eff}} - 2\,\dot{G}_{\text{eff}}^2\right) + 3\frac{\dot{a}}{a}\frac{\dot{G}_{\text{eff}}}{G_{\text{eff}}} + \frac{c^2}{l^2}\left(\frac{G_{\text{eff}}}{G(v)} - 1\right) = -\frac{8\pi G_{\text{eff}}}{3}\hat{q}\frac{\left(\varrho - 3p/c^2\right)}{(1 + 4\pi/3\breve{\alpha})}. \tag{3.66}$$

With (3.64) and (3.40), (3.50) leads to the generalized Friedmann–Lemaître equations in the forms (independent of the source parameter $\hat{q}$)

$$\begin{aligned}\frac{\dot{a}^2 + Kc^2}{a^2} &= \frac{1}{1+\xi}\left[\frac{8\pi G_0}{3}(\varrho + V(\xi)) - \frac{\dot{a}}{a}\dot{\xi} + \frac{\pi}{3\breve{\alpha}}\frac{\dot{\xi}^2}{1+\xi}\right] \\ &= \frac{8\pi G_{\text{eff}}}{3}\varrho + \frac{1}{3}\Lambda(\xi)c^2 + \frac{\dot{a}}{a}\frac{\dot{G}_{\text{eff}}}{G_{\text{eff}}} + \frac{\pi}{3\breve{\alpha}}\frac{\dot{G}_{\text{eff}}^2}{G_{\text{eff}}^2} \\ &= (1+\xi)^{-1}\left[\frac{8\pi G_0}{3}\varrho + \left(\frac{c^2}{4l^2}\xi^2\left(1 + \frac{4\pi}{3\breve{\alpha}}\right) - \frac{\dot{a}}{a}\dot{\xi} + \frac{\pi}{3\breve{\alpha}}\frac{\dot{\xi}^2}{1+\xi}\right)\right],\end{aligned} \tag{3.67}$$

and

$$\begin{aligned}\frac{\ddot{a}}{a} + \frac{\dot{a}^2 + Kc^2}{a^2} &= -\frac{1}{1+\xi}\left[8\pi G_0\left(\frac{p}{c^2} - V(\xi)\right) + \ddot{\xi} + 2\frac{\dot{a}}{a}\dot{\xi} + \frac{\pi}{\breve{\alpha}}\frac{\dot{\xi}^2}{1+\xi}\right] \\ &= -8\pi G_{\text{eff}}\frac{p}{c^2} - (1+\xi)^{-1}\left[\ddot{\xi} - \frac{3c^2}{4l^2}\xi^2\left(1 + \frac{4\pi}{3\breve{\alpha}}\right)\right] + 2\frac{\dot{a}}{a}\frac{\dot{G}_{\text{eff}}}{G_{\text{eff}}} - \frac{\pi}{\breve{\alpha}}\frac{\dot{G}_{\text{eff}}^2}{G_{\text{eff}}^2},\end{aligned} \tag{3.68}$$

with the density distribution ϱ and pressure p and the cosmological function $\Lambda(\xi)$, known from (3.41).

Further, the second generalized Friedmann–Lemaître equation may be written as

$$2\frac{\ddot{a}}{a} + \frac{\dot{a}^2 + Kc^2}{a^2} = -\,8\pi G_{\text{eff}}\frac{p}{c^2} + \Lambda(\xi)c^2 - \frac{1}{G_{\text{eff}}^2}\left[\dot{G}_{\text{eff}}^2\left(2 + \frac{\pi}{\breve{\alpha}}\right) - \ddot{G}_{\text{eff}}G_{\text{eff}}\right] + 2\frac{\dot{a}}{a}\frac{\dot{G}_{\text{eff}}}{G_{\text{eff}}}. \tag{3.69}$$

Both the generalized Friedmann–Lemaître equations give the usual ones for $\xi = 0$. If the scalar field excitation does not vanish, however, there appears a cosmological function $\Lambda(\xi) = \Lambda(G_{\text{eff}})$ which is dependent on the relation between the gravitational coupling in the Higgs ground state $(G(v))$ and its effective term. Furthermore, other correction terms arise from the time dependence of the gravitational coupling itself.

Here, the solution for the density is

$$\varrho = \frac{M_\alpha}{a^{3(1+\alpha)}}\left(\frac{G(v)}{G_{\text{eff}}}\right)^{(1/2)(1-3\alpha)(1-\hat{q})} , \tag{3.70}$$

with a barotropic pressure parameter $\alpha = p/\left(\varrho c^2\right)$ and an integration constant M_α and a current Hubble parameter $(\alpha = 0)$ that may be written as

$$H_0 = -\frac{1}{3}\frac{a_0^3}{M_0}\dot{\varrho}_0(1+\xi_0)^{-(1/2)(1-\hat{q})} + (1-\hat{q})\,\frac{1}{6}\,\dot{\xi}_0\,(1+\xi_0)^{-1} \tag{3.71}$$

$$= -\frac{1}{3}\frac{a_0^3}{M_0}\dot{\varrho}_0\left(\frac{G(v)}{G_{\mathit{eff}0}}\right)^{-(1/2)(1-\hat{q})} - (1-\hat{q})\,\frac{1}{6}\,\frac{\dot{G}_{\mathit{eff}0}}{G_{\mathit{eff}0}} . \tag{3.72}$$

It is easily noticed that an increase (decrease) in the density is related to a contraction $(\dot{a} < 0)$ (expansion $(\dot{a} > 0)$) of the cosmos and that for $\hat{q} = 0$ the time variation of the gravitational coupling plays a role in cosmic expansion, too (higher derivatives reduce the value of H_0). In the same way, this variation in the coupling leads to a screening of the density parameter ϱ in (3.70) in relation to the case where the G-coefficient is negligible. The value of the density for $q = 1$ is smaller if $\xi < 0$, i.e. $G_{\text{eff}} > G(v)$ (anti-screening of the gravitational constant).

At the same time, following the Friedmann–Lemaître equations (3.67) and (3.68), the deceleration parameter q, defined by $q = -1/H^2\ddot{a}/a$, is given by

$$\frac{\ddot{a}}{a} = -\frac{4\pi G_{\text{eff}}}{3}\left(\varrho + 3\frac{p}{c^2} - 2V(\xi)\right) + f(G)$$

$$= -\frac{4\pi G_{\text{eff}}}{3}\left(\varrho + 3\frac{p}{c^2}\right) + \frac{1}{3}\Lambda(\xi)c^2 + f(G), \tag{3.73}$$

with

$$f(G) = \frac{1}{2}\left[\frac{1}{G_{\text{eff}}}\left(\ddot{G}_{\text{eff}} + \frac{\dot{a}}{a}\dot{G}_{\text{eff}}\right) - 2\,\frac{\dot{G}^2_{\text{eff}}}{G^2_{\text{eff}}}\left(1 + \frac{\pi}{3\breve{\alpha}}\right)\right]. \tag{3.74}$$

Apart from the fact that G and Λ are functional, it is $f(G)$ which makes (3.73) formally different from the usual equation in GR, where there is acceleration ($q < 0$) merely for $\Omega_\Lambda > \Omega/2$. This new term gives the changes of dynamics caused by the time dependence of the effective coupling constant and (together with the correction to the first Friedmann–Lemaître equation) it can be compared with an analog function derived within Modified Gravity (MOG) by Moffat [97], but here with a functional cosmological term Λ and defining a scalar field $\check{\xi} = 1 + \xi$.

The deceleration parameter q is a dimensionless measure of the cosmic acceleration of the expansion of the universe (seee Section 3.1.1), which is related to dark energy as measured in Super Novae of type Ia (SNeIa) [31–33]. Furthermore, through $q = \Omega/2 - \Omega_\Lambda - (\ddot{a}a/\dot{a}^2)f(G)$ (the way the deceleration parameter looks within this model), $f(G)$ is related to the cosmological density parameters Ω_i,[21] the form of which is derivable from the Friedmann–Lemaître equations, so that the following can be written, using (3.67):

$$a^2 = \frac{\dot{a}^2 + Kc^2}{8\pi G_{\mathrm{eff}}/3\,(\varrho + 3p/c^2) + 1/3\Lambda(\xi)c^2 + 1/3\Lambda_I(\xi)c^2}\,, \tag{3.75}$$

with ($\alpha \approx 0$ is valid for matter dominance $\varepsilon \gg p$)

$$\Omega = \frac{8\pi G_{\mathrm{eff}}}{3\,H^2}\varrho\,(1 + 3\alpha)\,, \quad \Omega_\Lambda^* = \frac{1}{3}c^2\,(\Lambda(\xi) + \Lambda_I(\xi))\;H^{-2} \equiv \Omega_\Lambda + \Omega_I\,, \tag{3.76}$$

and with

$$\Lambda_I(\xi) := \frac{3\,H}{c^2}\,\frac{\dot{G}_{\mathrm{eff}}}{G_{\mathrm{eff}}} + \frac{\pi}{c^2\,\breve{\alpha}}\,\frac{\dot{G}^2_{\mathrm{eff}}}{G^2_{\mathrm{eff}}} \tag{3.77}$$

as part of an effective cosmological function $\Lambda_{\mathrm{eff}}(\xi)$ which derives from the correction to the first Friedmann–Lemaître equation and which depends on the Hubble parameter $H = \dot{a}/a$ and on the time variation of the gravitational coupling. Further, the density of the cosmological term

$$\varrho_\Lambda = V(\xi) - \frac{3\,H}{8\pi G_0}\dot{\xi} + \frac{v^2}{8}\,\frac{\dot{\xi}^2}{1 + \xi} \tag{3.78}$$

$$= V(\xi) + \frac{3\,H}{8\pi}\,\frac{\dot{G}_{\mathrm{eff}}}{G^2_{\mathrm{eff}}} + \frac{\dot{G}^2_{\mathrm{eff}}}{8\breve{\alpha}\,G^3_{\mathrm{eff}}} \tag{3.79}$$

may be defined, so that

$$\dot{a}^2 = \frac{8\pi G_{\mathrm{eff}}}{3}(\varrho + \varrho_\Lambda)a^2 - Kc^2 \tag{3.80}$$

is valid for the scale factor a (with an effective value $\varrho_{\Lambda\,\mathrm{eff}} = \varrho_\Lambda/(1 + \xi)$). The same applies for a pressure term, so that for

$$\ddot{a} = -\frac{4\pi G_{\mathrm{eff}}}{3}\left(\varrho + 3\frac{p}{c^2} + 3\frac{p_\Lambda}{c^2}\right)a\,, \tag{3.81}$$

we have

$$p_\Lambda = -\frac{2}{3}V(\xi)\,c^2 - \frac{c^2}{8\pi G_{\mathrm{eff}}}\left[\frac{\ddot{G}_{\mathrm{eff}}}{G_{\mathrm{eff}}} + H\frac{\dot{G}_{\mathrm{eff}}}{G_{\mathrm{eff}}} - 2\frac{\dot{G}^2_{\mathrm{eff}}}{G_{\mathrm{eff}}}\left(1 + \frac{\pi}{3\breve{\alpha}}\right)\right]\,. \tag{3.82}$$

[21] These parameters may differ from the standard ones. For instance, Ω_i in the standard approach represent observed quantities based on a screened value of the gravitational constant (or of density), so that $\Omega_i = (G_{\mathrm{eff}}/G_0)\,\Omega_i^{\mathrm{std}}$, whereas the geometry of the universe is determined by the constant's "bare" value .

This pressure term is dependent on G_{eff} and its derivatives, as is the case for ϱ_Λ. One can see that high effective values of the coupling parameter or, in general, high values of the scalar field excitations, as well as positively valued (second) derivatives of the gravitational coupling G_{eff}, may lead to negative pressures. These strengthen the cosmological acceleration, whereas $\dot{G}^2_{\text{eff}}$ may act as deceleration factor. The latter term, however, may be considered as negligible under normal circumstances, and relevant only for the primordial universe, possibly in relation with primeval inflation. The concept of primeval, cosmic inflation, first proposed by Alan Guth in 1981 [98] and later improved by Albrecht, Steinhardt [99] and Linde [100], assumes a phase of very highly accelerated expansion in the early universe to explain horizon and flatness problems of cosmology. Often, an hypothetical scalar field, namely the inflaton field, is proposed in this context. “New” and “chaotic” inflation differ from the original, old one, due to the initial conditions of this scalar field.

An inflationary universe with induced gravitation can be derived within this context [86,87,101]. This model can indeed lead to primeval *new* or *chaotic* inflation. As a matter of fact, the Penrose–Hawking energy condition $3p + \varrho c^2 \geq 0$ [109,110] may be broken for chaotic inflation, for which a Big Bounce would be expected (that means no initial singularity before inflation). This case can be compared with the case of the works in [111], according to which Yukawa interactions of the magnitude of the nuclear density can lead to negative pressures that might play an important role in early stages of the universe, so that the Penrose–Hawking condition may not be satisfied. This Yukawa interaction in the primordial universe would be related to a pressure like p_Λ (coming from the potential $V(\xi)$ and the scalar-field derivatives, translated as the variable gravitational coupling), possibly contributing to the mechanism of inflation and dark energy as part of the cosmological term Λ.

Acknowledgments

This work was supported by the Graduate School “Mathematical Analysis of Evolution, Information and Complexity” and the Institute of Theoretical Physics of the University of Ulm. Furthermore, NMBR would like to thank Prof. Dr. Heinz Dehnen of the University of Konstanz and Dr. Hemwati Nandan of the Centre for Theoretical Studies of the Indian Institute of Technology for their important help in the realization and development of this work.

References

1 Sartre, J.-P. (**1989**) *Wahrheit und Existenz*, Rowohlt 22378, Reinbeck, 2nd. Ed., pp. 17ff.

2 Einstein, A. (**1944**) Bertrand Russel und das philosophische Denken, in *The Philosophy of Betrtrand Russel*, Library of Living Philosophers, vol. V, Cambridge University Press, London, pp. 35ff.

3 Kitcher, P. (**1982**) *Abusing Science: The Case against Creationism*, MIT Press, Cambridge, p. 45.

4 Popper, K. *Conjectures and Refutations*, Routledge and Kegan Paul, London, 33ff.

5 Einstein, A. (**1915**) Die Feldgleichungen der Gravitation. Sitzungsbericht der Preussischen Akademie Wissenschaften 844; reprint in: *The Collected Papers of Albert Einstein, vol. 6*, Princeton University Press, London, **1996**, pp. 244ff.

6 Einstein, A., Podolsky, B. and Rosen, N. (**1935**) Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, **47**, 777.

7 Bennet, C.H. *et al.* (**1993**) Teleporting an Unknown Quantum State via Dual Classical and Einstein-Podolsky-Rosen Channels. *Physical Review Letters*, **70**, 1895.

8 Bouwmeester, D. *et al.* (**1997**) Experimental Quantum Teleportation. *Nature*, **390**, 575.

9 Brassard, G., Braunstein, S.L. and Cleve, R. (**1998**) Teleportation as a Quantum Computation. *Physica D*, **120**, 43.

10 Laiho, R., Molotkov, S.N. and Nazin, S.S. (**2000**) On Teleportation of a Completely Unknown State of Relativistic Photon. *Physics Letters A*, **278**(1–2), 9.

11 Yukawa, H. (**1935**) On the Interaction of Elementary Particles. *Proceedings of the Physical-Mathematical Society of Japan*, **17**, 48.

12 Lattes, C.M.G. *et al.* (**1947**) Processes Involving Charged Mesons. *Nature*, **160**, 453.

13 Hofstadter, R. (**1957**) Nuclear and Nucleon Scattering of High-Energy Electrons. *Annual Review of Nuclear Science*, **7**, 231.

14 Friedman, J.I. and Kendall, H.W. (**1972**) Dep Inelastic Electron Scattering. *Annual Review of Nuclear Science*, **22**, 203.

15 Gell-Mann, M. (**1964**) A Schematic Model of Baryons and Mesons. *Physics Letters*, **8**, 214.

16 Zweig, G. (**1964**) An SU(3) Model for Strong Interaction Symmetry and its Breaking I. CERN-8182-TH-401.

17 Yang, C.N. and Mills, R.L. (**1954**) Conservation of Isotopic Spin and Isotopic Gauge Invariance. *Physical Review*, **96**, 191.

18 Higgs, P.W. (**1964**) Broken Symmetries and the Masses of Gauge Bosons. *Physical Review Letters*, **13**, 508.

19 t'Hooft, G. (**1971**) Renormalization of Massless Yang-Mills Fields. *Nuclear Physics B*, **33**, 173.

20 CDF Collaboration (**1995**) Observation of Top Quark Production in Pbar-P Collisions. *Physical Review Letters*, **74**, 2626.

21 Pauli, W. (**1930**) Offener Brief an die Gruppe der Radioaktiven bei der Gauvereins-Tagung zu Tübingen.

22 Reines, F. and Cowan, C.L. Jr. (**1956**) The Neutrino. *Nature*, **178**, 446.

23 Martin, S.P. (**1997**) A Supersymmetry Primer, arXiv:hep-ph/9709356v4.

24 Gallis, M.R. and Flemming, G.N. (**1990**) Environmental and Spontaneous Localization. *Physical Review A*, **42**, 38.

25 Kolb, E.W. and Turner, M.S. (**1990**) *The Early Universe*. Addison-Wesley, series Frontiers in Physics, Massachusetts.

26 Zwicky, F. (**1933**) Die Rotverschiebung von Extragalaktischen Nebeln. *Helvetica Physica Acta*, **6**, 110.

27 Einstein, A. (**1917**) Kosmologische Betrachtungen zur Allgemeinen Relativitätstheorie. Sitzungsbericht der Preussischen Akademie der Wissenschaften 142; reprint in: *The Collected Papers of Albert Einstein, vol. 6*, Princeton University Press, London, **1996**, pp. 540ff.

28 Wetterich, C. (**1988**) Cosmologies with variable Newton's 'constant'. *Nuclear Physics B*, **302**, 668.

29 Peebles, P.J.E. and Ratra, B. (**1988**) Cosmological consequences of a rolling homogeneous scalar field. *Physical Review D*, **37**, 3406.

30 Peebles, P.J.E. and Ratra, B. (**2003**) The Cosmological Constant and Dark Energy. *Reviews of Modern Physics*, **75**, 559.

31 Garnavich, P.M. *et al.* (**1998**) Constraints on Cosmological Models from Hubble Space Telescope Observations of High-z Supernovae. *Astrophysical Journal*, **493**, L53.

32 Perlmutter, S. *et al.* (**1998**) (SCP), Measurements of Ω and Λ from 42 High-Redshift Supernovae. *Nature*, **391**, 51.

33 Riess, A.G. *et al.* (**1998**) Observational Evidence from Supernovae for an Accelerating Univrse and a Cosmological Constant. *Astronomical Journal*, **116**, 1009.

34 Blome, H.J. and Priester, W. (**1991**) Big Bounce in the Very Early Universe. *Astronomy and Astrophysics*, **50**, 43.

35 Hoell, J. and Priester, W. (**1991**) Void Structure in the Early Universe – Implications for a $\Lambda > 0$ Cosmology. *Astronomy and Astrophysics*, **251**, L23.

36 Spergel, D.N. *et al.* (**2007**) Wilkinson Microwave Anisotropy Probe (WMAP) Three Year Results: Implications for Cosmology, astro-ph/0603449.

37 Fujii, Y. (**2000**) Quintessence, Scalar–Tensor Theories, and Non-Newtonian Gravity. *Physical Review D*, **62**, 044011.

38 Matos, T. and Guzman, F.S. (**2001**) On the Spacetime of a Galaxy. *Classical and Quantum Gravity*, **18**, 50055.

39 Bezares-Roder, N.M. and Dehnen, H. (**2007**) Higgs Scalar-Tensor Theory for Gravity and the Flat Rotation Curves of Spiral Galaxies. *General Relativity and Gravitation*, **39**(8), 1259.

40 Bezares-Roder, N.M., Nandan, H. and Dehnen, H. (**2007**) Horizon-less Spherically Symmetric Vacuum-Solutions in a Higgs Scalar-Tensor Theory of Gravity. *International Journal of Theoretical Physics*, **46**(10), 2420. Partially presented at the 24th IAGRG meeting on the Recent Advances in Gravitation and Cosmology, New Delhi, India (2007).

41 Rodríguez-Meza, M.A. and Cervantes-Cota, J.L. (**2004**) Potential-Density Pairs for Spherical Galaxies and Bulges: The Influence of Scalar Fields. *Monthly Notices of the Royal Astronomical Society*, **350**(2), 671.

42 Veltman, M. (**1977**) Second Threshold in Weak Interactions. *Acta Physica Polonica B*, **8**, 475.

43 Lee, T.D. (**1981**) *Particle Physics and Introduction to Field Theory*, Harwood Academic, New York.

44 Peskin, M.E. and Schroeder, D.V. (**1995**) *An Introduction to Quantum Field Theory*, Addison Wesley Reading, Massachusetts.

45 Bezares-Roder, N.M. and Nandan, H. (**2008**) Spontaneous Symmetry Breakdown and Critical Perspectives of Higgs Mechanism. *Indian Journal of Physics*, **82**(1), 69.

46 Nandan, H., Bezares-Roder, N.M. and Chandola, C. Screening Current and Dielectric Parameters in Dual QCD. *Accepted for publication in the Indian Journal of Pure and Applied Physics*.

47 Einstein, A. (**1913**) Zum gegenwärtigen Stande des Gravitationsproblems. *Physikalische Zeitschrift*, **14**, 198.

48 Brans, C. and Dicke, R.H. (**1961**) Mach's Principle and a Relativistic Theory of Gravitation. *Physical Review*, **124**, 925.

49 Carloni, S. and Dunsby, P.K.S. (**2006**) A Dynamical System Approach to Higher Order Gravity, gr-qc/0611133v1. Talk given at IRGAC, July 2006.

50 Caldwell, R.R., Dave, R. and Steinhardt, P.J. (**1998**) Cosmological Imprint of an Energy Component with General Equation of State. *Physical Review Letters*, **80**, 1582.

51 Boisseau, B. *et al.* (**2000**) Reconstruction of a Scalar–Tensor Theory of Gravity in an Accelerating Universe. *Physical Review Letters*, **85**, 2236.

52 Amendola, L. (**2001**) Dark Energy and the BOOMERANG Data. *Physical Review Letters*, **86**, 196.

53 Dehnen, H., Ghaboussi, F. and Schröder, J. (**1990**) Wissenschaftliche Zeitschrift der Friedrich-Schiller-Universität Jena **39**, 41.

54 Dehnen, H. and Frommert, H. (**1991**) Higgs Field Gravity Within the Standard Model. *International Journal of Theoretical Physics*, **30**(7), 985.

55 Dehnen, H. and Frommert, H. (**1990**) Higgs Field Gravity. *International Journal of Theoretical Physics*, **29**(6), 537.

56 van der Bij, J.J. (**1994**) Can Gravity Make the Higgs Particle Decouple?. *Acta Physica Polonica B*, **25**(5), 827.

57 van der Bij, J.J. (**1995**) Can Gravity Play a Role at the Electroweak Scale?. *International Journal of Physics*, **1**, 63. Based on talks at the DPG meeting, Dortmund, 1–4 March, 1994 and Bad Honnef, 7–10 march, 1994.

58 Feynman, R.P., Morignio, F.B. and Wagner, W.G. **1995** *Feynman Lectures on Gravitation*, edited by Brian Hatfield (Addison-Wesley, Reading).

59 Damour, T. (**2000**) Experimental Tests of Relativistic Gravity. *Nuclear Physics B*, **80**, 41.

60 Fierz, M. (**1956**) *Helvetica Physica Acta*, **29**, 128.

61 Jordan, P. *Nature* **164** (**1949**) 112; *Schwerkraft und Weltall*, Friedr. Vieweg & Sohn Verlag, 2. Aufl., Braunschweig, **1955**.

62 Fauser, B. (**2001**) Projective Relativity: Present Status and Outlook. *General Relativity and Gravitation*, **33**, 875.

63 Cotsakis, S. (**1997**) Mathematical Problems in Higher Order Gravity and Cosmology, gr-qc/9712046v1. Talk presented at the Eight Marcel Grossmann Meeting, Jerousalem, June 22-27, 1997.

64 Green, M.B., Schwarz, J.H. and Witten, E. (**1998**) *Superstring Theory*. Cambridge Universtiy Press, Cambridge.

65 Jordan, P. (**1968**) Bemerkungen zu der Arbeit von H. Hönl und H. Dehnen: 'Erlaubt die 3° K-Strahlung Rückschlüsse auf eine konstante oder veränderliche Gravitationszahl?'. *Zeitschrift für Astrophys.*, **68**, 201.

66 Penzias, A.A. and Wilson, R.W. (**1965**) A Measurement of Excess Antenna Temperature at 4080 Mc/s. *Astrophysical Journal*, **142**, 419.

67 Hönl, H. and Dehnen, H. (**1968**) Erlaubt die 3° K-Strahlung Rückschlüsse auf eine konstante oder veränderliche Gravitationszahl?. *Zeitschrift für Astrophysik*, **68**, 181.

68 Bergmann, P.G. (**1968**) Comments on the Scalar-Tensor Theory. *International Journal of Theoretical Physics* **1**(1), 25.

69 Wagoner, R.V. (**1970**) Scalar-Tensor Theory and Gravitational Waves. *Physical Review D*, **1**(12), 3209.

70 Assmann, J. (**1997**) *Das kulturelle Gedächtnis. Schrift, Erinnerung und politische Identität in frühen Hochkulturen*, C.H. Beck, München.

71 Catena, R., Pietroni, M. and Scarabello, L. (**2007**) Einstein and Jordan Frames Reconciled: A Frame-Invariant Approach to Scalar–Tensor Cosmology. *Physical Review D*, **76**, 094039.

72 Acharya, R. and Hogan, P.A. (**1973**) *Lettere al Nuovo Cimento* **6**, 668.

73 O'Hanlon, J. (**1972**) Intermediate-Range Gravity: A Generally Covariant Model. *Physical Review Letters*, **29**, 137.

74 Fujii, Y. (**1974**) Scalar-Tensor Theory of Gravitation and Spontaneous Breakdown of Scale Invariance. *Physical Review D*, **9**(4), 874.

75 Zee, A. (**1979**) Broken-Symmetric Theory of Gravity. *Physical Review Letters*, **42**(7), 417.

76 Weinberg, S. (**1974**) Gauge and global symmetries at high temperature. *Physical Review D*, **9**, 3357.

77 Reasenberg, R.D. *et al.* (**1979**) Viking Relativity Experiment: Verification of Signal Retardation by Solar Gravity. *Astrophysical Journal*, **234**, L219.

78 Hill, A. and van der Bij, J.J. (**1987**) Strongly Interacting Singlet-Doublet Higgs Model. *Physical Review D*, **36**(11), 3463.

79 van der Bij, J.J. (**1999**) Large Rescaling of the Scalar Condensate, Towards a Higgs-Gravity Connection?. Freiburg-THEP 99/09. Presented at the EPS-HEP999 meeting, July 1999, Tampere, Finland [hep-ph/9908297].

80 Dehnen, H., Frommert, H. and Ghaboussi, F. (**1992**) Higgs Field and a New Scalar-Tensor Theory of Gravity. *International Journal of Theoretical Physics*, **31**(1), 109.

81 Dehnen, H. and Frommert, H. (**1993**) Higgs Mechanism Without Higgs Particle. *International Journal of Theoretical Physics*, **32**, 1135.

82 Bronnikov, K.A. (**2002**) Scalar-Tensor Gravity and Conformal Continuations. *Journal of Mathematical Physics*, **43**(12), 6096.

83 Bronnikov, K.A. *et al.* (**1997**) Cold Black Holes in Scalar-Tensor Theories, gr-qc/9710092.

84 Bronnikov, K.A. and Shikin, G.N. (**2002**) Spherically Symmetric Scalar Vacuum: No-Go Theorems, Black Holes and Solitons. *Gravitation and Cosmology*, **8**, 107.

85 Bezerra, V.B. *et al.* (**2004**) Remarks on Some Vacuum Solutions of Scalar-Tensor Cosmological Models. *Brazilian Journal of Physics* **34**(2a), 562.

86 Cervantes-Cota, J.L. and Dehnen, H. (**1995**) Induced Gravity Inflation in the SU (5) GUT. *Physical Review D*, **51**, 395.

87 Cervantes-Cota, J.L. and Dehnen, H. (**1995**) Induced Gravity Inflation in the Standard Model of Particle Physics. *Nuclear Physics B*, **442**, 391.

88 Pimentel, L.O. and Mora, C. (**2000**) Quantum Cosmology in Bergmann–Wagoner Scalar-tensor Gravitational Theory, gr-qc/0009027.

89 Padmanabhan, T. (**2007**) Dark Energy and Gravity. arXiv:0705.2533v1 [gr-qc]. Invited Review for a special General Relativity and Gravitation isuue on Dark Energy.

90 Lee, T.D. and Yang, C.N. (**1956**) Question of Parity Conservation in Weak Interactions. *Physical Review*, **104**, 254.

91 Wu, C.S. *et al.* (**1957**) Experimental Tests of Parity Conservation in Beta Decay. *Physical Review* **105**(4), 1413.

92 Frommert, H. (**1991**) *Higgsmechanismus und Gravitation*, PhD thesis, Universität Konstanz, Fakultät für Physik, Konstanz.

93 Bento, M.C. *et al.* (**2000**) Self-Interacting Dark Matter and the Higgs Boson. *Physical Review D*, **62**, 041302.

94 Dehnen, H., Rose, B. and Amer, K. (**1995**) Dark Matter Particles and the Flat Rotation Curves of Spiral Galaxies. *Astrophysics and Space Science*, **234**, 69.

95 Gessner, E. (**1992**) A New Scalar Tensor Theory for Gravity and the Flat Rotation Curves of Spiral Galaxies. *Astrophysics and Space Science*, **196**, 29.

96 Adelberger, E.G. *et al.* (**2007**) Particle-Physics Implications of a Recent Test of the Gravitational Inverse-Square Law. *Physical Review Letters*, **98**, 131104.

97 Moffat, J.W. (**2007**) Non-Singular Cosmology in Modified Gravity, gr-qc/0610059v3 [gr-qc].

98 Guth, A. (**1981**) Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems. *Physical Review D*, **23**, 347.

99 Albrecht, A. and Steinhardt, P.J. (**1982**) Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking. *Physical Review Letters*, **48** 1220.

100 Linde, A.D. (**1982**) A new Inflationary Universe Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy and Primordial Monopole Problems. *Physics Letters B*, **108**, 389.

101 Cervantes-Cota, J.L. (**1996**) *Induced Gravity and Cosmology*, Hartung-Gorre-Verlag, Konstanzer Dissertationen 506, Konstanz.

102 Coles, P. and Lucchin, F. (*2003*) *Cosmology. The Origin and Evolution of Cosmic Structure*, John Wiley & Sons, West Sussex.

103 Nandan, H. and Bezares-Roder, N.M. Singularities and Black Holes in a Scalar–Tensor Theory of Gravity with Higgs Potential, in preparation.

104 Cervantes-Cota, J.L. *et al.* (**2007**) Newtonian Limit of Scalar-Tensor Theories and Galactic Dynamics: Isolated and Interacting Galaxies. *Revista Mexicana de Física S*, **53**(4), 22.

105 Bezares-Roder, N.M. and Steiner, F. work in progress.

106 Dehnen, H. (**1964**) Über den Energieinhalt statischer Gravitationsfelder nach der allgemeinen Relativitätstheorie in Newtonscher Näherung". *Zeitschrift für Physik*, **179**, 96.

107 Dehnen, H. (**1967**) Über den Energieinhalt statischer Gravitationsfelder nach der allgemeinen Relativitätstheorie in Newtonscher Näherung. *Zeitschrift für Physik*, **199**, 360.

108 Cervantes-Cota, J.L., Rodríguez-Meza, M.A. and Núñez, D. (**2007**) Flat rotation Curves Using Scalar-Tensor Theories. *Journal of Physics: Conference Series*, **91**, 012007.

109 Hawking, S. and Ellis, G. (**1968**) Cosmic Black-Body Radiation and the Existence of Singularities in Our Universe. *Astrophysical Journal* **152**, 25.

110 Penrose, R. (**1965**) Gravitational Collapse and Space-Time Singularities. *Physical Review Letters* **14**, 57.

111 Dehnen, H. and Hönl, H. (**1975**) The Influence of Strong Interactions on the Early Stages of the Universe. *Astrophysics and Space Science*, **33**, 49.

4
Relating Simulation and Modeling of Neural Networks

Stefano Cardanobile, Heiner Markert[1], Delio Mugnolo, Günther Palm, Friedhelm Schwenker

The discontinuity between cells, the role of the axons, the functioning of the synapses, the existence of synaptic microfissures, the leap each message makes across these fissures, make the brain a multiplicity immersed in its plane of consistency or neuroglia, a whole uncertain, probabilistic system ("the uncertain nervous system").

G. Deleuze and F. Guattari [11]

4.1
Introduction

We give an overview of neuron models used in theoretical neuroscience for "biologically realistic" modeling of single neurons, as well as several simplified models used for simulations of large-scale neural networks. The article emphasizes the connection between these models. It aims at describing how sophisticated biologically realistic neuron models can be simplified and reduced in order to end up with simpler models often used in computer science, neuroinformatics or theoretical physics. Our own specific compromise between simplicity and complexity is presented that combines several neurons into neural populations working as associative memories. The approach uses the so-called *spike counter model*, a particular simple model of spiking neurons. This architecture permits one to build artificial neural networks by using associatively connected populations of neurons as building blocks and then interconnecting these subnetworks into a larger network.

Modeling complex physical and biological structures and systems, such as brains, always involves some degree of simplification. In theoretical neuroscience there is no general agreement on the degree of simplification that is appropriate. It may depend on the taste of the modeler, on the amount of details of the experimental observations that are available on a particular functional or anatomical part of a particular brain, or mostly, on the animal or the function to be modeled

[1] Corresponding author.

(e.g. the bending of the worm, the flying of the fly, the fly-catching of the frog, the maze-running of the rat or the poem-writing of the human).

Usually, the unit of modeling is the single neuron, but it may be more microscopical (a small patch of neural membrane, a single synapse, an ion-channel) or more macroscopic (a column of cortical neurons that are all supposed to behave in a similar way, an area of the cortex or a nucleus of the brain).

We are dealing here with models on the level of the single neuron, and perhaps on a scale that is slightly below or above. If one wants to model interesting behavioral achievements of higher mammals, or even humans, these usually involve more than half of the brain and in the order of 10^9 neurons (at least). In such models, it may seem desirable to use larger units for modeling than the single neuron. In experimental recordings from single neurons one may get the impression that their responses to the same repeated stimuli are rather sluggish, imprecise or "random". For this reason experimentalists use time averaging for the characterization of neural responses and for some time it has been believed that the brain may use space averaging across small populations of neurons with a similar response, instead of time averaging, in order to pass the signal to the next level. This idea has led to the modeling unit of a "column" or a small group of neurons, and to the description of the unit's response in terms of firing rates instead of single spikes [41]. However, there is evidence that this kind of "mean field approximation" is not valid for behaving brains i.e. in awake animals. In fact, the exact timing of single spikes of individual neurons (for example, their response delay relative to the stimulus onset) can be used effectively for a number of interesting tasks [8, 12, 22, 23].

4.2
Voltage-Based Models

To begin with, we review some of the minimal properties of a model that describe the time evolution of a cell membrane potential (a so-called *voltage-based model*). To this aim, we first ought to recall the structure of the components of a neural network – the individual neurons. Subsequently, one can try to understand how neurons communicate with each other – a task which has been extensively pursued, but, almost a century after the first seminal investigations on synapses, still not yet fully accomplished.

In the present overview we do not aim for a complete physiological description of neurons and brain functions. An enormous number of books and research papers have already been devoted to this subject, see e.g. [29] for a beautiful historical overview. We will only point out some structural properties that are important for the mathematical modeling of neurons and neural networks.

As one can see in Figure 4.1, neurons are complex ramified structures that share some basic features. Several linear elements, the *dendrites*, compose the *dendritic tree*, that converges into the cell body, the *soma*. The tiny fiber leaving the soma is called the *axon*. Neural activity in these structures is represented by the voltage fluctuation (of usually less than 100 mV) across the membrane of the neuron. All these

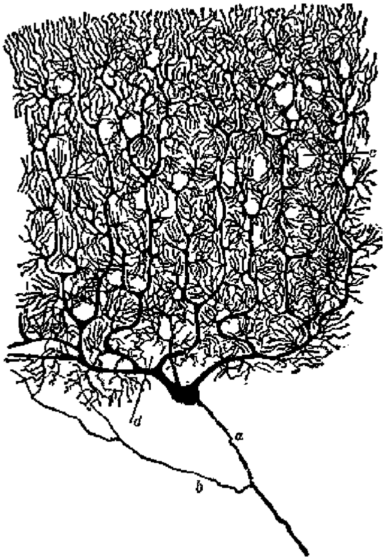


Figure 4.1 Drawing of a Purkinje cell in a pigeon's cerebellum by Santiago Ramón y Cajal, pioneer of neuroscience.

components possess transmembrane channels, that are responsible for the separation of ions. This results in a different concentration of ions inside the membrane compared to outside.

In the case of so-called *passive fibers* i.e. if the transmembrane channels display a behavior that is independent of the voltage, it is possible to derive differential equations for the time evolution of the transmembrane voltage in a formal way. A thorough discussion of this derivation can be found in [24].

While such dynamic laws were often studied on ideal fibers of infinite length, determining correct boundary conditions is crucial whenever real fibers of finite length are considered. The correct formulation of boundary conditions in each branching point of the dendritic tree, as well as in the soma, is W. Rall's main theoretical legacy. In a series of classical papers beginning with [33, 34] he suggested a model for the transmembrane voltage v of a whole dendritic tree.

Dendritic trees satisfying Rall's strong geometric assumptions are usually referred to as *equivalent cylinders*: they are represented as simple linear structures whose behavior is governed by the *cable equation*[2]

$$\dot{v}(t,x) = v''(t,x) - v(t,x) ,$$

where $v(t,x)$ denotes the membrane potential at time t and position x.

Even if equivalent cylinders are only fictitious objects, Rall showed that their time evolution is tightly related to that of the actual dendritic tree. Several attempts have

[2] Here and in the following we denote by $\dot{v}$ and v' the derivatives of the function v representing the cellular potential with respect to the time and spatial variables, respectively.

been made in order to extend Rall's ideas and his equivalent cylinder representation to the case of a more extended and heterogeneously branched dendritic tree: see e.g. [13, 25].

The investigation of *excitable* nerve fibers like axons is technically more demanding. The presence of different voltage-dependent transmembrane channels conducting specific ions – such as sodium, potassium, or calcium – is a distinctive feature of axons. Such *voltage-gated channels* add considerable difficulties to the study of biological neural networks, since the resulting system of partial differential equations is nonlinear. In most cases, it can only be treated numerically, in order to show that it correctly reproduces the dynamics observed in the experiments. It was A.L. Hodgkin and A.F. Huxley who first performed the experimental work leading to the formulation of a complete model for the most important individual channels, [17], a work for which they were awarded the Nobel Prize in Medicine in 1963. Their thorough, heuristic approach is still the main basis for research activity in theoretical neurobiology.

Up to renormalization, Rall's linear cable model for dendrites states the transmembrane voltage v of the nerve fiber at time t and point x of the dendritic tree (or rather, of its equivalent cylinder of length ℓ_1) satisfies

$$\left\{\begin{array}{rcl} \dot{v}_d(t,x) &=& v_d''(t,x) - v_d(t,x) \\ v_d'(t,0) &=& 0 \\ \dot{v}_d(t,\ell_1) &=& -v_d'(t,\ell_1), \end{array}\right. \tag{LC}$$

where the time-dependent boundary condition in ℓ_1 describes the leaky integration properties of the neuron (the second equations simply prescribe that the other end of the dendritic tree is sealed).

Rall's equations are essentially a specialization of a general Hodgkin–Huxley-like (H–H in the following) model of the form (see part 3 in [17])

$$\left\{\begin{array}{rcl} \dot{v}_a(t,x) &=& v_a''(t,x) + F(v_a(t,x)) - f(r(t,x)) \\ \dot{r}(t,x) &=& g_1(v_a(t,x)) + g_2(r(t,x)), \end{array}\right. \tag{HH}$$

which is formulated for biological structures of infinite length, therefor not equipped with boundary conditions. Without going into detail, r denotes in equation (HH) a generic *activation variable* (i.e. an $\mathbb{R}^n$-valued function) that models the activation of the voltage-gated channels (e.g. the dynamics of sodium and potassium ions); $g_2(r)$ represent the internal dynamics of the activation variables; $g_1(v_a)$ and $f(r)$ are *ad hoc* terms that account for the voltage-recovery feedback. Finally, $F(v_a)$ models a nonlinear threshold mechanism that lets the voltage converge asymptotically toward the experimentally observed peak potential value of approximately 40 mV during the transmission of *action potentials* (a.k.a. *spikes*, i.e., of traveling peaks of electrical discharge) along axons, and toward the resting value of approximately –70 mV after the action potential has moved forward. As we will see later, action potentials are the method used by neurons to mutually communicate across long distances.

In fact, it is possible to formulate boundary conditions satisfied by the voltage in the soma, and consequently to formulate a biologically realistic model of a whole neuron. We avoid going into detail and refer to [5, 9] for more detail, where a combination of (LC) and (HH) has been discussed.

As should be clear from the above anatomical description, neurons are not – biologically speaking – pointwise, dimensionless structures. On the contrary, their spatial structure plays a key role in the delays associated with the transmission of synaptic signals, as well as in synchronization and other nonlinear phenomena. In order to study realistic neural networks, it is necessary to understand computational features of single neurons. For that, the spatial structure of the latter cannot be neglected, thus leading to models similar to (LC) and (NC). This has already been recognised by Hodgkin and Huxley in [17].

Due to the great complexity of the nonlinear terms involved and of the dendritic geometries, it is simply not possible to solve explicitly hybrid models that feature Rall's and Hodgkin–Huxley-like dynamics. A possible, more realistic, approach is to study qualitative properties of these mathematical models by means of abstract tools from operator theory, linear algebra, and graph theory. This approach has been followed in [9], where a theoretical explanation of several experimental observations has been obtained. The main results state that:

- neuronal activity does not converge to an equilibrium point,
- solutions regularly transcend voltage thresholds set by initial conditions (a consequence of the lack of so-called L^∞-contractivity of the system), and
- neurons hyperpolarize during the undershoot period following an action potential (possibly because the system is not positivity-preserving).

The mathematical terminology and background have been discussed in detail in [9].

A further feature of a system of the type (HH) that can be investigated by mathematical methods is the existence of *traveling waves*. In fact, several experiments indicate that action potentials spread along the axon with constant velocity. Moreover, the shape of the action potential does not change during the transmission. A common *Ansatz* in the mathematical neuroscience is therefore that solutions $v(t, x)$ and $r(t, x)$ of (HH) satisfy

$$v(t, x) = v(x - ct) \qquad \text{and} \qquad r(t, x) = r(x - ct)$$

for some transmission velocity c. Such solutions are called *traveling waves* and they are a relevant research subject, both in the mathematical analysis of network equations (see [6]) and in the theoretical neuroscience (see e.g. [31]). It can be said that if the system (HH) correctly describes the spread of an action potential in an excitable fiber (i.e., if the phenomenological functions F, f, g_1, g_2 properly fit the experimental data), then it should be possible to find traveling wave solutions to (HH).

In the above discussion we have not even tried to model the interaction between two neurons. The reason for this is the extremely complex behavior of *synapses*, the junction points of two different neurons. Synapses are highly nonlinear objects. They can be classified in several groups that display different behavior, and a description of their activity is often only possible at a statistical level. This is why very few biologically realistic voltage-based models that also feature synaptical junctions have been proposed in the literature. However, the formulation and mathematical analysis of mixed problems implementing both the synaptic and neuronal level of information processing is necessary, thus justifying the use of theoretical models on the level of neuron populations. In the next section we discuss the heuristic motivations leading to the introduction of such simplified models of single neurons.

4.3 Changing Paradigm – From Biological Networks of Neurons to Artificial Neural Networks

Although most researchers nowadays agree with the *Ansatz* leading to models of Hodgkin–Huxley type (and, to some extent, also of Rall type), the model is not satisfactory due to its lack of simulation efficiency. The complexity and instability of the nonlinear HH-equations have motivated researchers to propose different approaches with the aim of simplifying the modeling and at the same time of facilitating neurophysiological tests. Morphological details have been neglected over and over again, eventually performing cortical representations based on lattices of point-like neurons, where functional investigation has been favored over anatomical realism.

Computational methods could be derived from the models described in the previous section by solving the equations presented in Section 4.2 numerically and then analyzing their behavior. However, numerical analysis of a biologically realistic neural network is hardly feasible, due both to the large size of cortical populations and to the typical lack of regularity exhibited by solutions of nonlinear equations, thus making such an approach quite inconvenient.

A first approach to the derivation of *efficient* computational methods from the models described in the previous section could be to coarsely discretize space parameters that appear in the equations of Section 4.2 and then try to formulate equivalent equations on this discretized space. This has led to the introduction of so-called *compartmental models*, where neurons are artificially disassembled in comparatively few (two up to several hundred) sections or compartments, to be analyzed by means of numerical methods. However, this requires extremely elaborate data-fitting for each compartment. Moreover, in order to obtain reasonably precise results one needs a large number of compartments, which in turn leads to extremely high computational costs.

Despite experimental measurements which suggest that a manifold of nonlinear operations may be performed by dendrites, in order to further reduce complexity of the models it is commonly assumed that postsynaptic potentials add linearly within the dendritic tree. Thus, synapses and dendritic trees are ideally collapsed into dimensionless structures whose only function is to transform presynaptic activity into an input that will be processed by the soma or axon hill (this simplification is also known as a point neuron model,). As an additional simplification, somatic output is then assumed to be further transmitted to proximal computational units with a fixed velocity and amplitude that are characteristic for each couple of connected neurons. Thus, signal processing in axonal trees is neglected for both computational and modeling purposes: the whole neuron has been ideally shrunken to a single soma whose dynamics is described by a system of ordinary differential equations for the different ion channels. This class of simplified model may be referred to as the *multiple-channel models*.

With the aim of further simplifying the model, a coarse quantitative analysis of the system can be performed if one neglects the differences in dynamics of individual channels and only considers the time evolution of the membrane potential as in so-called *leaky integrate and fire (LIF) models*. In spite of their apparent lack of biophysical accuracy, such models possess a competitive computational edge. They permit a thorough functional analysis of information processing in dependence on the connectivity implemented in the architecture of the neural network. The brain's well-documented robustness against noise, transmission failures, and further distorsions suggests that these simplifications may be admissible in the simulation of large networks.

Models based on such *networks of spiking neurons* are nowadays quite common in neuroscience. We will thoroughly discuss them in the next section. For a mathematically more precise reduction of Hodgkin–Huxley-type equations to computationally simple models, see e.g. [1].

Numerical simulations of neural networks are useful in many areas of brain research. While it is not (yet) possible to create truly intelligent machines, simulations already allow one, for example, to verify theoretical models by comparing their performance with measurements in real neural networks. Numerical simulation is often used to verify the performance of simplified models (e.g. the leaky integrate and fire model, see [21]).

The simulation of neural networks also has a background in computer science, as artificial neural networks are often able to efficiently solve certain kinds of problems. For example, artificial neural networks typically use very simplified, technical, and biologically unrealistic neuron models and only allow for very simple connectivity possibilities (e.g. only binary synapses). Typical problems that are solved by artificial neural networks come from the area of pattern analysis, like visual and auditive object recognition. Recent examples include recognizing music samples or cheating CAPTCHA-tests [10]. The property of *generalization* and the easy possibility of learning from large example databases makes them ideally suited to solve such tasks. Other examples are the management of gearboxes in automotive engineering, robot motion planning, data mining, financial prediction, marketing, and recently very successful predictions in the modern energy market.

4.4 Numerical Simulation of Neural Networks

Typically, Hodgkin–Huxley-type neurons are used for biologically plausible simulations. These kinds of models however are very complex to simulate and it requires exceptionally fast high-end machines in order to simulate networks the size of small mammalian brains and, even then, simplified neuron models have to be used or the size of the network has to be reduced [27]. Faced with such a huge amount of computational requirements, one obviously wants to use simpler models that can be efficiently simulated: we mention e.g. the widely known and accepted integrate-and-fire-type neuron with different modifications (e.g. [21]), the Spike Response model ([14]) or the recently introduced model of E. Izhikevich (see [19] for an overview or [18] for full details).

All these approaches share the property of trying to reduce the computational cost of the simulation while still aiming to be as biologically realistic as possible. This means that as many physiological properties of the single neurons *and* as many functional properties of networks of neurons as possible should be retained. Of course, some tradeoffs have to be made, but even with rather strongly simplified models, like the integrate-and-fire-type with adaptation, very good results can be achieved (see e.g. [21]). However, the goal of such models is usually to reproduce or predict the activity of biological neural networks. An alternative approach is to use neural networks to solve certain tasks or problems in an elegant way. These kinds of models focus on functional properties of larger networks of neurons much more than on the physiological features of single neurons, which are sacrificed in favor of faster computation. Examples for such models are Hopfield-type networks, Willshaw's associative memory [42], the perceptron [35] and the multi-layer perceptron [36], Radial Basis Function networks [32, 37], the SpikeNet approach [12], etc.

Already, from the above mentioned examples, it becomes clear that it is hard to draw a sharp line of separation between technical (or artificial) neural networks and biologically realistic networks. There are models available to almost any degree of biological realism, ranging from very technical to almost natural.

Usually for simulations of larger networks at least multiple-channel models as introduced in Section 4.3 are used. We now introduce such a model explicitly and propose a cascade of simplifications that can be performed in order to achieve enhanced computational efficiency. A set of N interconnected neurons is considered. The synaptic couplings are represented in a coupling matrix (see (4.3) below).

In the multiple channel model, neurons are completely classified according to the biophysical properties of their membranes (decay rate of excited fibers and duration of refractory time) as well as to the kind of effect on postsynaptic neurons. This classification results in a grouping of neurons labeled by indices $l \in L$ and $k \in K$, respectively.

Let u_j be the membrane potential of the j-th neuron of type $l(j) \in L$. For the sake of notational simplicity, the dependence on j is dropped and we write $l := l(j)$ in the following. Then, the membrane potentials dynamic is given by

$$\tau_l \cdot \dot{u}_j(t) = -u_j(t) + \sum_{k \in K} a_j^k(t) \left(U_k - u_j(t) \right) + x_j(t) \; . \tag{4.1}$$

Here a_j^k is the total input activity for the neuron j propagated through the input channel type k at time t, the term $\tau_l > 0$ is a time constant, and the value U_k is called the *reversal potential* of the channel $k \in K$ i.e. the potential where the effect of the channel $k \in K$ is zero. Crossing this potential reverses the effect of the channel.

Additional external (sensory) input from outside the simulated neurons enters the network in the variable $x_j(t)$.

Because of the linear input addition property, which we discussed before, the total input activity of channel k is given by

$$a_j^k(t) = \sum_{i \in I_k} a_{ij}(t) \ , \tag{4.2}$$

where I_k is the set of all indices of neurons belonging to channel type $k \in K$ and $a_{ij}(t)$ is the activity transferred from neuron i to neuron j. For a neuron i of class $k \in K$, this is given by

$$a_{ij}(t) = \sum_{s \in T^i} r_k(t - (s + d_{ij}))c_{ij} \ . \tag{4.3}$$

Here, $r_k : \mathbb{R}^+ \to \mathbb{R}^+$ is the response function of the channel type k i.e. the form of the effective postsynaptic potential. The term $r_k(t)$ models the effect that all synaptic, axonal and dendritic activity has on the postsynaptic neuron for an input of channel type k at time t when the presynaptic spike occurred at time $s = 0$. The parameter c_{ij} is the synaptic coupling strength from neuron i and neuron j, d_{ij} is the (non-negative) delay of the signal transfer from neuron i to neuron j.

The output activity of the j-th neuron is given by

$$y_j(t) = \mathbb{1}_{[u_j(t) \geq \theta_j(t)]} \ , \tag{4.4}$$

where the threshold function $\theta_j : \mathbb{R} \to \mathbb{R}$ for the neuron j of membrane type l is defined by

$$\theta_j(t) = \vartheta_l(t - s_j^*) = \max_{s \in T_t^j} \vartheta_l(t - s) \ . \tag{4.5}$$

Here, $T_t^j = \{s < t : y_j(s) = 1\}$ is the set of spike times for the neuron j, $s_j^* = \max T_s^j$ denotes the time where the last spike occurred and $\vartheta_l : \mathbb{R}^+ \to \mathbb{R}^+$ is the monotonically decreasing threshold function for the neuron type with index l. The threshold function $\vartheta_l(t)$ adjusts the firing threshold of the neurons of membrane type l and gives the value of the threshold if the last spike of the neuron was at time t in the past. In order to model an absolute refractory period during which spikes are very unlikely and almost impossible, $\vartheta_l(t - s_j^*)$ is set to a very high value whenever $t - s_j^* < \Delta_l$, where $\Delta_l > 0$ is the duration of the absolute refractory period. For larger values of $t - s_j^*$, the threshold decreases towards its resting value, modeling a relative refractory period.

Some of the equations of this model can be further simplified, resulting in several commonly known models which will be summarized below in Table 4.1.

First, in (4.2), the spike response functions r_k can be simplified to a total effect $w_{ij} \in \mathbb{R}$ for the connection from neuron i to neuron j, giving

$$a_j^k(t) = \sum_{i \in I_k} w_{ij} \cdot y_i(t - d_{ij}) \ . \tag{4.6}$$

Again, I_k is the index set of all neurons belonging to channel type k.

Another simplification is given by replacing the reversal potential term $U_k - u_j(t)$ in (4.1) by U_k, because the membrane potential $u_j(t)$ is small compared to U_k most of the time. This yields

$$\tau_l \cdot \dot{u}_j(t) = -u_j(t) + \sum_{k \in K} a_j^k(t) \cdot U_k + x_j(t) \, . \tag{4.7}$$

Here, it is possible to use either definition of $a_j^k(t)$. If the simplified variant from (4.6) is used, the membrane potential's equation reads

$$\tau_l \cdot \dot{u}_j(t) = -u_j(t) + \sum_{i=1}^{N} w_{ij} \cdot y_i(t - d_{ij}) + x_j(t) \, . \tag{4.8}$$

Note that the sum now is over all neurons, instead of being over all channel types. Further, the calculation of the total input activity per channel is not necessary because this simplification completely neglects channels.

Sometimes, an explicit reset of the membrane potential of a neuron j of class l is performed allowing

$$u_j(t) \leftarrow 0 \quad \text{if} \quad u_j(t) \geq \theta_l(t) \quad , \tag{4.9}$$

i.e. the membrane potential is forcefully set to zero whenever the threshold $\theta_l(t) \in \mathbb{R}^+$ is reached. In many models with reset, a constant threshold is chosen, yielding

$$\theta_l(t) = \vartheta_l \, . \tag{4.10}$$

Let us briefly mention the further class of so-called *firing-rate* models, based on the assumption that the output function y in (4.4) satisfies the simplistic dynamic law

$$y_j(t) = f_l(u_j(t)) \, , \tag{4.11}$$

where $f_l : \mathbb{R} \to \mathbb{R}$ is typically a sigmoidal function for the neuron type $l \in L$, describing the *spike rate* of the neuron. A variation of this simple model is sometimes used to generate quite realistic spike trains: in the so-called varying spike-rate *Poisson*

Table 4.1 Common neuron models in neuroscience and the corresponding equations in the text above.

Model name	Potential u	Input a_j^k	Output y	Threshold θ_j
Multiple-channel model	4.1	4.2	4.4	4.5
Dynamic threshold model	4.1	4.6	4.4	4.5
Simple dynamic threshold model	4.7	4.6	4.4	4.5
Spike response model	4.7 + 4.9	4.2	4.4	4.5
Integrate-and-fire model	4.8 + 4.9		4.4	4.10
Firing-rate model	4.1	4.2	4.11	
Simple firing-rate model	4.8		4.11	

model the output value $y_j(t)$ in (4.11) is used as the spike- or event-rate in a Poisson process that models the output of the neuron [3, 4].

Table 4.1 shows which combination of equations results in which commonly known model in the neurosimulation literature. In fact this literature is distributed across several disciplines and much more chaotic than this simple table suggests. Many additional small varieties and all sorts of different names are used for the models (e.g. [15,19,24,39]). Clearly, the model variants can behave differently under certain circumstances. Compared with the simple firing rate model, a spiking model can show qualitatively different dynamics under otherwise identical conditions. Wennekers and Palm [40, 41] have created a simple architecture in which a population of 1024 identically and fully connected neurons can be simulated with three different models (firing rate model, Poisson model and simple dynamic threshold model) that have the same input-output behavior on average or in the mean field

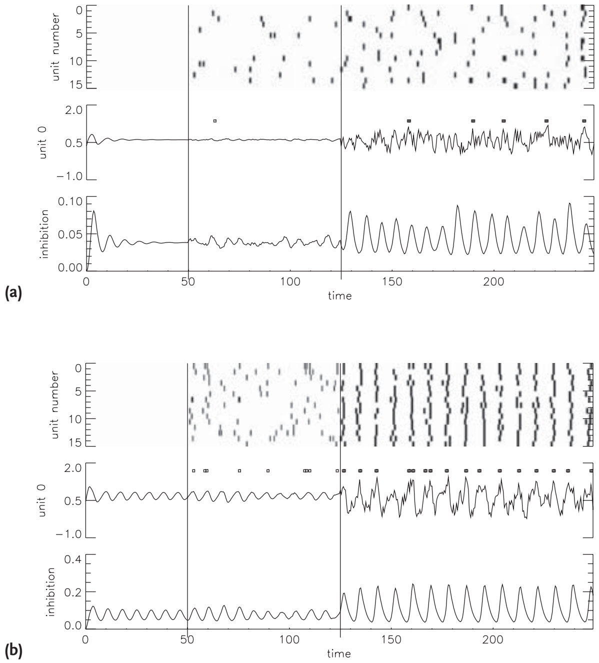


Figure 4.2 Qualitative behavior of the three single-neuron variants described in the main text in a fully connected excitatory network with inhibitory activation control. In (a) and (b) firing rate model neurons (type I) are used during the initial 50 steps, from 50–150 Poisson model neurons (type II), and afterwards simple dynamic threshold neurons (type III). The upper plots in (a) and (b) reveal spikes (black dots) of 16 of the 1024 simulated cells. The middle plot shows the membrane potential and spikes (dots) of one individual cell. At the bottom the time-course of the inhibitory neuron, reflecting the average activity of the whole population, is seen. In (a) the activity of type I and II neurons is stationary (up to fluctuations in the membrane potentials), whereas for type III neurons it is oscillatory. For higher external excitation as in (b) also networks with type I and II cells oscillate. Now type III shows a different frequency and shape. Using plausible parameter values these oscillations fall into the EEG-gamma range.

approximation. One result is shown in Figure 4.2, where the simulation switches between the three different models revealing qualitatively different dynamics.

Consider the cases depicted in (a) and (b). Note, that the temporally and spatially constant mean value of noisy input to the model neurons is increased in (b). One observes a breakdown of the mean field approximation due to emerging oscillations in (a) and due to almost complete spike synchronization in (b).

4.5
Population-Based Simulation of Large Spiking Networks

A relevant speedup of the simulation can be achieved by partitioning the neural network into subnetworks that contain several neurons and share the same connectivity and functionality. If the whole network is decomposed into "building blocks" of this kind, a structural similarity *Ansatz* can be deployed in order to make the simulation more efficient. Whenever populations of spiking neuron models are used, the simulation is still carried on the level of single neurons, i.e., the model keeps track of the precise spike timing of the basic units.[3)]

A very simple form of neural networks of spiking neurons is the binary neuron model of Willshaw [42], which is a suitable building block for the above approach. Although the model is very technical and the neurons and connections are only binary, there is a strong relationship with brain research. Many models of the human brain suggest that associative memory structures, which feature feedback in an essential way, are massively present throughout the neocortex. Examples of such theories reach back to the 1980s (see [30]) but can be found also in more recent research, although somewhat hidden in the explanation (see [16]). Furthermore, associative memories even in the simple binary form reflect the process of *synaptic plasticity*, a concept that is basically used to model changing synaptic weights in the neural network. Associative memories learn patterns by adjusting synaptic weights, even in the binary model, where synaptic weights are adjusted between the two possibilities zero and one using a simplified *Hebbian learning rule* i.e. the connection is strengthened (set to one) when the pre- and postsynaptic neuron are active at almost the same time.

Autoassociative memories can be considered as memories that can be addressed with erroneous versions of stored patterns and, in the ideal case, respond with the completed, error-corrected pattern. Willshaw's model of autoassociative memory implements such a device using binary neurons, binary synapses and a huge amount of feedback connections. We will not describe Willshaw's model in detail, for which we refer to Willshaw's original work [42]. Instead, the so called "spike counter" model will be introduced. In its very basic form, it is a slightly modified

[3)] This is in contrast to firing rate models, which study the behavior of whole populations under the assumption that they can be described by a joint firing rate (i.e. detailed information about spike timing of single neurons is neglected) as we have shown in Figure 4.2.

variant of the Willshaw model: the external input of a neuron influences the time evolution of its membrane potential in a linear fashion instead of directly adding to the potential itself as in the Willshaw model.

The spike counter model is a model for a population of neurons that work together to build one autoassociative memory. These memories can then be connected to other memories modeled in the same manner in order to construct a large neural network. Each memory makes heavy use of local feedback connections, setting the spike counter model apart from layered networks such as the perceptron which has one dedicated direction for information flow. The spike counter model is completely characterized by its pattern storage and retrieval features.

Pattern storage works as in Willshaw's binary model of associative memory. For a population of N neurons, consider the set of vectors

$$W = \{w^1, w^2, \dots, w^K\} . \tag{4.12}$$

These vectors represent patterns that should be stored in the system. Note that the vectors need to be binary and of length N i.e.

$$w^i \in \{0, 1\}^N \quad \forall i \in \{1, \dots, K\} . \tag{4.13}$$

Optimal patterns for the spike counter model are sparse i.e. $\sum_{j=1}^{N} w_j^i$ is small for all i, and of equal "size" i.e.

$$\sum_{j=1}^{N} w_j^i = \sum_{j=1}^{N} w_j^l \quad \text{for all} \quad i, l \in \{1, \dots, K\} .$$

The patterns are stored in a memory matrix $A \in \{0, 1\}^{N \times N}$ by

$$A = \bigvee_{i=1}^{k} w^i \otimes (w^i)^\top \tag{4.14}$$

where $\vee$ means the entrywise maximum operation and $\otimes$ the outer product of two vectors with N components.

For a population consisting of N neurons, the membrane potential of the i-th neuron is given by

$$\begin{cases} \dot{x}_i(t) &= a \cdot c_i^H(t) + b \cdot \left(c_i^A(t) - \alpha \cdot c^\Sigma\right), \\ x_i(0) &= 0, \end{cases} \tag{4.15}$$

where $\alpha \in [0, 1]$, $a, b \in \mathbb{R}$, $a, b \geq 0$. Typically, α is close to one and a is much smaller than b e.g. $a = 0.005$ and $b = 1$. The values c^H, c^A, and c^Σ are the "spike counter" variables:

c^A: It holds an individual value for each neuron in the population, i.e. $c^A \in \mathbb{R}^N$. For each neuron, it counts how many autoassociative feedback spikes the neuron has received during the current retrieval. The value of c^A changes during the retrieval i.e. it increases by one if a feedback spike is retrieved for a neuron.

c^{Σ}: It is constant for all neurons within the population i.e. $c^{\Sigma} \in \mathbb{R}$. It counts the total number of spikes that have already been fired during the current retrieval step. The value of c^{Σ} changes during the retrieval i.e., it increases by one with each occurring spike.

c^{H}: It holds an individual value for each neuron, i.e. $c^{H} \in \mathbb{R}^{N}$. It counts the number of external input spikes that a neuron receives during the retrieval. In the basic spike counter model, this value is constant throughout the retrieval.

Define the output Z_i of the i-th neuron by

$$Z_i(T) = \begin{cases} 1 & \text{if neuron } i \text{ fired before time } T, \\ 0 & \text{if neuron } i \text{ did not fire until time } T. \end{cases} \tag{4.16}$$

The autoassociative spike counter is given by

$$c_i^A(t) = \sum_{j=1}^{N} A_{ij} \cdot Z_j(t - d_{ij}) \ , \tag{4.17}$$

where $d_{ij} \geq 0$ is the delay of the autoassociation from neuron i to neuron j and A_{ij} is the i, j entry of the matrix A.

The c^{Σ}-counter is defined as

$$c^{\Sigma}(t) = \sum_{j=1}^{N} Z_j(t) \ . \tag{4.18}$$

The external input counter c_i^H accounts for the strength of the external input to the population.

During the retrieval, the membrane potentials of neurons that receive external input start to grow linearly towards the firing threshold. If the first neuron has fired, the spike is fed back through the autoassociation and affects the neurons that are connected to the one that fired via c^A. At the same time, c^{Σ} counts how many spikes have occurred during the retrieval. If $\alpha \approx 1$ and several neurons have fired, the term $c_i^A - c^{\Sigma}$ becomes negative for a neuron that is not connected to the firing neurons. Hence, the neuron will be inhibited and potentially, depending on the value of b, will not fire at all. Neurons that are connected to the ones that have actually fired, and thus probably belong to the same assembly, will receive positive feedback through $c_i^A - \alpha \cdot c^{\Sigma}$, driving their membrane potential to the firing threshold.

For a thorough discussion of the (extended) spike counter model and how to apply it to the understanding of simple languages, see [26].

In order to build large networks consisting of several autoassociative memories, it is necessary to forward activation from one memory to the next. This can be achieved by heteroassociative connections. Heteroassociations map an input vector $w \in \{0, 1\}^M$ to an output vector $\tilde{w} \in \{0, 1\}^N$, corresponding to populations of M and N neurons, respectively, and are thus describes by a nonquadratic $M \times N$-matrix H.

A set of K pattern pairs $\left((w^1, \tilde{w}^1), \ldots, (w^K, \tilde{w}^K)\right)$ is stored using

$$H = \bigvee_{i=1}^{K} w^i \otimes (\tilde{w}^i)^\top \,, \tag{4.19}$$

where $\otimes$ is the outer product and $\bigvee$ denotes the entrywise maximum function. The retrieved pattern $\tilde{u} \in \{0,1\}^N$ with an address pattern $u \in \{0,1\}^M$ is given by

$$\tilde{u}_i = \Theta_{\sum_{j=1}^{M} u_j} \left(\left((u^\top \cdot H)^\top \right)_i \right) \,, \tag{4.20}$$

where

$$\Theta_t(x) = \begin{cases} 1 & \text{if } x \geq t \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{4.21}$$

is the threshold function. In the case of the spike counter networks, heteroassociations neglect spike time. The input pattern u is equivalent to the output Z of the whole retrieval i.e. $u = Z(T)$, where T is the duration of the retrieval. When heteroassociations are used to connect different populations of autoassociative memories, often the threshold function Θ is not applied in (4.20). This forwards all pattern activations unmodified into the autoassociative memory which is then responsible for the processing of the inputs using its own retrieval algorithm.

Heteroassociations can be used to translate between different representations that might be used in different autoassociative memories. If autoassociative memories are viewed as basic processing units, heteroassociations can be used as connections between different autoassociative units. We call an autoassociative memory, together with the heteroassociations responsible for its input, a *module*. Two of these modules can now be connected by connecting the output of one module to the heteroassociative input of another module and vice versa. Using this kind of architecture, it becomes possible to simulate rather large neural networks in a relatively short time.

Biological neurons cannot transmit signals in arbitrary short time intervals. Instead, a spike needs some time to be transmitted along an axon, to pass a synapse and travel down the dendrite until it reaches the next neuron. The time depends on several factors, including the lengths of the axon and dendrites and the type of synapse in between. This fact needs to be reflected in simulations of neural networks, both for the sake of biological realism and in order to exploit its functional properties. Usually, it is represented by "synaptic delays" i.e. the signal is delayed for a certain amount of time when it passes a synapse, whereas dendritic and axonal delays are often not explicitly modeled. Positive delays are helpful to allow for bidirectional but causal connections between two modules: without synaptic delays, the output of the first module would determine the output of the second module and, simultaneously, the second module would feed back to the first module and modify its output.

While delays are easy to implement in numeric simulations of neural networks, they are usually very hard to deal with in analytical models. Numerical simulations can thus at least be used to give a hint about the actual behavior of neuronal dynamics. To conclude, let us summarize what we have seen in the previous sections. The building blocks of a neuron are the dendritical tree, the soma, and the axon: each of these anatomical entities is associated with a functional activity both at a modeling and a computational level. On one hand, Rall's model of the dendritical tree is a linear partial differential equation, corresponding to time integration of inputs. Moreover in the model presented in Section 4.2 the soma is assigned an essential filtering role. If the gathered potential falls into the attraction basins of traveling-wave solutions of equations (HH), then a traveling wave is started on the axon. This behavior is represented in the threshold mechanisms discussed in Section 4.4. Since traveling waves spread on the axon with constant velocity, the signal transmission on the axon may be regarded as a purely linear transport. This corresponds to the delay terms included in the computational model.

4.6
Synaptic Plasticity and Developing Neural Networks

As we have seen, the biophysics of spike transmission is nowadays comparatively well-described and understood. On the contrary, much less is known about the way in which a real cortex evolves during an animal's learning phase.

In fact, many learning models have been introduced and are commonly used in the neuroscientific literature. While most approaches favor feasibility in artificial neural networks over biological realism, some rules aim at describing learning processes in actual cortical areas. They reflect an evolutionary behavior involving reversible weakening or strengthening of synaptic coupling between single neurons. These phenomena go under the name of synaptic plasticity.

We mention the set of heuristic rules that have been developed by the neuropsychologist D. Hebb in the 1940s, based on the assumption that strengths of synaptic connections can vary over time. Hebbian rules were briefly introduced in Section 4.4. They prescribe that simultaneous activity of pre- and postsynaptic neurons results in the strengthening of their synaptic connection.

More recently, the alternative mechanism of so-called Spike-Timing Dependent Plasticity (STDP) [2, 7, 20, 28, 38] has become increasingly popular in the neuroscience community. In contrast to the simple Hebbian paradigm, STDP allows for the weakening of a connection depending on the chronological order of neuronal activity: if activity in a postsynaptic neuron *follows* in rapid succession upon a spike in the presynaptic one, then the connection will be strengthened; if, conversely, activity in a postsynaptic neuron *precedes* a presynaptic spike, then the connection will be weakened. Observe that the biochemically natural direction of information flow of (chemical) synapses, which is respected in the framework of STDP, is not explicitly used in simple Hebbian rules. Implementing such plasticity rules in biologically realistic neural network models may simplify the study of asymptotics of such systems, leading to a better theoretical understanding of firing patterns in cortical networks that are commonly recorded in experiments.

In conclusion, we observe that the standard theory of synaptic plasticity allows the strength of mutual neuronal connections to vary, but assumes that the basic architecture of the neural network remains constant over time. While this considerably simplifies the analysis, it is in contrast to the observed behavior of neurons and synapses in the developing animal cortex, whose topology irregularly evolves during an animal's maturation phase, providing a steady *rewiring* of the cortical tissue. In recent years, this issue has aroused a great deal of attention in developmental neuroscience. However, as yet there is no convincing approach for a mathematical analysis of these phenomena. It is an open problem whether and how they can be put into a framework of dynamic graph theory, allowing theoretical investigations of rewiring phenomena.

References

1 Abbott, L.F., Kepler, T.B. (**1990**) Model Neurons: From Hodgkin–Huxley to Hopfield, in *Statistical Mechanics of Neural Networks*, (ed L. Garrido) Springer.

2 Abbott, L.F. and Nelson, S.B. (**2000**) Synaptic plasticity: taming the beast, *Nat. Neurosci.*, **3**(Suppl), 1178–1183.

3 Aertsen, A., Erb, M. and Palm, G. (**1994**) Dynamics of functional coupling in the cerebral cortex: an attempt at a model-based interpretation, *Physica D*, **75**, 103–128.

4 Erb, M., Aertsen, A., in *Information Processing in the Cortex: Experiments and theory*, (eds A. Aertsen, V. Braitenberg) Springer, 1992.

5 Baerm, S.M. and Tier, C. (**1986**) An analysis of a dendritic neuron model with an active membrane site, *J. Math. Biol.*, **23**, 137–161.

6 von Below, J. Front propagation in diffusion problems on trees, in *Calculus of variations, applications and computations*, (eds C. Bandle *et al.*) Pitman Res. Notes Math. Ser. **326**, Longman Scientific & Technical, Harlow, 254–265.

7 Bi, G.Q. and Poo, M.M. (**1998**) Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type, *Journal of Neuroscience*, **18**, 10464–10472.

8 Borst, M., Knoblauch, A., Palm, G. (**2004**) Modelling the Auditory System: Preprocessing and Associative Memories Using Spiking Neurons, *Neurocomputing*, **58–60**, 1012–1018.

9 Cardanobile, S. and Mugnolo, D. (**2007**) Analysis of a FitzHugh–Nagumo–Rall model of a neuronal network, *Math. Meth. Appl. Sci.*, **30**, 2281–2308.

10 Chellapilla, K. and Simard, P. (**2004**) Using Machine Learning to Break Visual Human Interaction Proofs (HIPs), *Advances in Neural Information Processing Systems*, **17** (NIPS2004), MIT Press.

11 Deleuze, G. and Guattari, F. (**1987**) *A Thousand Plateaus*, University of Minnesota Press, Minneapolis.

12 Delorme, A., Gautrais, J., VanRullen, R., and Thorpe, S.J. (**1999**) SpikeNET: A simulator for modeling large networks of integrate and fire neurons, *NeuroComputing*, **26–27**, 989–996.

13 Evans, J.D. (**2000**) Analysis of a multiple equivalent cylinder model with generalized taper, IMA, *J. Math. Appl. Medic. Biol.*, **17**, 347–377.

14 Gerstner, W. (**2001**) A framework for spiking neuron models: the spike response model, in *The Handbook of Biological Physics, Vol. 4*, (eds F. Moss and S. Gielen) Elsevier, New York, pp. 469–516.

15 Gerstner, W. (**1999**) Spiking Neurons, in *Pulsed Neural Networks*, (eds W. Maass, C. Bishop) MIT Press.

16 Hawkins, J. (**2004**) On Intelligence, *Times Books*, New York.

17 Hodgkin, A.L. and Huxley, A.F. (**1952**) A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.*, **117**, 500–544.

18 Izhikevich, E.M. (**2007**) *Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting*, MIT Press, Cambridge.

19 Izhikevich, E.M. (**2004**) Which model to use for cortical spiking neurons?, *IEEE Transactions on Neural Networks*, **15**, 1063–1070.

20 Izhikevich, E.M. and Desai, N.S. (**2000**) Relating STDP to BCM, *Neural Computation*, **15**, 1511–1523.

21 Jolivet, R., Rauch, A., Lüscher, H.R. and Gerstner, W. (**2006**) Integrate-and-fire models with adaptation are good enough, *Advances in Neural Information Processing Systems*, **18**, 595–602.

22 Knoblauch, A. and Palm, G. (**2001**) Spiking Associative Memory and Scene Segmentation by Synchronization of Cortical Activity, in *Emerging Neural Architectures Based on Neuroscience*, (eds S. Wermter, J. Austin, D. Willshaw) Springer.

23 Knoblauch, A. and Palm, G. (**2001**) Pattern Separation and Synchronization in Spiking Associative Memories and Visual Areas, *Neural Networks*, **14**, 763–780.

24 Koch, C. (**1999**) Biophysics of Computation: Information Processing in Single Neurons, *Oxford University Press*, New York.

25 Major, G., Evans, J.D. and Jack, J.J. (**1993**) Solutions for transients in arbitrarily branching cables: I. Voltage recording with a somatic shunt, *Biophys. J.*, **65**, 423–449.

26 Markert, H., Knoblauch, A. and Palm, G. (**2007**) Modelling of syntactical processing in the cortex, *BioSystems*, **89**, 300–315.

27 Markram, H. (**2006**) The blue brain project, *Nat. Rev. Neurosci.* **7**, 153–160.

28 Markram, H., Lubke, J., Frotscher, M. and Sakmann, B. (**1997**) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs, *Science* **275**, 213–215.

29 Meunier, C. and Segev, I. (**2001**) Neurons as physical objects: dynamics and function, in *Neuro-Informatics and Neuronal Modelling*, (eds F. Moss and S. Gielen) North-Holland, Amsterdam, 353–467.

30 Palm, G. and Shaw, G.L. (**1988**) Brain Theory, *World Scientific*, Singapore.

31 Pinto, D.J. and Ermentrout, G.B. (**2001**) Spatially structured activity in synaptically coupled neuronal networks: I. Travelling fronts and pulses, *SIAM Journal on Applied Mathematics*, **62**, 206–225.

32 Poggio, T. and Girosi, F. (**1990**) Networks for approximation and learning, *Proceedings of the IEEE*, **78**(9), 1481–1497.

33 Rall, W. (**1959**) Branching dendritic trees and motoneurone membrane resistivity, *Exp. Neurol.*, **1**, 491–527.

34 Rall, W. (**1960**) Membrane potential transients and membrane time constant of motoneurons, *Exp. Neurol.*, **2**, 503–532.

35 Rosenblatt, F. (**1958**) The perceptron: a probabilistic model for information storage and organization in the brain, *Psychol. Rev.*, **65**(6), 386–408.

36 Rumelhart, D.E. (**1986**) Learning Internal Representations by Error Propagation, *MIT Press*, Cambridge.

37 Schwenker, F., Kestler, H. and Palm, G. (**2001**) Three learning phases for radial-basis-function networks, *Neural Networks*, **14**, 439–458.

38 Sjostrom, P.J., Turrigiano, G.G. and Nelson, S.B. (**2001**) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity, *Neuron*, **32**, 1149–1164.

39 Stein, R.B. (**1967**) Some models of neuronal variability, *Biophysical Journal* **7**, 37–68.

40 Wennekers, T., Sommer, F.T., Palm, G. (**1994**) Iterative retrieval in associative memories by threshold control of different neural models, in *Supercomputing in Brain Research: From Tomography to Neural Networks*, (eds Herrmann, H.J., Wolf, D.E., Pöppel, E.) World Scientific.

41 Wennekers, T., Palm, G. (**2007**) Modelling generic cognitive functions with operational Hebbian cell assemblies, in *Neuronal Network Research Horizons*, (ed Weiss, M.L.) Nova Science Publishers Inc.

42 Willshaw, D. (**1971**) *Models of Distributed Associative Memory*, PhD thesis, University of Edinburgh.

5
Boolean Networks for Modeling Gene Regulation

Christian Wawra, Michael Kühl, Hans A. Kestler[1]

5.1
Introduction

Cells of living organisms and their interactions are highly complex biochemical systems. By mutual interference genes build connected networks, called *gene regulatory networks*, to coordinate the behavior of cells. Using feedback mechanisms, transcription factors, which are a certain class of proteins, regulate the synthesis of other genes. Mathematical models of such networks can help one to understand these complex mechanisms that are responsible for all living organisms. Almost four decades ago Kauffman [17] studied random Boolean networks. His computer simulations revealed an analogy between these discrete artificial networks and the gene regulation of living organisms. He associated different types of cells to cycle states of Boolean networks. Table 5.1 shows the analogies on which these models are based.

Figure 5.1 shows the circuit diagram that represents the single Boolean functions and the corresponding state transition table of a simple network.

In terms of biological systems, Boolean networks are mainly applied in two ways. First, by building large artificial random networks to investigate the global behavior and to draw conclusions from these findings, second, through modeling of a certain gene regulatory network based on various measurements and observations to predict the behavior of the system.

This work is structured as follows. First we introduce the biological background and the goals of modeling in the field of biology. In Section 5.4 we discuss different modeling approaches with a focus on Boolean networks. Section 5.6 addresses seminal characteristics of large randomly constructed networks and outlines numerical as well as analytical results. Different algorithms to infer Boolean networks from measurements are explained in Section 5.7. In the end we discuss the approaches presented and give a short outlook.

[1] Corresponding author.

Table 5.1 Analogies between discrete random Boolean networks and Gene Regulatory systems.

Gene regulatory system	**Boolean network**
Genes that can be *expressed* or *not expressed*	Nodes x_i with one of the values 1 or 0
Promoter region that regulates the transcription of a gene dependent on the availability of transcription factors	Boolean function f_i which determines the next value for node x_i dependent on the current state.
Cell type	Attractor
Cell cycle length	Attractor cycle length

(See Section 5.2 and 5.5 for details.)

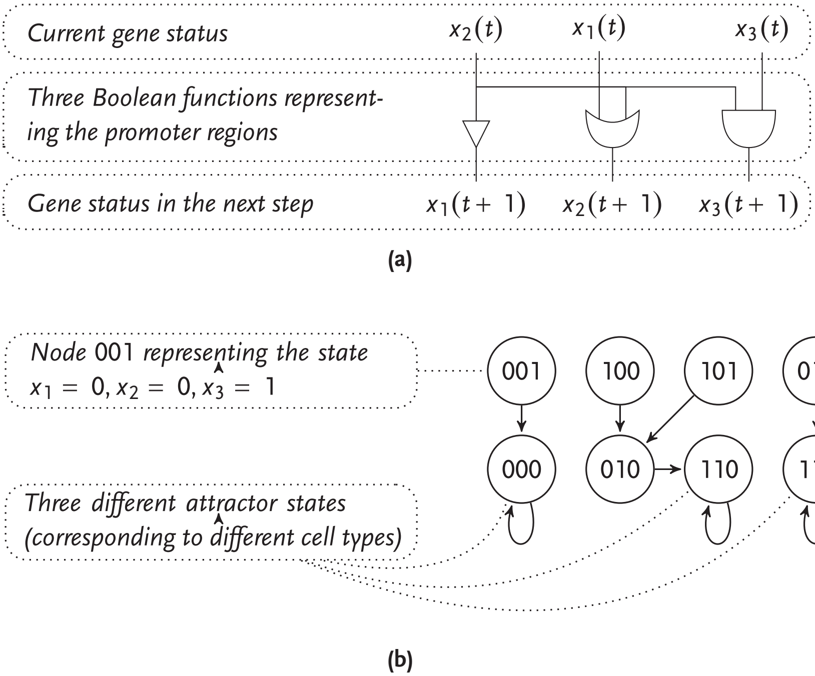

Current gene status
$x_2(t)$
$x_1(t)$
$x_3(t)$
Three Boolean functions representing the promoter regions
Gene status in the next step
$x_1(t+1)$
$x_2(t+1)$
$x_3(t+1)$
(a)
Node 001 representing the state $x_1 = 0, x_2 = 0, x_3 = 1$
001
100
101
011
000
010
110
111
Three different attractor states (corresponding to different cell types)
(b)

Figure 5.1 Circuit diagram (a) and state transition diagram (b) are two different but equivalent representations of a Boolean network. The circuit diagram is an abstract model of the promoter region of a gene and determines their synthesis based on the Boolean functions and their current availability. The state transition diagram represents all possible states.

5.2 Biological Background

This section briefly explains essential steps of gene regulatory systems. Comprehensive work concerning this topic can be found in many textbooks about molec-

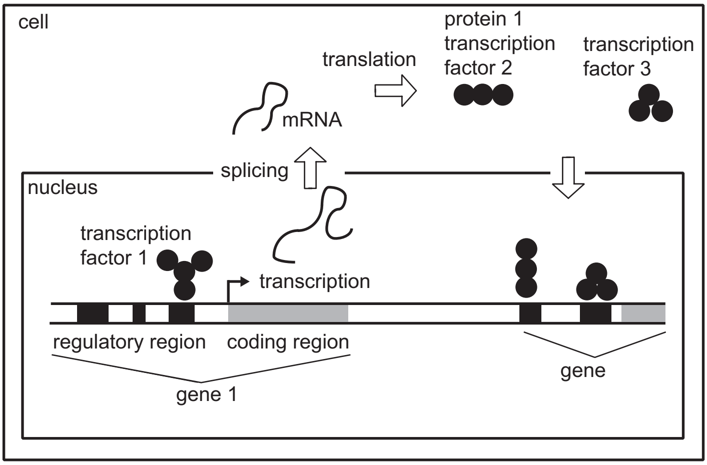


Figure 5.2 The regulatory region of gene 1 has three transcription factor binding sites. In this example, only transcription factor 1 has to bind to the third binding site to activate the transcription. The mRNA is spliced and translated into protein 1, which acts as transcription factor 2 and binds together with transcription factor 3 at the regulatory region of a further gene.

ular biology, for example Alberts *et al.* [5]. A human body, just as many animals and plants, consists of millions of cells with many different cell types. Every single cell has a nucleus[2] containing the *deoxyribonucleic acid* (DNA) which stores all the genetic information. This DNA is a long macromolecule made of four different sequentially arranged nucleotides. It includes the genes which code for certain proteins. A gene comprises the regulatory and the coding region. If one or more specific molecules bind to the regulatory region the coding region is transcribed to *messenger RNA* (mRNA) which can be translated into proteins. These procedures, transcription and translation, are called the central dogma of molecular biology. Since the mRNA is translated outside of the nucleus it must leave the nucleus and in doing so, regions not contributing to the entire protein are cut off in a process called splicing. This can take place at different locations on the mRNA allowing for different proteins originating from one single gene. Proteins are specialized macromolecules and act as enzymes, antibodies, or structural proteins to name just a few. A particular class of proteins which are able to bind to the DNA and regulate the transcription are called *transcription factors*. They can bind to specific transcription factor binding sites, called *cis-regulatory elements* and can either enhance or suppress the transcription. Normally several such binding sites are contained within the regulatory region of a gene, and multiple transcription factors, acting somehow together, are necessary to start the transcription of a gene (see Figure 5.2). It is assumed that strong cooperation and also competition exist between regulatory elements. Transcription factors can bind directly to the DNA or after the forma-

[2] There are also organisms without a nucleus.

tion of small complexes. Furthermore, other modifications like phosphorylation can alter a transcription factor's mode of action (Figure 5.2). Proteins and mRNA molecules are degraded by several cellular mechanisms. Therefore, transcription factors are only available for a certain period of time unless the corresponding genes are permanently activated. Thus a living cell is a dynamic system with genes temporarily switched on or off in a coupled manner. Although nearly all cells of an organism have the same DNA they can develop into different specialized units performing various tasks. To understand the gene regulatory mechanism underlying these processes is one of the main tasks in developmental biology. Furthermore, the interactions of many genes also play an important role in adult cells. Thus cells are able to maintain a certain function and can undergo distinct programs like the ordered cell death (apoptosis) or cell division. They are also able to react adequately to perturbations of the environment.

The decoding of this complex mechanism would not only give insight into the development of life, it would also help one to understand the reasons for complex diseases like cancer and also to facilitate drug design.

5.3 Aims of Modeling

Due to the fact that gene regulation plays such an important role it is desirable to fully understand living organism at the molecular level. Sequencing projects like the human genome project revealed the whole genome of humans and other organisms and now thousands of genes and their regulatory elements have been identified [10, 22]. Presently the major challenge is to discover which genes are expressed at different cell states and how they interact. The complex and dynamic feedback network of interwoven genes makes an intuitive understanding virtually

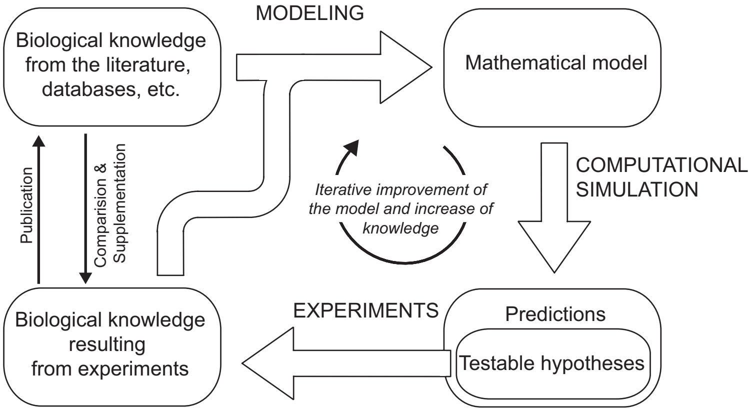

MODELING
Biological knowledge from the literature, databases, etc.
Mathematical model
Publication
Comparision & Supplementation
Iterative improvement of the model and increase of knowledge
COMPUTATIONAL SIMULATION
EXPERIMENTS
Biological knowledge resulting from experiments
Predictions
Testable hypotheses

Figure 5.3 Modeling of gene regulatory systems based on experiments and data from the literature. After the delineation of a basic model, alternating modeling, simulations and experiments improve the mathematical model.

impossible and formal methods are required, since they can describe the structure of gene regulation unequivocally. Simulation of mathematical models can create hypotheses in a systematic manner and provide biologists with useful predictions in the case of expensive, ethically unjustifiable, or even impossible, lab experiments.

The establishment of a model normally starts with the collection of data from the literature or from new experiments. Dependent on the accuracy of the data and the modeling goal, a first model is built and adapted until it fits to the data. Computer simulations yield predictions which should be tested in further experiments. Falsified predictions require a revision of the model and, in an iterative process, the desired model is built. Figure 5.3 outlines this procedure.

5.4
Modeling Techniques

In the following we present common techniques used to model gene regulatory systems. Comprehensive introductions can be found in the books of Bower and Bolouri [8] or Klipp *et al.* [22], a concise review focusing on mathematical methods is given by de Jong [10]. Markowetz *et al.* [29] provide an overview on inference algorithms with a main focus on statistical methods. The choice of an appropriate modeling approach can depend on several factors like the amount and accuracy of the given data, the objectives which the model should address, and the computational power available.

Different modeling techniques can be divided into (1) dynamic models, which allow for changes over time, or static ones, which do not, (2) quantitative or qualitative systems, (3) deterministic models or stochastic ones, which account for random processes ubiquitously present in biochemical reactions, (4) discrete or continuous systems.

Bayesian networks (Friedman *et al.* [14]) are used to represent static dependencies of gene interactions. They can be advantageously applied on incomplete and noisy data and learning techniques allowing for the inference of such networks from gene expression data. These models cannot represent dynamic aspects but dynamic Bayesian networks (Dojer *et al.* [12]) address this problem.

Another static approach used to reconstruct network structures and capable of dealing with indirect observations are nested effect models introduced by Markowetz and coworkers [28]. Boolean networks are presumably the simplest dynamic models, where only one of two logical values can be assigned to a system's component. They are explained together with probabilistic Boolean networks below.

An extension of Boolean networks are generalized logical networks (see Thomas [39]) in which not only two but an arbitrary number of discrete values are allowed for one modeled component. Furthermore, a transition to the next state can take place in an asynchronous manner. Differential equations are frequently utilized to model dynamical systems in engineering or science. In terms of gene regulatory networks each component (e.g. protein, protein complex, mRNA) is represented by a real number representing its concentration at a certain point in time. Dependent on these concentration values the rate equations describe the fluxes between components. Differential equations are suitable for precise dynamic modeling but they require accurate kinetic data. In general, these systems of differential equations are not linear and thus can only be solved numerically by appropriate solvers. All approaches discussed so far neglect spatial dimensions and thus assume an equal concentration of the components at any location, which sometimes is an oversimplification. Compartmental models solve this problem by introducing a spatial separation into two or more compartments, for example, the cell nucleus and the remaining part of the cell. The use of partial differential equations allows continuous modeling of the spacial dimension. In principle, many systems can be simulated very accurately by differential equations but the assumption that components change their concentrations deterministically and continuously is not appropriate if there are only a few interacting molecules. In these cases discrete stochastic simulations, as proposed by Gillespie [15] for example, are more convenient. Such algorithms model the exact number of molecules and randomly (based on probability distributions) select a reaction and a time period after which this reaction takes place.

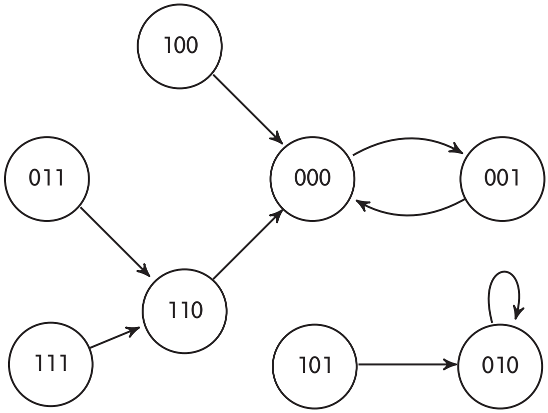
Here we have discussed only some important approaches – a multiplicity of further approaches exist, for example, piecewise linear differential equations and qualitative differential equations, two methods based on differential equations, rule-based methods, Petri nets, or models based on process algebra, to mention just a few.

5.5
Modeling GRNs with Boolean Networks

For consistency of notation with related work we define a Boolean network (BN) in a similar way to Akutsu *et al.* [4]. A Boolean network $G(V, F)$ consists of n nodes $V = \{x_1, \dots, x_n\}$ and n corresponding *Boolean functions* $F = \{f_1, \dots, f_n\}$. Each node can take the value 0 or 1 resulting in one discrete state $\boldsymbol{x}$ out of the state space $S = \{0, 1\}^n$ with $|S| = 2^n$. The Boolean function $f_i(x_{i_1}, \dots, x_{i_k}) \in F$ depends on k input nodes and its result determines the value assigned to the node x_i in the following time step:

$$x_i(t+1) = f_i(x_{i_1}(t), \dots, x_{i_k}(t)) \ , \quad 1 \le i \le n \tag{5.1}$$

The number of input nodes k may depend on the function f_i and in the following we define k as the maximum number of input nodes (*indegree*) among all functions in F. The connection between a node and its input nodes is also called *wiring*. There are 2^k different input settings for the k input nodes of a Boolean function. For each of these 2^k settings the output value of this function can either be 0 or 1 resulting in 2^{2^k} possible Boolean functions. Starting with a state $\boldsymbol{x} \in S$ the network updates



(a)

(t)			$(t+1)$		
x_1	x_2	x_3	x_1	x_2	x_3
0	0	0	0	0	1
0	0	1	0	0	0
0	1	0	0	1	0
0	1	1	1	1	0
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	0	0	0
1	1	1	1	1	0

(b)

$$\begin{aligned}
x_1(t+1) &= x_2(t) \wedge x_3(t) \\
x_2(t+1) &= x_1(t) \wedge x_3(t) \vee \overline{x_1(t)} \wedge x_2(t) \\
x_3(t+1) &= \overline{x_1(t)} \wedge \overline{x_2(t)} \wedge \overline{x_3(t)}
\end{aligned}$$

(c)

Figure 5.4 State transition diagram (a), state transition table (b), and the Boolean functions (c) are equivalent descriptions of a Boolean network. This network has two basins of attraction. One consists of the states 101 and the singleton attractor 010. The other one comprises the remaining six states and has an attractor cycle of length two.

its state synchronously at discrete time steps $t = 0, 1, 2, \ldots$, that is the value of each node at time $t + 1$ is derived based on the values of these nodes at time t. Thus a trajectory in the state space S is passed as shown in the example of Figure 5.4. Starting at an arbitrary state the system returns to a state visited before after at most 2^n states. This implies that the system has at least one cycle. Cycles are called attractor cycles and their length is the number of states visited possibly more than once within a trajectory. Attractor cycles with length one are called singleton attractors. Each attractor is assigned a basin of attraction comprising the cycle states, as well as states yielding the system into this cycle. Consequently, each of the 2^n network states belongs to exactly one basin of attraction which in turn is associated with exactly one attractor. States that are not part of an attractor are called transient states and the number of states between such a state and the first state belonging to a cycle is called the transient length. In terms of gene regulatory networks, a single gene is associated with a Boolean variable where the values 1 and 0 imply the dichotomized gene states expressed and not expressed, respectively. According to Kauffman [18], attractors are related to cell types and the trajectory before an entire attractor is associated with cell development or the cells repair mechanism after injury (see also Somogyi and Sniegoski [38]). The use of Boolean networks to model gene regulatory systems involves two questionable simplifications. First, the continuous expression of genes is reduced to two discrete states; second, the Boolean functions are updated synchronously at discrete time steps. However, cooperative behavior within biochemical reactions, induced or intensified by positive feedback mechanisms, results in sigmoidal input output relations and resembles switch-like behavior rather than linear relations [10, 38] as sketched in Figure 5.5. Due to different kinetic behavior and further protein modifications, the availability of transcription factors varies. Thus the assumption of synchronous updates, including similar lifetimes of different components, may be an idealization which is not able to render all aspects of cellular behavior [10].

Despite the fact that they are only an approximation [27], Boolean networks are assumed to be an appropriate technique, or at least a good starting point, to model gene regulatory systems [10, 41] for the following reasons. The models are simple and thus enable both the concentration on major network features as well as the computation of large, probably randomly constructed, networks to analyze their general properties [16–18, 38, 41]. Biological data obtained by measurements like microarray analysis are often binary and modeling large systems with thousands of components with alternative continuous approaches like nonlinear differential equations is barely manageable [18].

An extension of Boolean networks are probabilistic Boolean networks [34, 35] where each function f_i is chosen from a given set of possible functions.

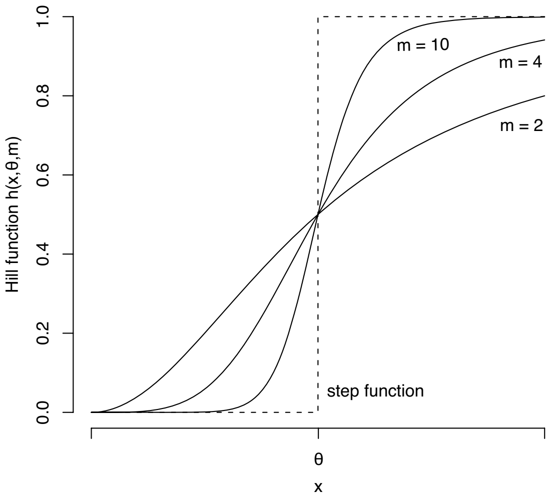


Figure 5.5 *Hill function* $h(x, \theta, m) = x^m/(x^m + \theta^m)$ for various m (solid lines) and the *step function* (dashed line). The Hill function becomes sigmoidal for $m \geq 2$ and is often used to model gene activation. The Boolean model corresponds to the step function and approximates the sigmoidal behavior.

5.6 Dynamic Behavior of Large Random Networks

Randomly constructed Boolean networks are used to identify the global behavior of large interconnected networks. Such models mainly address the long-term behavior revealing periodicity, self-organization, robustness, or redundancy. Furthermore, the relation between these global characteristics and local properties, that is the rules under which single nodes are constructed, are of great interest. It is unclear to which extent random networks with a distinct number of nodes or analytical results, where this number is infinite, can serve as a model for real gene regulatory networks found in living systems. However, Kauffman [17], who introduced random Boolean networks in terms of gene regulation [36], pointed out many similarities between the behavior of such artificially constructed and real gene regulatory networks.

We define a random Boolean network (RBN) as follows. Unless stated otherwise, the Boolean functions $f_i \in F$ are chosen randomly and uniformly distributed from the set of all 2^{2^k} Boolean functions. Models where the single functions are selected according to an arbitrary probability distribution are also feasible. Hence models where only a subset of all Boolean functions appear, for example, only *canalizing functions*,[3] are also possible. The k input nodes to every function are chosen independently among all n nodes. After this two-stage random selection the networks remain fixed. Kauffman [18] calls these nets random *NK* Boolean networks.[4] It is clear that all RBN with $k < n$ are a subset of all totally connected nets with $n = k$. A further subset of RBNs are networks were the single nodes are arranged in a spatial order and the input to each node is restricted to their immediate neighbors. These networks are called cellular automata. In general, they are too restrictive to represent the potential wiring of gene regulatory networks [38].

Kauffman not only introduced RBNs to study gene regulatory behavior but also gave some interesting links between his networks and living organisms [17] which can be seen as a kind of justification for the great interest in Boolean networks which have now existed for almost 40 years. Therefore we summarize the main results of Kaufmann's pioneering work in the following, although his conclusions concerning the length and number of attractor cycles were improved in later years. He first focused on RBNs containing Boolean functions connected to two other nodes ($k = 2$). Kauffman numerically simulated networks of various length up to approximately 8000 nodes. He observed that the median of the cycle lengths is rather short with a majority of even values and growth approximately with $\sqrt{n}$. Kauffman's simulations suggest also that the expected number of cycles is about $\sqrt{n}$.

[3] In canalizing Boolean functions at least one state of one input variable determines one state of the function's output, regardless of other input variables.

[4] Somogyi and Sniegoski [38] use the term randomly wired Boolean network or general Boolean network if the number of inputs is not restricted ($k = n$).

If we assume a network consisting of 25 000 nodes, which is about the current estimate for the number of protein coding genes of a human [9], the median cycle length and also the number of attractors in such a random network would be about 160. Compared to the state space, which is $2^{25\,000}$, this is an extremely small number, indicating the origin of a highly ordered system. Thus Kauffman established a relationship between these findings and various biological organisms such as bacteria, protazoans, or vertebrates (including humans) by comparing the cell cycle time subjected to the number of genes, on the one hand, and the estimated cycle time based on cycle lengths of RBNs, on the other hand. The function describing the median cell cycle time of several organisms lies between two functions describing the estimated cell cycle time of RBNs either with all 16 functions ($k = 2$) or with 14 Boolean functions where constant ones are not used. The predicted numbers of cell types were compared to data from organisms where these numbers are known and they are of the same magnitude. The hypothesis that attractors can be associated to cell types is substantiated by these findings and is probably one of the reasons why it has been accepted widely. However, the number of cycles in $k = 2$ RBNs, which has been assumed to be $\sqrt{n}$ for a long time, was revised by Bilke and Sjunnesson [7] and assumed to grow linearly with increasing network size n. Bilke and Sjunnesson applied a decimation algorithm that allowed one to neglect variables which are irrelevant for the long-term behavior of the network. Together with more computational power (the study was published approximately 30 years after Kauffman performed his numerical simulations) they were able to run simulations that enumerated the whole state space for a network size of up to 32, which is still small compared to networks comprising thousands of nodes as do gene regulatory networks. For larger networks they were forced to go back to the following sampling method already applied by Kauffman. Starting with arbitrarily chosen initial states the trajectories are observed and the number of different cycles is counted. Simulations by Socolar and Kauffman [37] suggested that the medium number of attractors grows “faster than linear” with n. A recent analytical study by Samuelsson and Troein [32] showed that the number of cycles grows faster than any power law with the network size n. It turned out that the sampling techniques underestimate this number. This can be explained by the fact that cycles with a small basin of attraction or those which occur rather seldom are not detected [13, 32]. Recently Klemm and Bornholdt [21] proposed RBNs with slightly modified updating rules that abolishes the strict synchronous update paradigm proposed previously (see (5.1)). For networks with indegree $k = 2$ their simulations indicated that the majority of attractors in parallel updated RBNs are artifacts as a result of this synchronous update. For a certain subset of nodes they delay the effects of these nodes for a short time period ε (smaller than the updating period), that is, some nodes which would change their states at a given point in time stay unchanged until the time ε has elapsed. Now an attractor is called *stable* if the system resides in this originally chosen attractor after applying all possible perturbations mentioned above (i.e. delaying every possible subset of nodes), otherwise the attractor is termed *unstable*. Simulations for a network size up to $n = 40$ reveal that, with increasing n, almost all attractors are unstable in the above sense and that the number of stable ones grows approximately with $\sqrt{n}$. Furthermore, the basin of attraction of unstable attractors are much smaller than in case of stable attractors.

Also, in the case of the attractor cycle length it seems that the network size of Kauffman's first simulations [17, 18] was too small to specify the behavior of larger networks. Lynch [26] could show that the mean cycle length is superpolynomial in the network size n but they further state that this large average cycle size is due to few RBNs with very large cycles and that most of the networks could have rather small cycles. Furthermore, the primarily assumed cycle length of $\sqrt{n}$ still holds as an approximation for small n and thus the interesting link between random networks and living organisms is not invalid at all [6].

Kauffman [17] also investigated the activity and robustness of RBNs and found that the number of nodes which change their value within a cycle in nets consisting of 100 nodes is at most 35, implying similar states within one cycle. Nodes that stop to change their values at some time are called *frozen nodes*. For systems with up to 2000 nodes Kauffman could also show that flipping one node of an attractor cycle's state lets the network return to the initial cycle with a probability between 0.85 and 0.95, demonstrating a strong robustness against such perturbations. The nodes which, when flipped, do not influence the entire cycle are called weak nodes. By simulations he also observed that networks where functions have three input nodes behave in a similar manner, although he only made simulations with at most 50 elements. He therefore concluded that these characteristics of RBNs are not restricted to functions with two input nodes. However, analytical studies showed a substantial difference between $k = 2$ nets and those where $k \geq 3$ (see below). These results reinforce the hypothesis that living organisms are based on randomly constructed gene interactions with a rather low connectivity.

Subsequently Kauffman and other authors discussed many further aspects of Boolean networks and a comprehensive summary is given in his book [18]. We will only discuss some of these results briefly. If every node in a RBN is connected to all other nodes ($k = n$) the number of attractors is of the order of n/e. However, these attractors have a median cycle length of $\sqrt{2^n}/2$. This exponential length can be calculated as follows. If every node receives input of all other nodes the successor state to an initial one can be considered as a random choice and thus the probability for a cycle length of 1 is $1/2^n$. Similarly, the probabilities for longer cycle lengths can be stated and thus the above-mentioned cycle length follows. The fact that we can regard successor states as randomly chosen also shows a vulnerability against small perturbations. Compared to the $k = 2$ networks, where a temporal flip of one node rarely leads the system into another basin of attraction, such a perturbation can lead the $k = n$ system to any other attractor.

The opposite to these fully connected networks are networks where each node has only one single input from another node ($k = 1$). These systems are rather modular, building many single cycles whose number increases exponentially as n increases. With a median cycle length of $\sqrt{\pi/2}\,\sqrt{n}$ the attractor cycles are quite short.

Derrida and Pomeau [11] approximated these RBNs by an annealed model where the functions and the wiring are re-chosen randomly in every single time step. For

this model they could proof an existing phase transition between $k = 2$ and $k \geq 3$ nets. Here we follow Kauffman [18] giving a simplified sketch of the proof.
Consider two arbitrary states $\boldsymbol{u}_{t_0}$ and $\boldsymbol{v}_{t_0}$ at time t_0 and let h_{t_0} be the Hamming distance between them, that is the number of nodes which have different values. With A_{t_0} we denote the set of all $n - h_{t_0}$ nodes which are equal in the two states $\boldsymbol{u}_{t_0}$ and $\boldsymbol{v}_{t_0}$. Let p_t be the probability that one input node is contained in the set A_t at a certain time point t. Since the input nodes for the functions are chosen arbitrarily p_{t_0} equals $(n–h_{t_0})/n$, and the probability that all k input nodes of one function belong to A_{t_0} is $(p_{t_0})^k$. The expected number of nodes receiving all their inputs from A at t_0 is therefore $n(p_{t_0})^k$. Clearly, these nodes are equal in the next time step t_1. Each of the remaining $n\left(1 - (p_{t_0})^k\right)$ nodes has at least one input node with different values and the probability that it becomes equal in the successor states $\boldsymbol{u}_{t_1}$ and $\boldsymbol{v}_{t_1}$ is 1/2. Thus the expected number of equal nodes after one time step p_{t_1} is

$$E(|A_{t_1}|) = n(p_{t_0})^k + \frac{1}{2}n(1 - (p_{t_0})^k) \,. \qquad (5.2)$$

For the next time step we have to distinguish between the RBNs introduced earlier and the approach by Derrida and Pomeau. In RBNs the Boolean functions and the input nodes are chosen at random and remain fixed thereafter. This is also called the quenched model. The proof here uses the annealed model where both the functions and their input nodes are re-chosen every single time step. Of course, this is a rigorous assumption compared to the networks introduced by Kauffman. In the annealed model we do not have to consider the fact that the states $\boldsymbol{u}$ and $\boldsymbol{v}$ are correlated to the fixed functions f_i and thus we can write:

$$p_t = (p_{t-1})^k + \frac{1}{2}(1 - (p_{t-1})^k) \,. \qquad (5.3)$$

Equation (5.3) reveals a crucial difference between networks with $k = 2$ and $k > 2$ since for $k = 2$ the two states $\boldsymbol{u}$ and $\boldsymbol{v}$ align and their distance becomes zero over time. If k is greater than 2 the distance approaches to a certain value, that is given by (5.3). Figure 5.6 compares the fraction of equal values at time t and $t+1$. Despite the assumptions made in the annealed model, numerical simulations comparing these results to the RBNs show very good accordance between both models [11] and it is assumed that the annealed model approach shows a phase transition for RBNs with a critical value of $k = 2$ [18].

A further analytical approach used to analyze the conclusions Kauffman drew from simulation results is given by Lynch [27]. Dependent on the probability distribution of the Boolean functions used to construct a network, he derived an algebraic parameter λ. If $\lambda \leq 1$ almost all nodes are weak nodes and become frozen ones very quickly. Thus the network exhibits ordered behavior. He applied his results to $k = 2$ networks with all possible 16 Boolean functions and to networks with only 14 functions where the two constant ones are not used. For the first case his analyses predict an ordered behavior but not for the second class of random Boolean networks. This is a slight contradiction to the computational simulation results of Kauffman who supposed the high fraction of constant functions in $k = 2$

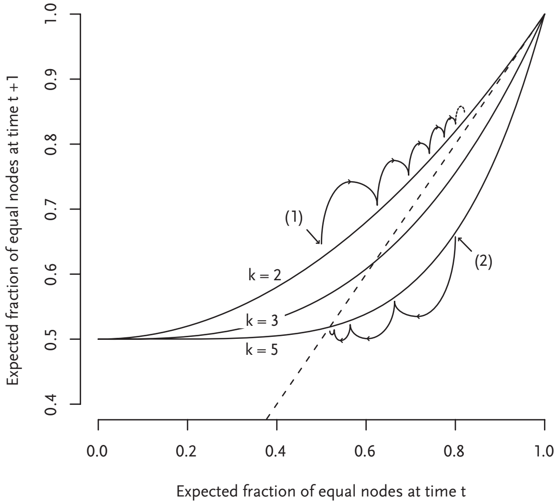


Figure 5.6 Expected fraction of equal values at time t and $t + 1$ for various k. For $k = 2$ this number approaches 1, i.e. the arbitrarily chosen initial states become the same (1). If $k > 2$ the equal values of these states approach a certain value given by the intersection with the straight line, where the number of equal values of successive states remains constant (2).

networks (two out of 16) to be responsible for the ordered behavior but received similar results for both cases.

