

Hartmann Römer

Theoretical Optics

An Introduction



WILEY-
VCH

WILEY-VCH Verlag GmbH & Co. KGaA

Author

Prof. Dr. Hartmann Römer
University of Freiburg
Institute of Physics
hartmann.roemer@physik.uni-freiburg.de

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.

Cover Picture

Courtesy of Jos Stam (Alias wavefront)

**Library of Congress Card No.: applied for
British Library Cataloging-in-Publication Data:**

A catalogue record for this book is available from the British Library

**Bibliographic information published by
Die Deutsche Bibliothek**

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <<http://dnb.ddb.de>>.

© 2005 WILEY-VCH Verlag GmbH & Co. KGaA,
Weinheim

All rights reserved (including those of translation into other languages). No part of this book may be reproduced in any form – nor transmitted or translated into machine language without written permission from the publishers. Registered names, trademarks, etc. used in this book, even when not specifically marked as such, are not to be considered unprotected by law.

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

Printing betz-druck GmbH, Darmstadt
Bookbinding Großbuchbinderei J. Schäffer GmbH
& Co. KG., Grünstadt

ISBN 3-527-40429-5

Contents

Preface to the German edition	IX
Preface to the English edition	XIII
1 A short survey of the history of optics	1
2 The electrodynamics of continuous media	15
2.1 Maxwell's equations	15
2.2 Molecular vs. macroscopic fields	18
2.3 A simple model for the electric current	20
2.4 Dispersion relations and the passivity condition	23
2.5 Electric displacement density and magnetic field strength	27
2.6 Index of refraction and coefficient of absorption	33
2.7 The electromagnetic material quantities	35
2.8 The oscillator model for the electric susceptibility	39
2.9 Material equations in moving media	40
3 Linear waves in homogeneous media	45
3.1 Elastic waves in solids	45
3.2 Isotropic elastic media	48
3.3 Wave surfaces and ray surfaces	51
4 Crystal optics	55
4.1 The normal ellipsoid	55
4.2 Plane waves in crystals	58
4.3 Optically uniaxial crystals	62
4.4 Optically biaxial crystals	65
4.5 Reflection and refraction at interfaces	66
4.6 Fresnel's equations	69
4.7 The Fabry–Perot interferometer	72
5 Electro-, magneto- and elasto-optical phenomena	75
5.1 Polarization effects up to first order – optical activity	75
5.2 Polarization effects of higher order	79
5.2.1 Dependence on distortions	80

5.2.2	Dependence on shear flows	80
5.2.3	Influence of electric fields	80
5.2.4	Dependence on magnetic fields	81
6	Foundations of nonlinear optics	83
6.1	Nonlinear polarization – combination frequencies	83
6.2	Nonlinear waves in a medium	85
6.3	Survey of phenomena in nonlinear optics	89
6.4	Parametric amplification and frequency doubling	91
6.5	Phase matching	93
6.6	Self-focussing, optical bistability, phase self-modulation	95
6.7	Phase conjugation	98
6.8	Fiber optics and optical solitons	101
7	Short-wave asymptotics	107
7.1	Introductory remarks	107
7.2	Short-wave expansion of Maxwell's equations	109
7.3	The scalar wave equation	111
7.4	Phase surfaces and rays	113
7.5	Fermat's principle	115
7.6	Analogy between mechanics and geometrical optics	116
8	Geometrical optics	121
8.1	Fermat's principle and focal points	121
8.2	Perfect optical instruments	122
8.3	Maxwell's fish-eye	123
8.4	Canonical transformations and eikonal functions	125
8.5	Imaging points close to the optic axis by wide spread ray bundles	128
8.6	Linear geometrical optics and symplectic transformations	131
8.7	Gaussian optics and image matrices	134
8.8	Lens defects and Seidel's theory of aberrations	139
9	Geometric theory of caustics	143
9.1	Short-wave asymptotics for linear partial differential equations	143
9.2	Solution of the characteristic equation	146
9.3	Solution of the transport equation	151
9.4	Focal points and caustics	153
9.5	Behavior of phases in the vicinity of caustics	156
9.6	Caustics, Lagrangian submanifolds and Maslov index	158
9.7	Supplementary remarks on geometrical short-wave asymptotics	161
10	Diffraction theory	167
10.1	Survey	167
10.2	The principles of Huygens and Fresnel	167
10.3	The method of stationary phases	171

10.4	Kirchhoff's representation of the wave amplitude	175
10.5	Kirchhoff's theory of diffraction	179
10.6	Diffraction at an edge	184
10.7	Examples of Fraunhofer diffraction	186
10.7.1	Diffraction by a rectangle	187
10.7.2	Diffraction by a circular aperture	188
10.7.3	Arrangements of several identical structures	189
10.8	Optical image processing in Fourier space	191
10.9	Morse families	195
10.10	Oscillatory functions and Fourier integral operators	198
11	Holography	203
11.1	The principle of holography	203
11.2	Modifications and applications	205
11.2.1	Observing small object deformations	206
11.2.2	Holographic optical instruments	206
11.2.3	Pattern recognition	207
11.3	Volume holograms	207
12	Coherence theory	211
12.1	Coherent and incoherent light	211
12.2	Real and analytical signals	213
12.3	The light wave field as a stochastic process	217
12.4	Gaussian stochastic processes	220
12.5	The quasi-monochromatic approximation	222
12.6	Coherence and correlation functions	224
12.7	The propagation of the correlation function	227
12.8	Amplitude and intensity interferometry	230
12.8.1	Amplitude interferometry: Michelson interferometer	230
12.8.2	Photon correlation spectroscopy	231
12.9	Dynamical light scattering	232
12.10	Granulation	236
12.11	Image processing by filtering	237
12.12	Polarization of partially coherent light	239
13	Quantum states of the electromagnetic field	245
13.1	Quantization of the electromagnetic field and harmonic oscillators	245
13.2	Coherent and squeezed states	251
13.3	Operators, ordering procedures and star products	259
13.4	The Q , P , and Wigner functions of a density operator	266
14	Detection of radiation fields	273
14.1	Beam splitters and homodyne detection	273
14.2	Correlation functions and quantum coherence	279
14.3	Measurement of correlation functions	281
14.4	Anti-bunching and sub-Poissonian light	285

15 Interaction of radiation and matter	289
15.1 The electric dipole interaction	289
15.2 Simple laser theory	294
15.3 Three-level systems and atomic interference	296
15.3.1 Electromagnetically induced transparency	299
15.3.2 Refractive index enhancement	301
15.3.3 Lasing without inversion	301
15.3.4 Correlated emission laser	301
15.4 The Jaynes–Cummings model	302
15.5 The micromaser	308
15.6 Quantum state engineering	310
15.7 The Paul trap	313
15.8 Motion of a two-level atom in a quantized light field	320
16 Quantum optics and fundamental quantum theory	323
16.1 Quantum entanglement	323
16.2 Bell’s inequalities	328
16.3 Quantum erasers and measurement without interaction	332
16.4 No cloning and quantum teleportation	337
16.5 Quantum cryptography	342
16.6 Quantum computation	343
Selected references	351
Index	355

Preface to the German edition

The idea for writing this book crystallized from an interest in methods of short-wave asymptotics and symplectic geometry. Later followed a course about theoretical optics, which gave me great pleasure. This pleasure turned out to be lasting, and so the present book arose from several revisions and extensions of the original manuscript.

Indeed, there are many reasons to present the venerable and traditional field of optics in a new form. The discovery of lasers, the steady progress of powerful data-processing systems, and the development of new materials with unusual optical properties have all revolutionized the field of optics.

Lasers are light sources of unprecedented intensity and coherence, without which such new branches as nonlinear optics and holography would have been impossible. Computer technology allows the processing of optical signals and the construction of diffractive optical devices, which have already taken their place next to common lenses and mirrors. Glass-fiber cables, compact disks, and simple holograms are proof that the products of the new kind of optics have already entered our daily life.

However, not only have experimental physics and technology experienced rapid progress, but also we have seen the development and application of new theoretical methods in optics:

The modern theory of nonlinear dynamical systems has found many applications in nonlinear optics. Here we meet bifurcation and chaotic behavior as well as dispersion-free solutions of nonlinear integrable systems.

The adoption of the concepts and methods from the theory of stochastic processes for the description of fluctuating light wave fields turned out to be extremely fruitful. In this way, stochastic optics became a new branch of theoretical optics; the notion of coherence found a more profound formulation, and today applications for image processing and correlation spectroscopy have become standard routines.

Progress in the theory of short-wave asymptotics within the framework of symplectic geometry not only led to an improvement of the WKB method and to a better understanding of the quasi-classical limit of quantum mechanics, but “symplectic optics” also allows a deeper insight into the geometrical structures of the realm between wave optics and ray optics, into the nature of caustics, and into the theory of diffraction.

These overwhelming developments in the applications and the theory of optics have led to a considerable number of publications in recent years. However, these presentations, often written in the form of an experimental textbook, either provide a summary of methods, phenomena, and applications of modern optics, or they describe, in the form of a monograph, special parts of experimental, applied, or theoretical optics.

The present book attempts a coherent and concise presentation of optics, emphasizing the perspective of a theoretician. However, it is not meant to be a textbook on mathematical physics, but tries instead to mediate among different positions:

For the experimentalist, the applied physicist, and the theoretician, this book aims to provide a unification and a deeper understanding of the theoretical background, and, for those who are interested, a first access to the corresponding mathematical literature. The theoretically or mathematically inclined reader is introduced to the “applications” and the manifold phenomena related to the theory of light. What, in my opinion, makes optics so particularly attractive is that in this field the path between theory and phenomena is shorter and more straightforward than in other physical areas.

In order to remain comprehensible for the above-mentioned group of readers, I have assumed as little as possible previous knowledge. The book requires only basic knowledge about Maxwell’s equations and the underlying elementary vector analysis, a certain familiarity with the essential properties of Fourier transformations, and the simplest phenomena of the propagation of waves and wave packets, in particular the notion of group velocity.

In short, this book pursues the following three aims:

- provision of a theoretical overview;
- theoretical extension and introduction to the fundamental mathematics;
- presentation and interpretation of many important optical phenomena.

A look at the table of contents reveals that, in particular, this third aim has not been neglected: the description of manifold phenomena in crystal optics, nonlinear optics, geometrical optics, diffraction theory, diffraction optics, as well as statistical optics and coherence optics will provide this book with a solid frame.

Unfortunately, certain choices and limitation of the material turned out to be unavoidable. I decided to include only a short presentation of those optical phenomena which depend mainly on quantum theory or the particle picture of light. In particular, this refers to quantum optics and the theory of lasers, as well as the interaction of light with matter, like the photoelectric effect, the Compton effect, pair creation and bremsstrahlung, and, finally, applications of optics in atomic and molecular spectroscopy.

In this way, the presentation of the material in this book could be structured as a straightforward development of the content of Maxwell’s equations:

Chapter 1 contains some historical remarks, with an emphasis on the development of the wave theory of light and the description of Newtonian optics.

Chapter 2 develops in a brief and concise form the electrodynamics of media and it is shown how the influence of polarizable media can be taken into account by the introduction of the additional fields $\mathbf{D}$ and $\mathbf{H}$. We will describe the causality and passivity conditions of media and their influence on the conductivity and susceptibilities.

The following four chapters are devoted to the propagation of waves in homogeneous but not necessarily isotropic media. Chapter 3 serves as an illustrative introduction to the general theory of wave propagation in elastic media; in particular, we will explain the notions of a wave surface and a ray surface, which will turn out to be essential for the coming chapters.

Chapter 4 summarizes the basic concepts of the theory of crystal optics. Among other things, we will discuss double refraction, conical refraction, and reflection and refraction at interfaces between homogeneous media.

Chapter 5 deals with the interesting phenomena related to electro-, magneto-, and elasto-optical properties, while Chapter 6 is devoted to nonlinear optics. It contains a summary of the most important nonlinear optical phenomena, the theory of nonlinear waves, and the coupling of three waves. We will describe the phenomena of frequency doubling, parametric amplification, self-focussing, momentum contraction, phase conjugation, wave conduction in glass fibers, and optical solitons.

The following five chapters deal with the propagation of waves in isotropic but not necessarily homogeneous media. In general, this is a very difficult problem, which includes, amongst other things, the theory of most optical elements. The necessary tools will turn out to be the short-wave asymptotics and the theory of diffraction. Although the presentation will remain on an elementary level, we also want to provide the necessary foundations for the advanced mathematical theories.

In Chapter 7, we describe the transition from wave optics to geometrical optics and develop a formal analogy between classical mechanics and geometrical optics.

Chapter 8 is devoted to geometrical optics. Special emphasis is put on the presentation of matrix methods and the relation to linear symplectic transformations. Additional subjects are the impossibility of perfect optical instruments, and Seidel's theory of aberrations.

Chapter 9 contains a general geometrical treatment of short-wave asymptotics, in particular a geometrical theory of caustics. In this context we will explain the notions of characteristic equations, transport equations, focal points, Lagrangian submanifolds, and Maslow index. There the interested reader will find a short introduction to symplectic geometry.

In Chapter 10, we will discuss the theory of diffraction, in particular the principles of Huygens and Fresnel, Fraunhofer diffraction, and image processing in Fourier space. Two sections at the end of this chapter contain an introduction to the theory of Morse families and Fourier integral operators.

Chapter 11 describes briefly the foundations and applications of holography, a particularly attractive branch of diffraction optics.

Chapter 12 concludes this book with a summarizing presentation of statistical optics. The central subject is a description of the wave field by stochastic processes, which, in a natural way, leads to the notions of the various correlation functions. As applications, we will explain correlation spectroscopy, dynamical light scattering, speckle effect, and image processing by filtering.

It is my sincere hope that this book will give the beginner a comprehensible introduction to the subject of theoretical objects and, at the same time, will convey to the advanced reader many interesting insights.

Finally, it is my pleasant duty to thank all those who have helped me, often essentially, during work on this book: First of all I should like to thank the students who attended my lectures. Their attention and interest has always been encouraging for me, and I owe them thanks for many helpful suggestions. Furthermore, I am very grateful to former members of the Institute for Theoretical Physics for their constructive criticism and comments, in particular J. Barth, W. Bischoff, C. Emmrich, M. Forger, T. Filk, D. Giulini, P. Glöbner, C. Kiefer, M. Koch, A. Münnich, K. Nowak, H. Steger, and A. Winterhalder, as well as E. Binz from

the University of Mannheim. C. Heinisch from Kaiserslautern has, with great patience, set the manuscript in $\text{\LaTeX}$.

Two persons I should like to mention with special thanks: E. Meinrenken was one of the attendees at my lectures. Later, he dealt intensively with the theory of short-wave asymptotics. The final mathematical sections in chapters 9 and 10 owe much to the presentations in his diploma thesis. Furthermore, he carefully looked through the preliminary versions of chapters 7 to 10. I am also grateful to G. Jerke from the VCH publishing company. His constant interest in the development of this book, his personal efforts, and countless valuable suggestions went far beyond what one can usually expect during such a book project. Finally, I should like to thank the VCH publishing company, in particular R. Wengenmayr, for their pleasant and confidential cooperation.

Freiburg, April 1994

Hartmann Römer

Preface to the English edition

With the appearance of the English edition, I am glad to see my book becoming accessible for a wider public. On the other hand, being cut off from my mother tongue in a matter of personal importance, I cannot conceal a feeling of alienation and loss of power. Fortunately, the English edition provided an occasion for a substantial augmentation by four new chapters on quantum optics, a subject of rapidly increasing fundamental and technical importance.

In the presentation of quantum optics, it is not my ambition to compete with available comprehensive monographs on this subject. Rather, and in accordance with the spirit of this book, I aim at a concise and coherent account, bringing out the basic structures in a clear and conceptually (not necessarily mathematically) rigorous way.

Chapter 13 deals with the quantization of the electromagnetic field and with a description of its quantum states, in particular coherent and squeezed states. For the theory of operators on quantum states, I use the opportunity to introduce the notion of star products, one of the central subjects of research in our group in Freiburg. I think that star products are a very useful and simplifying tool in quantum optics, which should be more widely known.

Chapter 14 is concerned with the detection of light fields: homodyne detection, interferometry and photon count statistics.

Chapter 15 is devoted to the interaction of light and matter. It contains central subjects like the Jaynes–Cumming model, three-state systems, the theory of the micromaser and the Paul trap, and of forces enacted on matter by light.

In chapter 16, I give a brief introduction into the fascinating subject of testing basic features of quantum theory by methods of quantum optics. “Gedanken” experiments have become feasible in this field. Quantum computers are a particularly interesting emerging application.

It is my pleasant duty to thank those people who have helped me to realize this English edition. In the first place I thank Thomas Filk, who not only translated the twelve chapters of the German edition and did the typing of the whole book, but also contributed valuable advice. Thanks are also due to the attendees at my lectures on quantum optics for their enthusiasm and valuable feedback.

Very helpful contributions also came from Kerstin Kunze, who was responsible for the exercises accompanying my course. She, and also Stefan Jansen and Svea Beiser, carefully proofread the last four chapters. I had many useful conversations with Stefan Waldmann. Finally, I would also like to thank my colleagues H. P. Helm, H. Reik and M. Weidemüller for kind interest, encouragement and helpful suggestions.

1 A short survey of the history of optics

Until the beginning of the seventeenth century, our understanding of the nature and properties of light evolved rather slowly, although, in contrast to electrical phenomena, for example, optical phenomena are straightforward to observe.

Ever since ancient times we have known about the straight-line propagation of light, which is most obvious in the shadows cast from a light source. From this observation, the ancient Greeks already developed the concept of straight light rays. This idea, however, was intimately entangled with the theory of “seeing rays”, which were emitted from the eyes and palpated, with the help of light, visible objects like feelers.

Another well-known fact was the equality of the angles of incidence and reflection for light rays, and Hero of Alexandria was able to attribute this law to the more general principle of a shortest path for light. Common optical devices were the gnomon as well as plane and curved mirrors and lenses. At least since ancient Roman times, magnifying glasses were in use.

Concerning the nature of light, different concepts were put forward. The “atomists” following Democritus (about 460–371 BC) and his less well-known predecessor Leucippus (about 480–??? BC) believed that all objects consisted of atoms traveling through empty space. All changes could be attributed to the movements and rearrangements of atoms. In this view, light rays were considered to be a flux of light particles traveling in straight lines and freely through empty space and which could penetrate transparent bodies. The different types of colors were explained by different shapes or sizes of these atoms of light. Later the Roman Lucretius (about 96–55 BC) formulated a systematic summary of the atomistic viewpoint in his savant poem “De Rerum Natura”.

In clear contrast to these ideas was the opinion of Aristotle (384–322 BC), which he formulated in his treatments on the soul (“De Anima” II,7) and the senses (“De Sensu” III). For ontological reasons the concept of empty space as an existing non-being was unacceptable for Aristotle. For him, light was not a substance or body but a quality; to be more precise, the actuality of the quality of the transparent. The mere potentiality of the transparent is darkness. The transition from the potentially transparent to the actually transparent happens under the influence of fire or the shining bodies of Heaven. The primarily visible quality of objects is their color. Colors are qualities which, like light, become actual in bodies that are not completely transparent but only participate in the nature of transparency to a certain degree. Therefore, it seems natural that Aristotle calls light the “color of the transparent”.

As is commonly known, the mainstream thinking in Europe did not follow the atomistic view for many centuries. In particular, the simple and plausible color theory of Aristotle was of far-reaching influence and initiated, amongst others, the color theory developed by Goethe.

Throughout the Middle Ages, the atomistic philosophy was only known from discussions among the Aristotelians and as an object of polemics for the Early Fathers. It was not until 1417, after Poggio Bracciolini (1380–1459) had recovered a hand-written document of Lucretius, that a first-hand presentation of the atomistic ideas was available.

During the long era following the end of the ancient world up to the beginning of the seventeenth century, the efforts for a better understanding of the nature of light saw but a few highlights.

In the moslem Near East one of the outstanding scientists was Ibn al Haitham (963–1039), in Europe known by the name Alhazen. He found that the assumption of Ptolemaeus (about 100–170) about the proportionality of the angles of incidence and reflection was an approximation that only holds for small angles. Furthermore, he succeeded in giving a precise description of the functioning of the human eye.

Roger Bacon (1215–1294) knew quite well about the properties of lenses and concave mirrors and is regarded as the inventor of the camera obscura. Even more important is the invention of spectacles around 1299 by Salvino degli Armati from Florence. At the end of this chapter, the reader will find a table of important names and dates related to the history of optics.

The breakthrough in the seventeenth century was initiated by the invention of new optical instruments. In 1600 the Dutch maker of spectacles, Zacharias Janssen from Middelburg, built the first microscope. This instrument was continuously improved during the seventeenth century and allowed a glimpse into a world that remained out of reach for the naked eye. The same holds for the telescope, which, according to most sources, was invented by the Dutch Hans Lippershey around 1608. News of this discovery spread fast throughout Europe and caused Galileo Galilei (1564–1642) in 1609 to construct his telescope. In 1611 Johannes Kepler (1571–1630) published the drawings of his telescope using a convex lens ocular.

Finally, in 1621, Willebrord Snell (1591–1626) found the correct law of refraction and thereby solved one of the fundamental problems in geometrical optics after more than 1500 years. Even Kepler had tried to attack this problem and failed. In its present form the law of refraction was published by René Descartes (1596–1650) who, in 1644, derived it from his bizarre theory of light, which we will discuss later. He also found the right explanation for the rainbow. In 1657, Pierre de Fermat (1601–1665) derived the law of refraction from the principle of least time, which assumed the propagation of light inside optically dense media to be slower than outside.

After the discovery of the law of refraction, mathematicians like Carl Friedrich Gauss (1777–1855), William Rowan Hamilton (1805–1865), and Ernst Abbe (1840–1905) took over and continuously improved the theory of geometrical optics. Hamilton based his theory on Fermat's principle, and it was only later that he applied his methods to the realm of analytical mechanics. During his lifetime it was in particular the discovery of conical refraction that made him famous.

The scientifically very fruitful seventeenth century saw also the discovery of many fundamental optical phenomena: In his book “*Physicomathesis de Lumine, Coloribus et Iride*”, which appeared in 1665 after his death, the Jesuit clergyman Francesco Maria Grimaldi (1618–1663) described the phenomenon of refraction; double refraction was found in 1669 by Erasmus Bartholinus (1625–1698); and the phenomenon of polarization was discovered by Christiaan Huygens (1629–1695), who described the effect in his “*Traité de la Lumière*” in

1690. However, in contrast to the phenomenon of double refraction, he was not able to explain it. Finally, we should emphasize Olaus Rømer (1644–1710), who in 1676 determined the velocity of light from the delayed appearances of the eclipses of the moons of Jupiter during a period of increasing distance between Jupiter and the Earth.

Concerning the understanding of the natures of light and color, three different concepts competed during the seventeenth century. Still of great influence and part of the canonical teaching at universities was the philosophy of Aristotle, which coined, for example, the ideas of Kepler.

Others, like Robert Boyle (1627–1691), followed the ancient atomistic view and vehemently propagated a corpuscular theory of light. Boyle, for example, related the various colors to different velocities of the light particles.

A third group, the so-called “plenists”, differed from the atomists in their objection to the existence of empty space. For them, light was in some obscure way related to flows, vortices, and waves; they assumed the global existence of a continuous, space-filling medium for the propagation of an action. These ideas laid the foundations for the wave theory of light, which later, in the nineteenth century, became the dominant view in science.

Because of its great influence on the philosophy of the seventeenth century, the plenistic theory of Descartes is of special importance. Any form of corporeal substance is a “*res extensa*”, characterized by its principal property of extension and occupation of space. A physical vacuum would be a contradiction in itself.

Descartes distinguished three types of matter, which differentiated from the primordial homogeneous bulk substance: a very tiny, subtle matter that one may identify with an ether; a so-called spherical matter consisting of small impenetrable spheres; and finally the third type of bulk matter from which the large material bodies in our world are formed. All three forms of matter fill space completely, and any action between material substances can be traced back to a “mechanical origin”, which for Descartes consisted only of collisions or direct contacts.

According to Descartes, light constitutes an action of pressure exerted onto the spherical matter by the subtle matter, which is in constant motion. Because the spheres are in direct contact and of infinite rigidity, light propagates with infinite velocity. Refraction can produce colored light, if the spheres of the second type of matter are set into rotation through a deflection of the propagation direction. From his model, Descartes was able to derive Snell’s law of refraction. Descartes explained the planetary rotations about their own axes as well as around the Sun “mechanically” as vortex-like movements of the subtle matter, which in turn would carry along the third type of matter.

Descartes did not trust in the truth of sensations but was convinced that thinking and reason were the only sources of reliable knowledge. The general influence of his program to mathematize the natural sciences was great. However, especially in England, people like Francis Bacon and Isaac Newton objected to his way of gaining knowledge by pure speculations. His mechanical models were considered to be artificial, lacking justification based on experience, and without any power of explanation or prediction. On many occasions Newton emphasized his disapproval of “hypotheses”, which to a large extent may be seen as his response to the Cartesian philosophy.

In his “*Physicomathesis*”, Grimaldi comes much closer to a wave theory of light than Descartes. He not only describes the phenomena of diffraction on small objects, including color fringes and the brightening in the center of the shadow, but he also mentions the phe-

nomenon of interference (although he did not use this word) behind two closely neighboring apertures, and the colors appearing in the reflection of light from surfaces with densely spaced grooves. He also observed deviations from straight-line propagation of light and called them “diffractions”. Grimaldi compares the propagation of light with the floating of a fine liquid, but he does not draw far-reaching conclusions from this picture. The long-winded, roundabout style of his writing strongly diminished the influence of his work. Often the distinction of flow, vortex formation, and real propagation of light waves is not obvious, but he makes the comparisons with spherically spreading water waves and the propagation of sound. He interprets the variety of colors as the undulations of a fluid, which are as versatile as the handwriting of different people.

It was Robert Hooke (1635–1703) who formulated a real wave theory of light. He considered light as the propagation of a longitudinal wave in an ether. In his “*Micrographia*” from 1665, he describes the appearance of color at thin films, and he seeks an explanation in the reflection of light at both sides, the front as well as the back side of the film. As he and his contemporaries were lacking the notion of interference, he was not able to complete this explanation. The appearance of color in the refraction of light is explained by wave fronts, which after refraction are no longer parallel to the direction of propagation. In general, colors were described as modifications of white light.

In his “*Traité de la Lumière*”, Christiaan Huygens developed the well-known principle now bearing his name, according to which at every point of space the passing light excites elementary waves (or wavelets). Using this picture he derived a simple and convincing explanation of Snell’s law of refraction, which to the present day is contained in almost every school textbook. The most brilliant triumph of this model was the quantitative description of double refraction in calcite by assuming spherically and ellipsoidally shaped wavelets.

Most surprisingly, Huygens model is not really a wave theory. His wavelets are rather to be compared with wave fronts or shock waves, from which the resulting wave fronts may be reconstructed as the envelope. They are not related to some kind of periodic motion.

In his *Traité*, he does not elaborate on the problem of colors, although his comments on Newton’s color theory indicate his preference for a color theory based on two or three primary colors.

We should now emphasize the contributions of Isaac Newton (1643–1724) who, for a long time, eclipsed all other natural philosophers not only because of his deciphering of the laws of motion and gravity but also for his contributions to the field of optics. Almost all lines of developments enter and sometimes even interfere in his works on optics, and nowhere else are the contradictory aspects of the wave and corpuscular theory of light more obvious than in his person.

With full justification, Newton is generally considered to be the representative for the corpuscular theory of light. In this aspect, he rather represented a minority of those scholars who were involved in the progress of optics at this time. Newton has often been blamed that under the pressure of his authority the breakthrough of the wave theory was delayed until the nineteenth century. However, a deeper look reveals that such an oversimplified judgment of Newton’s role is inappropriate.

Like his development of the infinitesimal calculus and theory of gravitation, one can trace back the works of Newton on optics to his earlier years, in particular to the eminently fruitful time between 1665 and 1667, which he spent in his birth town Woolsthorpe after the University

of Cambridge had to be closed in the aftermath of the plague. Newton mainly published his results related to optics during the first four decades of his life.

As is well known, the starting point of his investigations of prismatic colors was the problem of chromatic aberrations in lens telescopes. He was convinced that it was impossible to correct the chromatic aberrations of lenses, which led him to the construction of a mirror telescope. This brought him the Membership of the Royal Society.

In 1672 his revolutionary article “New Theory about Light and Colors” appeared in the “Philosophical Transactions”, the publication organ of the Royal Society. In this article Newton reports in a concise but lively style about his famous experiments on the decomposition of white light. His conclusion was that white light is composed of components of different colors and refractivities. These components cannot be further decomposed by a prism and cannot be modified by other means. However, it is possible to compose white light from these components. The colors of bodies can be explained by their varying reflection or absorption of the different colors contained in white light. His theory of an infinity of basic colors that cannot be viewed as a modification of the original white light was considered as revolutionary and led to fierce and controversial discussions.

Never again in his writings did Newton formulate his ideas as frankly and unprotected as he did in those thirteen pages. In his later publications, in particular in his main works, Newton preferred the unassailable but rigid form of a mathematical treatment with definitions, axioms, and propositions.

Three reasons may have been responsible for this caution: First, Newton tried to avoid scientific arguments, which he hated but with which he nevertheless saw himself confronted all the time. His relationship with Robert Hooke, for instance, consisted of a decade-long chain of wearing quarrels. Second, Newton had a general and constantly growing distrust in argumentation of the Cartesian type. Third, his position on the interpretation of physical phenomena often remained open because he himself could not make up his mind regarding a final opinion.

However, in a series of papers and lectures presented to the Royal Society around 1675/76, Newton cautiously commented on his view about the nature of light, emphasizing his aversion to hypotheses and clearly distinguishing facts from explanations.

The most important physical content of this works lies in the description of Newton’s rings, as they are called today. These are concentric colored rings that become visible when a slightly curved convex lens is placed onto a plane glass plate. Only much later did it become clear that these rings result from the interference of light waves that are reflected at the boundaries of the thin layer of air between the two glasses. For the coming 130 years, the precision of Newton’s observations was unparalleled; they later allowed the precise determination of the wavelengths of different colors. In this context Newton even surpassed Hooke whose special field was the color phenomena at thin layers.

From Newton’s explanations it seems obvious that at that time he considered the existence of an ether, through which the waves could propagate, as very likely. “If I had to assume a hypothesis as true,” he says, “it would be this one.” The ether is a medium with a large density outside of material bodies but only a diluted density inside where it is displaced by other forms of substance. The ether not only mediates the electric and magnetic forces but also the excitations traveling in nerve fibers and – particularly surprising for Newton – the gravitation between material bodies. As we shall see in a moment, the ether also has an influence on the

propagation of light. In order to fulfill all these different functions, the ether was thought to be a complicated mixture of several components.

Although the ether is able to vibrate, we should not, according to Newton's opinion, identify the nature of light with these ether undulations. The reason Newton mentions this is that for him the straight propagation of light is not compatible with a wave theory. Indeed, this was a fundamental problem at a time when a mathematical treatment of wave theory, the principle of interference, and, in particular, the method of Fourier analysis were not at hand. It was only 150 years later that Fresnel found a way out of this dilemma using his method of zone constructions.

Therefore, Newton saw himself forced to ascribe a corpuscular character to light. He explained the phenomena of refraction and total reflection by assuming that the light particles were dragged into spatial areas of diluted ether. According to his opinion, the velocity of light should be larger in transparent matter as compared to the ether between the bodies. In addition, violet light should consist of smaller particles because they are easier to deviate and therefore refracted more strongly. With these ideas, Newton opposed Fermat and Huygens.

When light particles traversed the interfaces between bodies or passed near such an interface, they would trigger undulations of the ether "like a stone thrown into water", and these denser and thinner parts of the ether, which propagate faster than light itself, had somehow to react back onto the motions of the light particles. Newton explicitly mentions the work of Grimaldi and explains the colored borders and fringes when light is refracted at edges or small objects as the interaction of light particles with excited ether waves.

This sophisticated theory of ether waves as a kind of "guiding waves" for light provided a qualitative and quantitative explanation for Newton's rings. Newton's only assumption was that the light particles, after hitting the interface, were either predominantly reflected or they passed through the interface, depending on whether they had hit a density peak or a density minimum of the excited ether waves. This also explained why light was sometimes reflected and sometimes diffracted at interfaces. The radius of Newton's rings allowed a precise determination of the wavelengths of the ether waves that were triggered by light particles of different colors.

Newton was well aware of the advantages inherent in the wave theory of light and he even proposed to relate the wavelengths of the ether undulations with the color of light. He knew, and he clearly states, that the color and dispersion phenomena of light as well as Newton's rings could be explained by both a wave theory and a corpuscular theory. As we have seen, it was the problem of understanding the straight-line propagation of light that led Newton to favor a corpuscular theory.

In 1704, almost 30 years later and one year after the death of Robert Hooke, Newton published his extensive work on the nature of light in his book "Opticks". The first Latin version followed in 1706, and the second English edition in 1717. Compared to his earlier writings, one finds characteristic changes in the form of representation as well as in his opinion.

For his "Opticks" he chooses the form of a mathematical treatment, like he did in the "Principia" from 1687. However, in contrast to the "Principia", he derives his results more by induction, starting from the phenomena and proceeding to their mathematical description. When Newton developed his theory of gravitation and experienced the sometimes nasty discussions about fundamental principles, his aversion to Cartesian-like explanations by hypotheses grew. In his famous "Scholium Generale" at the end of the "Principia", Newton refused

to supplement a hypothetical interpretation for his action at a distance, which describes the gravitational forces. The second book of the “*Principia*” is mainly devoted to a disproof of Descartes’ vortex theory of the motion of planets. The ether itself appears as an, at best, unnecessary auxiliary hypothesis, which he rather considers to be disproved by the observed phenomena. Already the first sentence in “*Opticks*” reads: “It is not my intention to explain in this book the properties of light by hypotheses, but only to state them and to confirm them by calculation and experiment.”

For the interaction between light and matter, observed in the reflection, diffraction, and refraction of light, he prefers a description “free of hypotheses” in terms of an action at a distance between the particles of light and matter. A derivation of Snell’s laws of refraction, assuming an attractive force between light and matter, is already contained in the fourteenth section of part I of the “*Principia*”. Once more, the velocity of light must be larger inside a medium than for the vacuum.

“*Opticks*” consists of three parts: Book I contains an extended and systematic description of Newton’s theory of colors, including his quantitative explanation for the colors of the rainbow.