Based on Lynch's findings Schober and Bossert [33] analyze random networks with *biased Boolean functions* for arbitrary indegree k. The *bias* p of a Boolean function is the number of **1** in its truth table divided by the number of total entries (2^k). In case of randomly selected Boolean functions one can distinguish two possibilities. (1) The truth table entries are selected according to the probability p. (2) Functions where the number of **1**-entries in their truth table is exactly $p2^k$ are selected with equal probability, other function are never chosen. This is called *fixed bias*. Schober and Bossert showed that, for random networks where the functions are chosen in this way, the expectation of the average sensitivity equals the parameter λ introduced by Lynch. Interestingly, their analysis for example shows that also networks with $k = 2$ can be unstable, that is $\lambda > 1$, if the bias is fixed.

5.7
Inference of Gene Regulatory Networks from Real Data

The previous section dealt with large randomly constructed networks and their properties. Needless to say, Boolean networks can also be constructed to model a concrete gene regulatory network or a part of it, as mentioned in the Introduction. Models can be based on measurements combined with literature studies and reliable assumptions whereas rather small networks can be constructed manually.

However, in recent years high-throughput technologies, such as microarray analysis, have improved, allowing for simultaneous measurements of gene expressions of thousands of genes within one experiment. This progress comes with the need for fast inference algorithms to determine the underlying biological mechanism. Boolean networks provide a good framework for reconstructing a gene regulatory network from time series data for the same reasons as discussed in Section 5.4.

5.7.1 Problem Definition

The problem, which can be assigned to the field of *reverse engineering*, can be defined as follows (similar to Akutsu *et al.* [4]). Given a pair of states $(\boldsymbol{u}, \boldsymbol{v})$ where $\boldsymbol{u}$ and $\boldsymbol{v}$ are states of the state space S, a Boolean network $G(V, F)$ is consistent with this pair if the following holds

$$v_i = f_i(u_{i_1}, u_{i_2}, \dots, u_{i_k}) \;, \quad 1 \le i \le n \;. \tag{5.4}$$

That is, the network G in state $\boldsymbol{u}$ (input state) switches to state $\boldsymbol{v}$ (output state) after one state transition. Given a set of such input-output pairs $\{(\boldsymbol{u}^{(1)}, \boldsymbol{v}^{(1)}), (\boldsymbol{u}^{(2)}, \boldsymbol{v}^{(2)}), \dots, (\boldsymbol{u}^{(m)}, \boldsymbol{v}^{(m)})\}$ the consistency problem (also called extension problem) is to state a Boolean network G consistent with all pairs $(\boldsymbol{u}^{(l)}, \boldsymbol{v}^{(l)})$ if one exists. The enumeration problem is to list all such networks G consistent with all input-output pairs and identification problem is to decide whether there is a unique Boolean network with this property and to state it, if it exists. Clearly, if we have solved the enumeration problem the other problems are close at hand.

Often we are given time series data in the form of an $(m + 1) \times n$ matrix where each of the $m + 1$ rows corresponds to one observed network state (with n nodes) at a given time point. In this case we have m input-output pairs where, for every such pair, the input state $\boldsymbol{u}$ corresponds to the j-th row and the output state $\boldsymbol{v}$ corresponds to the $j + 1$-th row $(1 \le j \le m)$. In the following we assume that m input-output pairs are given, no matter whether they origin from time series data or from single measurements. To reduce the computation time the indegree k of Boolean functions is often reduced to small values (often $k = 3$). This may be justifiable since the mean connectivity is approximately 2 or 3 in *Escherichia coli* and 4 to 8 in higher metazoa [23]. Also used to reduce the algorithm's time complexities and to limit the number of results, only a subset of all 2^{2^k} Boolean functions with indegree k can be allowed. This subset may contain only biologically reasonable functions. All algorithms described below begin with the inference of one out of n possible functions. The whole network is derived by applying the methods n times.

5.7.2 Identifying Algorithms

An obvious algorithm derives Boolean functions which are consistent with the given data for every single node by testing the consistency of each Boolean function and each possible wiring with the given data. There are $\binom{n}{k}$ possibilities to connect a Boolean function with indegree k to n nodes. Each such wiring is combined with all 2^{2^k} possible functions and has to be tested if it is consistent with all input-output pairs according to (5.4). Note that we do not have to consider functions with less than k connections (which is equal to the multiple selection of input nodes) because the test of all possible Boolean functions with indegree k already implies such circumstances. The function of x_1 in Figure 5.4, for example, has only two effective inputs which would be covered by the $k = 3$ Boolean function that corresponds to the (reducible) truth table expressed by column four of Figure 5.4 (b). If the test for one such pair and a given function is done in $O(k)$ time, the algorithm performing this test for all n nodes has a time complexity of [4, 20]

$$O\left(2^{2^k} \cdot \binom{n}{k} \cdot m \cdot n \cdot k\right) . \tag{5.5}$$

If k is fixed, which is justifiable for biological systems, the algorithm works in polynomial time and it is often stated that this algorithm has a time complexity of $O(n^{k+1}m)$. Akutsu and colleagues who proposed this simple algorithm reduced the time complexity to $O(n^k m)$ by using a *trie* data structure [1] and achieved a further improvement by using a randomized algorithm based on matrix multiplication [2] which we will sketch in the following. They presented an algorithm for the counting problem but claimed that the order of the time complexity does not change if modifications are made to solve the consistency and identification problem. The algorithm counts the number of consistent functions of the form $x_1 \wedge \overline{x_2}$ similar to the algorithm proposed in Kearns and Vazirani [19] and Valiant [40] for learning Boolean conjunctions (PAC learnable) and by using fast algorithms for the matrix multiplication. Other *AND* and *OR* functions are treated likewise, *XOR* functions are counted by applying the *trie* data structure. It is explained how to extend the algorithm to an indegree greater than 2 and the total time complexity is given as $O(m^{\omega-2}n^k + mn^{k+\omega-3})$, where ω is the exponent of the matrix multiplication.[5]

5.7.3
Noisy Data and the Data First Approach

Although the inferring methods discussed so far can be very interesting from an algorithmic point of view, as the algorithm based on matrix multiplication impressively demonstrates, their application to time series data from biological experiments is limited. The reasons are, first, the uncertainty of gene regulatory systems, and secondly, the measurement errors involved in biological methods [23, 31]. Therefore it is necessary to use methods that allow for inconsistency. Algorithm 1 shows a simple algorithm for the identification problem very similar to that proposed by Akutsu and coworkers [3]. The algorithm allows for a fraction θ of the m state transitions not to be consistent with Boolean functions f. Clearly, if we set the

[5] $\omega < 2.376$, at the time the algorithm was published.

threshold θ to 0 the algorithm allows no inconsistent input-output pair and can be seen as the simple identification algorithm proposed previously. If the threshold θ is set too high probably two or more functions are found and thus the algorithm stops without a result. Therefore it would be better to select that function f which has a minimum number of inconsistencies among all input-output pairs. To do so, one has to store the input combination and function which cause the lowermost mismatches while running the algorithm.

Algorithm 1 Identifying noisy Boolean networks

```
for $i = 1$ to $n$ do
    $count = 0$
    for all combinations of $k$ nodes $(i_i, i_2, \cdots, i_k)$ do
        for all Boolean functions $f$ do
            $mismatch = 0$
            for $j = 1$ to $m$ do
                if $v_i^{(j)} \neq f(u_{i_1}^{(j)}, u_{i_2}^{(j)}, \ldots, u_{i_k}^{(j)})$ then
                    $mismatch = mismatch + 1$
            if $mismatch \leq \theta \cdot m$ then
                output $f$ as a function for node $i$
                $count = count + 1$
    if $count \neq 1$ then
        output "Not identified"; halt
```

An algorithm to find Boolean networks which minimize the error rate according to weighted states has been proposed by Lähdesmäki *et al.* [23]. They were also able to solve the consistency problem with a significant improvement in the time complexity. In terms of machine learning, the problem of inferring a function f_i for one node x_i can be defined as follows [23]. Separate all states $\boldsymbol{u}^{(l)}$ of the input-output pairs $(\boldsymbol{u}^{(l)}, \boldsymbol{v}^{(l)})$ into the sets T and F such that $\boldsymbol{u}^{(l)}$ belongs to T if the output corresponding to node x_i equals 1, that is $v_i^{(l)} = 1$, otherwise $\boldsymbol{u}^{(l)}$ belongs to F. These two sets define a *partially defined Boolean function* pdBF(T, F). A function f_i is now called a consistent extension of pdBF(T, F) if $f_i(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = 1$ for $\boldsymbol{s} \in T$ and $f_i(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = 0$ if $\boldsymbol{s} \in F$. That is we want to find a perfect Boolean classifier.

Assume that we are given positive weights $W(\boldsymbol{s})$ for each $\boldsymbol{s} \in T \cup F$ and the weight of a subset of $T \cup F$ is the sum of the single weights. The error size of a function f_i is defined as the weight of all misclassified states, that is the weight of all states $\boldsymbol{s}$ in T for which $f_i(s_{i_1}, s_{i_2}, \ldots, s_{i_k}) = 0$ and vice versa. The best-fit extension problem is now to find a function f with minimal error size. Obviously, the consistency and enumeration problem are special cases of the best-fit extension problem where the error size is zero.

For unbound indegree ($k = n$) Lähdesmäki and colleagues [23] proposed the following inferring algorithm which is also called the "data-first approach" by Nam *et al.* [31] who stated the same idea. A so-called generalized truth table with 2^n entries maps each possible state out of the state space $S = \{0, 1\}^n$ to a vector $\boldsymbol{f}'$. Initially its entries are filled with the symbol ?. Let $\boldsymbol{f}'_s$ be the entry of $\boldsymbol{f}'$ that corresponds to $\boldsymbol{s}$. The algorithm now processes all states $\boldsymbol{s} \in T \cup F$ and updates the vector $\boldsymbol{f}$ as follows:

```
**if** $s \in T$ **then**
    **if** $f'_s = 1$ or $f'_s =?$ **then**
        $f'_s := 1$
    **else**
        output “No solution”; halt
**if** $s \in F$ **then**
    **if** $f'_s = 0$ or $f'_s =?$ **then**
        $f'_s := 0$
    **else**
        output “No solution”; halt
```

The algorithm outputs “No solution” if and only if there is at least one state which is contained in T and in F. In this case no classifier can be found. If all entries in $\boldsymbol{f}'$ contain 0s or 1s there exists one Boolean classifier f and the corresponding truth table is $\boldsymbol{f}'$. However $\boldsymbol{f}'$ may contain some ?s indicating an under-determined system. For the consistency problem these entries can be filled arbitrarily with 0 or 1 resulting in one consistent Boolean function. Let h be the number of remaining ?s. In case of the enumeration problem all possible settings ($\{0, 1\}^h$) can be assigned to the entries with a ? resulting in 2^h solutions.

If the time for the initialization of $\boldsymbol{f}'$ is ignored, the time complexity is

$$O((|T| + |F|) \cdot n) = O(m \cdot n) \ . \tag{5.6}$$

Note that we consider unbound indegrees and thus the processing of one state takes $O(n)$ time. For bound indegree of k variables this algorithm can be applied to all $\binom{n}{k}$ variable combinations. Up to now we considered the search for a perfect Boolean classifier, that is we determined one Boolean function. To infer a Boolean network we simply have to apply the algorithm n times. Thus the consistency problem can be solved in time

$$O\left(\binom{n}{k} \cdot m \cdot n \cdot k\right) \ . \tag{5.7}$$

which reduces the complexity by a factor of 2^{2^k} compared to (5.5).

In the case of the best-fit extension problem a similar algorithm can be applied. Instead of the vector $\boldsymbol{f}'$ two vectors, $\boldsymbol{c}^{(0)}$ and $\boldsymbol{c}^{(1)}$, initialized with zeros, are used to store the weights of the single states $\boldsymbol{s} \in T \cup F$. If $\boldsymbol{s}$ is contained in T (F) the corresponding entry of the vector $\boldsymbol{c}^{(0)}$ ($\boldsymbol{c}^{(1)}$) is set to the weight of $\boldsymbol{s}$. Finally, the vector entries are compared pairwise and the corresponding truth table entry of the resulting function f is set to 0 or 1 depending on which of the values is greater. Thus an optimal function f with minimal error size is determined. Clearly, the time complexity is the same as for the consistency problem. The restriction to an indegree of k and the extension to a Boolean network is the same as in the consistency problem. From a biological point of view it may be desirable not only to search for the

Boolean function with the lowest error size but also for a set of functions with limited error size. Among this set of functions a biologist could select the most reliable one based on experiments or verifications by means of literature research. Adapted from the above sketched method Lähdesmäki *et al.* [23] proposed a recursive greedy algorithm to perform the search of functions with a limited error.

An additional gain in the time complexity by a factor of $(\log(m))^{(k-1)}$ was achieved by Nam and coworkers [31]. The authors proposed a randomized algorithm which they called "top down search algorithm". It is based on the fact that if a set G of input nodes is consistent with an output node, any superset of G still has the same property. The algorithm randomly selects a consistent set of more than k genes and determines whether it contains a consistent subset of size k.

5.7.4
An Information Theoretical Approach

Based on the mutual information between input-output pairs Liang *et al.* [24] proposed an inference algorithm, called REVEAL[6]. To determine one Boolean function f_i assume we are given m input-output pairs. Let $u_h^{(l)}$ be the value of the h-th input node and let $v_i^{(l)}$ be the corresponding output node of the l-th input-output pair with respect to the function f_i. Let U_h and V be the series of input and output values, that is $U_h = (u_h^{(1)}, u_h^{(2)}, \ldots, u_h^{(m)})$ and $V = (v_i^{(1)}, v_i^{(2)}, \ldots, v_i^{(m)})$. Liang *et al.* determine the wiring of a Boolean function f_i based on the mutual information as follows. If $M(V, U_p) = H(V)$ then V is exactly determined by U_p, whereas $H(..)$ and $M(..)$ denote the entropy and the mutual information, respectively. Since $M(V, U_p) = H(V) + H(U_p) - H(V, U_p)$ it is sufficient to check if $H(U_p) = H(V, U_p)$. The entropy $H(V, U_p)$ is calculated from the frequency of how often one of the four input-output settings $((u_p = 0, v_i = 0), (u_p = 0, v_i = 1), (u_p = 1, v_i = 0), (u_p = 1, v_i = 1))$ occurs. If V is exactly determined by U_p the rule table can be directly stated from the input-output pairs, otherwise the function f_i has more than one input node and the mutual information between V and pairs of input nodes is tested. Consequently, if $M(V, [U_p, U_q]) = H(V)$ (i.e. $H(U_p, U_q) = H(V, U_p, U_q)$) the pair $[U_p, U_q]$ determines V. The entropy $H(V, U_p, U_q)$ is determined from the frequency of the eight input-output pairs $(u_p = 0/1,\ u_q = 0/1\ , v_i = 0/1)$. Repeating this analysis with increasing number of input nodes allows the inference of networks with any indegree. Liang *et al.* made simulations to test their algorithm and observed that a small numbers of input-output pairs were sufficient to identify a network. Akutsu and colleagues [4] could explain this observation analytically. They proved that for fixed k only $O(log(n))$ state transitions are necessary and sufficient to infer a Boolean network.

[6] Abbreviation of *Reverse Engineering Algorithm.*

5.7.5
Using the Chi-Square Test to Find Relationships Among Genes

Kim and coworkers [20] proposed a method to infer large gene regulatory networks using the chi-square test. Based on this test they implemented a variable selection method which efficiently selects likely candidates as input nodes. Thus, there is no need for an exhaustive search testing each possible wiring among all n nodes. Their program, however, is restricted to functions with an indegree of $k = 3$ and focuses only on the Boolean relations AND, OR, and NOT. We will briefly outline the idea of this selection method in the following. Let U_h and V be defined as in the former section. A 2×2 contingency table comprising the four cells $\{0, 0\}$,$\{0, 1\}$,$\{1, 0\}$, and $\{1, 1\}$ is constructed. The entries correspond to how often each input-output setting was observed within U_h and V. By analyzing all pairs of nodes with the chi-square test based on this contingency table, only significant ones are selected and examined in the following. Assume two nodes are significant $2 \times 2 \times 2$ contingency tables for these two nodes and each of the remaining nodes are constructed. Assume that after applying the second chi-square test the set C of nodes are selected as significant candidates. Now every combination of two nodes of C are considered together with the node selected in the first step to get three candidates that arise as input nodes for x_i, which corresponds to V_i. Two significance levels α_1 and α_2 are used to vary the sensitivity of the two chi-square tests.

This method is applied together with the best-fit extension problem. They constructed an artificial network with 40 nodes, compared their inferring method with the algorithm proposed by Akutsu *et al.* [4], and observed that their algorithm is approximately 6.9 times faster.

5.8
Conclusion

Here we have reviewed the interesting field of Boolean networks – discrete dynamical systems that can display remarkable properties. The first part focused on random Boolean networks and how microscopic properties, namely the way of connecting single functions, can affect the macroscopic dynamic behavior of the whole network. The first application of such networks in terms of gene regulation dates back some decades but interesting theoretical results have emerged only recently. From a theoretical or mathematical point of view, the analytical study of random Boolean networks and also familiar systems is without doubt a fascinating field. However, it is questionable to what extent an asymptotic result – the number of nodes is often assumed to be infinite – can be useful for real biological systems where the number of genes is limited. On the other hand, the number of genes of a human, for example, is so high that a systematic processing of each possible state is absolutely impossible. This problem becomes obvious in the case of the number of cycle attractors in random Boolean networks as discussed in Section 5.6. Further simulations, with more computational power, yield improved estimations.

Since there is currently little known about how genes really interact, some of the theoretical results obtained so far seem to have no connection to biological reality, such as the findings for biased Boolean functions. However, if biological methods improve at the same rate as they did in recent years there could one day be enough knowledge of gene regulatory systems – perhaps supported by reverse engineering methods – to apply mathematical results which up to now have had only a theoretical impact.

In the second part, we addressed the problem of reverse engineering to reconstruct gene regulatory networks. We presented some interesting algorithms to infer Boolean networks that are consistent with the given data. A machine learning approach discovers a perfect Boolean classifier to obtain networks that best fit to the data, based on individually assigned weights. One drawback of these algorithms is their efficiency since they exhaustively test all wiring combinations. For an analysis with thousands of genes the method based on the chi-square test, introduced by Kim *et al.* [20], might be more suitable. In the case of numerous genes this algorithm probably has to test many combinations in its last step and a combination of this algorithm and the methods proposed by Nam *et al.* [31] may result in a reduced runtime. Furthermore, the algorithm of Kim and coworkers is restricted to $k = 3$ and an extension to more than three input nodes could be useful in some cases.

In our work we inferred a Boolean network from time series data [25] but for some genes this resulted in hundreds of consistent functions, while for other genes no consistent function could be found. Clearly, the problem that there is no function at all can be approached by using the best-fit extension paradigm but there is still the challenge to cope with many consistent functions for certain genes. One solution would definitely be the collection of further biological data, although this involves expensive and time-consuming experiments. Another way to reduce the number of possible Boolean networks obtained by a reverse engineering method might be a selection of networks that show a certain and biologically reasonable long-term behavior like small attractor cycles or a robust behavior against small perturbations – which are again the properties exhibited by many random Boolean networks. Such an algorithm, that uses additional dynamic characteristics of the consistent networks found is given by Martin *et al.* [30].

Until now we have assumed that the data resulting from experiments are binary values representing one of two states of the genes. In fact, the biological methods are based on the measurement of the amount of mRNA, which is assumed to be proportional to the number of corresponding proteins. However, the impact of transcription factors depends not only on their amount but also on their affinity towards binding partners. Therefore, it is not clear at which mRNA level a component should be assumed to be active or inactive and, due to different affinities, a reasonable threshold can vary for every single gene. Thus inferring algorithms that incorporate this uncertainty could be a further step towards robust and reliable reverse engineering methods.

Acknowledgments

We thank Uwe Schöning, Martin Bossert, Steffen Schober, and Martin Hopfensitz for valuable discussions and improvements.

References

1 Aho, A.V. (**1990**) Algorithms for finding patterns in strings, in *Handbook of theoretical computer science (vol. A): algorithms and complexity*, MIT Press, Cambridge, MA, USA, pp. 255–300.

2 Akutsu, T., Miyano, S. and Kuhara, S. (**2000**) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computational Biology*, **7**(3–4), 331–343.

3 Akutsu, T., Miyano, S. and Kuhara, S. (**2000**) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**(8), 727–734.

4 Akutsu, T., Miyano, S. and Kuhara, S. (**1999**) Identification of genetic networks from a small number of gene expression patterns under the boolean network model, in *Pacific Symposium on Biocomputing*, (eds R.B. Altman and K. Lauderdale) World Scientific, pp. 17–28.

5 Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (**2002**) *Molecular Biology of the Cell*. Garland Science, New York, fourth edition.

6 Bastolla, U. and Parisi, G. (**1998**) The modular structure of kauffman networks. *Physica D*, **115**(3–4), 219–233.

7 Bilke, S. and Sjunnesson, F. (**2002**) Stability of the Kauffman model. *Physical Review E*, **65**(1 Pt 2):016129.

8 Bower, J.M. and Bolouri, H. (**2001**) *Computational Modeling of Genetic and Biochemical Networks (Computational Molecular Biology)*. The MIT Press.

9 International Human Genome Sequencing Consortium (**2004**) Finishing the euchromatic sequence of the human genome. *Nature*, **431**(7011), 931–945.

10 de Jong, H. (**2002**) Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *Journal of Computational Biology*, **9**(1), 67–103.

11 Derrida, B. and Pomeau, Y. (**1986**) Random networks of automata: A simple annealed approximation. *Europhysics Letters*, **1**(2), 45–49.

12 Dojer, N., Gambin, A., Mizera, A., Wilczynski, B. and Tiuryn, J. (**2006**) Applying dynamic Bayesian networks to perturbed gene expression data. *BMC Bioinformatics*, **7**, 249.

13 Drossel, B., Mihaljev, T. and Greil, F. (**2005**) Number and length of attractors in a critical Kauffman model with connectivity one. *Physical Review Letters*, **94**(8), 088701.

14 Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (**2000**) Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, **7**(3–4), 601–620.

15 Gillespie, D.T. (**1977**) Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, **81**(25), 2340–2361.

16 Kauffman, S. (**1974**) The large scale structure and dynamics of gene control circuits: an ensemble approach. *Journal of Theoretical Biology*, **44**(1), 167–190.

17 Kauffman, S.A. (**1969**) Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, **22**(3):437–467.

18 Kauffman, S.A. (**1993**) *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press.

19 Kearns, M.J. and Vazirani, U.V. (1994) *An introduction to computational learning theory*. MIT Press, Cambridge, MA, USA.

20 Kim, H., Lee, J.K. and Park, T. (**2007**) Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*, **8**, 37.

21 Klemm, K. and Bornholdt, S. (**2005**) Stable and unstable attractors in boolean networks. *Physical Review E*, 72:055101.

22 Klipp, E., Herwig, R., Kowald, A., Wierling, C. and Lehrach, H. (**2005**) *Systems Biology in Practice: Concepts, Implementation and Application*. Wiley-VCH.

23 Lähdesmäki, H., Shmulevich, U. and Yli-Harja, O. (**2003**) On learning gene regulatory networks under the boolean network model. *Machine Learning*, **52**(1–2), 147–167.

24 Liang, S., Fuhrman, S. and Somogyi, R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures, in *Proceedings of the Pacific Symposium on Biocomputing*, volume 3, (eds R.B. Altman, A.K. Dunker, L. Hunter and T.E.D. Klein), World Scientific, pp. 18–29.

25 Liu, Y., Asakura, M., Inoue, H., Nakamura, T., Sano, M., Niu, Z., Chen, M., Schwartz, R.J. and Schneider, M.D. (**2007**) Sox17 is essential for the specification of cardiac mesoderm in embryonic stem cells. *The Proceedings of National Academy of Sciences USA*, **104**(10), 3859–3864.

26 Lynch, J.F. (**1995**) On the threshold of chaos in random boolean cellular automata. *Random Structures and Algorithms*, **6**(2–3), 239–260.

27 Lynch, J.F. (**2007**) Dynamics of random boolean networks, in *Current Developments in Mathematical Biology, Proceedings of the Conference on Mathematical Biology and Dynamical Systems*, (eds K. Mahdavi, R. Culshaw and J. Boucher) World Scientific, pp. 15–38.

28 Markowetz, F., Bloch, J. and Spang, R. (**2005**) Non-transcriptional pathway features reconstructed from secondary effects of rna interference. *Bioinformatics*, **21**(21), 4026–4032.

29 Markowetz, F. and Spang, R. (2007) Inferring cellular networks – a review. *BMC Bioinformatics*, **8**(Suppl 6), S5.

30 Martin, S., Zhang, Z., Martino, A. and Faulon, J.-L. (**2007**) Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, **23**(7), 866–874.

31 Nam, D., Seo, S. and Kim, S. (**2006**) An efficient top-down search algorithm for learning boolean networks of gene expression. *Machine Learning*, **65**(1), 229–245.

32 Samuelsson, B. and Troein, C. (**2003**) Superpolynomial growth in the number of attractors in Kauffman networks. *Physical Review Letters*, **90**(9), 098701.

33 Schober, S. and Bossert, M. (**2007**) *Analysis of random boolean networks using the average sensitivity*. Preprint, available online at ArXiv, arXiv:nl.cg/0704.0197.

34 Shmulevich, I., Dougherty, E.R., Kim, S. and Zhang, W. (**2002**) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*, **18**(2), 261–274.

35 Shmulevich, I., Dougherty, E.R. and Zhang, W. (**2002**) From boolean to probabilistic boolean networks as models of genetic regulatory networks. *Proceedings of the IEEE*, **90**(11), 1778–1792.

36 Shmulevich, I., Yli-Harja, O. and Astola, J. (**2001**) Inference of genetic regulatory networks under the best-fit extension paradigm, in *IEEE-Eurasip Workshop on Nonlinear Signal and Image Processing*, (ed G. Arce) Baltimore, Eurasip.

37 Socolar, J.E.S and Kauffman, S.A. (**2003**) Scaling in ordered and critical random boolean networks. *Physical Review Letters*, **90**(6), 068702.

38 Somogyi, R. and Sniegoski, A. (**1996**) Modeling the complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation. *Complexity*, **1**, 45–63.

39 Thomas, R. (**1991**) Regulatory networks seen as asynchronous automata: A logical description. *Journal of Theoretical Biology*, **153**(3), 1–23.

40 Valiant, L.G. (**1984**) A theory of the learnable. *Communications of the ACM*, **27**(11), 1134–1142.

41 Wuensche, A. (**1998**) Genomic regulation modeled as a network with basins of attraction, in *Proceedings of the Pacific Symposium on Biocomputing*, (eds R.B. Altman, A.K. Dunker, L. Hunter and T.E. Klien), World Scientific, Singapore, pp. 89–102.

6
Symmetries in Quantum Graphs

Jens Bolte, Stefano Cardanobile, Delio Mugnolo[1]*, Robin Nittka*

6.1
Symmetries

In most branches of physics, symmetries play an eminent role. They are behind regular patterns that help one to understand the structure of a physical system, and identifying its symmetries often serves as an important step towards a correct description of a system. This is not only true on the level of the basic building blocks of a system, but also on a dynamical level where symmetries are relevant in finding the structure of forces that act in a given situation. Furthermore, symmetries can be used to reduce the complexity of a physical system and therefore help, for example, to solve the equations of motion describing the dynamics of a system. Moreover, they can also have an impact on the degree of regularity of the dynamics.

An important consequence of the presence of symmetries is its connection with conserved quantities. This connection can already be illustrated in the case of a point particle moving in euclidean three-dimensional space when it is described in terms of classical (Newtonian, Lagrangian, or Hamiltonian) mechanics. A symmetry then means a space–time operation (such as a translation or a rotation) that maps any possible motion of the particle to some other possible motion. For example, if the forces are such that any translation is a symmetry, linear momentum is conserved; in contrast, if rotations are a symmetry, angular momentum is conserved. Probably the best-known conserved quantity, however, is energy: its conservation follows from motions being symmetric with respect to time translation, meaning that any motion that is possible at a given time is also possible at any other time.

Within the framework of classical mechanics the connection between the "classical" symmetries described above and the associated conserved quantities can be established rather easily and has been known for a very long time. Roughly ninety years ago, however, Emmy Noether discovered that this connection extends far beyond what was known before [1]. She proved what later became known as Noether's Theorem and which, for example, plays a fundamental role in the theory of elemen-

[1] Corresponding author.

tary particles. Forces between elementary particles are described in terms of gauge theories in which an abstract "charge space" is attached to every point in space–time. The freedom to perform, typically unitary, transformations in these charge spaces is referred to as a gauge symmetry which, via Noether's Theorem, eventually enables the identification of conserved charges and currents.

The basic idea behind Noether's Theorem can, however, already be explained in the context of the classical mechanics of a single particle where it can be stated as follows. Let $x = (x_1, x_2, x_3)$ be the vector of the particle's position in three-dimensional euclidean space, and denote by $v = (v_1, v_2, v_3)$ the associated vector of velocity. When the particle has mass m, its momentum is $p = mv$. Then the state of the particle is specified in terms of (x, p) so that the state space of the system is $\mathbb{R}^3 \times \mathbb{R}^3$. This space is usually called phase space. A symmetry then is a transformation g on $\mathbb{R}^3 \times \mathbb{R}^3$ that leaves the form of Hamilton's equations of motion invariant (canonical transformation) and maps (x, p) to $(x', p') = g(x, p)$ such that any solution of the equations of motion is mapped to a solution of the same equations.

Often it is not only one such operation which is a symmetry. In such a case, typically, the set of symmetries is a group. This can be finite, infinite discrete, or continuous. An example of a finite symmetry group would be that of rotations about an axis with angles, say, $2\pi n/N$, $n = 0, 1, 2, \dots, N-1$. If arbitrary angles are permitted the group will be continuous; in fact, it will be the group SO(2). If, moreover, rotations about any axis are symmetries, the symmetry group will be SO(3). These rotation groups have additional properties that qualify them as Lie groups (see [2]). In the first case, SO(2) is a Lie group of dimension one since any rotation about a fixed axis is specified by a single parameter; its angle. In the second case, SO(3) is of dimension three since one requires three Euler angles to characterize a rotation about an arbitrary axis. Noether's Theorem now provides a connection between such Lie symmetry groups and conserved quantities. Before presenting the precise statement, however, we allow for somewhat more general cases. More precisely, the physical system should have d degrees of freedom. The phase space then is, in general, a (symplectic) manifold of dimension $2d$.

Theorem 6.1 (Noether, 1918) *Let G be a Lie group of dimension r acting on the phase space in terms of symmetries. Then there exist r independent conserved quantities.*

In the case of a single particle in three-dimensional euclidean space, where $d = 3$, rotations about any axis could be symmetries. Then $G = \mathrm{SO}(3)$, $r = 3$, and the associated conserved quantities are the three components of the vector of angular momentum. In the general case the conserved quantities are the generators of one parameter subgroups of the symmetry group.

An obvious consequence of Noether's Theorem is that the presence of a Lie symmetry group helps to solve the equations of motion (partially). To this end one uses the fact that the motion is confined to a submanifold of the phase space that is defined by the conserved quantities to take fixed values. This way the complexity of the problem is reduced since the relevant submanifold is of lower dimension. In particular, if the number of conserved quantities equals the number of degrees of freedom (and a further condition is fulfilled) the system is *integrable*. This does not necessarily mean that the integration of the equations of motion is a simple task, but it does imply that the motion is of a very regular type: in the phase space its trajectories regularly wind around tori of half the dimension of the state space itself. While the very concept of integrability goes back to Liouville [3], its full consequences – including the existence of the invariant tori – were only later recognized by Einstein [4]. The dynamical behavior of integrable systems is in stark contrast to that of chaotic systems, which possess no conserved quantities (apart from energy). A characteristic property of chaotic systems is ergodicity, implying that typical trajectories uniformly fill the entire surface of constant energy in the phase space.

When some obvious modifications are carried out, many of the ideas outlined above in the context of classical physics can be carried over to quantum physics. The modifications have to account for the fact that in quantum mechanics the state of a system is not specified by position and momentum, but rather in terms of a (complex-valued) function, often called the wave function of the system. This function typically is a function on position space, and is denoted as $\psi(x)$. The state space then is the space $L^2(\mathbb{R}^3)$ of square integrable functions and is a Hilbert space. More precisely, not every square integrable function may serve as a quantum state, but only normalized ones, fulfilling

$$\int_{\mathbb{R}^3} |\psi(x)|^2 \, \mathrm{d}x = 1 \; .$$

Then $|\psi(x)|^2$ is interpreted as a probability density for outcomes of position measurements. Obviously, $\psi(x)$ and $e^{i\alpha}\psi(x)$ yield the same probability density, so that quantum states are indeed equivalence classes of normalized wave functions under the multiplication with complex numbers of unit absolute value. Therefore, strictly speaking, the quantum state space is not a Hilbert space, but its associated projective Hilbert space. Symmetry operations must hence map equivalence classes to equivalence classes. However, Wigner showed that any such map can be realized in terms of either a unitary or an anti-unitary operator on the Hilbert space itself [5]. Moreover, when these operators provide realizations of symmetry transformations, they can be chosen to define a (projective) representation of the symmetry group, so that a composition of two symmetries fulfills

$$U_{g_1 g_2} = e^{i\omega(g_1,g_2)} U_{g_1} U_{g_2} \tag{6.1}$$

where ω is a suitable real phase.

Often it is more convenient to work with Hilbert spaces and only later refer to their projective spaces. For example, when a symmetry acts on position in the form $x \mapsto gx$, and is measure-preserving, an appropriate operation of the symmetry on the wave function is given by

$$(U_g\psi)(x) = \psi(g^{-1}x) \; . \tag{6.2}$$

That way U_g is a unitary operator on the Hilbert space and already provides the unitary realization of the symmetry as predicted by Wigner. In the example considered in (6.2) the phase ω occurring in (6.1) vanishes, hence yielding a proper unitary representation of the symmetry group. Furthermore, for an operation of this type to qualify as a symmetry it has to take solutions of the equations of motion again to solutions. In quantum mechanics the time evolution of a wave function is governed by the Schrödinger equation

$$i\hbar \frac{\partial}{\partial t} \psi(x, t) = H\psi(x, t) \tag{6.3}$$

where H is the so-called Hamiltonian operator which is a self-adjoint operator on the Hilbert space and represents the energy; $\hbar$ is Planck's constant divided by 2π. Solving this Schrödinger equation is equivalent to finding the spectral decomposition of H. In the case of a discrete spectrum this amounts to finding eigenvalues and eigenfunctions of the operator.

In this quantum mechanical context a unitary operator U represents a symmetry if it commutes with the time evolution,

$$[\mathrm{e}^{-itH/\hbar}, U] := \mathrm{e}^{-itH/\hbar} U - U \mathrm{e}^{-itH/\hbar} = 0\ . \tag{6.4}$$

When the symmetry group is a Lie group of dimension r, the statement of Noether's Theorem, if interpreted appropriately, can be carried over verbatim: there exist r constants of motion for the quantum system. Typically, these are the quantum analogs of the respective classical conserved quantities. In analogy to the classical case a Lie group symmetry in quantum mechanics can be used to solve the equations of motion partially, in this case by employing *complete reducibility* of unitary representations, i.e. the possibility of decomposing the unitary representation of the symmetry group into irreducible subrepresentations. The irreducible subspaces are then invariant under the dynamics and one only needs to solve the dynamics up to these subspaces.

In classical physics the presence of finite or discrete symmetry groups can be used to reduce the phase space to a certain relevant subset which, however, is of the same dimension as the phase space itself. Hence, the reduction of complexity is very modest. In quantum mechanics, however, a finite or discrete, nonabelian symmetry group does have noticeable consequences. In that case unitary irreducible representations and hence the invariant subspaces (typically) are of dimension greater than one. This situation is in close analogy to that of a Lie symmetry group and, in fact, a certain amount of regularity occurs in the dynamics although no conserved quantity exists. For the associated classical system this means that it is expected to behave "generically"; it may even be chaotic.

When a discrete symmetry group is present one often applies the procedure of a "desymmetrization" to the quantum system i.e. a restriction to one irreducible component. The reduced quantum systems are then expected to behave quantum chaotically, if they are sufficiently complex. In particular, this is conjectured to be the case when the associated classical system behaves chaotically [6]. The only symmetries that are exempt from desymmetrization are those that are represented through anti-unitary operators, the most prominent example being time reversal. The (anti-unitary) time reversal operator squares to $(-1)^{2s}$, where $s \in \{0, 1/2, 1, 3/2, \dots\}$ is the spin quantum number characterizing the behavior of the quantum system under rotations.

The plan of this article is as follows. In Section 6.2 we introduce quantum graphs and then, in Section 6.3, we summarize the precise definition of networks and quantum graphs by specifying an appropriate abstract Cauchy problem. Subsequently, in Section 6.4 we identify symmetries and invariant subspaces. Section 6.5 contains an extension to quantum graphs with magnetic forces and, finally, in Section 6.6 we summarize our findings.

6.2 Quantum Graphs

Typically, the configuration space of a physical system is a manifold (such that the associated phase space is the cotangent bundle of that manifold). The Hilbert space of quantum states then is that of the square integrable functions ψ on the configuration manifold. The quantum Hamiltonian that appears in the Schrödinger equation (6.3) is often of the form

$$H = (i\hbar\nabla + A)^2 + V\,.$$

Here, A, V are suitable potentials representing forces that are applied to the system. In the absence of magnetic fields, when $A = 0$, the magnetic Laplacian $(i\nabla + A)^2$ reduces to the positive definite Laplacian $\Delta = -\nabla^2$.

Obviously, via the Schrödinger equation the quantum dynamics are closely related to spectral properties of the Hamiltonian. Details of the latter are, however, notoriously difficult to obtain. One approach, which plays a prominent role in the field of quantum chaos, makes use of a trace formula. This relates the spectrum of H to the periodic orbits of a classical dynamic system associated with the given quantum system [7]. Usually, such a trace formula is an asymptotic relation that is valid in the limit of short wavelengths (semiclassical limit). However, there are a few exceptional cases in which the relevant trace formula is an identity. Laplacians on flat tori or on manifolds of constant negative sectional curvatures are the most prominent examples, in which the trace formula is essentially the Poisson summation formula or the Selberg trace formula [8], respectively.

Quantum graphs were introduced by Kottos and Smilansky [9] to provide examples for the above setting in which the trace formula takes a particularly simple form, while the spectrum of H is still sufficiently complex. In particular, in extensive numerical calculations it was found that the distribution of eigenvalues in quantum graphs is the same as that of the eigenvalues of random hermitian matrices (with respect to appropriate probability measures i.e. the Gaussian orthogonal, unitary, or symplectic ensemble). Such behavior is generally viewed as one essential characteristic of quantum chaos, since it is usually observed in quantum systems whose classical counterparts show chaotic dynamics. Therefore, quantum graphs have since become popular models in the field of quantum chaos, in which a central issue is to identify and understand fingerprints of classical chaos in the associated quantum systems (see [10]). For example, recently quantum graphs have been among the first model systems in which the random matrix behavior of eigenvalue statistics has been confirmed analytically to a certain extent by making extensive use of the trace formula [11]. In another application of the trace formula, Gutkin and Smilansky were able to answer the question "Can one hear the shape of a graph?" [12][2]. In order to solve this inverse problem they identified a class of quantum graphs that are uniquely characterized through their spectra.

In this context, a quantum graph is usually a finite metric graph (a *network*) equipped with a suitable self-adjoint realization of the differential Laplacian, such that the Hamiltonian is $H = \Delta$. Here, for convenience, the value of Planck's constant is set to $\hbar = 1$. Thus, the network is viewed as a caricature of a manifold and serves as a simplified model for the configuration space of a physical system as it typically occurs in quantum mechanics. The trace formula for quantum graphs relates the spectrum of H to the closed paths on the network. It thus appears that the classical counterpart of the dynamics provided by a quantum graph is given as a motion of a particle along the edges of the network, with nondeterministic decisions taken at the vertices.

As noticed already in [15], quantum graphs can possess discrete symmetry groups. For example, if some edges are of equal length a permutation of these edges may be a symmetry. Since quantum graphs are generally expected to possess typical quantum chaotic properties, in this article we identify situations in which invariant subspaces of quantum graph Hilbert spaces arise, and to relate them to symmetries. One aim is to identify cases in which no invariant subspaces can occur. This question is also related to the inverse problem, since symmetries and invariant subspaces would obstruct a unique specification of a quantum graph in terms of its spectrum.

6.3 Energy Methods for Schrödinger Equations

We now want to discuss at an intuitive level what are the main ideas for the mathematical analysis of the Schrödinger equation on a network. In particular, we will focus our attention on *connected networks*.

We will always consider a network as a (possibly infinite) set of vertices $V := \{v_i : i = 1, 2, \ldots\}$ connected by edges chosen from a set $E := \{e_j : j = 1, 2, \ldots\}$. The connections within the network are encoded in the incoming connectivity function $\Gamma^+ : V \to 2^E$, that expresses which edges $e_j \in \Gamma^+(v_i)$ are directed into the i-th node, and in the outgoing connectivity function Γ^-, defined analogously. These two functions contain the complete information about the topological structure of the oriented network (V, E). Further, all graph-theoretical features of the nonoriented network (V, E) are encoded in the connectivity mapping $\Gamma := \Gamma^+ \cup \Gamma^-$, since for

[2] This problem is related to a famous general question first addressed by Kac in [13] and also discussed in [14].

each i, $\Gamma(v_i) := \Gamma^+(v_i) \cup \Gamma^-(v_i)$ is the set of all edges emerging from v_i. We assume throughout that the network is uniformly locally finite i.e. the number $|\Gamma(v_i)|$ of edges emerging from the vertex v_i is bounded from above by a uniform constant.

We also want to define a *metric structure* of the network, allowing edges to have (possibly different) lengths ℓ_j. After rescaling we will, however, parametrize all edges as intervals $[0, 1]$.

Consider now on each of these edges the one-dimensional, time dependent Schrödinger equation

$$i\frac{\partial}{\partial t}\psi_j(t,x) = -\frac{1}{\ell_j^2}\frac{\partial^2}{\partial x^2}\psi_j(t,x)$$

where the function ψ_j denotes the wave function on the j-th edge e_j. Observe that we introduced the factor ℓ_j^{-2} in order to compensate the edge-dependent scaling. Our aim is to give a rigorous mathematical meaning to this equation, if considered on a network. To do that, one has to prescribe the transition behavior in the nodes. The first (quite natural) assumption is that at each moment the wave functions ψ_j, ψ_ℓ attain the same value (denoted by $\psi(t, v_i)$), if the edges e_j, e_ℓ meet in the node v_i. As a second condition we want to impose conservation of the quantum mechanical probability current in each node. This can, for example, be realized by imposing in each vertex v_i nodal conditions

$$\sum_{j\in\Gamma^+(v_i)} \frac{\partial\psi_j}{\partial x}(t, v_i) = \sum_{j\in\Gamma^-(v_i)} \frac{\partial\psi_j}{\partial x}(t, v_i)\ . \tag{6.5}$$

of Kirchhoff type.

Summarizing, the Schrödinger equation on a network (V, E, Γ) has the form

$$\left\{\begin{array}{rcll} i\partial\psi_j(t,x)/\partial t &=& -\left(1/\ell_j^2\right)\partial^2\psi_j(t,x)/\partial x^2 & e_j \in E, x \in (0,1), t \in \mathbb{R} \\ \psi_j(t,v_i) &=& \psi_\ell(t,v_i) & j,\ell \in \Gamma(v_i), v_i \in V, t \in \mathbb{R} \\ \sum_{j\in\Gamma(v_i)} \partial\psi_j(t,v_i)/\partial x &=& 0 & v_i \in V, t \in \mathbb{R}. \end{array}\right. \tag{6.6}$$

There are several ways to investigate the above differential problem.

First, one can reformulate the system (6.6) as an abstract Cauchy problem

$$\left\{\begin{array}{rcll} i\,\mathrm{d}\psi/\,\mathrm{d}t(t) &=& H\psi(t) & t \in \mathbb{R} \\ \psi(0) &=& \psi_0 & \end{array}\right. \tag{6.7}$$

in the Hilbert space $L^2(0, 1; \ell^2(E))$ of square integrable wave functions taking a value in $\ell^2(E)$. That is, $\psi(t) = \left(\psi_1(t), \dots, \psi_m(t), \dots\right) \in L^2(0, 1; \ell^2(E))$ for all $t \in \mathbb{R}$, a vector of square summable wave functions defined over the edges of the network. Moreover, the action of the Hamiltonian H is given by

$$H\psi := H\begin{pmatrix} \psi_1 \\ \vdots \\ \psi_m \\ \vdots \end{pmatrix} := -\begin{pmatrix} 1/\ell_1^2\,\frac{\mathrm{d}^2\psi_1}{\mathrm{d}x^2} \\ \vdots \\ \frac{1}{\ell_m^2}\frac{\mathrm{d}^2\psi_m}{\mathrm{d}x^2} \\ \vdots \end{pmatrix}.$$

Notice that H does not map the whole space $L^2(0,1;\ell^2(E))$ into itself i.e. it is an *unbounded* operator. In fact, we incorporate the boundary conditions into the domain of H, which is a suitable subspace of $H^2(0,1;\ell^2(E))$. Here $H^2(0,1;\ell^2(E))$ denotes the Sobolev space of twice weakly differentiable, square integrable functions.

More precisely, the domain of H has to be defined as the space of all functions $\psi \in H^2(0,1;\ell^2(E))$ that are continuous on the whole graph and satisfy the Kirchhoff condition (6.5) in the nodes. A direct computation shows that the operator $(H, D(H))$ is self-adjoint and positive definite. In particular, iH generates a unitary group on $L^2(0,1;\ell^2(E))$, which we denote by e^{itH}. That is, there exists a family e^{itH}, $t \in \mathbb{R}$, of (linear, bounded) unitary operators on $L^2(0,1;\ell^2(E))$ such that for all $t, s \in \mathbb{R}$

$$\mathrm{e}^{itH}\mathrm{e}^{isH} = \mathrm{e}^{i(t+s)H} \tag{6.8}$$

and such that the mapping

$$t \mapsto \mathrm{e}^{itH}x \tag{6.9}$$

is continuous for all $x \in L^2(0,1;\ell^2(E))$. For each initial value $\psi_0 \in L^2(0,1;\ell^2(E))$ this group yields the solution of (6.7) in the form $\psi(t,x) = \mathrm{e}^{itH}\psi_0(x)$.

The abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}\psi}{\mathrm{d}t}(t) &= -H\psi(t) \quad t \geq 0 \\ \psi(0) &= \psi_0 \,, \end{cases}$$

is formally obtained by dropping a coefficient $-i$ from the first equation of (iACP). In this way one obtains a parabolic system of heat equations over a network, whose solution is given by the C_0-semigroup e^{-tH}, see [16,17]. By a C_0-semigroup we mean a family of bounded linear operators e^{tH}, $t \geq 0$, on $L^2(0,1;\ell^2(E))$ such that conditions (6.8)–(6.9) are only satisfied for $t \geq 0$.

It is known that the mathematical behavior of parabolic and Schrödinger differential problems is fundamentally different. Nevertheless, we will develop a theory that permits to deduce symmetry properties of the system of linear Schrödinger equations introduced above as an application of so-called energy methods[3] that are typical of parabolic problems.

The basic tool we will use is the following result, [18, § 5.2].

Theorem 6.2 *Let Y be a closed subspace of $L^2(0,1;\ell^2(E))$. Then the following assertions are equivalent.*

(a) The operators of the group e^{itH} map Y into Y;

(b) The operators of the semigroup e^{-tH} map Y into Y.

[3] These are a class of results that allow one to check a property of the evolution equation by showing that a certain condition is satisfied by the energy form.

In other words, the subspace Y is invariant under the time evolution of the diffusion problem if and only if it is invariant under the time evolution of the Schrödinger equation.

Thus, it suffices to characterize those closed subspaces that are invariant under the action of the semigroup e^{-tH} in order to find a class of *symmetries* of the quantum graph.

To this end we consider the space $H^1(0,1;\ell^2(E))$ of weakly differentiable functions and its subspace

$$\mathcal{V} = \left\{\psi \in H^1(0,1;\ell^2(E)) : \psi \text{ is continuous in the nodes}\right\} . \tag{6.10}$$

For this class of functions we introduce an energy form $E : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ by

$$E(\psi,\phi) := \sum_{e_j \in E} \frac{1}{\ell_j^2} \int_0^1 \frac{d\psi_j}{dx}(x) \overline{\frac{d\psi_\ell}{dx}(x)}\, dx . \tag{6.11}$$

Observe that

$$E(\psi,\psi) = \sum_{e_j \in E} \frac{1}{\ell_j^2} \int_0^1 \left| \frac{d\psi_j}{dx}(x) \right|^2 dx ,$$

defines the energy of our system. For functions ψ that vanish in the nodes one easily sees that the value $E(\psi,\psi)$ of the energy functional coincides with the expectation value $(H\psi,\psi)$ of the Hamiltonian H in the state ψ i.e. with the expected outcome of energy measurements if the quantum system is prepared in the state ψ. More precisely, a tedious computation shows that the Hamiltonian satisfies

$$\begin{aligned} D(H) &= \{\psi \in \mathcal{V} : \exists \varphi \in L^2(0,1;\ell^2(E)), E(\psi,\phi) = (\varphi \mid \phi) \quad \text{for all} \quad \phi \in \mathcal{V}\} \\ H\psi &:= \varphi \end{aligned}$$

i.e. it is the *operator associated with E*. By abstract methods going back to Kato and Lions that are, for example, described in [19], one can prove that H is self-adjoint.

All invariance results for e^{tH}, and hence all symmetry properties of (6.6) presented in this note, will be deduced from the following characterization.

Lemma 6.1 *Consider a closed subspace Y of the space $L^2(0,1;\ell^2(E))$. Denote by P the orthogonal projection onto Y. The following assertions are equivalent.*

(a) The solution $\psi(t)$ of the abstract Cauchy problem (6.7) remains in Y for all initial values $\psi_0 \in Y$.

(b) For every $\psi \in \mathcal{V}$ the invariance condition $P\psi \in \mathcal{V}$, and for all $\psi \in Y, \phi \in Y^\perp$ the orthogonality condition $E(\psi,\phi) = 0$ is fulfilled.

The proof of Lemma 6.1 is purely functional analytic. It is based on the self-adjointness of the Hamiltonian and a result due to E.M. Ouhabaz, see [19, Theorem 2.2]. Notice that if Y is invariant under the time evolution of the quantum graph, by Lemma 6.1, the complement $Y^\perp$ is also invariant.

6.4 Symmetries in Quantum Graphs

Typical potential symmetry operations in quantum graphs are of the form

$$\psi \mapsto U\psi \,, \quad (U\psi)_j(x) := \sum_{e_\ell \in E} \Pi_{j\ell}\psi_\ell(g^{-1}x) \tag{6.12}$$

where Π is a permutation matrix interchanging the edges and g is a measure preserving map of the interval $(0,1)$. Hence (6.12) is the equivalent to (6.2) in networks. This clearly is a unitary operator on the Hilbert space $L^2(0,1;\ell^2(E))$, and is indeed a symmetry if it commutes with the time evolution, see (6.4). A simple example, which is a symmetry for every quantum graph, is given by choosing Π as the identity and g the orientation-reversing map $gx = 1-x$. In this case, consider the subspace of $L^2(0,1;\ell^2(E))$ consisting of functions that are invariant under U i.e. their components ψ_j are even with respect to the reflection g. Then, since (6.4) holds, this subspace is invariant under the symmetry U.

Often several symmetries are present, and the associated unitary operators $U_1, U_2, \ldots$ form a group or, more precisely, they are operators of a unitary representation of the symmetry group. This representation can then be decomposed into irreducible subrepresentations. Since the unitary representation operators commute with e^{itH}, the time evolution respects the decomposition of the total Hilbert space into the invariant subspaces.

In order to cover a larger class of examples we now approach the problem from a different perspective, in that we aim to identify invariant subspaces directly. For this purpose we first introduce a particularly relevant class of closed subspaces of $L^2(0,1;\ell^2(E))$ that can be constructed as follows. For a linear closed subspace X of $\ell^2(E)$ consider

$$Y := \left\{ f \in L^2(0,1;\ell^2(E)) : f(x) \in X \quad \text{for a.e.} \quad x \in (0,1) \right\} . \tag{6.13}$$

Thus, a function $f \in L^2(0,1;\ell^2(E))$ belongs to Y if and only if f satisfies pointwise the linear relation defined by X.

Denoting by K the orthogonal projection from $\ell^2(E)$ onto X, the orthogonal projection P_K from $L^2(0,1;\ell^2(E))$ onto Y satisfies

$$(P_K f)(x) = K(f(x)) \quad \text{for almost every} \quad x \in (0,1). \tag{6.14}$$

Recall that orthogonal projections are self-adjoint, hence P_K is an observable of the system. It is also relevant to note that Y is isomorphic to $L^2(0,1;\mathbb{C}^d)$, if X is finite-dimensional with dimension d. Thus, also in this context, finding symmetries helps to reduce the complexity of the system.

Whenever considering the above orthogonal projections and the energy form E defined in (6.11), condition (b) of Lemma 6.1 reads

- $P_K\psi$ is a continuous function over the network whenever ψ is continuous on the network;
- $E(P_K\psi, (I-P_K)\psi) = 0 \quad$ whenever $\quad \psi \in \mathcal{V}$.

In the following these two conditions will be referred to as *admissibility of P_K* and *orthogonality with respect to the lengths ℓ_j*, respectively.

In the following, $\tilde{\mathcal{I}}$ and $\tilde{\mathcal{K}}$ will denote the operators

$$\tilde{\mathcal{I}} := \begin{pmatrix} (I^+)^T \\ (I^-)^T \end{pmatrix} \qquad \text{and} \qquad \tilde{\mathcal{K}} := \begin{pmatrix} K & 0 \\ 0 & K \end{pmatrix} \tag{6.15}$$

where I^+ and I^- are the incoming and outgoing incidence matrices of the network, respectively, describing its connectivity. That is, I^+ is a (possibly infinite) matrix satisfying $I^+_{ij} = 1$ if $e_j \in \Gamma^+(v_i)$, and $I^+_{ij} = 0$ otherwise; I^- is defined accordingly replacing Γ^+ by Γ^-. The following statement is [18, Theorem. 3.4].

Theorem 6.3 *The projection P_K is admissible if and only if the range of $\tilde{\mathcal{I}}$ is $\tilde{\mathcal{K}}$-invariant, i.e., $\tilde{\mathcal{K}}\,\mathrm{Range}\,\tilde{\mathcal{I}} \subset \mathrm{Range}\,\tilde{\mathcal{I}}$.*

Let L now be the diagonal matrix with entries ℓ_j^{-2} on the diagonal and consider the sesquilinear form on $L^2(0,1;\ell^2(E))$ defined as in (6.11). The following is [18, Prop. 3.9].

Theorem 6.4 *The orthogonality condition with respect to the lengths ℓ_j is satisfied, i.e.*

$$E(P_K\psi, (I - P_K)\psi) = 0 \qquad \textit{for all} \quad \psi \in \mathcal{V}$$

if and only if the matrix L leaves the range of K invariant,

$$L\,\mathrm{Range}\,K \subset \mathrm{Range}\,K\,.$$

Thus, the following characterization of time evolutions leaving invariant pointwise proportions is a direct consequence of the above results.

Corollary 6.1 *Let K be an orthogonal projection onto a closed subspace of $\ell^2(E)$ and define a closed subspace Y as in (6.13). Then the following assertions are equivalent.*

(i) The solution $\psi(t)$ of the time evolution (6.7) of the quantum graph belongs to Y for all $t \in \mathbb{R}$ provided that the initial data ψ_0 belongs to Y.

(ii) The lengths ℓ_j, the incidence matrices I^+, I^-, and the proportion matrix K satisfy the conditions

$$\tilde{\mathcal{K}}\,\mathrm{Range}\,\tilde{\mathcal{I}} \subset \mathrm{Range}\,\tilde{\mathcal{I}} \qquad \textit{and} \qquad L\,\mathrm{Range}\,K \subset \mathrm{Range}\,K\,.$$

Corollary 6.1 can be directly applied in several cases, reducing the investigation of an (infinite dimensional) system of partial differential equations to checking that two purely algebraic conditions are satisfied. As an easy, yet nontrivial, example we mention the following.

Example 6.1 *Consider a finite oriented star, i.e. a graph with a single "central" node and m outgoing edges of lengths $\ell_1, \ell_2, \ldots, \ell_m$. The projection P_K is admissible if and only if the vector $\mathbb{1} = (1, 1, \ldots, 1)^T$ is either in Y or in $Y^\perp$, as the matrices I^+ and I^- have range span$(\mathbb{1})$ and $\mathbb{C}^m$, respectively.*

Moreover, there exist nontrivial[4] *subspaces X with $\mathbb{1} \in X$ or $\mathbb{1} \in X^\perp$ such that $\tilde{\mathcal{K}}$ Range $\tilde{\mathcal{I}} \subset$ Range$\tilde{\mathcal{I}}$, if and only if at least two lengths in the graph are equal, i.e. there exist $i \neq j$ such that $\ell_i = \ell_j$.*

We hence conclude that, on a finite star graph, nontrivial invariant subspaces of the kind discussed above exist if and only if at least two edges have the same length.

6.5 Schrödinger Equation with Potentials

We now briefly discuss the generalization to Hamiltonians associated with the forms $\tilde{E} : \mathcal{V} \times \mathcal{V} \to \mathbb{C}$ defined by

$$\tilde{E}(\psi, \phi) = -\sum_{e_j \in E} \frac{1}{\ell_j^2} \int_0^1 \left(-i\frac{d\psi_j}{dx} + a_j\psi_j\right)\overline{\left(-i\frac{d\psi_j}{dx} + a_j\psi_j\right)} + v_j(x)\psi_j(x)\overline{\phi_j(x)}\, dx\,.$$

These correspond to a Schrödinger equation on a quantum graph with electric and magnetic potentials v_j, a_j, respectively. Notice that we allow for different potentials on different edges of the graph. We can extend the invariance results of the preceding section to this more general setting, which is also discussed in [20]. For convenience, assume the potentials a_j, v_j of class $L^\infty(0, 1; \mathbb{R})$ and $L^2(0, 1; \mathbb{R})$, respectively, for all j. In fact, we only have to check the conditions of Lemma 6.1 for this form. Again, one can prove that the operator $\tilde{H}$ associated with the symmetric form $\tilde{E}$ is self-adjoint. Thus, it can be considered as a new Hamiltonian and it generates a unitary group, denoted by $e^{it\tilde{H}}$.

In the statement of the following theorem L, $A(x)$ and $V(x)$ denote the diagonal matrices with entries ℓ_j^{-2}, $a_j(x)$ and $v_j(x)$, $x \in (0, 1)$, respectively.

Theorem 6.5 *Let K be an orthogonal projection of $\ell^2(E)$ and define a closed subspace Y as in (6.13). Assume the lengths ℓ_j, the incidence matrices I^+, I^-, the potential functions v_j, and the proportion matrix K satisfy the conditions*

$$\tilde{\mathcal{K}}\,\text{Range}\,\tilde{\mathcal{I}} \subset \text{Range}\,\tilde{\mathcal{I}}, \qquad L\,\text{Range}\,K \subset \text{Range}\,K, \quad \textit{as well as}$$
$$A(x)\text{Range}\,K \subset \text{Range}\,K \quad \textit{and} \quad V(x)\text{Range}\,K \subset \text{Range}\,K$$

for all $x \in (0, 1)$. Then the solution $\psi(t)$ of the Schrödinger equation with magnetic and electric potentials on the quantum graph belongs to Y for all $t \in \mathbb{R}$ provided that the initial data ψ_0 belongs to Y.

[4] That is, different from $X = \mathbb{C}^m$ and $X = \{0\}$.

Proof We first have to check admissibility of the orthogonal projection P_K associated with K. Let $\psi \in \tilde{\mathcal{V}}$. By the theorem of Pythagoras

$$|\psi(x)|^2 = |K\psi(x)|^2 + |(I-K)\psi(x)|^2 \qquad \text{for all} \quad x \in (0,1)$$

and therefore

$$\sum_{e_j \in E} \int_0^1 v_j(x) \left|(P_K\psi)_j(x)\right|^2 \mathrm{d}x \le \sum_{e_j \in E} \int_0^1 v_j(x) \left|\psi_j(x)\right|^2 \mathrm{d}x < \infty$$

since $v_j \in L^2(0,1)$ for all j. Similarly one shows that the assumption $a_j \in L^\infty(0,1)$ for all j implies integrability of the magnetic potential term. Thus, also due to the inclusion $\tilde{\mathcal{K}}$ Range $\tilde{\mathcal{I}} \subset$ Range $\tilde{\mathcal{I}}$, the admissibility condition $P\psi \in \tilde{\mathcal{V}}$ follows. Furthermore, since P_K commutes with $\mathrm{d}/\mathrm{d}x$, due to the self-adjointness of P and V we have

$$\begin{aligned}
&\sum_{e_j \in E} \frac{1}{\ell_j^2} \int_0^1 \left(-i \left(P_k \frac{\mathrm{d}\psi}{\mathrm{d}x} \right)_j \overline{a_j((I-P_K)\psi)_j} + a_j(P_K\psi)_j \overline{\left(-i \left((I-P_K) \frac{\mathrm{d}\psi}{\mathrm{d}x} \right)_j \right)} \right) \\
&\quad = -i \left((I-P_K)LAP_K \frac{\mathrm{d}\psi}{\mathrm{d}x} \mid \psi \right)_{L^2(0,1;\ell^2(E))} + i \left(\psi \mid P_K LA(I-P_K) \frac{\mathrm{d}\psi}{\mathrm{d}x} \right)_{L^2(0,1;\ell^2(E))} \\
&\quad = 0
\end{aligned}$$

because, by assumption, $LAP_K \text{Range} K \subset \ker(I-K)$. Under the above assumptions it can thus be shown that

$$\tilde{E}(P_K\psi, (I-P_K)\psi) = 0$$

in a manner similar to that of Theorem 6.4. Again by Lemma 6.1 we conclude that Y is left invariant by $\mathrm{e}^{-it\tilde{H}}$. □

6.6 Concluding Remarks and Open Problems

Example 6.1 has provided us with a simple model whereby we can verify that an obvious symmetry of the network is preserved by solutions of the corresponding Schrödinger equation. In that particular situation it is very easy to apply the criteria we have introduced. But as a rule of thumb most common-sense symmetries can be represented and analyzed in term of some invariant subspace of the state space.

To give a slightly more complex example than a star, consider a rooted complete binary tree of finite depth. The left and the right subtree of the root are identical, hence interchangeable. Thus the subspace of functions whose values on the right subtree are the mirror image of the values on the left (i.e. functions symmetric with respect to the root node), is an invariant subspace, i.e. a symmetry in the sense of this article, which can easily be checked using the above characterizations.

There are a lot more invariant subspaces to find, as there obviously are a lot more symmetries in a complete binary tree.

Observe that strange phenomena occur when the graph fails to be locally finite. For example, consider an (outward oriented) graph with infinitely many incident edges. Since each function in V as in (6.5) is of class $H^1(0, 1; \ell^2(E))$, one has $f(0) = (f_1(0), f_2(0), \dots) \in \ell^2(E)$, and due to continuity in 0, $f_1(0) = f_2(0) = \dots$, hence $f(0) = 0$. That is, a function in the form domain $\mathcal{V}$ necessarily vanishes in the center of the star. In particular, if the support of the initial data ψ_0 of (6.7) lies in a single edge e_j of the infinite star, then the solution $\psi(t)$ will also have support contained in the edge e_j only. This suggests that the energy form E with domain $\mathcal{V}$ in the phase space $L^2(0, 1; \ell^2(E))$ does not lead to the correct solution *Ansatz* for the Schrödinger equation on such quantum graphs and motivates one to consider weighted spaces instead.

Further, infinite graphs are interesting from a spectral theoretic point of view. While the free Schrödinger equation on the real line has a purely essential spectrum, it has a purely discrete spectrum on a finite network. Nevertheless, the Schrödinger equation on an infinite network features essential spectrum and eigenvalues alike, see [21, Sections 4–5].

Let us also mention that two noncongruent, isospectral graphs have been exhibited in [12]. The construction exploits the existence of a discrete group of symmetries of a specific graph. This relates to the well-known question addressed by Kac, [13]. A more general theory linking discrete symmetries and isospectrality has recently been announced by Band, Shapira and Smilansky in [22].

We also remark that, in the previous sections, we have restricted ourselves to the most basic example of an energy form that is related to a Schrödinger operator. However, the techniques can be applied to a variety of related problems, for example, allowing nonlocal coupling of the edges, or different behavior at the nodes, like an absorption term or nonlocal interactions. Most of these modifications have in common that they do not change the domain of the energy functional, thus leading to the same notion of admissibility as for our model case. On the other hand, the form itself changes significantly, hence our interest then lies in the second condition which we above called "orthogonality". Some of these extensions have been investigated in [18].

A class of nodal conditions that does lead to the consideration of a different domain of the energy functional (and also a larger phase space) consists of *dynamic* ones like

$$\frac{\partial \psi}{\partial t}(t, v_i) = \sum_{j \in \Gamma^+(v_i)} \frac{\partial \psi_j}{\partial x}(t, v_i) - \sum_{j \in \Gamma^-(v_i)} \frac{\partial \psi_j}{\partial x}(t, v_i) \, . \tag{6.16}$$

These are, among others, motivated by the results concerning approximation of shrinking thin domains, [23, 24]. A variational approach to such problems based on energy methods has been performed in [25]: in fact, it can be shown that the results on symmetry discussed in Section 6.4 also hold when the Kirchhoff-type condition (6.5) is replaced by (6.16) (for example, whenever a star with a dynamic condition in the central node is considered).

Another interesting issue can be formulated as follows. *Does a quantum graph admit local symmetries?* Adopting the notation of Section 6.4, if we consider a family of orthogonal projections K_x, $x \in [0,1]$, then we also have a family of closed subspaces X_x of $\ell^2(E)$, $x \in [0,1]$. If these families are constant, and hence we have a single projection K, then we end up with the theory of global symmetries presented in Section 6.4. In fact, it is possible to show that, if the subspace

$$Y := \{f \in L^2(0,1;\ell^2(E)) : f(x) \in X_x \quad \text{for a.e.} \quad x \in (0,1)\}$$

is invariant under time evolution of the quantum graph, then necessarily $X_x \equiv X$. Thus, we cannot investigate this more general class of symmetries by means of Lemma 6.1. Searching for local symmetries of a graph requires sophisticated methods in order to determine the covariant derivative associated with a suitable gauge field. This is work in progress.

References

1 Noether, E. (**1918**) Invariante variationsprobleme. *Gött. Nachr.*, 235–237.

2 Bourbaki, N. (**1975**) *Elements of mathematics. Lie groups and Lie algebras.* Adiwes International Series in Mathematics. Hermann, Paris.

3 Liouville, J. (**1855**) Note sur l' intégration des équations différentielles de la dynamique. *J. Math. Pure Appl.*, **20**, 137–138.

4 Einstein, A. (**1917**) Zum Quantensatz von Sommerfeld und Epstein. *Verh. der Deutsch. Phys. Ges.*, **19**, 82–92.

5 Wigner, E.P. (**1959**) Group theory: And its application to the quantum mechanics of atomic spectra. Expanded and improved ed. Translated from the German by J.J. Griffin. *Pure and Applied Physics*, Vol. 5. Academic Press, New York.

6 Bohigas, O., Giannoni, M.-J. and Schmit, C. (**1984**) Characterization of chaotic quantum spectra and universality of level fluctuation laws. *Physical Review Letters*, **52**(1), 1–4.

7 Gutzwiller, M.C. (**1990**) Chaos in classical and quantum mechanics, volume 1 of *Interdisciplinary Applied Mathematics.* Springer-Verlag, New York.

8 Selberg, A. (**1956**) Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J. Indian Math. Soc. (N.S.)*, **20**, 47–87.

9 Kottos, T. and Smilansky, U. (**1999**) Periodic orbit theory and spectral statistics for quantum graphs. *Annual Physics*, **274**(1), 76–124.

10 Haake, F. (**2001**) *Quantum Signatures of Chaos.* Springer Series in Synergetics. Springer-Verlag, Berlin, second edition. With a foreword by H. Haken.

11 Berkolaiko, G., Schanz, H. and Whitney, R.S. (**2003**) Form factor for a family of quantum graphs: an expansion to third order. *J. Phys. A*, **36**(31), 8373–8392.

12 Gutkin, B. and Smilansky, U. (**2001**) Can one hear the shape of a graph? *J. Phys. A*, **34**(31), 6061–6068.

13 Kac, M. (**1966**) Can one hear the shape of a drum? *The American Mathematical Monthly*, **73**, 1–23.

14 Arendt, W., Nittka, R., Peter, W. and Steiner, F. (**2009**) Weyl's formula: Spectral properties of the Laplacian in mathematics and physics, in *Mathematical Analyisis of Evolution, Information and Complexity*, (eds W. Arendt and W. Schleich) Wiley-VCH, Weinheim.

15 Exner, P. and Šeba, P. (**1989**) Free quantum motion on a branching graph. *Rep. Math. Phys.*, **28**, 7–26.

16 von Below, J. (**1985**) A characteristic equation associated with an eigenvalue problem on C^2-networks. *Lin. Algebra Appl.*, **71**, 309–325.

17 Kramar Fijavž, M., Mugnolo, D., and Sikolya, E. (**2007**) Variational and semigroup methods for waves and diffusion in networks. *Appl. Math. Optim.*, **55**, 219–240.

18 Cardanobile, S., Mugnolo, D. and Nittka, R. (**2008**) Well-posedness and symmetries of strongly coupled network equations. *J. Phys. A*, **41**, 055102.

19 Ouhabaz, E.-M. (**2004**) *Analysis of Heat Equations on Domains*, volume 30 of *LMS Monograph Series*. Princeton University Press, Princeton.

20 Kuchment, P. (**2005**) Quantum graphs II: Some spectral properties of quantum and combinatorial graphs. *J. Phys. A*, **38**, 4887–4900.

21 Cattaneo, C. (**1997**) The spectrum of the continuous Laplacian on a graph. *Monatsh. Math.*, **124**, 215–235.

22 Band, R., Shapira, T. and Smilansky, U. (**2006**) Nodal domains on isospectral quantum graphs: the resolution of isospectrality? J. Phys. A: Math. Gen. 39 13999–14014.

23 Kuchment, P. and Zeng, H. (**2003**) Asymptotics of spectra of Neumann Laplacians in thin domains, in *Advances in differential equations and mathematical physics (Proc. Birmingham 2002)*, (ed Y. Karpeshina *et al.*) volume 327 of *Contemp. Math.*, Providence. American Mathematical Society, 199–213.

24 Exner, P. and Post, O. (**2005**) Convergence of spectra of graph-like thin manifolds. *J. Geom. Phys.*, **54**, 77–115.

25 Mugnolo, D. and Romanelli, S. (**2007**) Dynamic and generalized Wentzell node conditions for network equations. *Math. Meth. Appl. Sci.*, **30**, 681–706.

7
Distributed Architecture for Speech-Controlled Systems Based on Associative Memories

Zöhre Kara Kayikci[1], Dmitry Zaykovskiy[1], Heiner Markert[1], Wolfgang Minker, Günther Palm

7.1
Introduction

Distributed speech recognition enables people to have easy and flexible access to the full range of computer systems, without the need to be able to type or to be in front of a computer. Around this objective, we have developed a distributed speech recognition architecture that can easily be adapted to various speech-controlled systems. It can be divided into four parts: feature extraction, Hidden Markov Model based subword unit recognition, associative memory based word recognition and semantic parsing. In the architecture, the feature extraction part can be carried out on a mobile device or on a server, depending on the application. The architecture was designed by combining the advantages of distributed speech recognition and associative memories such as flexible usage in different environments, handling of ambiguities, and simple and fast incrementing of task vocabulary.

Starting in the early 50s, speech understanding made significant progress from the isolated word recognizers to modern speaker-independent systems capable of dealing with natural language. The rapid development of hardware and software technologies during the last decades gave rise to such concepts as speech-to-text processing, speech-controlled appliances, spoken language dialogue systems, and speech-enabled services. Common to all these systems is the use of the human voice for human–machine communication.

In the present work we focus on a wide class of applications involving *remote* access to the robotic or information systems using speech. Among others, typical examples of such applications are remote home automation, pedestrian navigation, or ticket-reservation systems. These are scenarios where the user has to operate a device which is too weak to perform speech recognition itself and the paradigm of remote speech recognition has to be employed. Our approach merges interdisciplinary technologies from engineering science, communication technology and neuro-informatics.

[1] Corresponding authors.

State-of-the-art automatic speech understanding systems are commonly installed on powerful computers and make use of the Hidden Markov Models (HMMs), a very flexible way of modeling speech data vectors of variable length, which are typically used to model the speech signals for every word known to the system. In particular, recognition systems based on HMMs are effective and allow for good recognition performance under many circumstances, but suffer from some limitations concerning increasing dictionary size and robustness to environmental conditions. Starting from the late 1980s, HMMs and artificial neural networks (ANNs) have been combined within a hybrid architecture and a variety of different approaches have been proposed in the literature in order to overcome these limitations.

An ANN operates by creating connections between many processing elements like neurons. These neurons can be simulated by a digital computer. Each neuron takes many input signals and produces a single output signal that is typically sent as input to other neurons. The neurons can be fully or partly interconnected and are typically organized into different layers. The input layer receives the input and the output layer produces the output. Usually one or more hidden layers are used in between. An ANN realizes a mapping between an input space and an output space, which can be specified by learning from a finite set of patterns. Due to their pattern-matching and learning capabilities, artificial neural networks have proven useful in a variety of real-world applications that deal with complex, often incomplete data. The first of these applications were in pattern recognition and speech recognition, in particular.

Early approaches to hybrid speech recognition were based on ANN architectures that attempted to emulate HMMs [1]. In some ANN/HMM hybrids, ANNs are used to estimate the HMM state-posterior probabilities from the acoustic observations [2]. In other approaches [3], the ANN is used to extract observation feature vectors for a HMM.

In contrast to these approaches, in the proposed architecture, HMMs are used on the elementary phonetic level and the neural associative memories (NAMs) are used on higher levels such as words and sentences. A NAM is the realization of an associative memory in a single-layer artificial neural network. The proposed architecture yields both the advantages of distributed speech recognition and of artificial neural networks (ANNs).

In our contribution we concentrate on two aspects, which are different from the classical systems. First, we will show how the generic architecture can be modified to support remote speech processing. Secondly, we demonstrate how HMM-based systems can be extended by NAMs.

The chapter is organized as follows. In Section 7.2 we briefly review the conventional state-of-the-art automatic speech-understanding systems, provide insights into the building blocks of our system and introduce the proposed system architecture. The third section targets speech processing on mobile devices. Considering the feature extraction process on handhelds we review the required processing steps and discuss implementation issues. The results of experiments with real mobile devices are also given there. Section 7.4 presents a speech recognition system based on associative memories. The system based on HMMs for features-to-phonemes conversion and on Willshaw's model of neural associative memory for phoneme-to-word mapping is introduced. Handling ambiguities like unclear sequence of phonemes generated by Hidden Markov Models and online learning of new patterns are also presented here. The following section focuses on the extraction of the semantic meaning from recognized text, i.e. words-to-semantic conversion. Unlike Willshaw's simple model of associative memory from the previous section, the so-called spike counter model of associative memory is used for words-to-semantic conversion. Section 7.6 demonstrate the functionality of the architecture on some sample tasks. The presented architecture is compared with Hidden Markov Models on word level for different speech corpora. Finally, the last section presents some conclusions.

7.2 System Architecture

In order to understand the architecture of the proposed system, let us first consider the structure of a generic speech-controlled system as shown in Figure 7.1. The goal of a speech-controlled system is to initiate certain actions triggered by the human voice. To make this possible, a multi-stage analysis of the spoken utterance is performed most commonly as follows.

The first stage is the *feature extraction*. At this stage the digitalized speech signal is converted into a sequence of vectors carrying a transformed version of the original data suitable for further processing. Feature extraction may be viewed as a special form of dimensionality reduction, that is the methodology which allows one to shorten the dimension of the classifier (speech recognizer in our case) input without much loss in classification (recognition) accuracy.

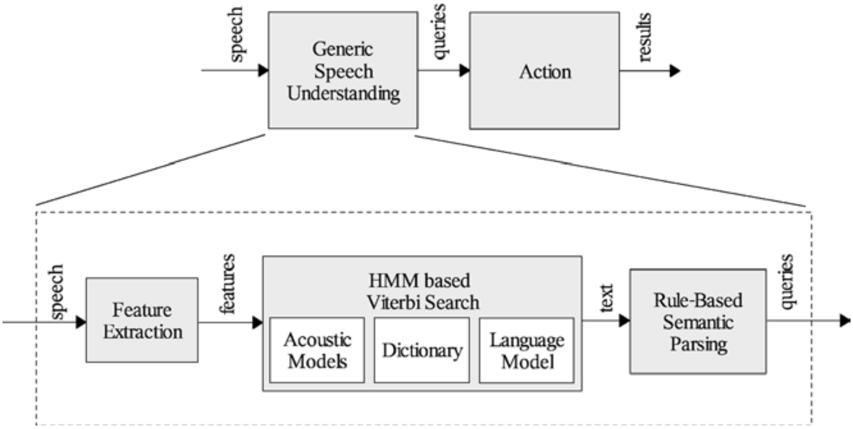


Figure 7.1 Generic architecture of a speech-controlled system.

At the second stage the sequence of feature vectors is mapped into a sequence of words. This process is called automatic speech recognition (ASR) or *ASR decoding*. It is commonly performed by searching for the most probable sequence of words on the graph of all possible alternatives. The search graph is built using a language model (which describes the probability of the given combination of words) and a dictionary (which specifies how words are composed of phonemes). The optimal path is found by applying the Viterbi algorithm, which scores the observed features on the search graph by the acoustic model. The latter describes each phonetic unit as Hidden Markov Models (HMMs) [4]. This type of system attempts to match subword units, words and sentences in one step.

The final stage of the speech understanding process is *semantic parsing*. Here certain linguistic analysis is performed to extract the semantic content of the utterance based on the recognized text. Commonly some set of predefined grammatical rules is used. The results of this stage are data structures (queries) containing spoken information in a form suitable for processing.

Figure 7.2 shows the structure of the speech-controlled system based on associative memories (AM), which we suggest. As in the case of generic architecture, first the feature extraction takes place. However, the process of the features-to-text conversion is now performed in two steps: subword unit matching and word matching.

First of all the sequence of the feature vectors is mapped into a sequence of some subword units. Generally these units can be any phonetic entities, for example, phonemes, diphones, triphones or syllables. In order to perform features-to-subword unit conversion we use the HMM-based ASR system as before. However, the Viterbi search is performed not on the graph containing the possible combinations of words, but on the graph having subword units as nodes. This can be easily implemented if we define a new, trivial dictionary containing single subword units as words. The language model should then describe the probability for a subword

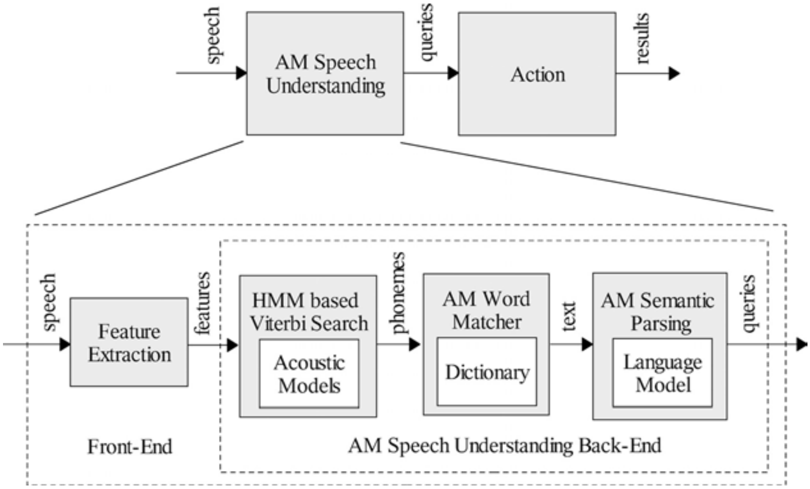


Figure 7.2 Architecture of a speech-controlled system based on associative memories.

unit combination to occur.[2] The acoustic models undergo no change compared to the former system.

The obtained sequence of subword units then serves as an input for the second module – an associative memory-based subword units-to-words converter. This word recognizer consists of a number of interconnected neural auto- and hetero-associative memories. The associative memories are trained such that they reflect the information contained in the original dictionary. Using the activation levels of neurons, the recognizer outputs the most probable word for the given sequence of phonetic units.[3] The detailed procedure is described in Section 7.4.2.

The sequence of words generated by the word recognizer is forwarded to the semantic parser, which is also built on using the neural associative memories. In this case the neural associative memories store the information concerning the language model and syntax. After parsing, the system is able to assign the relevant recognized words to the task-specific categories like color, date, destination, action, etc. Using such category-value pairs allows the initiation of the required actions.

The proposed system architecture uses a speech-understanding engine based on the associative memory. However, contrary to the previous case the overall setup is now distributed (see Figure 7.3), by spreading over the client and the remote server. The following will give a short description of the system.

First, the speech signal is captured on the client device. Then the acoustic features are extracted from the speech signal. Finally, the compression of the feature vectors takes place. This results in a bit-stream to be packaged and transmitted. Since via transmission over the wireless channel some data can be corrupted, some mitigation algorithm against the effect of transmission errors has to be applied on the server side [5].

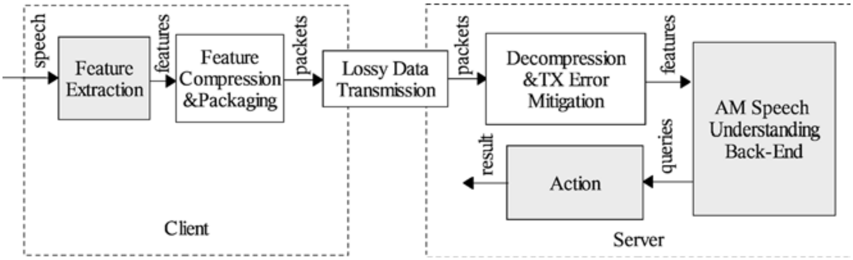


Figure 7.3 Architecture of a distributed speech-controlled system based on the associative memories.

[2] Note that this language model and dictionary are not the same as those pictured in Figure 7.1.

[3] Since the sequence of words produced on the previous step is generated without use of a language model, it can generally be grammatically inconsistent.

7.3 Feature Extraction on Mobile Devices

Our system adopts a client-server architecture. However, the successful deployment of distributed speech-controlled systems is only possible in practice, if both the client and the server assume the same standardized procedure for the feature extraction and feature compression. A number of such standards [6–9] was developed by the Aurora working group established within the European Telecommunications Standards Institute (ETSI). In our system we use the front-end conformed with ETSI standard ES 201 108.

7.3.1 ETSI DSR Front-End

The ETSI standard ES 201 108 defines the feature extraction algorithm, feature compression algorithm, the mechanism for transmission error protection and the feature packaging process [6].

7.3.1.1 Feature Extraction

After capturing the analog speech signal the relevant features have to be extracted. The waveform itself is not suitable for speech recognition since it is rather redundant. In order to obtain the feature vectors, the signal is divided into sections (frames) of several milliseconds. For each frame a spectral analysis is performed. This results in a parameteric representation of the spectral content of the signal.

The process of feature extraction is shown in Figure 7.4. First, the speech signal is captured from the devices' microphone as a 8 kHz 16 bit/sample data stream. After segmenting the speech data into overlapping frames of 25 ms, an offset filter is applied to remove the constant offset. For each frame, the logarithm of the energy is computed. Then the high-frequency signal components are emphasized using a digital filter. For better spectral resolution the frame is weighted with a Hamming

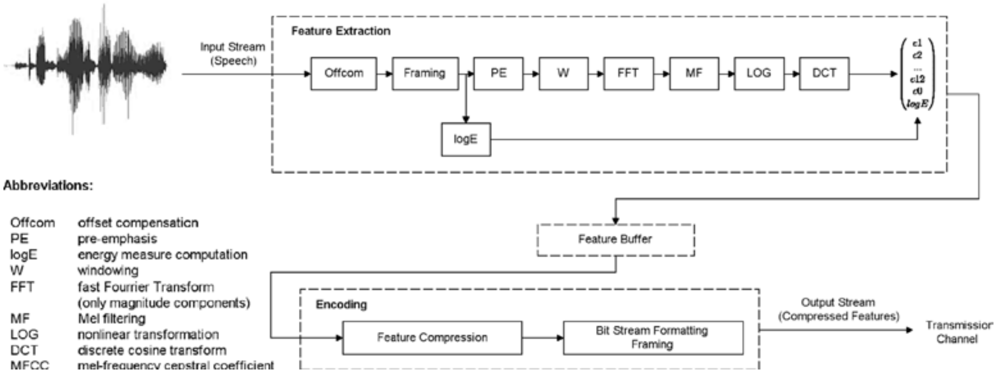


Figure 7.4 Detailed architecture of the client side.

window. In order to compute the magnitude spectrum of the frame, an FFT of length 256 is applied. The mel filter block divides this spectrum into 23 equidistant, weighted, half-overlapping channels in the mel frequency domain using triangular-shaped frequency windows in the range between 64 Hz and 4000 Hz. A nonlinear transformation is applied by taking the logarithm of the 23 channel-values. The obtained values are subject to a discrete cosine transform resulting in 13 cepstral coefficients. The cepstral coefficients together with the log-energy value build up a feature vector.

7.3.1.2 Feature Compression

After feature extraction, one second of the speech is represented by 100 feature vectors, while a single feature vector consists of 14 elements described by four bytes each. Hence the feature vectors can be represented by 44.8 kbit s^{-1}, which already constitutes an impressive reduction in the data rate compared with 128 kbit s^{-1} for the raw speech signal. However, further reduction of the data rate is possible by means of the split-vector quantization algorithm.

The principle of two-dimensional vector quantization is illustrated in Figure 7.5. Prior to the actual quantization a set of reference points (codebook) has to be defined in order to quantize some point from the two-dimensional space and the distance to each reference vector is computed. The closest reference point will be used instead of the original vector. Note that, if the codebook is also known on the server side, it is sufficient to transmit only the index of the corresponding reference point. This means that with a codebook of size 64, quantization of a couple of cepstral coefficients is possible using only 6 bits. The standard ES 201 108 specifies

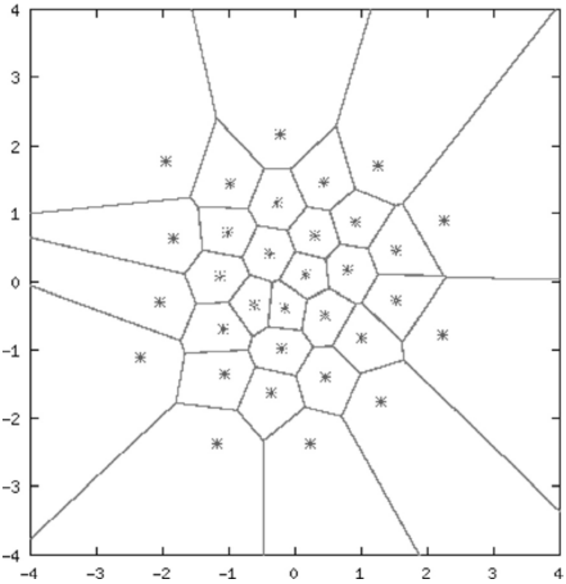


Figure 7.5 Vector quantization based on space clustering.

codebooks of size 64 for cepstral coefficients $(c_1, c_2) \ldots (c_{11}, c_{12})$ and of size 256 for the pair $(c_0, \ln E)$. The final data rate constitutes 4800 bit s^{-1}.

7.3.2 Implementation of the Front-End on Mobile Phones

As a framework for the implementation of the client side we have used the Java Platform Micro Edition also known as Java ME. Even though there are some alternatives, for example, Symbian OS, Windows Mobile, .NET Compact Framework, de facto Java ME is the only solution for the development of software for consumer cell phones.

In addition to the fact that Java by itself is meant to be a slow programming language, the development of the application for mobile phones is further challenged by the limited resources of the handhelds. The most critical issues are the small amount of available memory and the slow processors of cellular phones. In this regard the use of multi-threading and fixed-point arithmetic has been shown to be beneficial.

7.3.2.1 Multi-Threading

The client front-end shown in Figure 7.4 consists of two main blocks: feature extraction and encoding. Both modules can be launched either sequentially (single-threading) or in parallel (multi-threading).

In a single-threaded architecture the encoder waits until all the feature vectors are computed and saved in a buffer. Thus, for long utterances, the feature buffer may become very large. In the case of multi-threading the encoder works in parallel with the feature extraction block and processes the feature vectors upon availability. This might require some additional processing power, but saves memory usage considerably.

7.3.2.2 Fixed-Point Arithmetic

In general, mobile phone processors do not have a floating-point unit. This implies that the operations with real values are software-emulated, which leads to an additional processing time.

This problem can be resolved by using fixed-point arithmetic, where a fixed number of digits before and after the radix point is used. Since most programming languages, including Java, do not support a native data type for fixed-point real values, this has to be simulated by using integer values. The floating-point values are first scaled up by a certain factor, then the remaining fractional part is truncated and the time-efficient operation on the integer values is performed. Finally, the result is scaled down. Example: $12.345+0.11|\cdot 100$ and truncate $\Rightarrow 1234+11 = 1245|\cdot 1/100 \Rightarrow 12.45$. Even though this may lead to a precision loss, our experiments revealed that the use of fixed-point arithmetic for the feature extraction does not effect recognition accuracy.

7.3.2.3 Processing Time on Real Devices

In order to assess the processing time required for feature generation we have performed a number of tests, where the same spoken utterance was processed by different mobile appliances. Some of the results are presented in Table 7.1. The processing time is expressed as a real-time factor – the time required for the processing divided by the utterance duration.

As it is evident from the table the multi-threaded architecture requires only minor additional processing time. It is also clear that on most of the tested devices the use of the fixed-point arithmetic can significantly shorten the processing time (by a factor of four in some cases). This makes real-time feature extraction possible on devices like the Nokia N70 or the Nokia 6630. It is also interesting that on some Nokia phones the front-end using floating-point arithmetic outperforms the one with fixed-point. This may be explained either by an installed FPU on those devices or our fixed-point algorithm has some operation being computationally costly on these devices.

Table 7.1 Time required for feature extraction (FE only) and compression (FE+VQ) related to the utterance duration.

Name of cellular phone	FE		FE+VQ			
	only		Single-Thread		Multi-Thread	
	Float	Fixed	Float	Fixed	Float	Fixed
Nokia 6630, N70	1.3	0.7	1.8	0.9	2.0	1.4
Nokia E70	1.3	0.9	1.8	1.2	1.9	1.3
Nokia 7370	1.2	2.7	1.6	3.7	1.7	3.8
Nokia 7390	0.9	1.6	1.3	2.2	1.4	2.3
Nokia 6136, 6280, 6234	1.1	2.2	1.5	3.0	1.5	3.1
Siemens CX65, CX75	3.1	2.1	4.4	2.7	5.0	3.8
Sony-Ericsson W810i	7.9	2.0	12.5	2.9	13.4	3.1

7.4 Speech Recognition Systems Based on Associative Memory

Unlike conventional back-end processing of speech recognition, the feature for word conversion on the remote server are split into two parts as shown in Section 2; features to subword units (such as context-dependent phonemes, syllables) conversion and subword units to words conversion. The features to subword units conversion is performed by Hidden Markov Models (HMMs) [4], a statistical speech-modeling method, whereas the subword units to words conversion is done using neural associative memories (such as Willshaw's model) [10]. The last part of the speech recognition system is a parsing process in which word sequences are analyzed with respect to a set of grammar rules extracting the semantic information of spoken commands, which can then be used in control systems. The speech recognition system also has ability to handle ambiguities at the subword unit and word levels due to the fault-tolerance of neural associative memories. Further learning of new command words in real time is possible by adjusting the synaptic connections in the associative memories accordingly, given that the HMMs generate a plausible subword-level representation for the new command word.

7.4.1 Features to Subword Units Conversion using HMMs

Given a sequence of feature vectors, the next stage in speech understanding is mapping between the sequences of feature vectors and the underlying sub-symbol sequences (such as context-dependent phonemes or syllables). This is implemented using HMMs. For a given sequence of feature vectors, the subword unit recognition system should determine the best sequence of sub-symbols that maximizes the a posteriori probability:

$$\underset{S}{\operatorname{argmax}}\{P(S|O)\}\ . \tag{7.1}$$

Using Bayes' Rule, (7.1) can be written as

$$P(S|O) = \frac{P(O|S)P(S)}{P(O)}\ , \tag{7.2}$$

where S is the sequence of sub-symbol models and O is the sequence of feature vectors. The first term in (7.2), $P(O|S)$, is generally called the *acoustic model*, as it estimates the probability of a sequence of acoustic observations. The second term, $P(S)$ is commonly referred to as the *language model*, describing the probability associated with a postulated sequence of sub-symbols [4].

7.4.1.1 Acoustic Models

In acoustic modeling, each sub-symbol is modeled with a HMM using a three-state left-right topology with no skips, shown in Figure 7.6.

The output probability is specified by a Gaussian distribution [4]. For this study, context-dependent phonemes (such as triphones) are used as basic acoustic models. While generating context-dependent phoneme (triphone) models, monophone HMMs are first created and their parameters are estimated with the flat-start Baum–Welch reestimation strategy [4]. The models are then cloned to yield the context-dependent phoneme models which are defined as $p_L - p + p_R$, where p_L is the phoneme preceding p and p_R is the phoneme following p. Based on this

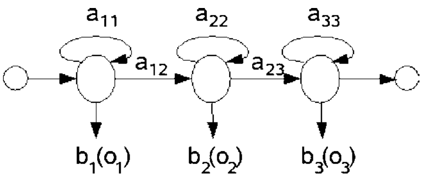


Figure 7.6 Left-right three-state HMM. a_{ij} is the transition from state i to state j and $b_j(o_t)$ is the output probability of being in state j at time t.

definition the first state of a HMM represents the transition from p_L to p, its middle state represents the center of p and its last state represents the transition from p to p_R. Due to insufficient data associated with many of the triphone states, the states within triphone models are then tied to share data to make robust parameter estimates using the Baum–Welch re-estimation strategy [4]. At the last stage of the triphone-model generation, the output distributions of the HMMs are approximated by eight Gaussians per state.

7.4.1.2 Language Model and Dictionary

The goal of the statistical language model is to provide an estimate of the probability of a subword unit sequence:

$$\widehat{P}(w_1, w_2, ..., w_m) = \prod_{i=1}^{m} \widehat{P}(w_i | w_1, ..., w_{i-1}) \ . \tag{7.3}$$

However, it seems to be impossible to reliably estimate the conditional probabilities for all sequence lengths in a given language. Therefore, n-gram language models, which predict each symbol in the sequence given that its $n - 1$ predecessors, are used. However, even n-gram probabilities are difficult to estimate reliably. Hence, in practice, the bi-gram or tri-gram models are applied.

A subword unit recognizer also requires a dictionary which contains an entry for each subword unit and a corresponding triphone-level transcription.

The most probable sequence of subword units is obtained by using a Viterbi search on the recognition network compiled from the language model, the dictionary and a set of acoustic models [4]. This sequence is then forwarded to the next stage where subword units to words conversion is performed.

7.4.2 Subword Units to Words Conversion using Neural Associative Memory

In this sub-section, we will introduce a neural associative memory-based approach to the subword units to words conversion. In this approach a subword unit sequence generated by HMMs is applied to the architecture composed of a number of binary neural associative memories to retrieve the best matching word sequence.

7.4.2.1 Neural Associative Memories

An associative memory is a system that stores patterns and associations between pairs of patterns, which can be represented as vectors with binary components [11]. A neural associative memory is the realization of an associative memory in a single-layer artificial neural network, where neurons are interconnected by binary synapses [12, 13]. In this framework, Willshaw's model, shown in Figure 7.7, is used as the basic architecture of binary neural associative memories [10, 14].

In *heteroassociative memories*, a mapping $m : x \rightarrow y$ is stored. This is called *pattern mapping*, where x is the input pattern and y is the content (output) pattern. The

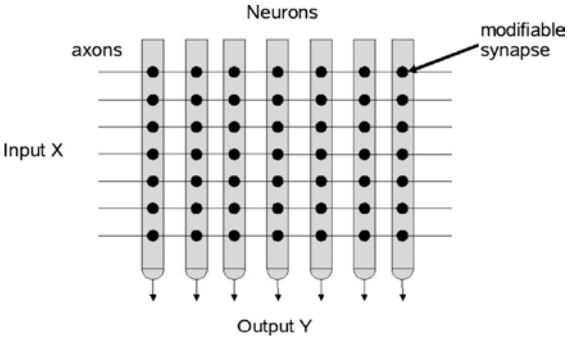


Figure 7.7 Willshaw's model of binary neural associative memory.

patterns need to be sparsely coded binary vectors in which the number of 1s is very low in comparison with the pattern vector length. The pattern activation a^k for the k-th pattern vector is

$$a^k = \sum_j x_j^k \, . \tag{7.4}$$

By using the sparse coding scheme, a larger number of memory patterns can be stored since the patterns do not have too much overlap. The set of pattern pairs $\{(x_k, y_k), k = 1, ..., M\}$ is stored in a binary memory matrix by using the *Hebbian learning rule* [15]:

$$w_{ij} = \bigvee_{k=1}^{M} x_i^k y_j^k \, , \tag{7.5}$$

where M is the number of patterns, x_k is the input pattern, y_k is the output pattern and w_{ij} corresponds to the synaptic weight of the connection from neuron i in the input population to neuron j in the address population. In (7.5), each synaptic weight w_{ij} in the matrix memory is set to 1 if x_i (input) and y_j (output) units are simultaneously active in the presentation of at least one pattern pair (x_k, y_k) [11]. In the case of *autoassociative memory*, the address pattern is assumed to be equal to the corresponding content (output) pattern. Such memory models can be used for *pattern completion*. The idea of pattern completion is that a noisy or incomplete version of a stored pattern should be completed to a pattern, which has been previously stored. The patterns are stored using (7.5), where y_j^k is equal to x_j^k. For retrieval of content patterns, we will use the so-called one-step retrieval strategy with threshold:

$$y_j^t = 1 \Leftarrow (Ax^t)_j = \Theta \, , \tag{7.6}$$

where the threshold Θ is set to a global value and y is the content pattern. A special case of this strategy is the Willshaw's strategy, where the threshold is set to the activity of the input pattern given.

7.4.2.2 The Neural Associative Memory-Based Architecture for Word Recognition

The architecture proposed for the conversion of subword units to words uses the neural associative memories. Given a subword unit sequence from HMMs, its task is to generate a sequence of the best matching words. The architecture consists of one auto- and four heteroassociative memories that are connected via hetero- and autoassociative connections. Figure 7.8 provides an overview of the architecture. Each box in Figure 7.8 denotes a binary associative memory.

The memories used in the architecture are based on subword units, such as context-dependent phonemes or syllables, and the words are processed in the architecture using their subword unit representations. The type of subword unit used in the architecture depends on the size of the vocabulary. The memory usage of the architecture is proportional to the number of subword units.

The memories used in the architecture are given as follows.

HM1. This heteroassociative memory is the input area of the architecture. It stores the subword units columnwise by using binary sparse representations. The memory is a matrix of dimension $L \times n$, where L is the length of the input vector and n is the number of subword units. During retrieval, the memory activates the subword unit received from HMMs and represents it to the architecture.

HM2. This is the same as the memory HM1, but it activates the subword unit that is currently expected by the architecture with respect to the subword unit and word hypothesis recognized in the previous step.

HM3. This heteroassociative memory is a matrix of dimension $n \times n$ and stores the subword units which typically follow each other with respect to the words in the vocabulary. Given an input subword unit, its output is a list of subword units that typically follow that specific input.

These three heteroassociative memories work together as a subword-unit recognizer by summing up their outputs and applying a global threshold. In this way, the input subword unit received from the HMMs may be corrected by using contextual internal information given by the architecture. In the subword-unit recognizer the total weight of the memories HM2 and HM3 equals the weight of HM1.

HM O. This is a memory that represents the output of the subword-unit recognizer consisting of HM1, HM2 and HM3.

DL. Its structure is the same as that of HM1. Its task is to hold a list of subword units that have been processed up to the current step.

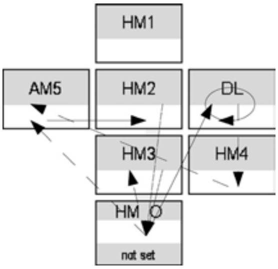


Figure 7.8 Neural associative memory-based architecture for the subword units to words conversion. Each box is an associative memory. The dashed arrows denote auto-associative and solid arrows denote hetero-associative connections.

HM4. This memory is a matrix of dimension $n \times m$, where m is the length of the word code vector. It stores heteroassociations between subsymbolic representations and the corresponding binary sparse word vectors. It is responsible for generating word hypotheses with respect to the subword units activated by DL.

AM5. The autoassociative memory is a matrix of dimension $n \times n$ and stores the words in the vocabulary using their subsymbolic representations. During retrieval it predicts the subword unit expected in the next step with respect to the input word(s) and subword unit(s) in the current step.

7.4.2.3
The Functionality of the Architecture

The architecture can be split into three modules: a subword-unit recognizer consisting of the memories HM1, HM2 and HM3, a word recognizer consisting of the memories DL and HM4 and a next-subword-unit predictor that is the memory AM5.

Given a sequence of subword units to the architecture, then each time, a subword unit from the input sequence is first presented to the memory HM1 in the subword-unit recognizer. The other memories HM2 and HM3 do not receive any input at the beginning of each word. Therefore, they do not activate any output neurons, consistent with the fact that no expectation can be generated in the beginning of the word-recognition process. Afterwards, the subword-unit recognizer generates the output subword unit.

The output of the recognizer is then forwarded to the memory DL in the word recognizer, where it is stored as the recognized subword unit. The subword unit(s) stored in DL is then sent to the memory HM4 which retrieves the words that best match the subword units stored in DL.

The subword units generated by the subword-unit recognizer and the word hypotheses generated by the word recognizer are forwarded together to the next-subword-unit predictor AM5 which predicts the subword unit(s) expected in the next step with respect to the word and subword-unit inputs.

In the same way, the next subword unit in the input sequence is processed in HM1, whereas HM2 activates the subword units from the next-subword-unit predictor and the memory HM3 activates its output subword units through a back-propagated connection from the subword-unit recognizer. The outputs of HM2 and HM3 represent the expectation generated by the architecture and this information may be used to correct erroneous input subword units that HMMs did not correctly recognize.

The iterations for the current word end when a small pause (silence) is detected in the input subword unit sequence. Note that the architecture cannot decide a unique word representation for a given input-subword-unit sequence that the HMMs did not correctly recognize. In this case, a superposition of word hypotheses matching the input sequence is generated by the architecture. After recognition of each single word in the input stream, it is forwarded to the network that is responsible for words to semantics conversion (see Section 7.5).

7.4.2.4 Learning of New Words

The system is able to learn new words without further training the HMMs or changing the structure of the system. The online learning performance of the system strongly depends on the performance of the HMMs that need to be trained with enough speech data and also need to have a comprehensive language model in order to allow for generating a plausible subword-unit representation for novel words. In one application [16], learning a new word is initiated by a sentence of the type “This is cup”, where “cup” is the word that has to be learned. “This is” arouses the system to learn a new object word. During learning of a novel word, the memories HM1 and HM2 in the subword recognizer are not updated, whereas HM3 is updated according to the subword-unit representation of the novel word. To store the new object word in HM4 and AM5, a new binary vector representation is randomly generated and stored in the associative memories.

7.5 Words to Semantics Conversion using Associative Memory

In order to extract the semantics from the stream of words that the system recognized using the methods introduced in the previous sections, a network of neural associative memories is used. Instead of Willshaw’s model of associative memory,

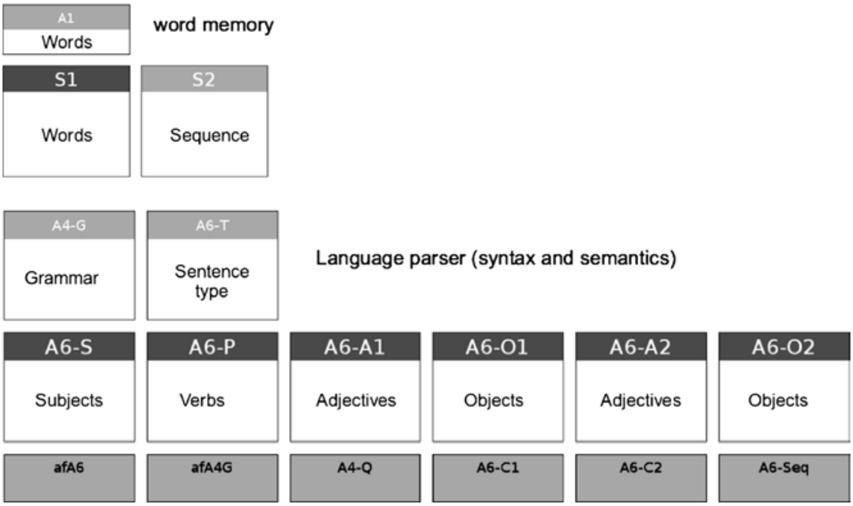


Figure 7.9 Overview of the neural network for semantic processing. Each box corresponds to one autoassociative memory. The memories are connected to each other via heteroassociative connections (not shown, see Figure 7.10). In each box the kind of information that the corresponding memory processes is displayed, for example, A6-S is dealing with subject words.

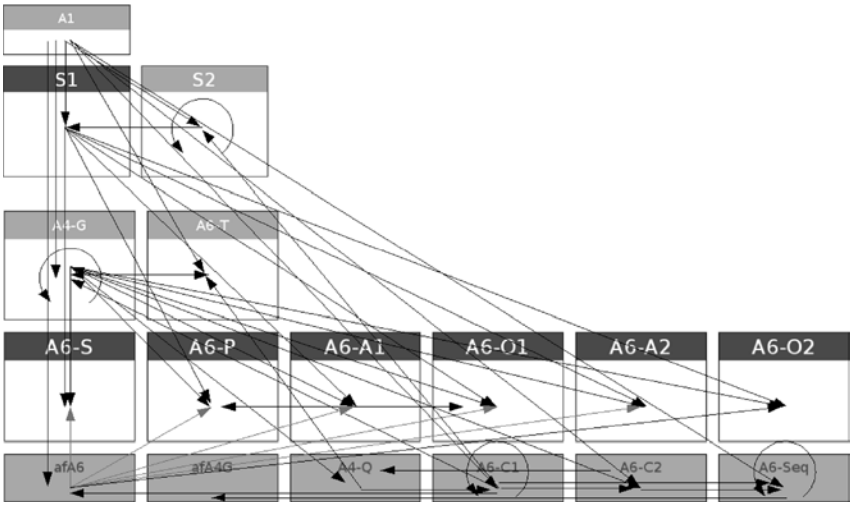


Figure 7.10 Overview of the networks connectivity. Each straight arrow corresponds to a heteroassociative connection between two autoassociative memories. Each circular arrow corresponds to short-term memory mechanism (see text for details).

a slightly version is used, the so-called Spike Counter model. It enhances Willshaw's model with the possibility of activating and dealing with superpositions of patterns, a feature that we use to represent ambiguities. For a detailed explanation of the model underlying the semantic parser, see the appendix in [16].

The language model in its current form is implemented using a total of 17 associative memories that are heavily interconnected with heteroassociative connections. Figure 7.9 gives an overview of the model's architecture and its connectivity. Each box depicts one associative memory, each straight arrow corresponds to a heteroassociative connection and circular arrows denote strong autoassociative feedback from one memory onto itself. Strong autoassociative feedback is generally used to keep patterns active longer than the input is available to the network (short-term memory).

The model shown in Figure 7.9 consists of two main parts; namely, the spoken word memory and the language parsing part. The spoken word memory just keeps track of the words that were input into the system in order to be able to look up later exactly what the input words were. The language parser is used to parse the input words with respect to a given grammar.

7.5.1 Spoken Word Memory

The spoken word memory, consisting of areas A1, S1 and S2, keeps track of the input word sequence. The connectivity of that area is pretty simple, A1 projects down into S1, S2 projects into S1. A1 serves as input area, it represents the output of the word-recognition software. S2 holds a fixed, arbitrary sequence of fixed length

(10 elements. See the appendix of [16] for a description of how sequence memories work) and the projection from S2 to S1 possesses fast hebbian synapses. If a word enters A1, it is associated with the sequence element that is active in S2 at the given moment through the fast synapses between S1 and S2. If the next input word enters A1, S2 is switched to the next sequence element and again an association is stored in the heteroassociative connection between S1 and S2. In the beginning, a dedicated start state is activated in S2. This allows the recall of the input words by activating the corresponding sequence elements in S2 while no input is active in A1. However, recalling the input is possible only for a certain amount of time, because the synapses between S1 and S2 also forget what they learned in several simulation steps.