Book II is devoted to the colors appearing in thin layers and he tries to use this theory also for an explanation of the colors of material bodies. The excellent observational data were already contained in the treatments of 1675/76. The explanation of these phenomena by guiding waves in the ether is replaced by a formal description, neutral with respect to hypothetical interpretations: The theory of “fits”, as he called them, the tendencies for a light particle to be reflected or to pass through an interface, is developed in propositions XII–XX of the second book. For instance, in proposition XII one finds: “Each light ray, when passing through a diffracting surface, acquires a certain property or disposition which reappears in the further course of the ray in equal intervals and by which it easily passes through the next diffracting surface, and is slightly reflected between each recurrence of this property.” Whether a light ray is reflected or diffracted at an interface depends on the momentary disposition of easy reflectibility or easy transmissivity. In proposition XIII Newton speculates that light might acquire these changing dispositions already with its emission. From the radius of Newton’s rings one can determine the recurrence intervals of these dispositions in air and water. For an explanation of these dispositions, Newton considers a hypothesis neither as necessary nor as proven, however, in proposition XII he mentions the model of ether undulations “for the reader”, beginning with a characteristic remark: “Those who are not willing to accept a new discovery unless it is explained by some hypothesis may assume, for the time being, that . . .” A description that is free of any hypotheses concerning the dispositions consists for Newton in undulations that are produced in material substances by forces at a distance between light and matter particles and then react onto the light.

In book III of “*Opticks*”, Newton describes his observations related to the refraction of light at small objects and edges. This is followed by an appendix where Newton poses some problems, which, in his opinion, are not yet sufficiently understood. He also includes general comments on the nature of light, its interaction with matter, and also on subjects of natural philosophy, which go far beyond the realm of optics. These famous “queries” allow a glimpse into Newton’s personal thoughts, interpretations, and doubts. From the way he poses the questions, it is usually not difficult to deduce the answers that Newton favored.

Newton's view about these subjects was everything else but settled, which manifests itself in the increasing number of queries added to each new edition, and by 1717 the edition of "Opticks" contained 130 queries. The first questions aim at an interpretation of the phenomena of reflection, refraction, and diffraction according to the ideas described above. Of special interest are the questions from 25 to 31, which were first included in the edition of 1706. The questions 25 and 26 describe double refraction of light in calcite. In this context Newton explains the phenomenon of polarization. For an explanation, Newton assigns a third quality to light, in addition to color and periodic dispositions, and which he calls the "sides" of a light ray.

The "sides" should be considered as a kind of orientation of a light ray that is perpendicular to the direction of propagation. A light ray hitting the surface of a calcite crystal may be diffracted in an ordinary or an extraordinary way, depending on the orientations of the sides relative to the crystal. A rotation by 90° around the direction of propagation will exchange ordinary and extraordinary diffraction. Newton explains his ideas by pointing out that the force between two magnets also depends on their relative orientation. His theory of sides already contains many essential ingredients of today's theory of polarization. However, with respect to a quantitative description of the extraordinary diffraction at calcite, Huygens' theory of wavelets remained superior.

In questions 27 and 28 Newton makes a thorough effort to falsify the wave theory of light. He emphasizes the straight-line propagation of light signals and compares it with the sound of a cannon or a bell, which could be heard also behind a small hill. And he summarizes all arguments disfavoring the existence of an ether. In question 29 he establishes his corpuscular theory of light: differences in color correspond to differences in the size of light particles; refraction, reflection, dispositions, and sides are explained by an action at a distance between light and matter.

Newton's enigmatic character reveals itself in that the 1717 edition of the "Opticks" contains new questions 17 to 24, where his ether theory from 1675 and 1676 reappears, including an ether hypothesis for gravitation, while at the same time the other questions rejecting and disproving the ether hypothesis still remain.

People have sometimes tried to mark Newton as an early pioneer of quantum mechanics, arguing with his aversion against hypotheses, his emphasis on the importance of observation, and his way to use the wave theory and corpuscular theory of light side by side. These attempts should be considered as inappropriate and unhistorical. But anyhow, Newton's free and cautious use of hypotheses differs much from the dogmatism of his successors, who certainly retarded the development of optics during the eighteenth century.

Today, in quantum theory, the relationship between wave and corpuscular theory has found a subtle solution in terms of complementarity. However, it would be wrong to consider the controversies between the respective proponents of wave and corpuscular theory as superfluous.

First, even from today's perspective we would use the wave picture for a proper description of reflection, refraction, diffraction, and double refraction; second, the propagation mechanism of light quanta, which is determined by absorption and re-emission, differs in many ways from the propagation of light particles, which should be slower in vacuum than in a medium; and third, the equation for the quantum field, which, within the framework of quantum electrodynamics, describes both the electromagnetic waves and the photons, is a wave equation.

With respect to the fundamentals of optics, nothing essential happened during the eighteenth century. However, we should mention two practical achievements: first, the construction of achromatic lenses, which Newton considered to be impossible, first in 1733 by the amateur Chester Moor Hall (1704–1770) and again in 1757 by John Dolland (1706–1761); and second, the discovery of aberration by James Bradley (1692–1762) in 1725. Aberration refers to the apparent change in the position of a star due to the movement of the Earth, and it was considered to be a confirmation of the corpuscular theory of light, because it found a simple and straightforward explanation within this framework.

The nineteenth century saw the second period of a rapid evolution of optics, at the end of which the nature of light was understood to be a transverse electromagnetic wave phenomenon. In 1801, the ingenious Thomas Young (1773–1829), who also contributed considerably to the deciphering of the hieroglyphs by suggesting the identification of the “cartouches” or “royal rings” around the kings’ names, formulated the principle of interference of waves and explained, as an immediate application, the diffraction of light. Building on the work of Newton and his at that time still unsurpassed accuracy of measurements, Young proposed a wave theoretic explanation of Newton’s rings together with a determination of the wavelengths of light. In 1809 Etienne Louis Malus (1775–1812) discovered the polarization of light in the reflections at mirrors, which led to a small crisis for the wave theory of optics. His observations led to the compelling conclusion that light, if really of wave-like character, must be transverse, which seemed to contradict the prevailing opinion according to which the propagation of light in a matter-free space was only possible for a longitudinal wave. For this reason even Malus himself judged his discovery to be a confirmation of the corpuscular theory. It took another eight years until Thomas Young, in 1817, ventured to interpret light as a transverse wave.

Essential progress in wave theory is due to Augustin Jean Fresnel (1788–1827). His zone construction, published in 1818, solved the long-standing problem of explaining the straight-line propagation of light. Together with Dominique Jean François Arago (1786–1853), he showed in 1819 that two light rays with perpendicular polarization planes do not interfere. Starting from a theory of transverse waves, not only was Fresnel able to derive the formulas which today bear his name and which allow the exact determination of the intensities for the reflected and refracted parts of light, but also he completed the subject of crystal optics as a theory of propagating transverse waves in anisotropic crystals. Of similar importance is his work on the theory of diffraction, which was further pursued by Joseph Fraunhofer (1787–1826), whose formal concepts were finally completed by Gustav Robert Kirchhoff (1824–1887). These successes of wave theory remained unmatched by any corpuscular theory of light. One of the last and tenacious proponents of corpuscular light theory was Jean Baptiste Biot (1774–1862). He looked for a mechanistic explanation of Newton’s fits in the form of prolonged and rapidly rotating light particles, which could penetrate through a surface if they hit the surface with their spiky heads, and which were reflected if they hit the surface with their flat sides. The final decision in favor of wave theory was the measurement of the velocity of light in water by Jean Bernard Léon Foucault (1819–1868), which in 1850 definitely proved that the speed of light inside a medium is slower than in the vacuum.

Based on the preliminary work of Michael Faraday (1791–1867), James Clerk Maxwell (1831–1879) derived his fundamental equations of electrodynamics, which imply the existence of transverse electromagnetic waves propagating with a fixed velocity, the velocity of light. The final experimental detection of these waves by Heinrich Rudolf Hertz (1857–1894) in 1888 made optics a branch of electrodynamics.

The theory of electrons developed by Hendrik Antoon Lorentz (1853–1928) allowed the explanation of optical properties of matter in terms of electromagnetic concepts. The derivation of Fresnel’s equations from electrodynamics is also due to Lorentz. In addition, we owe him essential contributions to the solution of the ether problem.

The final *coup de grâce* for the notion of an ether in its traditional form came in 1905, when Albert Einstein (1879–1955) formulated his theory of special relativity. The famous experiment by Albert Abraham Michelson (1852–1931) and Edward Williams Morley (1838–1923) did not reveal any measurable motion of the Earth with respect to the ether, as one would have expected according to the ether hypothesis.

Just when the final victory of the wave theory of light seemed complete, Max Planck (1858–1947) explained, in 1900, the spectral energy distribution of a black body using his quantum hypothesis. In 1905 Albert Einstein (1879–1955) took up the concept of energy quantization and applied it to the hitherto unexplained photoelectric effect. In his interpretation, Einstein went far beyond the ideas of Planck in that he described light as consisting of single energy quanta, so-called photons, thus assigning particle-like properties to light. The development of quantum theory at the beginning of the twentieth century finally led to a deeper understanding of the nature and the properties of light.

The discovery of the laser, the advance of the computer, the rapid development of holography and diffractive optics, the development of new materials, in particular materials with special nonlinear properties, as well as the sophistication and expansion of theoretical methods has led to a new revolution of optics during the past few decades. At present, optics can be considered as a particularly strong and growing branch of physics.

Table 1.1: Important people and events for the evolution of optics.

Euclid (about 300 BC)		“Katoptrik” (first scientific epos), general ideas about optics
Ibn al Haitham (AD 963–1039)		
R. Bacon (1214–1294)		discovery of the camera obscura
S. degli Armati	1299	discovery of spectacles
Z. Janssen	1600	the first microscope
H. Lippershey (1587–1619)	1608	construction of the first telescope
G. Galilei (1564–1642)	1609	telescope
J. Kepler (1571–1630)	1611	telescope

Table 1.1: continued

W. Snell (1591–1626)	1621	formulation of the law of refraction
R. Descartes (1596–1650)	1637	“La Dioptrique”, theory of the rainbow, law of refraction
P. de Fermat (1609–1665)	1657	derivation of the law of refraction, principle of temporally shortest path of light
F. Grimaldi (1618–1663)	1665	“Physicomathesis de Lumine, Coloribus et Iride”, discovery of diffraction
E. Bartholinus (1625–1698)	1670	discovery of double refraction
C. Huygens (1629–1695)	1678/90	“Traité de la Lumière”, wavelets, explanation of double refraction, discovery of polarization
R. Hooke (1635–1703)	1665	“Micrographia”, wave theory, colors of thin layers
I. Newton (1643–1727)	1668 1672 1675/76 1704	construction of the mirror telescope, “New Theory about Light and Colours”, lectures about Newton’s rings, “Opticks” (Latin 1706, 2nd English edn. 1717), mirror telescope, theory of colors, component theory of white light, colors of the rainbow, Newton’s rings, polarization, diffraction, corpuscular theory
O. Rømer (1644–1710)	1676	measurement of the speed of light
J. Bradley (1692–1762)	1725/28	stellar aberration of the light of fixed stars and its explanation
C. M. Hall (1704–1770)	1733	construction of achromatic lenses
J. Dolland (1706–1761)	1757	construction of achromatic lenses
F. W. Herschel (1738–1822)	1800	discovery of infrared radiation

Table 1.1: continued

J. W. Ritter (1776–1810)	1801	discovery of ultraviolet radiation
W. H. Wollaston (1766–1828)	1801	discovery of ultraviolet radiation
E. L. Malus (1775–1812)	1809	polarization by reflection
D. Brewster (1781–1858)	1815	Brewster's angle
T. Young (1773–1829)	1801 1817	development of interferometry, interpretation of light as a transverse wave
C. F. Gauss (1777–1855)		geometrical optics
J. Fraunhofer (1787–1826)	1814/15 1821/23	Fraunhofer lines, development of diffraction theory
J. A. Fresnel (1788–1827)	1818 1819 1821/22	Fresnel's zone construction, Fresnel's equations, development of diffraction theory
D. J. F. Arago (1786–1853)		polarization, color phenomena, interference
W. R. Hamilton (1805–1865)		geometrical optics, conical refraction
G. R. Kirchhoff (1824–1889)	1859 1883	spectral analysis, together with R. W. Bunsen, diffraction theory
R. W. Bunsen (1811–1899)	1859	spectral analysis, together with G. R. Kirchhoff
H. Fizeau (1819–1896)	1849	terrestrial measurement of velocity of light
J. B. L. Foucault (1819–1868)	1850	measurement of velocity of light in media
H. von Helmholtz (1821–1894)		theory of aberrations of optical instruments

Table 1.1: continued

E. Abbe (1840–1905)		theory of resolution of optical instruments
J. C. Maxwell (1831–1879)	1862	Maxwell's equations, fundamentals of electromagnetic light theory
L. G. Gouy (1854–1926)		phase change at caustics
A. Sommerfeld (1868–1951)		rigorous solution of diffraction problems
H. Hertz (1857–1894)	1888	detection of electromagnetic waves
W. C. Röntgen (1845–1923)	1895	X-rays
H. A. Lorentz (1853–1928)		electron theory of optical properties of matter
M. Planck (1858–1947)	1900	light quanta
A. Einstein (1879–1955)	1905	theory of photon effect, special relativity
A. A. Michelson (1852–1931)	1880	interferometer, Michelson–Morley experiment
D. Gabor (1900–1979)	1948	holography
T. H. Maiman (1927–)	1960	laser
C. H. Townes (1915–)	1964	Nobel prize for the development of lasers
A. M. Prochorow (1922–)		
N. G. Basow (1916–)		

2 The electrodynamics of continuous media

2.1 Maxwell's equations

The complete classical theory of electromagnetic fields is contained in Maxwell's equations. Along with Lorentz's equation for the electromagnetic force, they describe all the phenomena arising from interactions between electromagnetic fields and matter as long as quantum effects may be neglected. They also form the foundations for theoretical optics.

In a general form, independent of any given system of electromagnetic units, Maxwell's equations can be written in the form:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \quad (2.1)$$

$$\nabla \times \mathbf{E} = -\kappa \frac{\partial \mathbf{B}}{\partial t}, \quad (2.2)$$

$$\nabla \cdot \mathbf{B} = 0, \quad (2.3)$$

$$\nabla \times \mathbf{B} = \kappa \mu_0 \left(\mathbf{j} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right). \quad (2.4)$$

Here, $\mathbf{E}$ denotes the *electric field vector* (or *electric field strength*), $\mathbf{B}$ the *magnetic induction*, ρ the electric charge density, and $\mathbf{j}$ the electric current density. The force per unit volume acting on the electric charge density ρ and the electric current density $\mathbf{j}$ is given by the law for the *Lorentz force*:

$$\mathbf{f} = \rho \mathbf{E} + \kappa \mathbf{j} \times \mathbf{B}. \quad (2.5)$$

The constants ϵ_0 , κ , and μ_0 entering eqs. (2.1)–(2.18) can be understood in the following way.

- ϵ_0 has to be present in eq. (2.4), because, taking the divergence of eq. (2.4) and using eq. (2.1), one obtains

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{B} = 0. \quad (2.6)$$

- The constants κ in eqs. (2.2) and (2.5) have to be identical in order to guarantee the validity of the general law of induction. The induced tension in a conducting loop L is proportional to the change

$$\Phi = \frac{d}{dt} \int_L d\mathbf{f} \cdot \mathbf{B}$$

of the magnetic flux through the loop, and the change can be due either to a time variation of the magnetic induction or to a deformation of the loop. In this latter case, the induced tension is an effect of the Lorentz force.

- Looking at eqs. (2.1) and (2.5), we see that ϵ_0 fixes the units of the electric charge, charge density ρ , current density $\mathbf{j}$, and electric field $\mathbf{E}$.
- The unit of magnetic induction is either given via eq. (2.2) by fixing κ or via eq. (2.4) by fixing $\mu_0\kappa$.
- This double determination of the unit of $\mathbf{B}$ implies that there must be a relation between the constants ϵ_0 , μ_0 , and κ . Indeed, setting $\rho \equiv 0$ and $\mathbf{j} \equiv 0$, taking the rotation of eq. (2.2), and using eq. (2.4), one arrives at a wave equation:

$$\kappa^2 \mu_0 \epsilon_0 \frac{\partial^2}{\partial t^2} \mathbf{E} - \Delta \mathbf{E} = 0. \quad (2.7)$$

We see that

$$\kappa^2 \epsilon_0 \mu_0 = \frac{1}{c^2}, \quad (2.8)$$

where c denotes the speed of light in vacuum; ϵ_0 is called the *vacuum permittivity* and μ_0 the *magnetic permeability of vacuum*.

In particular, we should mention two commonly used systems of units:

- The SI system, also called the MKSA system, for which

$$\kappa = 1, \quad \mu_0 = 4\pi \times 10^{-7} \frac{\text{N}}{\text{A}^2}, \quad \epsilon_0 = \frac{1}{\mu_0 c^2}. \quad (2.9)$$

- The Gaussian system, also called the MKS system, for which

$$\kappa = \frac{1}{c}, \quad \mu_0 = 4\pi, \quad \epsilon_0 = \frac{1}{4\pi}.$$

In this system, $\mathbf{E}$ and $\mathbf{B}$ have the same dimension.