7.5.2 Language Parser

The language parsing part of the network mainly consists of the areas A4-G, A6-T and A6-XX. There are additional fields drawn in grey in Figure 7.9. They serve several control tasks, that is activating or inhibiting specific areas at certain times to, for example, switch sequences to the next state, initialize the S2 sequence memory in the word memory with its starting state, etc.

Area A4-G holds the grammar information. For each possible sentence type, a sequence of the corresponding word types is stored. There are not many sentence types in the current implementation, we have a total of 10 sentence types (e.g. SPO or SPAOO, standing for “subject predicate object” or “subject predicate adjective object1 object2”, respectively). However, expanding this to more sentence types is not problematic, although the memory requirements obviously increase.

While the input is entered into the word memory, A4-G continuously compares the entering words with respect to their possible grammatical functions with its sequences. To do this, A4-G follows all sequences that match the input in parallel (with a superposition of several active sequences). This is realized by a heteroassociative projection from S1 to A4-G. If a subject word (e.g. bot) is active in S1, it projects onto subject word representations in A4-G. Together with the current state of the grammatical sequence, A4-G can decide which sequences match the input. In the beginning, A4-G is in a starting state that all sequences have in common. If, for example, a subject is now entering S1, A4-G switches into a superposition of all sequences that start with a subject. If now a predicate is entered, A4-G activates only those sequences that start with a subject (these are the only ones that are active) followed by a predicate, and so on. If at the end of the input stream there is only one unique sentence type left, the sentence can be uniquely interpreted. Otherwise, there is either no valid interpretation if there is no activity left in A4-G (the sentence was grammatically incorrect) or a superposition remains active in A4-G (the sentence was ambiguous on a grammar level, e.g. “bot put orange orange orange”, which could either mean “bot put orange orange (to) orange” or “bot put orange (to) orange orange”. Note that the simplified grammar for testing does not use any prepositions).

A6-T stores the recognized sentence type after a sentence has been heard. If A6-T has a unique pattern activated (e.g. SPO), the system detects an SPO-type sentence and starts assigning the words from the input to the correct output boxes.

The A6-XX memories hold the final output of the system. After a sentence has been successfully parsed, A6-S holds its subject, A6-P its predicate, A6-A1 the adjective of the first object, A6-O1 the first object and A6-A2 and A6-O2 hold the second adjective/object pair, if applicable. Note that in our current grammar, only one attribute per object is allowed.

Further processing units can now use the sentence type available in A6-T together with the corresponding words in A6-XX to plan further actions.

7.5.3 Ambiguities

The language model is able to deal with ambiguities on the grammar and the single-word level. Section 7.5.2 gave an example for an ambiguity on the grammar level that cannot be resolved. Consider, however, the sentence “bot put orange orange orange plum”. This sentence is grammatically ambiguous until the last word “plum” is heard, meaning that no unique grammatical interpretation is possible until the last word enters the system. The complete sentence however is of “SPAOAO” type meaning “bot put (the) orange orange (to the) orange plum”. The system solves the ambiguity by keeping all possible states active concurrently in a superposition in the grammar area A4-G.

The same mechanism can be applied on the single-word level. If the word recognition network was not able to decide for a unique interpretation of one word, it forwards a list of alternatives to the language parser. These alternatives are activated concurrently as a superposition in A1 and S1. They project onto all possible word types in A4-G and all possible interpretations are processed in parallel due to the superpositions. If at the end a unique interpretation is remaining, it is activated as final output. If this is not possible, a superposition of all possible interpretations is kept. In that case, additional contextual input from other parts of the system can be used to resolve the ambiguities later.

For example, the sentence “bot show/lift green wall” (with an artificial ambiguity between “show” and “lift” in the input) is interpreted as “bot show green wall” because a wall is not liftable. Similarly, the sentence “bot lift/put green apple” is interpreted as “bot lift green apple” (put needs two object words) and “bot lift/put apple plum” is interpreted as “bot put apple (to) plum” (lift needs only one object). Obviously, the sentence “bot lift/put orange orange” cannot be disambiguated, because “orange” can refer to either an attribute or an object.

7.5.4 Learning of New Objects

It is possible to add new object words to the language model. If a sentence of the type “this is XX” is recognized by the system, where “XX” is a novel word, new object word representations are generated in areas A1, S1, A6-O1 and A6-O2 while the sentence is processed. This happens automatically: As soon as “this is” is understood, a special learn signal is activated in all areas dealing with object words. If the input now does not match well with a previously learned object word, the system considers this as a new input and generates a matching new pattern. Afterwards, corresponding heteroassociative connections are updated. After successful parsing of a “this is XX”-sentence, the new object word can be used as any other object word, for example the system can successfully parse sentences like “bot show XX” or “bot put red XX apple”.

7.6 Sample System/Experimental Results

The proposed speech recognition architecture can be applied to various speech-controlled systems either working in a distributed environment or not. In this section, we present two sample systems. In the first one, speech is captured via a mobile phone and features are extracted and transmitted to a remote server to recognize the spoken command, whereas in the other system, the whole speech recognition process is implemented in one device.

In the first system, the user utters a single command word or word phrase, which is a bus stop name, to a mobile phone and then the speech recognizer on the remote server, correctly recognizing the command, implements the corresponding task that is to send information about the spoken bus stop name from the remote server to the user. Due to the single word commands, the grammatical parser part of the speech recognition architecture is not used here. The system works on a set of 279 German bus stop names, which originated from the Institute for Information Technology, University of Ulm. The training set consists of 14 speakers, whereas the test set contains 5 speakers. The speakers both in training and test sets speak these 279 bus stop names. The number of word tokens in the test set is 1395 while it is 3906 in the training set.

In the system, triphones are used as subword units. Therefore, a triphone recognizer is designed by using HMMs, in which an acoustic model for each triphone is generated and trained on a training set and a triphone-based simple-task grammar is used [17]. The number of triphones is 1284 based on the training and test sets. The dictionary contains an entry for each “triphone” such that the triphone and the pronunciation are the same. The word-recognition architecture is then designed based on triphones and the words are stored using their triphone-level transcriptions.

The designed system has been tested on the test set, yielding a word-recognition accuracy of 98%. Compared with a HMM-based word recognizer, which achieves 99% word recognition accuracy, there is a slight difference between the presented architecture and the pure HMMs. This difference can decrease using larger word parts, such as syllables, instead of triphones and a more efficient language model.

In the second system, the commands are simple english sentences like “bot lift ball” or “bot put orange (to) red plum”. The system vocabulary consists of 43 words. The speech data consists of 105 different sentences, spoken by 4 speakers, 70 sentences of which were used for training and the remaining 35 sentences for testing. The test set is a total of 504 words. Beside our training set, we also used TIMIT training set without SA-type sentences [18]. In the architecture, the HMMs used triphones as subword units and words were stored in the word recognizer using their triphone-level transcriptions. To enable the learning of new words, a triphone-level bi-gram language model is used based on TIMIT training and test sets and our speech data. During the experiment on our test set, the proposed system recognized 98% of the words, whereas the HMM-based word recognizer with 8 Gaussian mixtures, recognized 96%.

For large-vocabulary continuous-speech recognition, a similar system to the presented one with a different language model was tested on TIMIT [18]. In comparison to an HMM-based triphone recognizer [19] achieving 91.9 ± 0.6% word accuracy, this system obtained a promising result of 92.97% word accuracy.

7.7 Conclusion

In this chapter we have presented a distributed architecture for speech-controlled systems based on HMMs and NAMs. The system works in a hierarchical manner such that the speech vectors are first mapped to subword units like context-dependent phonemes or syllables; at the next stage this subword unit sequence is converted to a word sequence and the word sequence is finally parsed to obtain the semantics.

Depending on the complexity of the task to be performed, the system can be trained by using the subword units that are most convenient. For small vocabularies, context-dependent phone-like units such as diphones or triphones can be appropriate, whereas syllables can be used for large vocabularies. It is also important to effectively choose the type of subword unit, due to the fact that it has a great effect on memory usage. The computational complexity of the system is heavily dependent on the size of vocabulary and the type of subword units.

We have also addressed the problems of handling ambiguities on the word level in case the user does not clearly speak a command, and also of learning new command words during performance.

The proposed system can be readily used with a mobile device such as a mobile phone. In this case, the feature extraction is done on the mobile device in the same way as in ASR systems. Then, the feature vectors are compressed and transfered to a remote server to recognize spoken commands.

The system is also well-suited for adaptation to different tasks due to its architecture where the associative memory-based modules are used on word and sentence levels. The performance of the system is comparable with other systems such as conventional HMM-based speech recognizers, but it has the advantages that preprocessing can be done on mobile devices, the task dictionary can be enlarged by learning during performance, and ambiguities can be transfered between modules to be resolved at the most appropriate level.

References

1 Bridle, J.-S. (**1990**) Alphanets: a Recurrent Neural Network Architecture with a Hidden Markov Model Interpretation. *Speech Communication*, **9**(1), 1167.

2 Bourlard, H. and Morgan, N. (**1994**) *Connectionist Speech Recognition. A Hybrid Approach*, Kluwer Academic Publishers.

3 Bengio, Y. (**1993**) A Connectionist Approach to Speech Recognition. *International J. Pattern Recognition Artificial Intelligence*, **7**(4), 647.

4 Rabiner, L. and Juang, B.-H. (**1993**) *Fundamentals of speech recognition*, Prentice-Hall, Inc., Upper Saddle River.

5 Tan, Z.-H., Dalsgaard, P. and Lindberg, B. (**2005**) Automatic Speech Recognition over Error-Prone Wireless Networks, *Speech Communication*, Elsevier, **47**(1–2), 220.

6 ETSI Standard ES 201 108, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm*, **September 2003**.

7 ETSI Standard ES 202 050, *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithm*, **October 2002**.

8 ETSI Standard ES 202 211, *Distributed Speech Recognition; Extended Front-end Feature Extraction Algorithm; Compression Algorithm, Back-end Speech Reconstruction Algorithm*, **November 2003**.

9 ETSI Standard ES 202 212, *Distributed Speech Recognition; Extended Advanced Front-end Feature Extraction Algorithm; Compression Algorithm, Back-end Speech Reconstruction Algorithm*, **November 2003**.

10 Willshaw, D., Buneman, O. and Longuet-Higgins, H. (**1969**) Non-holographic associative memory. *Nature* **222**, 960.

11 Buckingham, J. and Willshaw, D. (**1992**) Performance characteristics of the associative net. *Network* **3**, 407.

12 Palm, G. (**1982**) *Neural Assemblies*, Springer-Verlag, Berlin.

13 Schwenker, F., Sommer, F.-T. and Palm, G. (**1996**) Iterative retrieval of sparsely coded associative memory patterns. *Neural Networks*, **9**(3), 445.

14 Palm, G. (**1980**) On associative memory. *Biological Cybernetics* **36**, 19.

15 Hebb, D.-O. (**1949**) *The Organization of Behaviour*, John Wiley, Newyork.

16 Markert, H., Knoblauch, A. and Palm, G. (**2006**) Modelling of syntactical processing in the cortex, **BioSystems 89**, 300.

17 YOUNG, S., *et al.* (**2002**) *The HTK Book for HTK Version 3.2.1*, Cambrige University, Engineering Department.

18 TIMIT Acoustic-Phonetic Continuous Speech Corpus, National Institute of Standartsand Technology Speech Dics 1-1.1, NTIS Order No. PB91-505065, **1990**.

19 HÄMÄLÄINEN, A., DE VETH, J. AND BOVES, L. (**2005**) *Longer-Length Acoustic Units for Continuous Speech Recognition*, in *Proceedings EUSIPCO*, Turkey.

8
Machine Learning for Categorization of Speech Utterances

Amparo Albalate[1], David Suendermann, Roberto Pieraccini, Wolfgang Minker

8.1
Introduction

As a result of accelerated technological development and, particularly, due to the progressive advances in the field of automated speech recognition, first Spoken Language Dialog Systems (SLDSs) emerged in the mid 1990s as a new, important form of human-machine communication.

As their name suggests, SLDSs are interactive, voice-based interfaces between humans and computers, which allow humans to carry out tasks of diverse complexity (travel ticket reservations, bank transactions, information search or problem solving, etc.).

The typical architecture of an SLDS [1] is depicted in Figure 8.1. Input acoustic vectors generated from the speech signal are first processed by an Automatic Speech Recognizer (ASR), resulting in a raw text transcription[2] of the input utterance. Subsequently, the transcribed text is interpreted in a semantic analysis block which extracts the utterance meaning in the form of an appropriate semantic structure. This semantic representation is processed by the dialog manager which also communicates directly with an external application, namely a database interface. The dialog manager keeps control of the overall interaction progress towards task completion. During this process, the user may be queried for confirmations, disambiguations, necessary additional information, etc. Finally, the interaction result is presented to the user in the form of speech (text-to-speech synthesis or prerecorded prompts), text, tables or graphics.

Among the SLDS modules, speech recognition and semantic analysis play a decisive role for global system performance [2]. In particular, this chapter deals with the semantic analysis block, often referred to as *natural language understanding*. The extracted *semantics* from each user utterance can be viewed as an internal knowledge representation used (by the dialog manager) to trigger a certain action in the context of a particular task [3].

[1] Corresponding author.

[2] Most probable sequence of words detected.

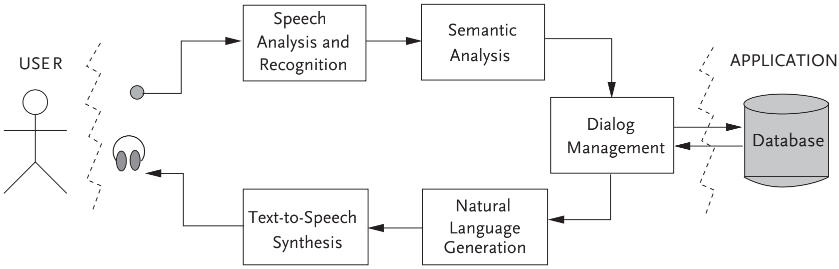


Figure 8.1 Overview of an SLDS.

In first- and second-generation SLDS, frequently used in applications such as banking and travel reservations, semantic analysis commonly relies on the definition of semantic or case-frame *grammars* [4]. A semantic gramar formalism provides a model for the sentence structure in terms of semantic constituents: words or phrases. The semantic analysis *decodes* the text of an input utterance by extracting the correspondences between the sentence constituents and their semantic labels. For example, in the framework of a flight booking application, the user utterance "I would like to fly from Munich to New York on July, 24th" may be decoded into the following semantic sequence: <book>(airport-origin)(airport-destination)(depart-day)(depart-month).

For the grammar implementation, two major tendencies exist: in a rule-based approach, a set of grammar rules is manually defined for a specific task or application. Rule-based methods provide the best performance for a restricted task for which they are originally designed. However, these methods turn out to be inflexible regarding their adaptation and portability to new application domains. Alternatively, in *data-oriented* approaches, stochastic models are used, such as Hidden Markov Models [5], which automatically infer the model parameters from training corpora of semantic representations. These techniques are more flexible and portable to different domains. Examples of systems using rule-based and stocastically-based parsing principles are the ATR translation system from Japanese to English (SL-TRANS) [6] and the AT&T-CHRONUS (Conceptual Hidden Representation of Natural Unconstrained Speech) speech understanding system [7], respectively.

However, third-generation SLDSs, deployed in applications dealing with problem solving, education and entertainment, have shown higher levels of complexity. In this chapter, we focus on the problem-solving domain, in particular on automated troubleshooting agents. These agents are specifically designed to perform customer care issues over the telephone in a similar way to human agents.

Today, natural language understanding is typically performed by a speech recognition module followed by a speech utterance classifier. Such classifiers are a sophisticated replacement of menu-based systems using dual-tone multi-frequency (DTMF) [8] technology (*...push 1 for billing, push 2 for sales ...*) or speech-recognition-based directed dialog (*...you can say billing, sales, or...*). These simple solutions are often impractical for several reasons.

- In certain applications, the number of classes can be too large to be handled in a single menu. Even a succession of menus hierarchically structured would prove unwieldy with hundreds of classes, not to mention the bad-caller experience when five or six menu levels are required to reach the correct routing point.
- Even when prompted with a clear menu, callers often describe the reason why they are calling in their own words, and that may not be covered by the rule-based grammar typically used with directed dialog systems.
- For complex domains, callers may not understand or be familiar with the terms used in the menu. For example in response to the prompt: *Do you have a hardware, software, or configuration problem?*, they may respond unexpectedly (*My CD-ROM does not work!*) or choose one of the options at random without really knowing if it applies to their case.

For these reasons, state-of-the-art troubleshooting agents [9] leave the dialog initiative to the users by presenting an open welcome message: *"please briefly describe the reason for your call"*. Unconstrained, natural language user responses describing the general problem or symptom they experience are then classified by a speech utterance classifier mapping the user utterance into one of a set of predefined categories [10].

Supervised statistical classifiers are algorithms trained with a corpus of transcribed utterances and their associated problem categories. The parameters learned in the training phase are applied to predict the classes of new utterances, not necessarily observed in the training corpus. A crucial factor on which a classifier's effectiveness depends is the size of available data for training.

However, the significant cost of hand-labeling a large amount of training data is one of the main problems associated with the use of such classifiers. Achieving appropriate classification performance even with small training sets [11] recently became the focus of research in the field. Also, the set of categories used for data labeling is subject to alteration. It is not rare to observe situations in which the set of problems handled by the automated agents needs to be updated. In such cases, algorithms which require only a few training data can be helpful to rapidly adapt the system.

In this chapter, we first provide an overview on the utterance categorization model and propose different schemes which use only one labeled example per category. With these minimal training data, considerable degradation of the categorization performance is expected with respect to categorizers that make use of large labeled corpora. One main reason is that semantic variability may not be adequately captured in small labeled sets. We therefore analyze word clustering as a means of extracting semantic relationships of words and, in consequence, boost the classification effectiveness. A similar task in the field of information retrieval is the efficient search of information on the Internet. In fact, one of the first applications of word clustering was the lexical term expansion of user queries to search engines with automatically discovered synonyms of the original query terms [12].

We also provide a comparison of formulations used in text-processing applications for estimating the different relevance of terms. Term scoring was applied to the categorization of utterances with different numbers of labeled examples.

The chapter is organized is as follows. An overview of general pattern recognition and its application to the categorization of texts is given in Sections 8.2 and 8.3. In Section 8.4 a description of the utterance corpora used in our experiments is provided. The utterance preprocessing is explained in Section 8.5. Details about feature extraction and term weighting are outlined in Sections 8.6 and 8.7 respectively. Finally, we evaluate the described algorithms in Section 8.8 and draw conclusions in Section 8.9.

8.2 An Overview of Pattern Recognition

Pattern recognition is an important problem addressed by scientists in a number of research fields: biology, geography, engineering, computer science, artificial intelligence, etc. [13]. In pattern recognition, patterns are defined as entities which can be subjected to classification. This is possible as long as their similarity can be calculated. Examples of patterns are genes, human faces, handwritten characters, or texts.

The classification task consists of (i) the mapping of patterns into one or more classes out of a pre-defined category set (supervised classification or discriminant analysis), or (ii) the grouping of patterns into previously unknown classes according to their affinities (unsupervised classification or clustering). In the latter case, the classes are also detected as a result of the classification process. A typical pattern recognition scheme is shown in Figure 8.2.

In supervised classifiers, pattern recognition operates in two separated modes: *training* and *classification* or *test*. In the case of unsupervised classification, the learning step is absent. In addition, one distinguishes between three phases: preprocessing, feature selection/extraction and classification.

Preprocessing, also known as preparation, aims at optimizing the representation and quality of the input observations in order to produce reliable data for statistical analysis [14]. This process involves operations such as segmentation, normalization and elimination of noise or irrelevant information.

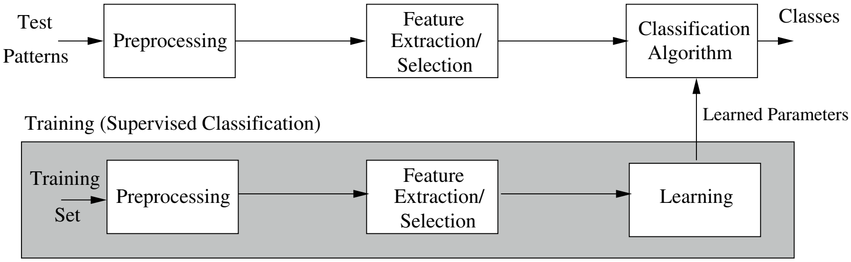


Figure 8.2 Pattern-recognition scheme. For supervised classification, a training module is required.

A segmentation stage decomposes the input data into pieces, thereby enabling the multi-dimensional representation of patterns. In certain cases, input objects are already presented segmented as a set of measurements captured by an array of sensors (for example, temperature and humidity in the classification of metereological phenomena). However, in many other situations, the objects to classify are the result of individual acquisitions. This is the case of images in computer vision. A data segmentation (sampling) may be used here to split continuous images into M pixels or blocks, so that an image can be observed, for example, as an M-dimensional array of pixel intensities. The output elements obtained after segmentation are also termed *classification features,* since they represent different properties of the objects to be classified. In consequence, patterns are also referred to as *feature vectors.*

Normalization procedures can be applied to features or patterns. Feature normalization is especially convenient and necessary if the classification features represent different object attributes in different scales. Common feature normalization techniques are linear scaling to unit variance, transformation to uniform [0-1] random variables and rank normalization, among other methods [15]. Moreover, pattern normalization applies to the feature values inside an individual pattern. An example is the normalization of image intensities for object recognition in images.

Feature selection and extraction techniques help in reducing the dimensionality of feature sets. As is broadly accepted, an optimal feature set should capture the relevant characteristics of the data in the most compact way.

Feature *selection* aims to retain the subset of the original features that best represents the input patterns. Typically, this process is carried out by sorting the initial features according to their relevance and filtering out those features which do not exceed a minimum relevance threshold. The resulting feature vectors are thus the projections of the original patterns over the selected feature sub-space. In contrast, feature *extraction* performs a transformation of the input pattern vectors into a different feature space through a statistical analysis of the input data. Examples of feature extraction techniques are principal components analysis (PCA), independent component analysis (ICA) or feature clustering [24, 32]. The feature selection/extraction module has proved to be very important for pattern recognition. A correct scheme may not only help in reducing computational costs associated with very high-dimensional data sets, but can also increase the classification effectiveness.

Finally, the *classification algorithm* maps input feature vectors to output classes. Supervised algorithms rely on the existence of training sets with labeled examples. The mapping is typically defined by a certain number of parameters whose values are usually adjusted to a training data set during learning. Some examples of supervised classifiers include, among others, the Naïve Bayes classifier, polynomial classifiers, neural networks or support vector machines.

Unsupervised techniques are suitable when no labeled examples are available. The output classes/groups are not known *a priori*, but are detected during the classification process. Hierarchical and partitioning clustering algorithms are used to group the input patterns according to distance-based criteria. Hierarchical approaches build the cluster solution gradually, resulting in cluster hierarchy structures or *dendograms*. Two kinds of hierarchical algorithms can be distinquished, depending on the dendogram construction methods: agglomerative (bottom up) and divisive (top down) [16–18]. In contrast, partitioning approaches learn a flat cluster structure, typically through an iterative search for the optimum of the criterion function (K-means, K-medoids, etc.) [19, 20]. More recent approaches have been developed to discover dense regions in a dataspace. Usually, the density notion is represented by two parameters, *minpts* and *epsilon*, denoting the minimum number of points to be enclosed in an *epsilon*-radius neighborhood of certain objects called core points (DBscan [21], Optics [22], Denclue [23], Clique [24], etc.). These algorithms are resistant to outliers[3] and more flexible than distance-based approaches, insofar clusters of irregular shapes and sizes can be detected [25]. Further, if the set of observations can be drawn from an underlying probabilistic distribution, model-based approaches can be applied in order to fit a probabilistic model to the input patterns. A common example is the Expectation Maximization algorithm, used to fit a mixture of Gaussians to a dataset [26].

A compromise between supervised and unsupervised techniques are semi-supervised approaches [27]. These methods make use of both labeled and unlabeled data for training. In *co-training* algorithms [11], two or more supervised classifiers are applied to different subsets of the original feature set. A new training data set is generated following a confidence evaluation of the classification results (e.g. agreement between classifiers). The main condition for the use of co-training approaches is the statistical independence between feature subsets used by the classifiers. Another kind of semi-supervised learning involves clustering algorithms in which certain constraints about the input data are manually defined (*Clustering with constraints*) [28]. The constraints specify whether two data instances must or cannot be linked together in a single cluster.

8.3 Utterance Classification as a Text-Classification Problem

Since speech utterances are transcribed into text by ASR, utterance-to-symptom categorization is a particular case of text classification, traditionally applied to documents. In this section, we describe how pattern recognition is applied to text and, in particular, to utterance classification.

During preprocessing, all words in a text corpus are reduced to units of semantic meaning: *stems* or *lemmas*. As a next step, an *n*-gram model[4] can be applied to extract and count subsequences of terms up to length *n*. A particular case is

[3] Outliers are *noise* patterns which do not belong to any cluster, but fall in the regions between two or more clusters. Outliers are often unreliable patterns which need to be discovered and accordingly treated.

[4] An *n*-gram model is a sentence structure specification based on the assumption that the probability of occurrence of a given word is conditioned upon the prior $N - 1$ terms. While the *n*-gram specification is of high relevance for the development of grammars and lexical parsers, it is less important for capturing the underlying semantics (meaning) of texts.

the uni-gram model where only single words are extracted, ignoring any possible order in which the words appear in the text. Due to their simplicity and adequate performance for classification, uni-grams are possibly the most frequently used approach for the representation of texts. When used for the representation of texts or utterances, uni-gram structures are commonly referred to as *bags of words*. Usually, texts are represented as vectors over a basis of terms or *n*-grams in what is called a *vector space model* [29]. A simplistic approach is to use binary vector components denoting the presence or absence of the respective terms in a text. Also, other vector components may be used to reflect term frequency counts in the text, or terms' discriminative significance estimated through relevance scores. A popular metric for estimating a word's relevance is the *term frequency – inverse document frequency (TF-IDF)* (Refer to Section 8.7.2 for more details about term scores).

Common feature-selection algorithms are based on the aforementioned relevance scores in order to filter out unimportant terms that do not exceed a relevance treshold. In contrast, feature-extraction approaches provide a transformation of the initial term features into a new feature space in which semantic effects related to terms can be mitigated, namely synonymy and polysemy. Synonymy refers to the fact that multiple terms can be used to denote a single concept – words with identical meaning.[5] Polysemy, on the other hand, indicates the existence of terms with multiple related meanings, which can therefore be observed in different contexts. These semantic artifacts are pointed out as one of the fundamental problems to be faced in text categorization, as they introduce a clear obstacle for capturing the semantic proximity between texts [30]. Attempts to address synonymy and/or polysemy have relied on Latent Semantic Analyis (LSA) [30, 31] and feature clustering [32], among other methods.

8.4 Utterance Corpus Description

For the experiments and results reported in the following sections, we used two corpora of transcribed and annotated caller utterances gathered from user interactions of commercial troubleshooting agents of the Internet and Cable TV domains. Some examples of transcribed utterances are:

- Internet troubleshooting:
  - *The Internet was supposed to be scheduled at my home today.*
  - *I'm having Internet problems.*
- Cable TV troubleshooting:
  - *I have a bad picture quality.*
  - *I don't get HBO channel.* (ChannelMissing).

[5] In text-processing applications, the synonymy concept is used in a general sense, to indicate not only terms with identical meaning but also terms with *similar* meaning (soft synonyms).

Table 8.1 Corpus definition. Number of categories (L) and number of utterances of test and training sets

Corpus	Number of Symtoms	Training (# utt.)	Test (# utt.)
Internet	28	3313	31 535
Cable TV	79	10 000	10 000

Further details about the corpora including the number of categories considered in this work as well as the dimensions of training[6] and test sets are shown in Table 8.1.

8.5 Utterance Preprocessing

The preprocessing module consists of part-of-speech (POS) tagging, morphological analysis, stop-word filtering, and bag-of-words representation.

First, the Stanford POS tagger [33] performs an analysis of each sentence and tags the words with their lexical categories (POS tags).

Subsequently, a morphological analyzer [34] is applied to reduce the surface word forms in utterances into their corresponding lemmas.

As a next step, stop words are eliminated from the lemmas, as they are judged irrelevant for the categorization. Examples are the lemmas *a, the, be, for*. In this work, we used the SMART stop-word list [35] with small modifications: in particular, we deleted confirmation terms (*yes* and *no*) from the list, whereas typical words for spontaneous speech (*eh, ehm, uh*) were treated as stop words.

For example, the input utterance *My remote control is not turning on the television* is transformed through the described steps (POS tagging, morphological analyzing and stop-word filtering) as follows:

- POS tagging: my/PRP remote/JJ control/NN is/VBZ not/RB turning/VBG the/DT television/NN[7]
- Morphological analysis: My remote control be not turn the television
- Stop-word filtering: remote control not turn television

The salient vocabulary is then defined as the set of distinct lemmas in the preprocessed training utterances: $W = (w_1, \dots, w_D)$. The vocabulary dimensions in Internet and Cable TV troubleshooting corpora are $D = 1614$ and $D = 1022$, respectively.

Finally, the lemmas for each utterance are combined as a bag of words, i.e. each utterance is represented by an D-dimensional vector, BW, whose binary elements,

[6] Note that, since the approaches described in this chapter make reference to small numbers of examples, we refer to the part of the available corpora used to select the categorizers examples as a training set and, if necessary, perform certain statistical analyzes which do not require the use of utterance labels.

[7] For a detailed inventory of POS tags used by the Stanford parser and their meanings, please refer to the parser homepage: *http://nlp.stanford.edu/software/lex-parser.shtml*

be, represent the presence/absence of the respective vocabulary element in the current utterance:

$$BW = (b_1, \ldots, b_D) . \tag{8.1}$$

8.6 Feature Extraction Based on Term Clustering

One of the simplest categorization algorithms is the nearest-neighbor (NN) approach. Given a set of M labeled examples per category (prototypes), the NN algorithm assigns each input pattern to the category of the closest prototype. In this work, we only use one prototype per category ($M = 1$), selected from the training corpus. One should therefore expect a degradation of the classifier performance with respect to categorizers making use of all utterance labels in the training set. This is partly due to the prevalence of synonymy and polysemy, which may be insufficiently represented in a small amount of prototypes. Also what is considered to belong to a class can be arbitrary and is up to the system design and to what the classification result is used for further down in the application.[8]

In effect, by using one labeled utterance per category, the effective vocabulary available to the categorizer is reduced to less than 10% of the vocabulary in the training set (W). This results in a large amount of utterances mapped to a *nomatch* class, provided the existence of out-of-vocabulary terms. As an example, we want to look at the category representing a problem related to sound (NoSound). One would select a typical caller utterance reporting this problem, *no sound*, as the category prototype. However, the user may utter other alternatives, such as *problem with volume* or *lost audio*, which cannot be matched to the desired prototype due to the bag-of-words' orthogonalities (absence of overlapping terms with the prototype). This problem could be partially solved if one could detect that *sound* has a similar meaning to *audio* or *volume*.

The feature-extraction methods described in the following paragraphs aim to capture semantic relationships between words. We analyze two approaches to the classification of words based on hard and fuzzy clustering.

In hard clustering, each input pattern is unequivocally allocated to one output cluster. This approach may be adequate for capturing semantically related terms (e.g. synonyms) in output *semantic* classes. In contrast, a soft-clustering algorithm associates the input patterns to all output classes through a matrix with membership degrees. If a considerable number of polysemous terms (with several related meanings) is present in the input data, fuzzy techniques should then be more

[8] The corpora used in this study contain a class for multiple symptoms (like *my picture is out, and I have no sound*) which is purposely omitted when the classifier is trained to catch such an utterance with one of the single-symptom classes (such as *NoPicture* and *NoSound* in the above example). It is extremely unlikely that such a class would be automatically isolated as it potentially contains contributions from all the other classes.

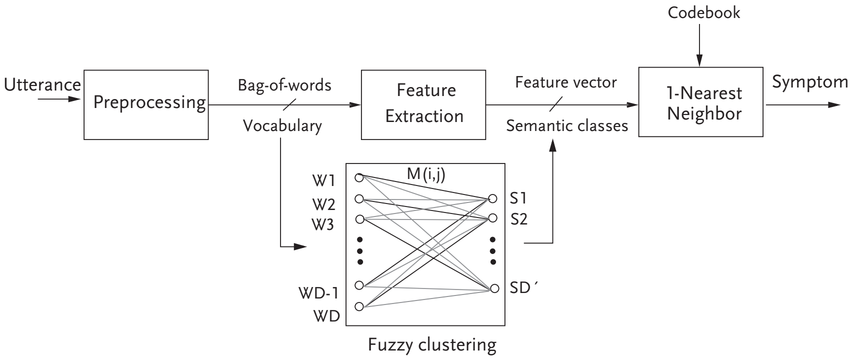


Figure 8.3 Utterance categorization components. For feature extraction, hard and fuzzy approaches to term clustering are compared. Hard clustering provides a hard mapping of each vocabulary term pattern into a single-output semantic class (bold traces). In contrast, a fuzzy clustering provides a *fuzzy* or *soft* association of each pattern to the output classes through a membership matrix (thin lines). Hard clustering can also be observed as a particular case of fuzzy clustering, where pattern memberships are either 1 or 0.

appropriate. An overview on utterance categorization based on term-clustering is shown in Figure 8.3.

After the feature extraction phase, each input bag of words (BW) is accordingly transformed into a feature vector F. Details of feature extraction based on hard and fuzzy clustering are discussed in the following sub-sections.

8.6.1
Term Vector of Lexical Co-occurrences

A frequently reported problem to word clustering is the adequate representation of word lemmas in vector structures so that mathematical (dis)similarity metrics applied to term vectors can reflect the terms' semantic relationships.

In this respect, among others, there are two criteria in the literature which attempt to explain the main characteristics of semantically related terms.

1. *First order co-occurrence.* Two words are similar to the degree that they co-occur or co-absent in texts [12, 36].
2. *Second order co-occurrence.* Two words are similar to the degree that they co-occur with similar words [37].

The first order co-occurrence criterion is adequate for text documents where a semantic variability can be observed inside a document. In contrast, semantically related terms rarely co-occur inside a sentence. Thus, a second-order co-occurrence criterion seems to be more appropriate for detecting terms' semantic proximities from an utterance corpus.

Consequently, each vocabulary term w_i is represented in a D-dimensional vector of lexical co-occurrences:

$$W_i = (c_{i1}, \dots, c_{iD}) \tag{8.2}$$

wherein the constituents c_{ij} denote the co-occurrence of the terms w_i and w_j, normalized with respect to the total sum of lexical co-occurrences for the term w_i:

$$c_{ij} = \frac{nc_{ij}}{\sum\limits_{k \neq i} nc_{ik}} \; . \tag{8.3}$$

Here, nc_{ij} denotes the total number of times that w_i and w_j co-occur. Finally, in order to extract the terms' semantic dissimilarities, we apply the Euclidean distance between term vectors.

8.6.2 Hard Term Clustering

A hard clustering algorithm places each input pattern into a single output cluster. Based on the complete-link criterion [17], the proposed term clustering produces a partition of the vocabulary terms given an input user parameter, the maximum intra-cluster distance d_{th}.

1. Construct a dissimilarity matrix U between all pairs of patterns. Initially, each pattern composes its individual cluster $c_k = \{w_k\}$.
2. Find the patterns w_i and w_j with minimum distance $U_{\min}$ in the dissimilarity matrix.
   – If the patterns which are found belong to different clusters, $c_a \neq c_b$, and $(U_{\max}(c_a, c_b)) \leq d_{th}$, where $U_{\max}(c_a, c_b))$ is the distance of the furthest elements in c_a and c_j, merge clusters c_a and c_b.

Table 8.2 Example utterances of semantic classes obtained by hard term clustering on a text corpus comprising 30 000 running words from the cable televison troubleshooting domain; average number of terms per cluster is 4.71; number of extracted features is 1458

speak, talk
operator, human, tech, technical, customer, representative, agent, somebody, someone, person, support, service
firewall, antivirus, protection, virus, security, suite, program, software, cd, driver
reschedule, confirm, cancel, schedule *remember, forget*
webpage, site, website, page, web, message, error, server
megabyte, meg
technician, appointment
update, load, download
boot, shut, turn
user, name, login, usb
area, room, day

  - Update U so that $U_{ij} = \infty$.
3. Repeat step 2) while $U_{\min} \leq d_{th}$ or until all patterns remain assigned to a single cluster.

As a result of the hard term clustering algorithm, different partitions of the vocabulary terms are obtained, depending on the input parameter d_{th}. Because the elements in each cluster should indicate terms with a certain semantic affinity, we also denote the obtained clusters as *semantic classes*. Table 8.2 shows examples of clusters produced by this algorithm.

After hard term clustering, the bag of words remains represented in a binary feature vector F_{hard}:

$$F_{\text{hard}} = (b_{f_1}, b_{f_2}, \ldots, b_{f_{D'}}) \tag{8.4}$$

where the b_{f_i} component denotes the existence of at least one member of the i-th extracted class in the original bag of words.

8.6.2.1 Disambiguation

If applied to bags of words or feature vectors extracted from hard term clusters, the NN classifier rejects a considerable number of ambiguous utterances for which several candidate prototypes are found.[9] A disambiguation module was therefore devised to resolve the mentioned ambiguities and map an ambiguous utterances to one of the output categories.

First, utterance vectors with more than one candidate prototype are extracted. For each pattern, we have a list of pointers to all prototypes. Then the terms in each pattern that cause the ambiguity are identified and stored in a competing term list.

As an example, let us consider the utterance *I want to get the virus off my computer* which, after pre-processing and hard term clustering, results in the feature set *computer get off virus*. Its feature vector has maximum similarity to the prototypes *computer freeze* (category *CrashFrozenComputer*) and *install protection virus* (category *Security*). The competing terms that produce the ambiguity are in this case the words *computer* and *virus*. Therefore, the disambiguation among prototypes (or clusters) is here equivalent to a disambiguation among competing terms. For that reason, as a further means of disambiguation, we estimate the *informativeness* of a term w_i as shown in (8.5):

$$I(w_i) = -\left(\log(Pr(w_i)) + \alpha \cdot \log\left(\sum_{\substack{j \\ L_j = N}} c_{ij} Pr(w_j)\right)\right) \tag{8.5}$$

[9] Candidate prototypes are those prototypes which share maximum proximity to the input utterance. This happens especially when the similarity metric between the vectors results in integer values, e.g. in the case of using the inner product of binary vectors as are the aforeintroduced bags of words and feature vectors.

where $Pr(w_i)$ denotes the maximum-likelihood estimation for the probability of the term w_i in the training corpus, and L_j refers to the part-of-speech (POS) tag of w_j.

As can be inferred from (8.5), two main factors are taken into account in order to estimate the relevance of a word for the disambiguation:

a) the word probability and
b) the terms' co-occurrence with frequent nouns in the corpus.

The underlying assumption that justifies this second factor is that words representative of problem categories are mostly nouns and appear in the corpus with moderate frequency. The parameter α is to control the trade-off between the two factors. Reasonable values are in the range of ($\alpha \in [1, 2]$) placing emphasis on the co-occurance term; for our corpus, we use $\alpha = 1.6$ which we found as best-performing in the current scenario.

Finally, the term with the highest informativeness is selected among the competitors, and the ambiguous utterance vector is matched to the corresponding prototype or class.

8.6.3
Fuzzy Term Clustering

The objective of the fuzzy word clustering used for feature extraction is a fuzzy mapping of words into semantic classes and leads to the membership matrix M representing this association.

8.6.4
Pole-Based Overlapping Clustering

In the PoBOC algorithm [38], two kinds of patterns are differentiated: poles and residuals.

Poles are homogeneous clusters located as distant as possible from each other. In contrast, residuals are outlier patterns that fall into regions between two or more poles. The elements in the poles represent monosemous terms, whereas the residual patterns can be seen as terms with multiple related meanings (polysemous).

The PoBOC algorithm is performed in two phases: (i) pole construction, and (ii) multi-affectation of outliers.

In the *pole construction* stage, the set of poles $\{P\} = \{P_1, \cdots, P_{D'}\}$ and outliers $\{R\}$ are identified and separated. Poles arise from certain terms with maximal separation inside a dissimilarity graph which are therefore known as the pole generators.

In the *multi-affectation* stage, the outliers' memberships to each pole in $\{P\}$ are computed. Finally, the term w_i is assigned a membership vector to each P_j pole as follows:

$$M_{ij} = \begin{cases} 1, & \text{if } w_i \in P_j \\ 1 - d_{\text{av}}(W_i, P_j)/d_{\max} & \text{if } w_i \in \{R\} \\ 0, & \text{otherwise} \end{cases} \tag{8.6}$$

where $d_{\mathrm{av}}(W_i, P_j)$ denotes the average distance of the w_i word to all objects in P_j, and $d_{\max}$ refers to the maximum of the term dissimilarity matrix.

To compute the semantic dissimilarity of terms, experiments with both Euclidean and cosine distances[10] were carried out.

8.6.4.1 PoBOC with Fuzzy *C*-medoids

The fuzzy *C*-medoids algorithm (FCMdd) [39] computes the fuzzy membership matrix M starting from an initial choice of cluster representatives or *medoids*. We initialize the algorithm with the D' pole generators ($C = D'$) obtained at the pole construction phase of the PoBOC scheme. The final solution for the membership matrix M is then reached through the iterative repetition of two steps: (i) (re)calculation of pattern memberships to the D' classes, and (ii) recomputation of the cluster medoids. The membership update of the term W_i to the j-th class is defined as:

$$M_{ij} = \frac{\left(1/d(W_i, C_j)\right)^{1/m-1}}{\sum_{k=1}^{C} \left(1/d(W_i, C_k)\right)^{1/m-1}} \tag{8.7}$$

denoting C_k, the k-th class medoid, $d(W_i, C_k)$, the dissimilarity between the term vector W_i and the medoid C_k, and m, a fuzzyfier factor, $m \in [1, \infty)$, denoting the smoothness of the clustering solution ($m = 2$ in this work). The procedure is iterated until either the updated cluster medoids remain the same, or a maximum number of iterations is reached.

8.6.4.2 Utterance Feature Vector

Finally, the feature vector obtained with soft term clustering, F_{soft}, is calculated as the normalized matrix product between the original bag of words BW and the membership matrix M:

$$F_{\mathrm{soft}} = \frac{BW_{(1xD)} \cdot M_{(DxD')}}{|BW \cdot M|} \tag{8.8}$$

8.6.5 Utterance Categorization

The objective of utterance categorization is to map an input utterance – represented as bag of words (BW) or feature vector after hard or soft word clustering – into one of the N categories, represented by the N prototypes supplied to the nearest-neighbor algorithm. The closeness of an input utterance vector to each one of the

[10] The cosine distance metric, $D_{\cos}$ is defined as the negative of the cosine score, $D_{\cos} = 1 - S_{\cos}$.

prototypes is quantified by means of the inner product between their feature vectors, F_i and F_j:

$$s(F_i, F_j) = F_i \cdot F_j^T \, . \tag{8.9}$$

8.7 Supervised Methods for Utterance Categorization

In this section, we describe two supervised approaches for utterance categorization: a probabilistic framework (Naïve Bayes classifier) and a vector model with term weighting. F is the number of labeled exampes per category, randomly selected.

8.7.1 Naïve Bayes Classifier

Naïve Bayes is a powerful and yet simple text categorization algorithm usually reporting adequate performance. It selects the most probable class $\hat{c}$ out of a set C given a test utterance u:

$$\hat{c} = \arg\max_{c \in C} (P(c|u)) \tag{8.10}$$

This expression cannot be computed directly, but it can be reformulated using the Bayes rule as:

$$\hat{c} = \arg\max_{c \in C} (P(c)(u|c)) \, . \tag{8.11}$$

By assuming conditional independence of the utterance terms, the Naïve Bayes solution can be expressed as:

$$\hat{c} = \arg\max_{c \in C} \left(P(c) \prod_{w_i \in u} (w_i|c) \right) \tag{8.12}$$

where $P(c)$ denotes the class prior probability estimated from the selected set of labeled samples.[11] To avoid zero probabilities, we applied Laplacian smoothing.

8.7.2 Vector Model with Term Weighting for Utterance Classification

In information retrieval, document classification and text summarization, documents are usually represented by means of term vectors, D

$$D = a_1, a_2, \cdots, a_N \tag{8.13}$$

[11] The generic variable F is used to reflect the number of examples per category randomly selected from a corpus of labeled utterances. However, the practical number of sample utterances in a given class may be lower than F if there are less than F labeled utterances available for that category. We use this information for the estimation of the category priors $P(c)$.

where the components a reflect the relative significance of terms in relation to the document at hand.

In vector model utterance categorisation, we use a Nearest Neighbour classifier. Test utterances are represented as bag of words vectors. The nearest neighbour prototypes are "virtual documents" builded up by merging F utterances per category randomly selected. Virtual category documents are further weighted by using term scoring methods as explained in the following paragraphs.

Term scores are generally computed as a contribution of two factors: (i) the absolute or relative frequencies of terms in the document; and (ii) the term dispersion over all documents. The second factor is also used for feature selection characterizing the "noisy" behavior of terms.

8.7.2.1 Term Frequency

In the literature, one finds different definitions for the term frequency. In this work, we use two formulations taken from [40] and [41]:

$$TF_1(w,d) = \frac{C(w_i,d)}{\sum_j C(w_j,d)} \tag{8.14}$$

$$TF_2(w,d) = \begin{cases} 1 + \log(C(w_i,d)) & \text{if } C(w_i,d) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8.15}$$

where $C(w,d)$ denotes the occurrence counts of the term w in the document d.

8.7.2.2 IDF, RIDF and ISCF Scores

We analyzed three relevance scores to capture the term distribution across documents: the inverse document frequency (IDF), the residual inverse document frequency (RIDF) and a new formulation, the inverse spectral crest factor (ISCF).

- *Inverse document frequency (IDF)*. This popular definition was proposed by [42]:

  $$IDF(w) = -\log\left(\frac{ND_w}{ND}\right) \tag{8.16}$$

  where ND_w denotes the number of documents in which the term w occurs and ND is the total amount of documents in the collection. In this work, the number of documents corresponds to the number of categories $ND = L$.
- *Residual inverse document frequency (RIDF)*. This is a variant of the inverse document frequency, proven to be effective for automatic text summarization [42]. It represents the difference between the IDF of a term and its expected value $\widehat{IDF}$ according to a Poisson model.

  $$RIDF(w) = IDF - \widehat{IDF} \tag{8.17}$$

  with

  $$\widehat{IDF}(w) = -\log(1 - \mathrm{e}^{-\lambda_w}) \tag{8.18}$$

where λ_w denotes the parameter of the Poisson distribution, calculated here as the average occurrence of the w term across all ND documents:

$$\lambda_w = \sum_j N_{w_j}/ND\,. \tag{8.19}$$

The main advantage of $RIDF$ compared to IDF is that rare terms are not assigned relevances.

- *Inverse spectral crest factor (SCF)*. We propose a third metric called the inverse spectral crest factor (ISCF). Motivation for the introduction of this formulation is to achieve a more accurate indicator of the term distribution over the categories. An IDF-based metric would place lower relevance on terms observed in more than one category. However, this metric does not reflect the possibility that terms may occasionally appear in several categories.

  The Spectral Crest Factor (SCF) is one of the measures used in audio processing [43] for determining the noisy character of the signal components through an analysis of their short time spectra. It provides an estimate of the *spectral flatness*, as the ratio of the arithmetic mean energy across spectral bands with respect to the maximum energy. We adopted this metric to estimate a term's dispersion across categories. The term relevance is given by the *inverse spectral crest factor*, defined as:

$$ISCF(w) = \frac{ND \cdot \max_i(TF_1(w, d_i))}{\sum_j TF_1(w, d_j)}\,. \tag{8.20}$$

8.8 Evaluation Methods and Results

In this section, we describe our methods to evaluate the performance of the utterance classification models described in previous sections.

This is done by comparing the output categories which the proposed algorithm assigns to a number of test utterances with manually assigned categories thereof (the reference). If both categories coincide, the automatic categorization is considered correct, otherwise it is counted as error. As overall accuracy, we define

$$\text{accuracy} = \frac{\text{\# correctly classified test utterances}}{\text{\# total utterances in test set}} \tag{8.21}$$

8.8.1 Classification with One Labeled Utterance and Feature Extraction

Table 8.3 shows the accuracy values reached on the Internet corpus by the nearest-neighbor classifier applied to bags of words and feature vectors in the case of feature extraction, with one sample utterance per category. In this case, the samples have been manually selected in such a way that the number of overlapping terms in different category samples is minimized.

Table 8.3 Utterance categorization with one labeled utterance per class using several feature extraction techniques and disambiguation

Term clustering	Disambiguation	Accuracy
–	no	45%
–	yes	57%
Soft (PoBOC)	no	50%
Soft (PoBOC + FCMdd)	no	47%
Hard	no	50.8%
Hard	yes	62.2%

Comparing classification performance without disambiguation to the baseline (no term clustering; at 45%), we see that both soft and hard term clustering perform in a very similar way: PoBOC and hard term clustering achieve around 50% outperforming the baseline by about 10%, relative.

As noted in Section 8.6.2.1, disambiguation partially overcame the sparseness of having only one example utterance per class shown by significant improvements from 45% to 57% on the baseline without term clustering (27% relative) and 50.8% to 62.2% on hard term clustering (22% relative). Hard term clustering with disambiguation outperformed the baseline by 38.2%, relative.

8.8.2 Classification Based on F Samples per Category

The following paragraphs show a comparative analysis of the Naïve Bayes classifier and the approach based on weighted document vectors. In particular, we investigate the dependence between classifier performance and number of (randomly chosen) samples per category ($F = 1, \dots, 100$). Tests are performed on the Cable TV troubleshooting corpus (10 000 test utterances and 79 problem categories). Figure 8.4 depicts the performance of the Naïve Bayes classifier and the nearest-neigbor, using term weighting against the number of samples/category F. Based on these experimental results, several observations can be made:

- Naïve Bayes outperforms NN with term weighting and term relevance scoring ($TF_2(w, d)$) for numbers of samples greater than 7. The poorer performance of Naïve Bayes in these cases may be attributed to the use of Laplacian smoothing. For small numbers of examples, the ratio of terms with a frequency of zero in the set of examples is rather large (Figure 8.5).
  Therefore, using Laplacian smoothing in conjunction with the Naïve Bayes classifier may produce inaccurate term probability estimates. Note that, without the use of Laplacian smoothing, the $TF_2(w|c)$ and $P(w|c)$ would be identical.
- We also observed a dependency of the classifier's performance on the specific term frequency metric (TF_1 or TF_2, respectively). The normalization with respect to the document lengths introduced in TF_2 seems to be a better strategy

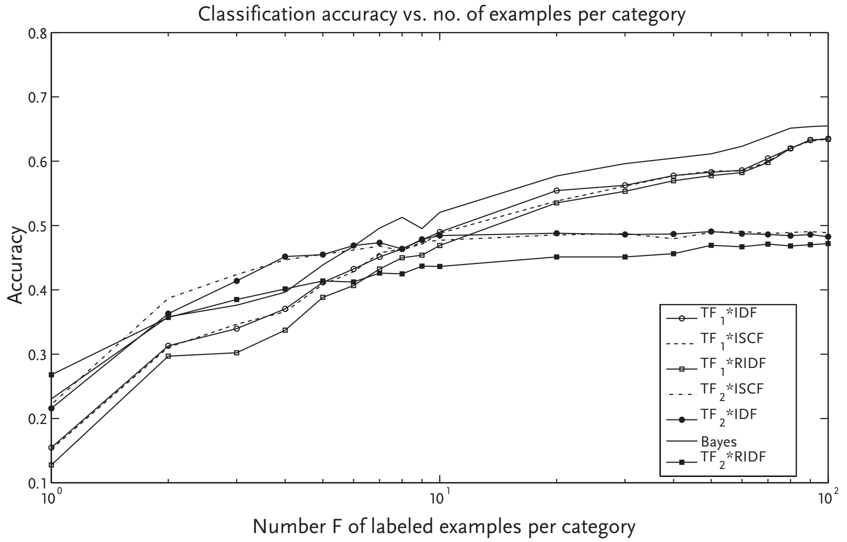


Figure 8.4 Mean accuracy rates achieved by the Naïve Bayes classifier and a vector model (nearest neighbor) with TFIDF, TFRIDF and TFISCF term weights in a logarithmic x-axis. Reported accuracy values refer to averaged results across 20 runs of the algorithm with different input sets of training utterances, randomly selected from a labeled corpus.

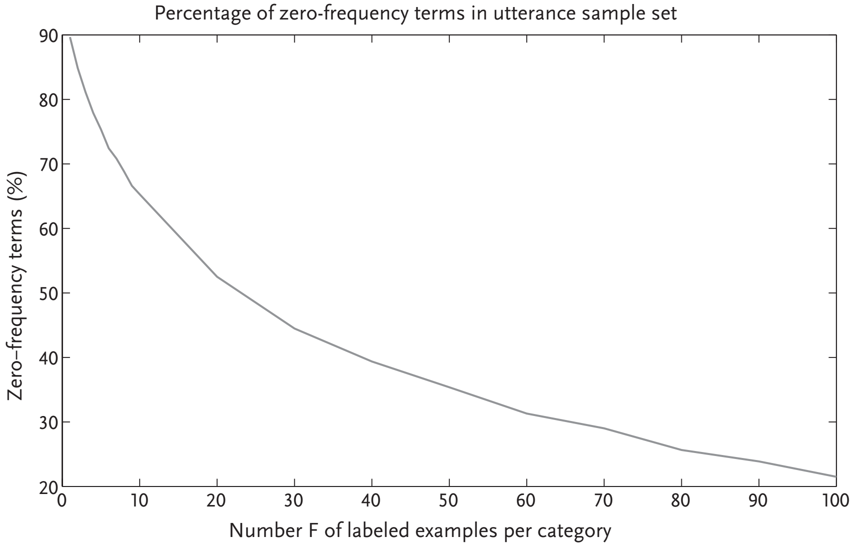


Figure 8.5 Ratio of terms in test utterances with zero frequency in sample utterance set vs number of sample utterances per category.

for few examples, but the classifier performance is stabilized after a number ($F = 7$) of samples. Also at this point, TF_1 starts to outperform TF_2. One possible reason for this phenomenon is the high sensitivity of classifiers to different utterance lengths when a small number of examples is provided.

- The contribution of *IDF*, *RIDF* and *ISCF*. Although *TFRIDF* was proposed as a more efficient solution in automatic text summarization, *TFIDF* has outperformed *TFRIDF* on this kind of data. This fact may be associated to certain characteristics of the utterance corpus and the way category documents are generated. On the one hand, there is a large number of terms which may be indicative of more than one category. This happens because the categories mentioned indicate different problems which can be experienced with a single device. For example, for problems related to the quality of the received image, utterances like *picture has poor quality* are commonly observed, and less frequently, utterances like *bad picture*. However, there also exist other categories to cover additional problems related to the picture. Here, we refer to terms such as *quality*, *poor* or *bad*, as specific category terms, in contrast to generic terms like *picture*. A generic term is descriptive for several categories simultaneoulsy, in which it occurs with quite high frequencies. Generic terms may be found to the extent that some underlying hierarchical category structure can be assumed. We also distinguish a third kind of term, referred to as *noisy terms*, which can be observed in many different categories, generally with low frequencies. It is desirable to emphasize specific terms with respect to generic terms, in order to "protect" utterances with a high probability of error like *bad picture*.

  In this respect, *IDF* scores capture a term's spreading over documents regardless of the term frequency in the documents. However, the average frequency of these terms (parameter λ_w of the Poisson model) considerably exceeds that of specific terms. According to a Poisson model, these terms (*picture*) should spread even more over documents in contrast to terms with low λ_w (specific terms), and, therefore, a part of the bias introduced by *IDF* appears compensated in the residual after subtracting the Poisson estimation $\widehat{IDF}$. Moreover, no significant differences can be observed between *TFISCF* and *TFIDF*. The use of *ISCF* scores was motivated to provide a more precise estimate of the term/category distribution which reflects the different frequences of the term in the category documents. However, one fact to be considered is that *IDF* and *ISCF* scores are here multiplied to *TF* scores. This may also explain why, despite its simplicity, *TFIDF* scores are among the most broadly used metrics in text processing. Whether *ISCF* can be effective for global feature selection remains an open question.

8.9
Summary and Conclusion

In this article, we have described different models for the categorization of caller utterances in the scope of automated troubleshooting agents. One of the goals of this research is to help overcome costs associated with the manual compilation of large training data sets. In the first part of the article, we proposed categorization schemes which make use of only one labeled sample per category. The proposed solution is based on feature-extraction techniques which automatically identify semantic word classes on a corpus of unlabeled utterances. Hard and fuzzy word-clustering methods were compared. The performance of feature extraction for utterance classification was experimentally evaluated on a test corpus of more than 3000 utterances and 28 classes. The most optimistic outcomes were achieved with hard word clustering in combination with a module for reallocating ambiguous utterances providing a maximum of 62.2% accuracy.

The second part of the paper provided an overview of supervised classifiers commonly used for the categorization of texts. A probabilistic framework (Naïve Bayes) and a vector model with term relevance scores were described. We experimentally compared these classifiers on a test corpus of 10 000 utterances and 79 classes. An analysis of the classifier's dependency on the number of labeled examples was carried out. Our experiments reported an inflection point in the classifier's behavior around seven training samples per category. For lower numbers of training samples, nearest-neighbor classification with term-weighting schemes achieved higher accuracies, whereas for larger numbers, Naïve Bayes outperfoms the other classifiers.

References

1 Minker, W., Albalate, A., Bühler, D., Pittermann, A., Pittermann, J., Strauss, P. and Zaykovskiy, D. (**2006**) Recent trends in spoken language dialogue systems, in *Proc. of the ICIT*, Cairo, Egypt.

2 Bühler, D., Minker, W. and Elciyanti, A. (**2005**) Using language modelling to integrate speech recognition with a flat semantic analysis, in *Proc. of the SIGdial Workshop on Discourse and Dialogue*, Lisbon, Portugal.

3 Minker, W. (**1998**) *Speech Understanding for Spoken Language Systems – Portability Across Domains and Languages*. Hänsel-Hohenhausen, Frankfurt, Germany.

4 Bach, E. and Harms, R. (**1968**) *Universals in Linguistic Theory*. Holt, Rinehart and Winston, New York, USA.

5 Rabiner, L. (**1989**) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, **77**(2).

6 Morimoto, T., Shikano, K., Iida, H. and Kurematsu, A. (**1990**) Integration of speech recognition and language processing in spoken language translation system (SL-Trans), in *Proc. of the ICSLP*, Kobe, Japan.

7 Levin, E. and Pieraccini, R. (**1995**) CHRONUS, the next generation, in *Proc.*

of the ARPA Workshop on Human Language Technology, Austin, USA.

8 Interactive Services Design Guidelines(**1995**) Technical Report ITU-T Recommendation F.902, ITU, Geneva, Switzerland.

9 Acomb, K., Bloom, J., Dayanidhi, K., Hunter, P., Krogh, P., Levin, E. and Pieraccini, R. (**2007**) Technical support dialog systems: issues, problems, and solutions, in *Proc. of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, Rochester, USA.

10 Evanini, K., Suendermann, D. and Pieraccini, R. (**2007**) Call classification for automated troubleshooting on large corpora, in *Proc. of the ASRU*, Kyoto, Japan.

11 Blum, A. and Mitchell, T. (**1998**) Combining labeled and unlabelled data with co-training, in *Proc. of the COLT*, Madison, USA.

12 Wulfekuhler, M. and Punch, W. (**1997**) Finding salient features for personal web page categories, in *Proc. of the International Web Conference*, Santa Clara, USA.

13 Jain, A. and Mao, J. (**2000**) Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **22**(1).

14 Pyle, D. (**1999**) *Data Preparation for Data Mining*. Morgan Kaufmann, Los Altos, USA.

15 Aksoy, S. and Haralick, R. (**2001**) Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, **22**(5).

16 Jain, A. and Dubes, R. (**1988**) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, USA.

17 Johnson, S. (**1967**) Hierarchical clustering schemes. *Psychometrika*, **32**(3).

18 Kauffmann, L. and Rousseeuv, P. (**1990**) *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley & Sons, New York, USA.

19 Hartigan, J. and Wong, M. (**1979**) Algorithm AS136: A k-means clustering algorithm. *Appl. Stat.*, **28**, 100–108.

20 Ng, R. and Han, J. (**1994**) Efficient and effective clustering methods for spatial data mining, in *Proc. of the 20th Conference on VLDB*, Santiago, Chile.

21 Ester, M., Kriegel, H., S, J. and Xu, X. (**1996**) A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proc. of KDD-96*.

22 Kriegel, H.-P., Ankerst, M., Breunig, M.M. and Sander, J. (**1999**) OPTICS: Ordering Points To Identify the Clustering Structure, in *Proc. of ACM-SIGMOD International Conference on Management of Data*, Philadelphia, Pennsylvania, United States.

23 Hinneburg, A. and Keim, D.A. (**1998**) An efficient approach to clustering in large multimedia databases with noise. *Knowledge Discovery and Data Mining*, 58–65.

24 Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. (**1998**) Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, 94–105.

25 Zhang, T., Ramakrishnan, R. and Livny, M. (**1997**) Birch. A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, **1**(2).

26 Dempster, A., Laird, N. and Rubin, D. (**1977**) Maximum likelihood from incomplete data via EM algorithm. *Journal of Royal Statistical Society*, **39**(1).

27 Chapelle, O., Schölkopf, B. and Zien, A. (**2006**) *Semi-Supervised Learning*. MIT Press, Cambridge, USA.

28 Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (**2001**) Constrained K-means clustering with background knowledge, in *Proc. of the ICML*, Williamstown, USA.

29 Salton, G., Wong, A. and Yang, C.S. (**1975**) A vector space model for automatic indexing. *Communication of the ACM*, **18**(11).

30 Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (**1990**) Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, **41**(6), 391–407.

31 Hofmann, T. (**1999**) Probabilistic latent semantic analysis, in *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm.

32 Li, Y.H. and Jain, A.K. (**1998**) Classification of text documents. *Comp. J.*, **41**(8).

33 Klein, D. and Manning, C.-D. (**2003**) Fast exact inference with a factored model for natural language parsing. *Advances in Neural Information Processing Systems*, **15**, 3–10.

34 Minnen, G., Carrol, J. and Pearce, D. (**2001**) Applied morphological processing of English. *Natural Language Engineering*, **7**(3).

35 Buckley, C. (**1985**) Implementation of the SMART information retrieval system. *Technical report*, Cornell University, Ithaca, USA.

36 Li, Y. and Jain, A. (**1998**) Classification of text documents. *Comp. J.*, **41**(8).

37 Picard, J. (**1999**) Finding content-bearing terms using term similarities, in *Proc. of the EACL*, Bergen, Norway.

38 Cleuziou, G., Martin, L. and Vrain, C. (**2004**) PoBOC: An overlapping clustering algorithm. Application to rule-based classication and textual data, in *Proc. of the ECAI*, Valencia, Spain.

39 Krishnapuram, R., Joshi, A., Nasraoui, O. and Yi, L. (**2001**) Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans. on Fuzzy Systems*, **9**(4).

40 Debole, F. and Sebastiani, F. (**2003**) Supervised term weighting for automated text categorization, in *Proc. of the SAC*, Melbourne, USA.

41 Mori, T., Kikuchi, M. and Yoshida, K. (**2001**) Term weighting method based on information gain ratio for summarizing documents retrieved by IR systems, in *Proc. of the NTCIR Workshop*, Tokyo, Japan.

42 Murrayand, G., Renals, S. (**2007**) Towards online speech summarization, in *Proc. of the Interspeech*, Antwerp, Belgium.

43 Peeters, G. (**2003**) A Large Set of Audio Features for Sound Description. *Technical report*, IRCAM, Paris, France.

9
Semi-Supervised Clustering in Functional Genomics

Johann M. Kraus, Günther Palm, Friedhelm Schwenker, Hans A. Kestler[1]

9.1
Introduction

Cluster analysis is a classical example of unsupervised learning. It is an important technique of explorative data analysis. It is used in different contexts such as text analysis, marketing, psychology and biology [1–8]. Cluster analysis is a collection of methods for generating hypotheses about the structure of the data by solely exploring pairwise distances or similarities in feature space. It is often applied as a first step in data analysis for the creation of initial hypotheses.

This chapter provides a brief introduction to the principles of cluster analysis, presents some basic algorithms, and mentions open challenges of clustering in functional genomics. In the next section we introduce functional genomics as our field of application and motivate the use of cluster analysis in this setting. Research and development of high-throughput technologies offer a powerful approach to the study of biological processes. For instance, microarray experiments give gene expression levels from thousands of genes simultaneously. In this context data-mining methods are paramount to the analysis of high-dimensional data. Unsupervised clustering can give an estimation of the functional grouping of genes or the pre-classification of profiles. As, for example, hierarchical clustering algorithms can predict a basic branching structure, they are often used in this context. Some cluster analysis basics are presented in Section 9.3. In consequence to the unknown a priori information about the data, validation of clustering results turns out to be critical. For instance, the use of different cluster methods on the same data set can lead to contradictory hypotheses. Up to now no general framework for the choice of the best suitable cluster method and validation procedure is known.

Clustering has been shown to be highly susceptible to minor changes on the data, for example, noise injection, feature selection or subsampling. Research on the robustness aspects of cluster algorithms is often neglected. But as we point out in Section 9.4, investigating robustness must be a common topic when handling data.

[1] Corresponding author.

A limitation in clustering microarray data concerns the reproducibility of clustering results after small modifications on the data, such as adding new points to the data set. Although cluster analysis is unsupervised and does not require an external teacher signal, recent work has shown the usefulness of incorporating additional information [9–16]. This knowledge can be provided as a set of labeled data items or a set of constraints between pairs of data items. Recent experiments have indicated that already sparse additional information may support the cluster analysis in building more stable partitions.

9.2 Biological Background

9.2.1 Functional Genomics

Functional genomics as part of molecular biology aims to provide a link between genomic information and biological functions, such as establishing a relationship between gene expression patterns and tumour status, see also Chapter 5, Figure 1.1. The term gene expression describes the process of translating the information encoded in the genome (DNA) into a biologically functional gene product, for example, proteins. Research on functional genomics grants access to a better understanding of molecular functions. Cell functions are supposed to be mostly coordinated by the expression and interaction of proteins. Although, recently, post-translational modification has been a major issue [17, 18]. Almost all cells in an organism contain the same set of genes, but they differ in their expression patterns depending on their specialization. The expression level and specification of proteins has been shown to be associated to the occurrence of many diseases [19, 20]. For instance, an increased expression level of the oncogene *HER2* is related to the emergence of mammalian carcinoma [21, 22]. To support biologists in generating new hypotheses about gene expression patterns mostly high-throughput technologies are used. As an example, DNA microarray technology enables a simultaneous analysis of thousands of genes.

9.2.2 DNA Microarray Technology

The central dogma of molecular biology forms the backbone of DNA microarray technology. All information encoded in DNA is transcribed into RNA and further may be translated into proteins [23]. As described in the previous section, one part of functional genomics is based on measuring gene expression levels. Quantifying proteins is difficult due to their chemical and structural differences. DNA microarray technology circumvents the problems of measuring proteins by quantifying RNA expression levels. According to the central dogma of molecular biology, knowledge about RNA expression leads to knowledge about protein expression. But assigning protein levels to RNA expression can be insufficient as the gene expression process is known to be much more complicated. For instance, gene regulation is affected by RNA interference. The small RNAs from both inside and outside the cell are processed by the RNA interference machinery to inhibit genes and proteins by cleaving messenger RNAs, blocking protein synthesis, or inhibiting transcription. Theses small RNAs are called small interfering RNA (*siRNA*) and micro RNA (*miRNA*) depending on their endogenous or exogenous origin [17, 18].

The raw data from microarray experiments usually needs preprocessing steps to reduce the influences of technical noise or experimental biases. These steps often include normalization and filtering methods. Further information about microarray technology and the research on preprocessing steps can be found in Allison *et al.* [24] and others, see [19, 25–27]. In this article we focus on unsupervised data-mining methods that follow the preprocessing step. Early studies on microarrays [28–30] applied standard data-mining tools. Recently more and more methods were adopted to the special issues arising from the analysis of microarray data [31–33].

A microarray chip typically covers a huge number of genes. Additionally the number of analyzed microarray chips per experiment may be insufficient for standard procedures because of the small number of examined individuals. As the high-dimensional microarray data sets usually provide only a few samples, most algorithms suffer from the *curse of dimensionality* [34]. Noisy data and unreliable a priori knowledge about the data may also bias the analysis. In the next section we describe basics of cluster analysis and focus on algorithms first used in microarray data analysis. In Section 9.4 we introduce our research on the robustness of cluster analysis.

9.3
Cluster Analysis

Let $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a set of gene expression profiles, where the feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$ is extracted from the gene expression of probe i. Cluster analysis is used to build a partition of a data set containing k clusters such that data points within a cluster are more similar to each other than points from different clusters. Figure 9.1 illustrates the idea of clustering data. As humans are rather good at identifying clusters of different shapes in up to three dimensions, one can easily recognize groups in the picture. Generally the precise number of clusters or a grouping of clusters depends mostly on the subjective evaluation of different observers.

A partition $P(k)$ is a set of clusters $\{C_1, C_2, \ldots, C_k\}$ with $0 < k < N$ and meets the following conditions:

$$\bigcup_{i=1}^{k} C_i = X, C_i \neq \emptyset \tag{9.1a}$$

$$C_i \cap C_j = \emptyset, i \neq j \tag{9.1b}$$

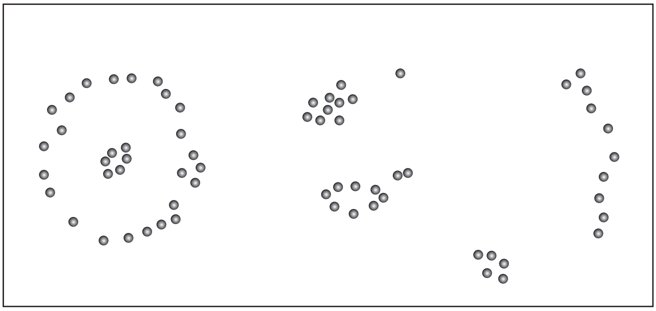


Figure 9.1 Illustration of different concepts of clusters. For instance, this picture includes clusters of regular shape, an elongated cluster, a cluster surrounding another cluster, a group of similar clusters and outlying clusters.

Another view on clustering is given by hierarchical cluster analysis. It builds a sequence of partitions which are usually displayed as dendrograms. Figure 9.2 shows both types of cluster analysis.

The basic clustering task can be formulated as an optimization problem.

Definition 9.1 *Partitional cluster analysis: For a fixed number of groups k find that partition $P(k)$ of a data set $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ out of the set of all possible partitions $\Phi(X, k)$ for which a chosen objective function $f\colon \Phi(X, k) \to \mathbb{R}^+$ is optimized.*

Solution: For all possible partitions with *k* clusters, compute the value of the objective function *f*. The partition with the best value is the set of clusters sought. This method is computationally infeasible as the cardinality of the set of all possible partitions is huge even for small *k* and *N*. The cardinality of $\Phi(X, k)$ can be computed by the Stirling numbers of the second kind [1]:

$$|\Phi(X, k)| = S_N^k = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} i^N \qquad (9.2)$$

As a consequence, existing algorithms provide a heuristic for this search problem.

9.3.1
Clustering Microarray Data

In the context of microarray analysis one can either aim to identify groups of genes or groups of individuals. A group of genes with similar expression measured under various conditions may imply a co-regulation of the genes (gene-based clustering). A clustering of individuals into groups with similar expression patterns may help to discover unknown subtypes of diseases (sample-based clustering). Further approaches known as biclustering and subspace clustering combine both clustering perspectives and allow to search for groups of genes with similar expression patterns only in a subset of the samples.

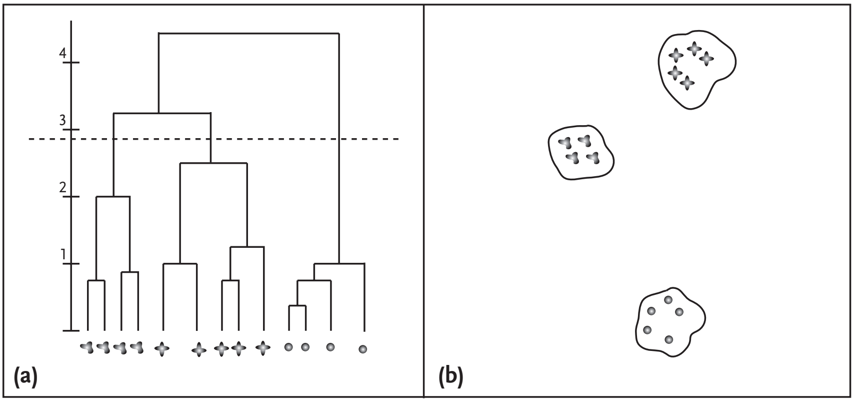


Figure 9.2 Results from a hierarchical and a partitional cluster analysis. The dendrogram (a) gives an overview of different possible subclusters. Splitting the dendrogram along the dotted line results in the same clustering as computed by a partitional cluster algorithm ($k = 3$), (b).

One of the first cluster algorithms applied to microarray data was used by Eisen *et al.* [28]. They performed a hierarchical clustering (*UPGMA*) and introduced a graphical display of the cluster results by building heat maps. Figure 9.3 illustrates this kind of data view.

Heat maps give an overview of the data. The expression levels of many genes are plotted against a number of samples. The columns depict the N-dimensional gene vectors (gene expression profile) of the samples. Each color on the grid represents a gene expression value in which similar values must attain similar colors. The samples and genes are ordered by their results from hierarchical cluster analysis. Recently, alternatives to this representation have been proposed incorporating confidence values and perceptually optimizing the red–green color scales [37].

In the following some other well known cluster algorithms were applied to microarray data, such as *k-means* [29] and self-organizing feature maps (*SOM*) [30]. Recently more and more complex cluster algorithms are proposed including variants from model-based clustering, spectral clustering and biclustering [38–40], although the above mentioned simple methods remain predominant. They benefit from their conceptual simplicity or easy availability through standard software.

There are no general guidelines for choosing the most appropriate cluster method. Cluster analysis aims to reveal an unknown hidden structure. But most cluster algorithms will even predict a structure in data sets sampled from a uniform distribution.

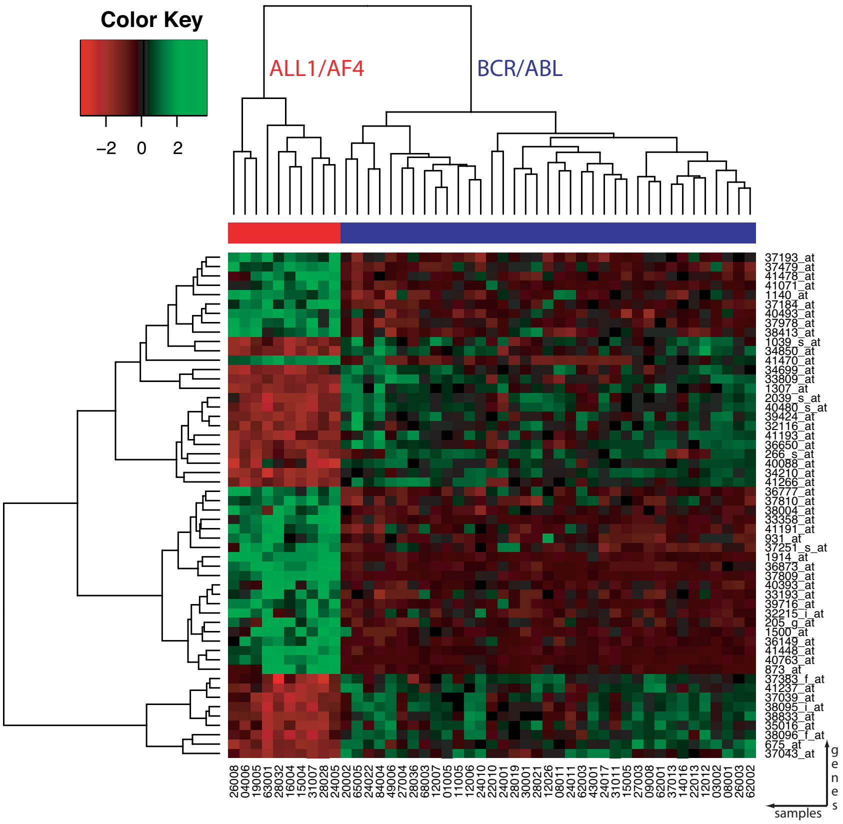


Figure 9.3 Heat map with two dendrograms. Graphical display of hierarchical clustering results in microarray data mining (created with the software R [35] and a freely available ALL (acute lymphatic leukemia) microarray data set from Chiaretti *et al.* [36]). Each sample is represented by a single column of colored boxes. Each box corresponds to the color-coded expression level of one gene in one sample. Two clusters of columns are indicated by the bars on the top. These clusters correspond to known groups of patients with a chromosomal translocation between chromosomes 4 and 11 (ALL/AF4) and chromosomes 9 and 22 (BCR/ABL), respectively, see [36].

9.3.2
Cluster Methods

9.3.2.1
Hierarchical Clustering

Hierarchical clustering builds a sequence of partitions $P^1, P^2, \ldots, P^N$ where each partition is a refinement of its predecessor. As mentioned before, hierarchical cluster methods build a dendrogram, that is a hierarchy of partitions. There are two ways to construct the dendrograms, top-down (divisive) and bottom-up (agglomerative). Divisive hierarchical clustering starts with the whole data set in one cluster.

At the refinement step one target cluster C_i out of $P = \{C_1, \ldots, C_k\}$ is selected and split into two clusters. This step is recursively repeated until all clusters contain only one element. Different methods are used to perform target cluster selection and splitting. One example is always splitting the cluster with the largest diameter. Splitting can be done by separating the element $\boldsymbol{x}_z$ with maximal dissimilarity to generate the new cluster. All other elements of the target cluster are then assigned to their nearest cluster, that is, they are either moved to the new cluster or remain in the target cluster.

Algorithm 2 Agglomerative hierarchical clustering

Data set: $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$

1. Start with partition $P^N = \{C_1, \ldots, C_N\}$, where each cluster C_i has exactly one element $\boldsymbol{x}_i$.
2. Identify those clusters C_i and C_j having the minimal distance $\operatorname{dist}(C_i, C_j)$.
3. Merge clusters C_i and C_j to cluster C_F.
4. Build the new partition by removing C_i and C_j and adding cluster C_F.
5. Repeat step (2) to (4) until partition $P_1 = \{C_1\}$ is reached.

The more popular agglomerative hierarchical clustering variants proceed by starting with a partition into N clusters and subsequently merging the nearest clusters, see Algorithm 2. In step (2) of this algorithm the distance calculation has to be extended from distances between data points to distances between clusters. Given a distance measure dist defined on the input domain X, the distance between two clusters can be computed in different ways. Everitt *et al.* [5] mention the following inter-cluster distance formula. Let C_i and C_j be the two clusters which are to be combined to cluster C_F in the agglomeration step. Let C_r be another cluster, the distance of cluster C_F to cluster C_r is then calculated by:

$$\begin{aligned}\operatorname{dist}(C_F, C_r) = \alpha_i \operatorname{dist}(C_i, C_r) + \alpha_j \operatorname{dist}\left(C_j, C_r\right) + \\ \beta \operatorname{dist}\left(C_i, C_j\right) + \gamma |\operatorname{dist}(C_i, C_r) - \operatorname{dist}\left(C_j, C_r\right)| \, . \end{aligned} \tag{9.3}$$

Varying parameters α_i, α_j β, and γ in (9.3) leads to different inter-cluster distances. For instance, choosing $\alpha_i = \alpha_j = 1/2, \beta = 0, \gamma = -1/2$ gives the single-linkage cluster distance:

$$\begin{aligned}\operatorname{dist}(C_F, C_r) =& \frac{1}{2}\left(\operatorname{dist}(C_i, C_r) + \operatorname{dist}\left(C_j, C_r\right)\right) \\ &- \frac{1}{2}|\operatorname{dist}(C_i, C_r) - \operatorname{dist}\left(C_j, C_r\right)| \\ =& \min\left(\operatorname{dist}(C_i, C_r), \operatorname{dist}\left(C_j, C_r\right)\right) \, .\end{aligned} \tag{9.4}$$

Here, at each agglomeration step the two clusters are combined which contain the two elements with the minimum inter-cluster distance. Following Everitt *et al.* [5],

Table 9.1 Inter-cluster distance methods in agglomerative hierarchical clustering. Substituting parameters $\alpha_i, \alpha_j, \beta, \gamma$ in (9.3) with the given values describes different cluster distances. e_x denotes the number of elements in cluster C_x.

Distance measure	α_i	α_j	β	γ
single-linkage	1/2	1/2	0	–1/2
complete-linkage	1/2	1/2	0	1/2
unweighted average	1/2	1/2	0	0
median	1/2	1/2	–1/4	0
group average	$e_i/(e_i + e_j)$	$e_j/(e_i + e_j)$	0	0
centroid	$e_i/(e_i + e_j)$	$e_j/(e_i + e_j)$	$-(e_i e_j)/(e_i + e_j)^2$	0
Ward's method	$(e_i + e_r)/(e_i + e_j + e_r)$	$(e_j + e_r)/(e_i + e_j + e_r)$	$-e_r/(e_i + e_j + e_r)$	0

Table 9.1 lists some more popular distance measures and the associated parameter setting.

9.3.2.2 Partitional Clustering

K-means [41] is probably the most popular partitional cluster algorithm. Given a number k, the algorithm splits the data set into k disjoint clusters. Here cluster centroids $\mu_1, \ldots, \mu_k$ are placed in the center of gravity of clusters $C_1, \ldots, C_k$. The objective function of k-means is:

$$F(\mu_j, C_j) = \sum_{j=1}^{k} \sum_{\boldsymbol{x}_i \in C_j} ||\boldsymbol{x}_i - \mu_j||^2 . \tag{9.5}$$

Minimizing this function amounts to minimizing the sum of squared distances of data points from their cluster centroids. K-means is implemented by iterating between two major steps that (1) reassign data points to nearest cluster centroids and (2) update centroids for the newly assembled cluster. The cluster centroid μ_j is updated by computing the centroid of all points in cluster C_j:

$$\mu_j = \frac{1}{|C_j|} \sum_{\boldsymbol{x}_i \in C_j} \boldsymbol{x}_i \; . \tag{9.6}$$

To show the convergence of k-means we restrict the argument to using the Euclidean distance. Given $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, let $\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y})^2$ denote the squared Euclidean distance between these points,

$$\mathrm{dist}(\boldsymbol{x}, \boldsymbol{y})^2 = ||\boldsymbol{x} - \boldsymbol{y}||_2^2 = \sum_{m=1}^{d} (x_m - y_m)^2 = (\boldsymbol{x} - \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y}) \; , \tag{9.7}$$

where $\boldsymbol{x} \cdot \boldsymbol{y}$ is the dot product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. K-means is started with k randomly chosen centroids μ_j. We show that the two steps, (1) reassigning data points and (2) updating centroids, will monotonically decrease the objective function F.

(1) Reassignment step:

A point $\boldsymbol{x}_i$ is assigned to cluster C_j **iff** $\|\boldsymbol{x}_i - \mu_j\|_2^2 \leq \|\boldsymbol{x}_i - \mu_l\|_2^2$, $\forall l \neq j$. Therefore F is monotonically decreasing.

(2) Update step:

Let $\mu_j^{'}$ be the new centroid of the modified cluster C_j. Then

$$\begin{aligned}
\Delta F &= \sum_{\boldsymbol{x}_i \in C_j} \|\boldsymbol{x}_i - \mu_j^{'}\|_2^2 - \sum_{\boldsymbol{x}_i \in C_j} \|\boldsymbol{x}_i - \mu_j\|_2^2 \\
&= \sum_{\boldsymbol{x}_i \in C_j} \left[\|\boldsymbol{x}_i - \mu_j^{'}\|_2^2 - \|\boldsymbol{x}_i - \mu_j\|_2^2 \right] \\
&= \sum_{\boldsymbol{x}_i \in C_j} -\|\mu_j^{'} - \mu_j\|_2^2 - \sum_{\boldsymbol{x}_i \in C_j} 2(\boldsymbol{x}_i - \mu_j^{'}) \cdot (\mu_j^{'} - \mu_j) \\
&= -|C_j| \, \|\mu_j^{'} - \mu_j\|_2^2 - 2(\mu_j^{'} - \mu_j) \cdot \sum_{\boldsymbol{x}_i \in C_j} (\boldsymbol{x}_i - \mu_j^{'}) \\
&= -|C_j| \, \|\mu_j^{'} - \mu_j\|_2^2 \leq 0 \, .
\end{aligned} \tag{9.8}$$

The last step follows from defining $\mu_j^{'}$ as being the centroid of point mass C_j, that is $\sum_{\boldsymbol{x}_i \in C_j} (\boldsymbol{x}_i - \mu_j^{'}) = \boldsymbol{0}$. Therefore F is monotonically decreasing.

As the objective function F is bounded by zero, and k-means is ensured to always monotonically decrease F, the previous deduction shows the convergence of the k-means algorithm. See Selim and Ismail [42] for a detailed proof of the convergence of k-means-type algorithms to a locally optimal partitioning of the data set X.

Bezdek [43] describes a modified version of the k-means algorithm. The algorithm *fuzzy-c-means* uses a fuzzifier to enable a so-called *soft clustering*. In this task no definite assignment of data points to one distinct cluster is given but a *membership matrix U* is computed. The entries of the membership matrix u_{ij} denote the degree of membership of each data point $\boldsymbol{x}_i$ to each cluster C_j. What is special about fuzzy-c-means is its ability to provide more information on the computed clustering decision. For instance, the membership matrix describes that a data point $\boldsymbol{x}_i$ belongs to cluster C_1 with the computed likelihood of 0.50 and to cluster C_2 with 0.50. The objective function of fuzzy-c-means is:

$$F(\mu_j, U) = \sum_{j=1}^{k} \sum_{i=1}^{N} u_{ij}^m \|\boldsymbol{x}_i - \mu_j\|^2 \, , \tag{9.9}$$

with $m > 1$. The entries u_{ij} of the membership matrix U have to satisfy the following restrictions:

$$u_{ij} \geq 0\,, \quad i = 1 \ldots N\,, \quad j = 1 \ldots k \tag{9.10a}$$

$$\sum_{j=1}^{k} u_{ij} = 1\,, \quad i = 1 \ldots N \tag{9.10b}$$

$$0 < \sum_{i=1}^{N} u_{ij} < N\,, \quad j = 1 \ldots k\,. \tag{9.10c}$$

The parameter m (fuzzifier) controls the influence of the membership matrix to the clustering. For m close to 1 the algorithm behaves similar to k-means. An implementation of fuzzy-c-means follows the k-means procedure as described above, including (1) update of the centroids:

$$\mu_j = \frac{1}{\sum_{i=1}^{N} u_{ij}^m} \sum_{i=1}^{N} u_{ij}^m \boldsymbol{x}_i\,, \quad j = 1 \ldots k \tag{9.11}$$

and (2) assignment of the cluster memberships:

$$u_{ij} = \frac{1}{\sum_{l=1}^{k} \left(||\boldsymbol{x}_i - \mu_j||^2 / ||\boldsymbol{x}_i - \mu_l||^2 \right)^{1/m-1}}\,, \quad i = 1 \ldots N\,, \quad j = 1 \ldots k\,. \tag{9.12}$$

9.3.2.3 Incremental Updates

Besides these batch-mode algorithms described so far, k-means and fuzzy-c-means can be implemented using an incremental update. In this variant, each time a data point $\boldsymbol{x}_i$ is presented, a centroid update is performed immediately. The update shifts the chosen centroid μ_j by a fraction (learning rate) of the distance to the presented data point. The amount of shift can be controlled by using a time varying and cluster specific learning rate $\alpha_j = 1/n_j$ instead of a global one. Here n_j denotes the number of updates which have already been computed for cluster C_j:

$$\mu_j^* = \mu_j^* + \alpha_j(\boldsymbol{x}_i - \mu_j^*)\,; \qquad \mu_j^* - \text{nearest (winning) prototype}\,. \tag{9.13}$$

It is possible not only to update the chosen cluster centroid (winner takes all), but also a number of neighboring cluster centroids. The algorithm *self-organizing feature maps* [44] (SOM) follows this approach. SOM uses an internal representation of cluster prototypes derived from a projection of the centroids to a low-dimensional grid, see Figure 9.4.

The neighborhood information is evaluated on this grid. An example for a neighborhood function on the grid is:

$$N(i,j) = \exp\left(-\frac{||p(i) - p(j)||^2}{2\sigma^2} \right) \tag{9.14}$$

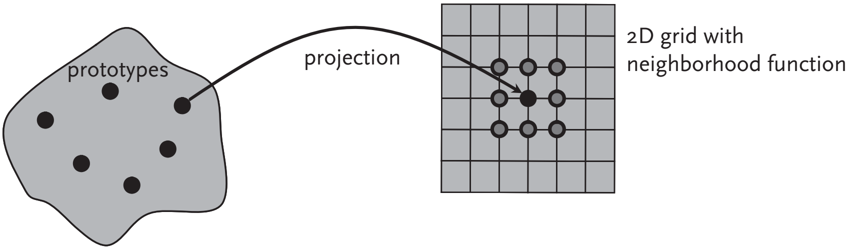


Figure 9.4 Prototypes and grids in SOM (self-organizing feature maps). Prototypes are projected onto a 2D-grid. On this grid a neighborhood function is defined. This neighborhood influences the update step of the algorithm, see (9.15).

with $p(i)$ and $p(j)$ denoting the projected position of centroids μ_i and μ_j on the grid. The learning rate α_t and the neighborhood function N are used to compute the amount of update for each cluster centroid μ_i after assigning point $\boldsymbol{x}_i$ to its nearest cluster centroid μ_j:

$$\mu_i = \mu_i + \alpha_t N(i,j)(\boldsymbol{x}_i - \mu_i) \ . \tag{9.15}$$

In each update step a whole neighborhood is changed, and subsequently similar data points are aggregating while the grid finally forms a low-dimensional map representing the high-dimensional data space.

9.3.2.4
Model-Based Clustering

Model-based methods [45–51] offer a statistical framework to model cluster structures. They are based on the assumption that the data set originates from a mixture of several underlying probability distributions. Each distribution with its associated parameters (mean, variance, and so forth) corresponds to one of the expected clusters. The aim of clustering under this perspective is to estimate the model parameters and the hidden cluster membership information. Suppose the data X consists of N independent multivariate observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$ sampled from k distributions. The likelihood of the mixture model is defined by:

$$\mathcal{L}(\theta_1, \dots, \theta_k, \tau_1, \dots, \tau_k | X) = \prod_{i=1}^{N} \sum_{j=1}^{k} \tau_j f_j(\boldsymbol{x}_i | \theta_j) \ , \tag{9.16}$$

where density f_j and parameters θ_j belong to the j-th component of the mixture. The probability that a data point $\boldsymbol{x}_i$ belongs to the j-th component is given by τ_j ($\tau_j \geq 0$; $\sum_{j=1}^{k} \tau_j = 1$). In the most commonly used Gaussian mixture model, each component j is defined by a multivariate normal density with parameters μ_j (mean vector) and Σ_j (covariance matrix):

$$f_j(\boldsymbol{x}_i | \mu_j, \Sigma_j) = \frac{\exp\{-1/2(\boldsymbol{x}_i - \mu_j)^T \Sigma_j^{-1} (\boldsymbol{x}_i - \mu_j)\}}{\sqrt{\det(2\pi\Sigma_j)}} \ . \tag{9.17}$$

The covariances determine geometric features of the components (clusters), as for instance shape, volume or orientation. The parameters θ_j and τ_j are usually estimated by the *EM algorithm* [52]. The EM algorithm iterates between the two major steps (1) expectation and (2) maximization. In the expectation step the hidden parameters τ_j are estimated from the data with the mixture model being based on the current parameters θ_j. In the maximization step the model parameters θ_j are updated to maximize the likelihood of the data given the estimated hidden parameters. When the EM algorithm converges, all data points are assigned to the cluster with the highest conditional probability. The k-means algorithm mentioned before is known to be equivalent to model-based clustering using the equal volume spherical model [45]. Also hierarchical clustering based on centroid methods or Ward's method can be designed by model-based clustering with underlying spherical Gaussians [45].

9.3.2.5 Spectral Clustering and Other Graph-Based Methods

Spectral clustering [53–56] is closely linked to a graph-based perspective of data, where a data set X is represented by its similarity graph $G(V, E)$. In this notation each vertex $v_i \in V$ stands for a data point $\boldsymbol{x}_i \in X$ and each edge e_{ij} provides information whether two data points x_i, x_j are connected. In similarity graphs an edge is weighted by the similarity between the corresponding data points. A graph is then given by its weighted adjacency matrix $W = (w_{ij})$, $i, j = 1 \ldots n$, where $w_{ij} \geq 0$ represents the similarity between two vertices v_i and v_j. It is assumed that w_{ij} is symmetric. The degree of a vertex v_i is defined as $d_i = \sum_{j=1}^{N} w_{ij}$. The degree matrix D is the diagonal matrix with entries $d_1, \ldots, d_N$.

Clustering is reformulated in terms of graph partitioning as finding groups of vertices such that the edges within a group of vertices have high weights and the edges between groups have low weights. Furthermore, a connected component is a subset of connected vertices that has no edge to the rest of the graph. Spectral clustering algorithms use graph Laplacians to solve the graph-clustering problem. The graph Laplacian is defined as $L = D - W$. A property of the graph Laplacian is that the multiplicity k of the eigenvalue $\lambda_1 = 0$ is equal to the number of connected components [57]. The least eigenvalue $\lambda_2 > 0$ can be used for a spectral bisection algorithm [58]. Therefore, the eigenvector u corresponding to the eigenvalue λ_2 is computed. A vertex $v_i \in V$ is put into cluster C_1 if $u_i < 0$ and into cluster C_2 if $u_i > 0$. This method can be extended to the use of k smallest eigenvectors or to normalized spectral clustering based on normalized Laplacians [59–61].

The eigenvalue λ_2 is also called the algebraic connectivity of $G(V, E)$ which gives rise to a graph cut point of view. For two subsets $A, B \in V$ a cut is defined as $\text{cut}(A, B) = \sum_{u_i \in A, v_j \in B} w_{ij}$. Then clustering is a means of solving the mincut problem, that is choosing a partition $C_1, \ldots, C_k$ that minimizes $\text{cut}(C_1, \ldots, C_k) = \sum_{i=1}^{k} \text{cut}(C_i, \overline{C_i})$ [62]. Additionally, the clusters C_i are often required to satisfy an additional size constraint. Two modified objective functions are RatioCut [63] and Ncut [59], which weight the cut by the number of vertices and the sum of the vertex degrees, respectively.

Furthermore, spectral clustering can be explained from a random-walk perspective. A random walk on a graph is a stochastic process that randomly decides which edge to follow whilst walking from vertex to vertex. Clustering can then be seen as finding a partition of the graph such that a random walk will stay in the same cluster as long as possible. The connection between mincut and random walk can be explained by the fact that a partition with a low cut has a reduced probability for random walks between the neighboring clusters. See Meila and Shi [64] for the formal equivalence of Ncut and the transition probabilities of a random walk.

Spectral clustering has also been shown to be equivalent to non-negative matrix factorization [65]. In non-negative matrix factorization a data matrix $X \in \mathbb{R}_+^{d \times N}$ is factorized into two non-negative matrices $F \in \mathbb{R}_+^{d \times k}$ and $G \in \mathbb{R}_+^{N \times k}$ for a given k such that $||X–FG^t||$ is minimal [66]. Non-negative matrix factorization methods have been shown to perform well in high-dimensional machine learning tasks [33, 67, 68].

Frey and Dueck [69] recently introduced a cluster algorithm which they call affinity propagation. Data points are modeled as nodes in a network recursively transmitting messages along edges of the network. The process follows a message-passing paradigm [70], that is data items negotiate their cluster and centroid preferences derived from proximity measurements to minimize an appropriately chosen cost function. The messages reflect the affinity of a data point for choosing data points as its cluster centroid. This controls the accumulation of data points around a centroid by taking into account feedback messages encoding the preferences of the points for their centroid.

9.3.2.6 Biclustering

The methods from the previous sections always use the whole feature space for the cluster analysis. Furthermore, the resulting clusters are exclusive and exhaustive, that is, there are no overlaps between two or more clusters and all samples have to be assigned to a specific cluster. When analyzing high-dimensional data one may wish to relax these restrictions. For instance, in an experiment only a small subset of features (genes) is known to be shared by a group of samples or only a small subset of samples may be of interest.

This leads to the concept of biclustering (also bi-dimensional clustering, simultaneous clustering, co-clustering, block clustering, conjugate clustering, distributional clustering, information bottleneck method, subspace clustering) [31, 39, 40, 71]. A bicluster is defined as a *block* built by a subset of genes and a subset of samples. A collection of biclusters extracted from a gene expression matrix characterizes different aspects of the data set at once where each bicluster may imply a subtype of phenotype–genotype relation. The problem in biclustering is to extract a meaningful set of biclusters out of the huge number of possible biclusters. Biclustering can be achieved by performing a two-way clustering, that is two independent cluster analyses by rows and by columns. It can also be done by simultaneous clustering in both dimensions.

Figure 9.5 illustrates an example of collections of biclusters. The data set A is given as $N \times d$ matrix where the elements a_{ij} denote the relation between sample

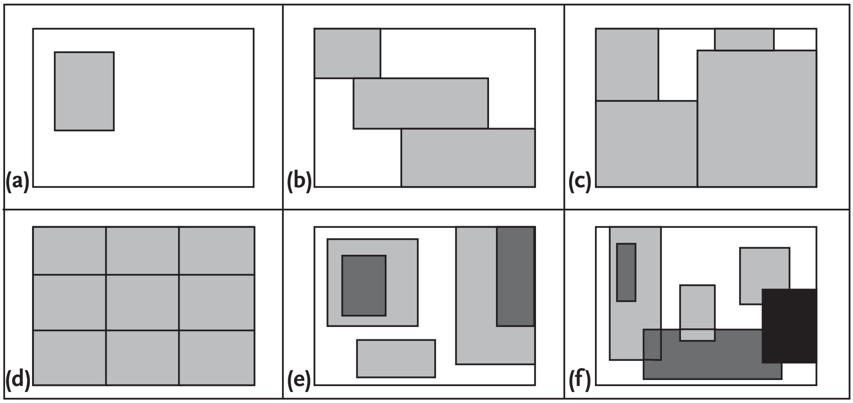


Figure 9.5 Examples of possible bicluster structures hidden in a data matrix. Each subfigure represents a data set given as a data matrix spanning rows of genes and columns of samples. Biclusters are drawn as small rectangles. The bicluster structures are (a) single bicluster, (b) exclusive row biclusters, (c) non-overlapping non-exclusive biclusters, (d) non-overlapping biclusters with checkerboard structure, (e) overlapping biclusters with hierarchical structure, (f) arbitrarily positioned overlapping biclusters.

i and feature j. A bicluster B_{RS} is a subset of rows $R \subseteq I$ and a subset of columns $S \subseteq J$ which satisfies a given characteristic of homogeneity. The simplest idea is to identify blocks with constant values [72–74]. Another approach looks for constant values along rows or columns of the data matrix [75–78]. More sophisticated methods search for biclusters with coherent values, where each row and column can be obtained by adding or multiplying a constant value to the others [32,79–82]. Figure 9.5 summarizes some possible bicluster structures in a data matrix.

The simplest structure to search for is one bicluster spanning some of the rows and some of the columns (a). Exclusive row biclusters (b) or the equivalent exclusive column biclusters present structures where every row or every column in the data matrix belongs exclusively to one of the k biclusters. Nonoverlapping nonexclusive biclusters (c) combine row and column biclusters and enable a relaxation to the constraint that rows or columns must belong exclusively to one bicluster. The checkerboard structure (d) restricts the arrangement of rows and columns as they all have to belong to exactly k biclusters. Overlapping biclusters can be explained by a hierarchical structure (e). This forms biclusters which are either disjoint or include one another. The most general biclusters are the arbitrarily positioned overlapping biclusters (f). These overlapping, nonexclusive and nonexhaustive biclusters are not restricted to form a hierarchy.

Lazzeroni *et al.* [82] proposed a method called the *plaid model* to handle most of these structures. In the plaid model the value of an element of the data matrix is defined as a sum of terms called layers, that is the data matrix is described by a linear function of layers corresponding to its biclusters:

$$a_{ij} = \sum_{l=0}^{k} \theta_{ijl} \varrho_{il} \kappa_{jl} \qquad (9.18)$$

with the weighting parameter θ. The binary indicator ϱ and κ denote the membership of row i and column j to the bicluster k. Then the plaid model is used to minimize the following objective function:

$$F(\varrho, \kappa) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{d} \left(a_{ij} - \theta_{ij0} - \sum_{l=0}^{k} \theta_{ijl} \varrho_{il} \kappa_{jl} \right)^2 \tag{9.19}$$

with an additional term θ_{ij0} that considers the existence of a single bicluster covering the whole data matrix.

9.3.3 Cluster Validation

Cluster validation provides methods to evaluate the results of cluster analysis in an objective way. The validation methods always measure a selected point of view, that is one has to predefine an expected characteristic of the data. Special attention must be turned on choosing a meaningful validation method. Clustering can be analyzed by evaluating the unusualness of the results in contrast to a baseline (random clustering). Several validation methods have been proposed. Following Jain et al. [1] they are grouped into three types of criteria:

Internal criteria. Measure the overlap between cluster structure and information inherent in data, for example silhouette, inter-cluster similarity.

External criteria. Compare different partitions, for example Rand-Index, Jaccard-Index, Fowlkes and Mallows.

Relative criteria. Decide which of two structures is better in some sense, for example, quantifying the difference between single-linkage or complete-linkage.

To demonstrate the superiority of cluster algorithms they are often applied on *a priori* labeled data sets and compared by an external criterion. In the following we introduce the basic concept of evaluating cluster results using external cluster indices. An external index describes to which degree two partitions of N objects agree. Given a set of N objects $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ and two different partitions $P = \{C_1, C_2, \ldots, C_r\}$ and $Q = \{D_1, D_2, \ldots, D_s\}$ into r and s clusters, respectively. The contingency table comparing these two partitions is given in Table 9.2. Here n_{ij} denotes the number of objects that are both in clusters C_i and D_j, $n_{i.}$ and n_j denote the total number of objects in cluster C_i and D_j respectively. Then a pair of objects is called concordant if they belong to the same cluster in both partitions or if they do not belong to the same cluster in both partitions.

Hubert and Arabie [83] define some indices based on the contingency table of two partitions, see Table 9.3.

Additionally, we have defined a measure of pairwise similarity between set partitions not based on the contingency table and counting each data point only once, in contrast to the Jaccard index, that can be interpreted as the mean proportion of

Table 9.2 Contingency table for comparing two partitions $P = \{C_1, C_2, \ldots, C_r\}$ and $Q = \{D_1, D_2, \ldots, D_s\}$. Here n_{ij} is the number of objects that are in both clusters C_i and D_j.

Q P	D_1	D_2	...	D_{r-1}	D_r	sums
C_1	n_{11}	n_{12}	$\ldots$	n_{1r-1}	n_{1r}	$n_{1.}$
C_2	n_{21}	n_{22}	$\ldots$	n_{2r-1}	n_{2r}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
C_s	n_{s1}	n_{s2}	$\ldots$	n_{sr-1}	n_{sr}	$n_{s.}$
sums	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.r-1}$	$n_{.r}$	N

Table 9.3 External cluster indices describe the agreement of two cluster results. All of the following indices can be derived from the contingency table of two partitions, see Table 9.2.

Name	Formula
Rand [84]	$1 + (\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^2 - (\sum_{i=1}^{r} n_{i.}^2 + \sum_{j=1}^{s} n_{.j}^2)/2)/\binom{n}{2}$
Jaccard [85]	$(\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^2 - n)/(\sum_{i=1}^{r} n_{i.}^2 + \sum_{j=1}^{s} n_{.j}^2 - \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^2 - n)$
Fowlkes and Mallows [86]	$(\sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}^2 - n)/(2(\sum_{i=1}^{r} \binom{n_{i.}}{2} \sum_{j=1}^{s} \binom{n_{.j}}{2}))^{1/2})$

samples consistent over different clusterings under the restriction $k = r = s$. The *MCA index* η is defined as the fraction of the number of data points in the intersection sets of the corresponding clusters to the overall number of data points.

$$\eta(P, Q) = \frac{1}{N} \max_{\pi} \sum_{i=1}^{k} |C_i \cap D_{\pi(i)}| \tag{9.20}$$

This score is based on solving the *linear assignment problem* (LAP) for the intersections between different partitions.

Let Π_k be the set of all permutations of the numbers $1 \ldots k$. Given a symmetric matrix $\boldsymbol{B} \in \mathbb{R}_+^{N \times N}$ with elements $b_{ij} \geq 0$ (representing the similarity between object i and j) try to find the permutation $\boldsymbol{\pi} \in \Pi_k$ that maximizes (or minimizes) the term $\sum_{i=1}^{k} b_{i\pi_i}$. This problem can be solved in $O(k^3)$ using the algorithm by Jonker and Volgenant [87].

In this investigation we applied the proposed robustness evaluation measure to three different prototype-based algorithms (self-organizing feature maps (SOM), k-means- and fuzzy-c-means clustering) and to CGH data, indicating loss and gain of genomic material, from Mattfeldt *et al.* [88] and gene expression profiles of leukemia patients from Golub *et al.* [89], see Figure 9.6.

The two-class scenario of the leukemia data is confirmed in the two-cluster result, see Figure 9.7. To correctly judge the results found the performance of a random

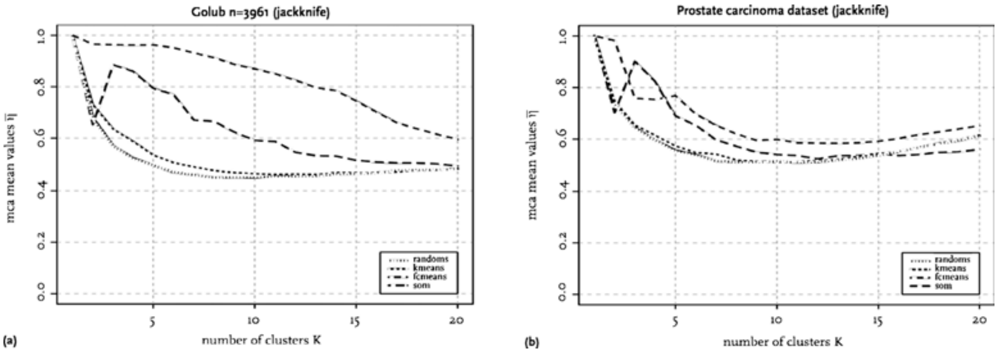


Figure 9.6 The graphs show the averaged robustness index (MCA) over a varying number of cluster centroids for the different cluster algorithms and the data sets. (a) the leukemia data from Golub *et al*. [89], (b) the prostate cancer data from Mattfeldt *et al*. [88]. In the presented results a jackknife method was used: for each run $d = \sqrt{n}$ elements were left out from the sample set, cluster and assign the remaining samples according to the nearest neighbor rule after the clustering is performed.

cluster algorithm is included as a base-line. This suggests, for the leukemia data two clusters and for the prostate carcinoma data, three clusters, as the difference of the MCA values of the fuzzy-c-means (leukemia) and the self-organizing feature map (prostate) to the respective random clusterings is maximal in these cases. Further analysis of the CGH data assigned to the three clusters of the prostate carcinoma cases showed a differentiation into different grades of malignancy based on their CGH profile.

9.4 Semi-Supervised Clustering

Unsupervised clustering methods do not make use of background knowledge about a partition of the data. Several modifications were proposed during recent years, to incorporate additional knowledge in cluster-analysis algorithms. Background knowledge is usually provided as a set of labeled data items [10] or a set of constraints between pairs of data items [90]. As we will see below in Results, the spare additional information seems to assist the algorithm in building more appropriate partitions.

In recent years various partitional cluster algorithms were adapted to make use of this kind of background information either by constraining the search process or by modifying the underlying metric. It has been shown that including background knowledge might improve the accuracy of cluster results, that is the computed clusters better match a given classification of the data [9–15]. Basu *et al*. [91] proposed a probabilistic framework for semi-supervised clustering based on Hidden Markov Random Fields. Kulis *et al*. [92] demonstrated how to generalize this model to optimize a number of different graph-clustering objectives like ratio cut or normalized cut described by Chan *et al*. [93] and Shi and Malik [59], respectively. Recently Yan

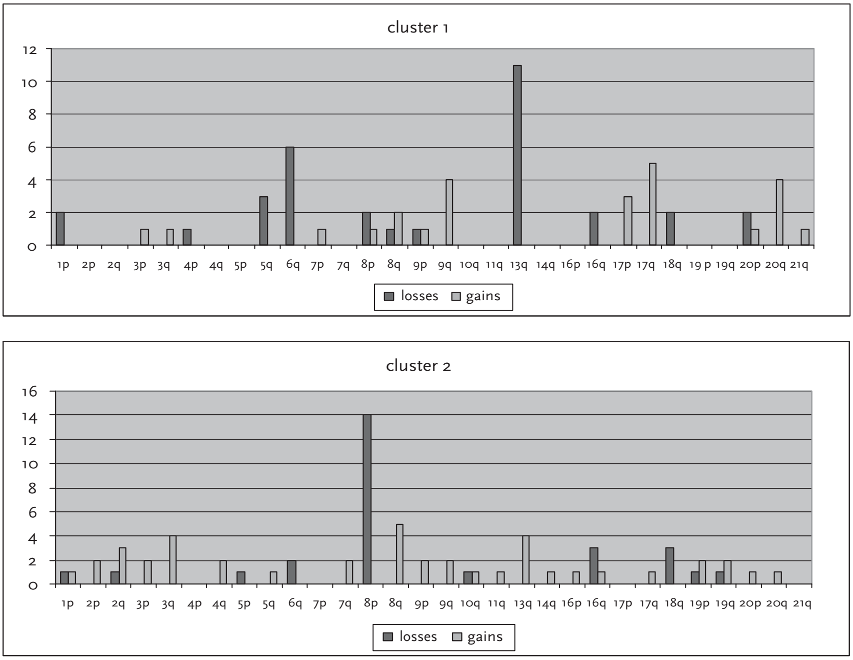


Figure 9.7 Losses and gains of the three-cluster SOM solution from Figure 9.6 are given. Cluster 1 contained 15 samples with a mean Gleason score of 5.8 and a mean WHO grading of 1.9 (Gleason score and WHO grading describe tumor malignancy). In cluster 1 there was 1 subject with tumor progression and 3 without progression. Cluster 2 contained 16 samples with a mean Gleason score of 7.3 and a mean WHO grading of 2.4. In cluster 2 there were 8 subjects with tumor progression and 1 without progression. Cluster 3 contained 29 data points with no gains and no losses (not shown). It contained one progression and 6 cases without progression. The mean Gleason score was 5.6 and the WHO score 1.9.

and Domeniconi [94] proposed an extension of this partitional-clustering method based on an adaptive kernel method. Davidson *et al.* [95] proposed a method to include background knowledge in agglomerative hierarchical clustering. Kestler *et al.* [16] analyzed the effects of constraints on hierarchical clustering and identified important limitations in the use of background knowledge on hierarchical clustering.

As hierarchical clustering algorithms can predict a basic branching structure they are often used in the context of microarray data mining [28]. Limitations in the reproducibility of clustering results after small modifications of the data set motivate the inclusion of background knowledge into the hierarchical clustering process. Figure 9.8 shows an example of this behavior. Analyzing a data set of pancreatic ductal adenocarcinoma (DAC) and normal pancreas (DUKTI) using an unsupervised hierarchical clustering results in a well-defined disjunction of cancer and normal samples. By removing only one data item (DAC.8), a clearly changed

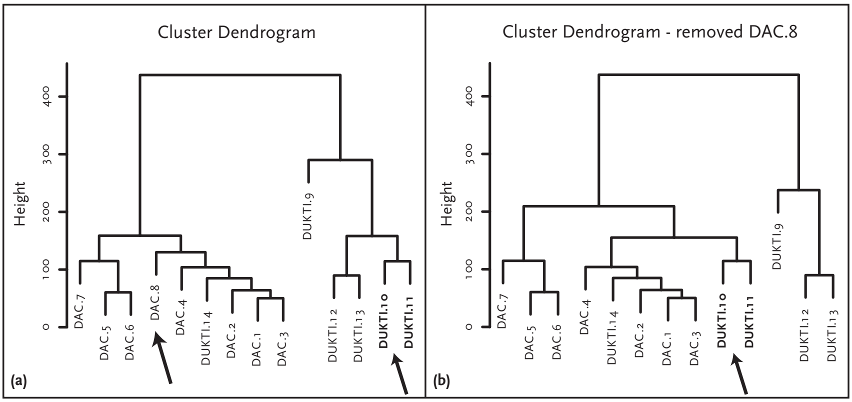


Figure 9.8 Example for the instability problem in hierarchical cluster analysis. The data is taken from expression profiles of pancreatic ductal adenocarcinoma (DAC) and normal pancreas (DUKTI), see [96]. (a) A dendrogram resulting from a hierarchical cluster analysis of the data set is shown. By removing one sample (DAC.8) the structure of the resulting dendrogram is clearly changed. (b) The resulting dendrogram is shown. DUKTI.10 and DUKTI.11 changed their position such that an even division into two clusters contradicts the left clustering.

dendrogram is built where two samples (DUKTI.10 and DUKTI.11) are supposed to be in the tumor class.

9.4.1 Modeling Background Knowledge

In microarray data analysis, background knowledge about the grouping of the data based on additional information may be available. For instance, an expert assumes a grouping of some samples after an analysis of cell images. This kind of background knowledge can be modeled in different ways; for example, as relations between two data items or as labels for the appointed items. The use of labels may be inferior, as they mostly indicate class memberships. Because a class can consist of different and distant clusters, a distance-based clustering method may be misled by the labeled information. Wagstaff *et al.* [90] suggested using cannot-link and must-link constraints instead. Known relationships between a pair of data points (x, y) are encoded as must-links $ml(x, y)$ to indicate two data items being arranged in one cluster and as cannot-links $cl(x, y)$ to indicate not to assign two data items to the same group, that is $ml(x, y) \Leftrightarrow x, y \in C_i$ and $cl(x, y) \Leftrightarrow x \in C_i, y \in C_j,\ l \neq k$. Additionally the set of must-link and cannot-link constraints can be extended using an approach to build the combined transitive closure introduced by Wagstaff *et al.* [9]. Figure 9.9 illustrates must-link and cannot-link relations between data points.

The use of background knowledge needs a modification of the validation methods. One has to ensure that the, correctly processed, *a priori* knowledge is not rated as an achievement of the cluster algorithm. The Constrained Rand Index serves as an example of a modified validation measure. The correction for constrained data is

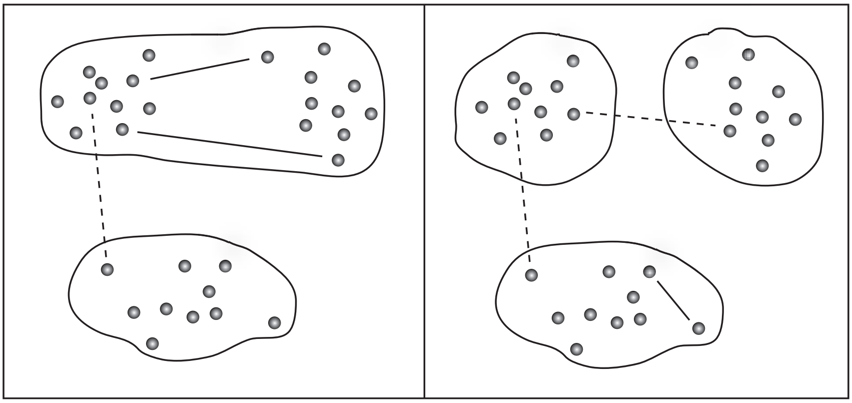


Figure 9.9 Example for must-link (solid lines) and cannot-link (dashed lines) constraints. On the left, constraints are included which refer to a clustering into two clusters. The must-links in the top cluster indicate that the pairs of points must stay in the same cluster. The cannot-link between the cluster points out that the corresponding points must not be clustered together. As it can be seen from the right subfigure a desired partition into three clusters must have another set of constraints; for example, the must-links are removed or replaced by cannot-links.

done by subtracting the number of constrained object pairs con_{ij} in both partitions from the entries in the contingency table, that is $\tilde{n}_{ij} = n_{ij} - \mathrm{con}_{ij}$, $\forall\, i = \{1, \ldots, r\}$, $j = \{1, \ldots, s\}$:

$$\mathrm{CRI} = 1 + \frac{\sum_{i=1}^{r} \sum_{j=1}^{s} \tilde{n}_{ij}^2 - (\sum_{i=1}^{r} \tilde{n}_{i.}^2 + \sum_{j=1}^{s} \tilde{n}_{.j}^2)/2}{\binom{\tilde{n}}{2}} \tag{9.21}$$

9.4.2 Results

In Kestler *et al.* [16] we proposed a semi-supervised divisive hierarchical clustering algorithm and presented results indicating an enhancement of the cluster stability. In the following we give the results of analyzing data from a study on the small round blue cell tumors of childhood [97]. It contains neuroblastoma samples (NB), rhabdomyosarcoma samples (RMS), non-Hodgkin lymphoma samples (NHL) and samples from the Ewing family of tumors (EWS). Gene expression data from glass cDNA microarrays containing 6567 genes were used.

To evaluate the stability of the clustering method several subsets of data are processed and the computed dendrograms are compared to one another. Because we are mostly interested in the top of the dendrograms we decided not to measure the overall agreement of two computed dendrograms, but only the agreement on the highest levels. We modified the data sets as described in the following. At first we built 10 different test sets by randomly leaving out ten times 5% the data items. Then we added for each test set 20 different sets of constraints, that is ten times 1% and 10% of the data items were constrained by a randomly chosen must-link or cannot-link according to the predefined labeling of the data. To evaluate the stabili-

Table 9.4 Improvements in stability of clustering microarray data after randomly removing and constraining data items. The table displays a summary of the Constrained Rand Index values from different runs. The percentage of held out and constrained data items is denoted with cut and constr.

Data set	min	1st qu.	median	mean	3rd qu.	max
Childhood cancer						
5% cut, unconstrained	0.49	0.50	0.65	0.63	0.76	0.84
5% cut, 1% constr.	0.49	0.51	0.88	0.76	1.00	1.00
5% cut, 10% constr.	0.48	0.50	0.87	0.76	1.00	1.00

ty we then compared each of the test sets with the same number of hold-out items and constraints. The more often a high agreement value is computed, the more stable are the partitions. The results from the cluster analyses (Constrained Rand Index) can be seen in Table 9.4.

Our results indicate that the maximum stability score may not be continuously improved by adding more background knowledge. In fact there seems to be a sudden phase transition between unstable and stable results leading to a saturation close to a value of 1. But the number of constraints needed to reach the phase transition may not be easy to predict as it seems to depend on the complexity of the data set and the chosen hold-out values. Constraints can also affect the cluster analysis adversely [98]. By randomly building constraint sets, some of the chosen relations guide the algorithm to a bigger variety of dendrograms, that is, a lower stability value is more often computed when increasing the number of constraints.

9.4.3 Challenges

Semi-supervised clustering has become an active field of research. It assists the unsupervised data-mining tools in generating most valuable hypotheses about the underlying data. Up to now many different cluster algorithms have been modified to enable semi-supervised clustering. Several authors demonstrated that the integration of background knowledge is able to produce more accurate [9–15] and robust cluster results [16]. However, there are still many important questions on the application of semi-supervised clustering. For instance, in some cases, constraining has an adverse impact on the performance independent of the correctness of the background knowledge [99]. A ranking of constraints based on a measure of applicability may help to reject misleading constraints.

9.5 Summary

As seen from the previous sections, the different cluster algorithms are motivated from different perspectives. Traditionally cluster algorithms are divided into partitional and hierarchical methods. Some authors prefer a distinction into model-based and heuristic methods. It is also possible to distinguish the goals of cluster algorithms into those that build compact clusters and those that generate connected clusters. Many other classifications of cluster algorithms seem possible according to the different aims in clustering data. Depending on the application, requirements, and the data distribution, each of the methods mentioned in this chapter may turn out to be superior to the others. However, sometimes claims concerning the overall superiority of a method are made. In that regard, Frey and Dueck [69] claim that their algorithm, is able to find clusters with much lower error than other methods and in less time. But in fact their algorithm, like all the others has advantages and disadvantages. Like k-means, affinity propagation uses an EM-like method to compute cluster association and centroid updates. In contrast, no fixed number of centroids is updated, but all data items are considered as being possible centroids. Like k-means, affinity propagation favors regularly shaped clusters surrounding their cluster centroids, strongly dependent on the chosen similarity measure. Also, all information on a possible hierarchical structure is lost. The algorithm is said to identify the number of hidden clusters itself, but on the other hand the search for a finer or coarser clustering often results in trial and error. Even the mentioned advantage in time can be adversely affected, as the construction of the input similarity matrix will be more time consuming when larger data sets are analyzed. The authors can report a massive time reduction compared to k-means as they re-ran k-means with different initial centroids several thousands of times. In fact a single run of k-means consumes a fraction of the runtime of affinity propagation and there are also different methods to loosen the requirement for repeated initializations of k-means [100] and variations that can change the number of centers, such as Isodata [101]. Nevertheless the authors proposed another useful tool which has been shown to provide satisfying clustering results.

Closely predefining the circumstances of the data analysis may guide the user to choose an appropriate cluster algorithm. Useful *a priori* knowledge may constrain the number of clusters, the need for hierarchical output or the eventuality of coping with outliers. The more complex the methods that are used, the more preparatory work must be invested. For example, in model-based clustering one has to check the assumption of whether the data is consistent with the model, such as whether it has been sampled from a mixture of Gaussians and which covariance matrix may best fit the data's geometric features.

Another point is the interpretability of the cluster results. Dendrograms, as computed by a hierarchical cluster analysis, offer many alternatives of defining different data partitions, for example cutting edges along/between branches. Biclustering often results in the presentation of a huge number of (possibly interleaving) clusters which may then have to be individually checked by the user for relevance. In addition one may favor an inappropriate validation measure wrongly emphasising the significance of the results. For instance, evaluating k-means using a measurement of inter-cluster similarity overrates the results as the clusters have just been optimized to fit this characteristic.

An important issue is the robustness of the cluster method. Cluster robustness can be measured by repeating a cluster procedure with fixed parameters, but under varying initial conditions. A cluster solution is then called robust when the fluctuations in the cluster assignments of the data points among the different runs are small. In Section 9.3.3 we introduced a cluster robustness validation method based on the pairwise similarity between partitions. This measure can be interpreted as the mean proportion of samples being consistent over different clusterings. A base-line is generated by measuring the performance of repeated random clusterings. The pairwise similarity measures are then averaged over all pairings and compared to the base-line. The most interesting cluster results are those with a high robustness value and a big distance to the base-line. In this way an estimation of the number of clusters k present in the data set can be assessed. A good value for k leads to more robust cluster results. Repeated clustering for various numbers of clusters, and evaluating the robustness with respect to the base-line, can give evidence for the true number of clusters.

In general, repeated clustering is associated with the possibility of ending up with contradictory results, see Section 9.4. Many algorithms are highly susceptible to minor changes in the clustering procedure. There are different sources which may affect the robustness of cluster results, e.g. different initializations, varying parameters, noisy data sets or different sub-samples of data. Some of these effects can be avoided in practice. For instance k-means is often re-run with different initial centroids to overcome, with its sensitivity for initialization. Moreover, many algorithms are extendable to handle noisy data. But the sensitivity of cluster methods to minor variations in the size of the data set, especially in settings with low cardinality and high dimensionality, is often neglected. Especially in clustering microarray data this aspect must be attentively investigated. As we have demonstrated in Section 9.4, the robustness of clustering solutions can be improved by incorporating additional background knowledge. Therefore, semi-supervised clustering may also assist the above mentioned resampling procedure for estimating the number of clusters.

Generally, clustering is a method for deriving hypotheses from data via grouping. Additional information can guide this process, stabilize the groups formed and substantiate the proposition.

References

1 Jain, A.K. and Dubes, R.C. (**1988**) *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

2 Kaufman, L. and Rousseeuw, P.J. (**1990**) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

3 Jain, A.K., Murty, M.N. and Flynn, P.J. (**1999**) Data clustering: A review. *ACM Computing Surveys*, **31**(3), 264–323.

4 Anderberg M.R. (**1973**) *Cluster Analysis for Applications*. Academic Press, London.

5 Everitt, B.S., Landau, S. and Leese, M. (**2001**) *Cluster Analysis*, 4th edn Oxford University Press Inc., New York.

6 Hartigan J.A. (**1975**) *Clustering Algorithms*. John Wiley & Sons, New York.

7 Berkhin P (**2002**) Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose.

8 Boutros, P.C. and Okey, A. (**2005**) Unsupervised pattern recognition: An introduction to the whys and wherefores of clustering microarray data. *Briefings in Bioinformatics*, **6**(4), 331–343.

9 Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S. (**2001**) Constrained k-means clustering with background knowledge. In Brodley, C.E. and Danyluk, A.P. (eds), *Proceedings of 18th International Conference on Machine Learning*, San Francisco, July 2001. Morgan Kaufmann, 577–584.

10 Basu, S., Banerjee, A. and Mooney, R.J. (**2002**) Semi-supervised clustering by seeding, in *Proceedings of 19th International Conference on Machine Learning*, (eds S. Claude and A.G. Hoffmann), San Francisco, July 2002. Morgan Kaufmann, 19–26.

11 Cohn, D., Caruana, R. and McCallum, A. (**2003**) Semi-supervised clustering with user feedback. Technical Report 1892, Cornell University, Ithaka.

12 Xing, E.P., Ng, A.Y., Jordan, M.I. and Russell, S. (**2003**) Distance metric learning, with application to clustering with side-information, in *Advances in Neural Information Processing Systems 15*, (eds S. Becker, S. Thrun and K. Obermayer). MIT Press, Cambridge, October 2003, 505–512.

13 Bar-Hillel, A., Hertz, T., Shental, N. and Weinshall, D. (**2003**) Learning distance functions using equivalence relations, in *Proceedings of 20th International Conference on Machine Learning*, (eds T. Fawcett and N. Mishra). Washington, August 2003. AAAI Press, 11–18.

14 Bilenko, M. and Mooney, R.J. (**2003**) Adaptive duplicate detection using learnable string similarity measures, in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (eds L. Getoor, T.E. Senator, P. Domingos and C. Faloutsos). New York, August 2003. ACM Press, 39–48.

15 Bilenko, M., Basu, S. and Mooney, R.J. (**2004**) Integrating constraints and metric learning in semi-supervised clustering, in *Proceedings of the 21st International Conference on Machine Learning*, (ed C.E. Brodley). New York, July 2004. ACM Press, 81–88.

16 Kestler, H.A., Kraus, J.M., Palm, G. and Schwenker, F. (**2006**) On the effects of constraints in semi-supervised hierarchical clustering, in *Artificial Neural Networks in Pattern Recognition*, (eds F. Schwenker and S. Marinai) volume 4087 of *Lecture Notes in Artificial Intelligence*, Berlin, September 2006. Springer, 57–66.

17 Hamilton, A. and Baulcombe, D. (**1999**) A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, **286**(5441), 950–952.

18 Ruvkun, G. (**2001**) Molecular biology. glimpses of a tiny rna world. *Science*, **294**(5543), 797–799.

19 Lander, E.S. (**1999**) Array of hope. *Nature Genetics (Supplement)*, **21**(1), 3–4.

20 Marx, J. (**2000**) Dna arrays reveal cancer in its many forms. *Science*, **289**(5485), 1670–1672.

21 Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A. and McGuire, W.L. (**1987**) Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science*, **235**(4785), 177–182.

22 Cho, H., Mason, K., Ramyar, K.X., Stanley, A.M., Gabelli, S.B., Denney Jr., D.W. and Leahy, D.J. (**2003**) Structure of the extracellular region of her2 alone and in complex with the herceptin fab. *Nature*, **421**(6924), 756–760.

23 Crick, F.H. (**1958**) On protein synthesis. *Symposia of the Society for Experimental Biology*, **12**, 138–163.

24 Allison, D.B., X. Cui, Page, G.P. and M. Sabripour (**2006**) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**(1), 55–66.

25 Lockhart, D.J. and Winzeler, E.A. (**2000**) Genomics, gene expression and dna arrays. *Nature*, **405**, 827–836.

26 Leung, Y.F. and Cavalieri, D. (**2003**) Fundamentals of cdna microarray data analysis. *Trends in Genetics*, **19**(11), 649–659.

27 Steinhoff, C. and Vingron, M. (**2006**) Normalization and quantification of differential expression in gene expression microarrays. *Briefings in Bioinformatics*, **7**(2), 166–177.

28 Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (**1998**) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, **95**(25), 14863–14868.

29 Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (**1999**) Systematic determination of genetic network architecture. *Nature Genetetics*, **22**(3), 281–285.

30 Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (**1999**) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, **96**(6), 2907–2912.

31 Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. (**1998**) Automatic subspace clustering of high dimensional data for data mining applications, in *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, New York, 1998. ACM Press, 94–105.

32 Cheng, Y. and Church, G.M. (**2000**) Biclustering of expression data, in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, (eds R. Altman, T.L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck and H. Weissig). Menlo Park, 2000. AAAI Press, 93–103.

33 Brunet, J.-P., Tamayo, P., Golub, T.R. and Mesirov, J.P. (**2004**) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, **101**(12), 4164–4169.

34 Bellman, R. (**1961**) *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton.

35 R Development Core Team (**2006**) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2006. ISBN 3-900051-07-0.

36 Chiaretti, S., Li, X., Gentleman, R., Vitale, A., Vignetti, M., Mandelli, F., Ritz, J. and Foa, R. (**2004**) Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, **103**(7), 2771–2778.

37 Kestler, H.A., Müller, A., Buchholz, M., Gress, T.M. and Palm, G. (**2006**) A perceptually optimized scheme for visualizing gene expression ratios with confidence values, in *Perception and Interactive Technologies, International Tutorial and Research Workshop*, (eds E. André, L. Dybkjær, W. Minker, H. Neumann and M. Weber) volume 4021 of *Lecture Notes in Computer Science*, Berlin, 2006. Springer, 73–84.

38 Jiang, D., Tang, C. and Zhang, A. (**2004**) Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, **16**(11), 1370–1386.

39 Madeira, S.C. and Oliveira, A.L. (**2004**) Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **1**(1), 24–45.

40 Parsons, L., Haque, E. and Liu, H. (**2004**) Subspace clustering for high dimensional data: a review. *SIGKDD Explorations Newsletter*, **6**(1), 90–105.

41 MacQueen, J.B. (**1967**) Some methods for classification and analysis of multivariate observations, in *Proceedings of the 5th Berkeley Symposium on Math, Statistics and Probability*, (eds J. Neyman and L. Le Cam) volume 1, Berkely, 1967. University of California Press, 281–297.

42 Selim, S.Z. and Ismail, M.A. (**1984**) K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**(1), 81–87.

43 Bezdek, J.C. (**1984**) Fcm: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, **10**, 191–203.

44 Kohonen, T. (**2001**) *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences* 3rd edn, Springer, Berlin.

45 Zhong, S. and Ghosh, J. (**2003**) A unified framework for model-based clustering. *Journal of Machine Learning Research*, **4**(11), 1001–1037.

46 Yeung, K., Fraley, C., Murua, A., Raftery, A. and Ruzzo, W. (**2001**) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**(10), 977–987.

47 McLachlan, G.J., Bean, R.W. and Peel, D. (**2002**) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413–422.

48 Ghosh, D. and Chinnaiyan, A.M. (**2002**) Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, **18**(2), 275–286.

49 Fraley, C. and Raftery, A.E. (**1998**) How many clusters? which clustering method? answers via model-based analysis. *The Computer Journal*, **41**(8), 578–588.

50 Banfield, J.D. and Raftery, A.E. (**1993**). Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**(3), 803–821.

51 Meila, M. and Heckerman, D. (**2001**) An experimental comparison of model-based clustering methods. *Machine Learning*, **42**(1/2), 9–29.

52 Dempster, A.P., Laird, N.M. and Rubin, D.B. (**1977**) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, **39**(1), 1–38.

53 Donath, W.E. and Hoffman, A.J. (**1973**) Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, **17**(5), 420–425.

54 Dhillon, I.S., Guan, Y. and Kulis, B. (**2005**) A unified view of kernel k-means, spectral clustering and graph cuts. Technical Report TR-04-25, The University of Texas at Austin, Department of Computer Sciences, 2005.

55 von Luxburg, U. (**2006**) A tutorial on spectral clustering. Technical Report TR-149, Max Plank Institute for Biological Cybernetics.

56 Chapelle, O. Schölkopf, B. and Zien, A. (eds) (**2006**) *Semi-Supervised Learning*. MIT Press, Cambridge.

57 Mohar B. (**1991**) The laplacian spectrum of graphs, in *Graph Theory, Combinatorics, and Applications*, (eds Y. Alavi, G. Chartrand, O.R. Oellermann and A.J. Schwenk) volume 2, John Wiley & Sohns, New York, 871–898.

58 Fiedler M. (**1973**) Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, **23**(98), 298–305.

59 Shi, J. and Malik, J. (**2000**) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

60 Ng, A., Jordan, M. and Weiss, Y. (**2002**) On spectral clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems 14*, (eds T.G. Dietterich, S. Becker and Z. Ghahramani). MIT Press, Cambridge, 849–856.

61 Dhillon, I.S. (**2001**) Co-clustering documents and words using bipartite spectral graph partitioning, in *Proceedings of the 7th ACM SIKDD international conferencec on Knowledge Discovery and Data Mining*. ACM Press, New York, 269–274.

62 Ding, C.H.Q., He, X., Zha, H., Gu, M. and Simon, H.D. (**2001**) A min-max cut algorithm for graph partitioning and data clustering, in *Proceedings of the 2001 IEEE International Conference on Data Mining*, (eds N. Cercone, T.Y. Lin and X. Wu). Los Alamitos, IEEE Computer Society, 107–114.

63 Hagen, L.W. and Kahng, A.B. (**1992**) New spectral methods for ratio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **11**(9), 1074–1085.

64 Meila, M. and Shi J. (**2001**) A random walks view of spectral segmentation, in *Artificial Intelligence and Statistics 2001: Proceedings of the Eighth Internatinoal Workshop*, (eds T.S. Jaakkola and T.S. Richardson). Elsevier, Oxford, 177–186.

65 Ding, C.H.Q. and He X. (**2005**) On the equivalence of nonnegative matrix factorization and spectral clustering, in *Proceedings of the 5th SIAM international conference on data mining*, (eds H. Kargupta, C. Kamath, J. Srivastava and A. Goodman). Philadelphia, 2005. Society for Industrial and Applied Mathematics, 606–610.

66 Lee, D.D. and Seung, H.S. (**2001**) Algorithms for non-negative matrix factorization, in *Advances in Neural Information Processing Systems 13*, (eds T.K. Leen, T.G. Dietterich and V. Tresp). MIT Press, Cambridge, 556–562.

67 Lee, D.D. and Seung, H.S. (**1999**) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791.

68 Kim, P.M. and Tidor, B. (**2003**) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research*, **13**(7), 1706–1718.

69 Frey, B.J. and Dueck, D. (**2007**) Clustering by passing messages between data points. *Science*, **315**(5811), 972–976.

70 Kschischang, F.R., Frey, B.J. and Loeliger, H.-A. (**2001**) Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, **47**(2), 498–519.

71 Tanay, A., Sharan, R. and Shamir, R. (**2006**) Biclustering algorithms: A survey, in *Handbook of Computational Molecular Biology*, (ed A. Srinivas), chapter 26. Chapman & Hall/CRC, London.

72 Hartigan, J.A. (**1972**) Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67**(337), 123–129.

73 Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D. and Brown, P. (**1999**) Clustering methods for the analysis of dna microarray data. Technical report, Stanford University.

74 Busygin, S., Jacobsen, G. and Kramer, E. (**2002**) Double conjugated clustering applied on leukemia microarray data, in *Proceedings of the 2nd SIAM International Conference on Data Mining, Workshop on Clustering High Dimensional Data*, 2002.

75 Getz, G., Levine, E. and Domany, E. (**2000**) Coupled two-way clustering analysis of gene microarray data. *Proceedings of the Natural Academy of Sciences*, **97**(22), 12079–12084.

76 Califano, A., Stolovitzky, G. and Tu, Y. (**2000**) Analysis of gene expression microarrays for phenotype classification, in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, (eds R. Altman, T.L. Bailey, P. Bourne, M. Gribskov, T. Lengauer, I.N. Shindyalov, L.F. Ten Eyck and H. Weissig). AAAI Press, Menlo Park, 75–85.

77 Segal, E., Taskar, B., Gasch, A., Friedman, N. and Koller, D. (**2001**) Rich probabilistic models for gene expression. *Bioinformatics*, **17**(1), 243–252.

78 Sheng, Q., Moreau, Y. and De Moor, B. (**2003**) Biclustering micrarray data by gibbs sampling. *Bioinformatics*, **19**(2), 196–205.

79 Tang, C., Zhang, L., Zhang, I. and Ramanathan, M. (**2001**) Interrelated two-way clustering: an unsupervised approach for gene expression data analysis, in *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, (ed N.G. Bourbakis). Washington, 2001. IEEE Computer Society, 41–48.

80 Yang, J., Wang, H., Wang, W. and Yu, P.S. (**2003**) Enhanced biclustering on expression data, in *3rd IEEE International Symposium on BioInformatics and BioEngineering*, (eds H. Jamil, and V. Magalooikonomou). Washington, 2003. IEEE Computer Society, 321–327.

81 Klugar, Y., Basri, R., Chang, J.T. and Gerstein, M. (**2003**) Spectral biclustering of microarray data: coclustering genes and conditions. *Genome Research*, **13**(4), 703–716.

82 Lazzeroni, L.C. and Owen, A. (**2002**) Plaid models for gene expression data. *Statistica Sinica*, **12**(1), 61–86.

83 Hubert, L.J. and Arabie, P. (**1985**) Comparing partitions. *Journal of Mathematical Classification*, **2**, 193–218.

84 Rand, W.M. (**1971**) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846–850.

85 Jaccard, P. (**1908**) Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des sciences naturelles*, **44**, 223–270.

86 Fowlkes, E.B. and Mallows, C.L. (**1983**) A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569.

87 Jonker, R. and Volgenant, A. (**1987**) A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, **38**, 325–340.

88 Mattfeldt, T., Wolter, H., Kemmerling, R., Gottfried, H.-W. and Kestler, H.A. (**2001**) Cluster analysis of comparative genomic hybridization (cgh) data using self-organizing maps: Application to prostate carcinomas. *Analytical Cellular Pathology*, **23**(1), 29–37.

89 Golub, T., Slonim, D., Tamayo, P., Huard, C., Gassenbeek, M., Coller, H., Loh, M., Downing, J., Caliguri, M., Bloomfield, C. and Lander, E. (**1999**) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

90 Wagstaff, K. and Cardie, C. (**2000**) Clustering with instance-level constraints, in *Proceedings of the 17th International Conference on Machine Learning*, (ed P. Langley). Morgan Kaufmann, San Francisco, 1103–1110.

91 Basu, S., Bilenko, M. and Mooney, R. (**2004**) A probabilistic framework for semi-supervised clustering, in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (eds R. Kohavi, J. Gehrke, W. DuMouchel and J. Ghosh). ACM Press, New York, 59–68.

92 Kulis, B., Basu, S., Dhillon, I. and Mooney, R. (**2005**) Semi-supervised graph clustering: A kernel approach, in *Proceedings of the 22nd International Conference on Machine Learning*, (L. De Raed and S. Wrobel). ACM Press, New York, 457–464.

93 Chan, P.K., Schlag, M.D.F. and Zien, J.Y. (**1994**) Spectral k-way ratio-cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, **13**(9), 1088–1096.

94 Yan, B. and Domeniconi, C. (**2006**) An adaptive kernel method for semi-supervised clustering, in *European Conference on Machine Learning*, (J. Fürnkranz, T. Scheffer and M. Spiliopoulou) volume 4212 of *Lecture Notes in Computer Science*. Springer, Berlin, 521–532.

95 Davidson, I. and Ravi, S.S. (**2005**) Agglomerative hierarchical clustering with constraints: Theoretical and empirical results, in *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, (eds A. Jorge, L. Torgo, P. Brazdil, R. Camacho and J. Gama) volume 3721 of *LNCS*. Springer, Berlin, 59–70.

96 Buchholz, M., Braun, M., Heidenblut, A., Kestler, H.A., Klöppel, G., Schmiegel, W., Hahn, S.A., Lüttges, J. and Gress, T.M. (**2005**) Transcriptome analysis of microdissected pancreatic intraepithelial neoplastic lesions. *Oncogene*, **24**(44), 6626–6636.

97 Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonesco, C., Peterson, C. and Meltzer, P. (**2001**) Classification and diagnostic prediction of cancer using gene expression profiling and artificial neural networks. *Nature Medicine*, **6**(7), 673–679.

98 Davidson, I., Wagstaff, K. and Basu, S. (**2006**) Measuring constraint-set utility for partitional clustering, in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, (eds J. Fürnkranz, T. Scheffer and M. Spiliopoulou) volume 4213 of *LNCS*. Springer, Berlin, 115–125.

99 Wagstaff, K. (**2007**) Value, cost, and sharing: Open issues in constrained clustering, in *Knowledge Discovery in Inductive Databases, 5th International Workshop*, (eds S. Džeroski and J. Struyf), volume 4747 of *Lecture Notes in Computer Science*. Springer, Berlin, 1–10.

100 Khan, S.S. and Ahmad, A. (**2004**) Cluster center initialization algorithm for k-means clustering. *Pattern Recognition Letters*, **25**(11), 1293–1302.

101 Ball, G.H. and Hall, D.J. (**1966**) A clustering technique for summarizing multivariate data. *Behavioral Science*, **12**(2), 153–155.

10
Image Processing and Feature Extraction from a Perspective of Computer Vision and Physical Cosmology

Holger Stefan Janzer[1], Florian Raudies[1], Heiko Neumann, Frank Steiner

10.1
Introduction

The main conjunction between Physical Cosmology and Computer Vision are images. Commonly structures and objects in those images should be detected and recognized. In this contribution we give a short survey of methods and assumptions used in both disciplines. Applications and illustrative examples of those methods are presented for the fields of Physical Cosmology and Medical Science.

In numerous scientific disciplines and applications areas high-dimensional sensory data needs to be analyzed for the detection of complex structures or for triggering special events. From the beginning the acquisition and analysis of image data formed the subject of *image analysis*. Nowadays many research disciplines work on the analysis of multi-dimensional images, namely engineering and computer science, physics, mathematics and even psychology. Together they formed the research discipline of *Computer Vision* (or *Computational Vision*) which accounts for the interpretation of images and image sequences rather than merely the raw processing of images [1, 2]. Computer Vision aims to be an umbrella for tasks that could be classified into:

(a) low-level vision, for example, image enhancement and filtering techniques for image processing;

(b) mid-level vision, for example, segmentation, feature extraction, and the detection of so-called intrinsic scene characteristics, in particular, the relative surface orientation or depth discontinuities with respect to the viewer direction; and

(c) high-level vision to generate and obtain, for example, descriptions of three-dimensional surfaces and volumes or the linking to steer a robot through complex terrain [3, 4].

The observation of similar approaches and computational methods that have been developed in different disciplines, namely in Computer Vision and Physical Cos-

[1] Corresponding authors.

mology, have motivated the writing of this contribution. This article highlights common assumptions and methods which are used in both fields of Physical Cosmology and Image Processing/Computer Vision, but which are often not well known in other research communities. At various stages we give insights into current research, beyond the scope of the good, and usual, textbooks as image processing or Physical Cosmology. Here we give a short survey of methods and assumptions utilized for stages of basic image processing and the extraction of meaningful features. Applications and illustrative examples of those methods are presented as images from Physical Cosmology and medical imaging to highlight the broad scope of applicability. The focus of this overview is restricted to the analysis of *single* images. It would be beyond the scope to discuss approaches of multi-image processing such as in stereo vision or motion analysis. For the interested reader, we refer to standard text books such as, for example, [5].

10.1.1 **Overall View**

The chapter is organized as follows. We start in Section 10.2 with a brief outline of some definitions and architectural issues in image processing and (low-level) Computer Vision. In addition, a short summary of the background of Physical Cosmology and its relation to image processing is presented in Section 10.3. In Section 10.4 properties of images are discussed, beginning with a sketch of the generation and representation of two-dimensional (2D) images. Image representations are often generated by a projective image acquisition process and involve a proper mapping of the scenic image onto an arbitrarily shaped surface. We briefly sketch some useful transformations often used in applications. Next, several main image properties are introduced. Then a brief overview of basic image characteristics is presented, including basic quantities such as, for example, distribution and correlation measures of image intensities as well as useful spectral measures such as the angular power spectrum and the two-point correlation function. Finally, a generalized framework for image registration is presented. Section 10.5 gives first an overview of the filtering process from systems theory, including a study of filters in Fourier space. Second, some simple methods for the analysis of structures inherent in images are discussed. In Section 10.6 invariance properties are introduced and representations accomplishing these properties are defined. From the perspective of image processing these are statistical moments dealing with continuously valued images. From the perspective of Physical Cosmology we present methods from stochastic geometry dealing with binary structures. We show feature extraction by means of Minkowski Functionals, their generalization Minkowski Valuations and we present several applications. In Section 10.7 some concluding remarks are given.

10.2 Background from Computer Vision

The goal of image analysis is the construction of meaningful descriptions of scenes (with their physical objects) from images and the subsequent interpretation of this description. The result aims at serving functional and behavioral system performances such as, for example, the navigation and collision avoidance of a mobile robot, the sensory-motor control in steering a gripper for object manipulation, or the generation of a scene description in natural language output. For intrinsically 2D scenes, that are scenes with negligible 3D layout, the processing could be depicted in terms of a cascade of sequential processing steps. One operational goal is motivated by the processing and feature extraction for pattern recognition. In a nutshell, such a processing sequence can be summarized by:

- image acquisition, projective mappings, and image enhancement;
- image pre-processing by linear or nonlinear filtering and signal restoration;
- image segmentation and grouping for item aggregation;
- feature extraction or generation of structural image descriptions; and finally
- the classification of shapes and objects.

We will present various examples which highlight the properties and characteristics of images. Our focus is on the primary steps of pre-processing to feature extraction. For example, we start with a display of simple properties based on the statistics of the intensity distribution as well as the joint distribution of intensities in multi-image representations and pairs of image locations and their values. Spectral properties of images are derived using basic integral transforms of image signals, such as the Fourier transform and variants of it. Issues of discrete representations and mappings for projection of planar images onto curved surfaces will be introduced. Basic methods of image processing will be discussed, such as linear and space-invariant filters, which are precursory to the extraction of features from images.

10.3 Background from Physical Cosmology

Cosmology is the scientific study of the large-scale properties of the Universe as a whole. Physical methods are used to understand the origin, evolution and ultimate fate of the entire Universe. The Universe is the entire space–time continuum in which we live, together with all its ingredients within it.

Modern cosmology began with Einstein's seminal paper from 1917 [6] in which he applied his general theory of relativity, published only two years earlier, to construct for the first time a relativistic model for the Universe. The Einstein universe is a static one and, furthermore, at the time was consistent with all available astronomical data. Thus it was a great surprise when in 1929 Edwin Hubble observed that distant galaxies fade away, which indicates an expanding Universe. Observations show a hierarchy of structures. There are galaxies similar to our Milky Way composed of billions of stars similar to our Sun. Several galaxies form galaxy clusters where gravitational attraction is still dominant over expansion. Further galaxy clusters form galaxy superclusters which form, via filaments, a net-like structure that has large cavities called voids. This structure is called the large-scale structure (LSS) of the Universe. On even bigger scales the Universe is, on average, homogeneous and isotropic. Thereby one can define a mean mass density $\bar{\varrho}(t)$.

On large scales only gravitational interaction is relevant. Thus Einstein's general relativity provides the appropriate theory to describe the Universe as a whole. Homogeneity and isotropy lead to solutions of the Einstein field equations corresponding to a class of universes called Friedmann–Lemaître universes. These solutions describe the evolution of the local space–time metric and depend on several cosmological parameters, in particular on the energy content of the Universe. Observations suggest, and theory states, that the Universe monotonously expanded since it was generated by the Big Bang 13.7×10^9 years ago.

Had the Universe once been perfectly homogeneous, there would be no structures today. However, the LSS of galaxies, galaxy clusters and galaxy superclusters shows fluctuations of the mass density $\varrho(t, \boldsymbol{x})$ about the mean $\bar{\varrho}(t)$ measured by the density contrast $\Delta\varrho(t, \boldsymbol{x}) = (\varrho(t, \boldsymbol{x}) - \bar{\varrho}(t))/\bar{\varrho}(t)$. These fluctuations are caused by primordial density fluctuations derived from the initial conditions at the Big Bang. After the Big Bang the ingredients of the Universe have undergone several phase transitions. 380 000 years after the Big Bang, called the decoupling age t_{rec}, matter and radiation decoupled. Thereby the detectable radiation background called the cosmic microwave background (CMB) resulted from free streaming photons. Note that the CMB is an almost perfect isotropic radiation on the celestial sphere which satisfies the quantum mechanical law of temperature radiation, that is Planck's law, with a mean temperature of $\bar{T} = 2.725$ K with an extraordinary precision and possessing tiny temperature deviations from isotropy of relative order of $\Delta T(\theta, \phi) = (T(\theta, \phi) - \bar{T})/\bar{T} \propto 10^{-5}$ only. These fluctuations are highly correlated[2] with the mass density contrast $\Delta\varrho(t_{\text{rec}}, \boldsymbol{x})$ of the entire early universe at the decoupling age t_{rec}. They are interpreted as the seed of todays observed structures and can be understood as a kind of projected snapshot from the entire early universe at decoupling age. Thereby the CMB represents one of the most powerful tools in cosmology and is the oldest accessible information with todays possibilities. The fluctuations of the CMB are shown in Figure 10.1.

Due to relativistic effects, that is, the finite speed of light, information of distant objects can only be received from past events. Furthermore, the expansion of the Universe prohibits the access to the entire Universe. That part which is accessible to observations is called the observable Universe. Since we cannot perform

[2] In the full relativistic description there is a dependence on the total energy content. In the standard model there is radiation, baryonic matter, dark matter and dark energy. One distinguishes between primary effects, that is, effects at decoupling age, and secondary effects, that are effects during propagation from decoupling age until the observation time.

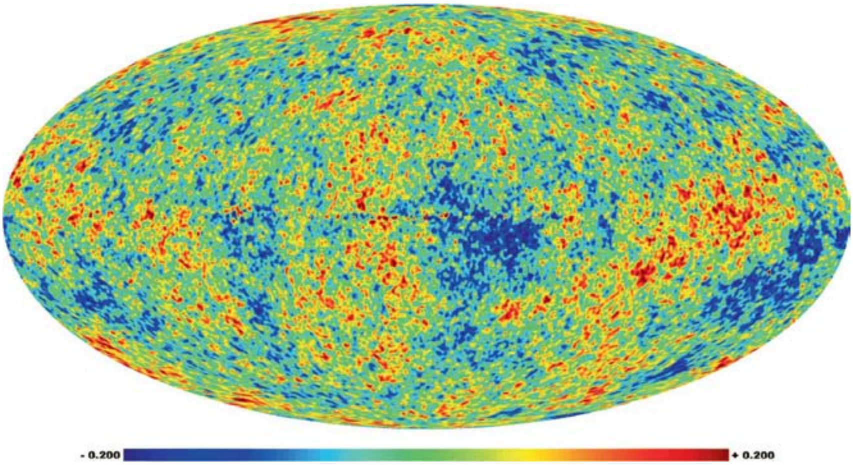


Figure 10.1 Cosmic microwave background (CMB) – Fluctuations $(T(\theta, \phi) - \bar{T})$ mK of the CMB measured by the NASA satellite Wilkinson Microwave Anisotropy Probe (WMAP) [7]. Dark indicates colder regions and bright hotter regions than the mean temperature of $\bar{T} = 2.725$ K. These structures detected on the celestial sphere are interpreted as the seeds of today's observed structures and can be understood as a kind of snapshot of the entire early Universe at the decoupling age. The image is displayed in Mollweide projection (see Section 10.4.1) and has a resolution of 12×512^2 pixels in HEALPix pixelization [8].

experiments with many universes, for example, we cannot repeat the Big Bang, cosmology is based on observations concerning just one universe, that is the one in which we happen to live. For that reason large-scale computations are performed simulating a large ensemble of universes. We are forced to extrapolate the laws of physics to extreme conditions where, in principle, they may no longer apply. For comparison and distinction of complex outputs, methods of image processing and feature extraction are of major interest in cosmology as will be illustrated below.

10.4
Image Formation and Characterization

Images are acquired in different applications for further analysis of their structural content for evaluation or steering and control purposes. At the sensory level of (discrete) signal generation an image can be described as a structural distribution of intensity or grey values. From a different point of view, namely statistical physics, the image values can be considered as an observations of stochastic variables. Both perspectives lead to essentially the same definitions of such basic image properties. Here, we will first present some basic formal methods to describe the generation and representation format for 2D images. Next, we sketch some basic image properties, namely their representation, the characteristics, and the registration of images. To study the characteristics, we model an image as an observation of a stochastic process (as sketched above), which is also the basic model in Physical Cosmology where it is assumed that the initial conditions of the Universe are described by a homogeneous and isotropic Gaussian random field. Throughout the article $u(\boldsymbol{x})$ denotes a field of scalar values which can be interpreted in two ways, namely, as intensity or probability distributions. Such scalar values in a field are addressed for the spatial position of a vector in an arbitrary space. Multiple intensity values, which are measured by different sensory channels or registration methods, but in a spatially registered way, can be represented as a vector field $\boldsymbol{u}(\boldsymbol{x})$. In other words, such an image, at a spatial position $u(\boldsymbol{x})$, contains a vector of dimension 2, 3, or n, depending on the number of registered sensory channels. Examples are images from magnetic resonance imaging (MRI) combining spin relaxation, T1 and T2, images taken by a color camera measuring in different bands of wavelength sensitivity as red, green, blue (RGB), or a scalar image field of intensity values combined with another field of derived features, such as local variances measured in a local image neighborhood.

10.4.1 Generation and Representation of 2D Images

Our focus in this article will be on two-dimensional (2D) images which are acquired from some environmental measurement space of arbitrary dimension. In the case of the three-dimensional (3D) space of an outdoor scene the image acquisition process can be formally described by a projective transform. Most often the projection results in a 2D cartesian coordinate system commonly named image plane. In Physical Cosmology, instead, an image can be defined on a unit sphere. In such cases the mapping of the (image) plane onto the curved surface can be described by a geometric transform. If the topology of the two such surfaces is identical then the mapping can be inverted. In the case when the image acquisition is distorted, for example due to some geometric lens distortion, the proper registration is also described by a proper warping transform. Here, we briefly sketch some projective as well as mapping transformations. A more complete overview is given in [9].

10.4.1.1 Perspective Projection

For the perspective projection, a point in 3D space $\boldsymbol{x} \in \mathbb{R}^3$ is projected in 2D space $\boldsymbol{y} \in \mathbb{R}^2$, for example, representing an image plane, by

$$(y_1, y_2) = (x_1, x_2) \cdot f/x_3 \ , \tag{10.1}$$

where the third component is omitted because it equals the constant f. The geometric interpretation is that arbitrary points (normally $x_3 > f$) are projected onto an image plane positioned at a positive distance $x_3 = f$ from the coordinate center $(0, 0, 0)$. If, instead, the distance from the projection center is taken to be negative, that is $-f$, the projection resembles a pinhole projection. A key characteristic is that the resulting image is upside down (negative image) while the former case yields a positive image. It should be noted that, in the extreme case of very distant scene points with $x_3 \gg f$, for small objects relative to the field of view, and with only minor relative depth variations, the perspective projection can be approximated by an orthographic projection, namely $(y_1, y_2) = (x_1, x_2)$.

10.4.1.2 Stereographic Projection

In various applications, specific surface properties, such as their 3D orientation, need to be represented in a flat space in order to build proper data structures. A useful approach is to map the spherical hemisphere of surface normals visible from the observer's viewpoint onto a tangential plane positioned in the sphere's north pole. If the center of projection is shifted into the south pole, then the upper hemisphere of the unit sphere is mapped into a circle of radius two (stereographic projection; [10]). Though this mapping is complete it also has some shortcomings, namely that only one-half of the sphere is projected and that the projection is not area-preserving.

10.4.1.3 Mollweide Projection

Generally, images can be defined on arbitrary surfaces (see also Section 10.4.4). Examples are the surface of the earth, which is ideally defined on a unit sphere, or satellites which measure over the celestial sphere. Therefore, images defined on the sphere play a special role. An approach that overcomes the limitations of the stereographic projection is the Mollweide projection,

$$y_1 = (2\sqrt{2}/\pi)\Phi\cos(\Psi/2) \tag{10.2}$$

$$y_2 = \sqrt{2}\sin(\Psi/2) \quad \text{with} \quad \Psi + \sin\Psi = \pi\sin\Theta \quad \text{and} \tag{10.3}$$

$$\Phi = \arg(x_1 + ix_2)\,, \quad \Theta = \arctan\left(x_3\Big/\sqrt{x_1^2 + x_2^2}\right)\,, \tag{10.4}$$

where $\Phi \in (-\pi, +\pi]$ denotes the longitude from the central meridian, $\Theta \in (-\pi/2, +\pi/2)$ denotes the latitude, and Ψ denotes only an auxiliary angle. This projects the surface of the total unit sphere onto a plane forming an ellipse, where the major axis is twice as long as the minor axis. Additionally, this projection is area preserving. This property can be easily checked if one integrates over an arbitrary

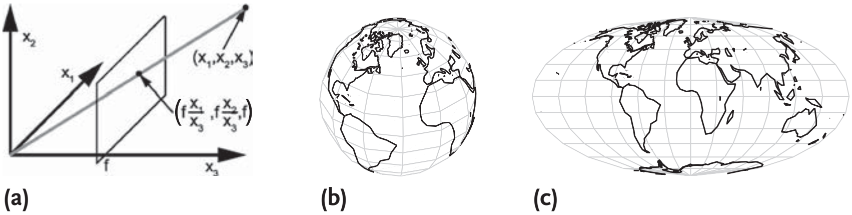


Figure 10.2 Projections: (a) Perspective projection of the point (x_1, x_2, x_3) onto the image plane positioned at f. (b) Continental contour lines of the earth. (c) Area-preserving Mollweide projection for the continental contour lines.

surface patch on the sphere and on the plane, and shows that the calculated areas are equal in size. Figure 10.2c shows the projection for the continental contours of the earth.

10.4.2 Image Properties

In each image representation the following key properties can be identified: (i) The space or surface where the image is defined (see methods for projections in Section 10.4.1); (ii) the number of quantization levels; (iii) the resolution; and (iv) the scale. Although, the scale of an image is intertwined with the resolution, we discuss these two properties separately. Figure 10.3 depicts all these properties.

10.4.2.1 Quantization Property

The quantization levels are defined by the range of the data values. This property is determined by two criteria. First, the range of the acquisition sensor, and

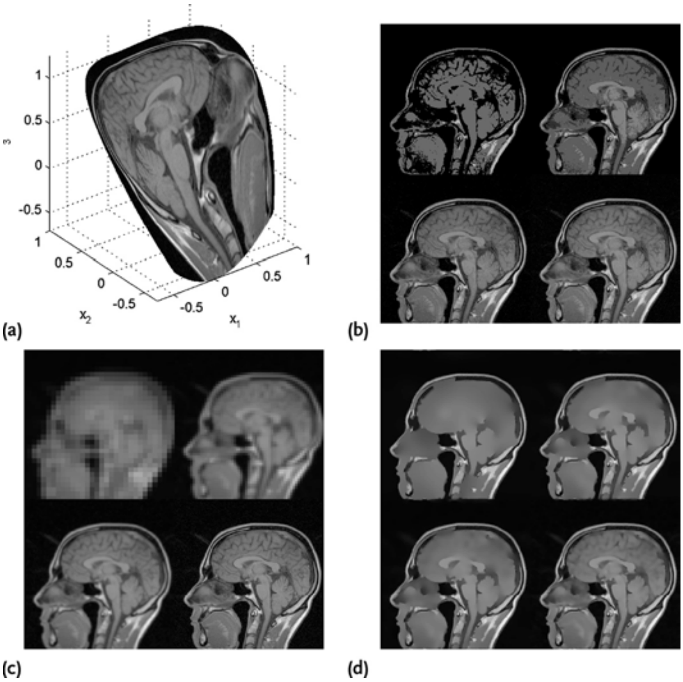


Figure 10.3 Image properties: (a) 2D magnetic resonance image mapped onto an arbitrary surface, thus the image coordinates depend on the surface geometry. (b) Same image visualized with 4, 8, 16 and 32 quantization levels. (c) Image with a resolution of 32 × 32 px, 64 × 64 px, 128 × 128 px, and 256 × 256 px (original). (d) Scales of nonlinear isotropic diffusion of the image for 1000, 100, 50, and 10 iterations ($\lambda = 0.02$, $\sigma = 1.5$, parameters referring to [11]).

second the storage format. Ordinary formats have 8 Bits, and thus 256 quantization levels. For color images each channel of the three channels in an RGB model could be quantized into 8 Bits, consequently having 24 Bits. Note, that the quantization always produces an error of discretization, which depends on the number of quantization levels and the sampling of the quantization interval.

10.4.2.2 Spatial Resolution Property

The upper spatial resolution is limited by the quality of the sensor, and the lower resolution is determined by the smallest object in the image which should be properly represented. Generally this depends on further processing tasks, for example, the successful recognition of a small object. Additionally the arrangement of pixels can be done in different ways for an image which is defined on a sphere the choice is not particularly obvious. Therefore, we refer to the HEALPix [8] pixelization technique on a sphere which possesses many features of a cartesian pixelization on a plane.

10.4.2.3 Scale Space Property

Scale spaces are used to represent the inner structure of an image with different levels of detail. On the finest level, all objects and structures are visible. In contrast, larger scales combine fine neighboring structures and subsume them to single objects. A general technique for the construction of scale spaces is the diffusion process [11],

$$\frac{\partial}{\partial t} u(\boldsymbol{x}, t) = \Delta[\varrho(u; \boldsymbol{x}, t)\nabla u(\boldsymbol{x}, t)] \ , \tag{10.5}$$

with u denoting the current diffused image at each location $\boldsymbol{x}$ for an internal time t, or scale. ϱ denotes the diffusion parameter, which determines the local rate of diffusion. In the simplest case, called homogenous diffusion, ϱ is a constant and does, not depend either on space or time, or on the actual solution $u(\boldsymbol{x}, t)$. For this type of diffusion an analytical solution can be derived based on the Green's function approach which leads to a Gaussian kernel with $\sigma = \sqrt{2t}$. An image constructed with this type of diffusion could also be defined by a corresponding resolution, which means smoothing the image with the same Gaussian kernel as for the diffusion process. Other types of diffusion for the construction of different scales, have results which cannot be derived by Gaussian smoothing. In particular, the results in Figure 10.3 (c) and (d) for resolution and scale are different. Here, a nonlinear isotropic diffusion is used for the construction of different scales, where ϱ depends on the actual solution $u(\boldsymbol{x}, t)$ and is a scalar. In image processing ϱ is regulated by the structure of the actual solution $u(\boldsymbol{x}, t)$, for example, the image gradient.

10.4.3 Basic Image Characteristics

For the analysis of image characteristics an image can be defined as the observation of a stochastic process in three ways. First, the image is modeled as the outcome of one random variable (RV); second, as one observation of a random field (RF), and third as a series of RVs. Each distinct modeling allows the study of distinct image characteristics which contain information of the structure and distribution of image intensities. An overview of characteristics and modeling is given in Table 10.1.

10.4.3.1 Histogram

Assume that all intensities contained in an image are continuous and the outcome of a single RV $\boldsymbol{X}$. Additionally, the probability distribution function of $\boldsymbol{X}$ is $u(x)$ for all $x \in \mathbb{R}$ of the random space. With this formalism the distribution of intensities in an image can be expressed by the normalized histogram

$$H_{\mathrm{N}}(B_\varepsilon) := P(\boldsymbol{X} \in B_\varepsilon) = \int_{x \in B_\varepsilon} u(x)\,\mathrm{d}x \,, \tag{10.6}$$

with $B_\varepsilon := \{x | b - \varepsilon/2 \le x < b + \varepsilon/2\}$, which equals the probability distribution function for $\varepsilon \to 0$. For images with continuous intensities the histogram has bins counting an interval of intensities which have only a finite number of levels. In this case ε defines the width of these bins. Figure 10.4a shows histograms of three image features. For the cumulative normalized histogram

$$H_{\mathrm{N,C}}(b) := F_u(b) = \int_{-\infty}^{b} u(x)\,\mathrm{d}x \tag{10.7}$$

these bins are not necessary and one integrates from the lowest possible intensity $-\infty$ to the intensity level b. This histogram equals the probability mass function F_u of the RV $\boldsymbol{X}$.

10.4.3.2 Covariance and Correlation

For two images modeled by RVs $\boldsymbol{X}$ and $\boldsymbol{Y}$ a comparison on the basis of their statistical behavior can be achieved by the covariance

$$\mathrm{Cov}_{\boldsymbol{X},\boldsymbol{Y}} = \langle (\boldsymbol{X} - \langle \boldsymbol{X} \rangle)(\boldsymbol{Y} - \langle \boldsymbol{Y} \rangle) \rangle = \langle \boldsymbol{XY} \rangle - \langle \boldsymbol{X} \rangle \langle \boldsymbol{Y} \rangle \;, \tag{10.8}$$

Table 10.1 Formalism for the modeling of an image and the achieved image characteristics.

Model	Single RV	RF	Series of RVs
One input	histogram	co-occurrence, Fourier transformation, two-point correlation	joint distribution, Fourier transformation
Two inputs	correlation	joint histogram, correlation	–

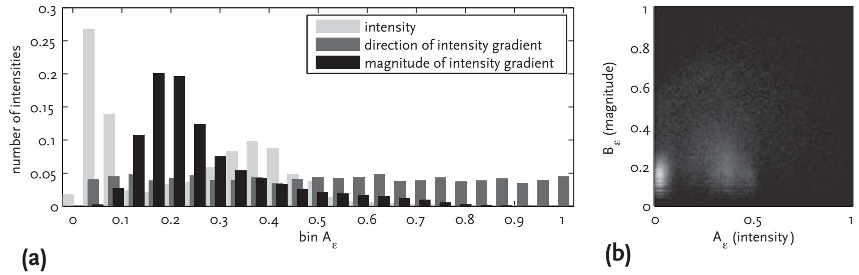


Figure 10.4 Image characteristics: (a) Histogram for image features, the intensity, direction, and magnitude of intensity gradient are from Figure 10.3c original (25 bins). Therefore the direction and magnitude are normalized to the unit interval. The magnitude has an unimodal histogram (one peak), the intensity a bimodal histogram (two peaks), and the direction is almost equaly distributed. (b) Joint histogram between intensity and gradient magnitude (256 bins). The two peaks in the histogram are the overlay between the unimodal and multimodal histograms, where the peaks are located at the same bin numbers as in the histograms. For visualization the square root of the joint histogram is shown.

where $\langle\cdot\rangle$ denotes the average over all observations of the RV. This definition is lacking for the dependence of the interval of Cov on the interval of outcomes from $\boldsymbol{X}$ and $\boldsymbol{Y}$. A normalized variant is the correlation

$$\mathrm{Cor}_{X,Y} = \mathrm{Cov}_{X,Y} / \sqrt{\mathrm{Var}(\boldsymbol{X})\mathrm{Var}(\boldsymbol{Y})} \in [-1, +1] \,, \tag{10.9}$$

where Var denotes the variance of the individual random variables. The division by the variances of $\boldsymbol{X}$ and $\boldsymbol{Y}$ normalizes the interval to $[-1, +1]$ independently of the input intervals. Correlation-based techniques are used in many fields of image processing; for example, for template matching, flow/stereo estimation and image registration [12, 13]. Generally, the goal is the determination of correspondences between the same sub-image in two images with temporal or spatial coherence.

10.4.3.3 Joint Histogram

Both characteristics (the normalized histogram and the correlation) do not include any information about the spatial distribution of intensities. All intensities are outcomes of the same single RV not bound on any spatial position in the image. On the contrary a dependence can be modeled by RFs $\boldsymbol{X}(x_1, x_2)$, containing one RV for each spatial position (x_1, x_2) in the image (here 2D). For two RFs $\boldsymbol{X}(x_1, x_2)$ and $\boldsymbol{Y}(x_1, x_2)$ with their corresponding probability distribution functions u and v, the joint normalized histogram is

$$\begin{aligned} H_\mathrm{J}(A_\varepsilon, B_\varepsilon) &= \int\!\!\int P(\boldsymbol{X} \in A_\varepsilon \wedge \boldsymbol{Y} \in B_\varepsilon)\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &= \int\!\!\int\!\!\int_{x\in A_\varepsilon}\!\int_{y\in B_\varepsilon} u(x_1, x_2; x)v(x_1, x_2; y)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}x_1\,\mathrm{d}x_2 \,, \end{aligned} \tag{10.10}$$

where the outer two integrals operate over the total image domain, and the inner two integrals over the bin intervals A_ε and B_ε. Note, that a separate probability distribution function u exists for each position (x_1, x_2) of the RF, and x denotes the argument of the function u, as in the 1D case $u(x)$. In the joint histogram, paired intensities according to the bins A_ε and B_ε are voted, where the intensity pair is located at arbitrary positions (x_1, x_2) in the two RFs $\boldsymbol{X}$ and $\boldsymbol{Y}$. In Figure 10.4 (b) a joint histogram between the feature channels intensity and magnitude of the intensity gradient is shown.

10.4.3.4 Co-occurrence

For images with multiple channels, as for example images acquired by multispectral sensors or color images, a joint histogram describes the correlation between the intensity distributions of different channels. The joint occurrence of values within *one* image or channel can be quantified by the co-occurrence matrix

$$\mathrm{Co}(A_\varepsilon, B_\varepsilon) = \int\int\int_{x\in A_\varepsilon}\int_{y\in B_\varepsilon} u(x_1, x_2; x)u(T(x_1, x_2); y)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}x_1\,\mathrm{d}x_2 \ . \quad (10.11)$$

Here, the first RF $\boldsymbol{X}(x_1, x_2)$ is defined by the image and the second RF $\boldsymbol{X}_T(x_1, x_2)$ by a spatial transformation $T(x_1, x_2)$ of the first RF. This characteristic highlights periodic structures and shifts of regions within one image.

10.4.3.5 Fourier Transformation of One Observation of an RF

Many applications in image processing profit from the analysis of images in the Fourier space, especially the study of the effectiveness of filters (see Section 10.5). The Fourier transformation

$$\hat{u}(k_1, \ldots, k_d; x) = \int_{\mathbb{C}} \ldots \int_{\mathbb{C}} u(x_1, \ldots, x_d; x) \exp\left(-i\sum_{l=1}^{d} k_l x_l\right) \mathrm{d}x_1 \ldots \mathrm{d}x_d \ , \quad (10.12)$$

$$u(x_1, \ldots, x_d; x) = \frac{1}{(2\pi)^d}\int_{\mathbb{C}} \ldots \int_{\mathbb{C}} \hat{u}(k_1, \ldots, k_d; x) \exp\left(i\sum_{l=1}^{d} k_l x_l\right) \mathrm{d}k_1 \ldots . \mathrm{d}k_d \ , \quad (10.13)$$

from the spatial domain u into the Fourier domain $\hat{u}$, is again based on the formalism of the RF, where one concrete observation of the RF is transformed. After the transformation into the Fourier domain, $\hat{u}$ is a complex number. Therefore, $\hat{u}$ can be analyzed in phase Φ and amplitude A

$$\Phi = \arg(\hat{u}) \ , \quad \text{and} \quad A = |\hat{u}| \ . \quad (10.14)$$

In Figure 10.5b the amplitude of the transformed input image a is shown. The corresponding inverse transformation of a filtered version is depicted in Figure 10.5d. For images the information about the spatial locality of structure is stored in the phase, and the information about the general periodicity is represented within the amplitude. To highlight this property consider a shift of the image in the spatial domain, therefore only the phase is influenced, not the amplitude. Thus, no information of the locality is included in the amplitude.

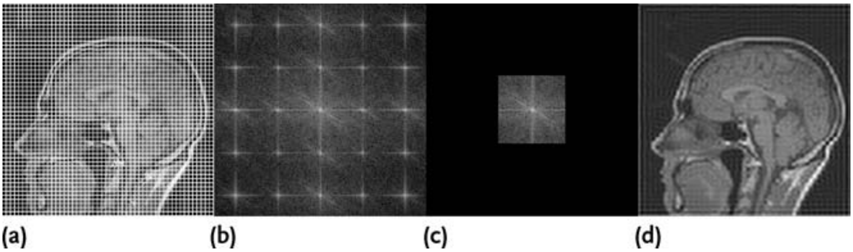


Figure 10.5 Fourier and inverse Fourier transformation: (a) Input image again from Figure 10.3b (original) with superimposed regular grid. (b) Fourier spectra (amplitude) of the image with characteristic peaks representing the grid and their multiples. (c) Boxed version of the fourier spectra. This realizes a box filter, where frequencies representing the first multitude of the grid are cut off. For visualization the square-root of the spectra is plotted. (d) Inverse Fourier transformation of boxed spectra.

10.4.3.6 Fourier Transformation of a Common Distribution Function

In some cases the stochastic process is defined by a series of RVs $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_d)$, and this models the dimensionality of the image. For example, an image is defined as the outcome of a dD normally distributed RV. Analogous to the Fourier transformation of RF, here the transformation is defined on the basis of the common density distribution function $u(x_1, \dots, x_d)$ which equals the product of the single distribution function $u(x_i)$ if the RVs are independent. The Fourier transformation

$$\hat{u}(\boldsymbol{k}) = \int_{\mathbb{C}^d} u(\boldsymbol{x}) \exp(-i\boldsymbol{k}\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,, \quad u(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{C}^d} \hat{u}(\boldsymbol{k}) \exp(i\boldsymbol{k}\boldsymbol{x})\,\mathrm{d}\boldsymbol{k}\,, \tag{10.15}$$

where $\boldsymbol{x}, \boldsymbol{k} \in \mathbb{R}^d$ is the joint characteristics function of the RVs $\boldsymbol{X}_1, \dots, \boldsymbol{X}_d$. In other terms this is the expected value $\hat{u} = \left\langle \exp(-i \sum_{l=1}^{d} k_l \boldsymbol{X}_l) \right\rangle$ of the series of RVs in respect to the Fourier base. $\hat{u}$ can be interpreted as the probability distribution function of new RVs $\hat{\boldsymbol{X}}_1, \dots, \hat{\boldsymbol{X}}_d$, which are also independent, in fact, of the orthogonality of the Fourier base.

10.4.3.7 Two-Point Correlation Function and Power Spectrum

In cosmology, images of homogeneous and isotropic RFs are often studied. Here, characteristics of length scale or separation distance are of special interest. The power spectrum is given by an average over the Fourier modes $\hat{u}(\boldsymbol{k})$ with a wave number $k = |\boldsymbol{k}|$ of the field[3] $u(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^3$. In configuration space, a field can be quantified by the two-point correlation function

$$\xi(r) := \langle u(\boldsymbol{x})u(\boldsymbol{x} + \boldsymbol{r}) \rangle \,, \tag{10.16}$$

[3] Note that we omitted the x for one realization of the RF $u(\boldsymbol{x}, x)$ at the position $\boldsymbol{x}$, because cosmological observations always display one realization.

where the average $\langle\cdot\rangle$ is taken over all positions $\boldsymbol{x}$ and all orientations of the separation vector $\boldsymbol{r}$, assuming homogeneity and isotropy. Hence there is only a dependence on $r = |\boldsymbol{r}|$. The two-point correlation function of a field on a sphere is

$$C(\vartheta) := \langle u(\theta, \phi) u(\theta', \phi') \rangle \ , \tag{10.17}$$

where now the average $\langle\cdot\rangle$ is taken over all pairs of (θ, ϕ) and (θ', ϕ') which are separated by the angle ϑ. Again a power spectrum

$$C_l := \left\langle |a_{lm}|^2 \right\rangle = \frac{1}{2l+1} \sum_{m=-l}^{l} |a_{lm}|^2 \tag{10.18}$$

can be defined, where a_{lm} are the complex coefficients obtained from an expansion into spherical harmonics $Y_{lm}(\theta, \phi)$ due to $u(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} a_{lm} Y_{lm}$. Here l parameterizes the separation scale and m the direction. In cosmological applications the so-called angular power spectrum

$$\delta T_l^2 := \frac{l(l+1)}{2\pi} C_l \tag{10.19}$$

is used. Note that, in the case of statistical homogeneity and isotropy, a two-point correlation function can be obtained by a transformation from its corresponding

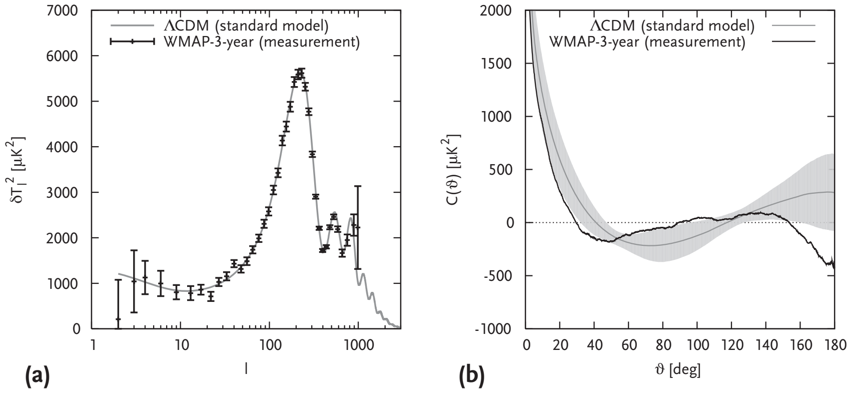


Figure 10.6 Statistical quantities of the CMB: (a) Angular power spectrum δT_l^2 of the observed fluctuations in the cosmic microwave background (CMB) measured by the Wilkinson Microwave Anisotropy Probe (WMAP) [7]. The error bars include measurement errors and the statistical variance. These characteristics are very sensitive, i.e. the peak positions and peak heights, to cosmological parameters. Independent cosmological determinations of the cosmological parameters are in excellent agreement with the best-fit standard model ΛCDM (gray line). The systematic deviations on largest scales (small l) cannot be explained by the standard model and are possible indications of a finite Universe [14]. (b) Two-point correlation function $C(\vartheta)$ of the best-fit standard model ΛCDM (dark gray line and the statistical variance as light gray area) and measurements of WMAP (black line). This characteristic highlights the largest scales ($\vartheta \approx 180°/l$) where the explanation by the standard model is limited. Going beyond the standard model, recent work shows [14], that beside the suppression of δT_l^2 on small l the shape of this characteristic can be reproduced with high confidence, by studying universes with a finite spatial extension.

power spectrum and vice versa. These characteristics carry the same information, but highlight different separation scales and thus different cosmological features. In Figure 10.6 the angular power spectrum and the two-point correlation function of the measured cosmological microwave background (CMB) is shown where now $u(\theta, \phi) = T(\theta, \phi) - \bar{T}$ (see Section 10.3). In addition to omitting the constant term (monopole) with $l = 0$, which is equivalent to $\bar{T}$, the dipole with $l = 1$ is, also omitted due to a superimposed dipole generated by the relative motion of the observer to the CMB.

10.4.4 Image Registration

The problem in image registration is to find a mapping between a reference image u^{ref} and a template image u^{tem}. Formally, the problem is to determine a transformation ϕ applied to the template minimizing the difference to the reference image. This is a minimization problem which could include additional constraints, for example, the exact mapping of specific parts within the template and corresponding parts in the reference image. Here, we describe a generalized framework formerly introduced in [15] which is based on variational techniques. Let us define the optimization problem

$$E(\phi) := D(u^{\text{ref}}, u^{\text{tem}}; \phi) + \alpha S(\phi) + \beta C^{\text{soft}}(\phi) \xrightarrow{\phi} \min\,, \quad C^{\text{hard}}(\phi) = 0\,, \quad (10.20)$$

which contains three main terms: (i) a data term D; (ii) a smoothness term S; and (iii) a (soft) constraint term C^{soft}. Additionally, hard constraints C^{hard} can be included by side conditions. The parameter α steers the smoothness and β controls the influence of additional constraints, respectively. In Figure 10.7 the functionality of the three main terms is depicted. The task is to find a transformation $\hat{\phi}$ such that $E(\hat{\phi})$ is minimal, considering the side conditions. A restricted class of possible transformations are affine transformations, including translation t, rotation r, scaling c, and shear s. For these transformations each spatial position $\boldsymbol{x} \in \mathbb{R}^3$ is transformed into projective space by $\Theta(\boldsymbol{x}) = \boldsymbol{y} = (x_1, x_2, x_3, 1)$. Inverse transformation Θ^{-1} is realized by $\boldsymbol{x} = (y_1, y_2, y_3)/y_4$, if the fourth component of $\boldsymbol{y}$ is not equal

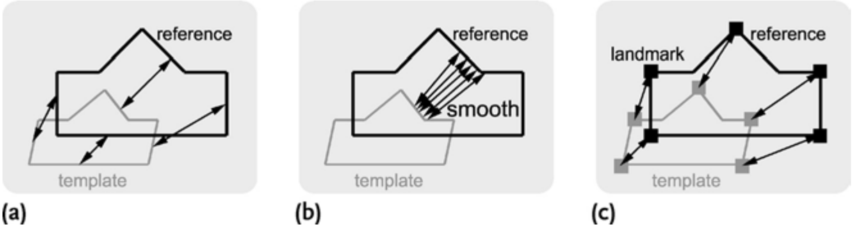


Figure 10.7 Image registration: (a) The problem in image registration is to find a mapping between the template and the reference image. (b) Additionally, smoothness for the solving transformation ϕ could be required, and (c) landmarks should be matched best for soft constraints and exact for hard constraints.

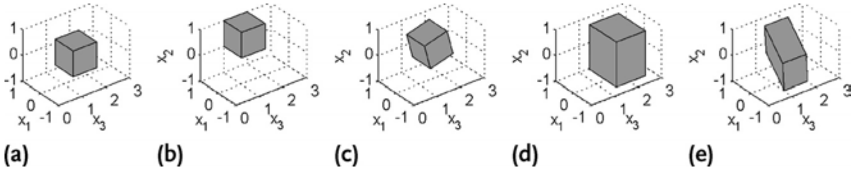


Figure 10.8 Affine transformations: (a) Reference for the transformations. (b) Translation for $\boldsymbol{t} = (0.5, 0.5, 0)$. (c) Rotation for $\boldsymbol{r} = (15, 0, 0)$ deg. (d) Scaling for $\boldsymbol{c} = (1.5, 1.75, 1.25)$. (e) Shear for $\boldsymbol{s} = (85, 0, 0)$ deg.

to zero. Affine transformations in the projective space are realized by a sequential chained multiplication of transformation matrices

$$\phi(t, r, c, s; \gamma) = \begin{pmatrix} 1 & 0 & 0 & t_1 \\ 0 & 1 & 0 & t_2 \\ 0 & 0 & 1 & t_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos r_1 & \sin r_1 & 0 \\ 0 & -\sin r_1 & \cos r_1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \ldots$$
$$\begin{pmatrix} c_1 & 0 & 0 & 0 \\ 0 & c_2 & 0 & 0 \\ 0 & 0 & c_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \cot s_1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \ldots \gamma \, , \tag{10.21}$$

where we indicated all three rotational matrices by the rotation around the x_1-axis, and all three shearing matrices by the shear in the x_1x_2-plane. Figure 10.8 depicts examples of those transformations. More advanced transformation techniques can be found in [9].

For *global* affine transformations the parameters $t, r, c, s \in \mathbb{R}^3$ are defined only once for the domain, while *local* affine transformations exist for each spatial position $t(\boldsymbol{x})$, $r(\boldsymbol{x})$, $c(\boldsymbol{x})$, and $s(\boldsymbol{x})$. This approach is very extensive introducing twelve unknowns for each spatial position. A simplification assumes constant transformations in small regions within the total spatial image domain.

10.4.4.1
Data Term

The data term D compares the (gray) values of the reference image u^{ref} (or some extracted feature measure, such as edges, corners, etc.) with the values of the transformed template image u^{tem}. In this case several distance measures could be applied. An intuitive measure of distances is the sum of squared distances (SSD):

$$D(u^{\text{ref}}, u^{\text{tem}}; \phi) = \int \left[u^{\text{ref}}(\boldsymbol{x}) - u^{\text{tem}} \left(\Theta^{-1} \left(\phi[\boldsymbol{t}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{s}; \Theta(\boldsymbol{x})] \right) \right) \right]^2 \mathrm{d}\boldsymbol{x} \, . \tag{10.22}$$

This distance assumes that the intensities of corresponding values in u^{ref} and u^{tem} are equal. If this assumption is not satisfied, correlation-based methods could be applied, which assume a unimodal distribution of intensity values. For images with a multimodal histogram, mutual information (MI) related measures could be used, which are based on the joint intensity histogram. In general, the data term D is also called an *external force*, because this term is mainly driven by the intensity values of the template and reference.

10.4.4.2 Smoothness Term

In contrast to the data term, the smoothness term S is defined by the mapping ϕ which constitutes an internal force by imposing a condition on the set of possible solutions. The key observation is that in this framework any smoother which is Gârteaux-derivable could be applied [15]. Because of its similarity with the diffusion equation in Section 10.4.2 we present as a smoothness condition, the diffusion smoothness

$$S(\phi) = \int \| \nabla \phi(t, r, c, s; \Theta(\boldsymbol{x})) \|_2^2 \, \mathrm{d}\boldsymbol{x} \, . \tag{10.23}$$

Here the integral, which ranges over the total image domain, induces global smoothness onto the mapping function ϕ, by squaring and summing up all first-order partial derivatives of ϕ according to the spatial change of the variables t, r, c, and s. Thus, each strong change in ϕ causes a high derivative, which is unwanted and therefore penalized.

10.4.4.3 Constraint Term

Finally, we discuss the inclusion of extra constraints that need to be achieved by the optimized solution. Assume two sets of landmarks, the first set defined in the reference image $\{x_l^{\text{ref}}\}_{l=1\ldots m}$, and the second set defined in the template image $\{x_l^{\text{tem}}\}_{l=1\ldots m}$, where correspondence between the landmarks is expressed by the same index. A soft constraint term can be formalized by

$$C^{\text{soft}}(\phi) = \sum_{l=1}^{m} \| \Theta^{-1}(\phi[t, r, c, s; \Theta(x_l^{\text{tem}})]) - x_l^{\text{ref}} \|_2^2 \quad . \tag{10.24}$$

This constraint enforces the transformed landmarks of the template to be closely matched with the landmarks of the reference x_l^{ref}, but deviations are possible. In contrast, for the hard constraint C^{hard} a match should be exact.

10.5 Methods of Image Processing

In this section we discuss some approaches for pre-processing image signals utilizing a filtering process. Many methods in image processing utilize mechanism that can be described in terms of linear systems theory [2, 9, 16]. Filters can be defined according to their functional purpose, for example, smoothing by elimination of high-frequency content, or discontinuity detection by extracting high-frequency content. We briefly summarize the properties of linear systems and display Gaussian smoothing filters and some related derivative operations. So-called Gabor filters define band-pass operations for combined smoothing and discontinuity detection having localized spectral windowing properties. We also show how a bank of filters can be constructed. Finally, we briefly present approaches to nonlinear filtering as well as approaches that help to detect localized key points which obey local 2D image structure properties.

10.5.1 Filtering Process

We consider here a specific class of system operators $\mathcal{H}$ to model the filtering stage, namely those that are linear and space invariant.[4] Such systems are commonly assumed in image processing, since the computations can be fully described using linear systems theory. A system is linear if the identity

$$\mathcal{H}\{a\, u(\boldsymbol{x}) + b\, w(\boldsymbol{x})\} = a\, \mathcal{H}u(\boldsymbol{x}) + b\mathcal{H}w(\boldsymbol{x}) \ , \tag{10.25}$$

holds. Further, a system is space or shift invariant if

$$\mathcal{H}\{u(\boldsymbol{x} - \boldsymbol{x}_0)\} = v(\boldsymbol{x} - \boldsymbol{x}_0) \ , \quad \text{for} \quad \mathcal{H}\{u(\boldsymbol{x})\} = v(\boldsymbol{x}) \ , \tag{10.26}$$

to denote that the system operator response is position invariant given identical input conditions. Taken together, the system response for an arbitrary input is fully defined by the correspondence

$$\mathcal{H}\{u(\boldsymbol{x})\} = H(\boldsymbol{x}) * u(\boldsymbol{x}) \ , \quad \vdash \quad \mathcal{H}\{\hat{u}(\boldsymbol{k})\} = \hat{H}(\boldsymbol{k}) \cdot \hat{u}(\boldsymbol{k}) \ , \tag{10.27}$$

where the left-hand side denotes the convolution of the input signal $u(\boldsymbol{x})$ by the system's impulse response function $H(\boldsymbol{x})$ ($*$ is the convolution operator). The correspondence (denoted by $\vdash$) establishes that the same result can be computed in the spatial as well as the spectral, or Fourier, domain. In the frequency domain $\boldsymbol{k}$ the convolution is equivalent to a multiplication of the Fourier transforms of the respective signals. This property is useful to study the characteristics of systems.

10.5.2 Linear and Space-Invariant Filters

10.5.2.1 Gaussian

Smoothing for noise suppression is a key operation in early signal processing. An ideal low-pass filter $T \cdot \text{rect}(kT)$ is defined by a specific cut-off frequency $1/(2T)$ in the spectral domain. Due to the similarity theorem the corresponding spatial

[4] Space invariance is the generalization of the time invariance property defined for time-series analysis.

filter si($\pi x/T$) (where si(x) = sin(x)/x) is of infinite extent with a damping that is proportional to the maximum frequency. In order to utilize a filter function that is localized in the spatial as well as the frequency domain, a Gaussian low-pass is often employed

$$H^{\text{Gauss}}(\boldsymbol{x}) = \frac{\exp\left(-\parallel \boldsymbol{x} \parallel_2^2 /2\sigma^2\right)}{(\sqrt{2\pi}\sigma)^d} \quad \vdash \quad \hat{H}^{\text{Gauss}}(\boldsymbol{k}) = \exp\left(-\frac{\parallel \boldsymbol{k} \parallel_2^2}{2\hat{\sigma}^2}\right) . \tag{10.28}$$

The Fourier transform pair results in two Gaussians which are related by their standard deviations, namely $\hat{\sigma} = 1/\sigma$. Therefore a sharp spatial Gaussian corresponds to a flat Gaussian in the Fourier space and vice versa. (compare Figure 10.9a and e). Due to the Gaussian damping of higher spatial frequencies the filter application reduces the effective resolution of an image, resulting in a coarser scale (see Section 10.4.2).

10.5.2.2
First-Order Derivative

Spatial derivative operations can also be formulated by filtering operations. For example, the first-order derivative is denoted by

$$u(\boldsymbol{x} + e_j\,\mathrm{d}x) = u(\boldsymbol{x})\,\mathrm{d}x^0 + \frac{\partial}{\partial x_j}u(\boldsymbol{x})\,\mathrm{d}x^1 + O(\mathrm{d}x^2), \quad j = 1 \ldots d\ , \tag{10.29}$$

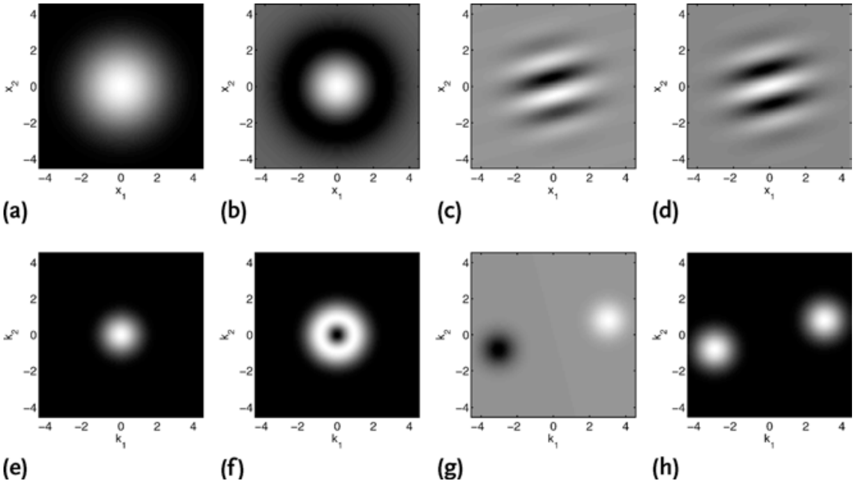


Figure 10.9 Filters and their spectra: (a) Gaussian filter. (b) Negative of Laplacian of Gaussian. (c) Odd part of Gabor filter (sine). (d) Even part of Gabor filter (cosine). In all Gaussians $\sigma = 1.5$, and for the Gabor filters $\lambda = 2$, $\psi = 15$ deg. The corresponding spectra of the filters are drawn in the lower part. (e) shows the lowpass filter characteristic of the Fourier transformed Gaussian. (f) is a characteristic bandpass filter. (g) visualizes the spatial selectivity of the odd part in the Fourier space, and (h) for the even part. Note that, for better visibility, all values are rescaled to the full intensity range.

where e_j denotes the j-th unit vector. Division by dx and rearranging terms results in the operator for the first-order derivative

$$\mathcal{H}\{u(\boldsymbol{x})\} = \frac{u(\boldsymbol{x} + e_j \cdot \mathrm{d}x) - u(\boldsymbol{x})}{\mathrm{d}x} + O(\mathrm{d}x) \ . \tag{10.30}$$

All first-order partial derivatives for $j = 1, \ldots, d$ together form the gradient of the input image. In Fourier space the derivatives

$$\mathcal{H}\{u(\boldsymbol{x})\} = \frac{\partial}{\partial x_j} u(\boldsymbol{x}) \quad \vdash \quad \mathcal{H}\{\hat{u}(\boldsymbol{k})\} = i\boldsymbol{k}\hat{u}(\boldsymbol{k}) \quad j = 1 \ldots d \ , \tag{10.31}$$

lead to a multiplication of the original spectrum with $i\boldsymbol{k}$, as the partial derivatives can be calculated within the second integral of (10.15). Multiplication of the spectrum with a linear or faster-than-linear function amplifies noise. This effect can be reduced by appliance of a Gaussian filter kernel before calculating the derivative of the input image.

10.5.2.3
Second-Order Derivative and Laplacian of Gaussian (LoG)

The second-order derivatives are defined on the basis of the Hessian matrix

$$\mathcal{H}\{u(\boldsymbol{x})\} = \frac{\partial^2}{\partial x_j \partial x_l} u(\boldsymbol{x}) \quad \vdash \quad \mathcal{H}\{\hat{u}\} = -k_j k_l \hat{u}(\boldsymbol{k}) \quad j, l = 1 \ldots d \ . \tag{10.32}$$

In Fourier space the Hessian matrix is the negative of the transformed image $\hat{u}(\boldsymbol{k})$ multiplied by the two frequency components k_j and k_l. In this case noise is amplified twice which makes the second-order derivatives highly sensitive, especially for high-frequency noise.

The trace of this Hessian matrix defines the Laplacian operator $\mathcal{L} = \mathrm{trace}(\{\partial^2/(\partial x_j \partial x_l)\}_{j,l=1\ldots n})$. For suppression of noise again the Gaussian filter could be applied, before the Laplacian. Due to the law of associativity the convolution of an image with the Laplacian operator can be applied directly to the Gaussian, resulting in the Laplacian-of-Gaussian (LoG)

$$\begin{aligned} H^{\mathrm{LoG}}(\boldsymbol{x}) = \mathcal{L}\{H^{\mathrm{Gauss}}(\boldsymbol{x})\} &= \frac{1}{(\sqrt{2\pi}\sigma)^d} \left(\frac{\| \boldsymbol{x} \|_2^2}{\sigma^4} - \frac{1}{\sigma^2} \right) \exp\left(-\frac{\| \boldsymbol{x} \|_2^2}{2\sigma^2} \right) \\ &\vdash \hat{H}^{\mathrm{LoG}}(\boldsymbol{k}) = - \| \boldsymbol{k} \|_2^2 \ \hat{H}^{\mathrm{Gauss}}(\boldsymbol{k}) \ , \end{aligned} \tag{10.33}$$

which is characteristic for a bandpass filter, where the frequency with maximal amplification is $\tilde{k}_j = \pm \sqrt{2}\hat{\sigma}$ for each dimension j. The 2D version of this filter defines a ring with radius $\sqrt{2}\hat{\sigma}$ of maximum spectral sensitivity (see Figure 10.9f).

10.5.2.4
Gabor Filter

While the LoG operator specifies an isotropic band-pass filter, often orientation sensitive filter devices are needed, for example, to separate oriented texture patterns of similar wavelength properties. Gabor filters specify an example in which a selected set of frequencies are passed which fall into the region of a pair of Gaussian windows positioned along a given axis of orientation shifted in opposite directions with respect to the origin (see Figure 10.9 g and h). The combined frequency/phase shift (defined by the wavelength λ and direction ψ) of the Gaussian filter in the frequency domain leads to a modulation of the space domain Gaussian by the wave function $\exp(i\boldsymbol{k}_0\boldsymbol{x})$.

$$H^{\text{Gabor}}(x) = \exp(i\boldsymbol{k}_0\boldsymbol{x})H^{\text{Gauss}}(\boldsymbol{x}) \;, \quad \vdash \quad \hat{H}^{\text{Gabor}}(\boldsymbol{k}) = \hat{H}^{\text{Gauss}}(\boldsymbol{k}-\boldsymbol{k}_0) \;. \tag{10.34}$$

Note, that the Gabor filter results in a quadrature filter with an odd (sine) shown in Figure 10.9 (c) and an even part (cosine) shown in (d),[5] with their corresponding Fourier transforms in (g) and (h), respectively. For the interpretation of these two parts, again the phase and amplitude as defined in (10.14) are considered. Here, the amplitude can be interpreted as the power of the filtered image. The phase has high responses for parts of the image which are coincident with this phase and the specified wavelength. A separated analysis of the two parts shows that the odd part (sine) behaves in a similar way to a bandpass filter, where the even part (cosine) has a Direct Current (DC) level component, due to the residual response

$$\text{DC}(H^{\text{Gabor,even}}) = \frac{\int_{\mathbb{R}^d} 1\cos(\boldsymbol{k}_0\boldsymbol{x})H^{\text{Gauss}}(\boldsymbol{x})\,\text{d}\boldsymbol{x}}{(\sqrt{2\pi}\sigma)^d} = \exp\left(-\frac{\|\ \boldsymbol{k}_0\ \|_2^2}{2\hat{\sigma}^2}\right) \;, \tag{10.35}$$

for a constant signal. For a DC-level free definition the constant value $\text{DC}(H^{\text{Gabor,even}})$ is subtracted from the even part of the Gabor filter, which can be recognized in Figure 10.9d by a slightly darker gray in the display of responses as in c, especially for high frequencies.

10.5.2.5
Gabor Filter Bank

Using properly scaled and shifted Gaussian window functions the whole frequency domain could be sampled using Gabor filters. This, in turn, leads to a signal representation by a population of Gabor filter responses (compare Figure 10.10). This sampling can be achieved in two ways. (i) The Gaussian envelope $\hat{\sigma}_l$ is constant in all rings; and (ii) the number of Gaussians in each ring is constant, meaning that $\Delta\psi_l$ is constant in all rings l. For this second approach the wavelengths and standard deviations are

$$\lambda_{l+1} = \lambda_l \frac{1-\sin(\Delta\psi/2)}{1+\sin(\Delta\psi/2)}, \quad \text{and} \quad \hat{\sigma}_l = \frac{2\pi}{\lambda_l}\sin(\Delta\psi/2) \;, \quad l \geq 0 \;, \tag{10.36}$$

where l denotes the number of the ring, given λ_0 the radius of the innermost ring. This scheme constructs Gabor wavelets, defined by a constant number of waves in

[5] The Hilbert transform $\hat{u}(x) = u(x) * 1/(\pi x)$ of the even part results in the negative odd part and the Hilbert transform of the odd part results in the even part.

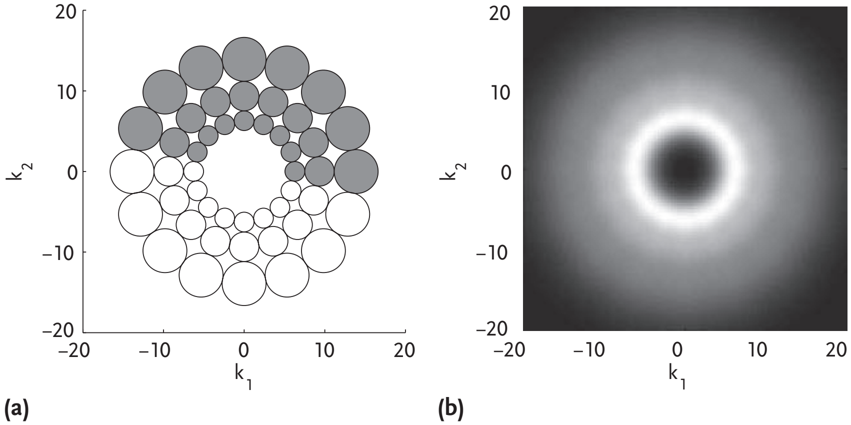


Figure 10.10 Construction of Gabor space: (a) Gabor filters drawn as circles with radius of standard deviation for three rings with $\Delta\psi = \pi/8$, and $\lambda_0 = 1$. Applications using only the energy of the quadrature Gabor filter need only a sampling of the half-space (drawn in gray). Here the whole space (including all circles) is sampled. (b) The superimposition of all filter spectra for the even part (cosine). The three filter rings in the total spectrum are visible through the three stepped gray valued regions. For better visibility the square-root of the spectra is shown.

the envelope of the Gaussian for each ring (self-similarity), due to $\lambda_{l+1}/\lambda_l = \hat{\sigma}_l/\hat{\sigma}_{l+1}$. In Figure 10.10 three rings of Gabor filters sampling the Fourier space are depicted. An application of Gabor filters is given in Section 10.5.4 for the extraction of contour lines from images.

10.5.3 Morphological Filtering

In addition to the linear position-invariant filtering, we briefly present a class of nonlinear operations based on mathematical morphology. Such nonlinear filters follow the general schema

$$\mathcal{H}\{u(x)\} = \mathcal{F}\{\{u(x)|x \in \mathcal{N}(x)\}\} , \tag{10.37}$$

where $\mathcal{F}$ operates on a set and returns a single element of this set, and $\mathcal{N}(x)$ is the neighborhood or support for the operator. These filters are also known as rank order or morphological operators, operating on the order of the elements in the set. For the first filter this set of input values is sorted and the central element is selected by the operator $\mathcal{F}$, which is the median of the input data set. This filter is appropriate for eliminating impulsive noise, visualized in Figure 10.11. In general, this filter obtains edges and transforms patches with continuous gray-level ramps into areas of a single gray-level, caused by the selection of the median element. Morphological operators select the minimum or maximum from the set. Therefore, closed objects become smaller according to their spatial extent for selecting the minimum and, respectively, wider for the maximum selection. These

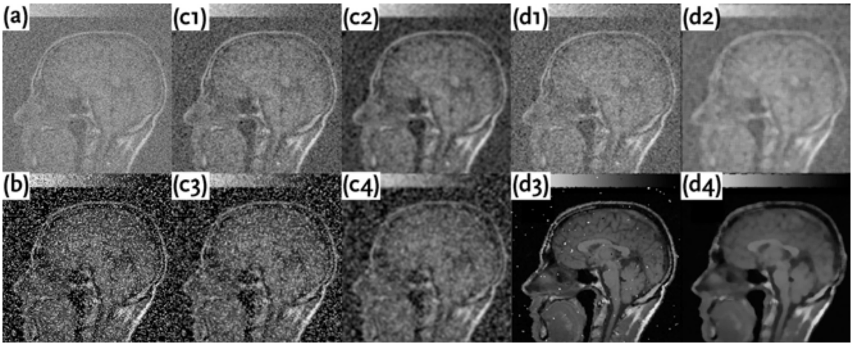


Figure 10.11 Gaussian and median filter: A magnetic resonance image is distorted by Gaussian noise with standard deviation $\sigma = 0.25$ and mean $\mu = 0.5$ (a), and impulsive noise where 25% of all pixels are disturbed (b). Results of filtering with a Gaussian kernel with $\sigma = 0.25$ and length of 3 px are shown in (c1) for Gaussian noise and (c3) for outlier noise, and with a Gaussian kernel with $\sigma = 1.75$ and length of 7 px in (c2) and (c4), respectively. Results for median filter with a neighborhood of 3 × 3 px for Gaussian noise are in (d1) and outlier noise in (d3), and with a neighborhood of 7 × 7 px for Gaussian noise in (d2) and outlier noise in (d4). Note that the median filter is appropriate to cope with impulsive noise, and the Gaussian filter is appropriate for handling Gaussian noise.

operations can be consecutively combined resulting in an opening or closing of structures. Further details for opening and closing are reported in Section 10.6.2.

10.5.4
Extraction of Image Structures

The gradient and higher-order derivatives of an image are key approaches for the extraction of image structures. For the gradient, first-order derivatives (as stated in (10.30)) are calculated. Each position in the image contains a gradient directed into the direction of the strongest increasing gray-value ramp (see Figure 10.12b). An analysis of the structure based on this gradient is not appropriate because uncorrelated noise and a constant gray value patch cannot be distinguished. Thus, the orientation which best fitts the gradients in a neighborhood (for example, defined by the size of the patches) should be calculated. This could be defined as an optimization problem,

$$\int_{\mathcal{N}(\boldsymbol{x}_0)} \| \boldsymbol{v}(\boldsymbol{x}_0)\nabla u(\boldsymbol{x}) \|_2^2 \, \mathrm{d}\boldsymbol{x} = v(\boldsymbol{x}_0)^t [\int_{\mathcal{N}(\boldsymbol{x}_0)} (\nabla \boldsymbol{u})^t \nabla \boldsymbol{u} \, \mathrm{d}\boldsymbol{x}] v(\boldsymbol{x}_0) \xrightarrow{v} \max . \qquad (10.38)$$

The vector product of the gradient integrated in the local neighborhood is the structure tensor

$$\{J(u)\}_{j,l} := \int_{\mathcal{N}(\boldsymbol{x}_0)} u_{x_j} u_{x_l} \, \mathrm{d}\boldsymbol{x} \, , \quad j, l = 1 \ldots d \, , \qquad (10.39)$$

for all positions $\boldsymbol{x}_0$. The eigenvalue decomposition of J is denoted by the eigenvalues λ_k and eigenvectors $\boldsymbol{v}_k \in \mathbb{R}^d$ for $k = 1 \ldots d$. The main direction in the neighborhood is the eigenvector corresponding to the largest eigenvalue. For the 2D case the full interpretation of the structure tensor is given in Table 10.2.

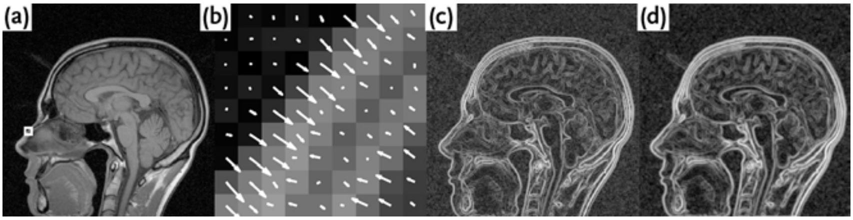


Figure 10.12 Image gradient, edges and contour: (a) Magnetic resonance image with marked region. (b) Intensity gradient in marked region. (c) Magnitude of intensity gradient. (d) Contour constructed of responses for oriented Gabor filters ($\sigma = 1$, $\Delta\psi = \pi/8$, $\lambda = \pi/3$). For clarity in (c) and (d) the square- root is shown.

Table 10.2 Interpretation of the structure tensor for a 2D image.

Condition	Interpretation
$\lambda_1 \approx \lambda_2 \approx 0$	constant gray-value in neighborhood
$\lambda_1 \gg 0, \lambda_2 \approx 0$	directed structure in neighborhood (hint for edge)
$\lambda_1 > \lambda_2, \lambda_2 \gg 0$	changes of gray-values in more directions (hint for corner)

Based on the structure tensor, measures for edges and corners in images can be defined. A corner is present if both eigenvalues are significantly greater than zero. A measure considering this condition is the Förstner corner detector [17]. An edge can be identified by the ratio between the eigenvalues. A contour line similar to edges is constructed with a small Gabor filter bank, consisting of one scale and eight orientations, only using the amplitude of the complex filter responses. From this ensemble of filter responses corresponding to each specific orientation the sum is calculated, resulting in the contour signal. This sum is depicted in Figure 10.12d.

10.6 Invariant Features of Images

Above, we have discussed some basic processing approaches for noise suppression, signal restoration, the detection of localized structures, and their proper coding. The main aim of signal processing is the extraction of relevant features from images to generate compact descriptions. Such descriptions need to possess certain invariance properties, mainly against transformations such as position, rotation, or scaling. Such descriptions serve as a basis for classification or matching different representations utilizing proper similarity measures. In this section we focus on features and descriptions derived thereof which are relevant for Physical Cosmology. We first address the aim to find matchings between objects using representations which are invariant to translation, rotation and scaling. Afterwards, we leave the direct description of scalar fields and switch to methods of stereography, using descriptions of binary fields. We present explicit relations of scalar fields to binary fields and discuss the connectivity of structures by topological classification. Then Minkowski Functionals are shown as a full set of shape descriptors which obey additivity and are invariant to translations and rotations. We also present their generalization, namely Minkowski Valuations. Finally, we end by illustrating applications in cosmology.

10.6.1 Statistical Moments and Fourier Descriptors

Representations with invariant properties are helpful for finding and matching objects and structures. Important invariance properties are translation, rotation, and scaling. These properties are depicted in Figure 10.8a–d. In the next paragraphs several representations which are invariant for at least some transformations are presented.

10.6.1.1 Statistical Joint Central Moments

The statistical joint central moments of the two random variables (RV)s $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with the joint probability distribution function $u(x_1, x_2)$

$$\begin{aligned}\mu_{p,q} &= \langle (\boldsymbol{X}_1 - \langle \boldsymbol{X}_1 \rangle)^p (\boldsymbol{X}_2 - \langle \boldsymbol{X}_2 \rangle)^q \rangle \\ &= \int\int (x_1 - \bar{x}_1)^p (x_2 - \bar{x}_2)^q u(x_1, x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ , \end{aligned} \tag{10.40}$$

are invariant with translation. If invariance with scale is additionally required we assume that $\tilde{u}(x_1, x_2) = u(x_1/\alpha, x_2/\alpha)$. Through simple substitutions in the integrals we see that $\tilde{\mu}_{p,q} = \alpha^{p+q+2}\mu_{p,q}$. If we divide $\tilde{\mu}_{p,q}$ through the zeroth moment to the power of $(p + q + 2)/2$ we obtain

$$\breve{\mu}_{p,q} = \alpha^{p+q+2}\mu_{p,q}/(\alpha^2\mu_{0,0})^{(p+q+2)/2} = \mu_{p,q}/\mu_{0,0}^{(p+q+2)/2} \ , \tag{10.41}$$

which is invariant with scaling. On the basis of the moments a tensor for the moment of inertia

$$J = \begin{pmatrix} \mu_{2,0} & -\mu_{1,1} \\ -\mu_{1,1} & \mu_{0,2} \end{pmatrix} \tag{10.42}$$

can be constructed. The orientation of the eigenvector corresponding to the smallest eigenvalue of J is $\Phi = 1/2 \arctan(2\mu_{1,1}/(\mu_{2,0} - \mu_{0,2}))$, which is the smallest moment of inertia. This criterion is invariant with scaling and translation, because of the invariance of $\mu_{p,q}$ with translation. The calculation of the ratio causes the invariance with scaling.

10.6.1.2 Fourier Descriptors

Fourier descriptors are a general method used for the compact description and representation of contours. Therefore, we assume that there exists a parameterized description $z(t) = z(t + lT) \in \mathbb{C}$, $t \in [0, \ldots, T]$, $l \in \mathbb{Z}$ of the contour, which is periodic in T, and t is the parameter for the actual position of the curve. For the Fourier transformation assume $z(t) = x_1(t) + ix_2(t)$, that the first component of the curve defines the real part and the second component defines the complex part. The Fourier transform

$$Z(\nu) = \frac{1}{T} \int_0^T z(t) \exp\left(\frac{-2\pi i \nu t}{T}\right) \mathrm{d}t \,, \quad \nu \in \mathbb{Z} \tag{10.43}$$

provides the Fourier coefficients $Z(\nu)$ for the curve. The first coefficient $Z(0)$ is the mean point or centroid of the curve. The second coefficient $Z(1)$ describes a circle. Including the coefficient $Z(-1)$, a arbitrarily ellipse can be constructed. For each pair of Fourier coefficients $Z(n)$ and $Z(-n)$ an ellipse is constructed, which is run through n times. The reconstruction of the original parameter curve with the Fourier coefficients is

$$z(t) = \sum_{\nu=-\infty}^{+\infty} Z(\nu) \exp\left(\frac{-2\pi i \nu t}{T}\right) . \tag{10.44}$$

Note that, in practical applications, for appropriate results only a small number of Fourier coefficients must be calculated. Now we consider the invariance properties of this representation. A translational shift in the contour only influences the Fourier coefficient $Z(0)$. A scaling of the contour line influences all coefficients. The same holds for a rotation of the curve. An invariant representation can be constructed in three steps. (i) Drop the coefficient $Z(0)$, which gives translational invariance. (ii) Set the norm of $Z(1)$ to unity, which gives invariance for arbitrary scaling. (iii) Set all phases in relationship to the phase of $Z(1)$, which gives rotational invariance.

In summary, Fourier coefficients are a good representation of contours and moments for the total intensities of objects.

In the following sections we leave the direct description of a scalar field and discuss methods of stereography, particularly binary fields which only have the field value 0 or 1. This leads to methods of shape description, which we shall discuss later.

10.6.2 Stereography and Topology

First, we present several definitions and basic methods of stereography. Then we discuss the topological classification of structures which measures their connectivity. Furthermore this gives a motivation for the next subsection.

10.6.2.1 Stereography

To analyze a scalar field $u(\boldsymbol{x})$ where $\boldsymbol{x} \in \mathbb{R}^d$ with methods of stereography one has to generate a binary image further called a structure Q. This can be done by thresholding. One gets the excursion set

$$Q_\nu := \{\boldsymbol{x}|u(\boldsymbol{x}) \geq \nu\} \;, \tag{10.45}$$

by discriminating between regions with a higher and lower field value than the threshold ν. The boundary ∂Q_ν of the excursion set Q_ν is then obviously given by $\partial Q_\nu := \{\boldsymbol{x}|u(\boldsymbol{x}) = \nu\}$. Varying the threshold ν causes in general a variation in the monochrome image Q_ν. So the threshold can be used as a diagnostic parameter to analyze the scalar field $u(\boldsymbol{x})$.

Given a structure Q or even only a point distribution, that is a union of points which can also be understood as a structure Q, one can generate the parallel set Q_ε. By putting a ball B_ε with fixed radius ε at every point of the structure Q one gets

$$Q_\varepsilon = Q \oplus B_\varepsilon \;. \tag{10.46}$$

The sum is understood as the Minkowski sum $C = A \oplus B$ of two sets A and B, which consists of all points $\boldsymbol{c} = \boldsymbol{a}+\boldsymbol{b}$ that can be written as a vector sum of two points $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$. The corresponding difference $Q \ominus B_\varepsilon$ is called the Minkowski difference. In image processing these operations are called dilation and erosion. Again varying the radius ε causes, in general, a variation of the generated structure Q_ε. Therefore the radius ε can also be used as a diagnostic parameter.

Note that, in general, $(Q \oplus B_\varepsilon) \ominus B_\varepsilon \neq Q \neq (Q \ominus B_\varepsilon) \oplus B_\varepsilon$ holds true. Both have an effect of smoothing on a length scale ε. Closing is understood as $Q_\varepsilon^{\oplus,\ominus} = (Q \oplus B_\varepsilon) \ominus B_\varepsilon$ and opening as $Q_\varepsilon^{\ominus,\oplus} = (Q \ominus B_\varepsilon) \oplus B_\varepsilon$ where, compared to Q, the structure $Q_\varepsilon^{\oplus,\ominus}$ loses small holes and $Q_\varepsilon^{\ominus,\oplus}$ loses small cusps. As discussed in Section 10.5.3, effects of closing and opening can also be achieved by applying nonlinear and space-invariant filters on scalar fields.

Another way to get a diagnostic parameter to analyze a scalar field $u(\boldsymbol{x})$ is to apply an appropriate filter before thresholding. Then the individual filtering parameters can be used as diagnostic parameters. In practice, it is useful to restrict oneself to one diagnostic parameter which reflects the feature of interest and hold the other parameters fixed. For filter processes we refer to Sections 10.5.1–10.5.3.

10.6.2.2 Topology

A useful feature to distinguish between different structures is their connectivity. To analyze the connectivity of a structure Q we use the topological measure called the Euler Characteristic (EC), denoted by χ, which is related to the genus g by $\chi = 1 - g$. The definition we present here is not only motivated by historical reasons from set theory and convex geometry, but also provides a good access to its interpretation. For a convex body K the EC is defined by

$$\chi(K) := \begin{cases} 1 & \text{for } K \neq \emptyset \\ 0 & \text{for } K = \emptyset \end{cases} \tag{10.47}$$

and obeys the functional equation for adding two convex bodys K_1 and K_2

$$\chi(K_1 \cup K_2) = \chi(K_1) + \chi(K_2) - \chi(K_1 \cap K_2) \quad \text{and} \quad \chi(cK) = \chi(K) \;, \tag{10.48}$$

the scaling property, for scaling a convex body K with a constant positive real number $c \in \mathbb{R}^+$.