Equations (2.1) and (2.4) contain the electric charge density and current density, which is why they are referred to as the *inhomogeneous Maxwell's equations*, while the other two are called the *homogeneous Maxwell's equations*.

Maxwell's equations assume a slightly simpler form if we introduce two new fields, the *electric displacement density* $\mathbf{D}$ and the *magnetic field vector* $\mathbf{H}$:

$$\begin{aligned} \mathbf{D} &= \epsilon_0 \mathbf{E}, \\ \mathbf{H} &= \frac{1}{\mu_0} \mathbf{B}. \end{aligned} \quad (2.10)$$

This leads to:

$$\nabla \cdot \mathbf{D} = \rho, \quad (2.11)$$

$$\nabla \times \mathbf{E} = -\kappa \frac{\partial \mathbf{B}}{\partial t}, \quad (2.12)$$

$$\nabla \cdot \mathbf{B} = 0, \quad (2.13)$$

$$\nabla \times \mathbf{H} = \kappa \left(\mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \right). \quad (2.14)$$

The homogeneous equations remain unchanged, while the inhomogeneous equations now contain the new fields $\mathbf{D}$ and $\mathbf{H}$.

We shall see later that this form of Maxwell's equations also remains valid when we look at electrodynamics in continuous media. In this case, however, the fields $\mathbf{E}$, $\mathbf{B}$, ρ , and $\mathbf{j}$ should be interpreted as quantities that are averaged (or smeared) over volumes that are small compared to macroscopic scales but large compared to molecular scales. Furthermore, the fields $\mathbf{D}$ and $\mathbf{H}$ will no longer be given by eq. (2.10).

Finally, we want to draw some helpful conclusions from Maxwell's equations. Derivations may be found in any textbook on electrodynamics.

1. The energy density ρ^E and the energy flux density $\mathbf{S}$ of the electromagnetic field are given by

$$\begin{aligned} \rho^E &= \frac{\epsilon_0}{2} \mathbf{E}^2 + \frac{1}{2\mu_0} \mathbf{B}^2 = \frac{1}{2} (\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H}), \\ \mathbf{S} &= \frac{1}{\kappa\mu_0} \mathbf{E} \times \mathbf{B} = \frac{1}{\kappa} \mathbf{E} \times \mathbf{H}. \end{aligned} \quad (2.15)$$

2. The density ρ_i^P and the current density j_{ik}^P of the i th component of the momentum of the electromagnetic field may be expressed as

$$\rho_i^P = \kappa\epsilon_0\epsilon_{ijk}E_jB_k = \kappa\epsilon_{ijk}D_jB_k, \quad (2.16)$$

$$\begin{aligned} j_{ik}^P &=: -T_{ik} = \epsilon_0 \left(\frac{1}{2} \mathbf{E}^2 \delta_{ik} - E_i E_k \right) + \frac{1}{\mu_0} \left(\frac{1}{2} \mathbf{B}^2 \delta_{ik} - B_i B_k \right) \\ &= \frac{1}{2} \mathbf{E} \cdot \mathbf{D} \delta_{ik} - E_i D_k + \frac{1}{2} \mathbf{B} \cdot \mathbf{H} \delta_{ik} - B_i H_k. \end{aligned} \quad (2.17)$$

3. The homogeneous equations (2.2) and (2.3) can be solved by introducing a scalar potential Φ and a vector potential $\mathbf{A}$:

$$\begin{aligned} \mathbf{B} &= \nabla \times \mathbf{A}, \\ \mathbf{E} &= -\nabla \Phi - \kappa \frac{\partial \mathbf{A}}{\partial t}. \end{aligned} \quad (2.18)$$

For any function χ , the transformed potentials

$$\mathbf{A}' = \mathbf{A} + \nabla\chi, \quad \Phi' = \Phi - \kappa \frac{\partial\chi}{\partial t} \quad (2.19)$$

define the same fields $\mathbf{B}$ and $\mathbf{E}$. These gauge degrees of freedom may be used to impose on $\mathbf{A}$ and Φ the Lorentz gauge condition:

$$\nabla \cdot \mathbf{A} + \frac{1}{\kappa c^2} \frac{\partial\Phi}{\partial t} = 0. \quad (2.20)$$

Inserting eq. (2.18) into the inhomogeneous Maxwell's equations (2.1) and (2.4) yields

$$\begin{aligned} \square\Phi &:= \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) \Phi = \frac{\rho}{\epsilon_0}, \\ \square\mathbf{A} &:= \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) \mathbf{A} = \kappa\mu_0\mathbf{j}. \end{aligned} \quad (2.21)$$

The retarded solutions to these equations are

$$\begin{aligned} \Phi(t, \mathbf{x}) &= \frac{1}{4\pi\epsilon_0} \int d^3x' \frac{\rho(t - |\mathbf{x} - \mathbf{x}'|/c, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} \\ &= \frac{1}{\epsilon_0} \int dt' d^3x' G_{\text{ret}}(t - t', \mathbf{x} - \mathbf{x}') \rho(t', \mathbf{x}') \\ \mathbf{A}(t, \mathbf{x}) &= \frac{\kappa\mu_0}{4\pi} \int d^3x' \frac{\mathbf{j}(t - |\mathbf{x} - \mathbf{x}'|/c, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} \\ &= \kappa\mu_0 \int dt' d^3x' G_{\text{ret}}(t - t', \mathbf{x} - \mathbf{x}') \mathbf{j}(t', \mathbf{x}'), \end{aligned} \quad (2.22)$$

where

$$G_{\text{ret}} = \frac{1}{4\pi} \frac{1}{|\mathbf{x}|} \delta\left(t - \frac{1}{c}|\mathbf{x}|\right) \quad (2.23)$$

is the retarded Green's function for the wave operator $\square$.

2.2 Molecular vs. macroscopic fields

The subject of optics is the investigation of the interactions between light and matter. However, in bulk matter there always exist a huge number of charged particles. The number of particles per cm^3 ranges between roughly 6×10^{20} and 6×10^{23} , depending on whether the matter is in a gaseous or condensed state. Furthermore, due to the fast motion of the particles, the electromagnetic fields vary strongly as a function of space and time. A microscopic description of these *molecular fields* seems hopeless. Any information processing system would be cracked by the task of storing the data of 10^{20} particles. Furthermore, nobody could possibly be interested in all these data, which is why such a description is even undesirable.

On the other hand, what is observable and of general interest are the *macroscopic fields* obtained by averaging the molecular fields over space-time regions with a space-like extension of L and a time-like extension of $T > L/c$. Such averaging procedures eliminate the rapid field fluctuations present on microscopic scales and lead to macroscopic fields that are easier to observe and easier to handle analytically. In general, spatial averaging already implies averaging with respect to time-dependent fluctuations. The wavelengths of visible light are centered around 5×10^{-5} cm. For $L = 10^{-6}$ cm, a volume of $V = L^3$ still contains between 10^3 and 10^6 particles. Hence, even for wavelengths corresponding to the ultraviolet region, it is meaningful to take averages. Not before approaching wavelengths in the X-ray region do we have to adopt a different form of description.

Formally, the averaging of some molecular quantity A_{mol} to a macroscopic quantity $A = \langle A_{\text{mol}} \rangle$ is achieved by the following operation

$$\langle A_{\text{mol}}(t, \mathbf{x}) \rangle = A(t, \mathbf{x}) = \frac{1}{2T|V|} \int_{-T}^T dt' \int_V d^3x' A(t-t', \mathbf{x}-\mathbf{x}'), \quad (2.24)$$

where V denotes the spatial volume of averaging. More generally, we can choose an arbitrary function $f(t', \mathbf{x}')$ with the following properties: it vanishes outside the space-time volume V , it is almost constant inside this volume, it rapidly drops to zero close to the boundary of the volume, and it is subject to the normalization condition

$$\int dt' \int d^3x' f(t', \mathbf{x}') = 1. \quad (2.25)$$

For such a function we may define the average of a quantity by the following prescription:

$$\langle A_{\text{mol}}(t, \mathbf{x}) \rangle = A(t, \mathbf{x}) = \int dt' \int d^3x' f(t', \mathbf{x}') A_{\text{mol}}(t-t', \mathbf{x}-\mathbf{x}'). \quad (2.26)$$

This operation of taking averages commutes with linear operations like differentiation and convolution.

In particular, Maxwell's equations as well as eqs. (2.18)–(2.22) also hold for the averaged electromagnetic fields, charge and current densities, and potentials.

In general, two types of charge carriers will be present in a medium:

1. freely moving charges like conduction electrons in metals or ions in electrolytic solutions;
2. bounded charges whose motion is confined to the vicinity of some resting position by some binding force.

If an additional electric field is applied, the freely moving charged particles will travel through the medium. This produces an electric current and leads to electrical conductivity. Those charges which are bounded to their ionic cores will experience only a small displacement, resulting in an electric and magnetic *polarization* of the medium.

In the following sections, we shall first consider the case of free charges and later the case of bounded charges. We shall find that the equations for averaged quantities are comparatively simple, and the presence of the material medium becomes manifest only in certain material functions like conductivity, or electric and magnetic polarizability.

The requirement of causality for the response of a material medium to a disturbance exerted by an external field as well as a so-called passivity condition will lead to severe constraints on the possible form of the material functions.

First, we shall assume the medium to be at rest, i.e. the mean velocity of the bounded charges vanishes. Only in the last section of this chapter shall we drop this assumption.

2.3 A simple model for the electric current

The theory of electrical conductivity is part of solid-state physics and statistical mechanics. In the present context, we will restrain ourselves to a simple model developed around 1900 by Paul Drude in his attempts to describe electronic conductivity. If we want to describe electrolytic solutions, the freely moving electrons are to be replaced by ions. We will not discuss the (largely analogous) theory of ionic conductivity.

For the i th conduction electron carrying a charge e and being in an electric field $\mathbf{E}$, we assume the following equation of motion to hold:

$$m \frac{d\mathbf{v}_i}{dt} + \zeta \mathbf{v}_i = e\mathbf{E}, \quad (2.27)$$

where the influences of the ionic cores and the other conduction electrons have been taken into account globally by a friction term $\zeta \mathbf{v}_i$. The microscopic current density is given by

$$\mathbf{j}_{\text{mol}}(t, \mathbf{x}) = \sum_i e \mathbf{v}_i(t) \delta(\mathbf{x} - \mathbf{x}_i(t)). \quad (2.28)$$

Averaging over a volume of size L^3 leads to the macroscopic charge density

$$\mathbf{j}(t, \mathbf{x}) = en\mathbf{v}_D(t, \mathbf{x}) = \rho\mathbf{v}_D(t, \mathbf{x}), \quad (2.29)$$

where n denotes the number of conduction electrons per unit volume. Here $\rho = en$ is the macroscopic charge density and

$$\mathbf{v}_D = \frac{1}{nL^3} \sum_k \mathbf{v}_k \quad (2.30)$$

is the so-called *drift velocity*, an average velocity that is superimposed onto the random motion of the conduction electrons. The sum in eq. (2.30) extends over all conducting electrons in the volume L^3 where we take the average. If $\mathbf{E}$ is almost constant for scales of order L , the average of eq. (2.27) leads us to the *generalized Ohm's law*:

$$\frac{\partial}{\partial t} \mathbf{j} + \frac{\zeta}{m} \mathbf{j} = \frac{n e^2}{m} \mathbf{E} = \omega_p^2 \epsilon_0 \mathbf{E}. \quad (2.31)$$

For reasons that will become clear soon, the quantity

$$\omega_p = \sqrt{\frac{n e^2}{m \epsilon_0}} \quad (2.32)$$

is called the *plasma frequency*. For copper we find $\omega_p \cong 10^{16} \text{ s}^{-1}$. For a static field $\mathbf{E}_0$ the solution of eq. (2.31) yields the static current density

$$\mathbf{j}(\mathbf{x}) = \frac{ne^2}{\zeta} \mathbf{E}_0(\mathbf{x}) = \sigma_0 \mathbf{E}_0(\mathbf{x}). \quad (2.33)$$

Here σ_0 is called the *static conductivity*. For copper $(1/\epsilon_0)\sigma_0 \cong 5 \times 10^{17} \text{ s}^{-1}$. A simple argument shows us a connection between the coefficient of friction ζ and the mean free flying time τ . We assume that the field $\mathbf{E}$ constantly accelerates the conduction electrons, where $\mathbf{a} = e/m\mathbf{E}_0$, until they are stopped in a collision with other particles. The overall directed part of their velocity, which is superimposed onto their thermal agitation, has the form shown in fig. 2.1 and we obtain $\mathbf{v}_D = (e\tau/2m)\mathbf{E}_0$. Therefore, $\zeta/m = 2/\tau$ and

$$\sigma_0 = \frac{ne^2\tau}{2m} = \frac{1}{2}\epsilon_0\omega_p^2\tau. \quad (2.34)$$

For copper one finds $\tau \cong 10^{-14} \text{ s}$.

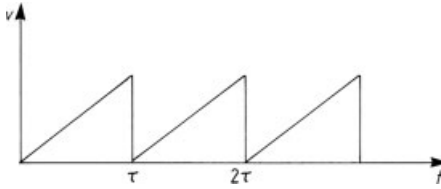


Figure 2.1: Illustration of the regular part of the velocity. The period of the graph determines the mean free flying time τ .

We obtain the standard form of Ohm's law for static potential differences $U = El$ as follows. For a wire with cross-section F and length l , the current I is given by

$$I = jF = \sigma_0 FE = \sigma_0 F \frac{U}{l}, \quad (2.35)$$

from which we get for the resistance $R = U/I = l/\sigma_0 F$. For currents and fields with a periodic time dependence, $\mathbf{E} = \mathbf{E}_0 e^{-i\omega t}$ and $\mathbf{j} = \mathbf{j}_0 e^{-i\omega t}$, the general form of Ohm's law (2.31) yields

$$\left(\frac{2}{\tau} - i\omega\right) \mathbf{j} = \epsilon_0\omega_p^2 \mathbf{E}$$

or

$$\mathbf{j} = \sigma(\omega) \mathbf{E} = \frac{\epsilon_0\tau\omega_p^2/2}{1 - i\omega\tau/2} \mathbf{E} = \frac{\sigma_0}{1 - i\omega\tau/2} \mathbf{E}. \quad (2.36)$$

For $\omega \ll 1/\tau \cong 10^{14} \text{ s}^{-1}$, the *dynamical conductivity* satisfies $\sigma(\omega) \cong \sigma_0$.

The physical meaning of the quantity ω_p will become clear when we take the divergence of Ohm's law in its generalized form

$$\frac{\partial \mathbf{j}}{\partial t} + \frac{2}{\tau} \mathbf{j} - \epsilon_0\omega_p^2 \mathbf{E} = 0.$$

We now use $\nabla \cdot \mathbf{j} = -\partial\rho/\partial t$ and $\nabla \cdot \mathbf{E} = \rho/\epsilon_0$ and find

$$\frac{\partial^2 \rho}{\partial t^2} + \frac{2}{\tau} \frac{\partial \rho}{\partial t} + \omega_p^2 \rho = 0.$$

Hence, the charge density inside the conductor dissolves exponentially with a decay time τ while it oscillates with the plasma frequency ω_p .

For not too large frequencies ($\omega \ll 10^{14} \text{ s}^{-1}$) we can write Maxwell's equations (2.1)–(2.4) describing the physics inside a conducting material in the form:

$$\begin{aligned} \nabla \cdot \mathbf{E} &= \frac{\rho}{\epsilon_0}, \\ \nabla \times \mathbf{E} &= -\kappa \frac{\partial \mathbf{B}}{\partial t}, \\ \nabla \cdot \mathbf{B} &= 0, \\ \nabla \times \mathbf{B} &= \kappa \mu_0 \sigma_0 \mathbf{E} + \frac{1}{\kappa c^2} \frac{\partial \mathbf{E}}{\partial t}. \end{aligned} \tag{2.37}$$

Application of the operator $\nabla \times$ for the case $\rho \equiv 0$ leads to

$$\begin{aligned} \square \mathbf{E} &:= \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) \mathbf{E} = -\frac{\sigma_0}{\epsilon_0 c^2} \frac{\partial \mathbf{E}}{\partial t}, \\ \square \mathbf{B} &:= \left(\frac{1}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) \mathbf{B} = -\frac{\sigma_0}{\epsilon_0 c^2} \frac{\partial \mathbf{B}}{\partial t}. \end{aligned} \tag{2.38}$$

This is the so-called *telegraph equation*.

For $\omega \ll \sigma_0/\epsilon_0$, we can neglect the term $(1/c^2)(\partial^2/\partial t^2)$ as compared to $(\sigma_0/\epsilon_0 c^2)(\partial/\partial t)$, and we obtain a diffusion equation:

$$\Delta \mathbf{E} = \frac{\sigma_0}{\epsilon_0 c^2} \frac{\partial \mathbf{E}}{\partial t}. \tag{2.39}$$

The solutions of these equations describe the behavior of field strengths inside conducting materials. As a simple example, we look for a solution of the form $E(t, x) = e^{-i\omega t} E(x)$ and find $\partial^2 E(x)/\partial x^2 = -i(\sigma_0 \omega / \epsilon_0 c^2) E(x)$, i.e.

$$E(x) = E_0 \exp \left(\pm \sqrt{\frac{\sigma_0 \omega}{\epsilon_0 c^2}} \frac{1-i}{\sqrt{2}} x \right).$$

The sign must be chosen in such a way that the solution decreases inside the conductor and we obtain

$$E(t, x) = E_0 \exp \left(-i\omega t - \frac{x}{d}(1-i) \right), \tag{2.40}$$

where

$$d = c \sqrt{\frac{2\epsilon_0}{\sigma_0 \omega}} \tag{2.41}$$

is the *penetration depth*. We see that the damping is stronger for large frequencies than for small frequencies. For very high-frequency fields, E and $\mathbf{j} = \sigma_0 \mathbf{E}$ can be neglected almost everywhere except at the surface of the conductor. This phenomenon is called the *skin effect*.