There is a demonstrative morphological interpretation of the value of the EC which is governed by the number $N(\diamond)$ of objects with the characteristic $\diamond$ in the structure Q. For 2D structures, there are $\chi(Q) = N(\text{components}) - N(\text{holes})$. Positive values are generated by isolated objects of a spot-like structure and negative values point to a mesh-like structure, where the absolute value reflects the strength. For 3D structures, there are $\chi(Q) = N(\text{components}) + N(\text{cavities}) - N(\text{tunnels})$. If there is a connected structure then positive values reflect a cheese-like structure and negative values a sponge-like structure. The absolute value again reflects the strength. Figure 10.13 illustrates the functional equation of the EC in (10.48) and its interpretation for 2D structures.

Given a smooth dD structure Q with $d > 1$ and a regular boundary ∂Q, then every point $\boldsymbol{x} \in \mathbb{R}^d$ on its hypersurface has $d - 1$ principal curvature radii $R_i(\boldsymbol{x})$ with $i = 1, \ldots, d - 1$. The local mean curvature H and Gaussian curvature G of the hypersurface are defined by

$$H := \frac{1}{d-1} \sum_{i=1}^{d-1} \frac{1}{R_i(\mathbf{x})} \quad \text{and} \quad G := \prod_{i=1}^{d-1} \frac{1}{R_i(\mathbf{x})} \,. \tag{10.49}$$

Its EC follows from the Gauss–Bonnet theorem after surface integration

$$\chi(Q) = \frac{\Gamma(d/2)}{2\pi^{d/2}} \int\limits_{\partial Q} G \, \mathrm{d}A \,, \tag{10.50}$$

where $\mathrm{d}A$ denotes the element of the hypersurface ∂Q and $\Gamma(x)$ the Γ-function. The EC of a dD excursion set Q_ν follows from the Morse theorem by counting the different stationary points of the thresholded function which lie in the excursion set. It is

$$\chi(Q_\nu) = \sum_{k=0}^{d} (-1)^k N_k(Q_\nu) \,, \tag{10.51}$$

where $N_k(Q_\nu)$ denotes the number of stationary points ($\nabla u = 0$) to the index k, where k is the number of negative eigenvalues of the matrix $\{\partial_i \partial_j u\}$ for every stationary point. For a 2D excursion set Q_ν, which was generated from a function $u(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^2$, we get $\chi(Q_\nu) = N(\text{maxima}) + N(\text{minima}) - N(\text{saddle points})$, where now $N(\diamond)$ denotes the number of stationary points of kind $\diamond$ which are in the excursion set [19].

In practice, the EC is a appropriate measure to study percolation, wetting and connectivity where only the topology is of interest. To study also the geometry of structures one needs more measures. This leads to Minkowski Functionals and their generalization, Minkowski Valuations, which will be studied in the next section.

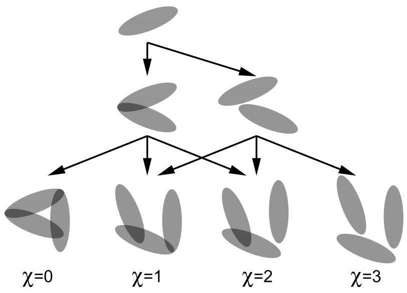


Figure 10.13 Euler Characteristic (EC). Multiple union (adding from top to bottom following the arrows) of convex bodies (light gray) illustrates the continuation of the functional equation of the EC in (10.48). The intersections are shown by dark gray coloration. Comparing the columns shows the interpretation of the EC for 2D structures, i.e. the value of the EC reflects the number of components minus the number of holes [18].

10.6.3 Minkowski Functionals and Minkowski Valuations

Minkowski Functionals (MFs) permit are to analyze the morphology, that is, the topology and geometry of structures. They map these structures on numbers or, in a generalized way, on tensors which will be discussed later in the case of Minkowski Valuations (MVs), with a well-known geometric and topological meaning. Historically, MFs were introduced in convex geometry by Hermann Minkowski and were called “Quermaßintegrale”. Appropriations in stochastic geometry and integral geometry followed [20–22].

10.6.3.1 Stochastic Geometry

If one wants to analyze the morphology of a *d*D structure, one needs appropriate measures which map the structure Q on a set of values $F(Q)$. For simplicity let these values be real numbers, then the mapping can be formulated by

$$Q \to F_j(Q) \in \mathbb{R} \quad \text{for} \quad j = 0, \ldots, n\,. \tag{10.52}$$

Because we are only interested in the morphology of the structure Q we can make a few assumptions for the mapping and, respectively, the functional $F(Q)$. Thus the number n of linear independent functionals $F_j(Q)$ may be quantified as follows.

(i) Additivity: the functional behaves as a volume, in a mathematical sense like the functional equation of the EC in (10.48),
(ii) Motion invariance: the mapping is independent of the position and orientation of the structure, that means that the functionals are independent on applying translations and rotations to the structure Q.
(iii) Conditional continuity: if a structure Q_1 continuously goes over into a structure Q_2 then also the functional $F(Q_1)$ is continuously goes over into the functional $F(Q_2)$.

One gets, from the Functional theorem of stochastic geometry which was found by Hugo Hadwiger [20], that a *d*D structure has $d+1$ linearly independent functionals which satisfy (i)–(iii). Others follow by linear combination. One full set of these descriptors are the MFs which represent intuitive parameters in common dimensions (see Table 10.3). Note that there is the freedom of scaling them by positive

Table 10.3 Geometrical and topological interpretation of the d+1 Minkowski Functionals of structures in common dimensions $d = 1, 2, 3$.

	$d = 1$	$d = 2$	$d = 3$
F_0	length	area	volume
F_1	Euler Characteristic	circumference	surface area
F_2	–	Euler Characteristic	total mean curvature
F_3	–	–	Euler Characteristic

real numbers $c \in \mathbb{R}^+$. In reasonable applications, there is a need for normalization. In any normalization the homogeneity of MF

$$F_j(cQ) = c^{d-j} F_j(Q) \quad \text{for} \quad j = 0, \ldots, d \tag{10.53}$$

of a dD structure Q holds true. This is consistent with the scaling property of the EC in (10.48).

10.6.3.2 Integral Geometry

With this interpretation in mind we can focus on an integral geometric approach which is suitable for the description of a smooth dD structure Q with $d > 1$ and a regular boundary ∂Q. This approach leads to a natural generalization of the framework by calculating higher moments. For this reason we add an extra upper index in the notation of the MFs in (10.54) to show that they are tensors of rank 0, namely scalars.

The MF of a structure Q for $j = 0$ follows by a volume integration and a set of d MFs for $j = 1, \ldots, d$ by a surface integration

$$F_j^0(Q) = \begin{cases} \mathcal{N}_0^0 \int\limits_Q \mathrm{d}V & \text{for} \quad j = 0 \\ \mathcal{N}_j^0 \int\limits_{\partial Q} S_j \, \mathrm{d}A & \text{for} \quad j = 1, \ldots, d \,. \end{cases} \tag{10.54}$$

dV denotes the hypervolume element of the dD structure Q and dA the hypersurface element of its $(d-1)$D surface. The integrands of a set of the d MFs for $j = 1, \ldots, d$ can be generated by the j-th elementary symmetric function S_j, which is defined by

$$\sum_{j=1}^{d} z^{d-j} S_j := \prod_{i=1}^{d-1} \left[z + \frac{1}{R_i(\mathbf{x})} \right] . \tag{10.55}$$

The functions S_j follow by comparing the coefficients of the polynomial in z and are functions of the $j-1$ principal curvature radii $R_{j-1}(\boldsymbol{x})$ at the position $\boldsymbol{x} \in \mathbb{R}^d$. With the definitions in (10.49) one can see that for $d = 2$ we have $S_1 = 1$ and $S_2 = G$. For $d = 3$ we have $S_1 = 1$, $S_2 = H$ and $S_3 = G$ and for $d > 3$ always $S_1 = 1$, $S_2 = H$ and $S_d = G$. Note the consistency between the statements in Table 10.3. The prefactors

$\mathcal{N}_j^0$ for $j = 0, ..., d$ are arbitrary but fixed normalization constants as explained before are caused by the freedom of normalization [22].

Compared to scalar MFs, one finds that tensor-valued MFs, further called Minkowski Valuations (MVs) to distinguish clearly between MFs, also obey additivity (i) and conditional continuity (iii). But motion invariance (ii) breaks down. Therefore, MVs obey motion covariance and the number n of linear independent functionals $F_j(Q)$ with $j = 0, ..., n$ can again be quantified. Note that, for rank $r > 1$ in d dimensions, n differs from $d + 1$ and $n \leq d + r - 1$ holds true [23].

By adding the position vector $\mathbf{x} \in \mathbb{R}^d$ in the integrands in (10.54) one gets

$$F_j^1(Q) = \begin{cases} \mathcal{N}_0^1 \int\limits_Q \mathbf{x}\, \mathrm{d}V & \text{for} \quad j = 0 \\ \mathcal{N}_j^1 \int\limits_{\partial Q} S_j\, \mathbf{x}\, \mathrm{d}A & \text{for} \quad j = 1, \ldots, d \end{cases} \tag{10.56}$$

for the first-order MVs. Due to the multiplication by a dD vector, the mapping in (10.52) now reads $Q \rightarrow F_j^1(Q) \in \mathbb{R}^d$ for $j = 0, \ldots, d$. Therefore these MVs become dD vectors, that are first-rank tensors.

As mentioned before, there is the possibility of constructing more than $d + 1$ linear independent MVs of rank $r > 1$. Second-order MVs can be constructed using

$$F_j^2(Q) = \begin{cases} \mathcal{N}_0^2 \int\limits_Q \mathbf{x}^2\, \mathrm{d}V & \text{for} \quad j = 0 \\ \mathcal{N}_j^2 \int\limits_{\partial Q} S_j\, \mathbf{x}^p \mathbf{n}^q\, \mathrm{d}A & \text{for} \quad j = 1, \ldots, d + r - 1 \end{cases} \tag{10.57}$$

where $\boldsymbol{n} \in \mathbb{R}^d$ denotes the normalized normal vector on the hypersurface ∂Q in the point $\boldsymbol{x}$ and $p, q \in \mathbb{N}_0$ with rank $r = p + q = 2$. The multiplication of the dD vectors in the integrands is understood as a symmetric tensor product, which for two dD vectors $\boldsymbol{a} = (a_1, ..., a_d)$ and $\boldsymbol{b} = (b_1, ..., b_d)$ leads to a $d \times d$ matrix with the elements $c_{ij} = a_i b_j$ $(i, j = 1, ..., d)$. The mapping in (10.52) now reads $Q \rightarrow F_j^2(Q) \in \mathbb{R}^{d \times d}$ for $j = 0, \ldots, d{+}r{-}1$. Therefore these MVs get second-rank tensors with $d \times d$ elements.

Although higher-rank tensors can be constructed, we will not consider them here. Next, we show several applications of MFs and MVs motivated by cosmological interest.

10.6.3.3 Applications

In practice, MFs and MVs turned out to be robust measures for a huge bandwidth of applications. Due to the intuitive interpretation, the individual measures can be related to some physical properties like the EC for percolation studies. Also full sets of theoretical expected values of MFs for several randomly generated structures are known [19, 21]. In cosmology, galaxy distributions were studied and compared to several Poisson point processes. Around every point a ball with radius ε was placed (see (10.46)), where the radius was varied and used as a diagnostic parameter. This technique is known as the boolean Germ–Grain-Model.

In cosmology random fields also play an important role. Density fluctuations of the very early universe as imprinted in the CMB, are assumed to be Gaussian.

To check the Gauss hypothesis, which is a fundamental aspect to allow highly-precision cosmology and is the basis for simulations, MFs are used [18, 22, 24]. MFs of a thresholded Gaussian random field Q_ν (see (10.45)) are analytically well known [19, 22] being

$$F_j^0(Q_\nu) = \begin{cases} \mathcal{N}_0^0\left[1 - \Phi\left((\nu-\mu)/\sqrt{2\sigma}\right)\right] & \text{for} \quad j = 0 \\ \mathcal{N}_j^0 H_{j-1}\left((\nu-\mu)/\sqrt{\sigma}\right) & \text{for} \quad j = 1, \ldots, d \,, \end{cases} \tag{10.58}$$

where

$$\Phi(x) = \left(2/\sqrt{\pi}\right)\int\limits_0^x \mathrm{d}t \exp(-t^2) \quad \text{and} \quad H_n(x) = \left((-1)^n/\sqrt{2\pi}\right)\left(\tfrac{\mathrm{d}}{\mathrm{d}x}\right)^n \exp\left(-x^2/2\right) \tag{10.59}$$

is the Gaussian error function $\Phi(x)$ and $H_n(x)$ the nth Hermite function. They only depend on parameters of the stochastic process, that is, the mean μ and the variance σ of the field $u(\boldsymbol{x})$ and not of the cosmological process, in particular, the cosmological parameters. Depending on the application it can be more convenient to simulate a huge set of realizations including additional numerical effects like discretization, pixelation and masking, which also experimental data is dealing with. Further, one also gets the statistical variance to perform likelihood analyses and one can define a confidence level. Applications to the CMB are shown in Figure 10.14.

Similarly, the evolution of the LSS of the universe, from its Gaussian origin to its current observed net-like structure, can be studied by MFs of the total field. Since high thresholds disentangle the cosmic web by yielding many isolated objects the concept of shapefinders, which provide measures to distinguish between different shape characteristics of single objects, is useful. An illustrative example in 2D, actually on the sphere, as used for an analysis of hot and cold spots in the CMB [18], is to quantify the elongation of structures being approximately ellipses by the ratio of their area to their squared circumference $\mathcal{E}(Q) := F_0^0(Q)/[F_1^0(Q)]^2$. This is a dimensionless, and thereby scale-invariant, shapefinder $\mathcal{E}$, which immediately provides an axis ratio for expected shapes.

Let us come back to the example of the LSS in 3D. Three independent ratios of MFs which have the dimension of length, namely thickness $\mathcal{T}$, width $\mathcal{W}$ and length $\mathcal{L}$, can be defined:

$$\mathcal{T}(Q) := \frac{F_0^0(Q)}{2F_1^0(Q)}\,, \quad \mathcal{W}(Q) := \frac{2F_1^0(Q)}{\pi F_2^0(Q)} \quad \text{and} \quad \mathcal{L}(Q) := \frac{3F_2^0(Q)}{4F_3^0(Q)}\,. \tag{10.60}$$

An appropriate normalization is $\mathcal{N}_0^0 \equiv 1$, $\mathcal{N}_1^0 \equiv 1/6$, $\mathcal{N}_2^0 \equiv 1/(6\pi)$ and $\mathcal{N}_3^0 \equiv 1/(2\pi)$. Then for every convex body K the property $\mathcal{L}(K) \geq \mathcal{W}(K) \geq \mathcal{T}(K)$ holds true. Further dimensionless shapefinders called the

$$\text{planarity} \quad \mathcal{P}(Q) := \frac{\mathcal{W}-\mathcal{T}}{\mathcal{W}+\mathcal{T}} \quad \text{and filamentarity} \quad \mathcal{F}(Q) := \frac{\mathcal{L}-\mathcal{W}}{\mathcal{L}+\mathcal{W}}$$

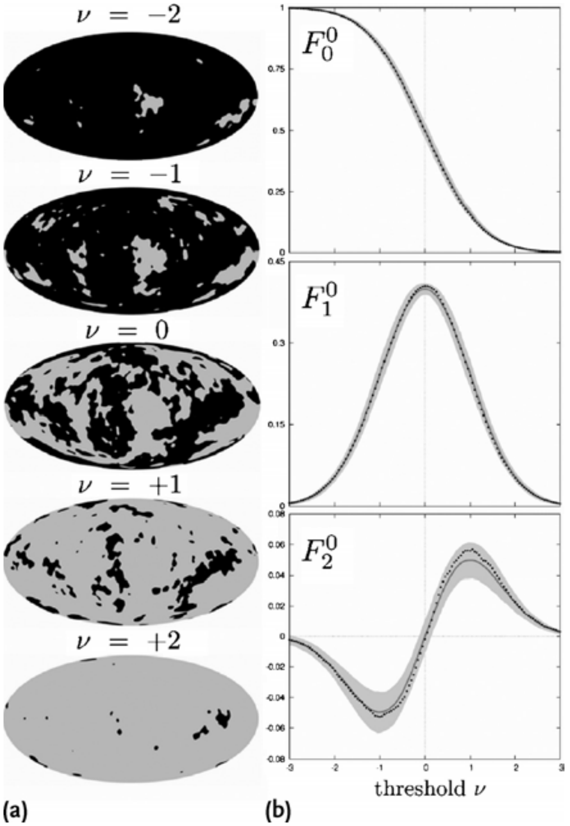


Figure 10.14 Minkowski Functionals (MFs). Test on Gaussianity of the fluctuations of the cosmic microwave background (CMB) with Minkowski Functionals (MFs). The normalized signal became thresholded (as illustrated in (a), where the analyzed structure is the black area). The corresponding normalized MFs F_0^0, F_1^0 and F_2^0 are calculated (black dots in graphs in (b)) and compared to the mean values (dark gray line) and statistical variances (3σ as light gray area) of a Gaussian random field. The area functional F_0^0 is equivalent to the cumulative normalized height distribution and therefore drops monotonously. By increasing the threshold, holes appear, get bigger and become connected. Isolated spots appear, get smaller and finally vanish. Therefore the length functional F_1^0 increases, reaches a maximum and decreases again. Negative values of the connectivity functional F_2^0 reflect the appearance of holes, and positive values the appearance of isolated spots, where in the intermediate both balance each other. With these results likelihood analysis showed a high confidence level for the CMB being a Gaussian random field [18].

$$\tag{10.61}$$

can be defined, providing appropriate measures to discriminate between LSS of different evolution scenarios of physical modeling and their comparison to real data. Their behavior is demonstrated in Figure 10.15. These shapefinders also provide the possibility of categorizing observed galaxy shapes. A real galaxy differs from a sphere by being an oblate spheroid, (like a pancake) known as the family of spiral

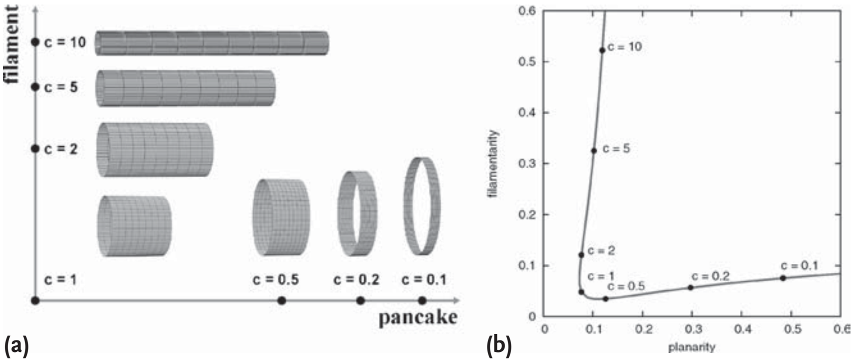


Figure 10.15 Shapefinders:. Different cylinders as test bodies (a). For illustration only their coat is shown. The ratio of their height to their diameter is denoted by c. Varying c from zero to infinity causes a transition from a pancake to a filament (gray arrows). Note the invariance of scaling. The scatter plot (b) of the corresponding shapefinders planarity $\mathcal{P}$ and filamentarity $\mathcal{F}$ (black points). The results of varying c from zero to infinity continuously are shown as a gray line. Building a pancake ($c \to 0$) the value of the shapefinder $\mathcal{F}$ stays stable, but $\mathcal{P}$ increases. On the other hand, by building a filament ($c \to \infty$) the value of the shapefinder $\mathcal{P}$ stays stable, but now $\mathcal{F}$ increases.

galaxies, or being a prolate spheroid, (like a cigar) known as the family of elliptical galaxies [22].

MFs are useful for analyzing structures of a stochastic origin or single objects with $\chi = 1$. Given more than one single object, MFs can have identical values for different structures (compare (5) and (6) in Figure 10.16). When the relative position of partial structures is important, as in analysis of the inner structure of galaxies [25], galaxy clusters or galaxy superclusters [26], one can use MVs for a reasonable description. Due to the concept of center of mass and moments of inertia, known from mechanics, a possible interpretation becomes obvious [23]. First-order MVs can be interpreted as the geometrical center of the scalar measure and one can define curvature centroids

$$\mathbf{p}_j(Q) = F_j^1(Q)/F_j^0(Q) \quad \text{with} \quad \mathcal{N}_j^0 \equiv \mathcal{N}_j^1 \equiv 1 \quad \text{for } j = 0, \ldots, d \,. \tag{10.62}$$

To fulfill the mentioned interpretation of Second-order MVs one considers only a subset of (10.57) with $r = p = 2$ and executes an appropriate transformation. Then one gets the elements

$$\left\{\mathbf{P}_j(Q)\right\}_{kl} = \begin{cases} \int\limits_Q (d_{\mathbf{p}_j}^2 \delta_{kl} - x_k x_l)\ \mathrm{d}V & \text{for} \quad j = 0 \\ \int\limits_{\partial Q} S_j (d_{\mathbf{p}_j}^2 \delta_{kl} - x_k x_l)\ \mathrm{d}A & \text{for} \quad j = 1, \ldots, d \end{cases} \tag{10.63}$$

of the curvature tensors $\mathbf{P}_j(Q)$ for every corresponding scalar measure. $d_{\mathbf{p}_j}^2$ is the distance to the corresponding curvature centroid $\mathbf{p}_j$, which is used as the origin of the coordinate system, and δ_{kl} is the Kronecker δ. These tensors satisfy the eigenvalue equation

$$\mathbf{P}_j(Q)\mathbf{v}_j^m(Q) = \lambda_j^m(Q)\mathbf{v}_j^m(Q) \quad \text{for } j = 0, \ldots, d \quad \text{and} \quad m = 1, \ldots, d \tag{10.64}$$

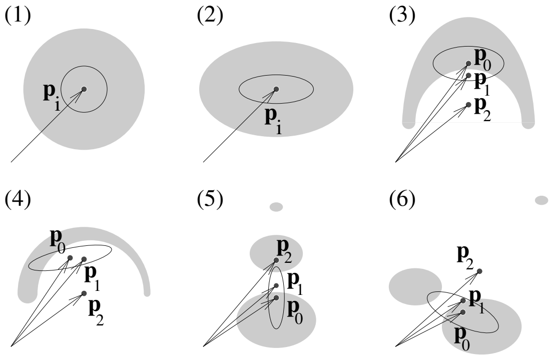


Figure 10.16 Minkowski Valuations (MVs). Some simple structures (gray), their curvature centroids $\mathbf{p}_i$ with $i = 0, 1, 2$ and their area tensor $\mathbf{P}_0$ (black ellipse, which indicates the ratio of eigenvalues and eigendirections). A circle (1) and an ellipse (2) with the same volume. Because of point symmetry, all centroids coincide, but due to orientation and elongation, the tensors differ. Given an axis-symmetric structure which is not point-symmetric (3) the centroids no longer coincide, but still remain in one straight line, which indicates, as well as the tensor, the symmetry axis. Breaking additional axis symmetry is shown in (4). The scalar Minkowski Functional cannot distinguish between structure (5) and (6). Tensor Minkowski Valuations on the other hand can. (Reprinted Figure 5 from [23] with kind permission of Springer Science and Business Media. We thank the author Claus Beisbart for providing the data for reproduction.).

with $\lambda_j^m(Q) \in \mathbb{R}$, $\mathbf{v}_j^m(Q) \in \mathbb{R}^d$ and $\mathbf{v}_j^m(Q) \cdot \mathbf{v}_j^n(Q) = 0$ for $m \neq n$. Thus the scalar measures of the structure Q were vectorized, which means parameterized, by the orientation and strength in orthogonal directions along the direction of the eigenvectors $\mathbf{v}_i^m(Q)$ and their corresponding eigenvalues $\lambda_j^m(Q)$. Figure 10.16 illustrates the power of the shape description with MVs on a number of simple 2D structures.

These measures immediately serve as descriptors or, adjusted to the application combinations, as in the case of MFs in the concept of shapefinders, are more advantageous. Let us restate the task of quantifying the elongation of structures being approximately ellipses. For an ellipse E, where a and b denote the two semi-axes, $a/b = \sqrt{\lambda_0^1(E)/\lambda_0^2(E)}$ when $a \geq b$ then $\lambda_0^1(E) \geq \lambda_0^2(E)$ holds true. Again an appropriate shapefinder is found. Combining this with the one stated before in the case of MFs, it even provides the possibility of defining quality measures for expected shapes [18].

10.7 Concluding Remarks

This work was initiated by the observation that, in both research areas, namely Computer Vision and Physical Cosmology similar tasks in *image analysis* are employed. Therefore, we have highlighted several methodological approaches concerning topics of Computer Vision and Physical Cosmology in the field of image processing. The aim was to give some examples that show how different disciplines arrive at related approaches that can be considered at a more systemic level of a classical processing hierarchy.

Motivated by the recording of the cosmic microwave background (CMB) on the celestial sphere, we started with projection methods, in particular, the Mollweide projection (see Figure 10.2). Next, we discussed the representation of images and characterized four main properties, namely the space an image can be defined on, the quantization, the resolution and the scale space of intensities. Scale spaces are known from physics, but here the properties of scale spaces were adapted to methods of image processing according to [11].

Images and their characteristics were not only defined in the plane, but also on arbitrary surfaces (see Figure 10.3a). For example, the two-point correlation function was defined on the sphere and, in addition, the angular power spectrum. This angular power spectrum is a measure used to compare the expected CMB for different models of the Universe with the measured CMB (see Figure 10.6).

After consideration of image characteristics, the basic methods in image processing, modeled by a filtering process, were discussed. Besides the analysis of simple filters like the Gaussian or the Laplacian of the Gaussian, we additionally discussed the Gabor filter and specified a scheme for the construction of a Gabor filter bank (see Figure 10.10). On the basis of partial derivatives of image intensities an intensity gradient was constructed. This gradient denotes the steepest gray-value ramp. A structure tensor based on the gradient was defined, containing information of the local gray-value distribution. For the 2D case, an interpretation of this structure tensor was given (see Table 10.2).

In the last part we focused on invariant descriptions of image features, where invariance was restricted to scaling, translation and rotation. First, statistical moments for continuously valued images and descriptions of contour lines were described. Therefore, the contour lines were assumed to be defined as periodic curves and a representation based on Fourier coefficients was employed. Then we changed to thresholded images, denoted by binary structures which could be characterized by the Euler Characteristic as a simple measure. But, for the analysis of the topology and geometry of structures, this measure was not sufficient, therefore scalar Minkowski Functionals were considered. These measures are suitable to analyze structures initiated by a random process, like investigating the statistical properties of the CMB (see Figure 10.14) or for shape recognition of single objects by means of shapefinders (see Figure 10.15). Also scalar Minkowski Functionals do not reveal the full description of structures as explained in Figure 10.16. Thus, we introduced tensor Minkowski Valuations offering a more detailed analysis of structures but also leaving motion invariance (translation and rotation).

In summary, we have outlined the primal methods and general concept of *image analysis*. Generally, these methods are located within the first steps of an image-processing hierarchy, providing image enhancement and basic feature extraction – normally seen as low- and mid-level vision tasks. Based on these first steps, further methods like 3D scene reconstruction, optic flow estimation, or classification and identification, can be realized. In particular, for the last two tasks the given invariant descriptions are essential. In conclusion, basic methods and general concepts will also be useful in other research areas besides image processing, but so far, a starting point has been given for an exchange of problems and existing solutions among researchers studying our observable world.

References

1 Ballard, D.H. and Brown, C.M. (**1982**) *Computer Vision*, Prentice-Hall

2 Forsyth, D.A. and Ponce, J. (**2003**) *Computer Vision – A Modern Approach*, 3rd ed, Pearson Education International, Upper Saddle River, New Jersey.

3 Ullman, S. (**1996**) *High-Level Vision*, MIT Press, Cambridge, Massachusetts.

4 Witkin, A.P. and Tenenbaum, J.S. (**1983**) On the role of structure in vision, in *Human and Machine Vision* (eds J. Beck, B. Hope and A. Rosenfeld), Academic Press.

5 Faugeras, O. (**1993**) *Three-dimensional Computer Vision*, MIT Press, Cambridge, Massachusetts.

6 Einstein, A. (**1917**) *Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie*. Sitzungsberichte der Preußischen Akademie der Wissenschaften 1917 S. 142–152 (also in: *The Collected Papers of Albert Einstein* Bd. 6 S. 540ff Princeton University Press 1996).

7 Spergel D.N. *et al.* (**2007**) Three-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Implications for cosmology. *The Astrophysical Journal Supplement Series* **170**, 377–408.

8 Górski, K.M., Hivon, E., Banday, A.J., Wandelt, B.D., Hansen, F.K., Reinecke, M. and Bartelmann, M. (**2005**) HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *The Astrophysical Journal* **622**, 759–771.

9 Gomes, J., Darsa, L., Costa, B. and Velho, L. (**1999**) *Warping and Morphing of Graphical Objects*, Morgan Kaufmann Publishers, San Francisco.

10 Hilbert, D., Cohn-Vossen, S. (**1952**) *Geomtery and Imagination*, Chelsea Publ. Company, New York.

11 Weickert, J. (**1998**) *Anisotrope Diffusion in Image Processing*, Teubner Verlag, Stuttgart.

12 Horn, B.K.P. (**1986**) *Robot Vision*, MIT Press, Cambridge, Massachusetts.

13 Trucco, E. and Verri, A. (**1998**) *Introductory Techniques for 3-D Computer Vision*, Prentice-Hall, Upper Saddle River, New Jersey.

14 Aurich, R., Janzer, H.S., Lustig, S. and Steiner, F. (**2008**) *Do we live in a 'small Universe'?*. Class. Quantum Grav. 25, 125006, (12pp).

15 Fischer, B. and Modersitzki, J. (**2006**) Image Fusion and Registration – a Variational Approach. In *Computational Science and High Performance Computing II, Notes on Numerical Fluid Mechanics and Multidisciplinatory Design*, Springer, 193–205.

16 Jähne, B. (**2005**) *Digital Image Processing*, 6th edn, Springer Verlag Berlin, Heidelberg, New York.

17 Förstner, W. (**1986**) A feature based correspondence algorithm for image matching. *ISP Comm. III, Rovaniemi 1986, International Archives of Photogrammetry*, 26-3/3.

18 Janzer, H.S. (**2006**) *Untersuchung der kosmischen Hintergrundstrahlung mit Minkowskifunktionalen*. Diploma thesis, Universität Ulm.

19 Tomita, H. (**1990**) Statistics and Geometry of Random Interface Systems. In *Formation, Dynamics and Statistics of Patterns* (Kawasaki, K., Suzuki, M., and Onuki, A. eds), volume 1, World Sientific.

20 Hadwiger, H. (**1957**) *Vorlesungen über Inhalt, Oberfläche und Isoperimetrie*. Springer.

21 Mecke, K.R. (**1996**) *Integralgeometrie in der statistischen Physik: Perkolation, komplexe Flüssigkeiten und die Struktur des Universums*. Verlag Harri Deutsch.

22 Schmalzing, J. (**1999**) *On Statistics and Dynamics of Cosmic Structure*. Doctoral dissertation, LMU München.

23 Beisbart, C., Dahlke, R., Mecke, K. and Wagner, H. (**2002**) Vector- and Tensor-valued Descriptors for Spatial Patterns, in *Morphology of Condensed Matter. Physics and Geometry of Spatial Complex Systems* (eds K. Mecke and D. Stoyan), Springer.

24 Eriksen, H.K., Novikov, D.I., Lilje, P.B., Banday, A.J. and Gorski, K.M. (**2004**) Testing for Non-Gaussianity in the Wilkinson Microwave Anisotropy Probe Data: Minkowski Functionals and the Length of the Skeleton. *Astrophys. J.* **612**, 64–80.

25 Rahman, N. and Shandarin, S.F. (**2003**) Measuring shapes of galaxy images – I. Ellipticity and orientation. *Monthly Notices of the Royal Astronomical Society* **343**(3), 933–948.

26 Beisbart, C., Buchert, T. and Wagner, H. (**2001**) Morphometry of spatial patterns. *Physica A* **293**, 592–604.

11
Boosting Ensembles of Weak Classifiers in High Dimensional Input Spaces

Ludwig Lausser, Friedhelm Schwenker[1], *Hans A. Kestler*[1]

11.1
Introduction

Classification is a fundamental method of data analysis, it has numerous applications in many disciplines. Categorizing of objects, situations, and other data into classes is a task required in a variety of areas; for example, medical diagnosis and human-machine interfaces. Humans accomplish this categorization quite easily in everyday situations such as classifying food, traffic signs, music styles and many other things. In biological and technical systems the following processing steps will always appear during classification. Collecting some data about an object, choosing some criteria (features) to judge it, and finally, predicting a class label for the object. In the simplest case this label only indicates whether or not the object belongs to a certain class. In this case the class labels are in the set $Y = \{+1, -1\}$. A function, which assigns such class labels to some data is called a concept. In general, the true concept of a class is unknown and has to be approximated. This approximating concept is called a hypothesis or classifier. Normally it is not clear which hypothesis should be used. In general, a class of parameterized functions (hypothesis space $\boldsymbol{H}$) has to be chosen and the parameters of the concrete hypothesis have to be determined during the learning process. This kind of learning is called supervised learning. Here the learning algorithm (learner) builds a hypothesis after receiving some correctly labeled examples from an external source (teacher). For complex pattern recognition tasks it might be difficult to get a highly accurate hypothesis from a given hypothesis space and learning algorithm. In such scenarios a single hypothesis might be insufficient, if it does not predict the class labels accurately. In this chapter a general method called Boosting is presented, which combines ensembles of weak hypotheses to a highly accurate (strong) hypothesis.

[1] Corresponding authors.

[2] It is assumed that hypotheses and concepts are measurable functions and fullfil some fairly weak measurability conditions [4].

11.2 Hypothesis Boosting Problem

The idea of hypothesis boosting was first introduced by Kearns and Valiant [1, 2] and was formulated for the distribution-free or probability approximately correct (PAC) learning model [3]. This model deals with the learning of target concepts c. Concepts are Boolean functions on some domain of instances X. A concept class $\mathbf{C}$ is a set of concepts, which sometimes can be divided into subclasses $\mathbf{C}_n$ by a parameter n ($\mathbf{C} = \cup_{n\geq 1}\mathbf{C}_n$). Concepts in $\mathbf{C}_n$ have the same domain X_n. It is assumed that encodings for instances of X_n are bounded by a polynomial in n. Given concept $c \in C_n$, a tupel (x, y) of an instance $x \in X_n$ and a Boolean label $y = c(x)$ is called an example. The source of such examples is the example oracle EX. Instances of X_n are drawn independently from an arbitrarily fixed probability distribution $\boldsymbol{D}$ which is typically unknown. According to such collected examples, a learning algorithm computes an hypothesis h as an approximation to a concept c (Figure 11.1). The prediction error[2] (also called the generalization error) of a hypothesis can be computed by

$$Pr_{\boldsymbol{D}}[h(x) \neq c(x)] \ .$$

Hypothesis h is called ε-close to concept c, if the prediction error is smaller than ε. Using this vocabulary we can define the main terms of the Boosting Hypothesis Problem.

Definition 11.1 (Strongly learnable [5])
A concept class $\mathbf{C}$ is (strongly) learnable, if there exists an algorithm A such that for all $n \geq 1$, for all target concepts $c \in \mathbf{C}_n$, for all distributions $\boldsymbol{D}$ on X_n and for all $0 < \varepsilon, \delta \leq 1$, an algorithm A, given parameters n, ε, δ, the size[3]s of c and access to oracle EX, runs in time polynomial in n, s, $1/\varepsilon$ and $1/\delta$, and outputs a hypothesis h that with probability at least $1 - \delta$ is ε-close to c under $\boldsymbol{D}$.

Definition 11.2 (Weakly learnable [6])
A concept class $\mathbf{C}$ is weakly learnable if there exists a polynomial p and an algorithm A such that for all $n \geq 1$, for all target concepts $c \in \mathbf{C}_n$, for all distributions $\boldsymbol{D}$ on X_n, and for all $0 < \delta \leq 1$, algorithm A, given parameters n,δ, the size s of c, and access to oracle EX, runs in time polynomial in n, s and $1/\delta$, and outputs a hypothesis h that with probability at least $1 - \delta$ is $(1/2 - p(n, s)^{-1})$-close to c under $\boldsymbol{D}$.

The main difference between these two definitions is the error which a learned hypothesis is allowed to have. A concept class $\boldsymbol{C}$ will be called strongly learnable if there is an algorithm which creates an ε-close hypothesis. It will be called weakly

[3] The size of a concept is a measure of the length of cs representation, for example in case of Boolean formulas it is the shortest Boolean formula computing c.

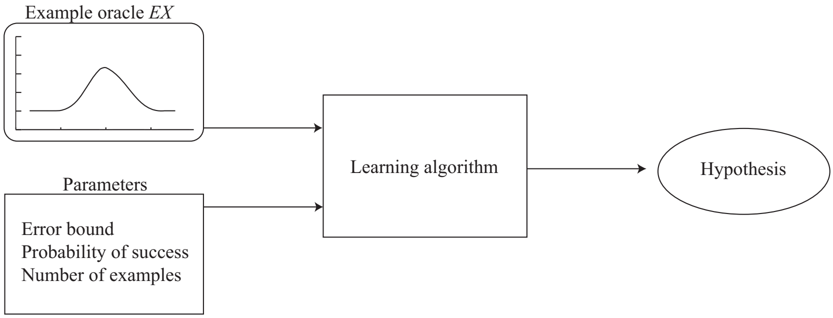


Figure 11.1 General scheme of learning used in this chapter. The learning algorithm accesses an example oracle *EX*. The oracle chooses training examples arbitrarily from an unknown distribution. The learning algorithm produces a hypothesis according to these examples and other training parameters.

learnable if there is an algorithm which performs slightly better on it than random guessing (prediction error slightly smaller than 0.5 in the two-class case). According to their performance, learning algorithms are separated into strong and weak learning algorithms.

The Hypothesis Boosting Problem is formulated by Kearns [1] as follows. "Is it the case that any C that is weakly learnable is in fact strongly learnable?"

A spectacular answer to this question was given 1990 by Schapire [5].

Theorem 11.1 ([5])
A concept class C is weakly learnable if and only if it is strongly learnable.

Since the equality of weak and strong learnability was proven in 1990, many different boosting algorithms were proposed. Although these algorithms differ in detail, they have much in common. A boosting algorithm is a meta learning algorithm, which is not able to generate hypotheses on its own. It needs access to a weak learning algorithm, which will be called WeakLearn in this context. A boosting algorithm calls WeakLearn many times and generates an ensemble of weak hypotheses. Each weak hypothesis is trained to counterbalance the misclassifications of its predecessor. The examples for a training set are chosen according to a distribution, which represents how difficult the single example is classified according to the previous weak hypothesis. The final strong hypothesis of a boosting algorithm is a combination of the weak hypotheses.

11.3 Learn

The proof of the important Theorem 11.1 is constructive. In it Schapire [5] describes an algorithm, which uses a weak learning algorithm to generate a hypothesis with high accuracy. In this context this algorithm will be called Learn in order to distinguish it from other approaches. A pseudocode description of Learn is given in Figure 11.7. The algorithm gets access to the example oracle *EX* and an weak learning algorithm WeakLearn(δ,*EX*) which outputs a $(1/2 - p(n,s)^{-1})$-close hypothesis to the target concept c with a probability of at least $1-\delta$. The method described by Schapire combines three hypotheses h_1, h_2, h_3 through a majority vote. Thereby a single hypothesis h_i is either built by a call of WeakLearn or a recursive call of Learn according to the performance which the final hypothesis should have. This leads to some kind of tree structured learning scheme (Figure 11.2). The basic idea of Learn could be seen in the training of a single stage, where for each classifier a different example oracle EX_i is used. While the first classifier h_1 is trained on EX_1, a direct copy of *EX*, h_2 receives examples from EX_2. 50% of the examples generated by EX_2 are examples which are misclassified by h_1. The oracle EX_3 for h_3 only returns examples where h_1 and h_2 mismatch. In this way the original distribution of the examples is modified, and h_2 will be confronted with more difficult examples than h_1, and hypothesis h_3 specializes in tie breaking. If the original hypothesis h_1 of stage produces an error α, the error ε of the stage can be bound by

$$g(\alpha) = 3\alpha^2 - 2\alpha^3 \ .$$

By using the inverse function $g^{-1}(\varepsilon)$ the maximal tolerable error of a single hypothesis can be determined.

The algorithm of Schapire described above was not widely used in real applications. The reason is the large amount of examples needed for training the weak

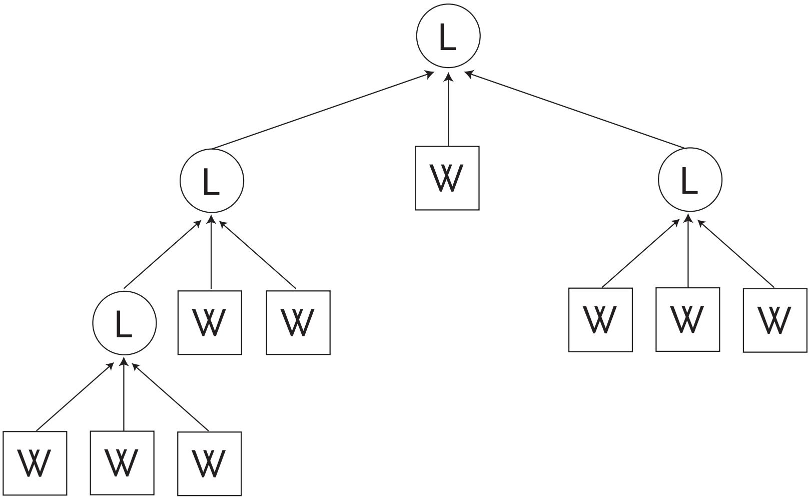


Figure 11.2 Schematic view of Learn. The procedure Learn returns a hypothesis, which is a majority vote of three single hypothesis h_1, h_2, h_3. Each single hypothesis can be either built by the weak learning algorithm WeakLearn (denoted as W) or by a recursive call of Learn (denoted as L). The three single hypothesis are trained on examples received from different example oracles EX_i. While EX_1 returns random examples, 50% of the examples returned by EX_2 are misclassified by h_1. EX_3 returns only examples, which are classified in different ways by h_1 and h_2.

hypotheses. The example oracles, as they are used here, filter examples from a potentially endless input stream. As the complexity of a hypothesis grows, more and more conditions have to be satisfied until an example is chosen by an oracle.

11.4 Boosting by Majority

Another boosting algorithm called Boosting by Majority (BBM) (Figure 11.8) was suggested by Freund [6]. It is a direct derivation from a game introduced by Freund (Figure 11.3) called the majority vote game . The game is played on a board containing a set of fields F. One player, the weightor, chooses weights w_i for each field ($w_i \geq 0$ for all i and $\sum_{i=1}^{|F|} w_i = 1$). These weights are unknown to the second player, the chooser, who has to select a subset $U \subseteq F$ such that $\sum_{F(i) \in U} w_i \geq 1/2 + \gamma$. After this step the fields of U are marked. The game is played for as many rounds as the weightor wants to play. The weightor's reward is the number of fields which were marked in more than 50% of the rounds. The goal of the weightor is to maximize his reward.

Freund derived an optimal weighting strategy for the weightor and used it to develop the BBM algorithm. The BBM version described here is a *boosting by*

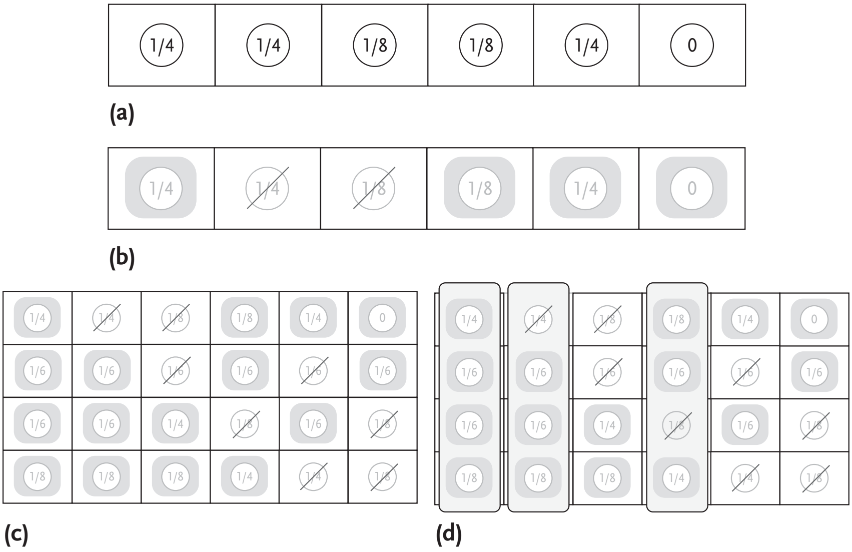


Figure 11.3 Majority vote game. The majority vote game is played by two players, the weightor and the chooser. First the weightor assigns some weights w_j to each fields of the board game (a) ($w_j \geq 0$ and $\sum w_j = 1$). Those weights are unknown to the chooser. In a second step the chooser selects some fields, until the sum of their weights is larger than $1/2 + \gamma$. In this figure the chosen fields are marked by a box (b). These steps will be repeated until the weightor decides to stop (c). The reward of the weightor is the number of fields, which have been marked in more than 50% of all cases. In this example the reward is three (d).

resampling algorithm in contrast to the *boosting by filtering algorithm* Learn described above. Here the example oracle *EX* is only used once to create an initial sample $S = (S_1, \dots, S_N)$, $S_j = \left(x_j, y_j\right)$. For each example S_j in S there exists a weight $D_j := D\left(x_j\right)$ describing the "difficulty" of S_j. The whole vector $\boldsymbol{D}$ describes a distribution over S. According to $\boldsymbol{D}$, subsamples of S are composed to train the single weak hypothesis by calling the subroutine FiltEX. The role of the BBM is the role of the weightor in the majority vote game. WeakLearn is the chooser which has to achieve an accuracy of at least $1/2 + \gamma$ on all N training examples. The number of used weak hypotheses k and the strategy how to adapt the weights D_j are derived from the majority vote game setting. The final hypothesis h_M is then the unweighted majority vote of all weak hypotheses.

Although the BBM was much more practical to use than its antecessor, it was not often applied to real data problems. The problem of this algorithm was the parameter γ, which influences the reweighting strategy of distribution $\boldsymbol{D}$. Parameter γ has to be an upper error bound of all hypotheses, otherwise BBM will fail. An adequate γ is hard to find or even unknown.

11.5 AdaBoost

The most popular boosting algorithm, called AdaBoost (= Adaptive Boosting) was introduced by Freund and Schapire [7] in 1995. A pseudocode description of AdaBoost for the two-class classification problem is given in Figure 11.9. The algorithm produces a threshold classifier, which simply computes the weighted sum of the weak classifiers' output. In each iteration t a weak hypothesis h_t is calculated by the chosen algorithm WeakLearn with respect to the training sample S and the distribution $\boldsymbol{D}_t$ (Figure 11.4). As for the BBM, the weighted error ε_t is calculated, but it is used in a different manner by the AdaBoost algorithm. First the $\boldsymbol{D}_{t+1}$ is adapted directly with respect to ε_t. In this way no concrete assumptions about the weak classifiers' accuracy are needed and the parameter γ of the BBM can be dropped. AdaBoost also combines the trained weak hypotheses h_t with respect to ε_t in order to create a weighted majority vote, this combined hypothesis is denoted by h_f. Adapting on the performance of the single weak classifier was the very new feature of AdaBoost which entitles the algorithm. Because of its simplicity AdaBoost has been used in many applications and became the basis of many new machine-learning algorithms (see the paper by Freund and Schapire [8] for an overview).

AdaBoost can be implemented as a *boosting by resampling* algorithm. But there is also a technique called *boosting by reweighting* which can be applied. It can only be chosen, if the selected WeakLearn algorithm is able to handle weighted training errors on its own. WeakLearn has this ability when the hypothesis produced by it is more likely to classify an example correctly if the example's weight is high. In this technique the whole sample S is used as training data and the single examples are

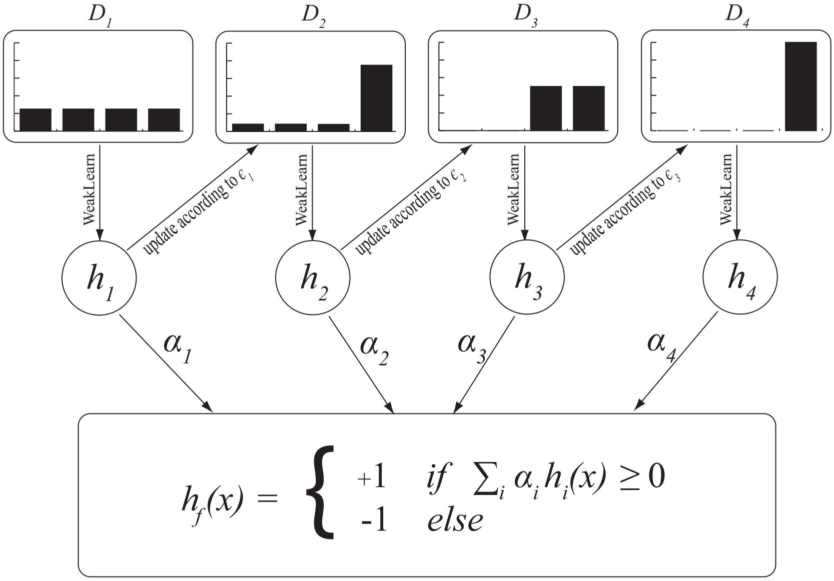

D_1
D_2
D_3
D_4
WeakLearn
WeakLearn
WeakLearn
WeakLearn
update according to ϵ_1
update according to ϵ_2
update according to ϵ_3
h_1
h_2
h_3
h_4
α_1
α_2
α_3
α_4
$h_f(x) = \begin{cases} +1 & if\ \sum_i \alpha_i h_i(x) \geq 0 \\ -1 & else \end{cases}$

Figure 11.4 Schematic view of AdaBoost which iteratively trains weak hypothesis h_i by using a weak learning algorithm WeakLearn. The single examples in the training sample are weighted after the distribution $\boldsymbol{D}_i$. The distribution $\boldsymbol{D}_{i+1}$ is an update of $\boldsymbol{D}_i$ according to the weighted training error ε_i of h_i. Examples, which are misclassifed by h_i will receive a higher weight in $\boldsymbol{D}_{i+1}$. The error ε_i also determines an weight α_i, which can be seen as the influence of h_i on the final weighted majority vote h_f.

weighted according to the distribution $\boldsymbol{D}_t$. In this way the weak hypothesis should be able to classify some hard learnable data correctly.

In the next sections, some theoretical bounds of the training and generalization error are reviewed.

11.5.1 Training Sample Error

The training error of a hypothesis $h(x)$ is the empirical error measured on the training sample. It is denoted by

$$\widehat{Pr}\,[h(x) \neq y] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{[h(x_i) \neq y_i]} ,$$

where $\mathbb{1}$ is the indicator function. Freund and Schapire [7] derived an upper bound for the training error of the hypothesis h_f built by AdaBoost. It depends on the weighted training errors ε_t of the ensemble's hypotheses h_t (see Figure 11.9) and the total number of AdaBoost iterations T

$$\widehat{Pr}\left[h_f(x) \neq y\right] \leq \prod_{t=1}^{T} \left[2\sqrt{\varepsilon_t\,(1-\varepsilon_t)}\right] .$$

If $\varepsilon_t < 1/2$ for all t, the AdaBoost algorithm decreases the overall training error of h_f exponentially fast to zero.

11.5.2 Generalization Error

The generalization error $Pr_{\boldsymbol{B}}(h(x) \neq y)$ is the probability that a hypothesis h misclassifies unseen examples with respect to distribution $\boldsymbol{B}$ on sample space X. Using techniques from structural risk minimization [9] the generalization error can be bounded with high probability (see Freund and Schapire [7]) by

$$Pr_{\boldsymbol{B}}(h_f(x) \neq y) \leq \widehat{Pr}\left[h_f(x) \neq y\right] + O\left(\sqrt{\frac{Td}{N}}\right) .$$

This bound depends on T, the number of weak hypotheses, d the VC-dimension [9] of the hypothesis space $\boldsymbol{H}$, and the sample size N. Although this bound increases with T and therefore overfitting should be observed quite often, some numerical studies report a different behavior of AdaBoost. In these studies AdaBoost has decreased the generalization error, even if the training error has dropped to zero after some iterations [10–14].

As a result of examining this phenomenon Schapire *et al.* found another way of bounding the generalization error of AdaBoost [15]. This bound depends on the margins of the single training examples. The margin of a training example (x, y) with respect to $\boldsymbol{\alpha}^t$ is defined as

$$\varrho(x, y, \boldsymbol{\alpha}^t) = \frac{y \sum_{i=1}^{t} \alpha_i h_i(x)}{\sum_i |\alpha_i|} .$$

Here $\boldsymbol{\alpha}^t$ is the vector of the t single weights α_i assigned to the weak classifiers h_i by the AdaBoost algorithm. The margin lies in the interval $[-1, +1]$ and can be used as a confidence measure for the quality of the single prediction. A high magnitude therefor shows high confidence in the prediction. Because the margin depends on the label y of x, it is positive if the example is classified correctly and negative otherwise. We will also denote the margin as $\varrho(x, y)$ if its clear which $\boldsymbol{\alpha}^t$ is meant. The generalization error can be bounded with high probability by the margin as follows:

$$\widehat{Pr}\left[\varrho(x, y) \leq \theta\right] + O\left(\sqrt{\frac{d}{N\theta^2}}\right) .$$

The term $\widehat{Pr}\left[\varrho(x, y) \leq \theta\right]$ denotes the empirical margin error measured on the training sample

$$\widehat{Pr}\left[\varrho(x, y) \leq \theta\right] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\left[\varrho(x_i, y_i, \alpha^t) \leq \theta\right]} .$$

Here $\theta > 0$ is an arbitrary threshold which can be seen as a minimum confidence value which must be achieved for correct classification. The main insight for this bound is that the generalization error of AdaBoost can be diminished by maximizing the margins of the single training examples.

Schapire *et al.* [15] show, that Adaboost is a rapid method for maximizing the margin. They show that the margin error can be bound in terms of the total number of iterations T and weighted training errors ε_t of the AdaBoost algorithm for any θ

$$\widehat{Pr}\left[\varrho(x,y) \leq \theta\right] \leq \prod_{t=1}^{T}\left[2\sqrt{\varepsilon_t^{1-\theta}\,(1-\varepsilon_t)^{1+\theta}}\right] .$$

So also after the training error drops to zero, ϱ can further be decreased by adding an h_t with $\varepsilon_t < 1/2$ and the generalization error will further decrease. Note that this argument does not prevent AdaBoost from overfitting behavior, it only delays this effect sometimes.

11.5.3 AdaBoost on Noisy Data

On noisy data AdaBoost does not perform as well as described in Section 11.5.2. Data is called noisy if some of the shown training examples are corrupted, for instance, if some examples were labeled incorrectly or single features were measured wrongly. Here we will talk of a noisy example, if it has a wrong label. A noisy data sample contains a certain amount of noisy examples.

Trained on a noisy data sample, the classifier generated by AdaBoost tends to overfit. This effect was investigated in several empirical studies (e.g. [16, 17]) and theoretical work (e.g. [18]). The experiments showed that the generalization ability of an AdaBoost classifier decreases, as the fraction of noisy examples in the training sample rises. This is not a phenomenon, which is only observed for AdaBoost classifiers. But, according to some comparative studies, AdaBoost suffers more from this problem than the other ensemble methods [16].

One explanation for this can be found in AdaBoost's selection strategy for combining the weights α_t and adapting distributions $\boldsymbol{D}_t$. Breiman [19] has shown, that AdaBoost minimizes a functional G in each iteration t which depends on $\boldsymbol{\alpha}^t = (\alpha_1, \cdots, \alpha_t)^T$:

$$G(\alpha_t, \boldsymbol{\alpha}^{t-1}) = \sum_{i=1}^{N} e^{-\phi(x_i, y_i, \boldsymbol{\alpha}^t)}$$

Here $\phi(x_i, y_i, \boldsymbol{\alpha}^t)$ is an unnormalized version of the margin

$$\phi(x, y, \boldsymbol{\alpha}^t) = y\sum_{i=1}^{t}\alpha_i h_i(x) .$$

Rätsch *et al.* [18] have prove that AdaBoost can be seen as a gradient descent procedure and that the reweighting scheme in the $(t+1)$-th iteration is equivalent to normalizing the gradient of $G(\alpha_{t+1}, \boldsymbol{\alpha}^t)$ with respect to $\phi(x_i, y_i, \boldsymbol{\alpha}^t)$

$$D_{t+1}(x_i) = \frac{\partial G(\alpha_{t+1}, \boldsymbol{\alpha}^t)}{\partial\phi(x_i, y_i, \boldsymbol{\alpha}^t)} \Bigg/ \sum_{j=1}^{N}\frac{\partial G(\alpha_{t+1}, \boldsymbol{\alpha}^t)}{\partial\phi(x_j, y_j, \boldsymbol{\alpha}^t)} .$$

By rewritting this expression, the following explicit formula for reweighting can be derived:

$$D_{t+1}(x_i) = \frac{\exp\left(-1/2\varrho(x_i, y_i, \boldsymbol{\alpha}^t)\right)^{|\boldsymbol{\alpha}^t|}}{\sum_{j=1}^{N} \exp\left(-1/2\varrho(x_j, y_j, \boldsymbol{\alpha}^t)\right)^{|\boldsymbol{\alpha}^t|}}$$

The most interesting part of this formula is its numerator. Here, the negative margin is the argument of an exponential function, which is raised to the power of $|\boldsymbol{\alpha}^t|$. For a constant exponent, examples with a large margin are assigned small weights in the next iteration and examples with small margins receive large weights. The denominator normalizes the single weights. It can be shown that $|\boldsymbol{\alpha}^t|$ increases at least linearly in t, if each weak classifier achieves a weighted training error smaller than 1/2. While this parameter increases, the weighting strategy of AdaBoost gets more and more selective in giving large weights to only a few examples with very small margins.

Let for example x_i and x_j be two data points with margins $\varrho(x_i, y_i, \boldsymbol{\alpha}^t) = 0.4$ and $\varrho(x_j, y_j, \boldsymbol{c}^t) = 0.3$. The numerator assigns to them for $|\boldsymbol{\alpha}^t| = 1$ the values 0.8187 and 0.8607. For $|\boldsymbol{\alpha}^t| = 1000$ this values are 1.3839×10^{-87} and 7.1751×10^{-66} and the difference between them is about 20 orders of magnitude. If enough iteration steps were carried out, the weighted error of a weak hypothesis will only be determined by the examples with the smallest margin.

Noisy examples tend to be surrounded by data points with different labels and will be more difficult to learn for AdaBoost. Therefore, these examples will receive lower margins than the ordinary data. So in later iterations the AdaBoost algorithm will only be concerned with learning the noisy data points and will lose its classification ability for the regular data, which will lead to overfitting.

11.6 BrownBoost

In order to create boosting algorithms, which are more robust against noise, it was not only derivates from AdaBoost (e.g. [18, 20–22]) which were suggested. The BrownBoost [23] algorithm, introduced by Freund, is an enhancement on his earlier BBM algorithm. Freund's intention was to create a variant of the BBM algorithm, which does not depend on any assumption about the training error of the weak hypotheses.

For this let $0 < \delta < \gamma$ and h' be a hypothesis with very low precision (error $> 1/2 - \gamma$) but with error $< 1/2 - \delta$, such that it can be achieved by almost all hypotheses. Such a hypothesis h' can be built from an ordinary hypothesis h whose error is $1/2 - \gamma$, $\gamma > \delta$ by creating a probabilistic hypothesis

$$h'(x) = \begin{cases} h(x), & \text{with probability } \delta/\gamma \\ 0, & \text{with probability } (1 - \delta/\gamma)/2 \\ 1, & \text{with probability } (1 - \delta/\gamma)/2\,. \end{cases}$$

Choosing such an hypothesis h' during the BBM training would not have much effect on reweighting the distribution of the single training examples. Testing h' on this modified distribution will return an error which is also lower than $1/2 - \delta$ with high probability. In this way h' can be used in many consecutive iterations until its error becomes larger than $1/2 - \delta$ and very close to 1/2. Used in this way, BBM creates a weighted majority sum, where each weak hypotheses h'_i is weighted by the number of iterations it has "survived". Although the noise level of the single hypotheses h'_i is very high, a final hypothesis with expected error ε could be achieved, if at least $O(\delta^{-2} \ln(1/\varepsilon))$ boosting iterations were carried out. Because of the very small value of δ and the dependence of the number of iterations on δ^{-2}, the run-time of this algorithm is not very attractive.

In a next step we can think of an algorithm which uses weak hypotheses with $\delta \to 0$. Defining time $t = \delta^2 i$ and "location"

$$r_\delta = \delta \sum_{j=1}^{\lceil t/\delta^2 \rceil} h'_j(x)$$

this can be interpreted in the limit $\delta \to 0$ as Brownian Motion with drift, a special kind of stochastic process (see e.g. [24]) with mean $\mu(t)$ and variance $\sigma^2(t)$

$$\mu(t) = \int_0^t \frac{1}{\gamma(s)} \, \mathrm{d}s, \qquad \sigma^2(t) = t \, .$$

Here $1/2 - \gamma(t)$ is the weighted error of the hypothesis h at time t.

An pseudocode of BrownBoost is given in Figure 11.10. The overall structure of BrownBoost is very similar to the structure of BBM or AdaBoost. One difference is that BrownBoost works continuously with time. Time t_i and its weight α_i are determined by solving the differential equation in step 3 of the algorithm. The procedure stops if there is no time left.

Besides the training sample S and the WeakLearn procedure, BrownBoost just needs a parameter c, which determines the amount of time BrownBoost is allowed to use, and a parameter ν, which is used to avoid degenerate cases.

It can be shown that c and the overall training error ε of the final hypothesis h_f are directly related:

$$\varepsilon = 1 - \mathrm{erf}(\sqrt{c}) = 1 - \frac{2}{\pi} \int_0^c e^{-x^2} \, \mathrm{d}x \, . \tag{11.1}$$

The algorithm therefore will run out of time before it learns the last εN examples. This fraction will consist of the hardest training examples. In a setting with noisy data this fraction will hopefully contain the noisy examples. If we can guess the percentage of noisy data in the training sample, we can find a c, which will pay more attention to the ordinary data than to the noisy ones.

We compared the sensitivity to noise of BrownBoost with AdaBoost and a 1-NN classifier by a four-fold cross-validation experiment [25]. To that end an artifical dataset was built containing 300 examples. Each example consists of 20 features in $[-1, +1]$. The values are chosen with respect to a uniform distribution and the data is separated by the concept con:

$$\operatorname{con}(\boldsymbol{x}) = \begin{cases} +1\,, & \text{if } 0.75\sin(\pi x_7) < x_{14} \\ -1\,, & \text{otherwise}\,. \end{cases}$$

The experiments were performed at noise levels of 0%, 10% and 20%, which means that a certain amount of class labels are toggled. To estimate sensible values for the number of hypotheses we ran BrownBoost on the data assuming a 0% noise level. We used this maximal number of hypotheses (200) in all subsequent experiments with AdaBoost. The amount of time for BrownBoost is determined by the noise level and (11.1). The WeakLearn algorithm used produces simple threshold classifiers

$$h_{i,j,t}(\boldsymbol{x}) = \begin{cases} \operatorname{sign}(\mathbb{1}_{[t \le x_j]} - 0.5), & \text{if } i = 1 \\ \operatorname{sign}(\mathbb{1}_{[t \ge x_j]} - 0.5), & \text{otherwise}\,. \end{cases}$$

The experiment was repeated ten times on newly sampled datasets. The results are given in Figure 11.5. The 1-NN classifer was outperformed by both boosting approaches. AdaBoost seems to perform better in the noise-free case. BrownBoost has lower error rates for the higher noise levels.

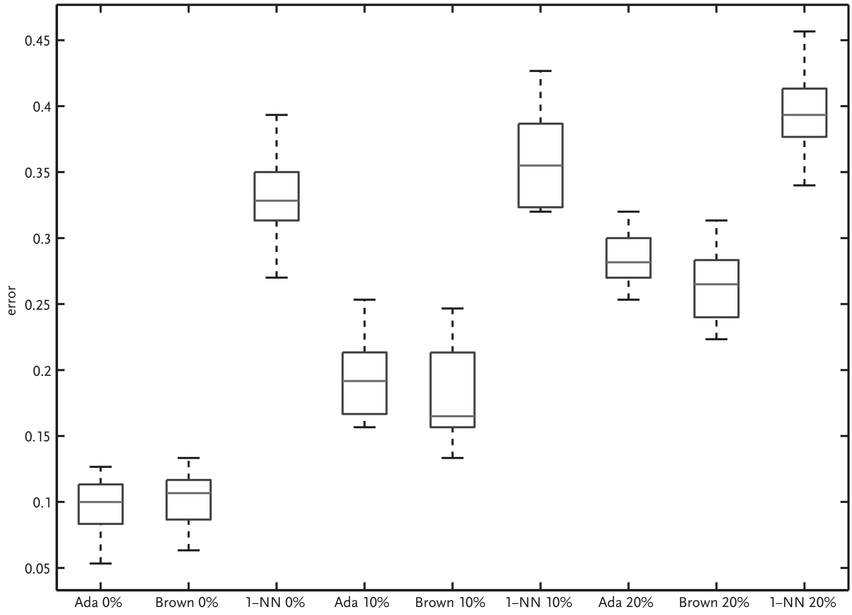


Figure 11.5 Comparison of AdaBoost, BrownBoost and 1-NN for different noise levels.

11.7 AdaBoost for Feature Selection

With some slight modifications, AdaBoost can be used for feature reduction in high-dimensional data. This was, for example, suggested by Viola and Jones [26]. Their work is about robust real-time object detection on visual data. In this kind of application several thousands of sub-images of an image have to be scanned and classified by a detector. In this data often only a few sub-images are of interest. Because of this imbalance a detector will be judged rather by its false positive rate and its detection rate than by its classification error. An accurate detector will have a high detection rate and a low false positive rate.

In order to achieve this in real time, the classifier should evaluate as few as possible features while classifying an example. Also, single features have to be computable quickly. By using a special representation of the image data, called integral images, Viola and Jones were able to create features, which can be computed in constant time for any scale of the input image. The set of possible features increases exponentionally with the size of the minimum sub-image shown to the detector. For a basic size of 24 × 24 pixels the set of all possible features already contains over 160 000 features. The overall structure of the detector is a conjunction of several ensembles h^i_{ada} trained by the AdaBoost scheme:

$$h_{\text{cas}}(x) = \bigwedge_K h^i_{\text{ada}}(x) \ .$$

The features which will be used in a single stage are chosen during the training of the AdaBoost ensemble. Each member of this ensemble is restricted to a single feature of the set. Usually these weak hypotheses are returned by WeakLearn, calling this procedure a weak hypothesis for each single feature is built. The weak hypothesis is returned which minimizes the weighted training error. In this way it is not only the best weak hypothesis, but also the best feature which is chosen. By iteratively increasing the number of weak hypotheses a small set of features can be found which achieves the predefined false positive rate and detection rate.

By evaluating the cascade iteratively, normally only a few stages of the cascade have to be evaluated. An example will be rejected with a negative label immediately, if it has received its first negative label by an AdaBoost classifier. Particularly in a detection task there will usually be a higher number of negative examples than positive. Additionally a classifier h^i_{ada} does not need to classify examples correctly which have been rejected by a previous stage $j < i$. So the training set can be adapted for each single stage. Let de_i and fa_i denote the detection rate and false positive rate of the classifier h^i_{ada} on such a training set

$$\begin{aligned} de_i = & \quad \widehat{Pr}_V\left[h^i_{\text{ada}}(x) | \textstyle\bigwedge_{j<i} h^j_{\text{ada}}(x) \wedge (y = 1)\right] , \\ fa_i = & \quad \widehat{Pr}_V\left[h^i_{\text{ada}}(x) | \textstyle\bigwedge_{j<i} h^j_{\text{ada}}(x) \wedge (y = -1)\right] . \end{aligned}$$

Here the class label $y = 1$ stands for an object x of interest and V is the distribution of the examples in the validation set. The cascade's false positive rate Fa_i and

Table 11.1 Experiments with a cascaded classifier for zebra crossing detection. The table shows the amount of sub-images which were rejected by the *i*-th stage of the cascade by several driving scenarios. The rows "false detections" and "correct detections" show the number of sub-images which received the label 1 from the cascade.

Stage	City	Highway	Crossways	Zebra crossing
1	67.49%	65.98%	66.83%	64.44%
2	22.14%	24.78%	23.90%	23.04%
3	4.80%	4.37%	4.71%	5.43%
4	0.77%	0.83%	0.76%	1.13%
5	1.55%	1.47%	1.34%	1.66%
6	0.35%	0.26%	0.27%	0.47%
7	1.97%	1.35%	1.43%	2.51%
False detections	0.92%	0.95%	0.75%	0.07%
Correct detections	–	–	–	0.06%

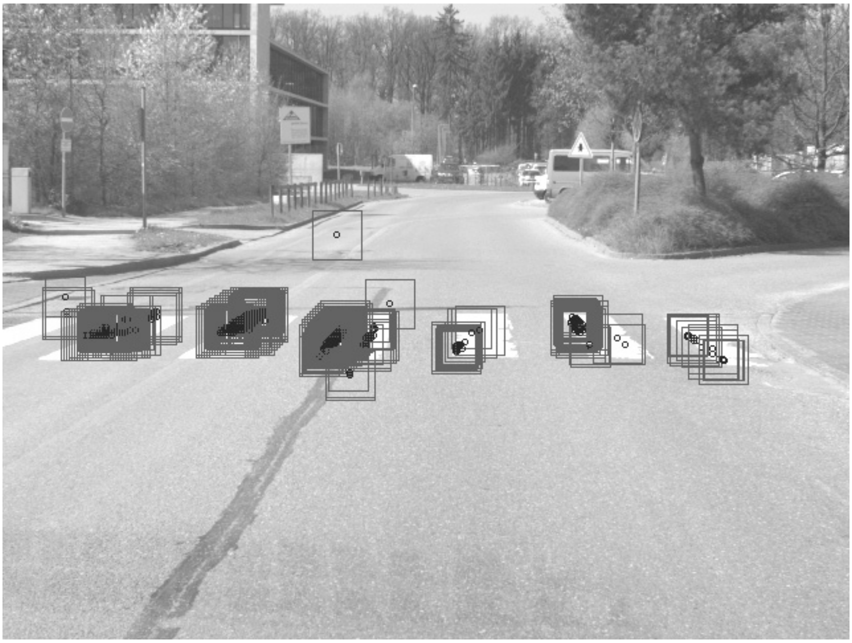


Figure 11.6 Image of the zebra crossing experiment.

detection rate De_i at stage i can be determined as

$$De_i = \prod_{j \le i} de_j$$

$$Fa_i = \prod_{j \le i} fa_j$$

The pseudocode of the cascade's training algorithm is given in Figure 11.11. The parameters of this procedure contain upper bounds for the false positive rates of a single stage fa and the whole cascade Fa_{target} and a bound for the detection rate of a single stage de. These three parameters determine the structure of the cascade. The algorithm adds at each iteration i an AdaBoost ensemble to the cascade until the detectors false positive rate Fa_i is smaller than Fa_{target}.

The parameters fa and de determine the number n_i of weak hypotheses in a single AdaBoost ensemble. This number is increased iteratively until both bounds are fulfilled. The AdaBoost algorithm was built to minimize a classification error. In order to do so it might happen that the detection rate decreases. By decreasing the threshold of h^i_{ada} both the detection rate and the false positive rate will increase. If a threshold can be found for which both bounds fa and de hold then this method will be preferred to increasing n_i. A single-stage classifier of a higher stage will be more complex than that of a lower one; n_i is also the number of features which have to be evaluated. By combining these techniques the expected number of feature, which have to be evaluated by a cascade N_{cas} is much smaller than that of a monolithic detector

$$N_{\text{cas}} = n_1 + \sum_{i=2}^{K} n_i p_{i-1} \ll \sum_{i=1}^{K} n_i = N_{\text{mon}} \,.$$

Here p_i denotes the positive rate that a random example passes stage i. Note that small values for fa_i not only improve the performance of the final detector but also increase the average speed of the final cascade.

In practical applications the cascaded structure is very advantageous. The results of an experiment with a cascade trained for detecting zebra crossings (Figure 11.6) are shown in Table 11.1. The table shows the fraction of sub-images, which were rejected in the i-th stage of the cascade. The tests were made for different driving scenarios. It can be seen that most of the sub-images can be rejected at early stages. These classifiers consist of a very small set of features and are therefore very simple. The more complex classifiers must only be evaluated for an small percentage of all sub-images. Note that the results of Table 11.1 come from a pure classifier trained by the algorithm of Viola and Jones. These results can be optimized by using additional pre- and postprocessing steps [27]. In this way over 95% of all sub-images can be rejected in the first stage of the cascade.

11.8 Conclusion

In this chapter an overview of the boosting approach has been given. The hypothesis boosting problem, introduced by Kearns in 1988 [1] and Kearns and Valiant in 1989 [2], was presented and a review on the very first boosting algorithms of Schapire [5] and Freund [6] was given (Learn and BBM). Some weaknesses of these algorithms were discussed, which motivated the development of the AdaBoost algorithm by Freund and Schapire in 1995 [7]. This boosting algorithm is the most popular one. Some theoretical work on this algorithm was currently demonstrated. It was also shown that AdaBoost is susceptible to noise. The algorithm BrownBoost of Freund developed in 2001 [23] was introduced as a possible way of dealing with noise. The way in which Boosting leads naturally to feature selection via a cascade of classifiers was demonstrated by Viola and Jones [26] and our own work [27].

```
Learn($\varepsilon$,$\delta$,WeakLearn,$EX$)

Input:
        error parameter $\varepsilon$
        confidence parameter $\delta$
        weak learning algorithm WeakLearn
        example oracle $EX$
        (implicit) size parameters $s$ and $n$

Procedure:
        if $\varepsilon \le 1/2 - 1/p(n, s)$ then return WeakLearn($\delta$, $EX$)
        $\alpha \leftarrow g^{-1}(\varepsilon)$
        $EX_1 \leftarrow EX$
        $h_1 \leftarrow$ Learn($\alpha, \delta/5, EX_1$)
        $\tau_1 \leftarrow \varepsilon/3$
        let $\hat{a}_1$ be an estimate of $a_1 = Pr_{v \in D}[h_1(v) \neq c(v)]$:
        choose a sample sufficiently large that $|a_1 - \hat{a}_1| \le \tau_1$ with probability $\ge 1 - \delta/5$
        if $\hat{a}_1 \le \varepsilon - \tau_1$ then return $h_1$

        defun $EX_2$()
            { flip coin
            if heads, return the first instance $v$ from $EX$ for which $h_1(v) = c(v)$
            else return the first instance $v$ from $EX$ for which $h_1(v) \neq c(v)$ }
        $h_2 \leftarrow$ Learn($\alpha, \delta/5, EX_2$)
        $\tau_2 \leftarrow (1 - 2\alpha)\varepsilon/8$
        let $\hat{e}$ be an estimate of $e = Pr_{v \in D}[h_2(v) \neq c(v)]$:
        choose a sample sufficiently large that $|e - \hat{e}| \le \tau_2$ with probability $\ge 1 - \delta/5$
        if $\hat{e} \le \varepsilon - \tau_2$ then return $h_2$

        defun $EX_3$()
            { return the first instance from EX for which $h_1(v) \neq h_2(v)$ }
        $h_3 \leftarrow$ Learn($\alpha, \delta/5, EX_3$)

        defun $h(v)$
            { $b_1 \leftarrow h_1(v)$, $b_2 \leftarrow h_2(v)$
            if $b_1 = b_2$ then return $b_1$
            else return $h_3(v)$ }
        return $h$

Output:
        a hypothesis that is $\varepsilon$-close to the target concept $c$ with
        probability $\ge 1 - \delta$
```

Figure 11.7 Pseudocode description of the algorithm Learn.

```
BBM($EX$,$\gamma$,WeakLearn,$N$)

Input:
    example oracle $EX$
    an algorithm WeakLearn which generates weak hypothesiss with training error smaller than $1/2-\gamma$
    with probability $1-\delta$
    sample size $N$

Procedure:
    Call $EX$ $N$ times to generate a sample $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$. To each example $(x_j, y_j)$ in $S$ corresponds a weight $D_1(j) = 1/N$ and a count $r_j = 0$.
    Find a (small) $k$ that satisfies

        $$\sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} (\frac{1}{2}-\gamma)^i (\frac{1}{2}+\gamma)^{k-i} < \frac{1}{N}$$

    (For example, any $k > 1/(2\gamma^2)\ln(N/2)$ is sufficient.)
    Do for $t = 1 \dots k$
    1. Do for $l = 1 \dots (1/(1-\delta))\ln(2k/\delta)$ or until a weak hypothesis is found
        (a) Call WeakLearn, referring it to FiltEX ($D_t$) as its source of examples, and save the returned hypothesis as $h_t$
        (b) If $\sum_{j=1}^{N} D_t(j)\mathbb{1}_{\left[h_t(x_j)\neq y_j\right]} < 1/2-\gamma$ then declare $h_t$ a weak hypothesis and exit loop.
    2. Increment $r_j$ by one for each example on which $h_t(x_j) = y_j$.
    3. Update the weights of the examples according to $D_t(j) = a^t_{r_j}$, where

        $$a^t_r = \binom{k-t-1}{\lfloor k/2 \rfloor - r} \left(\frac{1}{2}+\gamma\right)^{\lceil k/2 \rceil - r} \left(\frac{1}{2}-\gamma\right)^{\lceil k/2 \rceil - t - 1 + r}$$

    4. Normalize the weights by dividing each weight by $\sum_{j=1}^{N} D_t(j)$

Output:
    A hypothesis $h_M$ that is consistant on a random sample of size $N$

        $$h_M(x) = \text{sign}\left(\sum_{t=1}^{k} h_t(x)\right)$$

Subroutine FiltEX($D_t$)

    Choose a real number $x$ uniformly at random in the range $0 \le x < 1$.
    Perform a binary search for the index $j$ for which

        $$\sum_{i=1}^{j-1} D_t(j) \le x < \sum_{i=1}^{j} D_t(j) \text{ where } \sum_{i=1}^{0} D_t(j) := 0$$

    Return the example $(x_j, y_j)$
```

Figure 11.8 Pseudocode description of the algorithm Boosting by Majority (BBM).

```
AdaBoost($S$,WeakLearn,$T$)

Input:
    sequence $S$ of $N$ labeled examples $\langle (x_1, y_1), \ldots, (x_N, y_N) \rangle$ where $x_i \in X$ and $y_i \in \{-1, 1\}$
    weak learning algorithm WeakLearn
    integer $T$ specifying number of iterations

Init:
    distribution $\boldsymbol{D}_1$ with $D_1^i = 1/N$ for all $i \in \{1, \ldots, N\}$

Procedure:
    Do for $t = 1, 2, \ldots, T$
        1. Call WeakLearn, providing it with the distribution $\boldsymbol{D}_t$;
           get back a hypothesis $h_t : X \rightarrow \{-1, 1\}$.
        2. Calculate the error of $h_t : \varepsilon_t = \sum_{i=1}^{N} D_t(i) \mathbb{1}_{[h_t(x_i) \neq y_i]}$.
        3. Set $\alpha_t = \ln\left((1 - \varepsilon_t)/\varepsilon_t\right)$
        4. Update weights vector
               $D_{t+1}^i = \frac{D_t^i \exp\left(-\alpha_t \mathbb{1}_{[h(x_i)=y_i]}\right)}{Z_t}$
           where $Z_t$ is a normalization factor

Output:
    A hypothesis $h_f$
        $h_f(x) = \begin{cases} 1, & \text{if } \sum_{t=1}^{T} \alpha_t h_t(x) > 0 \\ -1, & \text{otherwise} \end{cases}$
```

Figure 11.9 Pseudocode description of the algorithm AdaBoost.

```
**BrownBoost($S$,WeakLearn,$c$,$\nu$)**

**Input:**
    sequence $S$ of $N$ labeled examples $\langle (x_1, y_1), \ldots, (x_N, y_N) \rangle$ where $x_i \in X$ and $y_i \in \{-1, 1\}$
    weak learning algorithm **WeakLearn**
    a positive real-valued parameter c. (Total amount of time)
    a small constant used to avoid degenerate cases $\nu > 0$

**Procedure:**
    Set initial margin $r_1(x_i, y_i) = 0$ for all $i \in \{1, \ldots, N\}$
    "remaining time" $s_1 = c$
    **Do for** $i = 1, 2, \ldots$
    1. Associate with each example a positive weight

           $W_i(x, y) = e^{-(r_i(x,y)+s_i)^2/c}$

    2. Call **WeakLearn** with the normalized distribution $\boldsymbol{D}_i = W_i(x, y) / \sum_{(x,y)} W_i(x, y)$ and receive from it
       a hypothesis $h_i(x)$ which has some advantage over random guessing $\sum_{(x,y)} D_i(x, y) h_i y = \gamma_i > 0$
    3. Let $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{t}$ be real-valued variables that obey the following differential equation:

           $\frac{\mathrm{d}\boldsymbol{t}}{\mathrm{d}\boldsymbol{\alpha}} = \boldsymbol{\gamma} = \frac{\sum_{(x,y)} \exp\left(-\frac{1}{c}\left(r_i(x, y) + \boldsymbol{\alpha} h_i(x)y + s_i - \boldsymbol{t}\right)^2\right) h_i(x)y}{\sum_{(x,y)} \exp\left(-\frac{1}{c}\left(r_i(x, y) + \boldsymbol{\alpha} h_i(x)y + s_i - \boldsymbol{t}\right)^2\right)}$

       Where $r_i(x, y)$, $h_i(x)y$ and $s_i$ are all constants in this context. Given boundary conditions $\boldsymbol{t} = 0$, $\boldsymbol{\alpha} = 0$
       solve the set of equations to find $t_i = \boldsymbol{t}^* > 0$ and $\alpha_i = \boldsymbol{\alpha}^*$ such that either $\boldsymbol{\gamma}^* \le \nu$ or $\boldsymbol{t}^* = s_i$
    4. Update the prediction value of each example to

           $r_{i+1}(x, y) = r_i(x, y) + \alpha_i h_i(x)y$

    5. update "remaining time" $s_{i+1} = s_i - t_i$
    **Until** $s_{i+1} \le 0$

**Output:**
    A hypothesis $h_f$

        $h_f(x) = \begin{cases} 1, & \text{if } \sum_i \alpha_i h_i(x) > 0 \\ -1, & \text{otherwise} \end{cases}$
```

Figure 11.10 Pseudocode description of the algorithm BrownBoost.

```
Cascade($EX$, WeakLearn, $fa$, $de$, $Fa_{\mathrm{target}}$)

Input:
          example oracle $EX$
          weak learning algorithm WeakLearn
          maximum acceptable false positive rate per layer $fa$
          minimum acceptable detection rate per layer $de$
          maximum acceptable overall false positive rate $Fa_{\mathrm{target}}$

Init:
          $P$ set of positive examples according to $EX$,
          $N$ set of negative examples according to $EX$,
          $Fa_0 = 1.0$, $De_0 = 1.0$, $i = 0$

Procedure:
          While $Fa_i > Fa_{\mathrm{target}}$
              $i = i + 1$, $n_i = 0$, $Fa_i = Fa_{i-1}$
              While $Fa_i > fa \times Fa_{i-1}$
                  $n_i = n_i + 1$
                  $h^i_{\mathrm{ada}}$ = AdaBoost($\{P, N\}$, WeakLearn, $n_i$)
                  Evaluate current cascaded classifier on a validation set to determine $Fa_i$ and $De_i$
                  Decrease threshold of $h^i_{\mathrm{ada}}$ until his detection rate $de_i \leq de \times De_{i-1}$ (determine $Fa_i$ again)
              $N = \emptyset$
              If $Fa_i > Fa_{\mathrm{target}}$ refill $N$ with negative examples from $EX$ which are misclassified by the current
              cascade

Output:
          A hypothesis $h_{\mathrm{cas}}$

                  $h_{\mathrm{cas}}(x) = \bigwedge_{K} h^i_{\mathrm{ada}}(x)$
```

Figure 11.11 Pseudocode description of the algorithm Cascade.

References

1 Michael Kearns (**1988**) Thoughts on hypothesis boosting. Unpublished manuscript, Dec.

2 Michael J (**1993**) Kearns and Leslie G. Valiant. Cryptographic limitations on learning boolean formulae and finite automata, in *Machine Learning: From Theory to Applications – Cooperative Research at Siemens and MIT*, London, UK. Springer-Verlag, 29–49.

3 Valiant, L.G. (**1984**) A theory of the learnable. *Commun. ACM*, **27**(11), 1134–1142.

4 Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M.K. (**1989**). Learnability and the vapnik-chervonenkis dimension. *J. ACM*, **36**(4), 929–965.

5 Schapire, R.E. (**1990**) The strength of weak learnability. *Machine Learning*, **5**(2), 197–227.

6 Freund, Y. (**1995**) Boosting a weak learning algorithm by majority. *Information and Computation*, **121**(2), 256–285.

7 Freund, Y. and Schapire, R.E. (**1995**) A decision-theoretic generalization of on-line learning and an application to boosting, in *Computational Learning Theory*, (ed P. Vitányi), volume 904 of *Lecture Notes in Artificial Intelligence*, Berlin, Springer, 23–37.

8 Freund, Y. and Schapire, R.E. (**1999**) A short introduction to boosting. *Journal of Japanese Society for Artifical Intelligence*, **14**, 771–780.

9 Vapnik, V.N. and Chervonenkis, A.Y. (**1974**) *Theory of Pattern Recognition [in Russian]*. Nauka, USSR.

10 Breiman, L. (**1996**) Bias, variance, and arcing classifiers. Technical Report 460, Statistics Department, University of California.

11 Drucker, H. and Cortes, C. (**1995**) Boosting decision trees, in *Advances in Neural Information Processing Systems 8, NIPS, Denver, CO, November 27–30, 1995*, 479–485.

12 Freund, Y. and Schapire, R.E. (**1996**) Experiments with a new boosting algorithm, in *International Conference on Machine Learning*, 148–156.

13 Quinlan, R.J. (**1996**) Bagging, boosting, and c4.5, in *AAAI/IAAI*, **1**, 725–730.

14 Schwenk, H. and Bengio, Y. (**1998**) Training methods for adaptive boosting of neural networks, in *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, Cambridge, MA, USA. MIT Press, 647–653.

15 Schapire, R.E., Freund, Y., Barlett, P., and Lee, W.S. (**1997**) Boosting the margin: A new explanation for the effectiveness of voting methods, in *ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning*, (ed D.H. Fisher), San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 322–330.

16 Dietterich, T.G. (**2000**) An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, **40**(2), 139–157.

17 Maclin, R. and Opitz, D. (**1997**) An empirical evaluation of bagging and boosting, in *AAAI/IAAI*, 546–551.

18 Rätsch, G., Onoda, T. and Müller, K.-R. (**2001**) Soft margins for adaboost. *Mach. Learn.*, **42**(3), 287–320.

19 Breiman, L. (**1999**) Prediction games and arcing algorithms. *Neural Comput.*, **11**(7), 1493–1517.

20 Servedio, R.A. (**2003**) Smooth boosting and learning with malicious noise. *J. Mach. Learn. Res.*, **4**, 633–648.

21 Domingo, C. and Watanabe, O. (**2000**) Madaboost: A modification of adaboost, in *COLT '00: Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 180–189.

22 Long, P.M. and Vega, V.B. (**2003**) Boosting and microarray data. *Mach. Learn.*, **52**(1–2), 31–44.

23 Freund, Y. (**2001**) An adaptive version of the boost by majority algorithm. *Mach. Learn.*, **43**(3), 293–318.

24 Breiman, L. (1992) *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

25 Bishop, C.M. (**2006**) *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

26 Viola, P. and Jones, M. (**2001**) Robust real-time object detection, in *Proc. of IEEE workshop on Statistical and Computational Theories of Vision*. Vancouver, Canada, July.

27 Lausser, L. (**2007**) Erkennung von Fahrbahnmarkierungen mit maschinellen Lernverfahren. Master's thesis, Institut für Neuroinformatik, Ulm, Germany, May.

12
The Sampling Theorem in Theory and Practice

Wolfgang Arendt, Michal Chovanec, Jürgen Lindner, Robin Nittka[1]

12.1
Introduction and History

Today the sampling theorem plays an important role in many areas, in theory as well as in practice. This contribution tries to shed some light on applications on the one hand and theory on the other, with two proofs being presented in the theoretical part. For a more detailed exposition of the sampling theorem and its importance in various fields we refer to the standard textbooks, for example [1–6].

In signal processing, sampling means to convert an analog signal (a function of a continuous-time variable) into a sequence of numbers (a function of a discrete time variable). A precondition is that the signal is band-limited, i.e. that its Fourier transform is zero for all frequencies outside a given frequency interval. By taking samples with a rate greater than the length of this interval in Hz, the theorem says that the exact reconstruction of the continuous-time signal from its samples is possible. The theorem also gives a formula for the reconstruction. Band-limitation indicates how fast the signal can change in time and hence also how much detail it can convey between two adjacent discrete instants of time.

From an historical point of view, the sampling theorem has two parts. The first part asserts that a band-limited function is completely determined by its samples, the second shows how to reconstruct the function from its samples. A few authors [7] suggested that the first part of the theorem goes back to a paper of Cauchy from 1841 [8]. However, the paper does not contain such a statement and it seems to be Borel [9] who stated the mathematical form of the theorem for the first time in 1897. Within the engineering community, this first part is contained in the work of Nyquist [10] in 1928. He demonstrated that $2B$ independent pulse samples per second could be sent through a system having a pass-band interval of $[-B, B]$. However, he did not consider the sampling and reconstruction of continuous signals. At the same time as Nyquist, Küpfmüller [11] obtained a similar result and, furthermore, discussed the sinc-function impulse response of a band-limiting filter. Here

[1] Corresponding author.

$\mathrm{sinc}(x) = \sin(x)/x$ is the sine cardinal. This band-limiting and reconstruction filter is sometimes called the *Küpfmüller filter*.

Both parts of the theorem (i.e. including reconstruction) were presented to the Russian communication community by Kotelnikov in 1933 [12]. In 1939, these were also described in the German literature by Raabe, an assistant of Küpfmüller, in his PhD thesis. Both parts of the sampling theorem were given in a somewhat different form by J.M. Whittaker [13] in 1935 and, slightly earlier, also by Ogura [14].

The theorem in the precise form in which we use it today was proved by C.E. Shannon [15] in 1949, hence it is often called *Shannon's theorem*. It was introduced to Japanese literature by Someya [16] at the same time. In the English literature, Weston [17, 18] proved it independently of Shannon at about the same time. The fact that so many people have independently contributed to the discovery of the theorem is best reflected in the various names one may find attached to the theorem in the literature, such as, for example, the Nyquist–Shannon sampling theorem, Whittaker–Kotelnikov–Shannon sampling theorem, Whittaker–Kotelnikov–Raabe–Shannon–Someya sampling theorem, and others.

As there have been various attempts to trace the origins of the sampling theorem, we should mention at least J.R. Higgins [19] (where, more generally, the history of the cardinal series can be tracked down), A.J. Jerri [20], H.D. Lüke [21], and E. Meijering [22].

12.2 The Sampling Theorem in Applications

There are many applications of the sampling theorem – in theory as well as in practice. "In theory" means that the sampling theorem is used as a basis for further theoretical derivations or descriptions, and "in practice" means that it forms the basis for real products or systems; for example, in the field of information transmission or information storage. Of course, the theoretical formulation of the sampling theorem is the basis for all these applications.

12.2.1 Applications in Theory

The general field dealing in a theoretical way with signals is called *signal theory*, and there is a more application-oriented counterpart called *signal processing*, which again contains a discipline described by *digital signal processing*. Signal theory is a special kind of mathematics dealing to a large extent with functions of one variable, which we will denote by s. We will denote the time variable by t, hence $s(t)$ stands for the signal value at time t. Because signal theory as a discipline was developed within the wider field of electrical engineering, there is much special terminology, definitions, descriptions and theorems which have no direct counterpart in classical mathematics. Because systems with signals as their input and output, especially linear time-invariant systems, can also be described with signals, a common name for this area is also *signals and systems*. *System theory* is also related to these disciplines, but here the system itself is the focus of attention. The sampling theorem belongs to all those areas.

Because signals are functions, we take the domain and range for classification. Domain and range are the sets of numbers from where the t-values or the corresponding s-values are taken. It is common to use two types of sets for domain and range to define four groups of signals: *continuous* and *discrete*. "Continuous" means intervals of the set of real numbers, and "discrete" stands for a subset of the set of integer numbers. Based on this, the three most important signal types are: continuous-time continuous range signals (also called analog signals), discrete-time continuous range signals (discrete-time signals for short), and discrete-time discrete range signals (also called digital signals). In theory and in corresponding applications it is also common to deal with complex-valued signals, vector-valued signals and multi-dimensional signals.

The sampling theorem states that there is a mapping of analog signals $s(t)$ to discrete-time signals $s_{\mathrm{discr}}(k)$ and vice versa. Here k ranges over the integers. The mapping is unique in both directions, as long as the analog signals are band-limited and the sampling rate $1/\Delta t$ is taken appropriately. "Band-limited" means that the Fourier transform (or spectrum) $S(f)$ of $s(t)$ equals zero for all frequencies f outside an interval $[-f_{\mathrm{c}}, f_{\mathrm{c}}]$. With the sampling rate $1/\Delta t > 2f_{\mathrm{c}}$ we define the corresponding discrete-time signal as follows:

$$s_{\mathrm{discr}}(k) = s(k \cdot \Delta t)$$

A handy formulation is common in connection with the sampling theorem. Sampling in the time domain with distance Δt leads to a periodic repetition with period $1/\Delta t$ in the frequency domain. Of course, if the periods in frequency do not overlap (we say there is no *aliasing*), the band-limited signal $s(t)$ can be reconstructed exactly by restricting the periodic repetition to $[-f_{\mathrm{c}}, f_{\mathrm{c}}]$, i.e. by applying the transfer function of an ideal low-pass filter with cut-off frequency f_{c}. In the time domain this means interpolation of $s_{\mathrm{discr}}(k)$ with the inverse Fourier transform of $\mathbb{1}_{[-f_{\mathrm{c}}, f_{\mathrm{c}}]}$ which is $2f_{\mathrm{c}} \cdot \mathrm{sinc}(2\pi f_{\mathrm{c}} t)$. It is straightforward to see that the following is also true. Sampling with Δf in the frequency domain leads to a periodic repetition with $1/\Delta f$ in the time domain. It is also easy to see that the *discrete Fourier transform* (DFT) includes both, periodicity and sampling in time and frequency as a precondition. If the DFT is taken together with the sampling theorem for analog signals this must be taken into account.

The examples above show that the sampling theorem leads to a more general theory of signals and systems, and it is also the theoretical background for connecting the DFT with the Fourier transform. Also, the understanding of sampling in frequency domain is enabled by the sampling theorem.

12.2.2 Applications in Practice

The application of the sampling theorem in practice concerns many fields: information transmission and storage, control and measurement systems, acoustic signal processing, to mention only a few examples. There are also applications which use a two-dimensional form of the sampling theorem with signals $s(x, y)$. Instead of time functions we then have functions which depend on the spatial coordinates x and y. In practice this might be related to digital photography, for example. With a time dimension t added we have $s(x, y, t)$, which is suitable for describing video signals.

In many applications in technology, digital signal processing (DSP) plays an important role. Because powerful, small and cheap DSP hardware is available today, there is a trend to process analog signals in a digital way wherever it is possible. The sampling theorem is the basis needed for that.

In reality, an analog signal may be a voltage, produced for example by a microphone. A signal $s(t)$ then stands for the variation of the voltage over time. The first step towards digital processing of such a voice signal is to pass it through a so-called *anti-aliasing low-pass filter* with cut-off frequency f_c. This guarantees that we have a band-limited signal at the output. So the basic precondition for the application of the sampling theorem is fulfilled. DSP requires a further step: *quantization*. This is because the samples $s_{\text{discr}}(k)$ are from the field of real numbers, but "digital" means that they have to be integer numbers. Quantization performs this mapping, with the disadvantage that it is not reversible. What numbers must be taken for the cut-off frequency f_c of the anti-aliasing low-pass filter and for the number of binary digits (bits) to get integer values in range? Both depend on the type of analog signal and the application – finally on what a user accepts as being "good enough". While for speech signals with telephone quality a sampling rate of 8 ksamples/s with 8 bit/sample is sufficient, we do accept music signals only with at least 32 ksamples/s and 14 bit/sample quantization. For music Compact Discs (CDs) we have 44.1 ksamples/s and 16 bit/sample in reality. For images there is no fixed number of samples/mm, but for the quantization 8 bit/sample is accepted as good enough. For color images the triple of samples for three basic colors, for example RGB, are called *pixels* – so we can also say 24 bit/pixel is good enough for ordinary images.

12.2.3
Special Case: Applications in the Field of Information Transmission

To transmit the sequence of quantized samples of an analog signal, for example of a speech signal, from a transmitter to a distant receiver, digital transmission methods must be used. For such digital transmission methods, analog signals play an important role again, because any physical transmission channel can only have physical quantities (like voltages or field strengths, for example) at its input and output. Sequences of numbers (digital signals) are artificial and must be represented by measurable physical quantities.

The basic principle of digital transmission is the following: The transmitter selects an analog signal $e_i(t)$ (elementary signal, basic waveform, transmit impulse) from a predefined set A_e of signals and sends it via the channel to the receiver. The receiver knows this set A_e, and for a given signal at its input it decides, with this knowledge, what has been sent. If M is the number of signals in A_e the number of information bits transmitted with a selection/reception of a certain $e_i(t)$ is $\log_2(M)$ (more precisely, this is only true if all signals are selected with the same probability). The number of information bits is the number of bits needed to represent the counting number i in binary form. This means that, for any digital transmission scheme, the numbering of the analog signals $e_i(t)$ is the information being transmitted. So again analog signals play an important role. At first glance there is no restriction for selecting the different $e_i(t)$. But in reality for every transmission only limited bandwidth is available, so the $e_i(t)$ have to be band-limited. This means that the precondition for the application the sampling theorem is fulfilled. Therefore it is not a surprise that in real digital transmission systems it is also common to apply DSP whenever possible. As mentioned above, analog signals must be used on any physical channel. So one of the final steps in a transmitter is digital to analog conversion (DAC), and in the receiver one of the first steps is analog to digital conversion (ADC).

The examples and explanations up to here are concerned with the application of the sampling theorem in its original form, i.e. sampling of signals (or time functions) and periodic repetition of their spectra. As mentioned in Section 12.2.1 there is a second form: sampling in the frequency domain and periodic repetition in the time domain. This form has also very important practical applications. All digital broadcast transmission systems like Digital Audio Broadcast (DAB), Digital Video Broadcast (DVB-T), and Digital Radio Mondiale (DRM) are based on it. Moreover, the same is true for many wireless local area network (WLAN) transmission methods, and also for internet connections via DSL (Digital Subscriber Line). Additionally, there is a trend to have it for all wireless transmission schemes in future. The digital transmission method used in all those systems is called Orthogonal Frequency Division Multiplexing (OFDM). As a generalization of the simpler scheme described before, we transmit here not only one basic waveform at a time, but M of them "in parallel", which is also termed *multiplexing*. For each $m = 1, \dots, M$ we define a set of waveforms by $e_m(t) = x_m \cdot u_m(t)$ with $x_m \in A_x$. The transmit symbol alphabet A_x contains M_x complex numbers. They are used as complex amplitudes for the signals $u_m(t)$, and they produce for each m the variety in the number of waveforms that we need for information transmission. The number of possible waveforms for fixed m is therefore M_x, and $\log_2(M_x)$ bits are transmitted for each m. Because we have M transmissions in parallel, the total number of bits transmitted in one *symbol period* is $M \cdot \log_2(M_x)$. DVB-T, for example, has a mode where $M = 6817$ (from 8192 possible) waveforms $u_m(t)$ and a transmit symbol alphabet A_x where up to 64 symbols are used.

Before we try to understand multiplexing and also OFDM a little bit better, we have to know that all waveforms $e_m(t)$ are summed up before leaving the transmitter. The transmit signal in one *symbol interval* is then

$$s(t) = \sum_{m=1}^{M} e_m(t) = \sum_{m=1}^{M} x_m \cdot u_m(t)$$

with $x_m \in A_x$. Up to this point we are dealing with what we call multiplexing based on linear modulation methods, which is more general than OFDM. For OFDM two tricky ideas play an important role. The first is that *eigenfunctions* of linear time invariant systems (LTI systems) are taken to build the set of functions $u_m(t)$. The background for this choice is that real physical transmission channels can be modeled as linear systems, and if the transmitter and/or the receiver does not move too fast, time invariance can also be assumed. With this precondition we can say that the transmitted signals do not vary in their shape while being transferred through the channel – this is the definition of eigenfunctions. The only effect the channel causes in the received signal is a complex factor, namely the eigenvalue λ_m which corresponds to the eigenfunction $u_m(t)$. So, for $u_m(t)$ being transmitted, the corresponding received waveform is $\lambda_m \cdot u_m(t)$.

What are the eigenfunctions of LTI systems? The answer to this question has been known for a long time and is basic knowledge in many disciplines: complex exponential functions. So we define $u_m(t) := \mathrm{e}^{2\pi i f_m t}$. Here we denote the imaginary unit by $i := \sqrt{-1}$ to be consistent with Section 12.3, even though usually it is denoted by j within the engineering community. The eigenvalues λ_m depend on the frequency f_m. In the area of signals and systems it is basic knowledge that the eigenvalues $\lambda_m(f_m)$ are identical with the Fourier transform of the impulse response $h(t)$ of the LTI system, which is called the transfer function $H(f)$ of the LTI system. As a result, the mapping from channel input to channel output is

$$\mathrm{e}^{2\pi i f_m t} \mapsto H(f_m) \cdot \mathrm{e}^{2\pi i f_m t}.$$

Different eigenfunctions (with different frequencies f_m) are orthogonal and the channel preserves this orthogonality. The orthogonality is a very suitable condition to separate the parallel transmissions at the receiving side without mutual interference (often called *crosstalk*). In OFDM the $u_m(t)$ are called *subcarriers* of the OFDM transmission.

One problem remains to be solved. We understand this by looking at the mechanism we need for a continuous digital transmission. We have to transmit $e_m(t)$ (and hence the $u_m(t)$) more than once. We have to do it successively in symbol time intervals T_S (the *symbol interval* has already been mentioned). A simple approach might help at a first glance. Let us define

$$u_m(t) := \mathrm{rect}\left(\frac{t}{T_\mathrm{S}}\right) \cdot \mathrm{e}^{2\pi i f_m t}$$

with the rectangular function $\mathrm{rect} = \mathbb{1}_{[-1/2,1/2]}$. For $\Delta f := \min_{i \neq k} |f_i - f_k| = 1/T_\mathrm{S}$ we have also orthogonality between the waveforms (or subcarriers), as before. The problem with this approach is that the $u_m(t)$ are no longer eigenfunctions of the channel. So the orthogonality will be lost at the output of the channel. Now the second tricky idea enters. We extend $u_m(t)$ periodically to a slightly larger interval, for example

$$u_m(t) := \mathrm{rect}\left(\frac{t}{T_\mathrm{S} + T_\mathrm{G}}\right) \cdot \mathrm{e}^{2\pi i f_m t} \,.$$

The value T_G is the so-called *guard time*. The effect is the following. If the channel impulse $h(t)$ response has a finite duration t_h, we can always find an interval of duration T_S in the received signal, which is identical to the case in which we transmit real eigenfunctions of infinite duration, as long as $T_G > t_h$. This is easy to prove and will not be done here. As a final result, we only waste a little bit of the energy of the transmitted $e_m(t)$ but with the consequence that the channel cannot destroy the orthogonality. To give real numbers, again for the DVB-T example from before, the fraction of $T_S + T_G$ used as a guard interval can be selected from 1/4, 1/8, 1/16 or 1/32. This is identical to the fraction of wasted symbol energy.

Back to the sampling theorem for this OFDM example. At the transmit side we have periodic signals with period T_S. This enables us to use the discrete Fourier transform (DFT) for the preparation of the whole transmit signal $s(t)$ in each symbol interval. In practice the inverse DFT (or inverse FFT) is taken. This leads to another interpretation: $s(t)$ may also be considered as a result of a Fourier synthesis with Fourier coefficients x_m. At the receiving side we have to remember the trick explained just before, for example, cutting out the proper period in the received signal. With this we have also periodic signals at the receiving side. Without additive noise or other interference, the DFT gives back the transmitted Fourier coefficients x_m, of course multiplied by the eigenvalue caused by the channel. Therefore the total transmission of the complex transmit symbols (or Fourier coefficients) from transmitter to receiver can be described by the mapping $x_m \mapsto H(f_m) \cdot x_m$ for $m = 1, \ldots, M$ and $x_m \in A_x$. Cutting out the proper period of the received signal and using the DFT is identical with a periodic extension of this cut-out. This again corresponds to sampling in the frequency domain. The samples are the values $H(f_m) \cdot x_m$.

12.3 Mathematical Formulation of the Sampling Theorem

We start by introducing some notation that will frequently be used throughout the rest of this article. Then we state the version of the sampling theorem considered here using precise mathematical notation. After that, two independent proofs are presented.

12.3.1 Notation

The $L^p(a, b)$-spaces $(1 \leq p < \infty)$ consist of (equivalence classes of) complex-valued measurable functions f such that $|f|^p$ is integrable over the interval (a, b); in the following we will primarily consider the case $(a, b) = \mathbb{R}$. If the reader is not familiar with this concept they might think of f as being a piecewise continuous function satisfying the integrability condition. For $p = 2$, $L^2(a, b)$ is a Hilbert space equipped with the inner product

$$(f\,|\,g) := (f\,|\,g)_{L^2} := (f\,|\,g)_{L^2(a,b)} := \int_a^b f(x)\overline{g(x)}\,\mathrm{d}x\,.$$

We need the (unnormalized) sine cardinal

$$\operatorname{sinc}\colon \mathbb{R} \to \mathbb{R}, \quad \operatorname{sinc}(x) := \begin{cases} (\sin x)/x & \text{if } x \neq 0\,, \\ 1 & \text{if } x = 0\,. \end{cases}$$

The characteristic function of a set $M \subset \mathbb{R}$ will be denoted by

$$\mathbb{1}_M\colon \mathbb{R} \to \{0,1\}, \quad \mathbb{1}_M(x) := \begin{cases} 1, & x \in M\,, \\ 0, & x \notin M\,. \end{cases}$$

Series over $\mathbb{Z}$ are always understood as

$$\sum_{k=-\infty}^{\infty} a_k := \lim_{n\to\infty} \sum_{k=-n}^{n} a_k\,,$$

if the limit exists.

12.3.2 The Sampling Theorem

We now state the main result. For an interpretation of the theorem, Ω is the band-width of the spectrum G of a signal g, and ω_s is the sampling frequency.

Theorem 12.1 (Sampling Theorem) *Let $\omega_\mathrm{s} > 0$ and $G \in L^1(\mathbb{R})$. Assume that there exists a positive number $\Omega < \omega_\mathrm{s}/2$ such that G vanishes outside the interval $[-\Omega, \Omega]$. Then*

$$g(t) := \int_{-\Omega}^{\Omega} \mathrm{e}^{2\pi i t y} G(y)\,\mathrm{d}y$$

satisfies the relation

$$g(t) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\omega_\mathrm{s}\pi\left(t - \frac{k}{\omega_\mathrm{s}}\right)\right) g\left(\frac{k}{\omega_\mathrm{s}}\right). \tag{12.1}$$

In particular, the series in (12.1) *converges for every $t \in \mathbb{R}$.*

This is an exact formula for reconstructing a band-limited function g on the whole real line from its values on a sufficiently narrow grid. In particular, a signal is uniquely determined by its band-width and samples which are taken at a rate that exceeds the Nyquist rate which is twice the bandwidth.

We remark that, in general, the theorem does not remain true if we relax the assumption $\Omega < \omega_\mathrm{s}/2$ to $\Omega \le \omega_\mathrm{s}/2$. Imagine a situation like $g(t) = \sin\big(\pi\omega_\mathrm{s} t/2\big)$, corresponding to Dirac impulses $G = \left(\delta_{\omega_s/2} - \delta_{\omega_s/2}\right)/2i$ in the frequency domain. Then all samples equal zero and we cannot distinguish g from the zero signal. Thus g cannot be reconstructed. But of course the condition $G \in L^1(\mathbb{R})$ is also not satisfied in this example. However, at least this gives an impression that in general oversampling is needed, i.e. we have to require the strict inequality. Still, as an alternative, we may only admit more regular functions G, for example only G in $L^2(\mathbb{R})$. Then indeed the theorem remains true for $\Omega = \omega_s/2$ as careful inspection of the proof in Section 12.3.4 shows.

Remark 12.1 *Note that under the hypotheses of Theorem 12.1, the signal g possesses a holomorphic extension to the whole complex plane. It is a consequence of the identity theorem for holomorphic functions that the signal has infinite duration unless it is identically zero. Thus the sum in* (12.1) *is infinite, but in practice we can only take a finite number of samples. Still, merely taking an approximation, the procedure works extremely well in applications.*

12.3.3
Efficient Proof

In this section we give a short proof which uses only a well-known result about the convergence of a Fourier series. Even though the calculations are easy to understand, the physical interpretation of the mathematical manipulations are not clear. In other words, it is not apparent what happens “behind the scenes”.

The key to the proof is to identify the sampling theorem in a special case as a Fourier series expansion and to extend this observation by uniform convergence to a wide class of functions. It seems that Vachenauer [23, Section 11.6.5] was the first author who made this idea into a rigorous proof.

12.3.3.1
Dirichlet’s Theorem

For a function $h\colon (-\omega_s/2, \omega_s/2) \to \mathbb{C}$ we define the Fourier coefficients

$$\hat{h}(k) := \frac{1}{\omega_s} \int_{-\omega_s/2}^{\omega_s/2} \mathrm{e}^{-ik2\pi x/\omega_s} h(x)\,\mathrm{d}x\,.$$

Under certain regularity conditions on h the (formal) Fourier series

$$\sum_{k=-\infty}^{\infty} \mathrm{e}^{ik2\pi x/\omega_s} \hat{h}(k)$$

represents $h(x)$ for all $x \in (-\omega_s/2, \omega_s/2)$. For our proof it is necessary to know that the Fourier series of a continuously differentiable function converges uniformly to the function on each interval of the form $[-\Omega, \Omega]$ with $0 < \Omega < \omega_s/2$. This is asserted by the following theorem. Notice that we allow $h(-\omega_s/2)$ to be different from $h(\omega_s/2)$.

Theorem 12.2 (Dirichlet) *Let $h\colon [-\omega_s/2, \omega_s/2] \to \mathbb{C}$ be a continuously differentiable function. Then for every $0 < \Omega < \omega_s/2$ the Fourier series of h converges to h uniformly on $[-\Omega, \Omega]$.*

Proof This is a special case of the Dini–Lipschitz test [24, Chapter II, Section 10]. In the literature there are also some versions of this theorem which are closer to the formulation above, see for example [25, Theorem 15.5]. □

From the preceding result the well-known fact can be deduced that the Fourier series of a continuously differentiable, ω_s-periodic function converges uniformly on the whole real line.

12.3.3.2
A First Attempt of a Proof

At this point we insert a proof of the sampling theorem using Fourier series which most mathematicians would probably come up with as their first idea on how to attack the problem. But it does not yield the result in the generality of Theorem 12.1, nor is it simpler than the next proof we will give. However, the approach is more natural and should not be missing in a compilation of proofs for the sampling theorem.

Assume, in addition to the assumptions in Theorem 12.1, that G is continuously differentiable. The relation

$$\hat{G}(-k) = \frac{1}{\omega_s}\int_{-\omega_s/2}^{\omega_s/2} e^{ik2\pi x/\omega_s} G(x)\,dx = \frac{1}{\omega_s} g\left(\frac{k}{\omega_s}\right) \tag{12.2}$$

is an immediate consequence from the definitions of g and the Fourier coefficients $\hat{G}(k)$ of G. From the remark after Theorem 12.2 we also know that the Fourier series of G converges to G uniformly on $[-\omega_s/2,\ \omega_s/2]$ since there exists a continuously differentiable, ω_s-periodic function on $\mathbb{R}$ which agrees with G on this interval. Due to this uniform convergence we are allowed to interchange summation and integration in the following calculation.

$$\begin{aligned} g(t) &= \int_{-\Omega}^{\Omega} e^{2\pi ity} G(y)\,dy = \int_{-\omega_s/2}^{\omega_s/2} e^{2\pi ity} \sum_{k=-\infty}^{\infty} \hat{G}(-k)\, e^{-ik2\pi y/\omega_s}\,dy \\ &= \sum_{k=-\infty}^{\infty} \hat{G}(-k) \int_{-\omega_s/2}^{\omega_s/2} e^{2\pi iyt}\, e^{-2\pi iyk/\omega_s}\,dy \\ &= \sum_{k=-\infty}^{\infty} g\left(\frac{k}{\omega_s}\right) \operatorname{sinc}\left(\pi\omega_s\left(t - \frac{k}{\omega_s}\right)\right) \end{aligned}$$

This is the sampling theorem. For the last transformation we have used formula (12.2) and the identity

$$\begin{aligned} \frac{1}{\omega_s}\int_{-\omega_s/2}^{\omega_s/2} e^{-ik2\pi y/\omega_s}\, e^{2\pi ity}\,dy &= \frac{e^{2\pi i(t-k/\omega_s)\omega_s/2} - e^{-2\pi i(t-k/\omega_s)\omega_s/2}}{\omega_s 2\pi i\,(t-k/\omega_s)} \\ &= \frac{\sin(\omega_s\pi\,(t-k/\omega_s))}{\omega_s\pi\,(t-k/\omega_s)} = \operatorname{sinc}\left(\omega_s\pi\left(t - \frac{k}{\omega_s}\right)\right) \end{aligned} \tag{12.3}$$

which holds at least if $k \neq t\omega_{\mathrm{s}}$. But note that, due to continuity, the left most and right most expressions are also equal for $k = t\omega_{\mathrm{s}}$.

12.3.3.3 The Efficient Proof

Now we demonstrate how the sampling theorem can be proved using an idea which is similar to the approach in the preceding section, but applies to a much wider class of functions. The difference is that we write the factor $\mathrm{e}^{2\pi i t \gamma}$ as a Fourier series instead of G.

We apply Dirichlet's theorem to the function

$$h\colon \left[-\frac{\omega_{\mathrm{s}}}{2}, \frac{\omega_{\mathrm{s}}}{2}\right] \to \mathbb{C}, \quad h(\gamma) := \mathrm{e}^{2\pi i t \gamma}$$

for fixed $t \in \mathbb{R}$. We have already computed the Fourier coefficients $\hat{h}(k)$ of this function in (12.3). Hence h can represented as

$$h(\gamma) = \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\omega_{\mathrm{s}}\pi\left(t - \frac{k}{\omega_{\mathrm{s}}}\right)\right) \mathrm{e}^{ik2\pi\gamma/\omega_{\mathrm{s}}} ,$$

where the series converges uniformly on $[-\Omega, \Omega]$ if $0 < \Omega < \omega_{\mathrm{s}}/2$. Note that this proves formula (12.1) for the special case $g(t) := \mathrm{e}^{2\pi i t \gamma}$ for fixed $\gamma \in (-\omega_{\mathrm{s}}/2, \omega_{\mathrm{s}}/2)$. The general case now follows. In fact, under the assumptions of Theorem 12.1 we can compute g, interchanging summation and integration by virtue of the uniform convergence.

$$\begin{aligned} g(t) &= \int_{-\Omega}^{\Omega} h(\gamma) G(\gamma)\,\mathrm{d}\gamma \\ &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\omega_{\mathrm{s}}\pi\left(t - \frac{k}{\omega_{\mathrm{s}}}\right)\right) \int_{-\Omega}^{\Omega} \mathrm{e}^{ik2\pi\gamma/\omega_{\mathrm{s}}} G(\gamma)\,\mathrm{d}\gamma \\ &= \sum_{k=-\infty}^{\infty} \operatorname{sinc}\left(\omega_{\mathrm{s}}\pi\left(t - \frac{k}{\omega_{\mathrm{s}}}\right)\right) g\left(\frac{k}{\omega_{\mathrm{s}}}\right) \end{aligned}$$

This is the sampling theorem.

Remark 12.2 *The assumption $G \in L^1(\mathbb{R})$ is stronger than necessary. The proof shows at once that we only need to be allowed to interchange summation and integration in order to prove the theorem. For example, for the Dirac delta function $G = \delta_0$ the proof remains valid, which shows that the sampling theorem is true for the constant function $g = 1$.*

More generally, (12.1) holds if g is the inverse Fourier transform of a signed, finite measure μ on $\mathbb{R}$ whose support is contained within an interval $[-\Omega, \Omega]$ for $0 < \Omega < \omega_{\mathrm{s}}/2$, i.e. for functions of the form

$$g(t) := \int_{\Omega}^{\Omega} \mathrm{e}^{2\pi i t \gamma}\,\mathrm{d}\mu(\gamma) .$$

The above proof carries over to this situation verbatim.

Note that, in the proof, the assumption of G being band-limited is applied where we have to assert uniform convergence of the Fourier series on the support of G.

12.3.4 Conventional Proof

In this section we demonstrate how the sampling theorem follows from the theory of Fourier transformations. The proof follows those found in the majority of the literature on this subject, particularly within the engineering community. For readers used to convolution theorems and with a solid knowledge of standard transformations this proof is straightforward and easy to understand. However, as we strive to make all arguments as mathematically precise as possible, we run into a delicate problem at one point of the proof; we will address it at the end of this section in more detail.

12.3.4.1 Tempered Distributions

We need a rather general concept of Fourier transforms. To this end we introduce the *space of Schwartz functions*

$$\mathcal{S} := \left\{\varphi \in C^{\infty}(\mathbb{R}) \mid x^k \varphi^{(m)}(x) \text{ is bounded for all } m, k \in \mathbb{N}_0\right\}$$

of all rapidly falling, smooth functions. Although we require much if we say that a function lies in $\mathcal{S}$, there exist many Schwartz functions, for example e^{-x^2}. Note that a Schwartz function multiplied by a polynomial remains a Schwartz function and that derivatives of Schwartz functions are also Schwartz functions.

We say that a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of Schwartz functions *converges in the space $\mathcal{S}$* to a function $\varphi \in \mathcal{S}$, if $x^k\varphi_n^{(m)}(x)$ converges to $x^k\varphi^{(m)}(x)$ uniformly on $\mathbb{R}$ for every choice of $k, m \in \mathbb{N}$. We remark that this is a very strong kind of convergence: it is hard for a sequence to converge in $\mathcal{S}$.

We call the space of all (sequentially) continuous linear maps $\mathcal{S} \to \mathbb{C}$ the *space of tempered distributions* and denote it by $\mathcal{S}'$. Note that it is easy for a linear map $S : \mathcal{S} \to \mathbb{C}$ to be in $\mathcal{S}'$ as only under the strong assumption $\varphi_n \to \varphi$ in $\mathcal{S}$ does the convergence $S(\varphi_n) \to S(\varphi)$ have to hold. We usually write $\langle S, \varphi\rangle$ instead of $S(\varphi)$. We say that a sequence $(S_n)_{n\in\mathbb{N}} \subset \mathcal{S}'$ converges in $\mathcal{S}'$ to $S \in \mathcal{S}'$, if it converges pointwise – in other words, if $\langle S_n, \varphi\rangle \to \langle S, \varphi\rangle$ for every $\varphi \in \mathcal{S}$.

Note that every integrable or bounded function f can be considered to be a tempered distribution J_f via the identification

$$\left\langle J_f, \varphi\right\rangle := \int_{\mathbb{R}} f(x)\varphi(x)\,dx \qquad (\varphi \in \mathcal{S}) . \tag{12.4}$$

In fact, this map is one-to-one in the sense that $J_f = J_g$ implies $f = g$ almost everywhere, see [26, Chapter I, Section 1.5]. Hence many function spaces, for example the space of integrable functions and the space of bounded functions, are identified with subspaces of $\mathcal{S}'$ and for this reason distributions are often called *generalized functions*.

The next four sections deal with operations involving tempered distributions and their properties. We closely follow the presentation of distributions as found in the Appendix of [27].

12.3.4.2 Fourier Transformation

For an integrable function $f \in L^1(\mathbb{R})$ it is possible to define the following functions as the integrals converge.

$$\mathcal{F}f(x) := \int_{\mathbb{R}} e^{-2\pi ixy} f(y)\, dy \qquad (x \in \mathbb{R})$$

$$\bar{\mathcal{F}}f(x) := \int_{\mathbb{R}} e^{2\pi ixy} f(y)\, dy = \overline{\mathcal{F}\bar{f}(x)} \qquad (x \in \mathbb{R})$$

The functions $\mathcal{F}f$ and $\bar{\mathcal{F}}f$ are called the Fourier transform and inverse Fourier transform of f, respectively.

It can be immediately seen that $\mathcal{F}f$ is bounded for every $f \in L^1(\mathbb{R})$. Moreover, it is an easy calculation involving integration by parts to obtain the following formulae for arbitrary $m \in \mathbb{N}$ and $\varphi \in \mathcal{S}$.

$$\mathcal{F}\left(\varphi^{(m)}\right)(x) = (-2\pi ix)^m \mathcal{F}\varphi(x)$$

$$\mathcal{F}\left((-2\pi i\cdot)^m \varphi\right)(x) = (\mathcal{F}\varphi)^{(m)}(x)$$

As a consequence, $\mathcal{F}\varphi \in \mathcal{S}$ if $\varphi \in \mathcal{S}$, using that $x^k\varphi^{(m)}(x) \in \mathcal{S}$ for all $m, k \in \mathbb{N}_0$. Furthermore, one can see that $\mathcal{F} : \mathcal{S} \to \mathcal{S}$ is continuous, using the above formulae and the theorem about interchanging integration and uniform convergence.

The Fourier transformation is well-behaved in the space $\mathcal{S}$, which is the main reason for introducing $\mathcal{S}$ in the first place. But also the larger space $L^2(\mathbb{R})$ which contains $\mathcal{S}$ is related to this operation as the following theorem shows. We will make use of this fact when we show that $\mathcal{F}$ can be extended even to tempered distributions in a sensible way.

Theorem 12.3 (Plancherel) *The Fourier transform $\mathcal{F}$ on $\mathcal{S}$ extends to a unitary operator on $L^2(\mathbb{R})$. In particular, $(f \mid g)_{L^2} = (\mathcal{F}f \mid \mathcal{F}g)_{L^2}$ for all $f, g \in \mathcal{S}$.*

Proof For the proof we use the inversion theorem (Theorem 12.4). We may do so because the proof of the inversion theorem does not depend on Plancherel's theorem.

Let $f, g \in \mathcal{S}$ be arbitrary. Since $\bar{\mathcal{F}}\mathcal{F}f = f$, we can proceed as follows.

$$\begin{aligned}(f \mid g)_{L^2} &= \int_{\mathbb{R}} \left(\bar{\mathcal{F}}\mathcal{F}f\right)(x)\overline{g(x)}\, dx = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{2\pi ixy} (\mathcal{F}f)(y)\, dy\, \overline{g(x)}\, dx \\ &= \int_{\mathbb{R}} (\mathcal{F}f)(y) \int_{\mathbb{R}} e^{2\pi ixy}\overline{g(x)}\, dx\ dy = \int_{\mathbb{R}} (\mathcal{F}f)(y)\overline{(\mathcal{F}g)(y)}\, dy \\ &= (\mathcal{F}f \mid \mathcal{F}g)_{L^2}\end{aligned}$$

The interchange of integrals is allowed by virtue of Fubini's theorem because f and g decrease sufficiently fast at infinity.

Using the density of $\mathcal{S}$ in $L^2(\mathbb{R})$, we conclude that $\mathcal{F}$ can uniquely be extended to a unitary linear operator on $L^2(\mathbb{R})$. This concludes the proof of the theorem. □

The operator can be extended even to a much larger space: We define the Fourier transform of tempered distributions S to be the functional $\mathcal{F}S$ acting as $\langle \mathcal{F}S, \varphi\rangle := \langle S, \mathcal{F}\varphi\rangle$ on $\varphi \in \mathcal{S}$. The linearity of $\mathcal{F}S$ is obvious, and by the continuity of $\mathcal{F}$ on $\mathcal{S}$ the functional $\mathcal{F}S$ is continuous, hence $\mathcal{F}S \in \mathcal{S}'$. Moreover, it is obvious from the definition of convergence in $\mathcal{S}'$ that the map $\mathcal{F}: \mathcal{S}' \to \mathcal{S}'$ is continuous. Note that all this can be done analogously for $\bar{\mathcal{F}}$.

It is a remarkable consequence of Plancherel's Theorem that this definition is consistent. For every $f \in L^2(\mathbb{R})$ and, in particular, for $f \in \mathcal{S}$, we have defined $\mathcal{F}f$ in two different ways, namely the definition for the function f as in Theorem 12.3 and for the distribution J_f as defined in (12.4). But in fact those two definitions agree.

Proposition 12.1 (Consistency of $\mathcal{F}$) *If $f \in L^2(\mathbb{R})$ then $\mathcal{F}J_f = J_{\mathcal{F}f}$.*

Proof Let $\varphi \in \mathcal{S}$ be arbitrary.

$$\left\langle \mathcal{F}J_f, \varphi\right\rangle = \left\langle J_f, \mathcal{F}\varphi\right\rangle = \int_{\mathbb{R}} f(x)\,(\mathcal{F}\varphi)\,(x)\,\mathrm{d}x = \left(f \mid \overline{\mathcal{F}\varphi}\right)_{L^2} = \left(f \mid \bar{\mathcal{F}}\,\bar{\varphi}\right)_{L^2}$$

$$= \left(\mathcal{F}f \mid \mathcal{F}\bar{\mathcal{F}}\,\bar{\varphi}\right)_{L^2} = (\mathcal{F}f \mid \bar{\varphi})_{L^2} = \int_{\mathbb{R}} (\mathcal{F}f)\,(x)\varphi(x)\,\mathrm{d}x = \left\langle J_{\mathcal{F}f}, \varphi\right\rangle$$

This shows $\mathcal{F}J_f = J_{\mathcal{F}f}$ in the sense of tempered distributions. □

12.3.4.3
Inversion Theorem

Now we are going to show that $\mathcal{F}$ is invertible, its inverse being $\bar{\mathcal{F}}$. For this purpose we need the Fourier transform of the constant function $\mathbb{1} := 1$. It is possible to compute $\mathcal{F}\mathbb{1}$ by an extensive and tricky, though very elegant calculation [27, Appendix (3.28)]. For our purposes, we mention it as an example of a Fourier transform without proof.

Example 12.1 *Let $\delta_0: \mathcal{S} \to \mathbb{R}$ denote the Dirac distribution $\langle \delta_0, \varphi\rangle := \varphi(0)$ for $\varphi \in \mathcal{S}$. Then $\mathcal{F}\mathbb{1} := \mathcal{F}J_{\mathbb{1}} = \delta_0$.*

We also need the following translation lemma for Fourier transforms.

Lemma 12.1 *Let $a \in \mathbb{R}$, and define $L_a\,\varphi \in \mathcal{S}$ by $(L_a\varphi)\,(x) := \varphi(x + a)$. Then $(\mathcal{F}L_a\varphi)\,(x) = \mathrm{e}^{2\pi ixa}\mathcal{F}\varphi(x)$ for every $\varphi \in \mathcal{S}$.*

Proof This follows from the substitution rule as follows.

$$(\mathcal{F}L_a\varphi)\,(x) = \int_{\mathbb{R}} \mathrm{e}^{-2\pi ixy}\varphi(y + a)\,\mathrm{d}y = \int_{\mathbb{R}} \mathrm{e}^{-2\pi ix(y-a)}\varphi(y)\,\mathrm{d}y = \mathrm{e}^{2\pi ixa}\mathcal{F}\varphi(x) \qquad □$$

Now we are able to prove the inversion theorem.

Theorem 12.4 (Inversion Theorem for $\mathcal{S}$) *The operator $\mathcal{F}\colon \mathcal{S} \to \mathcal{S}$ is bijective, its inverse being $\bar{\mathcal{F}}\colon \mathcal{S} \to \mathcal{S}$.*

Proof Let $\varphi \in \mathcal{S}$ and $a \in \mathbb{R}$ be arbitrary.

$$\begin{aligned}\varphi(a) &= \langle \delta_0, L_a\, \varphi\rangle = \langle \mathcal{F}\mathbb{1}, L_a\varphi\rangle = \langle \mathbb{1}, \mathcal{F} L_a\varphi\rangle \\ &= \int_{\mathbb{R}} \mathrm{e}^{2\pi iax}\mathcal{F}\varphi(x)\,\mathrm{d}x = \left(\bar{\mathcal{F}}\mathcal{F}\varphi\right)(a)\end{aligned}$$

This shows $\bar{\mathcal{F}}\mathcal{F}\varphi = \varphi$. The equality $\mathcal{F}\bar{\mathcal{F}}\varphi = \varphi$ can be proved analogously. □

An entirely different, more direct, approach to this theorem can be found in [28, Chapter II, Section 1.6].

From the definition of the Fourier transform of tempered distributions the corresponding statement in $\mathcal{S}'$ follows directly.

Corollary 12.1 (Inversion Theorem for $\mathcal{S}'$) *The operator $\mathcal{F}\colon \mathcal{S}' \to \mathcal{S}'$ is bijective, its inverse being $\bar{\mathcal{F}}\colon \mathcal{S}' \to \mathcal{S}'$.*

12.3.4.4 Examples

For later use, we give further examples of Fourier transforms.

Example 12.2 (Sine Cardinal sinc **and the Rectangle Impulse)** *For $\omega_\mathrm{s} > 0$ define $f_{\omega_\mathrm{s}}(x) := \omega_\mathrm{s}\,\mathrm{sinc}\,(\pi\omega_\mathrm{s}x)$. Then $\mathcal{F}f_{\omega_\mathrm{s}} = \mathbb{1}_{(-\omega_\mathrm{s}/2,\omega_\mathrm{s}/2)}$.*

Proof Let $x \neq 0$. Then

$$\begin{aligned}\left(\bar{\mathcal{F}}\mathbb{1}_{(-\omega_\mathrm{s}/2,\omega_\mathrm{s}/2)}\right)(x) &= \int_{-\omega_\mathrm{s}/2}^{\omega_\mathrm{s}/2} \mathrm{e}^{2\pi ixy}\,\mathrm{d}y = \frac{1}{2\pi ix}\left(\mathrm{e}^{\pi ix\omega_\mathrm{s}} - \mathrm{e}^{-\pi ix\omega_\mathrm{s}}\right) \\ &= \frac{\sin(\pi x\omega_\mathrm{s})}{\pi x} = f_{\omega_\mathrm{s}}(x)\ .\end{aligned}$$

Proposition 12.1 now asserts that $\bar{\mathcal{F}}\mathbb{1}_{(-\omega_\mathrm{s}/2,\omega_\mathrm{s}/2)} = f_{\omega_\mathrm{s}}$ in the sense of tempered distributions. Applying $\mathcal{F}$ to both sides of the identity, Corollary 12.1 proves the claim. □

Example 12.3 (Shah-Function Ш_T) *For $T > 0$ consider $\text{Ш}_T\colon \mathcal{S} \to \mathbb{R}$ defined as $\text{Ш}_T := \sum_{n\in\mathbb{Z}} \delta_{nT} \in \mathcal{S}'$, i.e., $\langle \text{Ш}_T, \varphi\rangle := \sum_{n\in\mathbb{Z}} \varphi(nT)$ for $\varphi \in \mathcal{S}$. Then Ш_T is a tempered distribution and $\mathcal{F}\text{Ш}_T = 1/T\,\text{Ш}_{1/T}$.*

Proof It is easy to see that Ш_T is linear and continuous, hence $\text{Ш}_T \in \mathcal{S}'$.
Fix $\varphi \in \mathcal{S}$ and define $\psi(x) := \sum_{n\in\mathbb{Z}} \varphi(x + n/T)$. Then ψ is continuously differentiable on $\mathbb{R}$ according to a theorem on differentiation of a series of differentiable

functions. The function ψ is $1/T$-periodic because a shift of its argument by a multiple of $1/T$ corresponds to a rearrangement of the summands of this absolutely convergent series.

Using the $2\pi i$-periodicity of the e-function and the substitution $y = x + n/T$, we can compute the Fourier coefficients of ψ. The interchange of summation and integration occurring here is easily justified by Fubini's theorem after noticing that the integrand is bounded by C/n^2 for a sufficiently large constant $C > 0$. Thus

$$\begin{aligned}\hat{\psi}(k) &= T\int_0^{1/T} \psi(x)\,\mathrm{e}^{-ik2\pi Tx}\,\mathrm{d}x = T\sum_{n\in\mathbb{Z}}\int_0^{1/T} \varphi\left(x+\frac{n}{T}\right)\mathrm{e}^{-ik2\pi Tx}\,\mathrm{d}x \\ &= T\int_{\mathbb{R}} \varphi(y)\,\mathrm{e}^{-2\pi ikTy}\,\mathrm{d}y = T\,(\mathcal{F}\varphi)\,(kT)\ .\end{aligned}$$

Applying Theorem 12.2 with $\omega_\mathrm{s} = 1/T$ we conclude in particular that the Fourier series of ψ represents ψ at 0, but a much weaker version of Dirichlet's theorem would have been sufficient for this.

$$\begin{aligned}\langle \mathcal{F}\text{Ш}_T, \varphi\rangle &= \sum_{k\in\mathbb{Z}} (\mathcal{F}\varphi)\,(kT) = \frac{1}{T}\sum_{k\in\mathbb{Z}} \hat{\psi}(k) = \frac{\psi(0)}{T} \\ &= \frac{1}{T}\sum_{n\in\mathbb{Z}} \varphi\left(\frac{n}{T}\right) = \left\langle \frac{1}{T}\text{Ш}_{1/T}, \varphi\right\rangle\end{aligned}$$

This proves the claim. □

Example 12.3 asserts that for a function $\varphi \in \mathcal{S}$ there is a relation between the values of φ on a lattice and the values of $\mathcal{F}\varphi$ on an inverse lattice which is known as the Poisson summation formula. In fact, this relation holds true for a much wider class of functions, see [24, Chapter II, Section 13].

12.3.4.5 Convolution

Another important concept in signal theory is the convolution

$$(\varphi * \psi)\,(x) := \int_{\mathbb{R}} \varphi(x-y)\psi(y)\,\mathrm{d}y \tag{12.5}$$

of two sufficiently fast decreasing functions, for example, $\varphi, \psi \in L^2(\mathbb{R})$. Under the Fourier transformation, the convolution becomes a multiplication, which is one of the reasons that make Fourier transforms particularly useful for calculations. Formulae of this kind are called convolution theorems. We prove a rather general one (Theorem 12.5) for use in our second proof of the sampling theorem, as we need this identity not only for functions for which the expression in (12.5) is defined, but also for tempered distributions. Unfortunately, it is not possible to obtain a convolution theorem for arbitrary $\varphi, \psi \in \mathcal{S}'$. In fact, it is not even possible to consistently define the convolution or multiplication of such general distributions.

First we note that the convolution defined as in (12.5) of a function in $L^1_c(\mathbb{R})$, i.e. an integrable function with bounded support, and a Schwartz function, is again a Schwartz function. More precisely, let $g \in L^1_c(\mathbb{R})$ and define $\check{g} \in L^1_c(\mathbb{R})$ by $\check{g}(x) := g(-x)$. Then $\check{g} * \varphi$ is differentiable, and its derivative is $\check{g} * \varphi'$. It is not too hard to see that indeed $\check{g} * \varphi \in \mathcal{S}$ and convolution with $\check{g}$ is a continuous operation on $\mathcal{S}$.

With this knowledge we may define the convolution of a tempered distribution U and a function $g \in L^1_c(\mathbb{R})$ by $\langle U * g, \varphi\rangle := \langle U, \check{g} * \varphi\rangle$. It is obvious that $U * g$ is linear. By continuity of the convolution with $\check{g}$ the functional $U * g$ is continuous, thus $U * g \in \mathcal{S}'$.

Example 12.4 *Let $h \in L^1_c(\mathbb{R})$ and $T > 0$. We define $H(x) := \sum_{k\in\mathbb{Z}} h\big(x - k/T\big)$, so that the function H is a sum of shifted copies of h. Then $h * Ш_{1/T} = J_H$, and in particular $J_H \in \mathcal{S}'$.*

For those familiar with the definition of the L^p-spaces, we point out that it is correct to define H as above. Different representatives of the same equivalence class in $L^1_c(\mathbb{R})$ give functions that agree almost everywhere, and each such H is locally integrable because on compact sets only finitely many summands do not vanish. Thus H is well-defined as an element in $L^1_{\text{loc}}(\mathbb{R})$.

Proof Let $\varphi \in \mathcal{S}$.

$$\begin{aligned}\left\langle h * Ш_{1/T}, \varphi\right\rangle &= \sum_{k\in\mathbb{Z}} \left(\check{h} * \varphi\right)\left(\tfrac{k}{T}\right) = \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}} h\left(x - \tfrac{k}{T}\right)\varphi(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}}\left(\sum_{k\in\mathbb{Z}} h\left(x - \tfrac{k}{T}\right)\right)\varphi(x)\,\mathrm{d}x = \left\langle \sum_{k\in\mathbb{Z}} h\left(\cdot - \tfrac{k}{T}\right), \varphi\right\rangle\end{aligned}$$

Here we have interchanged summation and integration. This is allowed because the intersection of the supports of $h\big(x - k/T\big)$ and $\varphi(x)$ leaves any compact set as $|k|$ becomes large. Thus we can exploit the fast decay of φ to show integrability of the product on the product space, and then Fubini's theorem justifies the calculation. □

For functions η with the property that $\eta\varphi \in \mathcal{S}$ for every $\varphi \in \mathcal{S}$ we define multiplication of a tempered distribution U with η by $\langle \eta U, \varphi\rangle := \langle U, \eta\varphi\rangle$. We denote the set of all such functions η by O_M and mention that O_M is the set of all infinitely differentiable functions such that the function and all of its derivatives are bounded by polynomials. It is easy to see that for such η the functional ηU is again a tempered distribution for every $U \in \mathcal{S}'$.

We remark that convolution and multiplication defined in the setting of tempered distributions are consistent with the usual definitions if the tempered distribution can be represented by a function in the sense of (12.4).

As mentioned before, we will now relate the convolution and multiplication of tempered distributions as defined above via the Fourier transformation.

Theorem 12.5 (Convolution Theorem) *Assume that $U \in \mathcal{S}'$ and $g \in L^1_c(\mathbb{R})$. Then $\mathcal{F}g \in O_M$ and $\mathcal{F}\,(U * g) = \mathcal{F}g \cdot \mathcal{F}\,U$.*

Proof According to a theorem about the differentiation of parameter integrals, the function $\mathcal{F}g$ is infinitely differentiable and its n-th derivative is given by

$$(\mathcal{F}g)^{(n)}(x) = \int_{\mathbb{R}} (-2\pi i y)^n \, \mathrm{e}^{-2\pi i x y} g(y) \, \mathrm{d}y \, .$$

Since g is integrable and the other factor of the integrand is bounded on the (compact) support of g, we see that for $n \in \mathbb{N}$ the function $(\mathcal{F}g)^{(n)}$ is bounded, hence $\mathcal{F}g \in O_M$.

Now we show $\mathcal{F}(U * g) = \mathcal{F}U \cdot \mathcal{F}g$. Let $\varphi \in \mathcal{S}$. It is a straightforward calculation involving Fubini's theorem and linear substitution to show that the identities $\bar{\mathcal{F}}(\check{g} * \psi) = \bar{\mathcal{F}}\check{g} \cdot \bar{\mathcal{F}}\psi$ and $\bar{\mathcal{F}}\check{g} = \mathcal{F}g$ hold for every $g \in L^1_c(\mathbb{R})$ and $\psi \in \mathcal{S}$. Setting $\psi = \mathcal{F}\varphi$ this proves $\check{g} * \mathcal{F}\varphi = \mathcal{F}(\mathcal{F}g \cdot \varphi)$ and hence

$$\langle \mathcal{F}(g * U), \varphi \rangle = \langle U, \check{g} * \mathcal{F}\varphi \rangle = \langle U, \mathcal{F}(\mathcal{F}g \cdot \varphi) \rangle = \langle \mathcal{F}g \cdot \mathcal{F}U, \varphi \rangle$$

for every $\varphi \in \mathcal{S}$. □

Note that the proof can be carried out analogously for $\bar{\mathcal{F}}$ to get the corresponding convolution theorem for the inverse Fourier transform under the same assumptions.

This is not the most general version of the convolution theorem that can be proved. In fact, there are many generalizations to the statement above, none of them being the single most general truth.

12.3.4.6
The Conventional Proof

Now we demonstrate our second proof of the sampling theorem utilizing the theory of Fourier transforms and the convolution theorem. Readers familiar with these concepts and formulae for transforms will find this approach easy to understand. They easily obtain a picture of what is happening, switching their point of view from the time domain to the frequency domain and back. Nevertheless, there are some very delicate mathematical problems involved in this calculation, which we will point out afterwards.

As in the sampling theorem we fix values $\omega_s > 0$ and $0 < \Omega < \omega_s/2$. Let G be a function in $L^1(\mathbb{R})$ vanishing outside $[-\Omega, \Omega]$. Define $g(t) := \bar{\mathcal{F}}G$ as in Theorem 12.1. Note that $g \in O_M$ which follows as in the proof of Theorem 12.5. We define $g_a := g\text{III}_{1/\omega_s}$, i.e.,

$$\langle g_a, \varphi \rangle = \sum_{k \in \mathbb{Z}} g\left(\frac{k}{\omega_s}\right) \varphi\left(\frac{k}{\omega_s}\right)$$

for all $\varphi \in \mathcal{S}$. Note that g_a only depends on samples of g taken on a grid with mesh size $1/\omega_s$. We define the distribution g_a because we want to put the sequence of measurements into the framework of tempered distributions in order to apply Fourier transformation and convolution.

From Example 12.4 we know that $G * \text{Ш}_{\omega_s} = \sum_{k\in\mathbb{Z}} G(\cdot - k\omega_s)$. We can obtain G from $G * \text{Ш}_{\omega_s}$ by restricting the function $G * \text{Ш}_{\omega_s}$ to the interval $(-\omega_s/2, \omega_s/2)$ because the shifted copies of G occurring in the series $\sum_{k\in\mathbb{Z}} G(\cdot - k\omega_s)$ do not overlap due to $\Omega < \omega_s/2$ (in other words, there is no aliasing). Thus

$$G = (G * \text{Ш}_{\omega_s})\, \mathbb{1}_{(-\omega_s/2,\omega_s/2)} \ . \tag{12.6}$$

We point out that here we require the signal to be band-limited. Additionally, we see from this argument that the condition $\Omega \leq \omega_s/2$ is optimal since $G * \text{Ш}_{\omega_s}$ does not uniquely encode the data of G if G has a broader spectrum. However, we can permit $\Omega = \omega_s/2$ here.

At this point we see what happens if the signal fails to be band-limited. In this case higher and lower frequencies contribute to the values that $G * \text{Ш}_{\omega_s}$ has on $(-\omega_s/2, \omega_s/2)$. If we try to reconstruct the signal as if it were band-limited, we arrive at the signal belonging to the spectrum $(G * \text{Ш}_{\omega_s})\mathbb{1}_{(-\omega_s/2,\omega_s/2)}$. From this point of view it is reasonable to assume that we maintain most of the original signal's information even if G is merely negligibly small outside $(-\omega_s/2, \omega_s/2)$, but does not vanish.

Now we use the convolution theorem for the inverse Fourier transform and Example 12.3 and obtain

$$\bar{\mathcal{F}}(G * \text{Ш}_{\omega_s}) = \bar{\mathcal{F}}G \cdot \bar{\mathcal{F}}\text{Ш}_{\omega_s} = g \cdot \frac{1}{\omega_s}\text{Ш}_{1/\omega_s} = \frac{1}{\omega_s} g_a \ . \tag{12.7}$$

Remember that g_a is determined by the samples of g. Thus the samples contain the same information as $G * \text{Ш}_{\omega_s}$. In view of identity (12.6) this allows us to obtain G and finally g. In other words we have shown that the samples contain all information about the signal.

We combine Example 12.2 with the formulae (12.6) and (12.7).

$$\begin{aligned} \mathcal{F}g = G &= (G * \text{Ш}_{\omega_s}) \cdot \mathbb{1}_{(-\omega_s/2,\omega_s/2)} \\ &= \mathcal{F}\left(\frac{1}{\omega_s} g_a\right) \cdot \mathcal{F}(\omega_s \operatorname{sinc}(\pi\omega_s \cdot)) \overset{(\star)}{=} \mathcal{F}(g_a * \operatorname{sinc}(\pi\omega_s \cdot)) \end{aligned} \tag{12.8}$$

We now apply $\bar{\mathcal{F}}$ to both sides.

$$g(t) = (g_a * \operatorname{sinc}(\pi\omega_s \cdot))(t) \overset{(\star)}{=} \sum_{k\in\mathbb{Z}} g\left(\frac{k}{\omega_s}\right) \operatorname{sinc}\left(\pi\omega_s\left(t - \frac{k}{\omega_s}\right)\right)$$

This last relation is the sampling theorem, and thus the proof is complete.

However, at the moment the identities marked with $(\star)$ can be understood only formally. This convolution does not fit into the framework of Section 12.3.4.5 because none of the functions has compact support. More seriously, none of the Fourier transforms is in O_M. This means that the products in (12.8) have to be understood as products of functions and cannot be seen as products of a tempered distribution and a function in the sense of Section 12.3.4.5. In fact, the authors do not know of any convolution theorem in the literature that covers this situation. Therefore, it is necessary to give an ad hoc justification of those manipulations which we will do in the next section under slightly more restrictive assumptions on G. Apart from this, the proof is complete at this point.

12.3.4.7 A Convolution Theorem for a Specific Function

The proof of the last section is complete once the following result is proved. We still use the same notation as in the last section.

Lemma 12.2 *Let $G \in L^1(-\Omega, \Omega)$. The following identity holds almost everywhere and (equivalently) in the space of tempered distributions.*

$$\mathcal{F} g_a \mathcal{F} \operatorname{sinc}(\pi\omega_s\cdot) = \mathcal{F}\left(\sum_{k\in\mathbb{Z}} g\left(\frac{k}{\omega_s}\right)\operatorname{sinc}\left(\pi\omega_s\left(\cdot - \frac{k}{\omega_s}\right)\right)\right) \tag{12.9}$$

The series on the right-hand side converges pointwise to a bounded function and thus can be considered to be an element of $\mathcal{S}'$.

We are not able to prove this lemma to this generality without simply paraphrasing the ideas of the first proof. Therefore we will show the result only for $G \in L^2(-\Omega, \Omega)$, but provide a direct proof. In practice this is no severe restriction. For example, bounded functions automatically are in L^2. The only cases where we lose generality is the case of functions having integrable poles that are not square-integrable.

Proof We now assume $G \in L^2(-\Omega, \Omega)$. The idea is to approximate $\operatorname{sinc}(\pi\omega_s\cdot)$ in the norm of L^2 by functions with bounded support to which the convolution theorem applies. We then prove the claim by taking the limit.

To this end, choose any sequence $s_n \in L^2_c(\mathbb{R})$ converging to $s := \operatorname{sinc}(\pi\omega_s\cdot)$ in $L^2(\mathbb{R})$ such that $|s_n| \leq |s|$, for example $s_n(t) := \operatorname{sinc}(\pi\omega_s t)\, \mathbb{1}_{(-n,n)}$. According to the convolution theorem $\mathcal{F} g_a \mathcal{F} s_n = \mathcal{F}(g_a * s_n)$.

First we show $\mathcal{F} g_a \mathcal{F} s_n \to \mathcal{F} g_a \mathcal{F} s$ in $\mathcal{S}'$. For this, fix $\varphi \in \mathcal{S}$. Note that $\mathcal{F} g_a \varphi \in L^2(\mathbb{R})$ by the same arguments as in Example 12.4 (remember that $\mathcal{F} g_a$ is a sum of shifted copies of G). Now we exploit the continuity of the scalar product and the operator $\mathcal{F}$ in L^2 to see that

$$\begin{aligned}\langle \mathcal{F} g_a \mathcal{F} s_n, \varphi\rangle &= \int_{\mathbb{R}} \mathcal{F} g_a \mathcal{F} s_n \varphi = \left(\mathcal{F} s_n \mid \overline{\mathcal{F} g_a \varphi}\right)_{L^2} \\ &\to \left(\mathcal{F} s \mid \overline{\mathcal{F} g_a \varphi}\right)_{L^2} = \int_{\mathbb{R}} \mathcal{F} g_a \mathcal{F} s \varphi = \langle \mathcal{F} g_a \mathcal{F} s, \varphi\rangle\end{aligned}$$

This shows the convergence in $\mathcal{S}'$ to the left-hand side of (12.9).

Next we are going to show that, indeed,

$$h(t) := \sum_{k\in\mathbb{Z}} g\left(\frac{k}{\omega_s}\right)\operatorname{sinc}\left(\pi\omega_s\left(t - \frac{k}{\omega_s}\right)\right)$$

converges for every $t \in \mathbb{R}$ and represents a bounded function. First we identify $g\left(-k/\omega_s\right)/\omega_s$ as being the k-th Fourier coefficient of $G \in L^2\left(-\omega_s/2, \omega_s/2\right)$. By

Bessel's inequality (see [24, Chapter I, Section 7]) the sequence $\left(g\left(k/\omega_{\text{s}}\right)\right)_{k\in\mathbb{Z}}$ is square-summable. Moreover, it can be seen that for every $t \in \mathbb{R}$ the sequence

$$(a_k(t))_{k\in\mathbb{Z}} := \left(\operatorname{sinc}\left(\pi\omega_{\text{s}}\left(t-\frac{k}{\omega_{\text{s}}}\right)\right)\right)_{k\in\mathbb{Z}}$$

is square-summable and that the corresponding series of squares is uniformly bounded for $t \in \mathbb{R}$. In fact, $|a_k(t)| \le 1$ for all k and t. Moreover, for fixed $t \in \mathbb{R}$ we pick $n \in \mathbb{Z}$ such that $n \le \omega_{\text{s}} t \le (n+1)$ and obtain

$$\left|a_k(t)\right| = \left|\frac{(-1)^k \sin(\omega_{\text{s}}\pi t)}{\pi(\omega_{\text{s}} t - k)}\right| \le \begin{cases} (\pi(k-(n+1)))^{-1}\ , & \text{if} \quad k > n+1\ , \\ (\pi(n-k))^{-1}\ , & \text{if} \quad k < n\ . \end{cases}$$

By this we can estimate $\sum_{k=-\infty}^{\infty} |a_k(t)|^2 \le 3 + 2/\pi \sum_{k=1}^{\infty} 1/k^2$ independently of $t \in \mathbb{R}$. Schwarz's inequality (see [24, Chapter I, §9]) now asserts (absolute) convergence of the series $h(t)$ and gives a uniform bound for $|h(t)|$.

Finally we show that $g_a * s_n \to h$ in $\mathcal{S}'$. Let $\varphi \in \mathcal{S}$.

$$\begin{aligned} \langle g_a \ * s_n, \varphi\rangle &= \langle g_a, \check{s}_n * \varphi\rangle = \left\langle g\text{III}_{1/\omega_{\text{s}}}, \int_{\mathbb{R}} s_n(t-\cdot)\varphi(t)\,\mathrm{d}t\right\rangle \\ &= \sum_{k\in\mathbb{Z}} \int_{\mathbb{R}} g\left(\frac{k}{\omega_{\text{s}}}\right) s_n\left(t-\frac{k}{\omega_{\text{s}}}\right)\varphi(t)\,\mathrm{d}t \end{aligned}$$

Note that $\left|g\left(k/\omega_{\text{s}}\right)s\left(t-k/\omega_{\text{s}}\right)\varphi(t)\right|$ is an upper bound of the above integrand. The series $\sum_{k\in\mathbb{Z}}\left|g\left(k/\omega_{\text{s}}\right)s\left(t-k/\omega_{\text{s}}\right)\right|$ is uniformly bounded according to the last paragraph, and of course φ is integrable. This shows that we are in the position to apply Fubini's and Lebesgue's theorems to deduce what we have claimed.

$$\langle g_a \ * s_n, \varphi\rangle \to \int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} g\left(\frac{k}{\omega_{\text{s}}}\right) s\left(t-\frac{k}{\omega_{\text{s}}}\right)\varphi(t)\,\mathrm{d}t = \langle h, \varphi\rangle$$

By continuity of $\mathcal{F}$ on $\mathcal{S}'$ we conclude that $\mathcal{F}(g_a * s_n) \to \mathcal{F}h$. This is the convergence to the right-hand side of (12.9). The lemma is proved. □

We remark that this section can be generalized to the case $G \in L^p(-\Omega, \Omega)$ for any $p > 1$ without much effort by using interpolation theorems to get some decay rate for the Fourier coefficients. However, this does not even cover the case $G \in L^1(-\Omega, \Omega)$ completely, let alone the measures in Remark 12.2.

References

1 Frey, T. and Bossert, M. (**2004**) *Signal- und Systemtheorie*. BG Teubner Verlag.

2 Lindner, J. (**2004**) *Informationsübertragung: Grundlagen der Kommunikationstechnik*. Springer.

3 Lüke, H.D. (**1995**) *Signalübertragung*. Springer.

4 Oppenheim, A.V. and Willsky, A.S. (**1996**) *Signals and Systems*. Prentice Hall.

5 Proakis, J.G. and Manolakis, D.G. (**1996**) *Digital signal processing*. Prentice Hall.

6 Proakis, J.G. (**2000**) *Digital Communications*. McGraw-Hill.

7 Black, H.S. (**1953**) *Modulation theory*. Van Nostrand New York.

8 Cauchy, A.L. (**1841**) Memoire sur diverses formules danalyse. *Comptes Rendus des Séances de l'Académie des Sciences*, **12**(6), 283–298.

9 Borel, E. (**1897**) Sur l' interpolation. *Comptes Rendus des Séances de l'Académie des Sciences*, **124**(13), 673–676.

10 Nyquist, H. (**2002**) Certain topics in telegraph transmission theory. *Proceedings of the IEEE*, **90**(2), 280–305.

11 Küpfmüller, K. (**1928**) Über die Dynamik der selbsttätigen Verstärkungsregler. *Elektrische Nachrichtentechnik*, **5**(11), 459–467.

12 Kotelnikov, V.A. (**1933**) On the carrying capacity of the ether and wire in telecommunications. *All-Union Conference on Questions of Communications. Izd. Red. Upr. Svyazi RKKA, Moscow*.

13 Whittaker, J.M. (**1935**) *Interpolatory function theory*. Cambridge University Press.

14 Ogura, K. (**1920**) On Some Central Difference Formulas of Interpolation. *Tohoku Mathematical Journal*, **17**, 232–241.

15 Shannon, C.E. (**1949**) Communication in the Presence of Noise. *Proceedings of the LRE*, **37**, 10–21.

16 Someya, I. (**1949**) *Waveform Transmission*. Tokyo, Japan: Shyukyoo.

17 Weston, J.D. (**1949**) The cardinal series in Hilbert space. *Proceedings of the Cambridge Philosophical Society*, **45**, 335–341.

18 Weston, J.D. (**1949**) A note on the theory of communication. *Philos. Mag. (303)*, **40**, 449–453.

19 Higgins, J.R. (**1985**) Five short stories about the cardinal series. *American Mathematical Society*, **12**, 45–89.

20 Jerri, A.J. (**1977**) The Shannon sampling theorem – its various extensions and applications: a tutorial review. *Proceedings of the IEEE*, **65**(11), 1565–1596.

21 Lüke, H.D. (**1978**) Zur Entstehung des Abtasttheorems. *Nach. tech. Zeit*, **31**(4), 271–274.

22 Meijering, E. (**2002**) A chronology of interpolation: from ancient astronomy to modernsignal and image processing. *Proceedings of the IEEE*, **90**(3), 319–342.

23 Meyberg, K. and Vachenauer, P. (**2001**) *Höhere Mathematik 2*. Springer.

24 Zygmund, A. (**1968**) *Trigonometrical Series*. Cambridge University Press.

25 Champeney, D.C. (**1989**) *A Handbook of Fourier Theorems*. Cambridge University Press.

26 Gel'fand, I.M. and Shilov, G.E. (**1964**) *Generalized functions, II. Spaces of fundamental and generalized functions*. Academic Press.

27 Dautraym, R. and Lions, J.-L. (**1988**) *Mathematical Analysis and Numerical Methods for Science and Technology, Volume 2: Functional and Variational Methods*. Springer.

28 Gelfand, I.M. and Shilov, G.E. (**1964**) *Generalized Functions: Properties and Operations, Vol. 1*. Academic Press.

13
Coding and Decoding of Algebraic–Geometric Codes

Martin Bossert, Werner Lütkebohmert[1], Jörg Marhenke

13.1
Introduction

Coding theory is a vast area of research and has numerous practical applications in the storage and transmission of digital data. By coding one adds redundancy to the data in order to detect or correct possible errors that occurred during transmission or during the read–write process. In this article we focus on algebraic–geometric codes; and in particular on Reed–Solomon codes. In addition, we describe the concept of forward error correction and give some specific algorithms for the correction of errors.

After an introduction to the concepts of coding theory and methods of decoding, some basic results about Goppa codes, here called algebraic–geometric codes, are explained. The coding procedure for the latter codes merely consists of computing Riemann–Roch spaces which can be done using methods from algebraic number theory. The decoding method for hard decoding up to half the minimum distance and the one beyond half the minimum distance, which are due to Feng-Rao and Sudan, respectively, are presented. A specific algorithm for the decoding of Reed–Solomon codes beyond half the minimum distance is sketched. Furthermore, an algorithm for the use of possibly existing reliability information, so called soft-decision, is described.

13.2
Introduction to Linear Codes

Let $\mathbb{F} := \mathbb{F}_q$ be a finite field with q elements; for example, let $\mathbb{F} = \mathbb{Z}/p\mathbb{Z}$ be the prime field with p elements. The Hamming weight on the $\mathbb{F}$-vector space $\mathbb{F}^n$

$$\mathrm{wt}(x) := \#\{\, i \,;\; x_i \neq 0 \,\} \quad \text{for} \quad x = (x_1, \ldots, x_n) \in \mathbb{F}^n$$

[1] Corresponding author.

gives rise to a distance function

$$d(x,y) := \mathrm{wt}(x-y) \quad \text{for} \quad x,y \in \mathbb{F}^n$$

For a nonempty subset $C \subset \mathbb{F}^n$ the minimum distance of C is defined by

$$d(C) := \min\{d(x,y)\ ; \quad \text{for} \quad x,y \in C \quad \text{with} \quad x \neq y\}$$

For any $a \in \mathbb{F}^n$ and $r \in \mathbb{R}$ we denote by

$$\mathbb{B}(a,r) := \{x \in \mathbb{F}^n\ ;\ d(a,x) \leq r\}$$

the ball at a with radius r. For $t \in \mathbb{R}$ with $1 + 2t = d(C)$ the balls

$$\mathbb{B}(c,t) \cap \mathbb{B}(c',t) = \emptyset \quad \text{for} \quad c,c' \in C \quad \text{with} \quad c \neq c'$$

are disjoint; this number t is called the *error-correction bound*.

Fixing the parameters (n,d) and the base field with q elements, one can look at the number

$$A_q(n,d) := \max\{\,M\ ; \quad \text{there exists} \quad C \subset \mathbb{F}^n \quad \text{with} \quad \#C = M\ ;\ d(C) \geq d\,\}$$

Quite often one is interested in the ratios

$$\begin{array}{ll} R(C) := (\log_q \#C)/n & \text{the information rate} \\ \delta(C) := d(C)/n & \text{the relative minimal distance .} \end{array}$$

Upper bounds are easy to prove. A simple but important bound is

Theorem 13.1 (Singleton Bound) *For a code with parameters* (n,d) *the following holds*

$$\log_q(\#C) + d \leq n + 1\ .$$

For a relative distance $0 \leq \delta \leq 1$ one can consider

$$a_q(n,\delta) := \limsup_{n\to\infty} \frac{1}{n} \log_q A_q(n,\delta n)\ .$$

This is the asymptotic maximal information rate with fixed relative distance. Due to the Singelton Bound one has

$$1 - \delta \geq a_q(\delta) \quad \text{for} \quad 0 \leq \delta \leq 1\ .$$

For lower bounds one considers the entropy function which is defined for $0 \leq t \leq (q-1)/q$ by

$$H_q(t) := t\log_q(q-1) - t\log_q t - 1 - (1-t)\log_q(1-t)\ .$$

By simple combinatorial arguments one can show:

Theorem 13.2 (Gilbert–Varshamov Bound)

$$a_q(\delta) \geq 1 - H_q(\delta) \quad \textit{for} \quad 0 \leq \delta \leq (q-1)/q$$

A linear code is a linear subspace $C \subset \mathbb{F}^n$*, its dimension is* $k = \log_q(\#C)$*, its length* n *is equal to the dimension of the ambient space* $\mathbb{F}^n$*, and its minimum distance is* $d := d(C) = \min\{\mathrm{wt}(x)\ ;\ x \in C - \{0\}\}$*. The Hamming bound for a linear* (n, k, d) *code over the alphabet of size* q *is given by*

$$q^k \left(\sum_{i=0}^{\lfloor d-1/2 \rfloor} \binom{n}{i} (q-1)^i \right) \leq q^n$$

and counts the number of vectors of the space which lie inside the spheres.

In the sequel we will consider linear codes only. For such codes we have defined parameters (n, k, d) above, hence, $R(C) = k/n$ and $\delta(C) = d/n$. It is known that linear codes exists with rate δ and information rate greater than $1 - H_q(\delta)$. More precisely one has the following result which, for $q \geq 49$, gives a lower bound which improves the Gilbert–Varshamov-bound in a certain range of δ.

Theorem 13.3 (Tsfasman–Vlădut–Zink) *If* q *is a square, then*

$$a_q(\delta) \geq \left(1 - \frac{1}{\sqrt{q}-1}\right) - \delta \quad \textit{for} \quad 0 \leq \delta \leq 1 - \frac{1}{\sqrt{q}-1}$$

This relies mainly on the idea of Goppa constructing linear codes by methods from algebraic geometry which we will describe in Section 13.4.

13.3 Introduction to Forward Error Correction

When digital data is transmitted, or when it is stored on a medium, errors occur due to statistical disturbances. The statistics of the errors only are important for forward error correction and therefore channel models are used in order to describe the error occurrence. Shannon has defined in his landmark paper *A mathematical theory of communication* [1] the capacity of a channel and has proved that reliable transmission is only possible when a code of rate less than the capacity is used. In the following we will describe a selection of channel models and the basic concepts of forward error correction methods.

The channel is characterized by the conditional probabilities $\mathrm{pr}(y|x)$ that a symbol y from a finite or infinite alphabet is received when a symbol x from the code alphabet is sent via the channel.

An *error* has occurred if $y \neq x$ in the case that the channel output alphabet is identical to the code alphabet.

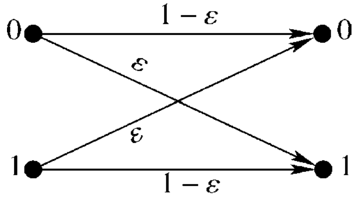


Figure 13.1 Binary symmetric channel.

For memoryless channels the reception of a vector $y = (y_1, y_2, \dots, y_n)$ when the codeword $x = (x_1, x_2, \dots, x_n)$ was transmitted has the probability

$$\mathrm{pr}(y|x) = \prod_{i=1}^{n} \mathrm{pr}(y_i|x_i)$$

because the transmissions of the individual symbols are independent of each other. In the following we introduce the two most common channel models.

13.3.1 Binary Symmetric Channel, BSC

The input and output alphabet of the BSC is $\{0, 1\}$ and the conditional probabilities are $\mathrm{pr}(0|0) = \mathrm{pr}(1|1) = 1 - \varepsilon$ and $\mathrm{pr}(0|1) = \mathrm{pr}(1|0) = \varepsilon$ which explains the name symmetric. In Figure 13.1 the BSC is depicted. If we transmit codewords of length n the probability $\mathrm{pr}(\tau)$ that τ errors occur in the codeword is

$$\mathrm{pr}(\tau) = \binom{n}{\tau} \varepsilon^{\tau} (1 - \varepsilon)^{n-\tau}$$

which follows the binomial distribution. The received vector y can be calculated by adding an error vector e to the transmitted codeword c, that is $y = c + e$ (note that the addition is in the field).

13.3.2 Additive White Gaussian Noise Channel, AWGN

The input alphabet is binary, namely $\{1, -1\}$. Usually binary codes are defined over the field $\{0, 1\}$ but due to signal transmission the mapping $[1 \mapsto 0\ , -1 \mapsto 1]$ is introduced. The channel output consists of real numbers $y \in \mathbb{R}$. Figure 13.2 shows the conditional probability densities $\mathrm{pd}(y_i|x_i = 1)$ and $\mathrm{pd}(y_i|x_i = -1)$ which determine the probability that $y_i \in [a, b]$ lies in the interval $[a, b]$

$$\mathrm{pr}(y_i \in [a, b]|x_i) = \int_a^b \mathrm{pd}(y_i|x_i)\,\mathrm{d}y_i = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i)^2/2\sigma^2}\,\mathrm{d}y_i\ .$$

The variance σ^2 determines the quality of the channel: the larger the variance the larger the number of errors. Note that an error occurs only if the sign of the received y is different from the sign of the transmitted symbol.

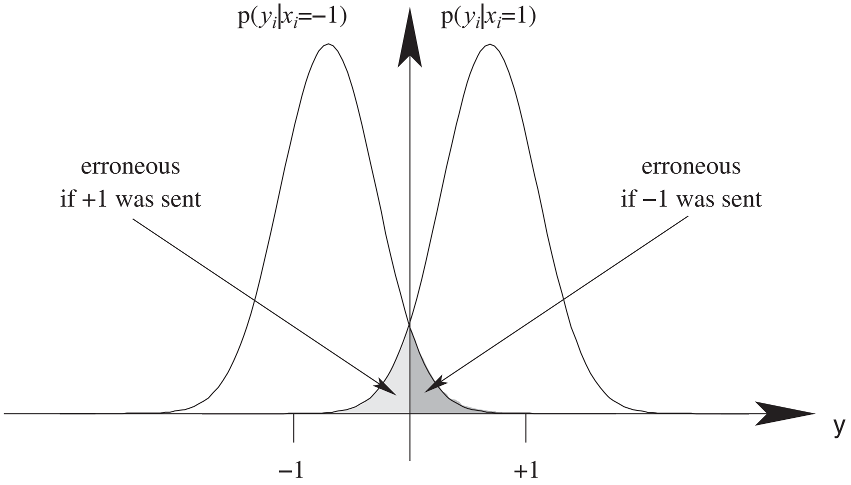


Figure 13.2 Additive white Gaussian noise channel with normalized binary phase shift keying.

Whenever the output alphabet is larger than the input or code alphabet, one can, in addition, use reliability information. Consider the so-called L-value

$$L_i = \log \frac{\mathrm{pr}(x_i = 1|y_i)}{\mathrm{pr}(x_i = -1|y_i)} \ . \tag{13.1}$$

The sign of L_i is the binary decision for the symbol. If the probability for –1 is larger, the fraction is smaller than 1 and the logarithm is negative. However, the absolute value of $|L_i|$ represents the reliability of the decision. The larger $|L_i|$ the larger the difference between the two probabilities.

13.3.3
Maximum A Posteriori (MAP) and Maximum Likelihood (ML) Decoding

A decoder should decide on the maximal probability $\mathrm{pr}(x|y)$ for an input word x and received word y. Clearly the assumption is that the decoder knows the channel statistics, observes y, and knows which code C is used. Applying Bayes rule we can calculate

$$\arg\max_{x\in C} \mathrm{pr}(x|y) = \arg\max_{x\in C} \frac{\mathrm{pr}(y|x)\cdot \mathrm{pr}(x)}{\mathrm{pr}(y)} = \arg\max_{x\in C} \mathrm{pr}(y|x)\cdot \mathrm{pr}(x) \ .$$

Now we can distinguish between two cases, first the decoder has knowledge about the a priori probability $\mathrm{pr}(x)$, a codeword x that is transmitted, and second, that this is not known. In the first case the decoder does a maximum a posteriori (MAP) decision according to the above equation, and in the second case it is called a maximum likelihood (ML) decision as follows

$$\arg\max_{x\in C} \mathrm{pr}(x|y) = \arg\max_{x\in C} \mathrm{pr}(y|x)$$

ML decoding is used when the a priori probabilities are not known or they are identical for all codewords.

13.3.4 Hard- and Soft-Decision Decoding

If the decoder cannot use reliability information, according to (13.1) the decoding method is called hard-decision decoding. Note that this can have two causes, first that the channel model does not give reliability information as, for example, the BSC; or that the algorithm used is not able to include the additional reliability information in its decision. The case when reliability information is used is called soft-decision decoding and is preferable, if possible, because the decoding performance is improved considerably.

Unfortunately, ML decoding is computationally too complex in many cases. Therefore one must use suboptimal decoding algorithms which are described in the following.

13.3.5 Bounded Distance Decoding

The name bounded distance is given to all decoding algorithms which put spheres around the codewords of a certain radius in the Hamming metric (hard-decision). All the received vectors which are in one of the spheres are decoded by the codeword which is the center of the sphere. Clearly, the radius t of the spheres can be less than half the minimum distance of the code, equal to or larger than, respectively. In the first case the decoding capability of the code is not used fully. The second case is called bounded minimum distance decoding and is the standard case. Note that this is not ML decoding because the number of vectors outside the spheres with radius half the minimum distance is larger than 0 for all code classes except the so-called perfect codes which fulfill the Hamming or sphere-packing bound with equality. For the case when we have a decoding algorithm which is able to decode beyond half the minimum distance, the decision might not be unique because two or more codewords may exist with the same distance to the received vector. If the decoder provides a list of possible codewords it is called a list decoder.

13.4 Algebraic–Geometric Codes

The famous Reed–Solomom codes (Example 13.1) are generalized by algebraic–geometric codes. The basic ideas of constructing linear codes by algebraic curves were invented by Goppa [2]. There are two approaches to algebraic–geometric curves.

The first one starts with a function field of transcendance degree 1; say $F := \mathbb{F}(\xi, \eta)$ with one relation $f(\xi, \eta) = 0$ for some absolutely irreducible polynomial $f \in \mathbb{F}[T_1, T_2]$. The geometric object behind it is the set of the discrete valuations $\nu : F \to \mathbb{Z}$ with $\nu|_{\mathbb{F}} = 0$. A morphism of function fields is a morphism of fields inducing the identity on the base field.

The second approach starts with an absolutely irreducible polynomial f as above. To describe the geometric object, one looks first at the vanishing locus $V(f) \subset \mathbb{A}^2_{\mathbb{F}}$; that is the set of all geometric points in the affine 2-plane $\mathbb{A}^2_{\mathbb{F}}$ which are zeros for f. Then one takes the topological closure $Y \subset \mathbb{P}^2_{\mathbb{F}}$ in the projective space of $V(f)$; that is $Y = V(\tilde{f}) \subset \mathbb{P}^2_{\mathbb{F}}$, where

$$\tilde{f}(T_0, T_1, T_2) := T_0^{\deg f} \cdot f\left(\frac{T_1}{T_0}, \frac{T_2}{T_0}\right) \in \mathbb{F}[T_0, T_1, T_2]$$

is the homogenization of f. If Y is smooth, then Y is already the geometric object. Unfortunately, Y may happen to have singular points. Therefore, one has to resolve the singularities by blowing-up or by performing the normalization of Y and, hence, gets the associated smooth projective curve $X \to Y$ which lives in some projective space $\mathbb{P}^N_{\mathbb{F}}$. The morphisms of curves are the mappings locally defined by polynomials.

There is an equivalence of categories between function fields and smooth projective algebraic curves. The valuations of a function field F correspond bijectively to the geometric points of the curve X. The *function field* associated to a curve is the field of rational functions on the curve. A smooth projective curve has genus $g := \dim_{\mathbb{F}} \Omega^1_X(X)$ being the dimension of the $\mathbb{F}$-vector space of regular differential 1-forms.

For the following, let us fix a smooth algebraic curve $X \subset \mathbb{P}^N_{\mathbb{F}}$ with function field F as above. A point $P \in X$ is called *rational* if its coordinates in the ambient space $\mathbb{P}^N_{\mathbb{F}}$ are numbers in the base field $\mathbb{F}$; in terms of the corresponding valuation $\nu : F \to \mathbb{Z}$, which means that the residue field of ν is equal to the base field $\mathbb{F}$. Moreover, let us fix an n-tuple $E := (P_1, \dots, P_n)$ of mutually different rational points of X and a divisor D on X whose support is disjoint from E. The divisor D determines a finite-dimensional $\mathbb{F}$-vector space of rational functions

$$L(X, D) := \{ \phi \in F \; ; \; \mathrm{div}(\phi) + D \geq 0 \} .$$

The ordered set E of rational points gives rise to a linear map

$$\mathrm{ev}_E : L(X, D) \to \mathbb{F}^n \; ; \; \phi \mapsto (\phi(P_1), \dots, \phi(P_n))$$

by evaluating functions at E. The image of ev_E is a linear code $C_L(E, D)$ of length n. The crucial point of this construction is that one can exactly determine the dimension of the code by the theorem of Riemann–Roch and estimate the minimum distance due to the fact that the number of zeros of a rational function is equal to the number of poles; the latter is bounded by $\deg(D)$. A further big advantage of algebraic–geometric curves is the geometric interpretation of the parity check matrix. Namely, there is a dual construction by taking residues

$$\mathrm{res}_E : \Omega^1(X, E - D) \to \mathbb{F}^n \; ; \; \omega \mapsto (\mathrm{res}_{P_1}(\omega), \dots, \mathrm{res}_{P_n}(\omega)) .$$

The image of res_E is a code $C_\Omega(E, D) \subset \mathbb{F}^n$ and its parameters can also be estimated. The canonical pairing $\mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F} \; ; \; (x, y) \mapsto \sum_{i=1}^n x_i y_i$ induces a commutative diagram

$$\begin{tikzcd}
L(X, D) \times \Omega^1(X, E - D) \arrow[r] \arrow[d, "\mathrm{ev}_E"', "\mathrm{res}_E"] & \mathbb{F}\,;\ (\phi, \omega) \mapsto \sum_{P \in E} \mathrm{res}_P(\phi \cdot \omega) \arrow[d, "\mathrm{id}"] \\
\mathbb{F}^n \times \mathbb{F}^n \arrow[r] & \mathbb{F}\ ;\ (x, y) \mapsto \sum_{i=1}^{n} x_i y_i
\end{tikzcd}$$

Since the upper horizontal map is zero due to the residue theorem, we see that $C_\Omega(E, D)$ is orthogonal to $C_L(E, D)$. Summarizing we state the result as follows.

Theorem 13.4 *Keep the above situation. Assume $n := \#E > \deg(D) \geq 2g - 1$. Then the following holds:*
(1.1) The map ev_E is injective and $\dim_{\mathbb{F}} C_L(E, D) = \deg(D) + 1 - g$.
(1.2) The minimum distance fulfills $d(C_L(E, D)) \geq n - \deg(D)$.
(2.1) The map res_E is injective and $\dim_{\mathbb{F}} C_\Omega(E, D) = n - \deg(D) + g - 1$.
(2.2) The minimum distance fulfills $d(C_\Omega(E, D)) \geq \deg(D) + 2 - 2g$.
(3) Under the canonical pairing $\mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}$ the codes $C_L(E, D)$ and $C_\Omega(E, D)$ are orthogonal complements to each other.

Example 13.1 *Let $X = \mathbb{P}^1_{\mathbb{F}}$ be the projective line with point P_∞ at infinity. For $k, n \in \mathbb{N}$ with $q \geq n > k \geq 1$ set $D := (k-1) \cdot P_\infty$ and let $E = (P_1, \ldots, P_n)$ be a set of rational points on the affine line. Then $g = 0$ and*

$$C_L(E, D) := \{(\phi(P_1), \ldots, \phi(P_n))\ ;\ \phi \in \mathbb{F}[T]\ ,\ \deg \phi < k\}$$
$$C_\Omega(E, D) := \left\{ \left(\frac{\phi(P_1)}{\Phi'(P_1)}, \ldots, \frac{\phi(P_n)}{\Phi'(P_n)} \right)\ ; \phi \in \mathbb{F}[T]\ , \deg \phi \leq \deg \Phi - k - 1 \right\}$$

where $\Phi(T) = (T - T(P_1)) \cdot (T - T(P_n))$ and $\Phi'(T)$ is the formal derivative. The code $C_L(E, D)$ is the famous Reed–Solomon code.

A linear code C has high density if $R(C) + \delta(C)$ is close to 1. For example, the Reed–Solomon code satisfies

$$\frac{\dim C_L(E, D)}{n} + \frac{d(C_L(E, D))}{n} = \frac{k}{n} + \frac{n - k + 1}{n} = 1 + \frac{1}{n}\ .$$

Unfortunately, n is bounded by the number $q := \#\mathbb{F}$ of elements of the base field. The advantage of more general algebraic–geometric codes is that the number of rational points is not bounded by q. In the case $g \geq 1$, Theorem 13.4 yields

$$R(C) + \delta(C) \geq \frac{\deg(D) + 1 - g}{n} + \frac{n - \deg(D)}{n} = 1 + \frac{1 - g}{n}$$

where one can choose $n = \#C(\mathbb{F})$ as the number of rational points of C. Due to a theorem of Weil, one knows $|\#C(\mathbb{F}) - (q + 1)| \leq 2g \cdot \sqrt{q}$. Thus, to produce codes with high density, of arbitrarily large length, amounts to constructing curves with many rational points compared to its genus. This is precisely what is done in proving Theorem 13.3.

Theorem 13.5 *If q is a square, then*

$$\limsup_{g\to\infty} \frac{\max\{\#X(\mathbb{F}_q)\ ; X/\mathbb{F}_q \text{ curve with } \mathrm{g}(X) = \mathrm{g}\}}{g} = \sqrt{q} - 1 \ .$$

Example 13.2 *Let $q = \ell^2$ be a square. The Hermite curve is defined by the locus*

$$X := V(T_0^{\ell+1} + T_1^{\ell+1} + T_2^{\ell+1}) \subset \mathbb{P}^2_{\mathbb{F}} \ .$$

It is a smooth curve of genus is $\ell(\ell-1)/2$ and it has $\ell^3 + 1$ points given explicitly by

$$X(\mathbb{F}) = \left\{ \begin{matrix} (1,\alpha,\beta)\ ;\ 1 + \alpha^{\ell+1} = -\beta^{\ell+1} & \textit{with}\ \ \alpha \in \mathbb{F}\,,\, \beta \in \mathbb{F}^{\times} \\ (1,\alpha,0)\ ;\ 1 = -\alpha^{\ell+1} & \\ (0,1,\beta)\ \ ;\ 1 = -\beta^{\ell+1} \quad \textit{with}\ \ \beta \in \mathbb{F} & \end{matrix} \right\} .$$

For example $P := (0,1,1)$ belongs to X if ℓ is a power of 2. A basis of $L(X, N \cdot P)$ is given by the elements

$$f_{i,j} := \frac{T_0^i T_1^j}{(T_1 + T_2)^{i+j}} \quad \textit{for} \quad 0 \le i \le \ell\,, 0 \le j\,, (\ell + 1)\cdot(i+j) - i \le N$$

The pole order of $f_{i,j}$ is $\ell \cdot i + (\ell + 1)\cdot j$. Evaluating the functions $(f_{i,j})$ at the rational points gives a basis of the Hermite code in $\mathbb{F}^n$ for $n := N + 1 - g$ if $N > \ell(\ell - 1) - 2$.

13.5 Computation of Riemann–Roch Spaces

Given a smooth projective curve X over $\mathbb{F}$ equipped with an n-tuple E of rational points and a divisor D on X with $\mathrm{supp}(D) \cap E = \emptyset$, the construction of the corresponding algebraic–geometric code amounts to finding a basis of the Riemann–Roch space $L(X, D)$. Namely, evaluating such a basis at E produces a generator matrix of $C_L(E, D)$. In this section we will briefly explain an effective method of producing such a basis. The tools are taken from algebraic number theory. The main steps are discussed in the following:

As in Section 13.4 we start with an affine coordinate ring

$$B := \mathbb{F}[\xi, \eta]/(f) \quad \text{where} \quad f \in \mathbb{F}[T_1, T_2] \quad \text{is monic in} \quad T_2$$

of $\tilde{X} := V(T_0^{\deg f} f(T_1/T_0, T_2/T_0)) \subset \mathbb{P}^2_{\mathbb{F}}$. Then $\mathbb{F}[\xi]$ is a free polynomial ring and B is a finitely generated free $\mathbb{F}[\xi]$-module of rank $N := \deg_{T_2}(f)$. First one computes the normalization A of B; this is exactly the ring of regular functions on $U := X - \mathrm{Pol}(\xi)$ where $\mathrm{Pol}(\xi)$ denotes the set of poles of ξ. One knows that B and A have the same field of fractions and that A is a free $\mathbb{F}[\xi]$-module of the same rank as B. Then one does the same for $A' := \mathcal{O}(U')$ where $U' := X - V(\xi)$ the complement of the zeros of ξ. Actually it suffices, to do this only in a neighborhood of $V(\xi)$.

Then it is easy to determine a basis of any ideal $\mathcal{I} \subset A$ resp. A' due to the theory of elementary divisors. Now consider a divisor D of X. As above, one can calculate a basis of

$$\begin{aligned}\mathcal{I}(D|_U) &= \mathbb{F}[\xi]f_1 \otimes \ldots \otimes \mathbb{F}[\xi]f_N \\ \mathcal{I}(D|_{U'}) &= \mathbb{F}[\xi]f'_1 \otimes \ldots \otimes \mathbb{F}[\xi]f'_N\end{aligned}$$

Again, by the theory of elementary divisors, one can arrange these bases in such a way that

$$f_i = \xi^{n_i} f'_i \quad \text{for} \quad i = 1, \ldots, N \,.$$

Finally an $\mathbb{F}$-basis of $L(\tilde{X}, D)$ is given by the system

$$\left(\xi^{\lambda_i} f_i \,; 0 \le \lambda_i \le -n_i \,, i = 1, \ldots, N\right)$$

For the computation of the normalization we can make use of the criterion for normality due to Grauert and Remmert [3, Appendix 3.3, Rule 7].

Lemma 13.1 *Let B be a reduced noetherian ring and let $\mathcal{I} \subset B$ be an ideal with the following properties*

(i) *$\mathcal{I}$ contains a nonzero divisor of B,*

(ii) *$\mathcal{I}$ is reduced,*

(iii) *the non-normal locus of B is contained in $V(\mathcal{I})$.*

Then B is normal if and only if $B = \mathrm{Hom}_B(\mathcal{I}, \mathcal{I})$.

If B is not normal we have $B \subsetneq B' := \mathrm{Hom}_B(\mathcal{I}, \mathcal{I})$. Moreover, B' is a commutative ring and canonically contained in A; that is $B \subsetneq B' \subset A$. Then we replace B by B'. By repeating this process, one obtains the normalization A since, due to the fact that A is a noetherian $R[\xi]$-module, the process produces a situation $B = B'$ which is then equal to the normalization. The representation of $\mathrm{Hom}_B(\mathcal{I}, \mathcal{I})$ as a B-algebra of finite type is slightly awkward, a full description of the entire algorithm can be found in [4].

Of course, this method to compute the normalization can be applied to the ring B defined above. A suitable ideal $\mathcal{I}$ is generated by the discriminant of B over $\mathbb{F}[\xi]$. Unfortunately, if X has many ramification points, the computations become long and complicated. To overcome this problem, Pohst and Zassenhaus have developed the Round-2 algorithm which carries out the above enlargement process locally for a prime factor of the discriminant. Thus the algorithm becomes more efficient. Further improvement comes from the use of a criterion due to Dedekind which simplifies the first step of enlargement for each prime.

Proposition 13.1 (Dedekind) *Let B and f be as before and let $p \in \mathbb{F}[T_1]$ be a prime element. Let $\bar{f} = \prod_{i=1}^{s} \bar{f}_i^{e_i}$ be the factorization of the reduction $\bar{f}$ of f modulo p and set*

$g := \prod_{i=1}^{s} f_i$ *where* f_i *are monic liftings of* $\overline{f_i}$. *Let* h *be a monic lifting of* $\overline{f}/\overline{g}$ *and set*

$$\psi = \frac{1}{p}(gh - f) \in (\mathbb{F}[T_1])[T_2] \ .$$

Then B *is normal above* p *if and only if* $\gcd(\overline{\psi}, \overline{g}, \overline{h}) = 1$.
If this is not the case, let U *be a monic lift of* $\overline{f}/\gcd(\overline{\psi}, \overline{g}, \overline{h})$. *Then*

$$B' := B + \frac{1}{p}UB \subseteq A$$

is a strict enlargement of B.

With this theorem, we can decide, for a given prime p, whether B is normal above p. If this is not the case, the theorem gives an enlarged algebra $B' \subseteq A$. Hence, the first step of enlargement for each prime can easily be computed. Then one continues with the Round-2 algorithm. Originally, the Round-2 algorithm was used in algebraic number theory. A detailed description of the Round-2 algorithm for this case can be found in [5].

13.6
Decoding up to Half the Minimum Distance

In this section we will give an overview of the algorithm of Feng and Rao for decoding an algebraic–geometric code if the weight of the error is below half the minimum distance; see [6]. We will construct a bounded minimum distance decoder; that is an algorithm to compute the retraction map

$$\varrho : \bigcup_{c \in C} \mathbb{B}(c, t) \to C \ , \quad c + x \mapsto c \quad \text{for} \quad x \in \mathbb{B}(0, t) \ .$$

Let us start with following data

X	smooth projective curve over $\mathbb{F}$ of genus g
$P, P_1, \ldots, P_n$	mutually different rational points of X
$E := (P_1, \ldots, P_n) \subset X(\mathbb{F})$	n-tuple of rational points
$D := (2g + 2t - 1) \cdot P$	$n > \deg(D) \geq 2g - 1$.

Due to Theorem 13.4 we know

$$\begin{array}{ll} \dim_{\mathbb{F}} C_L(E, D) = g + 2t & \dim_{\mathbb{F}}(C_\Omega)(E, D)) = n - g - 2t \\ d(C_L(E, D)) \geq n - \deg(D) & d(C_\Omega)(E, D)) \geq 2t + 1 \ . \end{array}$$

So we can repair t errors in a word of $C_\Omega(E, D)$. Consider a vector

$$a := c + e \in \mathbb{F}^n \quad \text{with} \quad c \in C_\Omega(E, D) \ , \ e \in \mathbb{B}(0, t) \ .$$

In order to determine the error e, we first try to find the positions

$$M(e) := \{\, i \,;\, e_i \neq 0 \,\}$$

of the defective components of a and define the corresponding divisor

$$M := \sum_{i \in M(e)} P_i \ .$$

Next we define the commutative diagram of bilinear forms

$$\begin{tikzcd}
\mathbb{F}^n \times L(X,D) \arrow[r] \arrow[d, "{(\mathrm{id},\mathrm{ev}_E)}"] & \mathbb{F} \arrow[d, "\mathrm{i}d"] & ;\ (a,\phi) \mapsto [a\,,\phi] := \sum_{i=1}^{n} a_i \phi(P_i) \\
\mathbb{F}^n \times C_L(E,D) \arrow[r] & \mathbb{F} & ;\ (x,y) \mapsto x \cdot y := \sum_{i=1}^{n} x_i y_i
\end{tikzcd}$$

We know $[a,\phi] = [e,\phi]$ for $\phi \in L(X,D)$ since $L(X,D)$ is orthogonal to $\Omega^1(X, E-D)$. The way to estimate the set $M(e)$ is to look at

$$\begin{aligned} L(\tilde{t}\cdot P - M) = \mathrm{Ker}\,(L(\tilde{t}\cdot P) &\to \mathrm{Hom}_{\mathbb{F}}(L((2g+t-1)\cdot P), \mathbb{F})) \\ \sigma &\mapsto (h \mapsto [e, h\sigma]) \end{aligned}$$

By the theorem of Riemann–Roch one knows that $L(\tilde{t}\cdot P - M) \neq 0$ for some $\tilde{t} \leq g + t$. Unfortunately, when computing this kernel, elements $h\sigma$ lying in $L((2g + t + \tilde{t} - 1)\cdot P)$ are involved, but we only know $[e,\phi]$ for elements $\phi \in L((2g+2t-1)\cdot P)$. So the required elements are known only for $\tilde{t} \leq t$. Therefore one proceeds stepwise starting with $\tilde{t} = t$. If $L(\tilde{t}\ \cdot P - M) \neq 0$, then we can compute an element in the kernel. If $L(\tilde{t}\ \cdot P - M) = 0$, one needs the algorithm of Feng and Rao to calculate the elements required to determine the kernel in the case $\tilde{t} + 1$.

In the following, we explain the idea of the algorithm. We consider functions $\phi_\lambda \in L(X, \lambda\cdot P)$ with pole order λ at P, if such a function exists, otherwise we set $\phi_\lambda := 0$. Then define the syndromes

$$S_{\lambda,\lambda'} := [e\,, \phi_\lambda \phi_{\lambda'}] \quad \text{for} \ \ \lambda + \lambda' \leq s := (\tilde{t}+1) + 2g + t - 1 \ .$$

The $S_{\lambda,\lambda'}$ in the given range are precisely the elements needed to compute the kernel $L((\tilde{t}+1)\cdot P - M)$. They are known for $\lambda + \lambda' \leq \tilde{t} + 2g + t - 1$ by assumption. We need to know $S_{\lambda,\lambda'}$ for $\lambda + \lambda' = s = \tilde{t} + 2g + t$. It is easy to see that one knows these elements if one of them is known. This means that from each element $S_{\lambda,\lambda'}$ one can calculate the syndrome $[e, \phi_s]$. Due to the condition $L(\tilde{t}\ \cdot P - M) = 0$, one can guess the elements $S_{\lambda,\lambda'}$ and, hence, $[e, \phi_s]$. For theoretical reasons, the majority of these guesses is correct. Thus the majority determines the whole diagonal $S_{\lambda,\lambda'}$ for $\lambda + \lambda' = s$.

Finally, after at most g steps, we end up with a function $\sigma \in L(\tilde{t} \ \cdot P - M)$ with $\sigma \neq 0$. The number of zeros of σ is bounded by $\tilde{t} \leq g + t$. So we have estimated the defective positions $M(e)$ by the zeros $Z(\sigma)$ of σ. Finally it remains to solve the linear system of equations

$$\sum_{j \in Z(\sigma)} X_j \cdot \phi_\lambda(P_j) = [e\,, \phi_\lambda] \quad \text{for } \lambda = 0, \ldots, s := \tilde{t} + 2g + t - 1 \ .$$

We know that there is a solution; namely $x = e$. This solution is unique. Namely, if $\tilde{e}$ is a further solution, the difference $e - \tilde{e}$ belongs to $C_\Omega(E, s \cdot P)$, since it is orthogonal to $C_L(E, s \cdot P)$. It has weight $\mathrm{wt}(e - \tilde{e}) \leq \tilde{t} + t$ which is smaller that the minimum distance of $C_\Omega(E, s \cdot P)$ due to Theorem 13.4 which is $s + 2 - 2g \geq \tilde{t} + t + 1$. This shows the uniqueness of the solution.

Example 13.3 *Consider the Hermite curve of Example 13.2 over the field with* $q := 64$ *elements which is given by the equation*

$$X := V(T_0^9 + T_1^9 + T_2^9) \subset \mathbb{P}^2_{\mathbb{F}} \ .$$

It has genus $g = 28$. *Let* $P := (0, 1, 1)$ *and* $E := (P_1, \ldots, P_{512})$ *be the set of all rational points of X different from P. For* $t = 114$, *we obtain the code* $C_\Omega(E, 283 \cdot P)$. *It has length* 512, *dimension* 256, *minimum distance* ≥ 229. *Its information rate is* 0.5 *and its relative minimal distance is* ≥ 0.447; *so it is of high density. It can repair up to* 114 *errors in a received word. Using the algorithm explained above, a modern computer will do the decoding in a second. If there are* $\tilde{t}$ *defective positions in a received word, the algorithm usually needs* $\max\{0, \tilde{t} - 86\}$ *steps to compute unknown diagonals of the matrix* $(S_{\lambda,\lambda'})$.

Example 13.4 *For Reed–Solomon codes, the decoding is much simpler and by using the Euclidean algorithm the computations can be done very quickly. To explain this, we transform the Reed–Solomon code defined in Example 13.1 by means of the discrete Fourier transformation in* $\mathbb{F}$

$$\mathcal{F} : \mathbb{F}^n \to L(\mathbb{P}^1, (n-1) \cdot P_\infty) \ ; (a_0, \ldots, a_{n-1}) \mapsto \frac{1}{n} \sum_{i=0}^{n-1} a_i T^i \ .$$

For natural numbers k, t with $n := q - 1 \geq k \geq 1$ *and* $n - k = 2t$ *we define*

$$RS(k) := \{C \in \mathbb{F}[T]/(T^n - 1) \ ; \ C(\beta^\ell) = 0 \quad \textit{for} \ \ 1 \leq \ell \leq 2t\}$$

where β *is a generator of the multiplicative group of* $\mathbb{F}$. *Then we set*

$$B := (\beta^1, \ldots, \beta^n) \ .$$

The Fourier transform gives rise to an isomorphism

$$\mathcal{F} \, : C_L(B, (k-1) \cdot P_\infty) \xrightarrow{\sim} RS(k)$$

of the Reed–Solomon code defined in Example 13.1 to $RS(k)$ preserving the weight. Namely, for any polynomial $f = f_0 + \ldots + f_{n-1}T^{n-1} \in \mathbb{F}[T]$ with evaluation $c := \mathrm{ev}_B(f)$, the Fourier transform $F(T) := \mathcal{F}(c)$ fulfills

$$F(\beta^{-i}) = \frac{1}{n}\sum_{j=0}^{n-1} f(\beta^j)\beta^{-ij} = \frac{1}{n}\sum_{j=0}^{n-1}\sum_{\ell=0}^{n-1} f_\ell \beta^{j\ell-ij} = \frac{1}{n}\sum_{\ell=0}^{n-1} f_\ell \sum_{j=0}^{n-1}\beta^{(\ell-i)j} = f_i \,. \tag{13.2}$$

Therefore $F(\beta^{-i}) = 0$ for $k \le i \le n-1$ iff $f \in L(\mathbb{P}^1, (k-1)\cdot P_\infty)$. Moreover, one has $(\beta^{-(n-1)}, \ldots, \beta^{-k}) = (\beta^1, \ldots, \beta^{2t})$ as $\beta^n = 1$.
For decoding, one prefers the representation of codewords by their Fourier transforms. Decoding a received word $R = C + E$ where the error polynomial $E = E_0 + \ldots + E_{n-1}T^{n-1} \in \mathbb{F}[T]$ satisfies $\mathrm{wt}(E) \le t$ and where $C \in RS(k)$ amounts to solving the congruence

$$\omega \equiv \sigma \cdot \sum_{\ell=1}^{2t} R(\beta^\ell)T^\ell = \sigma \cdot \sum_{\ell=1}^{2t} E(\beta^\ell)T^\ell \mod T^{2t+1} \tag{13.3}$$

with relatively prime polynomials $\sigma, \omega \in \mathbb{F}[T]$ satisfying $\sigma(0) = 1$ and $\deg\omega \le \deg\sigma \le t$, since $R(\beta^\ell) = E(\beta^\ell)$ as $C(\beta^\ell) = 0$ for $1 \le \ell \le 2t$; see [7]. The latter can easily be calculated by the generalized Euclidean algorithm to determine the greatest common divisor of T^{2t+1} and $\sum_{\ell=1}^{2t} R(\beta^\ell)T^\ell$. Then the error is given by

$$M := \{i \in \{0, \ldots, n-1\} \,;\, \sigma(i) = 0\}$$
$$E(T) := \sum_{i\in M} E_i T^i \quad \text{with} \quad E_i := \frac{-\omega(\beta^{-i})\beta^i}{\sigma'(\beta^{-i})} \,. \tag{13.4}$$

For more details see [8] or [9, Section 3.5].

13.7 Interpolation-Based Decoding

The last section showed how to repair errors of weight up to half the minimum distance t. That method fails completely if a received word a contains more defective positions. A different approach was invented by Sudan. It will produce a list of codewords within a certain radius τ of a which can be larger than t. If τ is larger than t the list may contain more than one element. Sudan's idea is to transform the decoding problem into a curve-fitting problem; see [10]. It improves the error-correcting capabilities of Reed–Solomon codes of rate below 1/3. Later Guruswami and Sudan enhanced this algorithm; see [11].

This idea can be applied to the case of general algebraic–geometric codes. This was shown by Shokrollahi and Wasserman in [12]. As in the original case, this algorithm is capable of correcting errors beyond half the minimum distance if the information rate of the code is not too high. In all cases the running time of the algorithm increases enormously compared to the bounded maximum-likelihood decoder.

Keep the situation of Section 13.6 with a divisor $D := N \cdot P$ for $N \geq 2g - 1$. To explain the idea of Sudan in decoding the code $C := C_L(E, D)$, we consider the ring A of regular functions on $X - \{P\}$

$$A := \bigcup_{\lambda=0}^{\infty} L(X, \lambda \cdot P) .$$

The valuation $-\mathrm{ord}_P$ on A extends to a valuation on the polynomial ring $A[Z]$ by setting $w(Z) := N$. So one obtains the valuation

$$w : A[Z] \to \mathbb{N} \; ; \; w\left(\sum_{\mu} q_\mu Z^\mu\right) := \max_{\mu \, ; \, q_\mu \neq 0} \{-\mathrm{ord}_P(q_\mu) + \mu \cdot N\} .$$

For a natural number b we define the $\mathbb{F}$-vector space

$$A[Z](b) := \{Q \in A[Z] \; ; \; w(Q) \leq b\} .$$

Proposition 13.2 *Let $b \in \mathbb{N}$ be an integer with $b \geq N$. Then*

$$\dim_{\mathbb{F}} A[Z](b) \geq \frac{1}{2N}\left[(b - g + 1)^2 + N(b - N - g)\right] .$$

Proof A natural number λ is called a gap with respect to the point P if $L(X, \lambda \cdot P) = L(X, (\lambda - 1) \cdot P)$. There are g gaps and each gap λ satisfies $\lambda \leq 2g - 2$. Furthermore, for each natural number λ the number of $\mu \in \mathbb{N}$ with $\lambda + \mu \cdot N \leq b$ is given by $[b - \lambda/N]+$ 1. Then it follows that

$$\begin{aligned}
\dim_{\mathbb{F}} A[Z](b) &\geq \sum_{\lambda=0\,,\,\lambda \text{not gap}}^{b} \left(\left[\frac{b-\lambda}{N}\right] + 1\right) \\
&\geq \sum_{\lambda=0}^{b} \left(\left[\frac{b-\lambda}{N}\right] + 1\right) - \sum_{\lambda=0}^{g-1} \left(\left[\frac{b-\lambda}{N}\right] + 1\right) \\
&= \sum_{\lambda=g}^{b} \left(\left[\frac{b-\lambda}{N}\right] + 1\right) = \sum_{\lambda=0}^{b-g} \left(\left[\frac{\lambda}{N}\right] + 1\right) \\
&\geq \sum_{\lambda=0}^{b-g} \left(\frac{\lambda}{N} + \left(\left[\frac{\lambda}{N}\right] + 1 - \frac{\lambda}{N}\right)\right) \\
&\geq \frac{(b-g)(b-g+1)}{2N} + \left[\frac{b-g}{N}\right] \cdot \frac{N \cdot (N+1)}{2N} \\
&\geq \frac{1}{2N}\left[(b-g)^2 + (N+1)(b - N - g + 1) + b - g\right] \\
&= \frac{1}{2N}\left[(b - g + 1)^2 + N(b - N - g)\right]
\end{aligned}$$

□

Definition 13.1 *Let $Q \in A[Z]$ and let (x, a) be a pair of a rational points x of X and an element $a \in \mathbb{F}$. Then write*

$$Q(Z) = \sum_{\mu=0}^{\deg Q} q_\mu (Z - a)^\mu$$

with $q_\mu \in A$. The pair (x, a) is a zero of Q of multiplicity at least m if q_μ vanishes at x of order at least $m - \mu$ for all $\mu = 0, \ldots, m$.
For $a \in \mathbb{F}^n$ and $E := (P_1, \ldots, P_n)$ we say (E, a) is a zero of Q of multiplicity at least m if (P_i, a_i) is a zero of Q of multiplicity at least m for $i = 1, \ldots, n$.

Remark 13.1 *For $f \in A$ the polynomial $(Z - f) \in A[Z]$ has a zero of multiplicity at least 1 at $(x, f(x))$ for any rational point x of X. Any multiple of $(Z - f)$ has a zero of multiplicity at least 1 at $(x, f(x))$ as well.*

Proposition 13.3 *Let $a \in \mathbb{F}^n$. Let $U \subset A[Z]$ be a linear subspace of dimension at least $n \cdot \binom{m+1}{2} + 1$. Then there exists a nonzero $Q \in U$ such that Q vanishes at (E, a) with multiplicity at least m.*

Proof The condition of vanishing at one (P_i, a_i) is equivalent to $\binom{m+1}{2}$ linear constraints on the vector space U. □

Theorem 13.6 *Let $f \in L(X, N \cdot P)$ with $c = \mathrm{ev}_E(f)$. Let $m \geq 1$ be an integer. Let $b \in \mathbb{N}$ and consider a polynomial $Q \in A[Z](b)$. Let $a \in \mathbb{F}^n$ and assume that Q vanishes at (E, a) of multiplicity at least m. If $d(a, c) < n - b/m$, then $(Z - f)$ divides Q.*

Proof Consider the function $h := Q(f) \in A$ and note that its pole order at P is at most b. On the other hand, h vanishes to order at least m at P_i if $a_i = c_i$. The latter happens $n - d(a, c)$ times. Thus the degree of the zero divisor of h is greater than b. Since the number of poles and zeros of h must be equal unless $h = 0$, the claim follows. □

Corollary 13.1 *In the above situation, let $b \in \mathbb{N}$ be minimal such that*

$$\dim_{\mathbb{F}} A[Z](b) \geq n \cdot \binom{m+1}{2} + 1$$

Set

$$\tau_m := n - \frac{b}{m} .$$

Let $a \in \mathbb{F}^n$ be a received word. Then, there exists a nonzero polynomial $Q \in A[Z](b)$ which vanishes at (E, a) of multiplicity at least m. If $c \in C$ satisfies $d(a, c) < \tau_m$, then $c = \mathrm{ev}_E(f)$ for an $f \in L(X, N \cdot P)$ and $(Z - f)$ divides Q.

Since N and, hence, $b \geq N + g$ are large compared to g, the dimension of $A[Z](b)$ is bounded below by $\dim_{\mathbb{F}} A[Z](b) \geq (b-g)^2/2N$ in order to ensure that the condition of Corollary 13.1 is fulfilled; see Proposition 13.2. The ratio $N/n = (k-1+g)/n = k/n + (g-1)/n$ is close to the information rate $R := k/n$. Setting approximately $b = g + \sqrt{2nN\binom{m+1}{2}}$, we see that this method can correct any error with weight less than

$$\tau_m = n - \frac{b}{m} = n\left[1 - \frac{g}{mn} - \sqrt{\left(R + \frac{g-1}{n}\right)\left(\frac{m+1}{m}\right)}\right]$$

errors.

For small values of m, the estimate for τ_m is bad, explicit computations will show much better values; see Example 13.5. For large m we have approximately

$$\tau_\infty/n = 1 - \sqrt{R + \frac{g-1}{n}} \,.$$

Let us briefly discuss some numerical results in order to compare the different methods; the bounded minimum distance decoding t/n and the interpolation-based decoding τ_∞/n. Here we assume that n is large compared to the genus g so that the above approximations $t/n := (1-R)/2$ and $\tau_\infty/n = 1 - \sqrt{R}$ are well chosen.

R	0.1	0.2	0.3	0.4
t/n	0.45	0.40	0.35	0.30
τ_∞/n	0.684	0.553	0.452	0.367

To decode a received word $a \in \mathbb{F}^n$ one has to compute a nonzero interpolation polynomial $Q \in A[Z]$ such that (E, a) is a zero of Q with multiplicity at least m. This computation is easily done, since it simply means solving a system of linear equations. In a second step, one has to find all linear factors $(Z - f)$ of Q such that $f \in L(X, N \cdot P)$ and $d(a, \mathrm{ev}_E(f)) < \tau_m$. Due to Corollary 13.1 this gives a list of all codewords within radius less than τ_m of a. Clearly, the crucial calculation which can be rather long is the factorization of Q.

This can be done in the following way. Choose a geometric point $\mathcal{P}$ on X such that its residue field $k(\mathcal{P})$ has degree $[k(\mathcal{P}) : \mathbb{F}] > N$. So $\mathcal{P}$ is not rational and, hence, it lies on $X - \{P\}$. So it corresponds to a maximal ideal of A which we denote by $\mathcal{P}$ also. Then the mapping

$$L(X, N \cdot P) \to A/\mathcal{P} = k(\mathcal{P})$$

is injective, since the degree of the zero divisor of a function $f \in L(X, N \cdot P)$ mapped to 0 would be greater than or equal to $[k(\mathcal{P}) : \mathbb{F}]$ which exceeds N. Let $\overline{Q} \in k(\mathcal{P})[Z]$ be the image of Q. Now look at linear factors $Z - f$ of $\overline{Q}$ with $f \in L(X, N \cdot P)$ regarded as a subset of $k(\mathcal{P})$. In any case, the factorization can be lengthy if the field $k(\mathcal{P})$ is very large, although there are good factorization algorithms over finite fields; see [5].

In the case of multiplicity $m = 1$, it is not necessary to perform such a large field extension $k(\mathcal{P})$. Namely, if one chooses a polynomial Q of minimal degree in Z

vanishing at (E, a) of multiplicity 1, there exists a pair (P_i, a_i) for some $i \in \{1, \ldots, n\}$ such that the derivative $\partial Q/\partial Z(P_i, a_i) \neq 0$ where $a_i = f(P_i)$ for some $f \in L(X, N \cdot P)$ with $d(a, \mathrm{ev}_E(f)) \leq \tau_1$. Namely, there exists an $f \in L(X, N \cdot P)$ with $d(a, \mathrm{ev}_E(f)) \leq \tau_1$ so that one can write $Q(Z) = (Z - f) \cdot R(Z)$. Since the degree of Q is minimal, $R(f) \neq 0$ and, hence $\partial Q/\partial Z(f) \neq 0$. Thus there exists an index i such that $\partial Q/\partial Z(P_i, a_i) \neq 0$. Now let $\mathcal{P}_i \subset A$ be the maximal ideal corresponding to P_i. Then the residue field fulfills $k(\mathcal{P}_i) = \mathbb{F}$. Now we apply Newton's method to lift the zero a_i to a zero $\bar{f} \in A/\mathcal{P}_i^{N+1}$ and one factors the polynomial $\overline{Q}$ over $A/\mathcal{P}_i^{N+1}$. Since the restriction $\varrho_i : L(X, N \cdot P) \rightarrow A/\mathcal{P}_i^{N+1}$ is injective, one will recover $\bar{f}$ as an element in the image of $f \in L(X, N \cdot P)$ under ϱ_i. For more details see [13].

If the numbers b, n, N are small, the estimates of Proposition 13.2 give too rough a value. In that case it is better to compute the dimension precisely. To conclude this section, we discuss several choices of parameter in the case of Reed–Solomon codes.

Example 13.5 *Consider again the Reed–Solomon code of Example 13.1. In particular, by choosing* $\mathbb{F} = \mathbb{F}_{64}$ *and* $D = 15 \cdot P_\infty$ *we obtain a* $[64, 16, 49]$*-code. Thus, it can correct up to* 24 *errors using a bounded minimum distance decoder. If we set* $m = 1$ *and use the algorithm above we are now able to correct up to* $\tau = 27$ *errors. Namely,* $b = 36$ *fulfills*

$$\dim_{\mathbb{F}} A[Z](36) = 37 + (37 - 15) + (37 - 30) = 66 \geq 64 \cdot \binom{2}{2} + 1 = 65$$

and, hence $\tau_1 - 1 = 64 - 36/1 - 1 = 27$.

The $[64, 32, 33]$ *Reed–Solomon code with* $D = 31 \cdot P_\infty$ *can correct up to* 16 *errors by maximum-likelihood decoding. By interpolation-based decoding for* $m = 1$ *we have for* $b = 48$

$$\dim A[Z](48) = 49 + (49 - 31) = 67 \geq 65$$

and, hence, $\tau_1 - 1 = 64 - 48 - 1 = 15$. *Thus, no gain over the classical decoding methods is achieved in this case. Increasing the parameter* m *to* $m = 3$ *will extend the error correction radius to* $\tau = 17$. *Namely, the degree is* $b = 139$ *as* $\dim A[Z](139) = 140 + 109 + 78 + 47 + 16 = 390$. *In this case* $\tau_3 = 64 - 139/3 > 17$. *Using Corollary 13.1 one shows that* $\tau \rightarrow 19$ *as* $m \rightarrow \infty$ *for this code. In other words, we cannot arbitrarily extend the error correction radius by means of this method.*

13.8
Power Decoding of Low Rate Reed–Solomon Codes

In this section we describe a simple method proposed in [14] to decode Reed–Solomon codes of rate $R < 1/3$ beyond half the minimum distance. We keep the notations of Example 13.4. The main idea can be deduced from the Fourier transform given by (13.2).

Consider some $f = (f_0, \dots, f_{n-1}) \in \mathbb{F}^n$ and let $f(T) = f_0 + \dots f_{n-1}T^{n-1}$ denote the associated polynomial. Moreover, denote by $F(T) := \mathcal{F}(\mathrm{ev}_B(f))$ the Fourier transform. The convolution theorem of the Fourier transform relates the componentwise multiplication to the polynomial multiplication modulo $(T^n - 1)$. Namely,

$$f^{\langle \nu \rangle}(T) := f_0^\nu + \dots + f_{n-1}^\nu T^{n-1} \text{corresponds to } \; F^\nu(T) \mod (T^n - 1)\,. \qquad (13.5)$$

Clearly, according to (13.2) we obtain $F^\nu(\beta^{-i}) = f_i^\nu$ for all $i = 0, \dots, n-1$. It is sufficient to consider powers $1 \leq \nu < n$. Now let $r = c + e \in \mathbb{F}^n$ be a received word consisting of a codeword $c \in C_L(B, (k-1) \cdot P_\infty)$ and an error $e \in \mathbb{F}^n$. Then we can consider the ν-th power of the i-th component

$$r_i^\nu = (c_i + e_i)^\nu = c_i^\nu + \tilde{e}_i(\nu)\,.$$

If $e_i = 0$ then it follows that $\tilde{e}_i(\nu) = 0$ for all $1 \leq \nu < n$. Moreover, we consider the words

$$r^{\langle \nu \rangle} := (r_0^\nu, \dots, r_{n-1}^\nu) := c^{\langle \nu \rangle} + \tilde{e}(\nu) \quad \text{where} \;\; c^{\langle \nu \rangle} := (c_0^\nu, \dots, c_{n-1}^\nu)\,.$$

In order to obtain $c^{\langle \nu \rangle}$ as a codeword of a Reed–Solomon code with minimum distance larger than 1 we have to restrict to all ν with $\nu \cdot (k-1) < n$. Due to (13.5) it follows that

$$c^{\langle \nu \rangle} \in C_L\left(B, (\nu \cdot (k-1) + 1) \cdot P_\infty\right)\,.$$

Clearly, the code $C_L\left(B, (\nu \cdot (k-1) + 1) \cdot P_\infty\right)$ is corrupted by the error $\tilde{e}(\nu)$. Each error $\tilde{e}(\nu)$ has the same error positions as the original error e, but most likely different error values. Therefore, more equations like (13.3) can be found to determine the error positions.

In [14] it is shown that the decoding performance of this power decoding is the same as for the Sudan [10] algorithm, which is computationally much more complex. However, there is a principal difference between the algorithm of Sudan and power decoding. When the Sudan algorithm gives a list of possible codewords of size larger than 1, power decoding yields a failure. Furthermore, the newly obtained equations may be linearly dependent. However, the probability for this is very small, hence the word error probabilities of the two algorithms are virtually identical.

13.9 Interpolation-Based Soft-Decision Decoding

Consider the situation of Section 13.7. In this section we want to work with a curve-fitting where the multiplicities at the interpolation points are not uniform. They will be adjusted by information known about the channel. Let us first explain the new curve-fitting problem. We start with a matrix

$$M = (m_{P,\alpha})_{P \in E, \alpha \in \mathbb{F}} \in M(n \times q, \mathbb{N})$$

where we took some numbering on $\mathbb{F}$ in order to arrange the numbers $m_{P,\alpha}$ as a matrix. For such a matrix we define the *cost of M* by

$$\gamma(M) := \sum_{P \in E} \sum_{\alpha \in \mathbb{F}} \frac{m_{P,\alpha}(m_{P,\alpha}+1)}{2} .$$

For $a = (a_P)_{P \in E} \in \mathbb{F}^n$ the *score of a* with respect to M is defined by

$$S_M(a) := \sum_{P \in E} \sum_{\alpha \in \mathbb{F}} m_{P,\alpha} \cdot \delta_{\alpha,a_P}$$

where $\delta_{\alpha,\beta}$ is the usual Kronecker delta. In a similar way to Theorem 13.6, one demonstrates the following.

Proposition 13.4 *Let $M = (m_{P,\alpha})_{P \in E, \alpha \in \mathbb{F}}$ as above. Let $Q \in A[Z](b)$ be a polynomial with multiplicity at least $m_{P,\alpha}$ at (P,α) for all $(P,\alpha) \in E \times \mathbb{F}$. Let $f \in L(X, N \cdot P)$ with $c := \mathrm{ev}_E(f)$. If $S_M(c) > b$, then $(Z-f)$ divides Q.*

Corollary 13.2 *Let $\gamma(M)$ be the cost of M. Let $b \in \mathbb{N}$ be the least integer such that*

$$\dim_{\mathbb{F}} A[Z](b) \geq \gamma(M) + 1$$

Then, for any $a \in \mathbb{F}^n$, there exists a nonzero polynomial $Q \in A[Z](b)$ vanishing at (P,α) with multiplicity at least $m_{P,\alpha}$ for all $(P,\alpha) \in E \times \mathbb{F}$.
Assume that $(b + g^2 - 3g - 2gN) > 0$. If $f \in L(X, N \cdot P)$ with $c := \mathrm{ev}_E(f)$ fulfills $S_M(c) \geq g + 1 + \sqrt{2N\gamma(M)}$, then the linear factor $(Z-f)$ divides Q.

Proof Vanishing at all (P,α) poses $\gamma(M)$ linear constraints on Q. So there exists a nonzero polynomial as asserted. Due to Proposition 13.2 we know $\dim_{\mathbb{F}} A[Z](b) > \gamma(M)$ if $(b-g)^2 \geq 2N\gamma(M)$, respectively $b \geq g + \sqrt{2N\gamma(M)}$. Since b is the least integer satisfying $\dim_{\mathbb{F}} A[Z](b) \geq \gamma(M) + 1$, we have $S_M(c) > b$. Thus the claim follows from Proposition 13.4. □

For the distribution of the multiplicities we make use of information we know about the channel. More precisely, we assume that the conditional probabilities $\mathrm{pr}(y|x)$ are known; see Section 13.3. So, for a received word $y = (y_1, \ldots, y_n) \in \mathbb{F}^n$ we try to concentrate the multiplicities $m_{P,\alpha}$ on words which were most likely sent. In view of Proposition 13.4 we have to maximize the expected value of the score with respect to a fixed cost γ. The expected value of the score under the condition that $y = (y_P)_{P \in E} \in \mathbb{F}^n$ is received, is given by

$$\mathbb{E}(S_M(X)|y) := \sum_{x \in \mathbb{F}^n} S_M(x) \cdot \mathrm{pr}(x|y) .$$

One can rewrite this in the following way

$$\begin{aligned}\mathbb{E}(S_M(X)|y) &= \sum_{x\in\mathbb{F}^n}\left(\sum_{P\in E}\sum_{\alpha\in\mathbb{F}} m_{P,\alpha}\cdot\delta_{\alpha,x_P}\right)\cdot \mathrm{pr}(x|y)\\ &= \sum_{P\in E}\sum_{\alpha\in\mathbb{F}} m_{P,\alpha}\sum_{x\in\mathbb{F}^n}\delta_{\alpha,x_P}\cdot \mathrm{pr}(x|y)\\ &= \sum_{P\in E}\sum_{\alpha\in\mathbb{F}} m_{P,\alpha}\pi_{P,\alpha}(y)\end{aligned}$$

where

$$\pi_{P,\alpha}(y) := \sum_{x\in\mathbb{F}^n}\delta_{\alpha,x_P}\cdot \mathrm{pr}(x|y)$$

is the probability for having sent a vector $x\in\mathbb{F}^n$ with component $x_P=\alpha$ under the condition that y is received. The matrix $\Pi(y) := (\pi_{P,\alpha}(y))$ is called the *reliability matrix of* y. Since the problem is discrete, it is hard to find the optimal multiplicity matrix M with $\gamma(M)$ equal to a fixed cost γ if we know the reliability matrix $\Pi(y)$. But, for certain values γ, there is an algorithm which produces the optimal M; see [15].

Fix a number s of interpolation points and a reliability matrix $\Pi = (\pi_{P,\alpha})$. One constructs a multiplicity matrix $M = (m_{P,\alpha})$ in the following way.

For $\ell = 0,\ldots,s$ set

$$\ell = 0: \quad \Pi(0) = (\pi_{P,\alpha}(0)) := \Pi \quad M(0) := (m_{P,\alpha}(0)) := 0\ .$$

Given $\Pi(\ell-1)$ and $M(\ell-1)$, find a position (P,α) maximal among the entries of $\Pi(\ell-1)$ and change only this position in the following way

$$\begin{aligned}\ell\ge 1:\quad \pi_{P,\alpha}(\ell) &:= \frac{\pi_{P,\alpha}(\ell-1)}{m_{P,\alpha}(\ell-1)+2} & m_{P,\alpha}(\ell) &:= m_{P,\alpha}(\ell-1)+1\\ \Pi(\ell) &:= (\pi_{P,\alpha}(\ell)) & M(\ell) &:= (m_{P,\alpha}(\ell))\ .\end{aligned}$$

Then $M(s)$ has maximal expected score $\mathbb{E}(S_M(X)|y)$ among all multiplicity matrices M with $\gamma(M) = \gamma(M(s))$. Namely, let $B_{P,\alpha}(\ell)$ be the rectangle with length ℓ and height $\pi_{P,\alpha}/\ell$. The total length and the area of the collection $B := (B_{P,\alpha}(\ell))$ is given by

$$\begin{aligned}\gamma(M) &= \sum_{P,\alpha}\frac{m_{P,\alpha}(m_{P,\alpha}+1)}{2} = \sum_{P,\alpha} m_{P,\alpha}\sum_{\ell=1}^{m_{P,\alpha}}\ell = \mathrm{length}(B)\\ E(S_M(X)|y) &= \sum_{P,\alpha} m_{P,\alpha}\pi_{P,\alpha} = \sum_{P,\alpha}\sum_{\ell=1}^{m_{P,\alpha}}\pi_{P,\alpha} = \mathrm{area}(B)\ .\end{aligned}$$

So to maximize the area of B by a given total length means to pick the s rectangles with the largest height. This is precisely what the algorithm does.