For copper the penetration depth d as a function of wavelength $\lambda = 2\pi c/\omega$ is summarized below:

λ	1 cm	1 m	100 m	10 km
d	2×10^{-4} mm	2×10^{-3} mm	0.02 mm	0.2 mm

If ω is large enough, such that the wire radius r satisfies $d \ll r$, the alternating current resistance becomes

$$\begin{aligned}
 E(x) &= E_0 \exp[-(x/d)(1-i)] \\
 \Rightarrow \quad I &\approx 2\pi r \sigma_0 \int_0^\infty dx E(x) = 2\pi r E_0 \sigma_0 d \frac{1}{1-i} \\
 &= U \frac{2\pi r \sigma_0 d}{l} \frac{1}{1-i},
 \end{aligned}$$

which implies

$$R = \frac{l(1-i)}{2\pi r \sigma_0 d} = R_0(1-i) \frac{r}{2d}, \quad (2.42)$$

where $R_0 = l/(\pi r^2 \sigma_0)$ is the *static resistance*; R is a complex quantity (with an additional inductivity), and, in contrast to the static case, $1/R$ is proportional not to the cross-sectional area but to the circumference.

The heat generated by the electric current per unit time and unit volume is given by $W = \mathbf{j} \cdot \mathbf{E}$, which for the static case becomes $W = \sigma_0 \mathbf{E}^2$.

2.4 Dispersion relations and the passivity condition

We used a simple model to derive the expression for the dynamical conductivity in eq. (2.36),

$$\sigma(\omega) = \frac{\sigma_0}{1 - i\omega\tau/2}. \quad (2.43)$$

In general, the relation between field strength $\mathbf{E}$ and current density $\mathbf{j}$ is more complicated. In the following, we will discuss a linear dependence between $\mathbf{E}$ and $\mathbf{j}$. Such behavior should be realized if the electric fields $\mathbf{E}$ are not too strong and if the conductivities are finite (i.e. not for the case of superconductors).

If $\mathbf{j}$ and $\mathbf{E}$ are spatially constant and the properties of the medium are time-independent, the most general linear relation between these two quantities is of the form

$$j_i(t) = \frac{1}{\sqrt{2\pi}} \int dt' \hat{\sigma}_{ik}(t-t') E_k(t'), \quad (2.44)$$

where $\hat{\sigma}_{ik}(t - t')$ is some matrix-valued function (or distribution). The time independence of the properties of the medium is expressed by the fact that $\hat{\sigma}_{ik}$ is only a function of the time difference, $t - t'$.

The current density $\mathbf{j}(t)$ can only depend on the field $\mathbf{E}(t')$ in the past, i.e. for $t' \leq t$, which leads to the *causality condition*:

$$\hat{\sigma}_{ij}(t - t') = 0 \quad \text{for} \quad t - t' < 0. \quad (2.45)$$

Furthermore, $\hat{\sigma}_{ij}$ is real-valued.

For the Fourier transforms,

$$\begin{aligned} j_i(t) &= \frac{1}{\sqrt{2\pi}} \int d\omega \tilde{j}_i(\omega) e^{-i\omega t}, \\ E_i(t) &= \frac{1}{\sqrt{2\pi}} \int d\omega \tilde{E}_i(\omega) e^{-i\omega t}, \\ \hat{\sigma}_{ik}(t) &= \frac{1}{\sqrt{2\pi}} \int d\omega \sigma_{ik}(\omega) e^{-i\omega t}, \end{aligned} \quad (2.46)$$

the convolution theorem yields the following relation between $\mathbf{j}$ and $\mathbf{E}$:

$$\tilde{j}_i(\omega) = \sigma_{ik}(\omega) \tilde{E}_k(\omega). \quad (2.47)$$

The reality condition for $\hat{\sigma}_{ik}$ implies, for the Fourier transform,

$$\sigma_{ik}(\omega) = \frac{1}{\sqrt{2\pi}} \int dt \hat{\sigma}_{ik}(t) e^{i\omega t}, \quad (2.48)$$

the condition

$$\sigma_{ik}^*(\omega) = \sigma_{ik}(-\omega). \quad (2.49)$$

Due to the causality condition, namely $\hat{\sigma}_{ij}(t) = 0$ for $t < 0$, the integral over $\hat{\sigma}_{ik}(t)$ extends only over positive time values, $t \geq 0$. This implies that, even for complex values of $z = \omega + i\eta$ with $\eta \geq 0$, the integral

$$\sigma_{ik}(z) = \frac{1}{\sqrt{2\pi}} \int_0^\infty dt \hat{\sigma}_{ik}(t) e^{i\omega t - \eta t} \quad (2.50)$$

exists and is a holomorphic function in the upper half-plane for $\eta = \text{Im } z > 0$. For the so-defined function $\sigma_{ik}(z)$, a generalized reality condition holds:

$$\sigma_{ik}^*(z) = \sigma_{ik}(-z^*). \quad (2.51)$$

Any function $\sigma_{ik}(z)$ holomorphic in the upper half-plane satisfies Cauchy's integral formula:

$$\sigma_{ik}(z) = \frac{1}{2\pi i} \int_C dz' \frac{\sigma_{ik}(z')}{z' - z}, \quad (2.52)$$

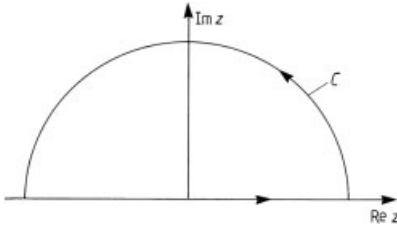


Figure 2.2: The integration contour C runs along the real axis and is closed by a semicircle in the upper half-plane, which in the limiting case becomes infinitely large.

where C denotes an integration contour as shown in fig. 2.2: C extends along the real axis and is closed by a semicircle in the upper half-plane, which in the end becomes infinitely large.

Furthermore, if $\sigma(z) \rightarrow 0$ for $|z| \rightarrow \infty$ (sufficiently fast) and $\text{Im } z \geq 0$, the contribution from the infinitely large semicircle vanishes and we obtain

$$\sigma_{ik}(z) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega' \frac{\sigma_{ik}(\omega')}{\omega' - z}. \quad (2.53)$$

On physical grounds the vanishing of $\sigma(z)$ for large values of z seems reasonable, because one would expect the current $\mathbf{j}$ to be negligible for infinitely high frequencies of $\mathbf{E}$, as the motion of the electrons can no longer follow the rapidly changing field. The expression we derived from our simple model,

$$\sigma(z) = \frac{\sigma_0}{1 - iz\tau/2},$$

has this feature and, in addition, for $\tau \geq 0$ satisfies the reality and the holomorphy conditions.

Writing $z = \omega + i\epsilon$ with $\epsilon > 0$ we recover the function $\sigma_{ik}(\omega)$ for real arguments in the limiting case $\epsilon \rightarrow 0$:

$$\sigma_{ik}(\omega) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} d\omega' \frac{\sigma_{ik}(\omega')}{\omega' - \omega - i\epsilon}. \quad (2.54)$$

(Expressions of this type are always to be understood in the limit $\epsilon \rightarrow 0$.)

The integrand contains the distribution

$$\frac{1}{\omega' - \omega - i\epsilon} = \frac{\omega' - \omega}{(\omega' - \omega)^2 + \epsilon^2} + i \frac{\epsilon}{(\omega' - \omega)^2 + \epsilon^2}. \quad (2.55)$$

It is easily verified that for an arbitrary test function φ the following identities hold:

$$\lim_{\epsilon \rightarrow 0} \int d\omega' \frac{\epsilon}{(\omega' - \omega)^2 + \epsilon^2} \varphi(\omega') = \pi \varphi(\omega) = \pi \int d\omega' \delta(\omega' - \omega) \varphi(\omega')$$

and

$$\begin{aligned} \lim_{\epsilon \rightarrow 0} \int d\omega' \frac{\omega' - \omega}{(\omega' - \omega)^2 + \epsilon^2} \varphi(\omega') &= \lim_{\epsilon \rightarrow 0} \left[\int_{-\infty}^{\omega - \epsilon} d\omega' \frac{\varphi(\omega')}{\omega' - \omega} + \int_{\omega + \epsilon}^{\infty} d\omega' \frac{\varphi(\omega')}{\omega' - \omega} \right] \\ &=: \int d\omega' \mathcal{P} \left(\frac{1}{\omega' - \omega} \right) \varphi(\omega'). \end{aligned}$$

The last equation defines the distribution $P(1/(\omega' - \omega))$, which is also called the *Cauchy principal value* of $1/(\omega' - \omega)$. For the divergent integral $\int d\omega' \varphi(\omega')/(\omega' - \omega)$, it tells us to take the integration limits symmetrically from both sides to the singularity at $\omega' = \omega$. The divergent part of the integral thus vanishes. The result of this calculation may be summarized in the well-known formula:

$$\frac{1}{\omega' - \omega - i\epsilon} = P\left(\frac{1}{\omega' - \omega}\right) + i\pi\delta(\omega' - \omega). \quad (2.56)$$

If we insert this formula into the integral representation of $\sigma_{ik}(\omega)$, eq. (2.54), and separate the real and imaginary parts, $\sigma_{ik}(\omega) = \text{Re } \sigma_{ik}(\omega) + i \text{Im } \sigma_{ik}(\omega)$, we obtain

$$\begin{aligned} \text{Re } \sigma_{ik}(\omega) &= \frac{1}{\pi} \int d\omega' P\left(\frac{1}{\omega' - \omega}\right) \text{Im } \sigma_{ik}(\omega') \\ \text{Im } \sigma_{ik}(\omega) &= -\frac{1}{\pi} \int d\omega' P\left(\frac{1}{\omega' - \omega}\right) \text{Re } \sigma_{ik}(\omega'). \end{aligned} \quad (2.57)$$

Because of these so-called *dispersion relations*, the function σ_{ik} is already determined by its real or imaginary part alone.

In the derivation of the dispersion relations we only made use of the causality condition and a sufficiently rapid decay of $\sigma(z)$ for large values of z in the upper half-plane. This last condition may even be relaxed when we divide through some function $f(z)$ (e.g. a polynomial function) and derive dispersion relations for $\sigma(z)/f(z)$.

Dispersion relations are of fundamental importance because they are intimately related to linear causal dependences. We shall see more examples soon. The dispersion relations are sometimes also called *Kramers–Kronig relations*.

The Joule heat $\mathbf{j} \cdot \mathbf{E}$ generated by the current per unit volume and unit time cannot be negative. Therefore, the integration over time yields

$$\int dt \mathbf{j}(t) \cdot \mathbf{E}(t) = \frac{1}{\sqrt{2\pi}} \int dt \int dt' E_i(t) \hat{\sigma}_{ik}(t - t') E_k(t') \geq 0. \quad (2.58)$$

This is the so-called *passivity condition* of the medium. It expresses the fact that, eventually, the medium can only absorb energy but not emit. The implication of the passivity condition for $\sigma_{ik}(\omega)$ is most easily expressed in terms of the Fourier transforms and follows when we chose the electric field to be periodic: $E_i(t) = \text{Re} \{E_i e^{-i\omega t}\}$. Taking the average over one period should result in a non-negative value for the Joule heat. Making use of the following identity,

$$\langle \text{Re} \{A e^{-i\omega t}\} \text{Re} \{B e^{-i\omega t}\} \rangle = \frac{1}{2} \text{Re} \{AB^*\}, \quad (2.59)$$

yields

$$\begin{aligned} \langle \text{Re} \{E_i e^{-i\omega t}\} \text{Re} \{\sigma_{ik}(\omega) E_k e^{-i\omega t}\} \rangle &= \frac{1}{2} \text{Re} \{E_i^* \sigma_{ik}(\omega) E_k\} \\ &= \frac{1}{4} \{E_i^* \sigma_{ik}(\omega) E_k + E_i \sigma_{ik}^*(\omega) E_k^*\} = \frac{1}{4} E_i^* (\sigma_{ik} + \sigma_{ki}^*) E_k \geq 0. \end{aligned} \quad (2.60)$$

Therefore, the matrix

$$\sigma_{ik} + \sigma_{ki}^* = \sigma_{ik} + \sigma_{ki}^*$$

must be positive. In general, one can show that because of *Onsager's symmetry relations* for irreversible thermodynamic processes, the matrix σ_{ij} has to be symmetric. This leads to the following form of the *passivity condition*:

$\text{Re } \sigma_{ik}$ is positive semidefinite.

The relation derived from our model,

$$\sigma_{ik}(\omega) = \delta_{ik} \frac{\sigma_0}{1 - i\omega\tau/2},$$

satisfies this condition for $\sigma_0 \geq 0$.

2.5 Electric displacement density and magnetic field strength

In addition to the conduction electrons, the influence of the molecular cores is also important. Although their charges cannot move freely, their presence gives rise to the phenomena of electric and magnetic polarization. Let

$$\phi_{\text{mol}}(t, \mathbf{x}) = \sum_{\nu} \phi_{\nu}(t, \mathbf{x}) \quad \text{and} \quad \mathbf{A}_{\text{mol}}(t, \mathbf{x}) = \sum_{\nu} \mathbf{A}_{\nu}(t, \mathbf{x}) \quad (2.61)$$

be the scalar and vector potentials, respectively, corresponding to the molecular cores. According to eqs. (2.22) and (2.23) the gauge potentials for a single molecular core, written in the Lorentz gauge, are given by

$$\begin{aligned} \phi_{\nu}(t, \mathbf{x}) &= \frac{1}{4\pi\epsilon_0} \int d^3x' \frac{\rho_{\nu}(t - |\mathbf{x} - \mathbf{x}'|/c, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}, \\ \mathbf{A}_{\nu}(t, \mathbf{x}) &= \frac{\kappa\mu_0}{4\pi} \int d^3x' \frac{\mathbf{j}_{\nu}(t - |\mathbf{x} - \mathbf{x}'|/c, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|}. \end{aligned} \quad (2.62)$$

Let $\mathbf{x}_{\nu}$ be the center of mass for molecule ν (see fig. 2.3). Expanding

$$|\mathbf{x} - \mathbf{x}'| = |\mathbf{x} - \mathbf{x}_{\nu} + \mathbf{x}_{\nu} - \mathbf{x}'| = |\mathbf{x} - \mathbf{x}_{\nu} - \mathbf{x}''|$$

up to first order in $\mathbf{x}'' = |\mathbf{x}' - \mathbf{x}_{\nu}|$, we obtain (see fig. 2.3)

$$\begin{aligned} 4\pi\epsilon_0\phi_{\nu}(t, \mathbf{x}) &= \frac{q_{\nu}}{|\mathbf{x} - \mathbf{x}_{\nu}|} - \mathbf{p}_{\nu}(t_{\nu \text{ ret}}) \cdot \nabla \frac{1}{|\mathbf{x} - \mathbf{x}_{\nu}|} + \frac{1}{c} \frac{\partial \mathbf{p}(t_{\nu \text{ ret}})}{\partial t} \cdot \frac{\mathbf{x} - \mathbf{x}_{\nu}}{|\mathbf{x} - \mathbf{x}_{\nu}|} + \dots, \\ \frac{4\pi}{\kappa\mu_0} \mathbf{A}_{\nu}(t, \mathbf{x}) &= \frac{\partial \mathbf{p}(t_{\nu \text{ ret}})/\partial t}{|\mathbf{x} - \mathbf{x}_{\nu}|} - \frac{1}{\kappa} \mathbf{m}_{\nu}(t_{\nu \text{ ret}}) \times \nabla \frac{1}{|\mathbf{x} - \mathbf{x}_{\nu}|} \\ &\quad + \frac{1}{\kappa c} \frac{\partial \mathbf{m}(t_{\nu \text{ ret}})}{\partial t} \times \frac{\mathbf{x} - \mathbf{x}_{\nu}}{|\mathbf{x} - \mathbf{x}_{\nu}|} + \dots, \end{aligned} \quad (2.63)$$

where we have used the following relations and definitions:

$$\begin{aligned}
 q_\nu &= \int d^3x' \rho_\nu(t, \mathbf{x}'), & \text{total charge of } \nu, \\
 \mathbf{p}_\nu(t) &= \int d^3x' (\mathbf{x}' - \mathbf{x}_\nu) \rho_\nu(t, \mathbf{x}'), & \text{electric dipole moment of } \nu, \\
 \mathbf{m}_\nu(t) &= \frac{\kappa}{2} \int d^3x' (\mathbf{x}' - \mathbf{x}_\nu) \times \mathbf{j}_\nu(t, \mathbf{x}'), & \text{magnetic dipole moment of } \nu, \\
 t_{\nu \text{ ret}} &= t - \frac{|\mathbf{x} - \mathbf{x}_\nu|}{c}, & \text{retarded time.}
 \end{aligned} \tag{2.64}$$

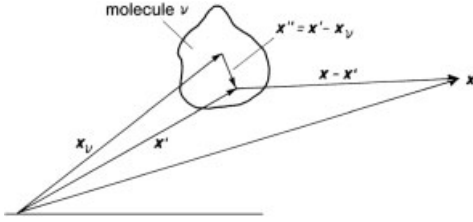


Figure 2.3: The coordinate system chosen for the determination of the electric displacement density and the magnetic field strength.