For the value $\gamma := \gamma(M)$, the multiplicity matrix M is optimal.

Finally, let us discuss the cost $\gamma(M)$. To produce a list of possible code words $c = \mathrm{ev}_E(f)$ one has to factorize a polynomial $Q(Z)$ of weighted degree b which is related to $\gamma(M)$; approximately $b = g + 1 + \sqrt{2N\gamma(M)}$. The difficulty of factoring Q can be measured by

$$\deg_Z(Q) \lessapprox \frac{b}{N} = \frac{g+1}{N} + \sqrt{\frac{2\gamma(M)}{N}} \ .$$

We see that complexity of the factorization task is growing proportionally to $\sqrt{\gamma(M)}$.

13.10 Soft-Decision Decoding with the Dorsch Algorithm

In order to illustrate the advantages of using the reliability information provided by the channel, we describe the Dorsch algorithm [16] proposed in 1979 by Dorsch. For simplicity, in the following we consider a code $C(n, k, d)$ over the binary field. The data is transmitted over the AWGN channel, see Section 13.3, and the decoder calculates the L-values according to (13.1). Without loss of generality we may assume that these L-values are in decreasing order, that is

$$|L_1| > |L_2| > \ldots > |L_n| \ .$$

Note that the probability that two L-values are identical is zero, therefore we can write > instead of ≥. The idea for the Dorsch algorithm is that in the k most reliable positions the probability for an error is smaller than in the remaining $n - k$ positions.

We recall that the encoding of a linear code is done by a mapping of k information symbols to n code symbols. The algorithm calculates the codeword $\hat{c}$ associated to the k most reliable received values $\operatorname{sgn}(L_i)$, $i = 1, \ldots, k$. We call a set of k positions a systematic set if they can be used as information positions, which means that these positions determine the codeword. However, the number of systematic sets is less than $\binom{n}{k}$ and, hence, arbitrary k positions may not uniquely determine a codeword. In this case we have to use the $k + l$ most reliable positions for some $l \geq 1$ such that a systematic set is included as a subset.

So the following mapping is calculated

$$(\operatorname{sgn}(L_1), \operatorname{sgn}(L_2), \ldots, \operatorname{sgn}(L_k)) \mapsto \hat{c} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n) \in C$$

where $\operatorname{sgn}(L_i) = \hat{c}_i$, $i = 1, \ldots, k$. In many cases the estimated codeword $\hat{c}$ will be the transmitted codeword. This is a decoding method with extremely low computational complexity. However, it has only good performance if the channel is quite good in the sense that it has a very small variance.

The Dorsch algorithm proposes to increase the decoding performance and also the decoding complexity by the following step. Rather than re-encoding the k most reliable positions only, we do another k re-encoding trials in which we successively change the sign of L_i for $i = 1, \ldots, k$. By this method we exclude all errors of weight 1 in the first k positions.

In order to check which of the solutions is the best we may use, for example, the squared Euclidean distance

$$d_\mathrm{E}(y, c) = \sum_{i=1}^{n} (y_i - c_i)^2 \; .$$

Note that we use the mapping $c_i = (-1)^{a_i}$ where $a_i \in \{0, 1\}$. Clearly the solution with the smallest Euclidean distance has the largest probability. For this case we have $k + 1$ re-encoding trials, which is still of small complexity.

If we use, in addition, another $\binom{k}{2}$ re-encodings where we change any possible two signs of the k most reliable received positions, again we increase the decoding complexity but also the decoding performance. A further extension to three or more positions is obvious. However, unfortunately the complexity grows exponentially. In [17] it is shown that, increasing the number of positions changed for re-encoding, the decoding approaches the performance of ML decoding, which is the best possible. Infact, for two positions, the decoding is already very good.

References

1 Shannon, C.E. (**1948**) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423 and 623–656.

2 Goppa, V.D. (**1975**) A new class of linear error-correcting codes. *Info. and Control*, **29**, 385–387.

3 Grauert, H. and Remmert, R. (**1971**) *Analytische Stellenalgebren*, Springer.

4 Greuel, G.-M. and Pfister, G. (**2002**) *A Singular Introduction to Commutative Algebra*, Springer, Berlin.

5 Cohen, A. (**2000**) *A Course in Computational Algebraic Number Theory*, GTM 138, Springer.

6 Feng, G.L. and Rao, T.R.N. (**1993**) *Decoding algebraic–geometric codes up to the designed minimum distance*, IEEE Transactions on Information Theory, vol. IT-39, 37–45.

7 Gorenstein, D.C. and Zierler, N. (**1961**) A class of error-correcting codes in p^m symbols. *Journal of the Society of Industrial and Applied Mathematics*, **9**, 207–214.

8 Forney, G.D. (**1965**) *On decoding BCH codes*, IEEE Transactions on Information Theory, vol. IT-11, 549–557.

9 Lütkebohmert, W. (**2003**) *Codierungstheorie*, Vieweg-Verlag.

10 Sudan, M. (**1997**) Decoding of Reed–Solomon codes beyond the error-correction bound. *Journal of Complexity*, **13**, 180–193.

11 Guruswami, V. and Sudan, M. (**1998**) *Improved decoding of Reed–Solomon and algebraic–geometric codes*, Foundations of Computer Science, Proceedings, 39th Annual Symposium, 28–37.

12 Shokrollahi, M.A. and Wasserman, H. (**1999**) *List decoding of algebraic–geometric codes*, IEEE Transactions on Information Theory, vol. IT-45, 432–437.

13 Augot, D. and Pecquet, L. (**2000**) *A Hensel lifting to replace factorization in list-decoding of algebraic–geometric and Reed–Solomon codes*, IEEE Transactions on Information Theory, vol. IT-46, 2605–2614.

14 Bossert, M., Schmidt, G. and Sidorenko, V.R. (**2006**) *Decoding Reed–Solomon codes beyond half the minimum distance based on shift-register synthesis*, Proceedings IEEE International Symposium on Information Theory, July 2006, Seattle, WA, USA, 459–463.

15 Kötter, R. and Vardy, A. (**2003**) *Algebraic soft-decision decoding of Reed–Solomom codes*, IEEE Transactions on Information Theory, vol. IT-49, 809–825.

16 Dorsch, B.G. (**1974**) *A decoding algorithm for binary block codes and J-ary output channels*, IEEE Transactions on Information Theory, May 1974, vol. IT-20, 391–394.

17 Fossorier, M.P.C. and Lin, S. (**1996**) *Computationally efficient soft-decision decoding of linear block codes based on ordered statistics*, IEEE Transactions on Information Theory, May 1996, vol. IT-42, 738–751.

14
Investigation of Input–Output Gain in Dynamical Systems for Neural Information Processing

Stefano Cardanobile, Michael Cohen, Silvia Corchs, Delio Mugnolo, Heiko Neumann[1]

14.1
Overview

The processing of sensory signals in the human cortex is currently the subject of numerous studies, both at an experimental and a theoretical level. These studies investigate the principles of interactions between pairs of cortical areas. Different theoretical models have been derived from the experimental results and then proposed, in order to describe the processing of stimuli in V1, the primary visual cortex.

Several models assume layers of discrete sets of neurons arranged in a regular grid structure to be coupled via feedforward and feedback connections. These local connection structures can be considered as local filters of parametrized spread, which implement local center-surround interactions as well as modulatory outer-surround ones. In such representations of two-layer architectures, at each location, pairs of neurons define a local recurrent system of two neurons that is driven by an external input to one of those neurons. In this contribution we provide an elementary mathematical investigation of the stability of a core model of feedforward processing that is modulated by feedback signals. The model essentially consists of a system of coupled nonlinear first-order differential equations. In this paper we will address the issues of the existence, asymptotics, and dependence on parameters of the corresponding solutions.

14.2
Introduction

The cerebral cortex is organized into different areas, each of them consisting of several layers of neurons that are connected via feedforward, feedback, and lateral connections. The human cortical architecure is organized in a hierarchical struc-

[1] Corresponding author.

ture of mutually interacting stages and the majority of cortical areas are connected bidirectionally, see [6], whereas *lateral connections* enable the interaction of neurons in the same layer.

Several empirical studies have investigated the principles of cortical interactions between pairs of bidirectionally coupled cortical areas, particularly at early stages in the processing hierarchy. For example, it has been shown that the analysis of input stimuli in the primate visual cortex is mainly driven by propagation and filtering along feedforward neural pathways. The output of such filtering processes are modulated by activations in a neighborhood in the spatial or feature domain (via lateral connections) as well by activations at stages higher up in the hierarchy (via fast-conducting feedback connections, see [8]).

Several authors have investigated the influence of higher cortical stages on responses in the primary visual sensory area of the cortex (commonly denoted by V1), and in particular, the shaping of feature selectivity of neurons at the earlier stage. Ample experimental evidence exists, [4, 9, 14, 15], which supports the view that the interaction between neural signals along different directions of signal flow can be characterized by basic properties such as driving feedforward processing and modulation of activation patterns by higher-order activities via top-down feedback processing. Several computational neural models that draw upon these empirical findings have been proposed. The feedforward and feedback processing along mutually excitatory connections raises the question of stability in such neural networks.

Most of the proposed developments have been derived on the basis of numerical and experimental investigations. In the present paper we consider the model proposed by Neumann and coworkers [1, 13, 17]. It features a modulatory feedback mechanism for feature enhancement along various dimensions of input signal features, such as oriented contrast, texture boundaries, and motion direction. These gain-enhancement mechanisms are combined with subsequent competitive mechanisms for feature enhancement. The architecture of this model is well-behaved, i.e. it complies with basic assumptions of the theory and, in particular, the neural activities appear stable for large ranges of parameter settings.

At a mathematical level, this model essentially consists of a coupled system of nonlinear first-order Ordinary Differential Equations (ODEs). No formal mathematical analysis of this neural model has yet been performed. The aim here is to discuss the existence and uniqueness of solutions to the systems, as well as to present an elementary stability analysis dependent on the settings.

The mechanisms of the considered model can be summarized as defining a cascade consisting of three computational stages, see Figure 14.1.

1. An initial (linear) feedforward filtering stage along any feature dimension.

2. A modulatory feedback activation from higher cortical stages (via top-down connections) or lateral intra-cortical activation (via long-range connections).

3. A final competitive center-surround interaction at the output stage.

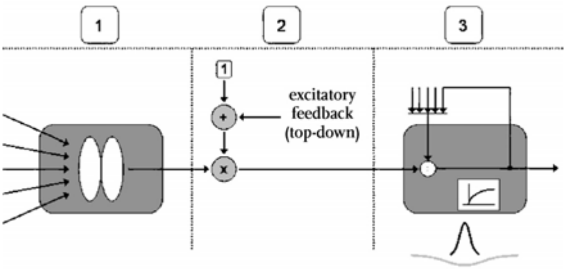


Figure 14.1 Scheme of the model circuit composed of three sequential stages. For a mathematical description of the computational stages, see text.

The different steps can be adequately modeled by a coupled system of ordinary differential equations. The core equations can be summarized as follows:

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = -c(t) + B\{s * \Gamma\}(t) \tag{14.1}$$

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -A\,x(t) + Bc(t)\left(1 + Cz^{\mathrm{FB}}(t)\right) \tag{14.2}$$

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = -A\,y(t) + Bf(x(t)) - (D + Ey(t))\{f(x) * \Lambda\}(t)\ . \tag{14.3}$$

The evolution of the input filtering is described by (14.1): s denotes the feedforward driving input, c is the filtered input or simply the input and Γ is a filter which is properly selected for the current processing scheme and goals (here $*$ denotes the convolution operation). The first negative term on the r.h.s of all three equations represents a decay term.

The second stage, modeled in (14.2), computes the activity x of a model neuron in terms of its potential. Such activity is assumed to be driven by the input c plus a feedback input. The feedback signal z^{FB} is a function of the output stage y, namely $z^{\mathrm{FB}} = g(y)$. The core nonlinearity for selective modulatory enhancement of inputs c is generated by the gating mechanism $(1 + Cz^{\mathrm{FB}}(t))$, i.e. the feedback z^{FB} will be controlled by the input c in a multiplicative form. If no input is present, available feedback should not generate new activation, whereas if available input is not supported by any feedback, the feedforward activity should not be suppressed. In case both input as well as feedback activation are available, this always leads to an enhancement of activation at time t by an amount $c(t) \cdot C \cdot z^{\mathrm{FB}}(t)$ which is proportional to the product between feedforward and feedback activity.

The third stage describes, in (14.3), the evolution of the activity y of a second interacting neuron. A feedforward excitatory term $f(x)$ and an inhibitory term $(D + Ey(t))\{f(x) * \Lambda\}(t)$ are present. The inhibitory term accounts for the lateral connections from the first layer of neurons (the lower neuron in Figure 14.2) through the convolution of $f(x)$ with the inhibitory kernel Λ. The term $Ey(t)$ normalizes the x-activities in the pool of cells covered by the inhibitory kernel.

The parameters A, B, C, D and E are constants. The simplest choice for the function f in (14.3) is the identity function: $f(x) = x$. Another possible choice that results in a faster increase of the activity is $f(x) = x^2$. The function g in (14.2) accounts for net computational effects of higher stages taking y activities as input and transforming them nonlinearly, before feeding the result into the feedback signal pathway. A common choice for g is a sigmoid function.

For stationary input luminance distributions, $\mathrm{d}c(t)/\mathrm{d}t = 0$, and the resulting equilibrium term for c can be treated as a constant c_0 for the remaining two equations to describe activities x and y, respectively. Throughout we will impose this (quite restrictive) assumption, which dramatically simplifies the analysis of the system's time evolution.

In the case of stationary input, c is commonly referred to as *tonic input*: the input is switched on and is kept constant during the evolution process of the dynamical system. Under this assumption, (14.2) simplifies to read

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -A\,x(t) + Bc_0\left(1 + Cz^{\mathrm{FB}}(t)\right) . \tag{14.4}$$

Under certain circumstances (as, e.g. in the case of spatially constant input), the inhibitory convolution term $\{f(x) * \Lambda\}(t)$ of (14.3) can be simply replaced by a point evaluation, i.e. by $f(x)(t)$, thus neglecting lateral connections.[2] In the absence of lateral connections, we are thus led to consider

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = -A\,y(t) + (B - D)f(x(t)) - Ey(t)f(x(t)) \tag{14.5}$$

instead of (14.3). Such a modified setting is clearly easier to investigate than a system with spatio-temporally varying external input. In order to fix the ideas, we will begin by studying this simplified model in Section 14.3. This will allow us to discuss elementary properties of the system using simple linear algebraical tools for the qualitative analysis of ODEs.

We can consider our model as a two-dimensional dynamical system with unknowns $x = x(t)$ and $y = y(t)$. A symbolic representation of this two-dimensional system given by (14.4)–(14.5) is shown in Figure 14.2. This basic unit, namely the recurrent connected pair of neurons x, y will be called *dipole* in the following. The dynamics of this two-dimensional system, which represents our basic unit, will be studied in Section 14.3.

It is, in fact, more realistic to consider (14.3), thus modeling (via the inhibitory kernel Λ) the lateral connections among neurons in the same hierarchical layer. A possible choice for this kernel is a Gaussian function. This can be interpreted as if the dipoles were mutually interconnected in a *ring* structure, as sketched in Figure 14.3. The mathematical properties of this ring architecture will be briefly discussed in Section 14.5.

[2] Please note that $f(x(t)) \equiv f(x)(t)$ holds, since we have assumed a spatially constant input distribution within the extent of a spatial kernel Λ.

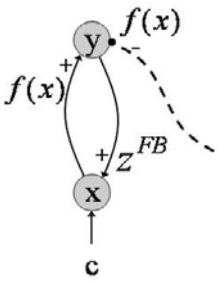


Figure 14.2 The basic unit of our model: the dipole. Two neurons are coupled via feedforward and feedback connections. The solid line indicates excitatory connections, the dashed line indicates inhibitory connections. See text for details.

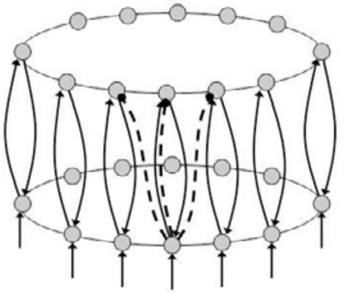


Figure 14.3 The basic units are coupled recurrently to form a ring structure. See Figure 14.2 and text where the basic unit is isolated and explained.

14.3 The Basic Unit: Analytical Study of the Dipole

The model we consider consists of interconnected rings. Each ring arises as the coupling of smaller basic units, namely of recurrently connected pairs of neurons, which we call a *dipole*. The aim of this section is to investigate the behavior of a single dipole. In particular, we neglect the lateral connections and discuss the problem given by (14.4)–(14.5). Observe that in our model neurons are schematized as point-like, lumped structures: there is no delay in their interactions and, what is more important, their spatial structure is neglected. Although not quite realistic, this assumption greatly simplifies the mathematical description of the system; see [2] for a discussion of the relations between lumped and nonlumped models. For the case of stationary input, the initial value problem associated with the two-dimensional system introduced in (14.4)–(14.5) can be written in the more general form as

$$\left\{\begin{array}{lcl} \dot{x}(t) & = & -\alpha x(t) + \beta\big(1 + \gamma g(y(t))\big) \\ \dot{y}(t) & = & -\eta y(t) + \big(\delta - \varepsilon h(y(t))\big) f(x(t) \\ x(0) & = & x_0 \in \mathbb{R} \\ y(0) & = & y_0 \in \mathbb{R}\,, \end{array}\right. \tag{14.6}$$

where $\dot{x}(t)$ and $\dot{y}(t)$ denote the time derivative of x, y. Here β depends on the input c_0, which we have assumed to be stationary. Moreover, the *activation parameters*

$\alpha, \gamma, \delta, \varepsilon, \eta$ are constants and f, g, h are real functions. In particular, g and f describe a feedback and feedforward activation, respectively. Thus, the initial value problem associated with (14.4)–(14.5) is actually only a special case of (14.6), after appropriate choice of parameters. We have preferred to perform qualitative analysis of this more general problem, instead on focusing of the special case introduced in Section 14.2.

In particular, the term $Bc(t)$ in (14.2) corresponds to β in the above problem. This is justified by the standing assumption that the input luminance distribution is stationary, and hence that $c(t) = c_0$, a constant. Therefore, $\beta = Bc_0$, and we can consider it as a *parameter* (which we are free to choose) depending on the (constant) input that we are feeding the system. We emphasize that our results do not hold in the general case when $\beta = \beta(t)$, i.e. if $c = c(t)$.

The above system of coupled first-order ODEs can be equivalently represented as a single Cauchy problem

$$\begin{cases} \dot{u}(t) &= F(u(t)) \\ u(0) &= u_0 \in \mathbb{R}^2 , \end{cases} \tag{CP}$$

to which we can apply standard mathematical results. Here $u := (x, y)$ and $u_0 := (x_0, y_0)$, while F is the nonlinear function on $\mathbb{R}^2$ defined by

$$F(x, y) := \begin{pmatrix} -\alpha x + \beta(1 + \gamma g(y)) \\ -\eta y + (\delta - \varepsilon h(y))f(x) \end{pmatrix} .$$

14.3.1 Well-Posedness Results

To begin with, we observe that the problem formulated above admits a solution, i.e. a pair $(x(t), y(t))$ satisfying (14.6).

Lemma 14.1 (The following assertions hold.)

(1) Let $f, g, h : \mathbb{R} \to \mathbb{R}$ be continuous. Then for all $x_0, y_0 \in \mathbb{R}$ there exists at least one solution of (14.6)*, locally in time.*

(2) Let moreover $f, g, h : \mathbb{R} \to \mathbb{R}$ be locally Lipschitz continuous. If $f(0) = h(0) = 0$, then there exists a unique solution of (14.6)*, locally in time.*

(3) Let finally $f, g, h : \mathbb{R} \to \mathbb{R}$ be globally Lipschitz continuous. If f or h is bounded, then the unique solution of (14.6) *is defined for all $t \in \mathbb{R}$.*

Proof

(1) The function F is clearly continuous. Thus, the assertion follows from Peano's existence theorem.

(2) Let $R > 0$ and $x, \tilde{x}, y, \tilde{y} \in B(0, R) := \{z \in \mathbb{R} : \|z\| \leq R\}$. Then one has

$$\begin{aligned}\left\|F(x,y) - F(\tilde{x},\tilde{y})\right\| &\leq |\alpha|\|x-\tilde{x}\| + |\eta|\|y-\tilde{y}\| \\ &\quad + |\gamma|\|g(y) - g(\tilde{y})\| + |\delta|\|f(x) - f(\tilde{x})\| \\ &\quad + |\varepsilon|\|h(y)f(x) - h(\tilde{y})f(\tilde{x})\| \\ &\leq (|\alpha| + |\delta|L_f)\|x-\tilde{x}\| + (|\eta| + L_g)\|y-\tilde{y}\| \\ &\quad + |\varepsilon|\|h(y) - h(0)\|L_f\|x-\tilde{x}\| \\ &\quad + |\varepsilon|L_h\|y-\tilde{y}\|\|f(\tilde{x}) - f(0)\| \\ &\leq (|\alpha| + |\delta|L_f)\|x-\tilde{x}\| + (|\eta| + L_g)\|y-\tilde{y}\| \\ &\quad + |\varepsilon|L_fL_hR\|x-\tilde{x}\| + |\varepsilon|RL_fL_h\|y-\tilde{y}\| ,\end{aligned}$$

where L_f, L_g, L_h denote the Lipschitz constants of f, g, h, respectively. This shows that F is locally Lipschitz continuous, and the uniqueness of a local solution is as a consequence of the Picard–Lindelöf theorem.

(3) It suffices to observe that F is globally Lipschitz continuous under the above assumptions. □

Having proved the well-posedness of the system, we now investigate its stability properties. Along the nullclines of the system we have

$$\alpha x(t) = \beta + \beta\gamma g(y(t)) \qquad \text{and} \qquad \eta y(t) = \delta f(x(t)) - \varepsilon h(y(t))f(x(t)) ,$$

respectively.

As already mentioned in Section 14.2, the activation functions f and h are commonly assumed to satisfy $f(0) = h(0) = 0$. Accordingly, we deduce that $(\bar{x}, \bar{y}) \equiv 0$ is an equilibrium point provided that $\beta(1 + \gamma g(0)) = 0$ (the converse is true if $\alpha \neq 0$). As already mentioned, in many models g is a sigmoid function and, in particular, $g(0) = 0$. Therefore, we have just observed that in absence of input (i.e. if $\beta = 0$) the only stationary state is the inactive state (i.e. $x(t) = y(t) \equiv 0$). The aim of the following section is to prove the existence of stationary solutions to (14.6) also in the case of nontrivial input.

14.3.2 Linearization

We show that equilibrium points also appear in the case of $\beta \neq 0$, corresponding to constant but nonvanishing inputs. For the sake of simplicity, in the remainder of this note we impose the realistic assumption that

$$f(0) = g(0) = h(0) = 0 . \tag{14.7}$$

We consider (14.6) as a dynamical system dependent on the parameters $\alpha, \beta, \gamma, \varepsilon, \delta, \eta$. We are going to show that the choice of the parameters $\gamma, \delta, \varepsilon$ is irrelevant for the existence of a stationary state, whereas specific conditions have to be imposed on parameters α and η. We then find a curve of stationary states in a neighborhood of the origin. This allows us to discuss some basic bifurcation properties of our system.

Theorem 14.1 *Let $\gamma, \delta, \varepsilon$ be fixed real numbers and f, g, h be given continuously differentiable activation function satisfying* (14.7)*. Then the following assertions hold.*

1. *For all parameters α_0, η_0 such that $\alpha_0\eta_0 \neq 0$ there exists a neighborhood U of $(\alpha_0, 0, \eta_0)$ and a continuous differentiable function $\kappa = (\kappa_1, \kappa_2)$ such that $(x, y) = (\kappa_1(\alpha, \beta, \eta), \kappa_2(\alpha, \beta, \eta))$ is a stationary state for all $(\alpha, \beta, \eta) \in U$.*
2. *The system has a bifurcation in $(\overline{x}, \overline{y}) = (0, 0)$ for $\beta = 0$ and all parameters α_0, η_0 such that $\alpha_0\eta_0 = 0$.*

Proof

(1) Fix $\gamma, \delta, \varepsilon \in \mathbb{R}$. Define a function $\Phi : \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}^2$ by

$$\Phi\begin{pmatrix}(x, y)\\(\alpha, \beta, \eta)\end{pmatrix} := \begin{pmatrix}-\alpha x + \beta + \beta\gamma g(y)\\-\eta y + \delta f(x) - \varepsilon h(y)f(x)\end{pmatrix} .$$

In other words, for fixed parameters $\gamma, \delta, \varepsilon$ and for given α, β, η one has

$$\Phi\begin{pmatrix}(\cdot, \cdot)\\(\alpha, \beta, \eta)\end{pmatrix} = F\big((\cdot, \cdot)\big) .$$

Then Φ is a continuously differentiable function, and its partial differential with respect to the first two variables is given by

$$D_{xy}\Phi\begin{pmatrix}(x, y)\\(\alpha, \beta, \eta)\end{pmatrix} = \begin{pmatrix}-\alpha & \beta\gamma g'(y)\\(\delta - \varepsilon h(y))f'(x) & -\varepsilon h'(y)f(x) - \eta\end{pmatrix} \tag{14.8}$$

For $\beta = 0$ and arbitrary coefficients α_0, η_0 the vector $(x, y) = (0, 0)$ is an equilibrium point of the system (14.6) since

$$\Phi\begin{pmatrix}(0, 0)\\(\alpha_0, 0, \eta_0)\end{pmatrix} = 0 .$$

In order to apply the implicit function theorem, compute the determinant of $D_{xy}\Phi$ in the point $((0, 0), (\alpha_0, 0, \eta_0))$. This is given by

$$det D_{xy}\Phi\begin{pmatrix}(0, 0)\\(\alpha_0, 0, \eta_0)\end{pmatrix} = \det\begin{pmatrix}-\alpha_0 & 0\\\delta f'(0) & -\eta_0\end{pmatrix} = \alpha_0\eta_0 , \tag{14.9}$$

because of the assumptions (14.7). By the implicit function theorem, there exist neighborhoods U of $(\alpha_0, 0, \eta_0) \in \mathbb{R}^3$ and V of $(0, 0) \in \mathbb{R}^2$ and a continuously differentiable function $\kappa : U \to V$ such that

- $\Phi(\kappa(\alpha_0, 0, \eta_0), \alpha_0, 0, \eta_0) = 0$ and
- for all $(x, y, \alpha, \beta, \eta) \in V \times U$ the point (x, y) is a stationary state with respect to parameters α, β, η if and only if $(x, y) = \kappa(\alpha, \beta, \eta)$.

(2) We have observed that the trivial state $(0, 0)$ is a stationary state if and only if $\beta = 0$. If $\alpha_0\eta_0 = 0$, then $\det D_{xy}\Phi((0, 0), (\alpha_0, 0, \eta_0)) = 0$ and the implicit function theorem fails to apply. Thus, the system has a bifurcation. □

By the above theorem it is possible to investigate the stability of our system by linearizing around the stationary state $(\overline{x}, \overline{y}) = (0, 0)$. To this end, we assume throughout that $\alpha\eta \neq 0$. By continuity, the asymptotic behaviour of the infinitely many stationary points whose existence has been proved in Theorem 14.1.(1) is the same of that of the stationary point $(0, 0)$. We can therefore restrict ourselves to investigating stability issues of (CP) for the case $(\overline{x}, \overline{y}) = (0, 0)$ and $\beta = 0$, only: we thus obtain the linearized Cauchy problem

$$\begin{cases} \dot{v}(t) &= DF(0,0)v(t) \\ v(0) &= u_0 \in \mathbb{R}^2 . \end{cases} \tag{lCP}$$

The Fréchet derivative $DF(0, 0)$ of F at $(0, 0)$ has been computed in (14.9): denoting by $T(t)$ the exponential of the 2×2 matrix $DF(0, 0)$, i.e.,

$$T(t) = \begin{pmatrix} e^{-t\alpha} & 0 \\ \delta f'(0)e^{-t\eta} & e^{-t\eta} \end{pmatrix} ,$$

the solution to (lCP) can be written as

$$v(t) = T(t)(x_0, y_0) .$$

This formula and the basic results from linear algebra yield interesting information about the asymptotic behaviour of the solution to (lCP) and hence, by the theorem of Hartman–Großman, also about the solution to (CP). Since the parameters have been assumed to be real, the relevant asymptotic behaviors are only determined by the sign of α and η.

This leads to the following, which holds for all parameters (α, β, η) inside the neighborhood U introduced in Theorem 14.1.

Theorem 14.2 *The following assertions hold for the linearized system* (lCP) *and all initial values* $u_0 \in \mathbb{R}^2$, *and hence also for the original system* (CP) *and all initial values in a neighborhood of the origin.*

1. *If* $\alpha \geq 0$ *and* $\eta \geq 0$, *then the solution is stable in the sense of Lyapunov, i.e.*

$$\|v(t)\| \leq \|u_0\| \quad \textit{for all} \quad t \geq 0 .$$

2. *If* $\alpha > 0$ *and* $\eta > 0$, *then the solution is uniformly exponentially stable, i.e.*

$$\|v(t)\| \leq e^{-\max\{\alpha,\eta\}t}\|u_0\| \quad \textit{for all} \quad t \geq 0 .$$

3. *If* $\alpha\eta < 0$, *then the solution with initial data* $u_0 = \xi_{\text{unst}}$ *(resp.* $u_0 = \xi_{\text{stab}}$*) is unstable (resp., exponentially stable), where* ξ_{stab} *and* ξ_{unst} *are the eigenvectors of* $DF(0, 0)$ *associated with* $\max\{\alpha, \eta\}$ *and* $\min\{\alpha, \eta\}$*, respectively.*
4. *If* $\alpha < 0$ *and* $\eta < 0$, *then the solution is unstable.*

In view of Theorem 14.2.(3), observe that the eigenspaces associated with α and η are spanned by $(1, \delta f'(0)/(\eta - \alpha))$ and $(0, 1)$, respectively, if $\alpha \neq \eta$.

In the above result we have denoted by $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^2$, i.e. $\|(x, y)\| = \sqrt{x^2 + y^2}$ for all $x, y \in \mathbb{R}$. While it is true that all norms on $\mathbb{R}^2$ are equivalent, and hence the stability properties we have obtained are norm-independent, it may still be interesting to characterize dissipativity of the system with respect to more relevant norms: in particular, with respect to the $\|\cdot\|_1$ and $\|\cdot\|_\infty$ norms defined by

$$\|(x, y)\|_1 = |x| + |y| \qquad \text{and} \qquad \|(x, y)\|_\infty = \max\{|x|, |y|\}\ .$$

These norms have a more intuitive interpretation than $\|\cdot\|$: at a given time t, $\|(x(t), y(t))\|_1$ and $\|(x(t), y(t))\|_\infty$ give the total potential of the inhibitory–excitatory system and the higher of both inhibitory and excitatory potentials, respectively. By [12, Lemma 6.1] we obtain the following: If $|\delta f'(0)| \leq \min\{\alpha, \eta\}$, then the solution of (lCP) satisfies

$$\|v(t)\|_1 \leq \|u_0\|_1 \quad \text{as well as} \quad \|v(t)\|_\infty \leq \|u_0\|_\infty \quad \text{for all} \quad t \geq 0\ .$$

14.4 The Basic Unit: Numerical Analysis of the Dipole

In this section we present some numerical simulations for the two-dimensional system introduced in (14.6). The analyis is made for a particular choice of functions f, g and h and for particular values of parameters. In particular, the constant input has been chosen in such a way that $\beta = 1$. Further, the parameters have been set to $\alpha = \eta = 1$, $\gamma = 10$, and $\varepsilon = 1$. The analysis has been made for different values of the gating parameter δ. The behavior for the cases $\delta = \pm 5$ has been plotted in the figures. The initial conditions are $x(0) = y(0) = 0$. The functions have been chosen as: $f(x) = x$, $g(y) = 1/(1 + \exp(-y)) - 0.5$ and $h(y) = y$. Given these functions, the nullclines of the system can be expressed analytically. The x-nullcline is given by

$$y = -\ln\left(\frac{10}{x + 4} - 1\right)$$

while the y-nullcline has coordinates

$$y = \frac{\delta x}{1 - x}$$

In Figure 14.4a the activities x and y as a function of t are shown for $\delta = 5$. The neural activities increase and converge to a stable value for $t > 2$ ms. In Figure 14.4b the trajectory of the system from the initial condition to the final state is shown together with the nullclines of the system. In this case, the corresponding eigenvalues $\lambda_{1,2}$ of the matrix $D_{xy}\Phi$ given by (14.8) are both negative ($\lambda_1 = -0.98$ and $\lambda_2 = -6.9$), indicating that the solution is stable; in fact, it is a stable *node*. The same simulations

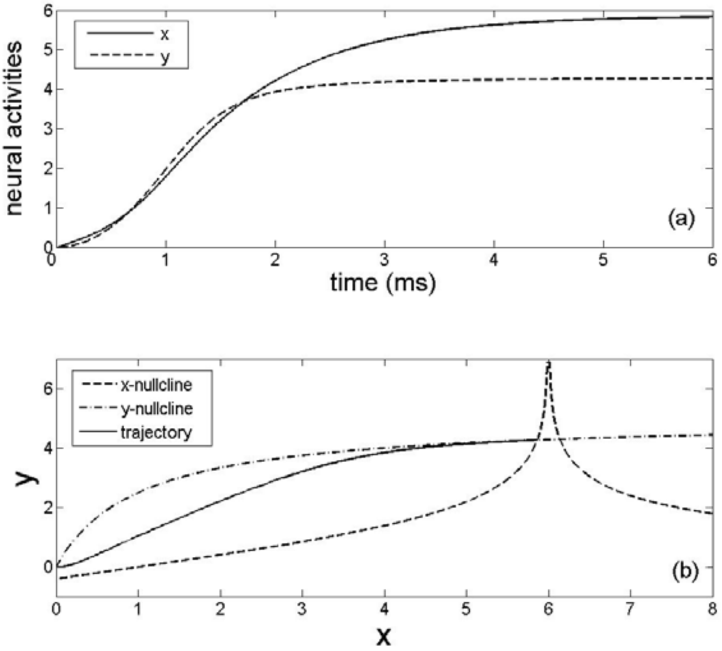


Figure 14.4 (a): Neural activities as a function of time. The parameters have been set as follows: $\alpha = \eta = 1, \gamma = 10, \varepsilon = 1$ and $\delta = 5$. (b) Nullclines and trajectory of the system.

have been carried out for $\delta = -5$ (see Figure 14.5). In this case the corresponding eigenvalues are complex conjugate numbers ($\lambda_1 = -1.4 + 1.5i$ and $\lambda_2 = -1.4 - 1.5i$) with negative real part. The activities present an oscillatory behavior for $t < 3$ ms and converge to a stable value for $t > 3$ ms. Therefore, the system evolves to a stable state and the trajectory corresponds to a stable *focus* as can be seen in Figure 14.5b.

14.5 Model of a Recurrent Network

In this section we briefly discuss the case of a recurrent neural network. We thus allow for lateral connections among neurons of the same hierarchical level, see Figure 14.3. We avoid technicalities and refer the interested reader to a later paper for mathematical details.

In order to describe the behavior of a ring, we denote a given dipole in the ring by its angle $\theta \in [0, 2\pi)$ with respect to some fixed reference direction: we will therefore denote by $x_\theta(t)$ and $y_\theta(t)$ the potentials of the dipole with angular coordinate θ at time t.

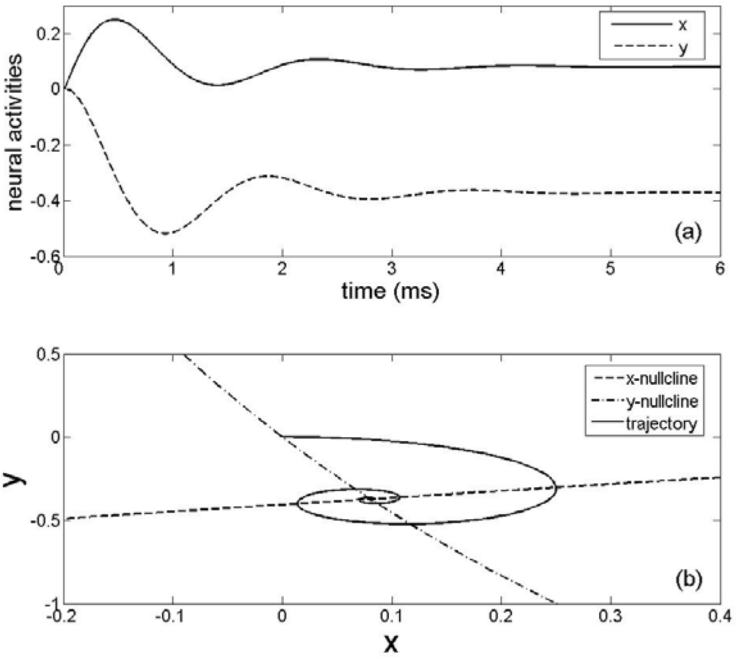


Figure 14.5 (a) Neural activities as a function of time. The parameters have been set as follows: $\alpha = \eta = 1, \gamma = 10, \varepsilon = 1$ and $\delta = -5$. (b) Nullclines and trajectory of the system.

As explained in Section 14.2, the correct mathematical setting is that of a system of integro-differential equations corresponding to (14.4 and 14.3). Formulating it in greater generality, such that the initial-value problem associated with this system reads as a Cauchy problem, we have

$$\left\{\begin{array}{rcl} \dot{x}(t) & = & -\alpha x(t) + \beta\big(1 + \gamma g(y(t))\big) \\ \dot{y}(t) & = & -\eta y(t) + \delta f(x(t)) \\ & & + (\varepsilon + \theta h(y(t))) \int_0^{2\pi} f(x(s)) \Lambda(t-s)\,\mathrm{d}s \\ x(0) & = & x_0 \\ y(0) & = & y_0\,, \end{array}\right. \tag{14.10}$$

where Λ is a suitable integral kernel. In this setting, we are considering the distribution of the neuronal activity on a whole ring. Thus, we have to consider a state space more involved than $\mathbb{R}^2$ (which we have used in Section 14.3), taking account of the angular distribution of the activity. In fact, there are several possibilities: we choose to take as the phase space the product space $L^2_{\text{per}} \times L^2_{\text{per}}$. Here L^2_{per} is the linear space of square integrable 2π-periodic real functions. Observe that the initial data are not numbers but functions representing the initial distribution of the neuronal activity, i.e. $x_0, y_0 \in L^2_{\text{per}}$.

As in Section 14.3 we rewrite (14.10) as the differential equation

$$\begin{cases} \dot{u}(t) &= F(u(t)) \\ u(0) &= u_0 \in \mathbb{R}^2 , \end{cases} \tag{DP2}$$

in the infinite-dimensional space $L^2_{\text{per}} \times L^2_{\text{per}}$, to which we can apply standard results on existence, uniqueness, and stability. As in Section 14.3, $u := (x, y)$ and $u_0 := (x_0, y_0) \in L^2_{\text{per}} \times L^2_{\text{per}}$, while F is now a nonlinear operator on $L^2_{\text{per}} \times L^2_{\text{per}}$ defined by

$$F(x, y) := \begin{pmatrix} -\alpha x + \beta(1 + \gamma g(y)) \\ -\eta y + \delta f(x) + (\varepsilon + \theta h(y))((f \circ x) * \Lambda) \end{pmatrix} .$$

We are still assuming that (14.7) holds. Accordingly, we again deduce that the vector $(0, 0)$ is an equilibrium point only in the case of $\beta = 0$, i.e. only in the case of zero input. In the case of nonzero input one can extend the trivial nullcline $(0, 0)$ following the same idea of Section 14.3.2. The somewhat technical proof of the following theorem is similar to that of Theorem 14.4 and it is based on the Banach space version of the implicit function theorem.

Theorem 14.3 *Let $\gamma, \delta, \varepsilon$ be fixed real numbers and f, g, h be given Fréchet differentiable operators on L^2_{per}. Let them satisfy* (14.7)*, where* 0 *now denotes the costant zero function. Then for all numbers α_0, η_0 such that $\alpha_0\eta_0 \neq 0$ there exists a neighborhood U of $(\alpha_0, 0, \eta_0)$ and a Fréchet differentiable function $\kappa = (\kappa_1, \kappa_2)$ such that $(x, y) = (\kappa_1(\alpha, \beta, \eta), \kappa_2(\alpha, \beta, \eta))$ is a stationary state for all $(\alpha, \beta, \eta) \in U$.*

We compute the Fréchet derivative of F at any vector $(r, s) \in L^2_{\text{per}} \times L^2_{\text{per}}$ and obtain

$$DF(r, s) = \begin{pmatrix} -\alpha & \beta\gamma g'(s) \\ \delta f'(r) + (\varepsilon + \theta h(s)) f'(r)((f \circ r) * \Lambda) & -\eta + \theta h'(s)((f \circ r) * \Lambda) \end{pmatrix} .$$

Letting $\beta = 0$, by (14.7), we thus obtain

$$DF(0, 0) = \begin{pmatrix} -\alpha & 0 \\ \delta f'(0) & -\eta \end{pmatrix} .$$

Thus, one can linearize around the origin and investigate asymptotics of the system in the same way as in Section 14.3. Performing bifurcation analysis in infinite-dimensional Banach spaces is technically more involved, and goes beyond the scope of this chapter. We refer, e.g., to [3] for details. We will address the complete mathematical analysis of the present case of a ring structure in a later paper (in preparation).

14.6
Discussion and Conclusions

In the present paper we have investigated stability issues for the neural network model presented by Neumann and coworkers in [1, 13, 17]. The basic unit of the model represents essentially two neurons coupled via feedforward and feedback connections. The model consists of three computational stages: (1) an input filtering stage; (2) a modulatory feedback stage; and (3) a final competititve center surround mechanism at the output stage. These stages can be modeled by a coupled system of three differential equations. In this chapter we addressed the case of input signals that were assumed temporally constant (i.e. they do not vary over time after onset). Furthermore, we have assumed a spatially homogeneous surround input to the center-surround competition stage. This fact reduces the general system to a two-dimensional system that can be represented as a single Cauchy problem. We have discussed the problem of existence of solutions and investigated their stability properties. In the trivial case of zero input, the only stationary state is the inactive state. The existence of solutions has also been proved for the case of nonzero input, where the conditions on the different parameters characterizing the equations have been derived. Two parameters play a crucial role for the stability of the system, namely α and η of (14.6). If these two parameters are positive, then the solution of the system (14.6) is uniformly stable. Let us recall that the ODE system (14.6) describes the neural activities x and y of the computational model given by (14.2 and 14.3). Within this framework, α and η represent the time constant of the decay terms and therefore the conditions $\alpha, \eta > 0$ are satisfied when a concrete neural model is being considered.

We stress that the mathematical analysis of the stability conditions here presented is valid in the case of nontrivial input but inside a certain neighborhood of zero input. This means that if the input increases beyond this neighborhood, the conclusions stated here may no longer be valid.

Numerical simulations for the neural activities and the trajectory of the system have been presented for two different sets of parameters (for nonzero input). For these particular choices, the neural activities converge to stationary solutions and the corresponding trajectories evolve to a stable node or stable focus, respectively.

Finally, the case where the basic units are recurrently coupled to form a ring structure has been briefly analyzed. This corresponds to the more realistic case where the neurons interact via lateral connections. The extension of the stability analysis to the ring structure has been sketched.

References

1 Bayerl, P. and Neumann, H. (**2004**) Disambiguating visual motion through contextual feedback modulation. *Neural Computation* **16**, 2041–2066.

2 Cardanobile, S., Markert, H., Mugnolo, D. and Palm, G. (**2008**) Relating simulation and modelling of neural networks, in *Evolution, Information and Complexity*, (eds W. Arendt and W. Schleich). Wiley, New York.

3 Da Prato, G. and Lunardi, A. (**1986**) Hopf bifurcation for fully nonlinear equations in Banach space. *Ann. Inst. Henri Poincaré* **3**, 315–329.

4 De Yoe, E. and Van Essen, D. (**1988**) Concurrent processing streams in monkey

visual cortex. *Trends in Neuroscience*, **11**, 219–226.

5 Ermentrout, G.B. and Cowan, J.D. (**1980**) Large scale spatially organized activity in neural nets. *SIAM J. Appl. Math.*, **38**, 1–21.

6 Felleman, D. and Van Essen, D. (**1991**) Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, **1**, 1–47.

7 Friedman, H.S., Zhou, H., von der Heydt, R. (**2003**) The coding of uniform colour figures in monkey visual cortex. *J. Physiol.*, **548**, 593–613.

8 Girard, P., Hupe, J. and Bullier, J. (**2001**) Feedforward and feedback connections between areas V1 and V2 of the monkey have similar rapid conduction velocities. *Journal of Neurophysiology*, **85**, 1328–1331.

9 Hupe, J., James, A., Payne, B., Lomer, S., Girard, P. and Bullier, J. (**1998**) Cortical feedback improves discriminatio between figure and background by V1,V2 and V3 neurons. *Nature*, **394**, 784–787.

10 Kärcher, J. (**2007**) *Continuous Attractor Networks Modeling Head Direction Cells.* Diploma Thesis, Universität Ulm.

11 Li, Z. (**1998**) A neural model of contour integration in the primary visual cortex. *Neural Computation*, **10**, 903–940.

12 Mugnolo, D. (**2007**) Gaussian estimates for a heat equation on a network. *Netw. Heter. Media*, **2**, 55–79.

13 Neumann, H. and Sepp, W. (**1999**) Recurrent V1-V2 interaction in early visual boundary processing. *Biological Cybernetics*, **81**, 425–444.

14 Salin, P. and Bullier, J. (**1995**) Corticocortical connections in the visual system: Structure and function. *Physiological Review*, **75**, 107–154.

15 Sandell, J. and Schiller, P. (**1982**) Effect of cooling area 18 on striate cortex cells in the squirrel monkey. *J. Neurophysiol.*, **48**, 38–48.

16 Seung, H. (**1996**) How the brain keeps the eyes still. *Proc. Nat. Acad. Sci. USA*, **93**, 13339–13344.

17 Thielscher, A. and Neumann, H. (**2003**) Neural mechanisms of cortico-cortical interaction in texture boundary detection: a modeling approach. *Neuroscience*, **122**, 921–939.

15
Wave Packet Dynamics and Factorization

Rüdiger Mack[1], Wolfgang P. Schleich, Daniel Haase, Helmut Maier

15.1
Introduction

Integer factorization is one of the major problems of algorithmic number theory. This topic is highly relevant for cryptography since public key cryptosystems [1] such as RSA draw their security from the supposed difficulty of this problem. Indeed, there is still no efficient classical algorithm to compute the factors of an integer.

The discovery of Shor's algorithm [2, 3] drew the attention of the cryptographic community to quantum computation [4, 5], which resulted in a boost in research on the subject. In the present paper we summarize the mathematical as well as the physical aspects of the Shor algorithm and draw analogies to familiar problems in scattering of atoms from phase gratings [6].

15.1.1
Central Ideas

Shor's method of factoring an integer $N = p \cdot q$ into its prime factors p and q consists of two parts: (i) a mathematical algorithm; and (ii) a quantum mechanical implementation.

The mathematical part contains [7,8] three essential ideas: (i) the construction of a periodic function f based on modular exponentiation; (ii) the use of the period of f to find integers which share a common factor with N; and (iii) the Euclidean algorithm to distill these common factors. For the understanding of the factorization algorithm rudimentary knowledge of elementary number theory [9] is sufficient. The method relies on the proper choice of the basis of the exponentiation. As a result, the algorithm does not always produce all factors. A deeper analysis taking advantage of elements of number theory such as modular arithmetic, the Chinese remainder theorem and the primitive root, shows that the approach is successful in half of the trials.

[1] Corresponding author.

The information about the factors p and q of N is contained in the period r of the function f based on modular exponentiation. However, in order to obtain r we need to know the values of f over one period. Since this period gets very large, we face a computationally extensive problem.

It is at this point that quantum mechanics [10] and, in particular, entanglement [11] of two quantum systems comes in useful. The states $|j\rangle_1$ of the first system encode the arguments j of the function f whereas the states $|f(j)\rangle_2$ of the second system yield the values $f(j)$ of the function f. An appropriate interaction between the two systems produces an entangled state which is a coherent superposition of the states $|j\rangle_1|f(j)\rangle_2$. A projective measurement on an arbitrary state $|f(k_0)\rangle_2$ of the second system creates a superposition of states $|k_0 + mr\rangle_1$ of the first system, where m assumes integer values. In this way we have mapped the periodicity of f onto the first quantum system. However, the value of k_0 is random. For this reason it is advantageous to consider this periodic state in the variable conjugate to j. In this representation the value of k_0 appears as a phase and the probability distribution displays clear maxima at multiples of the inverse period. As a result every measurement can only take on these values leading rapidly to r.

There is a close analogy between the Shor algorithm and the problem of the deflection of atoms from a standing light field in the Raman–Nath approximation [6]. Indeed, the distribution of scattered atoms in the far field consists of sharp peaks separated by the inverse of the period of the mode function. The derivative of the mode function plays the role of f. Needless to say, this scheme does not involve entanglement but only interference. In this sense, we could also realize it by scattering light off a phase grating.

The Shor algorithm takes advantage of the enormous size of Hilbert space. Indeed, when we consider two-level atoms, the dimensionality of Hilbert space grows exponentially with the number of atoms. It is this exponential growth, which leads to the exponential speedup of the Shor algorithm.

15.1.2
Outline of the Article

Ski slopes are usually categorized according to the experience of the skier. Experienced skiers can go straight to the most difficult routes usually marked by black diamonds. However, beginners should first familiarize themselves with the territory and practise on the easy slopes marked by green dots. Here, we follow the American notation. In this spirit our article contains two tracks summarized in Figure 15.1. The black run is the difficult one and assumes that the reader is familiar with quantum mechanics as well as number theory. This slope leads us straight to the very heart of the Shor algorithm.

We dedicate Section 15.2 to a brief summary of the mathematical aspects of the Shor algorithm. In Section 15.3 we then turn to the quantum mechanical implementation. Here we emphasize the mapping of the period of the modular exponentiation to the register with the help of entanglement and a subsequent readout in the conjugate variable using a Fourier transform. In Section 15.4

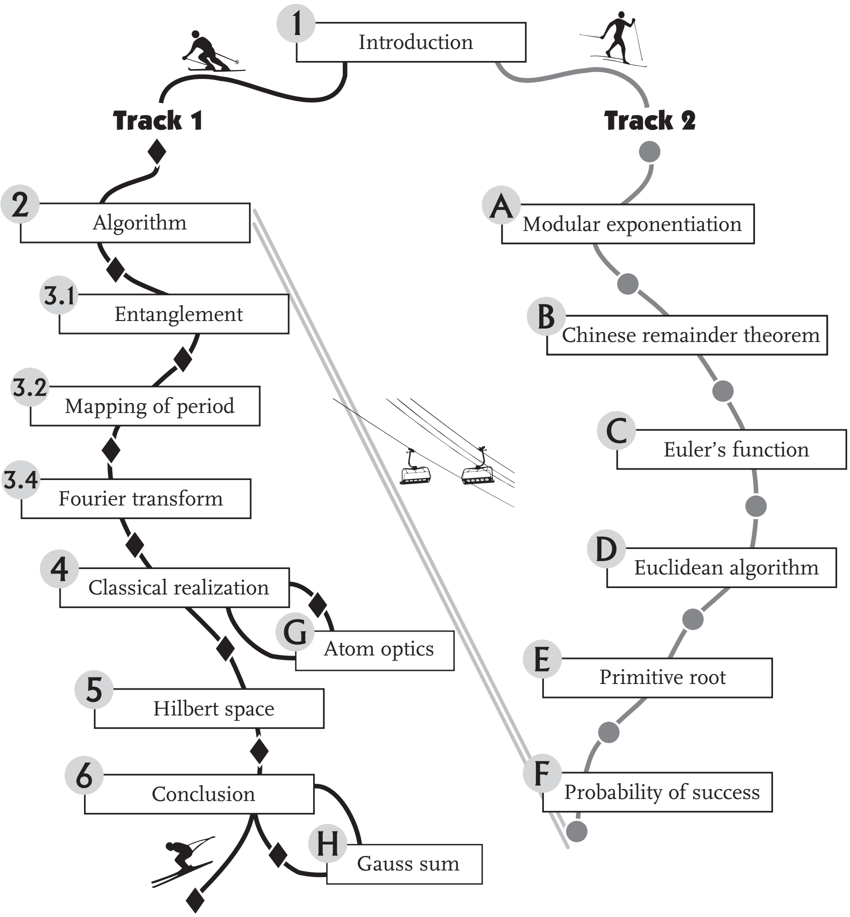


Figure 15.1 A guide through the problem of factorization with the Shor algorithm along two tracks highlighting the main themes of our article. The black diamond run (track 1) is recommended to a reader with a background in number theory. The run marked by the dot (track 2) prepares the reader for the diamond run and introduces the necessary concepts of number theory. Two appendices dedicated to specific problems in physics allow for small detours on the diamond run.

we elaborate on a classical realization based on the scattering of atoms from a standing light field. We show that the key step in the Shor algorithm, that is the projection onto the entangled state and the creation of the periodic structure in the first quantum system, is analogous to the projection of the quantum state $|\psi\rangle$ corresponding to the center-of-mass motion on the momentum eigenstate $|p\rangle$. This analogy stands out most clearly when we analyze the momentum probability amplitude $\psi(p) = \langle p|\psi\rangle$ with the method of stationary phase [12]. Our analysis of the Shor algorithm concludes in Section 15.5 with a brief discussion of the source of the exponential speedup. We summarize our results in Section 15.6.

A reader not acquainted with modular arithmetic is advised to first practise on the easy slopes, that is to follow the second track consisting of various appendices, before going onto the black route. Indeed, the appendices lay the foundations for the main body of the paper by studying various elements of number theory.

Since the Shor algorithm relies on finding the period of a function, we have selected several topics of number theory, which address this very point. We start in Appendix 15.A with an elementary introduction into modular arithmetic. Here we focus on the periodicity properties of modular exponentiation. Since these calculations quickly involve large numbers, it is useful to develop a tool, which allows us to work only with small numbers. In this context, the Chinese remainder theorem, discussed in Appendix 15.B, is a great help. For example, it yields the period of the modular exponentiation, when the module is the product of two coprime numbers. Moreover, the period of the modular exponentiation is a divisor of Euler's ϕ-function, which is defined by the number of coprime residues. As shown in Appendix 15.C we can calculate the values of ϕ with the help of the Chinese remainder theorem. At various stages of our analysis we need to determine the greatest common divisor of two positive integers. The Euclidean algorithm discussed in 15.D is an effective method of achieving this goal. Still, we have no analytic formula to calculate the period of the modular exponentiation. We have to find it recursively. This gap is filled by the concept of a primitive root introduced in Appendix 15.E. These elements of number theory are finally put to use in Appendix 15.F to estimate the probability of success for the Shor algorithm.

The last two appendices are of a completely different nature. They do not provide introductions into a field but rather contain detailed calculations of a physical problem. In Appendix 15.G we calculate the momentum distribution of atoms scattered off an electromagnetic field. In Appendix 15.H we show, that the constructive interference of the phase factors of Gauss sums [13] allows us to factor numbers [14, 15].

15.2 How to Factor Numbers

In this section we elaborate on the essential ideas of Shor's algorithm. After a brief introduction to the problem of factorization we outline the three steps of the mathematical algorithm: (i) modular exponentiation; (ii) factorization; and (iii) Euclidean algorithm. This analysis relies heavily on elements of modular arithmetic. A reader not familiar with this branch of number theory is well-advised to first work through Appendices 15.A–15.F before continuing here. Since at the stage of modular exponentiation one has to choose a number at random, the algorithm is probabilistic and does not work on every trial. We conclude by discussing these problems and present the probability of success.

15.2.1 Prime Numbers, a Primality Test and a Naive Approach to Factorization

A prime number p is a natural number that has exactly two factors 1 and p. Two integers m and n are called coprime if they have only 1 as a common factor. Any natural number N can be decomposed into a product of prime numbers which is unique up to ordering. It has been known from antiquity that there are infinitely many prime numbers. According to the prime number theorem of analytic number theory [13] the number $\pi(x)$ of prime numbers smaller than x is

$$\pi(x) \sim \frac{x}{\log(x)} \tag{15.1}$$

in the limit of large values of x.

It is not too difficult to check if a given number N is prime. For this purpose we consider the operation of modular exponentiation discussed in Appendix 15.A.3. For most composite numbers the compositeness may be shown by Fermat's little theorem [9]: If N is prime, then

$$a^{N-1} \equiv 1 \mod N \tag{15.2}$$

for all a coprime to N. For most composite N one can easily find a, such that (15.2) does not hold and thus prove the compositeness of N.

There are, however, exceptions, the so-called Carmichael numbers. Indeed, N is called a Carmichael number, if it is composite and if

$$a^{N-1} \equiv 1 \mod N \tag{15.3}$$

for each a coprime to N. It has been shown [16] that there are infinitely many Carmichael numbers. The smallest example is $N = 561 = 3 \cdot 11 \cdot 17$. One has $a^{560} \equiv 1 \mod 561$ for all a that are not divisible by 3, 11 or 17. The idea of considering Fermat's congruence may, however, be modified to decide for all numbers whether they are prime or composite. This method is known as Rabin's probabilistic primality test [17]. Instead of the single congruence (15.2) with the exponent $N-1$, one also considers congruences with exponents $(N-1)/2$ (and $(N-1)/4$, $(N-1)/8$, etc. if these quotients are integers).

The problem of finding the integer factors p and q of a number $N = p \cdot q$ is an even more complicated problem. For small numbers, such as 15, we can rely on our memory to find the prime factors 3 and 5. However, for large numbers this task is highly nontrivial. In principle one could try out all prime numbers from 2 to a given number if they divide N. In the worst case, when N consists only of two factors, we have to go up to $\sqrt{N}$. Since it suffices to try out prime numbers we have to perform $\sqrt{N}/\log\sqrt{N}$ divisions.

Computer science expresses the complexity of problems such as factorization in terms of the number k of digits of N. For example, the naive factorization approach based on division by all primes smaller than $\sqrt{N}$ scales exponentially with the number of digits of N. Moreover, the number of steps needed by all known classical algorithms to factor numbers cannot be bound by a polynomial in k. However, the Shor algorithm relying on quantum mechanics uses a number of steps polynomial in k, and is therefore efficient.

15.2.2
A More Sophisticated Algorithm

There exist more sophisticated schemes to find the factors of a number. The most prominent one is based on modular arithmetic. Here we calculate the powers of a given integer a modulo the number N that we want to factor, that is we evaluate

$$f(j) = a^j \bmod N\,. \tag{15.4}$$

Table 15.1 displays the function $f(j) = a^j \bmod N$ for the example of $a = 3$ and $N = 14$ and shows that $f(j)$ is periodic with the period $r = 6$. This periodicity property holds true for all pairs of numbers N and a. For a more detailed discussion we refer to the Appendices 15.A–15.F.

We can now use this periodicity to find the factors of N. In order to understand this claim, we note that if $r < N$ is the period we find

$$a^j \equiv a^{j+r} \mod N\,, \tag{15.5}$$

which is equivalent to

$$a^j(a^r - 1) \equiv 0 \mod N\,, \tag{15.6}$$

or

$$a^j(a^{r/2} - 1)(a^{r/2} + 1) = k \cdot N\,. \tag{15.7}$$

Equation (15.7) states that $a^{r/2} - 1$ or $a^{r/2} + 1$ share a common factor with N. As a consequence we have reduced the problem of factoring a number to the problem of finding factors common to numbers. The Euclidean algorithm, described in Section 15.D, is a very effective method to identify these common factors.

In summary, in order to factor a number, we have to first find the period of a function defined by modular exponentiation. This period leads us to two numbers which share a factor with the number to be factored. With the help of the Euclidean algorithm we determine these common factors.

15.2.3
Problems with this Algorithm and Probability of Success

For the example of Table 15.1 we have found the period $r = 6$ which for $a = 3$ yields $a^{r/2} - 1 = 3^3 - 1 = 26$ and $a^{r/2} + 1 = 3^3 + 1 = 28$. The common factors with $N = 14$ are 2 and 14. Thus we have found only one prime factor of $N = 14$, namely 2.

Table 15.1 For $N = 14$ and $a = 3$ the function $f(j) = a^j \bmod N$ displays a period of $r = 6$.

j	0	1	2	3	4	5	6	7	8	9
$f(j)$	1	3	9	13	11	5	1	3	9	13

This accident demonstrates that not every a is “useful”. But what are the criteria defining a useful value of a?

First of all a should not share factors with N. Otherwise we would already have found a factor and do not have to go to the trouble of searching for a period. To have such luck is very unlikely for a randomly chosen number a and we can always test if the a and N are coprime by running the Euclidean algorithm.

Moreover, a has to be chosen such that the period r of $f(j)$ defined in (15.4) is even. If r is odd, we cannot make the decomposition $a^r - 1 = (a^{r/2} - 1)(a^{r/2} + 1)$.

Another misfortune occurs, if $a^{r/2} + 1$ is a multiple of N. Then, $a^{r/2} - 1$ does not share a factor with N and we have not gained anything. As a consequence we have to choose another value for a and run the algorithm again.

Mathematically these requirements translate into the question: what is the chance that r is even and that each of the factors $(a^{r/2} - 1)$ and $(a^{r/2} + 1)$ is divisible by exactly one of the primes p and q? In Section 15.F we show that the probability for a suitable a is 50%, thus, if the trial – the random choice of a – is repeated sufficiently often, the factorization works with almost certainty.

15.3 How to Find the Period of a Function: The Magic Role of Entanglement

In the preceding section we have shown that the problem of factoring a number is closely related to finding the period of a function. In the present section we address the question of how to obtain the period in an efficient way in more detail. In particular, we use the entanglement of two quantum systems to map the periodicity of the function f defined by (15.4) encoded in one quantum system onto a second one, which we then read out using a Fourier transform.

15.3.1 Encoding in Quantum Systems

We consider a function f, which is defined for integer values j with a period r, that is,

$$f(j + r) = f(j) \; . \tag{15.8}$$

The goal is to determine r. For the sake of simplicity we assume that f can only take on integer values, and that these values are encoded by sequences of zeros and ones. To represent such a function by unitary operations we consider the unitary mapping U_f on product states defined by linear continuation of the rule

$$|a\rangle|b\rangle \mapsto |a\rangle|b \oplus f(a)\rangle \tag{15.9}$$

where $\oplus$ is binary addition. This operation is clearly bijective and we have $U_f^{-1} = U_f$. It is shown in [4] that it is indeed unitary for arbitrary f (which does not have to be injective itself).

We encode the argument j and the value of $f(j)$ in the quantum states of two quantum systems. Here j is always connected to the quantum system 1 whereas f is attached to the quantum system 2. We emphasize that neither the type of quantum system, nor the quantum state that we are considering, are of importance in this discussion. However, in order to guarantee that there is a one-to-one mapping between the arguments and the values of f with the corresponding quantum states, we consider two identical harmonic oscillators of frequency ω. Here the energy eigenvalues E_j are of the form [10]

$$E_j = \hbar\omega(j + 1/2) , \tag{15.10}$$

where $\hbar$ denotes the Planck constant and j runs from 0 to infinity. The corresponding energy eigenstates indicated by $|j\rangle$ are mutually orthogonal for different values of j, that is,

$$\langle j | j' \rangle = \delta_{jj'} . \tag{15.11}$$

We define the product state

$$|\Psi_j\rangle \equiv |j\rangle_1 |f(j)\rangle_2 \tag{15.12}$$

consisting of the quantum system 1 being in the energy eigenstate $|j\rangle_1$ and the quantum system 2 being in the energy eigenstate $|f(j)\rangle_2$. The state $|\Psi_j\rangle$ is created from the state $|j\rangle_1|0\rangle_2$ by the application of the unitary operation U_f. Since j runs through all integers the quantum state $|\Psi\rangle$ is the superposition

$$|\Psi\rangle = \sum_j |\Psi_j\rangle = \sum_j |j\rangle_1 |f(j)\rangle_2 \tag{15.13}$$

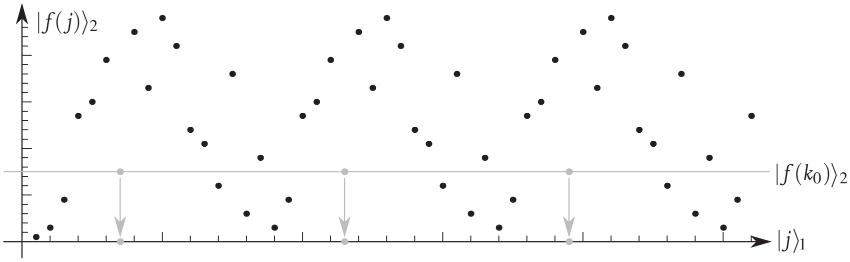


Figure 15.2 Mapping of periodicity using the entangled state $|\Psi\rangle = \sum_j |j\rangle_1 |f(j)\rangle_2$ given by (15.13) and a projective measurement onto $|f(k_0)\rangle_2$. The horizontal axis indicated by equally spaced ticks numbers the energy eigenstates $|j\rangle_1$ of one harmonic oscillator. The vertical axis marks the energy eigenstates $|f(j)\rangle_2$ of another harmonic oscillator. The entangled quantum state $|\Psi\rangle = \sum_j |\Psi_j\rangle$ is the "interference" of the product states $|\Psi_j\rangle = |j\rangle_1 |f(j)\rangle_2$. As a result we can view $|\Psi\rangle$ as the interference of all dots forming the graph $f(j)$. In this way we have obtained a geometrical representation of the entangled state $|\Psi\rangle$. A projection of $|f(k_0)\rangle$ onto $|\Psi\rangle$ can be depicted as the overlap between a straight line at $|f(k_0)\rangle_2$ parallel to the j-axis and the graph $f(j)$. The points common to both curves determine the quantum states $|j\rangle_1$ contributing to the so-created superposition $|\psi\rangle = \sum_m |k_0 + mr\rangle_1$. Obviously $|\psi\rangle$ inherits the periodicity of f. For this picture we have chosen the modular exponentiation $f(j) = a^j \bmod N$ for $a = 5$ and $N = 51$.

of all product states $|\Psi_j\rangle$. We emphasize that $|\Psi\rangle$ is a rather peculiar state. For specific choices of f, such as $a^j \bmod N$ it cannot be factored into the product of two quantum states describing the two individual quantum systems. Following Schrödinger [11] such quantum states are called entangled states.

We have not yet addressed the question of how to prepare $|\Psi\rangle$ in a real experiment. This question drives a very active field of research and has to be studied separately.

In Figure 15.2 we give an elementary representation of $|\Psi\rangle$. It is the graph of the function f. Indeed, we interpret the product state $|\Psi_j\rangle$ as a single point in a plane spanned by the argument j and the value $f(j)$ of the function encoded in the energy eigenstates of the two oscillators. The superposition of the product states $|\Psi_j\rangle$, that is, the interference of these dots represents $|\Psi\rangle$.

15.3.2 Mapping of Periodicity

The starting point of the method of finding the period of a function f is the entangled state $|\Psi\rangle$ given by (15.13). In order to take advantage of the periodicity of f it is useful to decompose the sum into a sum of periodic terms and a sum over an elementary cell consisting of one period. Figure 15.3 illustrates this principle summarized by the familiar summation rule

$$\sum_j a_j = \sum_{k=0}^{r-1} \sum_m a_{k+mr} \,. \qquad (15.14)$$

Indeed, this identity takes advantage of the constructive interference of identical terms and casts the quantum state $|\Psi\rangle$ into the form

$$|\Psi\rangle = \sum_{k,m} |k + mr\rangle_1 |f(k + mr)\rangle_2 \,. \qquad (15.15)$$

Due to the periodicity of f, given by (15.8), we find $f(k + mr) = f(k)$ and thus

$$|\Psi\rangle = \sum_k \sum_m |k + mr\rangle_1 |f(k)\rangle_2 . \qquad (15.16)$$

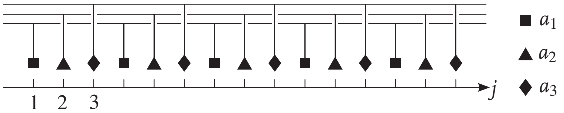


Figure 15.3 The geometrical representation of the summation formula, (15.14). In order to bring out the essential features we represent the individual contributions a_j to the sum by squares, triangles and diamonds. These symbols stand for state vectors, real or complex numbers. The sum of a_j over j can be performed by summing consecutive terms. However, when there is a periodicity in a_j it is more convenient to first sum over all terms that are equal and then add these results. Hence, we first sum over all squares, all triangles and all diamonds. When we add these three sums we obtain the final result for the total sum. In our example the period of a_j is 3. However, this method is even useful when the function a_j is not strictly periodic. For the example of a complex-valued a_j the phases may be periodic but not the amplitudes.

In this representation of $|\Psi\rangle$ we have two summations: (i) one over the periodic terms indicated by the summation index m; and (ii) one over the elementary cell denoted by the summation index k. We note that the second quantum system does not contain the summation over m. But the first system now involves the period r of f. In this way we have mapped the period of f from the second onto the first system.

15.3.3 Projection

Next we have to address the question of how to extract the period r from the first system. For this purpose we perform a measurement on the second quantum system, that is, we project [10] onto any state $|f(k_0)\rangle$ and we make use of the orthogonality Equation (15.11) of the energy eigenstates. As a result we now have to deal with the quantum state

$$|\psi\rangle := |\psi\rangle_1 := {}_2\langle f(k_0)|\Psi\rangle \; , \tag{15.17}$$

of a single system, which reads

$$|\psi\rangle = \sum_m |k_0 + mr\rangle_1 \; . \tag{15.18}$$

Obviously $|\psi\rangle$ is very different from $|\Psi\rangle$. First of all, it does not involve the second system; $|\psi\rangle$ represents a single harmonic oscillator. Moreover, we now do not have all integers present. The state $|\psi\rangle$ only contains quantum numbers j which are integer multiples m of the period r and start at some value k_0 determined by the measurement.

15.3.4 Phase State

In order to find the period r which is now stored in the selection of energy eigenstates given by $|\psi\rangle$, (15.18), we have to switch to the variable conjugate to a number, namely to the phase of the oscillator. By doing this, we can eliminate the dependence on k_0. Indeed, the value of k_0 is random. In each measurement we find a different value for k_0.

The concept of a Hermitian phase operator in quantum mechanics is a longstanding problem [18]. Notwithstanding all complications associated with the nonexistence of this operator, it is still possible to define the London phase states [19]

$$|\phi\rangle := \sum_j e^{ij\phi}|j\rangle \tag{15.19}$$

for a harmonic oscillator.

The phase distribution

$$P(\phi) = |\langle\phi|\psi\rangle|^2 \tag{15.20}$$

of the quantum state $|\psi\rangle$ is then determined by the phase probability amplitude

$$\langle\phi|\psi\rangle = \sum_m e^{-i(k_0+mr)\phi} = e^{-ik_0\phi} \sum_m e^{-imr\phi} \ . \tag{15.21}$$

In the last step we have recognized that the phase $k_0\phi$ is independent of the summation over m, and we can take it outside of the sum. As a result, the value of k_0 determined by the measurement of the second quantum system only enters as a phase.

When ϕ is an integer multiple of 2π the consecutive terms of the sum add up constructively and lead to a large value. When ϕ is different from an integer multiple of 2π, the phase factors interfere destructively and essentially lead to an almost vanishing result. Hence, the sum in (15.21) acts very much like a delta function located at integer multiples of $2\pi/r$, that is

$$\langle\phi|\psi\rangle = e^{-ik_0\phi} \sum_l \delta\left(l \cdot \frac{2\pi}{r} - \phi\right) \ . \tag{15.22}$$

The probability of finding a given phase ϕ in $|\psi\rangle$ is only nonzero at integer multiples l of the ratio $2\pi/r$ determined by the period r of the function f. Moreover, each term has equal weight.

15.3.5 Subtleties

In order to bring out the key ideas of the Shor algorithm, we have suppressed several subtleties of the calculation. In particular, we have neglected the fact that the sum over j is only a finite sum. In general it involves $\mathcal{N}$ states. Since the states of the two quantum systems must be normalized and each state contributes in a democratic way, that is, with equal weight, each state brings in a normalization factor $1/\sqrt{\mathcal{N}}$. This feature is quite important when we address scaling laws.

Moreover, the sum in (15.21) determining the phase distribution is not a Dirac δ-function but rather the square-root [6] of a δ-function. Indeed, the finite sum

$$\delta_{\mathcal{N}}^{(1/2)}(\phi) := \frac{1}{\sqrt{\mathcal{N}}} \sum_{m=0}^{\mathcal{N}-1} \exp(-imr\phi) \tag{15.23}$$

is a geometrical sum and can be performed. Since we only observe the probability, that is, the absolute value squared we find

$$\delta_{\mathcal{N}}(\phi) := \left|\delta_{\mathcal{N}}^{(1/2)}(\phi)\right|^2 = \frac{1}{\mathcal{N}} \frac{\sin^2(r/2\phi\mathcal{N})}{\sin^2(r/2\phi)} \ , \tag{15.24}$$

which, in the limit of large $\mathcal{N}$, approaches a δ-function.

As a result the phase distribution

$$P(\phi) = \delta_{\mathcal{N}}(\phi) \tag{15.25}$$

of the state ψ defined by (15.18) is a sequence of narrow peaks at integer multiples of $2\pi/r$. The width of each peak is determined by $1/\mathcal{N}$. In order to be able to resolve two neighboring peaks, their separation $2\pi/r$ has to be larger than their width $1/\mathcal{N}$, which gives rise to the condition

$$r < 2\pi\mathcal{N} \,. \tag{15.26}$$

We will find a similar condition in the analogy provided by atom optics discussed in the next section.

15.4 Analogy with Atom Optics

In the preceding section we have focused on the different steps of the quantum mechanical implementation of the Shor algorithm. The crucial element is the mapping of the periodicity of the function $f(j)$ encoded in one quantum system onto another one by entanglement and a subsequent projective measurement. In this section we analyze an alternative method of obtaining the period of a function. For this purpose we scatter atoms [6] from a standing light wave. Although this technique is classical it has many features in common with the Shor algorithm.

15.4.1 Scattering Atoms off a Standing Wave

We investigate the influence of a classical light field on the center-of-mass motion of an atom. In this treatment the atom is considered as a wave rather than a particle. This paradigm of atom optics is summarized in Figure 15.4.

A standing electromagnetic wave with a field distribution given by a mode function $u(x)$ is aligned along the x-direction. Here $u(x)$ is periodic with the wavelength $\lambda = 2\pi/k$, that is, $u(x + \lambda) = u(x)$. The gradient of the electric field exerts a force on the atom and deflects it from its original course. The initial wave function $\psi(x)$ of the atom along the x-axis is real and covers a large number $\mathcal{N}$ of wavelengths. As a result, the initial momentum p along the x-axis is negligible.

According to Appendix 15.G the distribution of the scattered atoms in the far field is determined by the momentum distribution

$$W(p) = \frac{1}{\hbar k}\delta_{\mathcal{N}}\left(\frac{p}{\hbar k}\right)|C_{p/\hbar k}(\beta)|^2 \tag{15.27}$$

at the exit of the field and consists of a comb

$$\delta_{\mathcal{N}}(\wp) = \frac{1}{\mathcal{N}}\frac{\sin^2(\wp\mathcal{N}\pi)}{\sin^2(\wp\pi)} \tag{15.28}$$

of peaks located at integer multiples $\wp$ of the elementary momentum $\hbar k$. Each peak has a width $1/\mathcal{N}$ and a weight determined by the function

$$C_{\wp}(\beta) = \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta \exp[-i(\wp\theta + \beta u(\theta/k))] \,. \tag{15.29}$$

Here β is the interaction parameter.

The periodicity of the mode function $u = u(x)$ manifests itself in the discreteness of the momentum distribution. The separation of neighboring peaks is determined by $\hbar k$, that is the wave vector $2\pi/\lambda$ of the light.

15.4.2 Method of Stationary Phase

The array of narrow peaks in the momentum distribution (15.27) is analogous to the comb of quasi-δ-functions in the phase distribution (15.25) of the periodic state $|\psi\rangle$, (15.18), created by the projective measurement. The analogy stands out most clearly, when we consider the probability amplitude $\psi(p)$ for the momentum p in the form

$$\psi(p) = \frac{1}{\sqrt{2\pi\hbar L}}\int_0^{L} \mathrm{d}x \exp[-iS(x)] \tag{15.30}$$

with the phase

$$S(x) = \frac{px}{\hbar} + \beta u(x) \,. \tag{15.31}$$

We derive this expression in appendix 15.G.

According to the stationary phase method, the main contributions to the integral in (15.30) arise [6] from the points of a stationary phase where the phase S oscillates

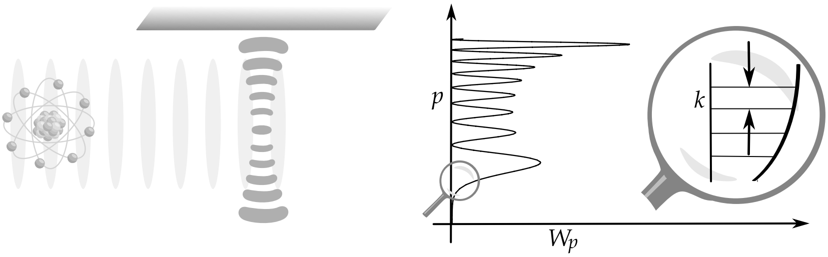


Figure 15.4 Determination of the period of a function with the help of the scattering of an atom from a classical standing electromagnetic wave. The function whose period we want to determine is encoded in the spatial distribution of the light field, that is in the mode function. The period emerges from the distribution of atoms in the far field, that is, from the momentum distribution of the atoms as they leave the wave. Due to the periodicity of the standing wave this distribution consists of a periodic array of narrow peaks whose envelope is determined by a Fourier transform involving the mode function. The separation of neighboring peaks yields the inverse of the period.

slowly. These points are determined by the condition

$$\frac{\partial S}{\partial x} = \frac{p}{\hbar} + \beta \frac{\mathrm{d}u(x)}{\mathrm{d}x} = 0 \,. \tag{15.32}$$

This equation can be represented geometrically in phase space spanned by position x and momentum p as the crossing point between a straight line parallel to the position axis and the derivative u' of the mode function with amplitude β. For an appropriate choice of p there will be at least one crossing $x_{\mathrm{c}}(p;\beta)$ in every period of the mode function. This notation expresses the fact that, for a fixed value of the interaction strength β, the crossing depends on the value of p. In addition the periodicity of the mode function leads to many such crossings $x_m = x_{\mathrm{c}} + m\lambda$ separated by λ, and m is an integer. Hence, a periodic array of positions contributes to the integral determining the momentum probability amplitude, in complete analogy to Figure 15.2.

As a result $\psi(p)$ given by the integral (15.30) can be approximated by

$$\psi(p) \approx \sum_m \frac{1}{\sqrt{i\hbar L S''(x_m)}} e^{-iS(x_m)} \,. \tag{15.33}$$

Since the mode function $u(x)$ is periodic, the phase S and its second derivative S'' evaluated at x_m reduce to

$$S(x_m) = \frac{p}{\hbar}(x_{\mathrm{c}} + m\lambda) + \beta u(x_{\mathrm{c}} + m\lambda) = S(x_{\mathrm{c}}) + 2\pi m \frac{p}{\hbar k} \tag{15.34}$$

and

$$S''(x_m) = \beta u''(x_{\mathrm{c}} + m\lambda) = \beta u''(x_{\mathrm{c}}) \,. \tag{15.35}$$

Hence, the momentum probability amplitude

$$\psi(p) = \mathcal{K}(p;\beta) \sum_m \exp\left(-2\pi i m \frac{p}{\hbar k}\right) \tag{15.36}$$

with

$$\mathcal{K}(p;\beta) = \frac{1}{\sqrt{i\hbar L \beta u''(x_{\mathrm{c}})}} e^{-iS(x_{\mathrm{c}})} \tag{15.37}$$

consists of an array of equally separated narrow peaks. They originate from the periodicity in space brought to light by the geometrical construction of the method of stationary phase.

15.4.3 Interference in Phase Space

Table 15.2 illuminates the analogy between the two methods of determining the period of a function using either entanglement or atom optics. Here the roles of the argument j of the function f determined by the modular exponentiation and its value $f(j)$ are played by the position variable x and the derivative of the mode

Table 15.2 Comparison between two methods determining the period of a function using either entanglement (left column) or scattering of atoms from a standing wave (right column).

Argument	j	x
Function	$f(j)$	$u(x)$
Periodicity	$f(j+r) = f(j)$	$u(x+\lambda) = u(x)$ $u'(x+\lambda) = u'(x)$
Encoding	$\lvert\Psi_j\rangle = \lvert j\rangle_1 \lvert f(j)\rangle_2$	$\lvert\psi_0\rangle = \lvert p = 0\rangle$
Quantum state	$\lvert\Psi\rangle = \sum_j \lvert\Psi_j\rangle$	$\lvert\psi\rangle = \exp[-i\beta u(\hat{x})]\lvert\psi_0\rangle$
Projection	$\lvert f(k_0)\rangle$	$\lvert p\rangle$
Periodic array	$\sum_m \lvert k_0 + mr\rangle$	$\sum_m \lvert x_c + m\lambda\rangle$

function $u(x)$, respectively. The projection onto $|f(k_0)\rangle$ corresponds to the projection of the momentum eigenstate $|p\rangle$ onto the state $|\psi\rangle$ of the center-of-mass motion, giving rise to the momentum probability amplitude $\psi(p) = \langle p|\psi\rangle$.

This analogy stands out most clearly in the semiclassical limit of quantum mechanics. Here, we can interpret scalar products of quantum mechanics such as $\langle p|\psi\rangle$ as interfering areas in phase space [6]. The momentum eigenstate $|p\rangle$ is a line parallel to the position axis and the quantum state $|\psi\rangle$ is the derivative of the mode function. The crossing points of the two curves interfere in phase space and provide the periodic array of narrow peaks in the momentum distribution. Hence, interference in phase space is at the heart of the Shor algorithm.

15.5
Exponential Growth of Hilbert Space as a Resource of Exponential Speedup

The analysis presented in the preceding section shows that scattering atoms from an electromagnetic wave reveals the period of the mode function $u = u(x)$ in an efficient way. It is interesting to note that this technique only takes advantage of interference. We emphasize that in this situation somebody has already created $u(x)$. For this purpose the preparator has to calculate $u(x)$ at every point x in space. This task is equivalent to calculating the modular exponentiation and determining the period of the function from these values.

In contrast, the quantum implementation of the modular exponentiation $f(j)$ does not calculate the values of f. It uses entanglement to map the period from one quantum system onto another. Moreover, there is a dramatic reduction in resources due to the large dimension of Hilbert space. Indeed, we do not encode the integers j in the energy eigenstates of a *single harmonic oscillator* but in an *array of two-level atoms*. Since the dimension of the Hilbert space of M two-level atoms grows as 2^M we can cover the integers from 1 to $\sqrt{N}$ by $M \sim \log N$ number of atoms. The exponential growth of Hilbert space is illustrated in Figure 15.5.

The unitary transformation U_f from Section 15.3.1 on appropriately prepared states encodes the values $f(j)$ for all values of j in the state by applying a single operation on $\log N$ atoms. The period of f is translated into a periodic amplitude distribution by the Fourier transform. The measurement therefore results in a col-

M	# states 2^M	
1	2	$\|0\rangle$
		$\|1\rangle$
2	4	$\|0,0\rangle$
		$\|1,0\rangle, \|0,1\rangle$
		$\|1,1\rangle$
3	8	$\|0,0,0\rangle$
		$\|1,0,0\rangle, \|0,1,0\rangle, \|0,0,1\rangle$
		$\|1,1,0\rangle, \|1,0,1\rangle, \|0,1,1\rangle$
		$\|1,1,1\rangle$

Figure 15.5 Power of Hilbert space illustrated by a collection of M two-level atoms consisting of an excited state $|1\rangle$ and a ground state $|0\rangle$. The dimension of the space grows as 2^M, that is, a single two-level atom has two states, whereas two two-level atoms already cover four states, three such atoms give eight states and therefore can be used to encode eight integers. As a result, we find an exponential gain in resources.

lapsed state which always belongs to a periodic point, in contrast to the classical case, where up to $N/2$ values $f(j)$ must be computed to find such a point. The probabilities of these periodic points are equidistributed, so an average of $\log N$ repetitions of the creation and measurement process suffice to infer the period. This concept is explained in detail for the general Fourier transform in the subsequent article by Dörn *et al.*

15.6 Conclusions

In the present paper we have discussed the problem of factoring a number. Here we have first discussed the mathematical algorithm underlying the celebrated Shor method and have then focused on its quantum mechanical implementation. Three essential ideas constitute our take-home message: (i) the translation of integer factorization into a period-finding problem; (ii) the use of entanglement to obtain the period; and (iii) optical interference as a classical substitute for entanglement. The last observation has led us to the question: what is the deeper origin of the speedup of a quantum algorithm compared to a classical one? Here we have identified the exponential growth of Hilbert space as a possible answer.

In this context it is interesting to compare and contrast the Shor algorithm with a recent proposal [14, 15] to factor numbers with the help of Gauss sums. Whereas the Shor algorithm relies on entanglement, the Gauss or exponential sum algorithms only take advantage of interference. For this reason, in the present form, they scale exponentially in the number of digits. However, it is interesting to note that both techniques are based on the summation rule (15.14) as shown in Appendix 15.H.

We conclude by noting that already two experiments have implemented the Shor algorithm. They rely either on methods of nuclear magnetic resonance [20] or optical interferometry using entangled photons [21]. These experiments with an enormous effort could factor the number $N = 15$. In contrast, the Gauss sum algorithm has been able to successfully decompose seventeen digit numbers [22, 23]. However, this impressive success will eventually be stopped by the exponential complexity. At that point only entanglement can bring an improvement.

15.A Modular Arithmetic

It was Carl Friedrich Gauss who started a new branch of mathematics, called modular arithmetic. The basic ingredient is periodicity. In this appendix we first briefly motivate the concept of residue classes [9] and then show two examples of modular arithmetic: modular multiplication and modular exponentiation.

15.A.1 Basic Idea

The principle of residues stands out most clearly for the example of time in our daily life. Time is measured by a clock with two hands. The large hand tells us the minutes whereas the small one points to the hours. In the present context we are only interested in full hours, that is, in the position of the small hand on the clock. Most old-fashioned clocks with a face display the hours from 1 to 12. Since the day has 24 hours, every number $1 \leq r \leq 12$ is visited by the hand twice in a day, corresponding to the hours r and $1 \cdot 12 + r$. When we extend our considerations from a single to many days, the number r indicated by the hand could represent hours of any integer multiple of 12 plus r. Hence, the clock only tells us the remainders with respect to integer multiples of 12.

This example of the clock suggests a generalization of numbers, called residue classes. Indeed, any integer number n can be represented as an integer multiple k of 12 plus a remainder r, that is,

$$n = k \cdot 12 + r\,. \tag{15.38}$$

We realize that there is an infinite number of integers n, which eventually lead to the same remainder r, for example, the numbers $34 = 2 \cdot 12 + 10$ and $58 = 4 \cdot 12 + 10$ both lead to the remainder $r = 10$. Hence, 34 and 58 are equivalent in the sense that they have the same remainder when considered with respect to multiples of 12. For this reason we call them congruent to each other. Since only the remainder of a number matters, it is useful to introduce a separate mathematical symbol for it. The expression

$$34 \equiv 58 \mod 12 \tag{15.39}$$

is a shorthand notation for the fact that 34 and 58 are equivalent in the sense that they lead to the same remainder 10 with respect to integer multiples of 12. The abbreviation mod 12 indicates the period and carries the name module. Since 34 and 58 are equivalent, they belong to the same class. All numbers with the same remainder form a so-called residue class. In the example of 12 there are the 12 residue classes 0, 1, . . . , 11.

15.A.2 Modular Multiplication

It is now possible to define arithmetic operations such as the addition and multiplication of two residue classes. Needless to say, due to the periodicity inherent in the definition of the residue classes, the outcomes of these operations lead to results that are dramatically different from those suggested by our experience with integer numbers.

In order to illustrate this point we consider the multiplication of the two residues 23 mod 35 and 29 mod 35. These classes contain the numbers $k \cdot 35 + 23$, that is, 23, 58, 93, . . . and $l \cdot 35 + 29$, that is, 29, 64, 99, ... We cast the product

$$(k \cdot 35 + 23)(l \cdot 35 + 29) = k \cdot l \cdot 35^2 + 23 \cdot l \cdot 35 + 29 \cdot k \cdot 35 + 23 \cdot 29 \qquad (15.40)$$

into multiples of 35, that is

$$(k \cdot 35 + 23)(l \cdot 35 + 29) = (k \cdot l \cdot 35 + 23 \cdot l + 29 \cdot k) \cdot 35 + 667\ . \qquad (15.41)$$

Since $667 = 19 \cdot 35 + 2$ we find

$$(k \cdot 35 + 23)(l \cdot 35 + 29) = (k \cdot l \cdot 35 + 23 \cdot l + 29 \cdot k + 19) \cdot 35 + 2\ , \qquad (15.42)$$

that is an integer multiple of 35 with the remainder 2. As a consequence, the product of the two residue classes 23 mod 35 and 29 mod 35 is 2 mod 35.

The above calculation clearly shows that the final result 2 mod 35 is only determined by the remainder of the product of the remainders 23 and 29 taken with respect to mod 35, that is

$$23 \cdot 29 = 667 \equiv 2 \quad \text{mod } 35\ . \qquad (15.43)$$

In summary, we have established the multiplication of residues as the multiplication of remainders with respect to the module.

15.A.3 Modular Exponentiation

Our second example of modular multiplication is modular exponentiation and plays a central role in the factorization of numbers. Here we calculate the powers of a given integer a modulo the number N, that is we consider the function

$$f(j) = a^j \text{ mod } N\ . \qquad (15.44)$$

We now derive a recurrence relation for $f(j)$ and show that this function is periodic.

15.A.3.1 Recurrence Relation

Table 15.3 displays the values of the function $f(j)$ for the example $a = 3$ and $N = 14$. Here we follow a rather naive approach. We first calculate the powers of a and then find the remainders by dividing with respect to N. Unfortunately in this way we have to deal with large numbers.

A more efficient method relies on a recurrence relation for $f(j)$. For its derivation we start from the relation

$$a^j = l \cdot N + r_j \tag{15.45}$$

where l is an integer and r_j is the residue of a^j mod N. As a result we find

$$a^{j+1} = a \cdot a^j = a \cdot (l \cdot N + r_j) = a \cdot l \cdot N + a \cdot r_j\ , \tag{15.46}$$

which together with the definition $f(j+1) = a^{j+1}$ mod N yields the desired recurrence relation

$$f(j+1) = a \cdot r_j \bmod N = a \cdot f(j) \bmod N\ . \tag{15.47}$$

As a consequence we obtain $f(j+1)$ by multiplying $f(j)$ by a and taking the product modulo N. For example, in order to calculate $f(3)$ we recall from Table 15.3 the value $f(2) = 9$ and arrive with the help of $a = 3$ at $f(3) = 3 \cdot 9$ mod $14 = 27$ mod $14 = 13$.

15.A.3.2 Periodicity

Table 15.3 brings out an important feature of the function $f(j)$. It is periodic, that is, $f(j + r) = f(j)$. In this particular example the period is $r = 6$.

In this context the basic concept is the order. If a and N are relatively prime, the order of a^j mod N abbreviated by $\text{ord}_N a$ is the smallest positive exponent r for which $a^r \equiv 1$ mod N. The period of the function $f(j)$ is $\text{ord}_N a$.

However, it is not always easy to find the period of $f(j)$. In Figure 15.6 we illustrate this statement for the case of $N = 2143$ and $a = 35$, where the values of $f(j)$ are quasi-random. This feature results from three facts: (i) a^x is a rapidly increasing nonlinear function of x; (ii) the values a^x are cut up into equidistant parts by the modular operation mod N; and (iii) the integer numbers $x = j$ are discrete and their separation is constant. The combination of these three ingredients leads to the rather irregular distribution shown in Figure 15.7.

Table 15.3 The function $f(j) = a^j$ mod N with $N = 14$ and $a = 3$ for increasing integer values of j displays a period of $r = 6$.

j	0	1	2	3	4	5	6	7
a^j	1	3	9	27 $1 \cdot 14 + 13$	81 $5 \cdot 14 + 11$	243 $17 \cdot 14 + 5$	729 $52 \cdot 14 + 1$	2187 $156 \cdot 14 + 3$
$f(j)$	1	3	9	13	11	5	1	3

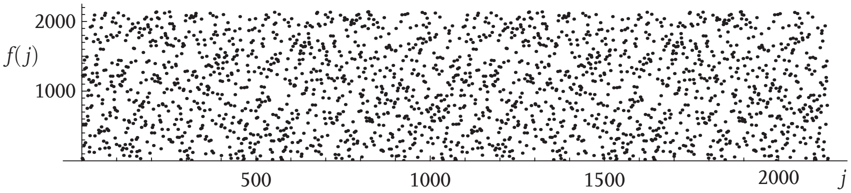


Figure 15.6 The function $f(j) = a^j \bmod N$ for $N = 2143$ and $a = 35$ is periodic. In contrast to the example of Table 15.3 now the period is not obvious since the values of f scatter over all integer numbers up to N.

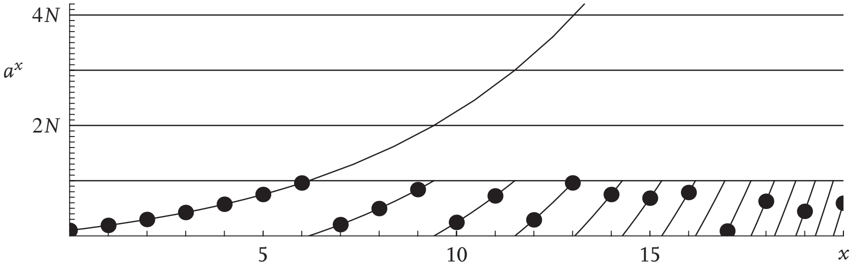


Figure 15.7 The modular operation cuts up the rapidly increasing function a^x into equidistant parts arranged along the horizontal axis. The discreteness and the equidistant separation of the integer numbers j leads to a quasi-random distribution of the function $a^j \bmod N$ indicated by the full circles.

Therefore, it is not straightforward to recognize the period of $f(j)$. In the worst case we have to evaluate $f(j)$ for $N/2$ arguments before we can tell the period. However, tools of elementary number theory, such as the Chinese remainder theorem and Euler's function will help us to find the period of the function in an efficient way as discussed in the next sections.

15.B
Chinese Remainder Theorem

In the preceding section we have recognized that modular multiplication or exponentiation can quickly involve large numbers. Obviously, it would be much more convenient to work with smaller ones. The Chinese remainder theorem [9] is an elegant way to achieve this goal.

15.B.1
Residues Represented as a Matrix

The Chinese remainder theorem has been known – in principle – from antiquity. The Chinese mathematician Sun-Tsu treated special cases of it in the first century

Table 15.4 The chinese remainder theorem illustrated for the example $N = 35 = 5 \cdot 7$. The residues 0, 1,. . . , 34 are arranged in a matrix with the rows and columns determined by the residue classes mod 5 and mod 7, respectively.

		mod 7						
		0	**1**	**2**	**3**	**4**	**5**	**6**
mod 5	**0**	0	15	30	10	25	5	20
	1	21	1	16	31	11	26	6
	2	7	22	2	17	32	12	27
	3	28	8	23	3	18	33	13
	4	14	29	9	24	4	19	34

A.D. ([24], [25]). These investigations became known in the west only in the 19th century ([24], [26]). Before their rediscovery, however, Gauss already had systematically developed the idea in his famous book on number theory – Disquisitiones Arithmeticae [27].

The Chinese remainder theorem expresses the residues mod N of a number N consisting of the product of pairwise coprime numbers by an array of residues of these numbers. In order to bring out the essential ideas we now discuss the Chinese remainder theorem for the special case $N = p \cdot q$, that is, we represent the residues mod N by two components, a residue mod p and a residue mod q. We illustrate this representation, using the example $N = 35 = 5{\cdot}7$, leading to the residues mod p = mod 5 and mod q = mod 7.

For $N = 35$ we have the residue classes 0, 1, 2, . . . , 34. These numbers build up the matrix shown in Table 15.4 and spanned by the residue classes 0, 1, . . . , 4 corresponding to mod 5 and the residue classes 0, 1, . . . , 6 corresponding to mod 7. The location of the number with respect to the row and the column of the matrix is determined by the residue classes with respect to mod 5 and mod 7, respectively. In order to bring out this matrix arrangement, we consider the example of 23. Since $23 = 4 \cdot 5 + 3$ and $23 = 3 \cdot 7 + 2$ the number 23 appears in the row corresponding to the residue 3 and the column corresponding to the residue 2. Obviously numbers which are multiples of 5, such as 0, 5, 10, . . . , 30, are in the first row corresponding to the residue 0. However, their ordering is determined by the remainder with respect to mod 7. For example $15 = 2 \cdot 7 + 1$ leads to the remainder of 1 with respect to mod 7, and is therefore in the column corresponding to 1. The first column contains the multiples of 7. Along the diagonal and the off-diagonals of the matrix, the numbers increase by unity, because both remainders increase by unity. When we come to the last row, the next higher number is located at the top of the next column. This feature reflects the periodicity of the modular operation.

15.B.2
Modular Multiplication

We now return to the problem of evaluating the product of the residues 23 mod 35 and 29 mod 35. In contrast to the treatment performed in 15.A we now take advantage of the Chinese remainder theorem.

From Table 15.4 we find the correspondences

$$23 \bmod 35 \sim (3 \bmod 5, 2 \bmod 7) \tag{15.48}$$

and

$$29 \bmod 35 \sim (4 \bmod 5, 1 \bmod 7) \ , \tag{15.49}$$

which lead to the correspondence

$$(23 \bmod 35)(29 \bmod 35) \sim (3 \bmod 5, 2 \bmod 7)(4 \bmod 5, 1 \bmod 7) \ . \tag{15.50}$$

Hence, in order to obtain the product of the two remainders, we have to multiply two pairs of numbers, that is

$$(23 \bmod 35)(29 \bmod 35) \sim (12 \bmod 5, 2 \bmod 7) = (2 \bmod 5, 2 \bmod 7) \ . \tag{15.51}$$

From Table 15.4 we recognize the correspondence $(2 \bmod 5, 2 \bmod 7) \sim 2 \bmod 35$, and thus

$$(23 \bmod 35)(29 \bmod 35) = 2 \bmod 35 \ , \tag{15.52}$$

in complete accordance with our earlier result, (15.43).

15.B.3
Period of Function from Chinese Remainder Theorem

We now return to the problem of determining the period of the function $f(j)$ defined by (15.44) and show that the Chinese remainder theorem is of great help. In order to bring out the essential ideas we again consider the example of $N = 35 = 5 \cdot 7$ with $a = 23$, and find the period of

$$f(j) = 23^j \bmod 35 \ . \tag{15.53}$$

According to Table 15.4 the residue 23 mod 35 corresponds to the pair of residues (3 mod 5, 2 mod 7). As a consequence we can write the powers of 23 as a pair of powers

$$23^j \equiv 3^j \quad \bmod 5 \quad \text{and} \quad 23^j \equiv 2^j \quad \bmod 7 \ . \tag{15.54}$$

In Table 15.5 we list the powers $3^j \bmod 5$ and $2^j \bmod 7$, which reveal that $3^j \bmod 5$ has a period of 4 and $2^j \bmod 7$ has a period of 3. From the multiplication rules of the Chinese remainder theorem we know, that $23^j \equiv 1 \bmod 35$ is only possible if $3^j \equiv 1 \bmod 5$ and $2^j \equiv 1 \bmod 7$. Hence, the period is the least common multiple of 4 and 3, that is 12.

Table 15.5 Determining the period of the function $23^j \bmod 35$ using the Chinese remainder theorem. The period of $23^j \equiv 3^j \bmod 5$ is 4 and the period of $23^j \equiv 2^j \bmod 7$ is 3. Therefore, the period of $23^j \bmod 35$ is the least common multiple of 4 and 3, that is 12.

j	1	2	3	4	5	6	7	8	9	10	11	12
23^j	23^1	23^2	23^3	23^4	23^5	23^6	23^7	23^8	23^9	23^{10}	23^{11}	23^{12}
3^j	3^1	3^2	3^3	3^4	3^5	3^6	3^7	3^8	3^9	3^{10}	3^{11}	3^{12}
$3^j \bmod 5$	3	4	2	1	3	4	2	1	3	4	2	1
2^j	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9	2^{10}	2^{11}	2^{12}
$2^j \bmod 7$	2	4	1	2	4	1	2	4	1	2	4	1

15.C Euler's Function

Euler's function [9] denoted by ϕ is an essential ingredient in the evaluation of the period of a function, such as modular exponentiation. It displays the multiplicative property which follows from the Chinese remainder theorem. In the present section we briefly review the properties of ϕ and give a prescription for the way to calculate it.

15.C.1 Definition

Euler's function $\phi(n)$ is defined as the number of residue classes $\bmod\, n$ that are coprime to n, that is, that have no common divisor with n other than 1. To illustrate this definition we consider the example $n = 12$. In Table 15.6 we list the residue classes 0 to 11 together with their greatest common divisor (gcd) with the module 12. We observe that this greatest common divisor always has to be a divisor of 12 and thus one of the numbers 1, 2, 3, 4, 6 or 12. Table 15.6 shows that only the residue classes corresponding to 1, 5, 7, 11 have the greatest common divisor 1 with 12 leading to four coprime residue classes, that is, $\phi(12) = 4$.

Table 15.6 Determination of the coprime residues and the value of Euler's function $\phi = \phi(n)$ for $n = 12$. Only the four residue classes 1, 5, 7, 11 marked by a shaded background have the greatest common divisor 1 and are therefore coprime, leading to $\phi(12) = 4$.

r mod 12	0	1	2	3	4	5	6	7	8	9	10	11
gcd with 12	12	1	2	3	4	1	6	1	4	3	2	1

15.C.2
Multiplicative Property

Euler's function satisfies the multiplicative property

$$\phi(mn) = \phi(m)\phi(n) \ . \tag{15.55}$$

for coprime integers m and n.

This feature follows directly from the Chinese remainder theorem, which identifies each residue class mod mn with a pair $(r \bmod m, s \bmod n)$ of components. For example, a residue class mod 35 is coprime exactly if its 7-component and its 5-component are both coprime. These are exactly the residue classes mod 35 that are neither divisible by 7 nor by 5, giving rise to $\phi(35) = \phi(5)\phi(7)$, which is the multiplicative property.

15.C.3
A Compact Formula for ϕ

We now discuss a method to compute $\phi(n)$. Here we make use of three properties: (i) it is easy to calculate $\phi(p)$ for prime numbers p and powers k of prime numbers; (ii) the Euler function is multiplicative; and (iii) every integer number can be represented in a unique way as a product of powers of prime numbers.

We start with the case when $n = p$ is a prime number. By definition p has only the two divisors 1 and p. We have a total number p of the residue classes $0, 1, \ldots, p-1$. The only residue class that is divisible by p is 0, all other ones can not be divided by p and hence are coprime. As a result, we have a total of $p - 1$ coprime residue classes and thus

$$\phi(p) = p - 1 \ . \tag{15.56}$$

Next we turn to the case of a prime power, $n = p^k$. The greatest common divisor also has to be a power of p. The only ones of the residue classes $0, 1, \ldots, p^k - 1$ that are divisible by a power of p are the p^{k-1} multiples of p: $0, p, 2p, \ldots, (p^{k-1} - 1)p$. As a result we find the expression

$$\phi(p^k) = p^k - p^{k-1} = p^k(1 - 1/p) \ . \tag{15.57}$$

In order to deal with an arbitrary integer n we first recall, that according to the fundamental theorem of arithmetic we can decompose every integer

$$n = p_1^{k_1} \cdot \ldots \cdot p_m^{k_m} \ , \tag{15.58}$$

in a unique way into powers $k_1, k_2, \ldots, k_m$ of prime numbers $p_1, p_2, \ldots, p_m$, leading to

$$\phi(n) = \phi(p_1^{k_1} \cdot \ldots \cdot p_m^{k_m}) \ . \tag{15.59}$$

When we apply the multiplicative property of ϕ, (15.55), we arrive at

$$\phi(n) = \phi(p_1^{k_1}) \cdot \ldots \cdot \phi(p_m^{k_m}) \,, \tag{15.60}$$

which with the help of (15.57) reduces to

$$\phi(n) = p_1^{k_1}(1 - 1/p_1) \cdot \ldots \cdot p_m^{k_m}(1 - 1/p_m) \,. \tag{15.61}$$

With the help of the definition (15.58) of n we find

$$\phi(n) = n \prod_{p|n} (1 - 1/p) \,. \tag{15.62}$$

It is interesting to note that the exponents k_i of the prime number p_i defining n do not enter into this expression.

We now apply the formula (15.62) for ϕ to reconfirm the result $\phi(12) = 4$ obtained in Section 15.C.1 by counting the coprime residues of 12. Since the prime numbers contained in 12 are 2 and 3, we find indeed

$$\phi(12) = 12 \cdot \left(1 - \frac{1}{2}\right)\left(1 - \frac{1}{3}\right) = 12 \cdot \frac{1}{2} \cdot \frac{2}{3} = 4 \,. \tag{15.63}$$

In Figure 15.8 we display Euler's function ϕ for $1 \le n \le 100$.

We conclude our brief introduction into Euler's function by returning to the example of f shown in Figure 15.6. Here we had chosen $N = 2143$. Since $N = 2143$ is a prime number, we find from (15.56) the result $\phi(2143) = 2143–1 = 2142 = 2{\cdot}1071$. Indeed, from the pattern of white and black spots in Figure 15.6 we recognize that the period is about 1000. A closer analysis shows that it is 1071. Hence, the period of f is a divisor of Euler's function. We will revisit this property in the context of the primitive root discussed in Section 15.E.

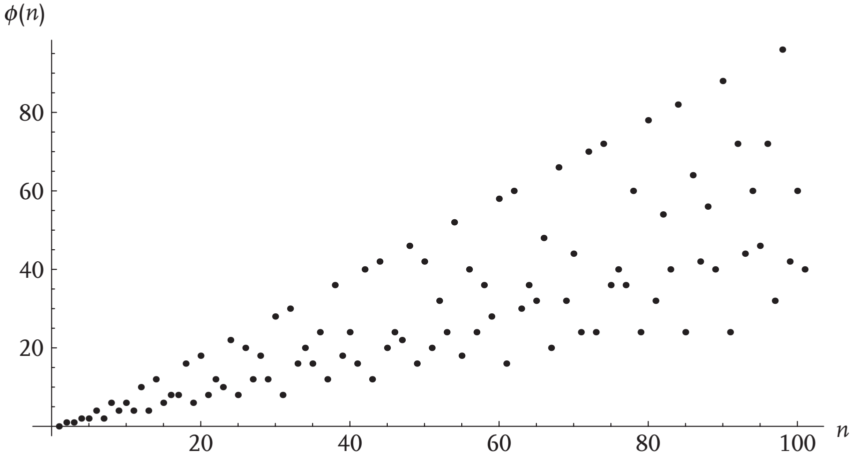


Figure 15.8 Euler's function $\phi(n)$ for $1 \le n \le 100$. The values of ϕ at prime numbers p are $p{-}1$ and form a straight line, providing an upper bound for ϕ. For nonprime values the function is rapidly varying.

15.D
Euclidean Algorithm

The greatest common divisor (gcd) of two positive integers a and b may quickly be determined by the Euclidean algorithm. It is based on division with a remainder. If $b < a$ are two positive integers, then there are integers q and r with $0 \leq r < b$ such that $a = qb + r$. The Euclidean algorithm [9] is the iteration of this division, that is

$$\begin{aligned} a &= q_1 b + r_1 \\ b &= q_2 r_1 + r_2 \\ r_1 &= q_1 r_2 + r_3 \\ &\vdots \\ r_{n-2} &= q_n r_{n-1} + r_n \\ r_{n-1} &= q_{n+1} r_n + 0 \,. \end{aligned} \tag{15.64}$$

The algorithm terminates if the remainder in the division is 0. This last step must occur in a finite number of steps, since the remainders r_k are decreasing in size.

One easily sees that

$$\gcd(a, b) = \gcd(b, r_1) = \gcd(r_1, r_2) = \cdots = \gcd(r_{n-1}, r_n) = r_n \,. \tag{15.65}$$

Thus the last nonzero remainder r_n is the gcd of a and b.

For the estimate of the running time we observe that, if $r_{k+1} > (1/2)r_k$, then $r_{k+2} \leq (1/2)r_k$. In any case $r_{k+2} \leq (1/2)r_k$. Thus the algorithm terminates in, at most, const. × log b steps.