Furthermore, in the derivation of the expressions for ϕ_ν and $\mathbf{A}_\nu$ we have made use of the following identities:

$$\nabla_k(x_i j_k) = j_i - x_i \frac{\partial \rho}{\partial t} \quad \text{and} \quad \nabla_k(x_i x_r j_k) = x_i j_r + x_r j_i - x_i x_r \frac{\partial \rho}{\partial t}, \tag{2.65}$$

and neglected quadrupole terms and terms of higher order.

The molecular densities for the charge, the electric dipole moment, and the magnetic dipole moment are given by

$$\begin{aligned}
 \rho_{\text{mol}} &= \sum_\nu q_\nu \delta^{(3)}(\mathbf{x} - \mathbf{x}_\nu(t)) \\
 \mathbf{P}_{\text{mol}} &= \sum_\nu \mathbf{p}_\nu(t) \delta^{(3)}(\mathbf{x} - \mathbf{x}_\nu(t)) \\
 \mathbf{M}_{\text{mol}} &= \sum_\nu \mathbf{m}_\nu(t) \delta^{(3)}(\mathbf{x} - \mathbf{x}_\nu(t)).
 \end{aligned} \tag{2.66}$$

If, in the formulae for the molecular potentials,

$$\phi_{\text{mol}}(t, \mathbf{x}) = \sum_\nu \phi_\nu(t, \mathbf{x}) \quad \text{and} \quad \mathbf{A}_{\text{mol}}(t, \mathbf{x}) = \sum_\nu \mathbf{A}_\nu(t, \mathbf{x}),$$

we replace the fields ϕ_ν and $\mathbf{A}_\nu$ by the expressions of eq. (2.62), the averaging procedure needs special care. We do not necessarily obtain the molecular potentials ϕ_{mol} and $\mathbf{A}_{\text{mol}}$ by applying the retarded Green's functions G_{ret} to the molecular densities in eq. (2.66). Two mutually related problems turn up in this context:

1. We have

$$\int d^3x' \frac{\rho_{\text{mol}}(t - |\mathbf{x} - \mathbf{x}'|/c, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} = \sum_{\nu} \int d^3x' q_{\nu} \frac{\delta(\mathbf{x}' - \mathbf{x}_{\nu}(t - |\mathbf{x} - \mathbf{x}'|/c))}{|\mathbf{x} - \mathbf{x}'|}.$$

Integration over the δ -function yields a contribution at a value of $\mathbf{x}'$ for which the argument of the δ -function vanishes: $\mathbf{x}' = \mathbf{x}_{\nu}(t'_{\text{ret}})$, where $t'_{\text{ret}} = t - |\mathbf{x} - \mathbf{x}'|/c$. Because of the identity

$$\int d^3x' \mathbf{g}(\mathbf{x}') \delta(\mathbf{h}(\mathbf{x}')) = \frac{1}{|\det(\partial h_i / \partial x'_j)(\mathbf{x}'_0)|} \mathbf{g}(\mathbf{x}'_0) \quad \text{with} \quad \mathbf{h}(\mathbf{x}'_0) = 0,$$

we also have to take into account a determinant, which in our case is of the form

$$D = 1 + \mathcal{O}(v/c),$$

where v denotes the typical velocity of the molecular cores. As we assumed the medium to be at rest, we may write $D \approx 1$. Furthermore, the next-to-leading-order term, $\mathcal{O}(v/c)$, will not contribute to the temporal average because the number of cores entering a given volume is, in the average, equal to the number of cores leaving it.

2. As the δ -function in eq. (2.66) is time-dependent, we cannot simply conclude that

$$\partial \mathbf{P}_{\text{mol}} / \partial t = \sum_{\nu} \partial p_{\nu}(t) / \partial t \delta^{(3)}(\mathbf{x} - \mathbf{x}_{\nu}(t))$$

and

$$\mathbf{M}_{\text{mol}} = \sum_{\nu} \partial m_{\nu}(t) / \partial t \delta^{(3)}(\mathbf{x} - \mathbf{x}_{\nu}(t)).$$

However, the additional terms depending on the derivative of the δ -function are again suppressed by factors of v/c and cancel in the average.

In view of problems 1 and 2, we may write

$$\begin{aligned} \phi_{\text{mol}}(t, \mathbf{x}) &= \frac{1}{4\pi\epsilon_0} \int d^3x' \frac{\rho_{\text{mol}}(t'_{\text{ret}}, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} \\ &\quad + \frac{1}{4\pi\epsilon_0} \int d^3x' \left[\mathbf{P}_{\text{mol}}(t'_{\text{ret}}, \mathbf{x}') \cdot \nabla' \frac{1}{|\mathbf{x} - \mathbf{x}'|} \right. \\ &\quad \left. + \frac{1}{c} \frac{\partial \mathbf{P}_{\text{mol}}}{\partial t}(t'_{\text{ret}}, \mathbf{x}') \cdot \frac{\mathbf{x} - \mathbf{x}'}{|\mathbf{x} - \mathbf{x}'|} \right] + \delta\phi \\ &= \frac{1}{4\pi\epsilon_0} \int d^3x' \frac{\rho_{\text{mol}}(t'_{\text{ret}}, \mathbf{x}') - (\nabla' \cdot \mathbf{P}_{\text{mol}})(t'_{\text{ret}}, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} + \delta\phi, \end{aligned} \quad (2.67)$$

and similarly

$$\begin{aligned} \mathbf{A}_{\text{mol}}(t, \mathbf{x}) &= \\ &\frac{\kappa\mu_0}{4\pi} \int d^3x' \frac{(\partial \mathbf{P}_{\text{mol}} / \partial t)(t'_{\text{ret}}, \mathbf{x}') + (1/\kappa)(\nabla' \times \mathbf{M}_{\text{mol}})(t'_{\text{ret}}, \mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|} + \delta\mathbf{A}. \end{aligned} \quad (2.68)$$

Here, $\delta\phi$ and $\delta\mathbf{A}$ denote contributions that are small or cancel in the average.

The last steps in eqs. (2.67) and (2.68) involve a partial integration. Note that in the final expressions the operator ∇' , occurring in the terms $(\nabla' \cdot \mathbf{P}_{\text{mol}})$ and $(\nabla' \times \mathbf{M}_{\text{mol}})$, only acts on the second argument $\mathbf{x}'$, but not on the retarded time.

In addition to the charge density ρ_{mol} , the scalar potential ϕ_{mol} also contains a polarization density $\rho_{\text{mol}} = -(\nabla \cdot \mathbf{P}_{\text{mol}})$, which is due to the fact that during the process of polarization some charges may be expelled out of a volume element (see fig. 2.4, left).



Figure 2.4: (Left) Charges are expelled out of a volume element of the medium due to polarization. (Right) Representation of the magnetic dipole moments by a ring current.

Two polarization terms contribute to the electric current: an electric polarization current $\partial \mathbf{P}_{\text{mol}} / \partial t$, which corresponds to the displacement of charges in time-dependent polarization processes, and a magnetic polarization current $(1/\kappa)(\nabla \times \mathbf{M}_{\text{mol}})$, whose origin becomes obvious when we represent a magnetic dipole moment by a ring current (see fig. 2.4, right).

Taking the average of eqs. (2.67) and (2.68), we finally obtain the macroscopic potentials Φ and $\mathbf{A}$, expressed in terms of the macroscopic densities ρ , $\mathbf{j}$, $\mathbf{P} = \langle \mathbf{P}_{\text{mol}} \rangle$, and $\mathbf{M} = \langle \mathbf{M}_{\text{mol}} \rangle$:

$$\Phi(t, \mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int d^3r' \frac{\rho(t'_{\text{ret}}, \mathbf{r}') - (\nabla' \cdot \mathbf{P})(t'_{\text{ret}}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \quad (2.69)$$

$$\mathbf{A}(t, \mathbf{r}) = \frac{\kappa\mu_0}{4\pi} \int d^3r' \frac{\mathbf{j}(t'_{\text{ret}}, \mathbf{r}') + (\partial \mathbf{P} / \partial t)(t'_{\text{ret}}, \mathbf{r}') + (1/\kappa)(\nabla' \times \mathbf{M})(t'_{\text{ret}}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}.$$

In these expressions, we have also taken into account further contributions to the average charge and current density due to the charges and currents of additional conduction electrons that have been brought into the medium.

From these potentials we can derive the averaged electric and magnetic field strengths,

$$\mathbf{E} = -\nabla\Phi - \kappa \frac{\partial \mathbf{A}}{\partial t} \quad \text{and} \quad \mathbf{B} = \nabla \times \mathbf{A},$$

which automatically satisfy the homogeneous Maxwell's equations:

$$\nabla \times \mathbf{E} = -\kappa \frac{\partial \mathbf{B}}{\partial t} \quad \text{and} \quad \nabla \cdot \mathbf{B} = 0. \quad (2.70)$$

Furthermore, because of eq. (2.69), we obtain in the Lorentz gauge (2.20)

$$\begin{aligned} \square \Phi &= \nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0} (\rho - \nabla \cdot \mathbf{P}), \\ \square \mathbf{A} &= \nabla \times \mathbf{B} - \frac{1}{\kappa c^2} \frac{\partial \mathbf{E}}{\partial t} = \kappa\mu_0 \left(\mathbf{j} + \frac{\partial \mathbf{P}}{\partial t} \right) + \mu_0 \nabla \times \mathbf{M}. \end{aligned} \quad (2.71)$$

Introducing new field quantities

$$\mathbf{H} := \frac{1}{\mu_0} \mathbf{B} - \mathbf{M} \quad \text{and} \quad \mathbf{D} := \epsilon_0 \mathbf{E} + \mathbf{P}, \quad (2.72)$$

we can write the inhomogeneous Maxwell's equations in the form:

$$\nabla \times \mathbf{H} = \kappa \left(\mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \right) \quad \text{and} \quad \nabla \cdot \mathbf{D} = \rho. \quad (2.73)$$

We can see that, phenomenologically, we can take into account the influence of a medium by introducing new field quantities $\mathbf{D}$, the *electric displacement density*, and $\mathbf{H}$, the *magnetic field strength* (not to be confused with the *magnetic induction* $\mathbf{B}$). $\mathbf{P}$ and $\mathbf{M}$ are called the *electric* and *magnetic polarization (density)*, respectively. Note that with our convention, $\mathbf{E}$ is not equal to $\mathbf{D}$ nor is $\mathbf{B}$ equal to $\mathbf{H}$, even for vacuum, although $\mathbf{M} = \mathbf{P} = 0$. For the vacuum these identifications are sometimes made if Gaussian units of measurement are used. However, our definitions have the advantage that they are independent of the chosen unit system and that the inhomogeneous Maxwell's equations assume their simplest form.

For given $\mathbf{j}$ and ρ , the field quantities are determined by Maxwell's equations (2.70) and (2.73) only if the functional dependences of $\mathbf{D}$ and $\mathbf{H}$ on $\mathbf{E}$ and $\mathbf{B}$ are known:

$$\mathbf{D} = \mathbf{D}[\mathbf{E}, \mathbf{B}], \quad \mathbf{H} = \mathbf{H}[\mathbf{E}, \mathbf{B}].$$

In general, $\mathbf{D}$ and $\mathbf{H}$ may have complicated dependences on the global space-time behavior of the fields $\mathbf{E}$ and $\mathbf{B}$. However, for $\mathbf{E}$ and $\mathbf{B}$ not too large, we can expect, within a good approximation, the relations between $\mathbf{D}$, $\mathbf{H}$ and $\mathbf{E}$, $\mathbf{B}$ to be linear.

Under a parity transformation $\mathbf{x} \mapsto -\mathbf{x}$, the fields $\mathbf{D}$ and $\mathbf{E}$ change sign: $\mathbf{E}(\mathbf{x}) = -\mathbf{E}(-\mathbf{x})$, $\mathbf{D}(\mathbf{x}) = -\mathbf{D}(-\mathbf{x})$; whereas for $\mathbf{B}$ and $\mathbf{H}$ we have: $\mathbf{B}(\mathbf{x}) = \mathbf{B}(-\mathbf{x})$, $\mathbf{H}(\mathbf{x}) = \mathbf{H}(-\mathbf{x})$. Therefore, in a linear approximation, $\mathbf{D}$ will only depend on $\mathbf{E}$ and $\mathbf{H}$ will only depend on $\mathbf{B}$, at least, if the medium itself when averaged over small spatial regions is invariant under parity transformation $\mathbf{x} \mapsto -\mathbf{x}$. As long as we are allowed to assume the applied fields to be homogeneous over regions containing many molecular cores, we can neglect their spatial dependence and write the linear relation, as described in section 2.4, in the form

$$\begin{aligned} P_i(t) &= \frac{\epsilon_0}{\sqrt{2\pi}} \int dt' \hat{\chi}_{ik}(t-t') E_k(t'), & D_i(t) &= \epsilon_0 E_i(t) + P_i(t) \\ M_i(t) &= \frac{1}{\sqrt{2\pi}} \int dt' \hat{\chi}_{Mik}(t-t') H_k(t'), & B_i(t) &= \mu_0 (H_i(t) + M_i(t)). \end{aligned}$$

After a Fourier transformation we get:

$$\begin{aligned} \tilde{P}_i(\omega) &= \chi_{ik}(\omega) \epsilon_0 \tilde{E}_k(\omega), \\ \tilde{D}_i(\omega) &= \epsilon_{ik}(\omega) \epsilon_0 \tilde{E}_k(\omega), \\ \epsilon_{ik}(\omega) &= \delta_{ik} + \chi_{ik}(\omega), \end{aligned} \quad (2.74)$$

and

$$\begin{aligned} \tilde{M}_i(\omega) &= \chi_{Mik}(\omega) \tilde{H}_k(\omega), \\ \tilde{B}_i(\omega) &= \mu_{ik}(\omega) \mu_0 \tilde{H}_k(\omega), \\ \mu_{ik}(\omega) &= \delta_{ik} + \chi_{Mik}(\omega). \end{aligned} \quad (2.75)$$

The quantities $\chi_{ik}(\omega)$ and $\chi_{Mik}(\omega)$ are called the *electric* and *magnetic susceptibilities*, respectively, and ϵ_{ik} and μ_{ik} are the *relative electric* and *magnetic permeabilities*, respectively. Sometimes the expression *relative dielectric tensor* is also used for ϵ_{ik} .

For a stronger spatial dependence of the fields, the inhomogeneity may be taken into account by including also the first spatial derivatives of $\mathbf{E}$ and $\mathbf{H}$. Before we discuss the properties of the material quantities introduced above (χ , χ_M , ϵ , and μ), we want to draw some simple conclusions from Maxwell's equations for media:

$$\begin{aligned} \nabla \cdot \mathbf{D} &= \rho, & \nabla \cdot \mathbf{B} &= 0, \\ \nabla \times \mathbf{H} &= \kappa \left(\mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \right), & \nabla \times \mathbf{E} &= -\kappa \frac{\partial \mathbf{B}}{\partial t}. \end{aligned} \quad (2.76)$$

1. At an interface (with normal vector $\mathbf{N}$) between two different media, the quantities

$$\mathbf{N} \cdot \mathbf{B} \quad \text{and} \quad \mathbf{N} \times \mathbf{E},$$

the normal component of $\mathbf{B}$ and the tangential component of $\mathbf{E}$, are continuous. If the interface carries neither a surface charge nor a surface current, then

$$\mathbf{N} \cdot \mathbf{D} \quad \text{and} \quad \mathbf{N} \times \mathbf{H}$$

are also continuous.

2. If $\rho \equiv 0$ and $\mathbf{j} \equiv 0$, the energy balance contained in the fields and the medium assumes the form

$$\nabla \cdot \mathbf{S} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \mathbf{H} \cdot \frac{\partial \mathbf{B}}{\partial t} = 0. \quad (2.77)$$

Here, $\mathbf{S} = (1/\kappa)\mathbf{E} \times \mathbf{H}$ denotes the Poynting vector. For this case the continuity conditions at the interface also imply the continuity of $\mathbf{N} \cdot \mathbf{S}$. No energy is assembled in the interface.

3. If $\rho \equiv 0$ and $\mathbf{j} \equiv 0$, the momentum balance reads

$$\frac{\partial}{\partial t} \rho_i^P - \nabla_j T_{ij} = 0, \quad (2.78)$$

where $\rho_i^P = \kappa(\mathbf{D} \times \mathbf{B})_i$ is the momentum density and

$$T_{ij} = E_i D_j + H_i B_j - \frac{1}{2}(\mathbf{D} \cdot \mathbf{E})\delta_{ij} - \frac{1}{2}(\mathbf{H} \cdot \mathbf{B})\delta_{ij} \quad (2.79)$$

is Maxwell's stress tensor. Here, we have assumed the relations between $\mathbf{D}$ and $\mathbf{E}$ and between $\mathbf{H}$ and $\mathbf{B}$ to be independent of $\mathbf{x}$.

2.6 Index of refraction and coefficient of absorption

This section contains a first discussion of the propagation of waves in media. The simplest but most important case is an isotropic medium, for which we have $\epsilon_{ij}(\omega) = \delta_{ij}\epsilon(\omega)$ and $\mu_{ij}(\omega) = \delta_{ij}\mu(\omega)$. Furthermore, we assume $j_i(\omega) = \sigma(\omega)E_i(\omega)$ and $\rho(\omega) = 0$. Given these conditions we look for wave-like solutions of Maxwell's equations (2.76). We make the following ansatz,

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}, & \mathbf{B} &= \mathbf{B}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}, \\ \mathbf{D} &= \mathbf{D}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}, & \mathbf{H} &= \mathbf{H}_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}, \end{aligned} \quad (2.80)$$

an insert it into Maxwell's equations, obtaining

$$\begin{aligned} \mathbf{k} \cdot \mathbf{D} &= 0, & \mathbf{k} \cdot \mathbf{B} &= 0, \\ \mathbf{k} \times \mathbf{H} &= -\kappa\omega\epsilon_0\eta\mathbf{E}, & \mathbf{k} \times \mathbf{E} &= \kappa\omega\mathbf{B}, \end{aligned} \quad (2.81)$$

where

$$\eta(\omega) = \epsilon(\omega) + \frac{i\sigma(\omega)}{\epsilon_0\omega}. \quad (2.82)$$

Now the conductivity contributes to the complex generalized electric permeability η . For very large frequencies, the distinction between free and bound charges becomes less obvious. In general, $\mathbf{k}$ will be complex for real ω , so that for some directions the waves will decay exponentially.

Furthermore, we have

$$\begin{aligned} \mathbf{k} \times (\mathbf{k} \times \mathbf{E}) &= \kappa\omega\mathbf{k} \times \mathbf{B} = \kappa\omega\mu\mu_0\mathbf{k} \times \mathbf{H} = -\mu\eta\frac{\omega^2}{c^2}\mathbf{E}, \\ \mathbf{k} \times (\mathbf{k} \times \mathbf{E}) &= \mathbf{k} \cdot (\mathbf{k} \cdot \mathbf{E}) - k^2\mathbf{E} = -k^2\mathbf{E}, \end{aligned}$$

from which

$$k^2 = \frac{\omega^2}{c^2}\eta\mu, \quad \text{i.e.} \quad k = \frac{\omega}{c}\sqrt{\eta\mu} = \frac{\omega}{c}p(\omega). \quad (2.83)$$

We decompose $p(\omega)$ with respect to its real and imaginary parts:

$$p(\omega) = \sqrt{\mu\eta(\omega)} = n(\omega) + i\kappa(\omega). \quad (2.84)$$

Without loss of generality, we can choose $\omega > 0$ and consider the propagation of waves in the z direction. The sign of the square root in (2.84) should be such that $n(\omega) > 0$. For this case $\mathbf{E}(t, z)$ assumes the form

$$\begin{aligned} \mathbf{E}(t, z) &= \mathbf{E}_0 \exp(-i\omega t + i\kappa z) \\ &= \mathbf{E}_0 \exp\left(-i\omega t + \frac{i\omega}{c}n(\omega)z - \frac{\omega}{c}\kappa(\omega)z\right) \\ &= \mathbf{E}_0 \exp\left[+i\frac{\omega}{c}n(\omega)\left(z - \frac{ct}{n(\omega)}\right)\right] \exp\left(-\frac{\omega}{c}\kappa(\omega)z\right). \end{aligned} \quad (2.85)$$

This describes an exponentially damped wave propagating in the positive z direction with a phase velocity of $c_{\text{ph}}(\omega) = c/n(\omega)$. The real part $n(\omega)$ of eq. (2.84)

$$n(\omega) = \frac{c}{c_{\text{ph}}(\omega)} \quad (2.86)$$

is called the *index of refraction* or *refractive index* of the medium. The imaginary part $\kappa(\omega)$ is related to the coefficient of absorption $\alpha(\omega)$ of the medium:

$$\alpha(\omega) = \frac{\omega}{c} \kappa(\omega). \quad (2.87)$$

Now we want to convince ourselves that indeed $\kappa(\omega) > 0$. Essentially, this is a consequence of the passivity condition for the conductivity,

$$\text{Re } \sigma(\omega) \geq 0,$$

as well as the passivity conditions

$$\text{Im } \omega \mu(\omega) \geq 0 \quad \text{and} \quad \text{Im } \omega \epsilon(\omega) \geq 0$$

for $\mu(\omega)$ and $\epsilon(\omega)$, which we shall derive in the next section. These conditions imply that the imaginary parts of both $\eta(\omega)$ and $\mu(\omega)$ are positive. Thermodynamic stability even allows us to deduce that $\eta(\omega)$ and $\mu(\omega)$ must lie in the first quadrant of the complex plane, and in this case $\sqrt{\eta\mu(\omega)}$ must also lie in the first quadrant. Many of the substances that are important in optics even have $\mu(\omega) \approx 1$, so that $p(\omega) = n(\omega) + i\kappa(\omega) \approx \sqrt{\eta(\omega)}$ lies in the first quadrant due to the passivity conditions.

These arguments may be generalized to anisotropic linear media, for which the conductivity and the permeabilities are tensors. Also in these materials the real part of the conductivity $\sigma_{ij}(\omega)$ and the imaginary parts of $\epsilon_{ij}(\omega)$ and $\mu_{ij}(\omega)$ contribute to the absorption.

In any resistor we find energy dissipation due to heating and, therefore, it seems obvious that conductivity implies absorption. A certain degree of absorption cannot be avoided in a material medium, because the real and imaginary parts of $\sigma_{ij}(\omega)$, $\mu_{ij}(\omega)$, and $\epsilon_{ij}(\omega)$ cannot be chosen arbitrarily but are related through dispersion relations – see also the next section for a discussion of $\mu_{ij}(\omega)$ and $\epsilon_{ij}(\omega)$.

Depending on the relevance of absorption one distinguishes two disciplines in optics:

- (A) **Metal optics:** Absorption due to conductivity is predominant and the imaginary parts of $\mu_{ij}(\omega)$ and $\epsilon_{ij}(\omega)$ are large. Wave propagation inside these media is not possible and, therefore, the subject of metal optics mainly deals with surface phenomena or thin layers of absorbing materials. In this book we will not enter into the subject of metal optics.
- (B) **Optics of transparent media:** In these media, conductivity and absorption are negligible and wave propagation is possible.

Looking at the optics of transparent media, the following distinction turns out to be useful:

- B1) *Homogeneous and isotropic media:* their treatment is simple and almost covered by the considerations in this chapter.

- B2) *Homogeneous and anisotropic media*: they are the subject of crystal optics, which we will discuss in chapter 4.
- B3) *Inhomogeneous and isotropic media*: they are the subject of geometrical optics and the theory of diffraction – their theoretical treatment is one of the central themes of this book.
- B4) *Inhomogeneous and anisotropic media*: they turn out to be of less practical importance, and their general theoretical treatment is difficult and still not completely worked out – in this book, we will only consider the most important special cases of reflection and refraction at an interface between two anisotropic homogeneous media.

2.7 The electromagnetic material quantities

In the following discussion of the quantities

$$\epsilon_{ik}(\omega) = \delta_{ik} + \chi_{ik}(\omega) \quad \text{and} \quad \mu_{ik}(\omega) = \delta_{ik} + \chi_{M ik}(\omega),$$

we will concentrate on the electric susceptibility $\chi_{ik}(\omega)$. The properties of the magnetic susceptibility $\chi_{M ik}(\omega)$ are similar and, as long as we neglect ferromagnetic substances, the values of $\chi_{M ik}(\omega)$ are much smaller than the values of the electric susceptibility. Ferromagnets are not transparent and are of no relevance for the optics of transparent media. For simplicity, we will assume in the following that $\mu(\omega) \equiv 1$.

The electric polarization

$$P_i(t) = \frac{\epsilon_0}{\sqrt{2\pi}} \int dt' \hat{\chi}_{ik}(t-t') E_k(t') \quad (2.88)$$

can only depend on $E_k(t')$ for $t' \leq t$. As the discussion about conductivity in section 2.4 revealed, this causality condition implies that

$$\chi_{ij}(z) = \frac{1}{\sqrt{2\pi}} \int dt \exp(izt) \hat{\chi}_{ij}(t)$$

is holomorphic in the upper z plane. Furthermore, we would expect that for large frequencies the polarization is small, which leads us to

$$\lim_{|z| \rightarrow \infty, \text{Im } z \geq 0} \chi_{ij}(z) = 0. \quad (2.89)$$

In section 2.4 we saw that from this condition we can derive the dispersion relations:

$$\begin{aligned} \text{Re } \chi_{ik}(\omega) &= \frac{1}{\pi} \int d\omega' P \left(\frac{1}{\omega' - \omega} \right) \text{Im } \chi_{ij}(\omega') \\ \text{Im } \chi_{ik}(\omega) &= -\frac{1}{\pi} \int d\omega' P \left(\frac{1}{\omega' - \omega} \right) \text{Re } \chi_{ij}(\omega'), \end{aligned} \quad (2.90)$$

and from eq. (2.90) we get

$$\begin{aligned}
 \chi_{ik}(\omega) &= \text{Re } \chi_{ik}(\omega) + i \text{Im } \chi_{ik}(\omega) \\
 &= \frac{1}{\pi} \int d\omega' \left[\text{P} \left(\frac{1}{\omega' - \omega} \right) + i\pi\delta(\omega' - \omega) \right] \text{Im } \chi_{ik}(\omega') \\
 &= \frac{1}{\pi} \int d\omega' \frac{\text{Im } \chi_{ik}(\omega')}{\omega' - \omega - i\epsilon}.
 \end{aligned} \tag{2.91}$$

As $\hat{\chi}_{ik}$ is real, we have $\chi_{ik}(\omega)^* = \chi_{ik}(-\omega)$. Therefore, $\text{Im } \chi_{ik}(\omega)$ is an odd function of ω and $\text{Re } \chi_{ik}(\omega)$ is an even function. If we make use of these properties, the dispersion relations assume the form:

$$\text{Re } \chi_{ik}(\omega) = \frac{1}{\pi} \int_0^\infty d\omega' \text{P} \left(\frac{2\omega'}{\omega'^2 - \omega^2} \right) \text{Im } \chi_{ik}(\omega'). \tag{2.92}$$

In section 2.6 we saw that the dispersion relations establish a connection between the frequency dependence of the refractive index and the absorption of a substance. This explains the expression “dispersion relation”.

In order to derive a passivity condition for $\chi_{ik}(\omega)$, we start from the balance relation satisfied by the energy

$$\nabla \cdot \mathbf{S} + \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} + \frac{1}{2\mu_0} \frac{\partial \mathbf{B}^2}{\partial t} = 0.$$

Because $\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$, we may rewrite this equation in the form

$$-\frac{d}{dt} \int_V d^3x \left(\frac{\epsilon_0}{2} \mathbf{E}^2 + \frac{1}{2\mu_0} \mathbf{B}^2 \right) = \int_{\partial V} \mathbf{S} \cdot d\mathbf{f} + \int_V d^3x \mathbf{E} \cdot \frac{\partial \mathbf{P}}{\partial t}. \tag{2.93}$$

From this equation we learn that the electromagnetic energy of the fields inside a volume V decreases either by emission of radiation or by polarizing the medium. We would expect that, if we apply a time-dependent electric or magnetic field that vanishes for $t \rightarrow \pm\infty$, the medium can only absorb but not release energy (otherwise it would not have been in a stable equilibrium state). This leads us to the passivity condition

$$\int_{-\infty}^{\infty} \mathbf{E}(t) \cdot \frac{\partial \mathbf{P}(t)}{\partial t} dt \geq 0. \tag{2.94}$$

The equality sign holds if and only if there is no absorption and/or dissipation of electromagnetic field energy. Written in terms of the periodic fields

$$E_i(t) = \text{Re}\{E_i e^{-i\omega t}\}, \quad \text{and} \quad P_i(t) = \text{Re}\{\chi_{ik}(\omega)\epsilon_0 E_k e^{-i\omega t}\}$$

so that

$$\frac{\partial P_i(t)}{\partial t} = \text{Re}\{\chi_{ik}(\omega)(-i\omega)\epsilon_0 E_k e^{-i\omega t}\},$$

we obtain the passivity condition in Fourier space after taking the time average,

$$\begin{aligned} \operatorname{Re}\{E_i^* \chi_{ik}(\omega)(-i\omega)E_k\} &= \operatorname{Im}\{E_i^* \omega \chi_{ik}(\omega)E_k\} \\ &= E_i^* \{(\omega/2i)[\chi_{ik}(\omega) - \chi_{ki}^*(\omega)]\}E_k \geq 0 \end{aligned} \quad (2.95)$$

(compare the derivation of eq. (2.60)). Hence, $(\omega/2i)[\chi_{ik}(\omega) - \chi_{ki}^*(\omega)]$ is a positive semidefinite matrix, and the matrix χ_{ik} is Hermitian for case that there is no absorption.

In an isotropic medium, we have $\chi_{ik}(\omega) = \chi(\omega)\delta_{ik}$, and the passivity condition reads

$$\operatorname{Im} \omega \chi(\omega) \geq 0. \quad (2.96)$$

This condition is compatible with the reality condition $\operatorname{Im} \chi(-\omega) = -\operatorname{Im} \chi(\omega)$, from which we can deduce the even stronger statement that

$$\operatorname{Im} z \chi(z) \geq 0 \quad \text{for} \quad \operatorname{Im} z \geq 0 \quad (2.97)$$

in the upper half-plane.

This result follows from the integral representation

$$\chi(z) = \frac{1}{\pi} \int d\omega' \frac{\operatorname{Im} \chi(\omega')}{\omega' - z}, \quad (2.98)$$

which leads to

$$\begin{aligned} \operatorname{Im} z \chi(z) &= \frac{1}{\pi} \int d\omega' \operatorname{Im} \chi(\omega') \operatorname{Im} \left\{ \frac{z}{\omega' - z} \right\} \\ &= \frac{1}{\pi} \int d\omega' \operatorname{Im} \chi(\omega') \frac{\omega' \operatorname{Im} z}{|\omega' - z|^2} \geq 0 \quad \text{for} \quad \operatorname{Im} z \geq 0. \end{aligned} \quad (2.99)$$

So, $z\chi(z)$ is holomorphic in the (open) upper half-plane H^+ and maps H^+ onto itself. Such functions are called *Pick functions*. If $f(z)$ and $g(z)$ are Pick functions, so are $f(z) + g(z)$ and $f(g(z))$.

So far we have concentrated on the behavior of $\chi_{ik}(z)$ as a function of the complex argument z . However, we can also make some general statements about the symmetries of the matrix $\chi_{ik}(z)$ with respect to the indices i and k . The static case $\omega = 0$ is particularly simple.

Due to the reality condition,

$$\chi_{ik}^*(\omega) = \chi_{ik}(-\omega),$$

$\chi_{ik}(0)$ is real. Furthermore, already from thermodynamic arguments we obtain the symmetry

$$\chi_{ik}(0) = \chi_{ki}(0). \quad (2.100)$$

Indeed, following our arguments in connection with the discussion of the passivity condition, we may conclude on the basis of the energy balance that the change in internal energy U of a medium, as a response to changes in the thermodynamic variable $\mathbf{P}$, the entropy S , and maybe further variables X_α , is of the form

$$dU = T dS + \mathbf{E} \cdot d\mathbf{P} + \sum_{\alpha} \xi_{\alpha} dX_{\alpha}. \quad (2.101)$$

For the free energy $F = U - TS$ we find

$$dF = -S dT + \mathbf{E} \cdot d\mathbf{P} + \sum_{\alpha} \xi_{\alpha} dX_{\alpha}. \quad (2.102)$$

This implies that

$$E_i = \left(\frac{\partial F}{\partial P_i} \right)_{T,X}. \quad (2.103)$$

The linear theory considered before yields

$$\epsilon_0 E_i = \chi_{ik}^{-1}(0) P_k$$

and therefore

$$\frac{1}{\epsilon_0} \chi_{ik}^{-1}(0) = \frac{\partial E_i}{\partial P_k} = \left(\frac{\partial^2 F}{\partial P_i \partial P_k} \right)_{T,X} = \frac{1}{\epsilon_0} \chi_{ki}^{-1}(0). \quad (2.104)$$

This proves the symmetry of χ_{ik}^{-1} and, hence, of χ_{ik} . Strictly speaking, one has to distinguish the susceptibility at constant temperature and the susceptibility at constant entropy, which is given by

$$\frac{1}{\epsilon_0} \chi_{ik}^{-1S}(0) = \left(\frac{\partial^2 U}{\partial P_i \partial P_k} \right)_{S,X}. \quad (2.105)$$

Both are symmetric. The requirement of thermodynamic stability implies that $\epsilon_{ik}(0)$ is positive definite.

For the non-static case we have $\omega \neq 0$, and the symmetry of $\chi_{ik}(\omega)$ can be derived from the more general symmetry principles for the kinetic coefficients of Onsager and Casimir, which follow from microscopic time reversion invariance. Therefore,

$$\chi_{ik}(\omega) = \chi_{ki}(\omega), \quad (2.106)$$

and the passivity condition simplifies to the statement that

$$\text{Im } \omega \chi_{ik}(\omega) \text{ is positive semidefinite.} \quad (2.107)$$

In particular, χ_{ik} has to be real for vanishing absorption.