We conclude by illustrating the Euclidean algorithm for the example of $a = 2457$ and $b = 553$. According to the prescription (15.64) we find

$$\begin{aligned} 2457 &= 4 \cdot 553 + 245 \\ 553 &= 2 \cdot 245 + 63 \\ 245 &= 3 \cdot 63 + 56 \\ 63 &= 1 \cdot 56 + 7 \\ 56 &= 8 \cdot 7 \,. \end{aligned} \tag{15.66}$$

Thus the greatest common divisor of 2457 and 553 is 7.

15.E
Primitive Root

The Shor algorithm relies on determining the period of the modular exponentiation. It is known [9] that the period is a divisor of the Euler function ϕ. Hence, ϕ gives us a hint for this period. Another approach to finding the period takes advantage of the concept of the primitive root [9]. In the present section we define and illustrate primitive roots and illuminate the connection to the period of modular exponentiation.

15.E.1
Definition

The concept of the primitive root stands out most clearly when we consider an example. In Table 15.7 we calculate the function $f(j) = a^j \bmod N$ for $a = 3$ and $N = 7$. We note, that $f(j)$ is periodic with the period 6. Obviously the residue classes mod 7 consist of the integers $1, \dots, 6$. Hence, all residue classes are coprime residues, that is they do not have a common divisor with 7. Thus we obtain all the coprime residues mod 7 as powers of 3. In this sense 3 is a primitive root. Indeed, a primitive root modulo N is an integer r such that the residues that have no common divisor with N – the coprime residues – are all powers of r.

In general, N does not have a primitive root. However, there always is one, if N is a prime number, as suggested by our example of $N = 7$. This fact was stated without proof by Lambert in 1769 [28]. Euler gave a defective proof in 1773 [29] and finally Legendre [30] provided a full proof in 1798.

15.E.2
Periods for Prime Numbers

Primitive roots can be used to calculate the period of $a^j \bmod N$ when N is a prime number. In the example of Table 15.7 the function $f(j) = 3^j \bmod 7$ has period 6. On the other hand we have exactly six coprime residues and hence, $\phi(7) = 6$. This example shows the connection between the period of f, the primitive root and Euler's function.

However, for other choices of a there exists a more sophisticated technique which we now illustrate using the example of $2^j \bmod 7$. From Table 15.7 we recognize that $2 \equiv 3^2 \bmod 7$ and therefore

$$2^j \equiv 3^{2\cdot j} \mod 7\ . \tag{15.67}$$

In order to find the period of $2^j \bmod 7$, we arrange the six powers of $3^j \bmod 7$ in a hexagon, shown in Figure 15.9. Then the powers $2^j \equiv 3^{2j} \bmod 7$ form a triangle consisting of each second vertex. This means that one needs three steps to get from $2^0 = 1$ back to 1, that is, $2^3 \bmod 7$. Thus the period of the function $2^j \bmod 7$ is 3.

We now turn to the general case of a prime number p with primitive root r and $a = r^\ell$. Again we arrange the powers of $r^j \bmod p$ in a polygon with $p - 1$ vertices, shown in Figure 15.10. Then we connect the vertices corresponding to the powers $a^j \equiv r^{j\ell} \bmod p$ by a path consisting of each ℓ-th vertex.

Table 15.7 The function 3^j mod 7 takes all integer values from 1 to 6, which are all coprime with 7. Therefore, 3 is a primitive root of 7.

j	1	2	3	4	5	6
3^j **mod 7**	3	2	6	4	5	1

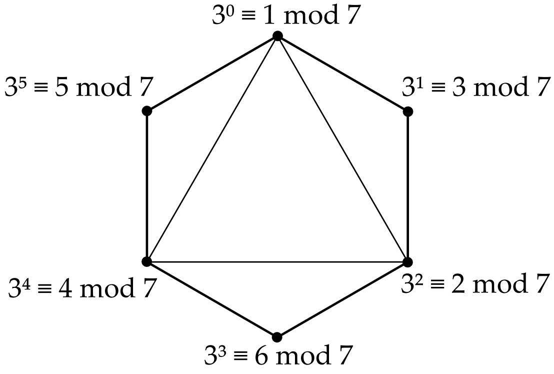


Figure 15.9 Determination of the period of $2^j \bmod 7$ using the primitive root 3 of mod 7. The six powers of $3^j \bmod 7$ are arranged in the form of a hexagon. Since $2^j \equiv 3^{2 \cdot j} \bmod 7$ only every second corner of the hexagon participates leading to a triangle and the period 3.

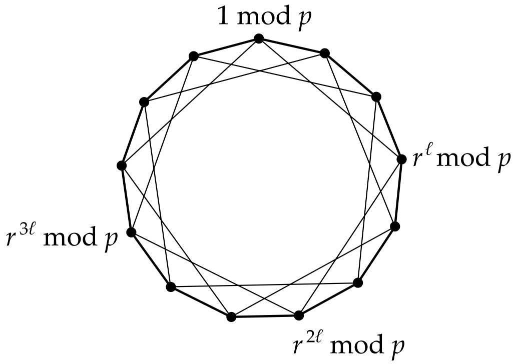


Figure 15.10 Determination of the period of the function $a^j \bmod p$ with the primitive root r. In this case we arrange the powers $r^j \bmod p$ in a polygon with $p-1$ vertices and connect every ℓ-th vertex. The period is then $(p-1)/\gcd(\ell, p-1)$.

How many vertices does this path contain? If $d = \gcd(\ell, p-1)$ is the *greatest common divisor* of ℓ and $p-1$, then this number is $(p-1)/d$, see [9]. Thus the period $\text{ord}_p a$ of the function $f(j) = a^j \bmod p$ reads

$$\text{ord}_p\, a = \frac{p-1}{\gcd(\ell, p-1)} \ . \tag{15.68}$$

We conclude by briefly mentioning the general problem of finding the period for N which is a product of prime numbers. In this case we can use the Chinese remainder theorem described in Section 15.B.3.

15.F Probability for Lucky Choice

In Section 15.2.3 we have seen that not every a leads to a period r that is useful for finding the factors of N. This fact is closely related to the question: what is the chance that r is even and that each of the factors $(a^{r/2} - 1)$ and $(a^{r/2} + 1)$ is divisible by exactly one of the primes p and q?

In the preceding sections we have become familiar with the tools needed to calculate the probability that our choice for a is suitable. They are the Chinese remainder theorem and the primitive root. In this section we put these concepts to use and derive the probability of success. Since these arguments are rather abstract we illustrate them for the specific example of $N = 35 = 5 \cdot 7$ in Table 15.8.

15.F.1 Expression for the Period

The number $N = p \cdot q$ to be factored consists of two prime numbers p and q. According to the Chinese remainder theorem the base a of the modular exponentiation can be represented by the pair $(a_p \bmod p, a_q \bmod q)$. For a_p we now determine the order mod p. Since p is a prime number there exists a primitive root ϱ_p and an exponent e_p so that $\varrho_p^{e_p} \equiv a_p \bmod p$. According to (15.68) the order is

$$r_p = \text{ord}_p\, a_p = \frac{p-1}{\gcd(e_p, p-1)} \,. \tag{15.69}$$

We follow the same procedure with the second prime q and find

$$r_q = \text{ord}_q\, a_q = \frac{q-1}{\gcd(e_q, q-1)} \,. \tag{15.70}$$

These calculations are only valid if a_p and a_q are nonzero, that is, a is coprime to both primes p and q.

In Section 15.B.3 we have seen that the period of $a^j \bmod N$ is the least common multiple of the two periods r_p and r_q, that is,

$$\text{ord}_N a = \text{lcm}(r_p, r_q) \,. \tag{15.71}$$

Since p and q are prime numbers and we assume $p, q \neq 2$ they must be odd. As a result $p-1$ and $q-1$ are even. We now factor out powers of 2 from $p-1$ and $q-1$, that is

$$p - 1 = 2^{s_p} t_p \tag{15.72}$$

and

$$q - 1 = 2^{s_q} t_q \tag{15.73}$$

with t_p and t_q odd. Then the pair of exponents is of the form $e_p = 2^{u_p} v_p$ and $e_q = 2^{u_q} v_q$ and the greatest common divisors read

$$\gcd(e_p, p-1) = 2^{w_p} \gcd(t_p, v_p) \tag{15.74}$$

and

$$\gcd(e_q, q-1) = 2^{w_q} \gcd(t_q, v_q) \tag{15.75}$$

with $w_p = \min(u_p, s_p)$ and $w_q = \min(u_q, s_q)$.

When we substitute (15.72)–(15.75) into the expressions (15.69) and (15.70) for the periods r_p and r_q we arrive at

$$r_p = 2^{s_p - w_p} \frac{t_p}{\gcd(t_p, v_p)} \tag{15.76}$$

and

$$r_q = 2^{s_q - w_q} \frac{t_q}{\gcd(t_q, v_q)} \ . \tag{15.77}$$

These two expressions constitute the main result of this appendix. They allow us now to estimate the probability of success.

15.F.2
Analysis of Different Cases

The factorization fails if: (i) the period r is odd; (ii) p and q both divide $a^{r/2} + 1$; or (iii) p and q both divide $a^{r/2} - 1$. We first show that each of these cases occurs only if $s_p - w_p$ and $s_q - w_q$ are equal. We then demonstrate that, with a probability of at least 50%, the exponents $s_p - w_p$ and $s_q - w_q$ are not equal, in which case the factorization works.

Indeed, the situation (i) of an odd value of r only appears when $s_p - w_p = s_q - w_q = 0$. For (ii) the heighest power of 2 dividing r is 2^m with $m = \max(s_p - w_p, s_q - w_q)$ since, according to (15.76) and (15.77), r is the least common multiple of r_p and r_q. When we now assume that $s_q - w_q < s_p - w_p$, r_q divides $r/2$ and thus $a^{r/2} \equiv 1 \bmod q$, that is, q does not divide $a^{r/2} + 1$. In the same manner it follows from $s_q - w_q < s_p - w_p$ that p does not divide $a^{r/2} + 1$. Thus case (ii) can indeed only occur if $s_q - w_q = s_p - w_p$. Case (iii) cannot occur at all, since if p and q both divide $a^{r/2} - 1$, the order of a would be at most $r/2$ and not r.

If $s_p = s_q$ all choices of the exponents e_p, e_q, in which one of u_p, u_q is equal to zero and the other one different from 0, give unequal values for $s_p - w_p$ and $s_q - w_q$. These are 50% of all choices. If $s_p < s_q$ all choices in which $u_q = 0$, u_p arbitrary, give unequal values of $s_p - w_p$ and $s_q - w_q$. These are again 50% of all choices. By symmetry this also holds for the case $s_q < s_p$. If the trial – the random choice of a – is repeated sufficiently often, the factorization almost certainly works.

Table 15.8 "Good" and "bad" choices for a in the factorization of $N = 35 = 5 \cdot 7$. The shaded rows indicate cases for which the algorithm does not produce a factor. We also indicate the numbers used in the argument of Section 15.F providing the success probability.

a	e_p	e_q	u_p	u_q	v_p	v_q	w_p	w_q	r	$\gcd(N, a^{r/2} \pm 1)$
1	4	6	2	1	1	3	2	1	1	35, 1
2	3	4	0	2	3	1	0	1	12	7, 5
3	1	5	0	0	1	5	0	0	12	7, 5
4	2	2	1	1	1	1	1	1	6	7, 5
6	4	3	2	0	1	3	2	0	2	5, 7
8	1	6	0	1	1	3	0	1	4	7, 5
9	2	4	1	2	1	1	1	1	6	7, 5
11	4	2	2	1	1	1	2	1	3	1, 1
12	3	1	0	0	3	1	0	0	12	7, 5
13	1	3	0	0	1	3	0	0	4	7, 5
16	4	4	2	2	1	1	2	1	3	1, 1
17	3	5	0	0	3	5	0	0	12	7, 5
18	1	2	0	1	1	1	0	1	12	7, 5
19	2	1	1	0	1	1	1	0	6	1, 35
22	3	6	0	1	3	3	0	1	4	7, 5
23	1	4	0	2	1	1	0	1	12	7, 5
24	2	5	1	0	1	5	1	0	6	1, 35
26	4	1	2	0	1	1	2	0	6	5, 7
27	3	3	0	0	3	3	0	0	4	7, 5
29	2	6	1	1	1	3	1	1	2	7, 5
31	4	5	2	0	1	5	2	0	6	5, 7
32	3	2	0	1	3	1	0	1	12	7, 5
33	1	1	0	0	1	1	0	0	12	7, 5
34	2	3	1	0	1	3	1	0	2	1, 35

We conclude by illustrating this argument for the case $N = 35 = 5 \cdot 7$ in Table 15.8. A choice of primitive roots is $\varrho_p = 3$ and $\varrho_q = 5$, and we have $s_p = 2$ and $s_q = 1$. The case $s_p - w_p = s_q - w_q$ occurs for six different values of a out of $\phi(N) = 24$ possible candidates for a coprime to N.

15.G Elements of Atom Optics

In this appendix, we analyze the scattering [6] of an atom from a classical electromagnetic wave of a periodic mode function. In particular, we derive an expression for the distribution of atoms in the far field. Our analysis is based on two assumptions: (i) the interaction time is short and the atoms do not move substantially while they are in the light field. This limit allows us to make the Raman–Nath approximation and neglect the operator of the kinetic energy in the Hamiltonian compared to the one of the interaction energy; and (ii) the initial distribution of atoms is broad compared to the period of the field. As a result, the momentum distribution of the atoms after the interaction consists of discrete momenta determined by the spatial period of the field. The envelope of this distribution is governed by the mode function.

15.G.1 Quantum State of Motion in Raman–Nath Approximation

The propagation of the de Broglie wave is governed by the time-dependent Schrödinger equation

$$i\hbar\frac{\mathrm{d}}{\mathrm{d}t}|\psi(t)\rangle = \hat{H}|\psi(t)\rangle \tag{15.78}$$

for the quantum state $|\psi\rangle = |\psi(t)\rangle$ of the center-of-mass motion. While the atom traverses the light field, the dynamics is governed by the Hamiltonian

$$\hat{H} = \frac{\hat{p}^2}{2M} + \hbar\kappa u(\hat{x}) \tag{15.79}$$

which contains the momentum operator $\hat{p}$ of the atom of mass M along the standing wave. The mode function $u(x)$ is periodic with the wavelength $\lambda = 2\pi/k$, that is $u(x+\lambda) = u(x)$ and the parameter κ denotes the interaction strength of the field.

Since the Hamiltonian is time independent a formal solution of the Schrödinger equation reads

$$|\psi(t)\rangle = \exp\left(-\frac{i}{\hbar}\hat{H}t\right)|\psi_0\rangle \tag{15.80}$$

where $|\psi_0\rangle$ denotes the quantum state of the center-of-mass motion before the interaction with the light.

When we assume that initially the atom had no momentum along the light field we can neglect the kinetic energy in $\hat{H}$ compared to the interaction term. In this so-called Raman–Nath approximation the quantum state of the system after the interaction time τ reduces to

$$|\psi\rangle \approx e^{-i\beta u(\hat{x})}|\psi_0\rangle\ . \tag{15.81}$$

with the interaction parameter $\beta = \kappa\tau$. This approximation is justified as long as the displacement of the atom due to the field is small compared to the wavelength.

15.G.2 Momentum Distribution

We are now interested in the distribution of atoms on a screen which is aligned parallel to the standing wave and far away from it. Once the atoms have left the field, they move freely, that is, in the absence of any force. Their ultimate position on the screen is only determined by: (i) the momentum they have gained due to their interaction with the light; and (ii) the duration of the free motion. As a consequence the position distribution on the screen is determined by the momentum distribution

$$W(p) = |\psi(p)|^2 = |\langle p|\psi\rangle|^2 \tag{15.82}$$

after the interaction with the field. Here

$$\psi(p) = \langle p|\psi\rangle = \int_{-\infty}^{\infty} \mathrm{d}x \, \langle p|x\rangle\langle x|\psi\rangle \tag{15.83}$$

is the probability amplitude in momentum space with $|x\rangle$ and $|p\rangle$ denoting position and momentum eigenstates, respectively.

According to the Born interpretation [10] of the wave function, the probability $W(p)$ of finding the momentum p to be between p and $p + \mathrm{d}p$ reads

$$W(p)\,\mathrm{d}p = |\psi(p)|^2 \,\mathrm{d}p \,. \tag{15.84}$$

From the scalar product [10]

$$\langle p|x\rangle = \frac{1}{\sqrt{2\pi\hbar}} e^{-ipx/\hbar} \tag{15.85}$$

between the position and the momentum eigenstates, together with the initial wave function $\psi_0(x) = \langle x|\psi_0\rangle$ in position space we obtain the expression

$$\psi(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{\infty} \mathrm{d}x \, e^{-ipx/\hbar} e^{-i\beta u(x)} \psi_0(x) \tag{15.86}$$

for the momentum probability amplitude.

15.G.3
Discreteness of Momentum due to Interference

We now consider a situation when the initial atomic wave ψ_0 covers many periods of the standing wave and only changes very slowly over one period. In this case we can approximate $\psi_0(x)$ by a constant, that is,

$$\psi_0(x) = \frac{1}{\sqrt{L}} [\Theta(x) - \Theta(x - L)] \tag{15.87}$$

where $\Theta(x)$ denotes the Heaviside step function.

As a consequence, (15.86) reduces to

$$\psi(p) = \frac{1}{\sqrt{2\pi\hbar L}} \int_0^L \mathrm{d}x \exp[-i(px/\hbar + \beta u(x))] \,. \tag{15.88}$$

When we introduce the dimensionless integration variable $\theta = kx$ and the dimensionless momentum $\wp = p/\hbar k$ together with the new probability amplitude

$$\tilde{\psi}(\wp) = \sqrt{\hbar k}\psi(\hbar k\wp) \tag{15.89}$$

we arrive at

$$\tilde{\psi}(\wp) = \frac{1}{\sqrt{\mathcal{N}}}\frac{1}{2\pi}\int_0^{2\pi\mathcal{N}} \mathrm{d}\theta \exp[-i(\wp\theta + \beta\tilde{u}(\theta))] \,. \tag{15.90}$$

Here $\tilde{u}(\theta) = u(\theta/k)$ and $kL = 2\pi L/\lambda = 2\pi\mathcal{N}$ denotes the number of periods.

Next we decompose the range of integration into intervals of 2π, that is

$$\tilde{\psi}(\wp) = \frac{1}{\sqrt{\mathcal{N}}}\sum_{\nu=0}^{\mathcal{N}-1}\frac{1}{2\pi}\int_{2\pi\nu}^{2\pi(\nu+1)} \mathrm{d}\theta \exp[-i(\wp\theta + \beta\tilde{u}(\theta))] \,. \tag{15.91}$$

The new integration variable $\bar{\theta} = \theta - 2\pi\nu$ and the periodicity of the mode function $\tilde{u}$ in the phase of the integral lead to

$$\tilde{\psi}(\wp) = \delta_{\mathcal{N}}^{(1/2)}(\wp)C_{\wp}(\beta) \tag{15.92}$$

with the abbreviations

$$\delta_{\mathcal{N}}^{(1/2)}(\wp) := \frac{1}{\sqrt{\mathcal{N}}}\sum_{\nu=0}^{\mathcal{N}-1}\exp(-2\pi i\nu\wp) \tag{15.93}$$

and

$$C_{\wp}(\beta) := \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\theta \exp[-i(\wp\theta + \beta\tilde{u}(\theta))] \,. \tag{15.94}$$

The sum defining $\delta_{\mathcal{N}}^{(1/2)}$ is a geometric sum and can thus be performed in closed form. According to (15.82) the distribution $\tilde{W}(\wp)$ of the dimensionless momentum $\wp$ is the absolute value squared of the probability amplitude $\tilde{\psi}(\wp)$. Hence, it is useful to establish the relation [6]

$$\delta_{\mathcal{N}}(\wp) := \left|\delta_{\mathcal{N}}^{(1/2)}(\wp)\right|^2 = \frac{1}{\mathcal{N}}\frac{\sin^2(\wp\mathcal{N}\pi)}{\sin^2(\wp\pi)} \,. \tag{15.95}$$

This expression also explains the notation $\delta_{\mathcal{N}}^{(1/2)}$. Indeed, for large values of $\mathcal{N}$ the function $\delta_{\mathcal{N}}$ consists of narrow peaks at integer values of $\wp$. The width of each peak in $\wp$ is essentially $1/\mathcal{N}$ and its height is proportional to $\mathcal{N}$ leading to an effective area which is independent of $\mathcal{N}$. In the limit $\mathcal{N} \to \infty$ the function $\delta_{\mathcal{N}}$ approaches a comb of Dirac δ-functions. In this sense $\delta_{\mathcal{N}}^{(1/2)}$ approximates the square-root of a delta function.

As a consequence, the momentum distribution

$$\tilde{W}(\wp) = |\tilde{\psi}(\wp)|^2 = \delta_{\mathcal{N}}(\wp)|C_{\wp}(\beta)|^2 \tag{15.96}$$

consists of narrow peaks at integer values $\wp$ with a weight given by $|C_{\wp}(\beta)|^2$.

We conclude by noting that the discreteness of the momentum, that is the emergence of the function $\delta_{\mathcal{N}}$ is due to the interference of identical behavior of the atom in each of the many periods of the standing wave. In order to obtain narrow structures, the interference of many periods is necessary.

15.H
Factorization with a Gauss Sum due to its Periodicity

Periodicity is a crucial element in finding factors of a number – this is the central lesson of the present paper. The Shor algorithm uses the periodicity of the modular exponentiation. However, we can also employ other periodic functions. Indeed, recently several ideas to factor numbers [14] based on Gauss sums have been proposed. These methods rely solely on interference. There exist many different versions of Gauss sum factorization. In this appendix we address the most elementary one. In particular, we identify the summation formula (15.14) as the central ingredient of this factorization technique.

To bring out this feature most clearly, we now consider the Gauss sum [13]

$$G(\ell, N) = \sum_{j=0}^{N-1} \exp(2\pi i j^2 \ell / N) \ , \tag{15.97}$$

where $N = p \cdot q$ is the product of the two integers p and q.

Such Gauss sums can be calculated analytically and are known since the days of Gauss. We now show that the periodicity of the phase factors defining the Gauss sum G, (15.97), allows us to factor numbers. This feature is a consequence of partial constructive interference.

Indeed, for $\ell = q$ we can cancel the factor q in N and the Gauss sum reads

$$G(q, p \cdot q) = \sum_{j=0}^{q \cdot p-1} \exp(2\pi i j^2 1/p) \ . \tag{15.98}$$

Since the phase factors in this sum have the period p, that is

$$\exp[2\pi i(j+p)^2 1/p] = \exp[2\pi i(j^2 + 2jp + p^2)1/p] = \exp[2\pi i j^2 1/p] \ , \tag{15.99}$$

we can apply the summation formula, (15.14), and find

$$G(q, p \cdot q) = \sum_{k=0}^{p-1} \sum_{m=0}^{q-1} \exp[2\pi i(k + mp)^2 1/p] = \sum_{k=0}^{p-1} \sum_{m=0}^{q-1} \exp[2\pi i k^2 1/p] \tag{15.100}$$

or

$$G(q, p \cdot q) = q \sum_{k=0}^{p-1} \exp(2\pi i k^2 1/p) \ , \tag{15.101}$$

that is,

$$G(q, p \cdot q) = qG(1, p) \ . \tag{15.102}$$

Hence, for a factor of N the individual phase factors contributing to the Gauss sum (15.97) interfere constructively and lead to a large significant value.

References

1 Koblitz, N. (**1994**) *A Course in Number Theory and Cryptography*, Springer, New York.

2 Shor, P. (**1994**) *Algorithms for Quantum Computation: Discrete Logarithms and Factoring*. Proceedings of the 35th Annual Symposium on Foundations of Computer Science, Santa Fe, NM, edited by S. Goldwasser (IEEE Computer Society Press, New York), 124–134.

3 Shor, P. (**1997**) Polynomial-Time Algorithm for Prime Factorization and Discrete Logarithms on a Quantum Computer. *SIAM Journal of Computing*, **26**, 1484.

4 Nielsen, M.A. and Chuang, I.L. (**2000**) *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge.

5 Stenholm, S. and Suominen, K.-A. (**2005**) *Quantum Approach to Informatics*, John Wiley, New York.

6 Schleich, W.P. (2001) *Quantum Optics in Phase Space*, Wiley & sons Ltd., Berlin.

7 Mermin, N.D. (**2007**) What has quantum mechanics to do with factoring? *Physics Today*, **4**, 8–9.

8 Mermin, N.D. (**2007**) Some curious facts about quantum factoring. *Physics Today*, **10**, 10–11.

9 Rosen, K. (**1988**) *Elementary Number Theory and Its Applications*, Addison Wesley, Reading.

10 Bohm, D. (**1951**) *Quantum Theory*, Prentice Hall, Englewood Cliffs.

11 Schrödinger, E. (**1935**) Die gegenwärtige Situation in der Quantenmechanik. *Die Naturwissenschaften*, **23**, 807–812; 823–828; 844–849.

12 Bleistein, N. and Handelsman, R. (**1975**) *Asymptotic Expansions of Integrals*, Dover, New York.

13 Maier, H. and Schleich, W.P. (**2009**) *Prime Numbers 101: A Primer on Number Theory*, Wiley & sons Ltd., New York.

14 Merkel, W., Averbukh, I.S., Girard, B., Mehring, M., Paulus, G.G. and Schleich, W.P. (**2007**) Factorization of Numbers with Physical Systems, in *Elements of Quantum Information*, (eds W.P. Schleich and H. Walther), Wiley & sons Ltd., Weinheim.

15 Stefanak, M., Haase, D., Merkel, W., Zubairy, M.S. and Schleich, W.P. (**2008**) Factorization with exponential sums. *Journal of Physics A*, **41**, 304024.

16 Alford, W.R., Granville, A. and Pomerance, C. (**1994**) There are infinitely many Carmichael numbers. *Annals of Mathematics* **140**, 703–722.

17 Rabin, M. (**1980**) Probabilistic algorithms for testing primality. *Journal of Number Theory*, **12**, 128–138.

18 Barnett. S.M. and Vaccaro, J.A. (**2007**) *The Quantum Phase Operator*, Taylor & Francis, London.

19 London, F. (**1927**) Über die Jacobischen Transformationen der Quantenmechanik. *Zeitschrift für Physik*, **37**, 915 (1926); Winkelvariable und kanonische Transformationen in der Undulationsmechanik, *Zeitschrift für Physik*, **40**, 193.

20 Vandersypen, L.M.K., Steffen, M., Breyta, G., Yannoni, C.S., Sherwood, M.H. and Chuang, I.L. (**2001**) Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature*, **414**, 883.

21 Lanyon, B.P., Weinhold, T.J., Langford, N.K., Barbieri, M., James, D.F., Gilchrist, A. and White, A.G. (**2007**) Experimental Demonstration of a Compiled Version of Shor's Algorithm with Quantum Entanglement. *Physical Review Letters*, **99**, 250505.

22 Weber. S., Chatel, B. and Girard, B. (**2008**) Factoring numbers with interfering random waves. *Europhysics Letters*, **83**(3), 34008.

23 Peng, X. and Suter, D. (**2008**) NMR implementation of Factoring Large Numbers with Gauß-Sums: Suppression of Ghost Factors. *Europhysics Letters*, **84**, 40006.

24 Dickson, L.E. (**1919**) *History of the Theory of Numbers*, AMS Chelsea Publishing, Providence.

25 Mikami, Y. (**1912**) *Abh. Geschichte Math. Wiss.*, **30**, 33.

26 Wylie, A. (**1852**) Jottings on the Science of the Chinese Arithmetic. *North China Herald*.

27 Gauss, C.F. (**1801**) *Disquisitiones Arithmeticae*, Springer, Berlin (1986).

28 Lambert, J.H. (**1769**) *Nova Acta Eruditorum, Leipzig*, http://www.mathematik.uni-bielefeld.de/~sieben/Rechnen/lambert.pdf.

29 Euler, L. *Novi Comm. Acad. Petrop*. **18**, 85 1773, *Comm. Arith*. **1**, 516–537.

30 Legendre, A.M. (**1798**) *Mem. Ac. R. Sc.*, Paris. *Théorie des nombres*, 471–73.

16
Isomorphism and Factorization – Classical and Quantum Algorithms

Sebastian Dörn, Daniel Haase[1], Jacobo Torán, Fabian Wagner

16.1
Introduction

The integer factorization problem (IF) consists of, being given an integer n, finding a prime factor decomposition of n. Graph isomorphism (GI) is the problem of deciding whether two given graphs are isomorphic, or in other words, whether there is a bijection between the nodes of both graphs, respecting the adjacency relation. These are two well known natural problems with many applications and with a long history in the fields of mathematics and computer science. Moreover, their decisional versions are the best known examples of problems in the class NP that are not known to be efficiently solvable (in the class P) or hard for NP. They have an intermediate complexity and this lack of an exact classification has attracted much attention to both problems in the past.

(The decisional version of) IF is one of the few examples of problems in NP ∩ coNP for which efficient algorithms (running in polynomial time over the length of the representation of n) are not known. It is believed to be a hard problem. In fact a considerable portion of modern cryptology relies on the supposition that IF cannot be efficiently solved. The best algorithm for IF runs in time $O(\exp(c \log(n)^{1/3} \log\log(n)^{2/3})$ for some constant c.

GI is not known to be in NP ∩ coNP but in a probabilistic generalization of this complexity class. The best known algorithm, testing the isomorphism of two unrestricted graphs with n nodes each, runs in time $O(\exp(c\sqrt{n\log(n)}))$. On the other hand, there are algorithms for the problem that work efficiently for “almost all graphs” and in fact a straightforward linear time algorithm can decide isomorphism for random graphs.

The best classical algorithms for IF and GI therefore run in exponential time in the input size. The situation is different in the field of quantum computation. In 1994 Shor gave an efficient quantum algorithm for factoring integers. His methods have been extended to more general algebraic problems and there is a hope that efficient quantum algorithms for graph isomorphism might also be developed.

[1] Corresponding author.

In this chapter we give an overview of several attempts to obtain efficient classical and quantum algorithms for IF and GI. In doing this we point out several similarities between both problems. Finally we review a result showing that IF and GI are, in fact, particular instances of a more general algebraic problem; the ring isomorphism problem.

16.2 Factorization of Integers: Classical Algorithms

A natural number p is called a *prime number*, if it is divided by exactly two natural numbers (1 and p). The number 1 is not a prime. Any natural number $n \in \mathbb{N}$ admits a decomposition

$$n = p_1^{c_1} \cdots p_r^{c_r}$$

into prime powers. The exponents $c_j \in \mathbb{N}$ are uniquely determined by n. A famous theorem of analytic number theory, the *prime number theorem* (see [1]), states that

$$\lim_{x\to\infty} \frac{\pi(x)}{x/\log(x)} = 1 \,, \quad \pi(x) = \#\{p \text{ prime} \mid p \le x\}$$

so the distribution of prime numbers among the natural numbers is asymptotically $x/\log(x)$. The *integer factorization* problem (IF) is defined as follows: given $n \in \mathbb{Z}$, find the primes p_j and exponents c_j of its prime factor decomposition.

No classical algorithm of polynomial complexity is yet known for this problem. Some important cryptographic systems (like RSA) draw their security from this fact. Factorization is a search problem. A decisional version of it (given $n, k \in \mathbb{Z}$ is there a prime factor of n grater than k?) belongs clearly to the class $\mathsf{NP} \cap \mathsf{coNP}$. Note that the inverse problem, i.e. the computation of n from its given factors, is extremely simple and is therefore used in various public key systems and key exchange protocols which draw their security from mathematical problems which are difficult to solve, but easy to verify.

The naive approach to the factorization problem (brute force checking of all possible prime factors) needs at most $O(\sqrt{n})$ operations to compute the prime factorization of n (the smallest prime contained in n is bounded by $\sqrt{n}$ if n is composite). The prime number theorem tells us that we do not gain a substantial speedup by restricting the search to prime numbers (not taking into account the cost to compute them). We mention the following algorithms which have been proposed in the last century to compute the prime factorization:

- Factorization using reduced *quadratic forms*, the ideas going back to C.F. Gauss. Today it is known that the computation of generators of these forms is actually as hard as classical factorization, and not practicable for large numbers. This method gave rise to several number theoretical generalizations of the factorization problem, all known to be in $\mathsf{NP} \cap \mathsf{coNP}$, but still without an efficient classical algorithm.

- *Pollard's ϱ-method*, using the birthday phenomenon to find pairs (x, y) with $\gcd(x - y, n) \neq 1$.
- The *elliptic curve method*, using the fact that noninvertible elements of $\mathbb{Z}/n\mathbb{Z}$ share a factor with n, and such elements can be found using the group law on elliptic curves defined over $\mathbb{Z}/n\mathbb{Z}$. This is a typical example of an algorithm using group structures to factor numbers.
- The various *sieve methods*, the most sophisticated algorithm for factoring numbers known today. They make heavy use of the number theoretic background of the factorization problem, especially the theory of number fields.

The best known algorithm today to solve this problem is the *general number field sieve* (see [2, 10.5]) with expected running time

$$O\left(\exp\left(\log(n)^{1/3} \cdot \log(\log(n))^{2/3} \cdot (C + o(1))\right)\right)$$

for a constant $C \approx 1{,}922$. We note that until 2002, there was no efficient deterministic algorithm to check if a given number is actually a prime. This is due to the fact that the prime number theorem gives only an asymptotic distribution of the prime numbers. If we pick a small interval $[a, b]$ for very large a, the distribution of primes in it is not structured at all. The fine-grained distribution is not known even today, it is connected to the famous Riemann Hypothesis which has been the field of research for many decades without a proof in sight.

16.3 Graph Isomorphism: Classical Algorithms

The *graph isomorphism* problem (GI) consists of, being given two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, deciding whether there is a bijection $f\colon V_1 \to V_2$ respecting the adjacency relations of the graphs. In symbols, for every $u, v \in V_1$, $(u, v) \in V_1 \iff (f(u), f(v)) \in V_2$. GI is clearly in NP and its complement is contained in AM, a probabilistic generalization of NP [3]. GI is not believed to be hard for NP. However, no polynomial time algorithms for the problem are known either.

The earliest significant algorithms for deciding isomorphism were restricticted to trees [4,5]. They provided a canonical enumeration of the input graphs that could be computed in linear time. The same technique was used for the isomorphism of planar graphs [6]. Several years later this result could be extended to graphs of bounded genus [7–9].

Babai used for the first time [10] a group-theoretic approach to the graph isomorphism problem. He was able to prove that the problem restricted to colored graphs (isomorphisms have to preserve the colors) for which the color multiplicity is bounded, can be solved in random polynomial time. Based on this work, Furst, Hopcroft and Luks [11] developed polynomial time algorithms for several permutation-group problems. They also were able to derandomize Babai's algorithm, making it deterministic.

Using associated results on the structure of permutation groups, a breakthrough was obtained by Luks in [12] when he gave a polynomial time algorithm for testing isomorphism of graphs of bounded degree. By providing a new degree-reduction procedure and using Luks result, Zemlyachenco [13] managed to give a moderately exponential procedure of $\exp(n^{(2/3+o(1))})$ for deciding isomorphism for unrestricted graph classes. Subsequent improvements in the bounded degree algorithm has brought this bound down to $\exp(c\sqrt{n \log n})$, (announced in [14]), which is still the algorithm for unrestricted graph isomorphism with the lowest worst-case complexity.

There are several algorithms based on vertex-classification schemes that work well in practice. This is not surprising since it is known that trivial algorithms perform well on randomly generated graphs. Babai, Erdős and Selkow in [15] gave a straightforward linear time canonical labeling algorithm for GI, proving that it works well for almost all graphs.

The existing algorithms for graph isomorphism could roughly be divided into two main groups: those based on vertex classification methods and those constructing canonical labelings of the graphs.

- *Vertex classification methods.* A natural technique to restrict the search space when looking for an isomorphism is to divide the vertices of the input graphs in certain classes so that the vertices in one class in the first graph can only be mapped to vertices of the corresponding class in the second graph. Some ways of doing this are, to divide the vertices acording to their degree, the degree of their neighbors or the number of vertices reachable by paths of a certain length, etc. This method can be used iteratively, refining the classifications of the vertices according to previous classifications.
- *Canonical labeling methods.* The idea here is to find canonical representatives for the different isomorphic graph classes. This problem is, in principle, harder than deciding isomorphism but in some known examples of restricted graph classes (like trees or planar graphs) the algorithms for isomorphism do in fact provide canonical representatives. For unrestricted graphs the best known labeling algorithm works in $\exp(n^{1/2+o(1)})$ steps [14].

For more facts about the graph-isomorphism problem and its structural complexity, we refer the reader to the textbook by Köbler, Schöning and Torán [16].

16.4
Quantum Algorithms for Integer Factorization

The first efficient algorithm used to factor integers on quantum computers was given by Peter W. Shor in [17]. His idea was to view the factorization problem as a period-finding problem. Such problems can be solved on quantum computers using the quantum Fourier transformation as we clearly explain in the next section. Here we briefly show how to encode the factorization problem into a period-finding problem.

Let N be the number to be factored and $a \in \{2, \dots, N\}$ be chosen randomly. We first compute $\gcd(a, N)$ using Euclids Algorithm. If $\gcd(a, N) \neq 1$ the gcd is already a factor of N, otherwise we define the function $f(n) = a^n \bmod N$, a mapping from $\mathbb{Z}$ to $\mathbb{Z}/N\mathbb{Z}$. Its smallest period is called the *order* of a mod N (because it is the order of the multiplicative group generated by a if multiplication is defined mod N). First, since $(\mathbb{Z}/N\mathbb{Z})^\times = \{a \bmod N \ : \ \gcd(a, N) = 1\}$ has at most $N - 1$ elements (0 mod N is never included in this set), the order of a is strictly less than N. We can compute the order as shown in the next section using $O(\log(N))$ measurements (it should be noted that already the computation of the gcd takes up to $\log(N)^3$ steps).

By definition of the order $a^n \equiv 1 \bmod N$, which means N is a multiple of $a^n - 1$. Suppose the order is even, then N is a multiple of $(a^{n/2} - 1)(a^{n/2} + 1)$, so we get a nontrivial factor of N by computing $\gcd(N, a^{n/2} + 1)$. Some elementary number theory shows that the probability of hitting a such that the order is odd is bounded by $\phi(N)2^{-m}$, where $\phi(N) = \#\{a \bmod N : \quad \gcd(a, N) = 1\} < N$ is Eulers Totient function, and m is the number of prime powers in the prime factor decomposition of N. The total number of measurements needed to find a proper factor of N with given error probability ε is polynomial in $\log(N)$ and $\log(\varepsilon^{-1})$, as was proved first by Shor.

16.4.1 The Quantum Fourier Transform and Period Finding

The concept of Fourier transformation is a fundamental tool in many areas of research. The quantum Fourier transform used in the field of quantum computation is, viewed mathematically, the Fourier transform on the finite abelian group $\mathbb{Z}/n\mathbb{Z}$. In general, the *Fourier transform* of a finite abelian group G is given by

$$\hat{f}(\chi) = \sum_{g \in G} f(g)\bar{\chi}(g)$$

where $f : \ G \to \mathbb{C}$ is any function and $\chi : \ G \to \mathbb{C}^\times$ is a homomorphism of groups from G to the multiplicative group $\mathbb{C}^\times = \mathbb{C}\backslash\{0\}$, and $\bar{z}$ is the complex conjugate of $z \in \mathbb{C}$. For finite cyclic groups, $\chi(g)^n = \chi(g^n) = 1$ for the order $n = \#G$, so $\chi(g)$ is actually a root of unity, necessarily of the form $\chi_a(g) = \exp(2\pi i(ag/n))$ for some $a \in \mathbb{Z}$. We may identify G with $\mathbb{Z}/n\mathbb{Z}$ and each $a \in G$ with the homomorphism $\chi_a(g)$, and let

$$\sum_{g \in G} f(g)\bar{\chi}_a(g) = \sum_{g \in G} f(g)e^{-2\pi i a g/n}$$

be the Fourier transform of f at the value $a \in G$. The *quantum Fourier transform* (QFT) is the normalized Fourier transform

$$\hat{f}(a) = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} f(j)e^{-2\pi i a j/n}$$

on $\mathbb{Z}/n\mathbb{Z}$. The Fourier operator $\hat{}$ is actually an automorphism of the $\mathbb{C}$-space of functions $f : G \to \mathbb{C}$, the inverse transformation is given by

$$f(j) = \frac{1}{\sqrt{n}} \sum_{a=0}^{n-1} \hat{f}(a)e^{2\pi i a j/n} \ .$$

The most interesting property of this transformation is the translation law. Let $f(a+k)$ be the function $a \mapsto f(a+k)$, the function f shifted by k, then its Fourier transform is

$$\widehat{f(a+k)} = \hat{f}(a) \cdot e^{2\pi iak/n}$$

for any $k \in \mathbb{Z}$, that is shifts are mapped to complex prefactors by the Fourier transform. Suppose f is periodic: $f(a + p) = f(a)$ for some $p \in \mathbb{Z}$ and all $a \in \mathbb{Z}$, then we get

$$\hat{f}(a) = \widehat{f(a+p)} = \hat{f}(a) \cdot e^{2\pi iap/n} \Rightarrow \hat{f}(a) \cdot \left(e^{2\pi iap/n} - 1\right) = 0 \quad \forall a \in \mathbb{Z}$$

which forces $\hat{f}(a)$ to be zero if a is not a multiple of n/p. The converse is also true, so the QFT defines an isomorphism

$$\left\{\ f\colon\ \mathbb{Z}/n\mathbb{Z} \to \mathbb{C} \text{ of period } p\ \right\} \leftrightarrow \left\{\ f\colon\ \mathbb{Z}/n\mathbb{Z} \to \mathbb{C} \text{ supported on } \frac{n}{p}\mathbb{Z}\ \right\} .$$

So if we want to compute the period of a function $f : \mathbb{Z}/n\mathbb{Z} \to \mathbb{C}$ which can be accessed only by evaluation, we compute $\hat{f}(a)$ for sufficiently many $a \in \mathbb{Z}/n\mathbb{Z}$ and calculate the greatest common divisor of those a for which $\hat{f}(a)$ does not vanish. Classically there is no speedup in finding the period this way, but computing a point in the support of a function is an easy task for a quantum computer, since this is what measurement actually does. Let us briefly note that the ket-notation $|k\rangle = |k\rangle(t)$ denotes the (column) vector whose k-th amplitude component is one. In the following we regard this

$$|k\rangle : \{0, \dots, n-1\} \to \mathbb{C}\ , \quad t \mapsto \begin{cases} 1 & \text{if } t = k \\ 0 & \text{otherwise} \end{cases}$$

as a function. The basic period-finding algorithm is as follows. Let $f\colon\ \mathbb{Z}/n\mathbb{Z} \to \mathbb{C}$ be a function of period p. We assume the period is primitive, that is f has no period which is smaller then p. First, initialize two registers in the state

$$\psi_1 = \psi_1(t) = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} |k\rangle|f(k)\rangle\ .$$

Measurement of the second register to a value y collapses the state to

$$\psi_2 = \frac{1}{\sqrt{\#\{k \in \mathbb{Z}/n\mathbb{Z} : f(k) = y\}}} \cdot \sum_{f(k)=y} |k\rangle|y\rangle = \frac{1}{\sqrt{p}} \cdot \sum_{j=0}^{p-1} |x + jp\rangle|y\rangle$$

since f is injective on the set $\{0, 1, \dots, p-1\}$ and $f(k) = y$ is equivalent to $k = x+jp$ for $x \in \{0, \dots, p-1\}$ uniquely defined by y. Now we apply the QFT to this state, which transforms the p-periodic state to an n/p-periodic supported state

$$\psi_3(t) = \frac{1}{\sqrt{p}} \sum_{j=0}^{p-1} \widehat{|x+jp\rangle}|y\rangle = \frac{1}{\sqrt{p}} \sum_{j=0}^{p-1} e^{2\pi i(x+jp)t/n}\widehat{|0\rangle}|y\rangle\ .$$

Since the equidistribution on all values is a function of period one, the Fourier transform of it is the singleton at 0. By the inversion formula the Fourier transform of $|0\rangle = |0\rangle(t)$ is the equidistribution $\sum |k\rangle(t)$. So actually we have the state

$$\psi_3(t) = \frac{1}{\sqrt{p}} \sum_{j=0}^{p-1} e^{2\pi i(x+jp)t/n} \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} |k\rangle|y\rangle \, .$$

Now, if t is a multiple of n/p, we have

$$\frac{1}{\sqrt{p}} \sum_{j=0}^{p-1} e^{2\pi i(x+jp)t/n} = e^{2\pi i x t/n} \cdot \sqrt{p}$$

while for other t the sum over the roots of unity $\exp(2\pi i(x+jp)t/n)$ cancels to zero. So this state has support on the set

$$p^{\perp} = \{0, \frac{n}{p}, 2\frac{n}{p}, \ldots, (p-1)\frac{n}{p}\} \, ,$$

and is (up to complex phases) equidistributed. The equidistribution allows us to bound the number of measurements needed to infer the period p. The probability of measuring two adjacent points in the set $p^{\perp}$ depends on the length p of the period but not on the original length n of the register.

16.4.2
Generalization of the Period-Finding Algorithm

There are several generalizations of the period-finding problem. The two most important are real periods, almost-periods, and hidden subgroups, of which we will explain the latter only. We briefly sketch the other generalizations. Periods of functions $f: \mathbb{R} \to S$ can be computed by approximation, that is using the period-finding algorithm on the values of f at $1/N, 2/N, 3/N, \ldots$ for sufficiently large N. The algorithm used is a modification of the integer period-finding algorithm, which employs the same method. The complexity of this algorithm depends on the choice of N and the smallest period of f. An application of this technique can be found in [18], which gives an efficient algorithm to compute the regulator and the class number of the quadratic number field (the complexity of this task is known to be at least as hard as integer factorization). Such algorithms can be generalized to functions $f: \mathbb{R}^n \to S$ having a period lattice, that is a discrete subgroup $L \subset \mathbb{R}^n$ such that $f(x+\lambda) = f(x)$ for any $x \in \mathbb{R}^n$ and $\lambda \in L$, which are approximated on a lattice of the form $(1/N)\mathbb{Z}^n$ for sufficiently large N. Applications occur again in algebraic number theory, see [19], for example. We will not consider real functions here, since the effort to prove the probability bounds is rather lengthy. However, the lattice-finding problem is closely related to the natural generalization of the period problem for finite groups, which is in general called the *hidden subgroup problem* (HSP).

Definition 16.1 *Given a group G and a function $f: G \to S$ into an arbitrary set S, find the subgroup $H \leq G$ such that f factorizes on H: $f(g+h) = f(g)$ for all $g \in G$ and $h \in H$, and $f(g) \neq f(g')$ whenever $g + H \neq g' + H$.*

This problem has been studied for many types of groups G. Again the property of the Fourier transform of mapping functions with periods to functions with periodic supports can be used to find H in the case of abelian groups G. This is done by extending the concept of Fourier transform to arbitrary finite abelian Groups. We need the concept of characters to introduce the Fourier transform.

Definition 16.2 *A character of a finite abelian group G is a homomorphism $\chi : G \to \mathbb{C}^\times$.*

As in the cyclic case, the Fourier transform of a function $f : G \to \mathbb{C}$

$$\chi \mapsto \sum_{g\in G} f(g)\bar{\chi}(g)$$

is defined on characters. By identifying the elements of G with these characters we retrieve the usual notation of Fourier transforms defined on G itself. We briefly show how this identification is established: By the fundamental theorem for abelian groups, any finite abelian group is of the form

$$G \cong \bigoplus_{j=1}^{k} \mathbb{Z}/n_j\mathbb{Z}$$

so any character of G is of the form

$$\chi(g) = \prod_{j=1}^{k} \chi_j(g_j)$$

where χ_j is a character on $\mathbb{Z}/n_j\mathbb{Z}$. The set of homomorphisms $\chi : \ G \ \to \mathbb{C}^\times$ is itself a group by multiplication, denoted by $\hat{G}$, and the above product representation shows $G \cong \hat{G}$, that is there is a bijection between elements of G and characters of G. Characters of the cyclic factors of G are always of the form $\chi^{(j)}(a) = \exp(2\pi i(ab/n_j))$, the bijection is then

$$\mathbb{Z}/n_j\mathbb{Z} \ni b \leftrightarrow \chi^{(j)} = \left[a \mapsto e^{2\pi iab/n_j}\right] \in \widehat{\mathbb{Z}/n\mathbb{Z}} \,.$$

Characters of arbitrary abelian groups are (multiplicative) linear combinations of these cyclic characters. Since each $\mathbb{Z}/n\mathbb{Z}$ also carries a multiplicative structure, we can define a multiplication in the additive group G artificially, by setting

$$g * h = (g_1, \ldots, g_k) * (h_1, \ldots, h_k) := (g_1h_1, \ldots, g_kh_k) \,,$$

obtaining the identification of $a \in G$ with the character $\chi_a : g \mapsto \chi(a * g)$ for any character χ which is the product of generating characters of the cyclic factors. Thus we can define the Fourier transform taken on elements of G instead of characters:

Definition 16.3 *The (normalized) Fourier transform on G by χ is defined as the transformation*

$$\hat{f}(a) = \frac{1}{\sqrt{\#G}} \sum_{g\in G} f(g)\bar{\chi}(a * g)$$

for any function $f : G \to \mathbb{C}$.

Now we have to show that the Fourier transform on G provides the needed properties. They follow from the known *orthogonality relation* for χ:

$$\sum_{g\in G}\chi(g) = \begin{cases} \#G & \text{if} \quad \chi = 1 \quad \text{is the trivial character} \\ 0 & \text{if} \quad \chi \neq 1\,. \end{cases}$$

The inversion formula for the general Fourier transform is

$$f(g) = \frac{1}{\sqrt{\#G}}\sum_{a\in G}\hat{f}(a)\chi(a * g)$$

which is proven by

$$\begin{aligned}\frac{1}{\sqrt{\#G}}\sum_{a\in G}\hat{f}(a)\chi(a * g) &= \frac{1}{\#G}\sum_{a\in G}\sum_{b\in G}f(b)\bar{\chi}(b * a)\chi(a * g)\\ &= \frac{1}{\#G}\sum_{b\in G}f(b)\sum_{a\in G}\chi(a * (g - b))\,.\end{aligned}$$

By the orthogonality relation the last sum selects the value $b = g$ and removes the normalization factor, and the value of the expression is $f(g)$ as asserted. So the Fourier transformation on G is still an automorphism of functions from G to $\mathbb{C}$ and by the normalization factor it is also a unitary transformation. It maps functions with period subgroup $H \leq G$ to functions having support in some group $H^{\perp}$, which is now more complex than the set $n/p\mathbb{Z}/n\mathbb{Z}$ in the cyclic case. We use the equation

$$\begin{aligned}\hat{f}(a) &= \frac{1}{\sqrt{\#G}}\sum_{g\in G}f(g)\bar{\chi}(g * a) = \frac{1}{\sqrt{\#G}}\sum_{g\in G}f(g + h)\bar{\chi}(g * a)\\ &= \frac{1}{\sqrt{\#G}}\sum_{g\in G}f(h)\bar{\chi}((g - h) * a) = \hat{f}(a)\cdot\chi(h * a)\end{aligned}$$

which is true for all $a \in G$ if and only if $\hat{f}(a) = 0$ or $\chi(h * a) = 1$ for all $h \in H$, that is, if $\hat{f}$ vanishes outside the *dual group*

$$H^{\perp} = \{a \in G : \forall h \in H : \chi(h * a) = 1\}$$

of H. We have proved:

Theorem 16.1 *The Fourier operator* $\hat{}$ *gives an isomorphism between these* $\mathbb{C}$*-spaces:*
- *The set of functions f with $f(g + h) = f(g)$ for all $g \in G$, $h \in H$.*
- *The set of functions f which vanish outside $H^{\perp}$.*

Finding the subgroup H therefore amounts to two tasks:

- Computation of $H^{\perp}$ by the period-finding algorithm (generalized to $k+1$ registers). This is again done by measuring the $(k+1)$-th register and applying the Fourier transform on G to the first k registers, and measurement of enough elements from $H^{\perp}$ to infer generators or coefficients for $H^{\perp}$.

- Computation of H from $H^{\perp}$. Note that we actually do not want to compute the set H, but its generators, or equivalently, coefficients d_j such that

$$H = \bigoplus_{j=1}^{k} \mathbb{Z}/d_j\mathbb{Z} .$$

  This is not difficult if the corresponding coefficients for $H^{\perp}$ are known. Important applications of the hidden subgroup problem, for example the class number problem, content themselves with the determinant of H, which is the product of the coefficients. There is an efficient quantum algorithm for group decomposition, that is an algorithm to compute the coefficients d_j given generators of H.

We note that both group decomposition and hidden subgroup computation still admit no efficient classical algorithms. We conclude by stating the complexity of the abelian hidden subgroup problem. Assume the following preconditions.

- Elements of G can be represented uniquely as a binary string, and it is possible to recognize a representation classically in a number of steps polynomial in the representation length.
- It is possible to compute (classically) the representation of the sum and the inverse of elements using a number of steps polynomial in the length of their representation.
- Both group operations and the evaluation of f can be implemented in a quantum circuit.

Then we have:

Theorem 16.2 *There is a quantum algorithm with the following properties. Given a finite abelian group G (represented by generators of cyclic factors of prime power) and $f : G \to S$ satisfying Definition 16.1, and some error probability ε, it computes a set of elements of G along with coefficients (or equivalently generator relations) for the subgroup H' generated by those elements, such that*

- *the probability that $H' = H$ is the period subgroup of f is $\geq 1 - \varepsilon$;*
- *the number of measurements performed is bounded by a polynomial in $\log(\varepsilon^{-1})$ and the input length for the group G.*

The proof uses the crucial fact that the Fourier transform of a H-period function is not only supported by $H^{\perp}$, but that function values are equidistributed among this set (up to complex phases). The Fourier theory of nonabelian groups is far more complicated, and we do not have the correspondence of group elements to characters implicitly used in the definition of the operation $*$. Currently there is no general quantum algorithm known to compute H in the case of nonabelian groups. There is some work on special group types, like dihedral groups or solvable groups, but these still cover a very small portion of all nonabelian groups.

16.5 Quantum Approach to Graph Isomorphism

16.5.1 The Hidden-Subgroup Problem and Graph Isomorphism

We observe that the graph-isomorphism problem can be solved with the help of the hidden-subgroup problem. Let $G = (V, E)$ be a graph with vertex set $V = \{1, \dots n\}$ and consider the symmetric group S_n of permutations over n elements. For a permutation $\pi \in S_n$, $\pi(G)$ is the graph resulting from permuting the labels of the vertices in G according to π. The set of automorphisms of G, $\mathrm{Aut}(G) = \{\pi \in S_n | \pi(G) = G\}$ is clearly a subgroup of S_n. Consider the function f_G acting on S_n defined as $f_G(\pi) = \pi(G)$. Observe that for every $\sigma \in S_n$ and $\pi \in \mathrm{Aut}(G)$,

$$f_G(\sigma \cdot \pi) = \sigma \cdot \pi(G) = \sigma(G) = f_G(\sigma) \, .$$

Moreover, f_G has different values for the different cosets of $\mathrm{Aut}(G)$ in S_n:

$$\sigma_1 \mathrm{Aut}(G) \neq \sigma_2 \mathrm{Aut}(G) \Rightarrow \sigma_1(G) \neq \sigma_2(G) \Rightarrow f_G(\sigma_1) \neq f_G(\sigma_2) \, .$$

In other words, $\mathrm{Aut}(G)$ is the hidden subgroup in S_n defined by f_G. With a generating set for $\mathrm{Aut}(G)$ it is possible to efficiently compute the order of the subgroup, $|\mathrm{Aut}(G)|$. With this, one can decide the graph isomorphism problem in the following way. Let G_1 and G_2 be the input graphs and consider the graph $G_1 \cup G_2$ defined by the vertices and edges of both G_1 and G_2. It is not hard to see that if G_1 and G_2 are not isomorphic then $|\mathrm{Aut}(G_1 \cup G_2)| = |\mathrm{Aut}(G_1)| \cdot |\mathrm{Aut}(G_2)|$. On the other hand, if the graphs are isomorphic then $|\mathrm{Aut}(G_1 \cup G_2)| = 2|\mathrm{Aut}(G_1)| \cdot |\mathrm{Aut}(G_2)|$ (in this case we have to count the automorphisms interchanging the vertices of G_1 and G_2).

Because of this observation, efficient algorithms for HSP would imply the existence of efficient algorithms for GI. But the symmetric group S_n needed here is nonabelian and therefore the methods explained in the previous section cannot be applied. There have been several attempts to extend the algorithms for HSP from abelian to nonabelian groups [20,21] but the solved cases are not sufficient for solving GI. Hallgren, Russel and Ta-Shma show in [22] how to solve the HSP efficiently in the cases where the hidden subgroup is normal. Observe that this extends the results presented in the previous section since every subgroup of an abelian group is normal. Bacon, Childs and van Dam [23] have proposed a new approach giving efficient quantum algorithms for various semi-direct product groups. Kuperberg [24] developed a sieve algorithm for the hidden-subgroup problem in the dihedral group D_n with running time $2^{O(\sqrt{n})}$. Based on this result a subexponential time algorithm for solving HSP on direct product groups was presented in [25].

Recently, some negative results pointing to the impossibility of obtaining efficient quantum algorithms for the HSP have been published. In [26] it is shown that strong Fourier sampling is insufficient to efficiently resolve the HSP on certain nonabelian groups and that multiregister Fourier sampling over $\Omega(\log |G|)$ registers is required to distinguish subgroups of certain groups, including the symmetric group. A good overview of these results can be seen in [27].

16.5.2 The Quantum Query Model and Graph Isomorphism

The difficulty of obtaining nontrivial upper or lower bounds for the graph isomorphism problem on classical or quantum computers motivates the study of more restricted models, in which it is possible to establish differences between both computing paradigms.

We consider here the quantum query model, a basic restricted model of quantum computation. In the query model, the input $x_1, \dots, x_N$ is contained in a black box or oracle and can be accessed by queries. In a query we give a position i as input to the black box and it outputs x_i. The goal is to compute a boolean function $f\colon \{0,1\}^N \to \{0,1\}$ on the input bits $x = (x_1, \dots, x_N)$ minimizing the number of queries. The classical version of this model is known as a decision tree. We can consider the query complexity of a concrete boolean function in trying to show that the quantum model presents advantages over the classical model.

The quantum query model was explicitly introduced by Beals *et al.* [28]. Unlike the classical case, the power of quantum parallelism can be used in order to perform queries in superposition. The state of the computation is represented by $|i, b, z\rangle$, where i is the query register, b is the answer register, and z is the working register. A quantum computation with T queries is a sequence of unitary transformations

$$U_0 \to O_x \to U_1 \to O_x \to \dots \to U_{T-1} \to O_x \to U_T \,,$$

where each U_j is a unitary transformation that does not depend on the input x, and O_x are query (oracle) transformations. The oracle transformation O_x can be defined as $O_x : |i, b, z\rangle \to |i, b \oplus x_i, z\rangle$. The computation consists of the following three steps.

1. Go into the initial state $|0\rangle$.

2. Apply the transformation $U_T O_x \cdots O_x U_0$.

3. Measure the final state.

The result of the computation is the rightmost bit of the state obtained by the measurement. The quantum computation determines f with bounded error, if for every x, the probability that the result of the computation equals $f(x_1, \dots, x_N)$ is at least $1 - \varepsilon$, for some fixed $\varepsilon < 1/2$. In the query model of computation each query counts as one computation step but all other computation steps are free.

We consider upper and lower bounds for the number of queries needed to compute a boolean function stored in the black-box. If the black-box contains N positions, then trivially N queries are sufficient. But in some cases less quantum queries are needed. Grover [29] showed that in order to compute the OR function of N inputs $(x_1, \dots, x_N)$, $O(\sqrt{N})$ quantum queries are sufficient. This supposes a quadratic speed-up over the number of classical queries for the same problem.

The idea that with quantum queries we could search more efficiently in an unordered search space than with classical procedures initially gave some hope for efficient quantum solution of NP problems. For example, we can consider that in a problem like graph isomorphism, each oracle position x_i encodes a 0 or a 1 depending on whether the i-th bijection is an isomorphism between two given graphs. Computing the OR of these bits is equivalent to solve the isomorphism problem. Since for graphs with n nodes there are $n!$ possible bijections, a naive application of Grover's method would compute the problem with $O(\sqrt{n!})$ queries, which is still a very large number of steps. Some other methods have been proposed in trying to improve the efficiency of the search. One of these methods is the quantum walk.

16.5.3 Quantum Walks and the Fix-Automorphism Problem

Quantum walks are the quantum counterpart of Markov chains and random walks. We present here some facts on quantum walks and their connection to isomorphism problems. A discrete quantum walk is a way of formulating local quantum dynamics on a graph. The walk takes discrete steps between neighboring vertices and is a sequence of unitary transformations. We present a recent scheme for quantum search, based on any ergodic Markov chain, given by Magniez *et al.* [30]. We then use this tool for the development of a quantum algorithm for a special isomorphism problem.

Aharonov *et al.* [31] introduced quantum walks on graphs. They showed how fast quantum walks spread and proved lower bounds on the possible speedup by quantum walks for general graphs. Ambainis [32] constructed a fundamental quantum walk algorithm for the element distinctness problem. This was the first quantum walk algorithm that went beyond the capability of Grover search. Magniez *et al.* [33] have used Ambainis algorithm for finding a triangle in a graph. Szegedy [34] generalized the element distinctness algorithm of Ambainis to an arbitrary graph by using Markov chains. He showed that, for a class of Markov chains, quantum walk algorithms are quadratically faster than the corresponding classical algorithms. Buhrman and Špalek [35] constructed a quantum algorithm for matrix multiplication and its verification. Recently, Magniez *et al.* [30] developed a new scheme for quantum search based on any ergodic Markov chain. Their work generalizes previous results by Ambainis [32] and Szegedy [34]. They extend the class of possible Markov chains and improve the quantum complexity. Dörn and Thierauf [36, 37] presented the first application of this new quantum random walk technique for testing the associativity of a multiplication table.

Let $P = (p_{xy})$ be the transition matrix of an ergodic symmetric Markov chain on the state space X. Let $M \subseteq X$ be a set of marked states. Assume that the search algorithms use a data structure D that associates some data $D(x)$ with every state $x \in X$. From $D(x)$, we would like to determine if $x \in M$. When operating on D, we consider the following three types of costs:

- *Setup cost s*. The worst-case cost to compute $D(x)$, for $x \in X$.
- *Update cost u*. The worst-case cost for transition from x to y, and update $D(x)$ to $D(y)$.
- *Checking cost c*. The worst-case cost for checking if $x \in M$ by using $D(x)$.

Theorem 16.3 ([30]) *Let $\delta > 0$ be the eigenvalue gap of an ergodic Markov chain P and let $|M|/|X| \geq \varepsilon$. Then there is a quantum algorithm that determines if M is empty or finds an element of M with cost*

$$s + \frac{1}{\sqrt{\varepsilon}}\left(\frac{1}{\sqrt{\delta}}u + c\right) .$$

In the most practical applications (see [32, 33]) the quantum walk takes place on the Johnson graph $J(n, r)$, which is defined as follows: the vertices are subsets of $\{1, \ldots, n\}$ of size r and two vertices are connected iff they differ in exactly one number. It is well known that the spectral gap δ of $J(n, r)$ is $\Theta(1/r)$ for $1 \leq r \leq n/2$.

Now we consider the *fix-automorphism* problem as an application of the quantum walk search procedure. We have given a graph $G = (V, E)$ with n vertices represented as adjacency matrix and an integer $k < n$. One has to decide whether G has an automorphism which moves, at most, k vertices of G.

Theorem 16.4 *The quantum query complexity of the fix-automorphism problem is $O(n^{2k/(k+1)})$.*

Proof We apply the quantum walk search scheme of Theorem 16.3. To do so, we construct a Markov chain and a database for checking if a vertex of the chain is marked.

Let $G = (V, E)$ be the input graph with n vertices represented as adjacency matrix. Let U be a subset of vertices of G of size r. We will determine r later. Our quantum walk takes place on the Johnson graphs $J(n, r)$. The database of the quantum walk is the induced subgraph on the set of vertices U, denoted by $G[U]$, has U as its vertex-set, and it contains every edge of G whose endpoints are in U. The marked vertices of $J(n, r)$ correspond to subsets $U \subset V$, such that $G[U]$ contains an automorphism which moves, at most, k vertices of G. In every step of our walk we exchange one vertex of U.

Now we determine the quantum query setup, update and checking cost. The setup cost to determine $G[U]$ is $O(r^2)$ and the update cost is $O(r)$. Checking if $G[U]$ contains such a fixed automorphism needs no queries, since we require only the database for checking if the vertex U is marked.

If there is at least one automorphism which moves at most k vertices, then there are at least $\binom{n-k}{r-k}$ marked vertices of the Johnson graph. Therefore, we have

$$\varepsilon \geq \frac{|M|}{|X|} \geq \frac{\binom{n-k}{r-k}}{\binom{n}{r}} \geq \Omega\left(\left(\frac{r}{n}\right)^k\right) .$$

Then the quantum query complexity of the fix-automorphism problem is

$$O\left(r^2 + \left(\frac{n}{r}\right)^{k/2} r^{1.5}\right) ,$$

which is minimized for $r = n^{k/(k+1)}$. □

16.6 Reductions of Integer Factorization and Graph Isomorphism to Ring Isomorphism

We review in this section a result from Kayal and Saxena [38] showing that GI and IF are instances of a more general question: the ring isomorphism problem. Ring isomorphism has received attention in recent years in connection with the efficient primality test algorithm from [39]. Other applications of this problem can be seen in [40] and [41].

Definition 16.4 *A finite ring with identity element 1 is a triple $(R, +, \cdot)$, where R is a finite set such that $(R, +)$ is a commutative group with identity element 0 and $(R, \cdot)$ is a semigroup with identity element 1, such that multiplication distributes over addition. The characteristic of a ring R is defined to be the smallest number of times one must add the ring's multiplicative identity element 1 to itself to get the additive identity element 0. Let I be an ideal of R, the factor ring is the ring $R/I = \{a + I : a \in R\}$ together with the operations $(a + I) + (b + I) = (a + b) + I$ and $(a + I)(b + I) = ab + I$.*

The polynomial ring $R[X]$ is the ring of all polynomials in a variable X with coefficients in the ring R.

Let n be the characteristic of the ring. The complexity of the problems involving finite rings depends on the representation used to specify the ring. We will use the following representation models of a ring.

- *Table representation.* A ring R is given as a list of all the elements of the ring and their addition and multiplication tables.
- *Basis representation.* A ring R is given by m *basis elements* $b_1, \dots, b_m$ and the additive group can be expressed as

  $$(R, +) = \bigoplus_{i+1}^{m} \mathbb{Z}_{n_i} b_i ,$$

  with $n_i | n$ for each i. The multiplication in R is given by specifying the product of each pair of basis elements as an integer linear combination of the basis elements: $b_i \cdot b_j = \sum_{k=1}^{m} a_{ij,k} b_k$ for $1 \leq i, j \leq m$ with $a_{ij,k} \in \mathbb{Z}_n$.

- *Polynomial representation.* A ring R is given by

  $$R = \mathbb{Z}_n[Y_1, \ldots, Y_m]/(f_1(Y_1, \ldots, Y_m), \ldots, f_k(Y_1, \ldots, Y_m)) \,,$$

  where $Y_1, \ldots, Y_m$ are basis elements and $(f_1(Y_1, \ldots, Y_m), \ldots, f_k(Y_1, \ldots, Y_m))$ is an ideal generated by the polynomials $f_1, \ldots, f_k$.

The table representation has size $O(|R|^2)$, which is a highly redundant representation. The size of the basis representation is $O(m^3)$, where m is the number of basis elements. This is in general exponentially smaller than the size of the Ring $|R| = \Pi_{i=1}^m n_i$. Often the polynomial representation is exponentially more succinct than the basis representation. For example $\mathbb{Z}_2[Y_1, \ldots, Y_m]/(Y_1^2, \ldots, Y_m^2)$ has 2^m basis elements and so the basis representation would require $\Omega(2^{3m})$ space.

Since the polynomial representation is of smaller size, for clarity of exposition, we will use it here to express the rings. However, for complexity issues, this representation is too succinct and we will consider the rings given as input to the problems given in basis representation. For a polynomial representation, say $R = \mathbb{Z}_n[Y_1, \ldots, Y_t]/\mathcal{I}$ an automorphism or isomorphism ϕ will be specified by a set of t polynomials $p_1, \ldots, p_t$ with $\phi(Y_i) = p_i(Y_1, \ldots, Y_t)$.

Definition 16.5 *An automorphism of ring R is a bijective map $\phi : R \mapsto R$ such that for all $x, y \in R$, $\phi(x + y) = \phi(x) + \phi(y)$ and $\phi(x \cdot y) = \phi(x) \cdot \phi(y)$. An isomorphism between two rings R_1, R_2 is a bijective map $\phi : R_1 \mapsto R_2$ such that for all $x, y \in R_1$, $\phi(x + y) = \phi(x) + \phi(y)$ and $\phi(x \cdot y) = \phi(x) \cdot \phi(y)$.*

We define some ring automorphism and isomorphism problems. All rings are given in basis representation.

- The *ring automorphism* problem (RA) consists of being in given a ring R, deciding whether there is a non-trivial automorphism for R.
- The *finding ring automorphism* problem (FRA) consists in given a ring R, find a non-trivial automorphism of R.
- The *ring isomorphism* problem (RI) consists in given two rings R_1, R_2, decide whether there is an isomorphism between both rings.

16.6.1
Factoring Integers and Finding Ring Automorphisms

We discuss the complexity of finding automorphisms in a ring and present a result from Kayal and Saxena [38] showing that this problem is at least as hard as factoring integers. Let $\leq_m^P$ denote a polynomial time many~one reduction between problems.

Theorem 16.5 ([40]) IF $\leq_{\mathrm{m}}^{\mathrm{P}}$ FRA.
The quadratic and number field sieve methods can be easily viewed as trying to find a nonobvious automorphism in a ring. Both methods aim to find two numbers u and v in $\mathbb{Z}_n$ such that $u^2 = v^2$ and $u \neq \pm v$ in $\mathbb{Z}_n$, where n is an odd square-free composite number to be factored. We will encode this in a ring such that finding its ring automorphisms gives us u and v. We consider the ring $R = \mathbb{Z}_n[Y]/(Y^2 - 1)$, which has an obvious

nontrivial automorphism mapping Y onto $-Y$. The problem is to find another nontrivial automorphism $\phi(R) \neq \pm Y$. Agrawal and Saxena give the following proof.

Proof Let $\phi(Y) = aY + b$, by definition of the factor ring R we have $\phi(Y^2 - 1) = 0$. Observe that $\phi(a + b) = \phi(a) + \phi(b)$, $\phi(ab) = \phi(a)\phi(b)$, and that $Y^2 - 1 \equiv 0$. Thus we have

$$\begin{aligned}\phi(Y^2 - 1) &= (aY + b)^2 - 1 = a^2Y^2 + 2abY + b^2 - 1\\ &= a^2(Y^2 - 1 + 1) + 2abY + b^2 - 1 = a^2 + b^2 - 1 + 2abY = 0\ .\end{aligned}$$

This gives $ab = 0$ and $a^2 + b^2 = 1 \in \mathbb{Z}_n$. Notice that a and n are relatively prime, that is $(a, n) = 1$, since otherwise

$$\phi\left(\frac{n}{(a,n)}Y\right) = \frac{n}{(a,n)}(aY + b) = \frac{a}{(a,n)}nY + \frac{n}{(a,n)}b = \phi\left(\frac{n}{(a,n)}b\right)\ ,$$

because $(a/(a, n))nY \equiv 0 \mod n$. Therefore, $b = 0$ and $a^2 = 1$. By assumption, $a \neq \pm 1$ and so $u = a$ and $v = 1$. Conversely, given u and v with $u^2 = v^2$, $u \neq \pm v$ in $\mathbb{Z}_n$ we get $\phi(Y) = (u/v)Y$ as an automorphism of R. □

As shown in [38], factoring integers can be reduced to a number of questions about automorphisms and isomorphisms of rings, that is, counting the number of automorphisms of ring $\mathbb{Z}_n[Y]/(Y^2)$ or finding the isomorphisms between rings $\mathbb{Z}_n[Y]/(Y^2 - a^2)$ and $\mathbb{Z}_n[Y]/(Y^2 - 1)$ for a randomly chosen $a \in \mathbb{Z}_n$. But for RA a polynomial time algorithm is known [38]. That means, some automorphisms are easy to compute while others are not.

16.6.2 Graph Isomorphism and Ring Isomorphism

An interesting fact is that there is also a connection between the graph isomorphism and the ring isomorphism problems.

Theorem 16.6 ([38]) GI $\leq_{\mathrm{m}}^{\mathrm{P}}$ RI.

Proof We present the reduction from Agrawal and Saxena [40]. Let $G = (V, E)$ be a simple graph on n vertices. Then define polynomial p_G as

$$p_G(x_1, \ldots, x_n) = \sum_{(i,j)\in E} x_i \cdot x_j\ ,$$

and define ideal $\mathcal{I}_G$ as

$$\mathcal{I}_G(x_1, \ldots, x_n) = \left(p_G(x_1, \ldots, x_n), \{x_i^2\}, \{x_i x_j x_k\}_{1\leq i,j,k\leq n}\right)\ .$$

Observe that for a polynomial ring with ideal $\mathcal{I}_G$ in basis representation the number of basis elements is bounded by $O(n^2)$ since any combination of three variables are zero. In detail, Agrawal and Saxena proved the following. Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be simple graphs over n vertices and let F_q be a field of odd characteristic. Then G_1 is isomorphic to G_2 iff

- either both graphs contain a clique of size m (K_m) and $n - m$ isolated vertices (D_{n-m}), each (in this case isomorphism testing is trivial),
- or the rings $R_1 = F_q[Y_1, \ldots, Y_n]/\mathcal{I}_{G_1}(Y_1, \ldots, Y_n)$ and $R_2 = F_q[Z_1, \ldots, Z_n]/\mathcal{I}_{G_2}(Z_1, \ldots, Z_n)$ are isomorphic.

R_1 and R_2 are polynomial rings with polynomials of degree at most two. If π is an isomorphism mapping G_1 onto G_2 then an isomorphism between both rings can be found as $\phi : R_1 \mapsto R_2$ with $\phi(Y_i) = Z_{\phi(i)}$, since $\phi(p_{G_1}(Y_1, \ldots, Y_n)) = p_{G_2}(Z_1, \ldots, Z_n)$.

We show now, that there is no further isomorphism. Suppose that $G_1 \cong G_2$ and that G_2 is not of the form $K_m \cup D_{n-m}$. Let $\phi : R_1 \mapsto R_2$ be an isomorphism with

$$\phi(Y_i) = \alpha_i + \sum_{1\le j\le n} \beta_{i,j} Z_j + \sum_{1\le j<k\le n} \gamma_{i,j,k} Z_j Z_k \ .$$

We will show now, which values for $\alpha_i, \beta_{i,j}$ and $\gamma_{i,j,k}$ in $\phi(Y_i)$, may occur. For any isomorphism ϕ on rings it must hold that $\phi(0) = 0$. In R_1, $Y_i^2 = 0$ and in R_2, $Z_i^2 = 0$ for any value of i it follows that

$$\phi(Y_i^2) = (\phi(Y_i))^2 = \alpha_i^2 + (\text{higher degree terms}) = 0 \ .$$

Thus $\alpha_i = 0$. Again looking at the same equation:

$$\phi(Y_i^2) = (\phi(Y_i))^2 = 2 \sum_{1\le j<k\le n} \beta_{i,j}\beta_{j,k} Z_j Z_k = 0 \ .$$

The other terms disappeared since $\alpha_i = 0$ or they become two degrees higher.

If more than one $\beta_{i,j}$ is nonzero, then we must have $\sum_{j,k\in J, j<k} \beta_{i,j}\beta_{i,k} Z_j Z_k$ divisible by $p_{G_2}(Z_1, \ldots, Z_n)$ with J the set of nonzero indices. Since p_{G_2} is also a homogeneous polynomial of degree two, it must be a constant multiple of the above expression implying that $G_2 = K_{|J|} \cup D_{n-|J|}$. This is not possible by assumption. Therefore, at most, one $\beta_{i,j}$ is nonzero.

If all $\beta_{i,j}$ are zero, then $\phi(Y_i, Y_l) = 0$ for all i, l which is not possible. Hence, exactly one $\beta_{i,j}$ is nonzero. Define $\pi(i) = j$ where j is the index with $\beta_{i,j}$ nonzero. We prove now, that π is not surjective. Suppose $\pi(i) = \pi(l)$ for $i \neq l$. Then $\phi(Y_i Y_l) = Z_{\pi(i)} Z_{\pi(l)} = 0$. This is not possible. Hence π is a permutation on $[1, n]$. Now consider $\phi(p_{G_1}(Y_1, \ldots, Y_n))$, then it follows that

$$0 = \phi(p_{G_1}(Y_1, \ldots, Y_n)) = \sum_{(i,j)\in E_1} \phi(Y_i)\phi(Y_j) = \sum_{(i,j)\in E_1} \beta_{i,\pi(i)}\beta_{j,\pi(j)} Z_{\pi(i)} Z_{\pi(j)}$$

The last expression must be divisible by p_{G_2}. This gives $\beta_{i,\pi(i)} = \beta_{j,\pi(j)}$ for all i, j and implies that the expression is a constant multiple of p_{G_2} or equivalently, that G_1 is isomorphic to G_2. □

Notice that the rings R_1 and R_2 constructed above have lots of automorphisms. For example, $Y_i \mapsto Y_i + Y_1 Y_2$ is a nontrivial automorphism of R_1. Thus, automorphisms of G_1 do not directly correspond to automorphisms of R_1.

References

1 Schwarz, W. (**1969**) *Einführung in die Methoden und Ergebnisse der Primzahltheorie*, Hochschultaschenbücher-Verlag.

2 Cohen, H. (**1993**) *A Course in Computational Algebraic Number Theory*, GTM 138, Springer Verlag.

3 Babai, L. (**1985**) *Trading group theory for randomness*, Proceedings of the 17th ACM Symposium on Theory of Computing (STOC), 424–429.

4 Hopcroft, J.E. and Tarjan, R.E. (**2005**) Dividing a graph into triconnected components, *SIAM Journal on Computing*, **2**(3), 136–158.

5 Zemlyachenko, V.N. (**1970**) *Canonical numbering of trees (Russian)*, Proc. Seminar on Comb. Anal. at Moscow State Univ., Moscow.

6 Hopcroft, J.E. and Wong, J.K. (**1974**) *Linear time algorithm for isomorphism of planar graphs*, Proceedings of Symposium on Theory of Computing (STOC), 172–184.

7 Filotti, I.S. and Mayer, J.N. (**1980**) *A polynomial-time algorithm for determining the isomorphism of graphs of fixed genus*, Proceedings of Symposium Theory of Computing (STOC), 236–243.

8 Lipton, R.M. (**1980**) The beacon set approach to graph isomorphism, *SIAM Journal on Computing*, **9**.

9 Miller, G.L. (**1980**) Isomorphism testing for graphs of bounded genus, *Proceedings of Symposium on Theory of Computing (STOC)*, 225–235.

10 Babai, L. (**1979**) *Monte-Carlo algorithms in graph isomorphism testing*, Preprint, University Toronto.

11 Furst, M., Hopcroft, J. and Luks, E. (**1980**) *Polynomial-time algorithms for permutation groups*, Proceedings of Symposium on Foundations of Computer Science (FOCS), 36–41.

12 Luks, E. (**1982**) Isomorphism of graphs of bounded valence can be tested in polynomial time, *Journal of Computer and System Science* **25**, 42–65.

13 Zemlyachenko, V., Kornienko, N. and Tyshkevich, R. (**1982**) Graph isomorphism problem, *The Theory of Computation I, Notes Sci. Sem. LOMI* **118**.

14 Babai, L. and Luks, E. (**1983**) *Canonical labeling of graphs*, Proceedings of the 15th ACM Symposium on Theory of Computing (STOC), 171–183.

15 Babai, L., Erdös, P. and Selkow, S.M. (**1980**) Random graph isomorphism, *SIAM Journal on Computing* **9**(3), 628–635.

16 Köbler, J., Schöning, U. and Torán, J. (**1993**) *The Graph Isomorphism Problem – Its Structural Complexity*, Birkhäuser.

17 Shor, P.W. (**1997**) Polynomial-time algorithms for factorization and discrete logarithms on a quantum computer, *SIAM Journal on Computing* **26**, 1484–1509.

18 Hallgren, S. (**2002**) *Polynomial-Time Quantum Algorithms for Pell's Equation and the Principal Ideal Problem*, Proceedings of Symposium on Theory of Computing, Montreal Canada.

19 Haase, D. and Maier, M. (**2006**) Quantum Algorithms in Number Fields, *Progress of Physics* **54**(8–10), 866–881.

20 Grigni, M., Schulman, L. and Vazirani, V. (**2001**) *Quantum mechanical algorithms for the nonabelian hidden subgroup problem*, Proceedings of Symposium on Theory of Computing (STOC), 68–74.

21 Ivanyos, G., Magnier, F. and Santha, M. (**2001**) *Efficient quantum algorithms for some instances of the non-abelian hidden subgroup problem*, Proceedings of Symposium on Parallel Algorithms and Architechtures, 263–270.

22 Hallgren, S., Russel, A. and Ta-shma, A. (**2000**) *Normal subgroup reconstruction and quantum computing using the group representation*, Proceedings of Symposium on Theory of Computing (STOC), 627–635.

23 Bacon, D., Childs, A. and van Dam, W. (**2007**) *From optimal measurement to efficient quantum algorithms for the hidden subgroup problem over semidirect product groups*, Proceedings of the 46th IEEE Symposium on Foundations of Computer Science (FOCS), 469–478.

24 Kuperberg, G. (**2005**) A subexponential time quantum algorithm for the dihedral hidden subgroup problem, *SIAM Journal on Computing* **35**(1), 170–188.

25 Alagic, G., Moore, C. and Russell, A. (**2007**) *Quantum algorithms for Simon's problem over general groups*, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1217–1224.

26 Hallgren, S., Moore, C., Rötteler, M., Russell, A. and Sen, P. (**2006**) *Limitations of quantum coset states for graph isomorphism*, Proceedings of Symposium on Theory of Computing (STOC), 604–617.

27 Alagic, G. and Russell, A. (**2007**) Quantum computing and the hunt for the hidden symmetry, *Bulletin of the EATCS*.

28 Beals, R., Buhrman, H., Cleve, R., Mosca, M. and de Wolf, R. (**2001**) Quantum lower bounds by polynomials, *Journal of ACM* **48**, 778–797.

29 Grover, L. (**1996**) *A fast mechanical algorithm for database search*, Proceedings of Symposium on Theory of Computing (STOC), 212–219.

30 Magniez, F., Nayak, A., Roland, J. and Santha, M. (**2007**) *Search via quantum walk*, Proceedings of Symposium on Theory of Computing (STOC), 575–584.

31 Aharonov, A., Ambainis, A., Kempe, J. and Vazirani, U. (**2001**) *Quantum walks on graphs*, Proceedings of Symposium on Theory of Computing (STOC), 50–59.

32 Ambainis, A. (**2004**) *Quantum walk algorithm for element distinctness*, Proceedings of Symposium on Foundations of Computer Science (FOCS), 22–31.

33 Magniez, F., Santha, M. and Szegedy, M. (**2005**) *Quantum Algorithms for the triangle problem*, Proceedings of Symposium on Discrete Algorithms (SODA), 1109–1117.

34 Szegedy, M. (**2004**) *Quantum speed-up of Markov chain based algorithms*, Proceedings of Symposium on Foundations of Computer Science (FOCS), 32–41.

35 Buhrman, H. and Špalek, R. (**2006**) *Quantum verification of matrix products*, Proceedings of Symposium on Discrete Algorithms (SODA), 880–889.

36 Dörn, S. and Thierauf, T. (**2007**) *The quantum query complexity of algebraic properties*, Proceedings of Symposium on Fundamentals of Computation Theory (FCT), 250–260.

37 Dörn, S. and Thierauf, T. (**2008**) *The quantum complexity of group testing*, Proceedings of Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM), 506–518.

38 Kayal, N. and Saxena, N. (**2005**) *On the ring isomorphism and automorphism problems*, Proceedings of Conference on Computational Complexity (CCC), 2–12.

39 Agrawal, M., Kayal, N. and Saxena, N. (**2004**) PRIMES is in P, *Annals of Mathematics, Second Series*, **160**, 781–793.

40 Agrawal, M. and Saxena, N. (**2005**) *Automorphisms of finite rings and applications to complexity of problems,* Proceedings of Symposium on Theoretical Aspects of Computer Science (STACS), 1–17.

41 Agrawal, M. (**2007**) *Rings and integer lattices in computer science,* Lecture Notes, the 2007 Barbados Workshop on Computational Complexity. http://www.cs.mcgill.ca/ denis/barbados07/barbados2007.ps.

17
QuickSort from an Information Theoretic View

Beatrice List, Markus Maucher[1], Uwe Schöning, Rainer Schuler

17.1
Introduction

The QuickSort algorithm was invented by C.A.R. Hoare in the sixties [4]. Knuth [6, page 115] cites this paper as one of the most comprehensive accounts of a sorting method that has ever been published. Later, Hoare received the ACM Turing Award in 1980, and he was knighted for his achievements in Computer Science by Queen Elizabeth II in 2000.

QuickSort is a classical divide and conquer method: the input sequence is divided into two subsequences which are both sorted by applying the QuickSort algorithm recursively. Then these sorted sequences are concatenated together to form the desired sorted sequence. Unlike other divide and conquer algorithms, the input sequence is not necessarily split into two parts of equal sizes. Actually the sizes depend on the input itself. In each recursive step, a splitting element (the "pivot") is selected, which, in many implementations, is the first element of the sequence to be sorted. The sizes of the subsequences depend on the rank of the pivot element within the sequence to be sorted (which is not known beforehand).

Every sorting algorithm which is based on pairwise comparisons of elements (like QuickSort does) has to identify, from an information theoretic point of view, which of the $n!$ many input permutations is actually present (and using this information, the algorithm has to rearrange the elements physically to form a sorted sequence). Each comparison of two elements gives the algorithm one bit of information. Therefore, for the entire sorting process the algorithm needs, in the worst case, at least $\log_2(n!) = n \log_2 n - \Theta(n)$ many bits of information, or comparisons.[2]

It is known that QuickSort's average number of pairwise element comparisons (averaging over all potential input permutations) is $(2 \ln 2) \cdot n \log_2 n - \Theta(n)$, so it is quite close to the ideal case, the lower bound. On the other hand, there are worst-

[1] Corresponding author.

[2] Here, $\Theta(g(n))$ denotes some function f which, for some constants $c < d$ and almost every n, satisfies $cg(n) \le f(n) \le dg(n)$.

case inputs where QuickSort does up to $n(n-1)/2$ comparisons (ironically, the already sorted sequence has this property.) Realizing this bad worst-case behavior, Hoare had suggested the variant called Random QuickSort. In Random QuickSort (see Figure 17.1) the pivot element is selected uniformly at random among the elements of the sequence to be sorted. A very similar analysis to the one mentioned above shows that the expected number of comparisons, *for each input sequence*, is $(2 \ln 2) \cdot n \log_2 n - \Theta(n)$. Here the expectation is taken over all random choices done in the course of the algorithm.

```
input:   finite sequence $A = (a[1], a[2], \ldots a[n])$ of distinct elements
output:  finite sequence $B$ that contains all elements from $A$ in
         increasing order
method:  if $A$ contains at most 1 element
             return $A$
         else
             Choose a random element $x$ from $A$
             Split $A$ into two subsequences $A_1$ and $A_2$ such that
                a) $A_1$ contains all elements from $A$ smaller than $x$
                b) $A_2$ contains all elements from $A$ greater than $x$
             $B_1 \leftarrow$ QuickSort($A_1$)
             $B_2 \leftarrow$ QuickSort($A_2$)
             return $B_1 \circ (x) \circ B_2$        ($\circ$ denotes concatenation)
```

Figure 17.1 Pseudo code of the randomized QuickSort algorithm.

As previously mentioned, this analysis uses ideal random numbers, i.e. those which are independent and uniformly distributed. Technically, such random numbers are difficult to produce, and in practice, one uses pseudorandom number generators instead, which start with some given "seed" x_0, and iteratively (and deterministically) compute successive values $x_{i+1} = f(x_i)$ according to some function f such that the obtained sequence of values $x_1, x_2, \ldots$ "looks random" (i.e. it passes some statistical tests). If the seed is fixed in advance, then the entire algorithm becomes a deterministic algorithm, and actually the above assertion about the existence of worst-case inputs, with $n(n-1)/2$ many comparisons, is still valid.

From a theoretical point of view, one might consider the seed of the pseudorandom generator as truly random. But still, under this theoretical model, when using a linear congruential generator, like $x_{i+1} = (ax_i + b) \mod c$ as suggested by D.H. Lehmer [7], Karloff and Raghavan [5] (see also [12]) have shown (under mild assumptions about the choice of the parameters a, b, c) that the expected number of comparisons can be, in the worst case, up to dn^2, for some constant d. Here the expectation is taken over the random choice of the seed, and the worst case refers to the choice of the input.

In this paper we follow this line of research and consider a random number generator for Random QuickSort, which is not ideal. We measure the deficiency of the random number generator in terms of C.E. Shannon's entropy function

$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$ (see [11]). Depending on the Shannon entropy of the random number generator we show a continous transition between the "ideal" case of an $(n \log n)$-behavior and the "bad" case of (n^2)-behavior.

17.1.1 Recursion for the Expected Number of Comparisons

Let $T_\pi(n)$ be the expected number of comparisons done by randomized QuickSort when operating on an input array $(a[1], \ldots, a[n])$ whose elements are distinct and permuted according to $\pi \in S_n$, that is,

$$a[\pi(1)] < a[\pi(2)] < \cdots < a[\pi(n)]$$

where S_n is the set of all permutations on $\{1, \ldots, n\}$.

Let X be a random variable taking values between 1 and n (not necessarily under uniform distribution) which models the random number generator that is used to pick out a pivot element $a[X]$. We say an element has rank i within the ordering of the array if there are exactly $i - 1$ smaller elements in the array. Let p_i be the probability that the pivot element has rank i within the ordering of the array, that is, $p_i = \Pr(\pi(X) = i)$.

We obtain the following recursion for the expected worst-case complexity (i.e. number of comparisons) $T(n) = \max_{\pi \in S_n} T_\pi(n)$. We have $T(n) = 0$ for $n \le 1$; and for $n > 1$ we get

$$\begin{aligned} T(n) &= \max_{\pi \in S_n} T_\pi(n) \\ &= (n-1) + \max_{\pi \in S_n} \sum_{i=1}^{n} p_i (T_\pi(i-1) + T_\pi(n-i)) \\ &\le (n-1) + \sum_{i=1}^{n} p_i \left(\max_{\phi \in S_{i-1}} T_\phi(i-1) + \max_{\psi \in S_{n-i}} T_\psi(n-i) \right) \\ &= (n-1) + \sum_{i=1}^{n} p_i (T(i-1) + T(n-i)) \ . \end{aligned}$$

That is, there are $n-1$ comparisons with the selected pivot element and, depending on the rank i of the pivot element within the array, there are at most $T_\pi(i-1)$ and $T_\pi(n-i)$ additional comparisons. If the rank of the pivot element is not uniformly distributed among the numbers 1 to n, a worst-case input permutation can be constructed such that the middle ranks receive relatively low probability and the extreme ranks (close to 1 or close to n) have relatively high probability, resulting in a large expected number of comparisons.

We give upper and lower bounds on the expected number $T(n)$ of comparisons. Lower bounds are given with respect to a fixed worst-case input sequence (e.g. the already sorted list of elements). These bounds are tight up to a logarithmic factor.

We can show (see Theorem 17.1) that $T(n) \le g(n) n \log_2 n$ for any function $g(n)$ greater than $1/\left(\min_\pi \sum_{i=1}^{n} p_i H(i/n)\right)$, where H is Shannon's binary entropy function. Note that $\min_\pi \sum_{i=1}^{n} p_i H(i/n)$ is independent of the permutation of the elements, i.e. is identical for all distributions p and q such that $p_i = q_{\pi(i)}$ for all i and some permutation π.

The lower bound (see Theorem 17.2) is derived for distributions on the ranks of the input elements. Therefore, the lower bound $T(n) \geq cng(n)$ (Theorem 17.5) is with respect to any function $g(n)$ less than $1/\sum_{i=1}^{n} p_i H(i/(n+1))$, where p_i is the probability of selecting the element of rank i within the input a as a pivot element.

17.2 An Upper Bound

Let $(P_1, P_2, \ldots)$ denote a sequence of probability distributions where $P_n = (p_{n1}, \ldots, p_{nn})$ is a distribution on $\{1, \ldots, n\}$.

Theorem 17.1 *Let $(P_1, P_2, \ldots)$ be a sequence of probability distributions on the indexes of the pivot elements used by Randomized QuickSort. Then $T(n) \leq g(n) n \log_2 n$ for any monotone increasing function g with the property*

$$g(n) \geq \left(\min_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} \cdot H\left(\frac{i}{n}\right) \right)^{-1}$$

where $H(x) = -x \log_2 x - (1-x) \log_2(1-x)$ is Shannon's binary entropy function.

Proof By induction on n. Using the above recursion for $T(n)$ we obtain

$$\begin{aligned}
T(n) &= (n-1) + \max_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} (T_\pi(i-1) + T_\pi(n-i)) \\
&\leq n + \max_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} (g(i-1)(i-1)\log_2(i-1) + g(n-i)(n-i)\log_2(n-i)) \\
&\leq n + g(n) n \max_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} \left(\frac{i}{n} \log_2 i + \left(1 - \frac{i}{n}\right) \log_2(n-i) \right) \\
&= n + g(n) n \max_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} \left(\frac{i}{n} \log_2 \frac{i}{n} + \left(1 - \frac{i}{n}\right) \log_2 \left(1 - \frac{i}{n}\right) + \log_2 n \right) \\
&= n + g(n) n \log_2 n - g(n) n \min_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} H\left(\frac{i}{n}\right) .
\end{aligned}$$

To finish the induction proof, this last expression should be at most $g(n) n \log_2 n$. This holds if and only if $g(n) \geq \left(\min_{\pi \in S_n} \sum_{i=1}^{n} p_{n\pi^{-1}(i)} H\left(\frac{i}{n}\right) \right)^{-1}$ as claimed. □

Example 17.1 *In the standard case of a uniform distribution $p_{ni} = 1/n$ we obtain $g(n) \geq \left(1/n \sum_{i=1}^{n} H(i/n)\right)^{-1}$. Asymptotically, this is $\left(\int_0^1 H(x)\,\mathrm{d}x\right)^{-1} = 2 \ln 2 \approx 1.38$, which is the known constant factor of QuickSort's average running time.*

Example 17.2 *In the median-of-three version of QuickSort (see [6, 9]), three different elements are picked uniformly at random and the median of the three is used as the pivot element. In this case*

$$p_{ni} = \frac{6(i-1)(n-i)}{n(n-1)(n-2)} \; .$$

Here the constant factor of the $n \log n$-term can be asymptotically estimated by

$$\left(6 \int_0^1 x(1-x)H(x)\,\mathrm{d}x\right)^{-1} = \frac{12 \ln 2}{7} \approx 1.18 \; .$$

This matches the average running time given in [9].

17.3 A Lower Bound

In a similar fashion, we can derive a lower bound for the number of comparisons. For the proof of Theorem 17.2, we need the following technical lemma:

Lemma 17.1 *For integers $n \ge 1$ and i with $0 \le i \le n$,*

$$\frac{(i-1)^2}{n^2} + \frac{(n-i)^2}{n^2} + H\left(\frac{i}{n+1}\right) \ge 1 \; .$$

Proof We use the known inequalities

$$-\ln(1-x) \ge x \quad \text{resp.} \quad -\log_2(1-x) \ge \frac{x}{\ln 2} \; ,$$

that hold for $0 \le x \le 1$. So we get

$$\begin{aligned}
&\frac{(i-1)^2}{n^2} + \frac{(n-i)^2}{n^2} + H\left(\frac{i}{n+1}\right) \\
&= \frac{i^2 - 2i + 1 + n^2 - 2in + i^2}{n^2} \\
&\quad - \frac{i}{n+1} \log_2 \frac{i}{n+1} - \left(1 - \frac{i}{n+1}\right) \log_2 \left(1 - \frac{i}{n+1}\right) \\
&= \frac{2i^2 - 2i + 1 + n^2 - 2in}{n^2} \\
&\quad - \frac{i}{n+1} \log_2 \left(1 - \frac{n-i+1}{n+1}\right) - \frac{n-i+1}{n+1} \log_2 \left(1 - \frac{i}{n+1}\right) \\
&\ge \frac{2i^2 - 2i + 1 + n^2 - 2in}{n^2} + \left(\frac{i}{n+1}\,\frac{n-i+1}{n+1} + \frac{n-i+1}{n+1}\,\frac{i}{n+1}\right) \frac{1}{\ln 2} \\
&\ge \frac{2i^2 - 2i + 1 + n^2 - 2in + 2in - 2i^2 + 2i}{n^2} = \frac{n^2+1}{n^2} \ge 1 \; .
\end{aligned}$$

For the second last inequality, we use that $(n+1)^2 \ln 2 \le n^2$ for $n \ge 1$. □

The running time derived in the upper bound theorem was independent of the actual input permutation and depended only on the distributions on the indices that are used to pick a pivot element from the input. Our lower bound however cannot be that flexible. For every distribution on the indices of the input, there exists an input that will be divided into two subarrays of approximately equal sizes, with high probability. Therefore, the theorem for the lower bound is formulated with respect to distributions on the ranks of the input numbers. Similar to Theorem 17.1 we get:

Theorem 17.2 *Let $(P_1, P_2, \ldots)$ be a sequence of probability distributions on the ranks of the chosen pivot elements, where $P_n = (p_{n1}, \ldots, p_{nn})$ is used to choose a pivot element from sequences of length n and the element of rank i is chosen with probability p_{ni}.*

(i) *$T(n) \geq cg(n)n - n$ for some constants $c > 0$ and n_0, if for all $n > n_0$, g satisfies the two conditions*

$$g(n) \leq \left(\sum_{i=1}^{n} p_{ni}\left(1 - \frac{(i-1)^2}{n^2} - \frac{(n-i)^2}{n^2}\right)\right)^{-1} \quad \textit{and}$$

$$\frac{g(i)}{g(n)} \geq \frac{i}{n} \quad \textit{for all} \quad 0 \leq i \leq n\,.$$

(ii) *Part (i) still holds if we replace the two conditions by*

$$g(n) \leq \left(\sum_{i=1}^{n} p_{ni} H\left(\frac{i}{n+1}\right)\right)^{-1} \quad \textit{and}$$

$$\frac{g(i)}{g(n)} \geq \frac{i}{n} \quad \textit{for all} \quad 0 \leq i \leq n.$$

Proof We prove (i) first, by induction. For $n \leq n_0$, just set the constant $c \leq 1$ small enough.

Now we look at the case $n > n_0$. Let $P_n = (p_{n1}, \ldots, p_{nn})$ be a distribution where p_{ni} is the probability that we choose as a pivot element the element with rank i. Using the induction hypothesis, it holds that

$$\begin{aligned}
&T(i-1) + T(n-i)\\
&\geq c(i-1)g(i-1) + c(n-i)g(n-i) - (n-1)\\
&= cng(n)\left(\frac{(i-1)g(i-1)}{ng(n)} + \frac{(n-i)g(n-i)}{ng(n)}\right) - (n-1)\\
&\geq cng(n)\left(\frac{(i-1)^2}{n^2} + \frac{(n-i)^2}{n^2}\right) - (n-1)\\
&= cng(n) - cng(n)\left(1 - \frac{(i-1)^2}{n^2} - \frac{(n-i)^2}{n^2}\right) - (n-1)\,.
\end{aligned}$$

Therefore,

$$\begin{aligned} T(n) &= n-1+\sum_{i=1}^{n} p_{ni}(T(i-1)+T(n-i)) \\ &\geq cng(n) - cng(n)\sum_{i=1}^{n} p_{ni}\left(1-\frac{(i-1)^2}{n^2}-\frac{(n-i)^2}{n^2}\right) \ . \end{aligned}$$

As $c \leq 1$, we can finish the induction if

$$g(n) \leq \left(\sum_{i=1}^{n} p_{ni}\left(1-\frac{(i-1)^2}{n^2}-\frac{(n-i)^2}{n^2}\right)\right)^{-1} \ .$$

The proof of part (ii) is quite similar. For $n \geq n_0$,

$$\begin{aligned} &T(i-1)+T(n-i) \\ &\geq cng(n)\left(\frac{(i-1)^2}{n^2}+\frac{(n-i)^2}{n^2}\right)-(n-1) \\ &= cng(n)\left(\frac{(i-1)^2}{n^2}+\frac{(n-i)^2}{n^2}+H\left(\frac{i}{n+1}\right)\right)-ng(n)H\left(\frac{i}{n+1}\right)-(n-1) \\ &\geq cng(n)-cng(n)H\left(\frac{i}{n+1}\right)-(n-1) \ . \end{aligned}$$

The last inequality uses Lemma 17.1. Now

$$\begin{aligned} T(n) &= n-1+c\sum_{i=1}^{n} p_{ni}(T(i-1)+T(n-i)) \\ &\geq cng(n)-cng(n)\sum_{i=1}^{n} p_{ni}H\left(\frac{i}{n+1}\right) \ . \end{aligned}$$

Again using $c \leq 1$, we can finish the induction if

$$g(n) \leq \left(\sum_{i=1}^{n} p_{ni}H\left(\frac{i}{n+1}\right)\right)^{-1} \ .$$

□

Remark 17.1 *In the second part of Theorem 17.2 the lower bound is given using the entropy function, similar to the upper bound in Theorem 17.1. This shows that, up to a logarithmic factor, we yield matching upper and lower bounds.*

Note that the condition $g(i)/g(n) \geq i/n$ is not actually a limitation. We already know that QuickSort's running time ranges from $n\log_2 n$ to n^2, so our function g will meet the condition anyway.

17.4 The δ-Random Source

A general model of a random bit source is the δ-random-source, which is sometimes also referred to as the *slightly random source*. Since the bias of each bit is a function of the previous output, it can be applied as an adversary argument and is particularly suited for worst-case analysis. See also [1, 8, 10].

Definition 17.1 (See [1]) *A δ-random-source is a random bit generator. Its bias may depend on the bits it has previously output, but the probability to output "1" must be in the range $[\delta, 1-\delta]$. Therefore, it has an internal state $\omega \in \{0,1\}^*$, denoting its previously output bits.*

To obtain a random number X in the range $1, \dots, n$ from the δ-random-source, we output $\lceil \log n \rceil$ bits and interpret them as a number Y. Then, we set $X := (Y \bmod n) + 1$.

Lemma 17.2 (See [2]) *For each p with $0 < p < 1/2$, there exists a constant c, such that for all $n \in \mathbb{N}$: $c\dfrac{2^{H(p)n}}{\sqrt{n}} \le \displaystyle\sum_{j=0}^{\lfloor np \rfloor} \binom{n}{j} \le 2^{H(p)n}$.*

Theorem 17.3 *For each δ-random-source, $0 < \delta < 1/2$, there exists $n_0 \in \mathbb{N}$, such that for each $n > n_0$, and each permutation π, Theorem 17.1 can be applied with*

$$g(n) = c_\delta \frac{1}{\sqrt{\log n}} n^{1-H(\delta)} ,$$

where the random bits are produced by a δ-random-source and c_δ is a constant that depends on δ.

Proof From the symmetry and monotony of the entropy function it follows that, for each s,

$$\sum_{i=1}^{n} p_{ni} H\left(\frac{i}{n}\right) \ge \left(1 - \sup_{\pi,\tilde{\omega}} \sum_{j=1}^{s-1} p_{n\pi^{-1}(j)}\right) H\left(\frac{s}{2n}\right) \tag{17.1}$$

where $p_{n\pi^{-1}(j)}$ depends on the internal state $\tilde{\omega}$ of the random source.

Now we examine the two factors on the right-hand side of (17.1) separately. We set $k := \lceil \log n \rceil$ and $s := 1/2 \sum_{j=0}^{\lfloor \delta k \rfloor} \binom{k}{j}$. Since

$$p_j = \begin{cases} \Pr[Y = \pi(j)], & \text{if } n + \pi(j) \ge 2^k \\ \Pr[Y = \pi(j)] + \Pr[Y = \pi(j) + n] & \text{otherwise} \end{cases}$$

we get for the first factor of (17.1)

$$\sup_{\pi,\tilde{\omega}} \sum_{j=1}^{s-1} p_{n\pi^{-1}(j)} \leq \sup_{\tilde{\omega}} \max_{M\subseteq\{0,1\}^k,|M|=2s} \Pr[Y \in M] \leq \sum_{j=0}^{\lfloor \delta k \rfloor} \binom{k}{j} \delta^j (1-\delta)^{k-j} .$$

Here we use the result from [1], that the maximum probability of hitting a set of a certain size can be achieved by an "extreme" δ-random-source that always outputs "0" with probability δ.

Since

$$\lim_{k\to\infty} \sum_{j=0}^{\lfloor \delta k \rfloor} \binom{k}{j} \delta^j (1-\delta)^{k-j} = \frac{1}{2}$$

(which follows from the DeMoivre–Laplace Limit Theorem, see [3]) there exists a constant c'_δ, such that

$$\sup_{\pi,\tilde{\omega}} \sum_{j=1}^{s-1} p_{n\pi^{-1}(j)} \leq c'_\delta .$$

Now we consider the second factor of (17.1). We use the monotony of $H(x)$ on the interval $[0, 1/2]$ and Lemma 17.2:

$$H\left(\frac{s}{2n}\right) \geq H\left(\frac{s}{2^{k+1}}\right) \geq H\left(\tilde{c}_\delta \frac{2^{(H(\delta)-1)k}}{4\sqrt{k}}\right) .$$

We consider $\delta < 1/2$ (so that $H(\delta) < 1$) and use that $H(x) \geq -x \log x$ to get

$$H\left(\frac{s}{2n}\right) \geq \tilde{c}_\delta \frac{2^{(H(\delta)-1)k}}{4\sqrt{k}} \left((1-H(\delta))k - \log \frac{\tilde{c}_\delta}{4\sqrt{k}}\right) .$$

For k big enough ($k > k_0$ corresponds to $n > n_0$), there is a constant c''_δ so that

$$H\left(\frac{s}{2n}\right) \geq c''_\delta \sqrt{k} 2^{(H(\delta)-1)k} .$$

Combining the results, there is a $n_0 \in \mathbb{N}$ and a c^*_δ, such that for all $n \geq n_0$, and all permutations π on $\{0, \ldots, n-i\}$ and all states $\tilde{\omega} \in \{0, 1\}^*$ of the generator, the following holds:

$$\begin{aligned}\sum_{i=1}^{n} p_{n\pi^{-1}(i)} H\left(\frac{i}{n}\right) &\geq c^*(\delta) \sqrt{\lceil \log n \rceil} 2^{(H(\delta)-1)\lceil \log n \rceil} \\ &\geq \frac{1}{c_\delta} \sqrt{\log n} n^{H(\delta)-1}\end{aligned}$$

which leads to the expected running time of $T(n) \leq c_\delta n^{2-H(\delta)} \sqrt{\log n}$. □

17.5 Conclusion

A new measure for sequences of probability distributions was given that can be used to bound the running time of the randomized QuickSort algorithm. For the upper bound, it can be applied to the distributions on array positions used to divide the input array. For the lower bound, however, we can only apply it to the distributions on the ranks of the elements that we use to divide the input.

There is still a gap of $\log n$ between the lower and upper bound. A more sophisticated analysis of the problem might close that gap, probably by raising the lower bound to something like $ng(n)\log n - n$.

References

1 Alon, N. and Rabin, M.O. (**1989**) Biased coins and randomized algorithms, in, *Advances in Computing Research 5*, (eds F.P. Preparata, S. Micali), JAI Press, 499–507.

2 Ash, R.B. (**1965**) *Information Theory*, Dover.

3 Feller, W. (**1961**) *An Introduction to Probability Theory and its Applications*, John Wiley.

4 Hoare, C.A.R. (**1962**) Quicksort. *Computer Journal*, **5**(1), 10–15.

5 Karloff, H.J. and Raghavan, P. (**1993**) Randomized algorithms and pseudorandom numbers. *Journal of the Association for Computing Machinery*, **40**, 454–476.

6 Knuth, D. (**1973**) *The Art of Computer Programming. Vol 3: Sorting and Searching*, Addison-Wesley.

7 Lehmer, D.H. (**1951**) *Mathematical Methods in Large-scale Computing Units*. Proc. 2nd Symp. on Large-scale Digital Calculating Machinery. Harvard University Press, 141–146.

8 Papadimitriou, C.H. (**1994**) *Computational Complexity*, Addison-Wesley.

9 Robert Sedgewick, R. and Flajolet, P. (**1994**) *Analysis of Algorithms*, Addison-Wesley.

10 Santha, M. and Vazirani, U.V. (**1984**) *Generating quasi-random sequences from slightly random sources*. Proceedings of the 25th IEEE.

11 Shannon, C. (**1948**) A mathematical theory of communication. *The Bell System Technical Journal*, July, October, **27**, 379–423, 623–656.

12 Tompa, M. (**1991**) *Probabilistic Algorithms and Pseudorandom Generators*. Lecture Notes.

Further Reading

1 Devroye, L. (**2001**) On the probabilistic worst-case time of "FIND". *Algorithmica*, **31**, 291–303.

2 List, B. (**1999**) Probabilistische Algorithmen und schlechte Zufallszahlen. PhD thesis, Universität Ulm.

Index