If χ_{ik} also depends on an external magnetic field or on some other parameter that changes sign under time reversion, the Onsager–Casimir relations require

$$\chi_{ik}(\mathbf{B}, \omega) = \chi_{ki}(-\mathbf{B}, \omega). \quad (2.108)$$

For the absorption-free case we find

$$\chi_{ik}(\mathbf{B}, \omega) = \chi_{ki}^*(\mathbf{B}, \omega) = \chi_{ik}^*(-\mathbf{B}, \omega),$$

and therefore

$$\begin{aligned} \text{Re } \chi_{ik}(\mathbf{B}, \omega) &= \text{Re } \chi_{ik}(-\mathbf{B}, \omega), \\ \text{Im } \chi_{ik}(\mathbf{B}, \omega) &= -\text{Im } \chi_{ik}(-\mathbf{B}, \omega). \end{aligned} \quad (2.109)$$

2.8 The oscillator model for the electric susceptibility

As we did for the electric conductivity, we now want to formulate a simple microscopic model for the electric susceptibility. For this purpose we imagine the electrons of a molecular core to be bounded elastically to their equilibrium position. Again, the interaction with other charge carriers is effectively described by a friction term. Under these conditions, the equation of motion of an electron in an external electric field $\mathbf{F}$ becomes

$$m \left(\frac{\partial^2 \mathbf{x}}{\partial t^2} + 2\gamma \frac{\partial \mathbf{x}}{\partial t} + \omega_0^2 \mathbf{x} \right) = e\mathbf{F}, \quad (2.110)$$

which for a periodic field, $\mathbf{F} = \mathbf{F}_0 e^{-i\omega t}$, leads to the solution

$$\mathbf{x} = \frac{e}{m} \frac{1}{\omega_0^2 - \omega^2 - 2i\omega\gamma} \mathbf{F}.$$

The total dipole moment, $\mathbf{p} = \sum_{i=1}^Z e\mathbf{x}_i$, of Z electrons in a molecular core is now given by

$$\mathbf{p} = \frac{Ze^2}{m} \frac{1}{\omega_0^2 - \omega^2 - 2i\omega\gamma} \mathbf{F} =: \alpha(\omega)\epsilon_0 \mathbf{F}. \quad (2.111)$$

Here $\alpha(\omega)$ is the *molecular polarizability*. For ω close to ω_0 , this quantity is large and purely imaginary. This rather crude oscillator model may be improved by a quantum-mechanical calculation where the periodic field $\mathbf{F}$ is treated as a time-dependent disturbance:

$$\alpha(\omega) = \frac{1}{\epsilon_0} \frac{e^2}{m} \sum_{r=1}^Z \frac{f_r}{\omega_r^2 - \omega^2 - 2i\omega\gamma_r}. \quad (2.112)$$

The resonance frequencies ω_r correspond to the excitation energies above the ground state:

$$\hbar\omega_r = E_r - E_0.$$

The quantum-mechanical *oscillation intensities* $f_r = (2m/\hbar)\omega_r |\langle E_r | Q_3 | E_0 \rangle|^2$ can be derived from the matrix elements of the position operator Q_3 . They are a measure for the coupling strength of an oscillation mode with frequency ω_r to the external field. They satisfy the *Thomas–Kuhn sum rule* $\sum f_r = Z$.

Let n be the number of molecular cores per unit volume. The polarization density is then given by

$$\mathbf{P} = n\mathbf{p} = n\alpha(\omega)\epsilon_0 \mathbf{F}. \quad (2.113)$$

If we were allowed to identify the field strength $\mathbf{F}$ at the position of a molecular core with the averaged field strength $\mathbf{E}$, we would find an electric susceptibility of $\chi = n\alpha$. However, $\mathbf{F}$ is the average field strength at the position of a molecule, while $\mathbf{E}$ is the average field strength within a small volume, independent of there being a molecule or not. It is only for very small densities n (e.g. in gases) that we may set approximately $\mathbf{E} = \mathbf{F}$.

In general, within the framework of a linear theory, the difference between $\mathbf{E}$ and $\mathbf{F}$ will be proportional to the polarization density $\mathbf{P}$:

$$\mathbf{F} = \mathbf{E} + r(\mathbf{P}), \quad (2.114)$$

where r denotes some linear mapping. If we also assume the molecular polarizability α to be a general linear mapping, we obtain

$$\mathbf{P} = n\epsilon_0\alpha(\mathbf{E} + r(\mathbf{P})),$$

i.e.

$$\mathbf{P} = (\mathbf{1} - n\epsilon_0\alpha r)^{-1}n\epsilon_0\alpha(\mathbf{E})$$

or

$$\chi = (\mathbf{1} - n\epsilon_0\alpha r)^{-1}n\alpha \quad (2.115)$$

and

$$n\alpha = \chi(\mathbf{1} + r\epsilon_0\chi)^{-1}. \quad (2.116)$$

For an isotropic medium, one can show that

$$r(\mathbf{P}) = \frac{1}{3\epsilon_0}\mathbf{P}, \quad (2.117)$$

so, for this case,

$$\begin{aligned} \chi &= \frac{n\alpha}{1 - \frac{1}{3}n\alpha} \\ n\alpha &= \frac{\chi}{1 + \frac{1}{3}\chi}. \end{aligned}$$

Setting $\chi = \epsilon - 1$, the last equation becomes

$$\frac{1}{3}n\alpha = \frac{\epsilon - 1}{\epsilon + 2}. \quad (2.118)$$

This is the well-known *Lorentz–Lorenz formula*.

For our simple model, $\alpha(\omega)$ is obviously a Pick function for $\gamma \geq 0$. As the function $f(z) = nz/(1 - \epsilon_0 rnz)$ is also a Pick function, χ satisfies the passivity and causality conditions. The reality condition is also fulfilled.

2.9 Material equations in moving media

Up to now the material equations for isotropic media,

$$\mathbf{j} = \sigma \mathbf{E} \quad (2.119)$$

and

$$\mathbf{D} = \epsilon\epsilon_0 \mathbf{E}, \quad \mathbf{H} = \frac{1}{\mu\mu_0} \mathbf{B}, \quad (2.120)$$

have only been formulated for the case of a medium at rest. The generalization to moving media is a surprisingly difficult problem, which is not only of fundamental interest. The solution of this problem is only possible within the relativistically covariant formulation of Maxwell's equations.

Maxwell's equations,

$$\nabla \cdot \mathbf{B} = 0, \quad \nabla \times \mathbf{E} = -\kappa \frac{\partial \mathbf{B}}{\partial t}, \quad (2.121)$$

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \quad \nabla \times \mathbf{B} = \mu_0 \kappa \left(\mathbf{j} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right), \quad (2.122)$$

can be written in the following relativistically covariant form:

$$\partial_\mu \hat{F}^{\mu\nu} = 0, \quad (2.123)$$

$$\partial_\mu F^{\mu\nu} = \kappa \mu_0 j^\nu, \quad (2.124)$$

where we have introduced the current 4-vector

$$j^\mu = (\rho c, \mathbf{j}), \quad (2.125)$$

the antisymmetric field tensor

$$F^{\mu\nu} = \begin{pmatrix} 0 & -E_1/\kappa c & -E_2/\kappa c & -E_3/\kappa c \\ +E_1/\kappa c & 0 & -B_3 & +B_2 \\ +E_2/\kappa c & +B_3 & 0 & -B_1 \\ +E_3/\kappa c & -B_2 & +B_1 & 0 \end{pmatrix}, \quad (2.126)$$

and the dual field tensor

$$\hat{F}^{\mu\nu} := \frac{1}{2} \epsilon^{\mu\nu\rho\sigma} F_{\rho\sigma}.$$

For a medium at rest, the inhomogeneous Maxwell's equations (2.122) have to be replaced by

$$\nabla \cdot \mathbf{D} = \rho, \quad \nabla \times \mathbf{H} = \kappa \left(\mathbf{j} + \frac{\partial \mathbf{D}}{\partial t} \right). \quad (2.127)$$

For a relativistic formulation we must introduce one more field tensor:

$$G^{\mu\nu} = \begin{pmatrix} 0 & -\kappa c D_1 & -\kappa c D_2 & -\kappa c D_3 \\ +\kappa c D_1 & 0 & -H_3 & +H_2 \\ +\kappa c D_2 & +H_3 & 0 & -H_1 \\ +\kappa c D_3 & -H_2 & +H_1 & 0 \end{pmatrix}. \quad (2.128)$$

The inhomogeneous equations now read

$$\partial_\mu G^{\mu\nu} = \kappa j^\nu. \quad (2.129)$$

The homogeneous equations (2.123) and (2.121) remain unchanged.

For a relativistically covariant formulation of eqs. (2.119) and (2.120), we first define the (local) velocity vector of the medium:

$$u^\mu = (\gamma c, \gamma \mathbf{v}) \quad \text{with} \quad \gamma = \frac{1}{\sqrt{1 - v^2/c^2}}. \quad (2.130)$$

We start with eq. (2.119), which generalizes to

$$j^\mu - \frac{1}{c^2} u^\mu u_\nu j^\nu = -\kappa \sigma u_\nu F^{\nu\mu}. \quad (2.131)$$

In the rest of the system we have $u^\nu = (c, \mathbf{0})$. For $\mu = 0$ we obtain an identity, and for $\mu = 1, 2, 3$ we recover the material equation $\mathbf{j} = \sigma \mathbf{E}$. Since eq. (2.131) is relativistically covariant, it holds in any inertial system.

For an arbitrary reference system, the $\mu = 0$ component of eq. (2.131) becomes

$$\rho c - \frac{\gamma}{c} u_\nu j^\nu = \frac{\sigma \gamma}{c} \mathbf{v} \cdot \mathbf{E},$$

and for $\mu = 1, 2, 3$ we obtain

$$\mathbf{j} - \frac{1}{c^2} \gamma \mathbf{v} u_\nu j^\nu = \gamma \sigma \mathbf{E} + \gamma \kappa \sigma \mathbf{v} \times \mathbf{B}.$$

Solving the first equation for $u_\nu j^\nu$ and inserting the result into the second equation yields

$$\mathbf{j} = \rho \mathbf{v} + \sigma \gamma \left[\mathbf{E} - \frac{\mathbf{v}}{c} \left(\frac{\mathbf{v}}{c} \cdot \mathbf{E} \right) \right] + \sigma \kappa \gamma \mathbf{v} \times \mathbf{B}. \quad (2.132)$$

The first term on the right-hand side of this equation is a contribution from convection, the second term generalizes the relation between $\mathbf{j}$ and $\mathbf{E}$ to moving media, and the third term is the contribution of the Lorentz force.

We now come to eqs. (2.120), which hold in the rest of the medium. With u^μ again denoting the velocity vector of the medium, these equations become

$$u_\mu G^{\mu\nu} = \kappa^2 c^2 \epsilon \epsilon_0 u_\mu F^{\mu\nu}, \quad u_\mu \hat{G}^{\mu\nu} = \frac{1}{\mu \mu_0} u_\mu \hat{F}^{\mu\nu}. \quad (2.133)$$

We solve these relations for $G^{\mu\nu}$ making the following ansatz, which takes into account the transformation and symmetry properties of $F^{\mu\nu}$:

$$G^{\mu\nu} = \alpha F^{\mu\nu} + \frac{\beta}{c^2} (u^\mu u_\rho F^{\rho\nu} - u^\nu u_\rho F^{\rho\mu}). \quad (2.134)$$

The dual tensor for this ansatz reads

$$\hat{G}^{\mu\nu} = \alpha \hat{F}^{\mu\nu} + \frac{\beta}{c^2} \epsilon^{\mu\nu\rho\sigma} u_\rho u^\tau F_{\tau\sigma}. \quad (2.135)$$

The coefficients α and β are determined by inserting eqs. (2.134) and (2.135) into eq. (2.133):

$$\begin{aligned}\alpha &= \frac{1}{\mu\mu_0}, \\ \beta &= \kappa^2 c^2 \epsilon \epsilon_0 - \frac{1}{\mu\mu_0} = \frac{1}{\mu_0} \left(\epsilon - \frac{1}{\mu} \right).\end{aligned}\tag{2.136}$$

(The last equation follows from $\kappa^2 c^2 \epsilon_0 \mu_0 = 1$.)

For vacuum we have $\epsilon = \mu = 1$, i.e. $\beta = 0$ and

$$G^{\mu\nu} = \frac{1}{\mu_0} F^{\mu\nu},\tag{2.137}$$

i.e. indeed

$$\mathbf{D} = \epsilon_0 \mathbf{E} \quad \text{and} \quad \mathbf{B} = \frac{1}{\mu_0} \mathbf{H}.$$

Finally, we rewrite eqs. (2.134) and (2.136) in a three-dimensional notation:

$$\begin{aligned}\mathbf{D} &= \epsilon_0 \left[\frac{1}{\mu} + \gamma^2 \left(\epsilon - \frac{1}{\mu} \right) \right] \mathbf{E} \\ &\quad - \frac{\gamma^2 \epsilon_0}{c^2} \left(\epsilon - \frac{1}{\mu} \right) \mathbf{v} (\mathbf{v} \cdot \mathbf{E}) - \gamma^2 \kappa \epsilon_0 \left(\epsilon - \frac{1}{\mu} \right) \mathbf{v} \times \mathbf{B}, \\ \mathbf{H} &= \frac{1}{\mu_0} \left[\epsilon - \gamma^2 \left(\epsilon - \frac{1}{\mu} \right) \right] \mathbf{B} \\ &\quad + \frac{\gamma^2}{\mu_0 c^2} \left(\epsilon - \frac{1}{\mu} \right) \mathbf{v} (\mathbf{v} \cdot \mathbf{B}) + \gamma^2 \kappa \epsilon_0 \left(\epsilon - \frac{1}{\mu} \right) \mathbf{v} \times \mathbf{E}.\end{aligned}\tag{2.138}$$

For the limiting cases $\mathbf{v} = 0$ or $\epsilon = \mu = 1$ we recover the well-known relations. Furthermore, we notice that for $\mathbf{v} \neq 0$ the medium becomes anisotropic and $\mathbf{B}$ contributes to $\mathbf{D}$, and $\mathbf{E}$ contributes to $\mathbf{H}$.

3 Linear waves in homogeneous media

In this chapter we will consider some fundamental aspects of linear wave phenomena, in particular the propagation of light waves in transparent media. Because of their intuitive appeal and the similarities with optical waves, we will first discuss elastic waves in solids.

3.1 Elastic waves in solids

We consider oscillations due to elastic deformations of a solid. The fundamental field quantity for such deformations is $\mathbf{u}(t, \mathbf{x})$, the displacement from the equilibrium position of a mass element of the body. After the deformation, the squared distance between two neighboring mass elements with equilibrium positions $\mathbf{x}$ and $\mathbf{x} + \Delta\mathbf{x}$ will be given by

$$\begin{aligned} (\Delta x'^i)^2 &= \left(\Delta x^i + \frac{\partial u^i}{\partial x^k} \Delta x^k \right)^2 \\ &= (\Delta x^i)^2 + \left(\frac{\partial u^i}{\partial x^k} + \frac{\partial u^k}{\partial x^i} \right) \Delta x^i \Delta x^k + \frac{\partial u^i}{\partial x^k} \frac{\partial u^i}{\partial x^r} \Delta x^k \Delta x^r. \end{aligned} \quad (3.1)$$

For small absolute deformations, the relative deformations will also be small and we have $\partial u^i / \partial x^k \ll 1$, and we can neglect the last term in the eq. (3.1). Hence, the distortions are described by the symmetric tensor

$$u_{ik} = \frac{1}{2} (\partial_i u_k + \partial_k u_i). \quad (3.2)$$

Instead of $\partial / \partial x^i$, we will often write simply ∂_i . (For Cartesian coordinates, the position of the indices is irrelevant, and we can position the indices freely.)

In order to derive the equation of motion for small elastic deformations, we start from the *Lagrangian density*

$$\mathcal{L} = \mathcal{T} - \mathcal{U},$$

where the kinetic energy density is given by

$$\mathcal{T} = \frac{1}{2} \dot{u}^r M_{rs} \dot{u}^s = \frac{1}{2} \dot{\mathbf{u}} \cdot M \dot{\mathbf{u}}, \quad (3.3)$$

and the density of the potential energy by

$$\mathcal{U} = \frac{1}{2} \partial_i u^r K_{rs}^{ij} \partial_j u^s = \frac{1}{2} \partial_i \mathbf{u} \cdot K^{ij} \partial_j \mathbf{u}. \quad (3.4)$$

Here, M and K^{ij} are constant 3×3 matrices, which should satisfy the condition that the potential energy density is positive semidefinite. The quadratic ansatz for the potential energy density also takes into account the following aspects:

- (a) *Translational invariance*: it is not the displacement u^i itself that enters the equations but the relative displacement $\partial_i u_j$.
- (b) *Hooke's law*: for small displacements, the elastic stress is proportional to the distortion.

In complete analogy to the Hamiltonian principle in classical mechanics with finitely many degrees of freedom, where one starts from a Lagrange function $L(q, \dot{q})$ and an action functional $S[q] = \int dt L(q, \dot{q})$, we also obtain in this case the equation of motion from the Lagrange density $\mathcal{L}$ by a variation of the action functional

$$S[u] = \int dt \int d^3x \mathcal{L}(u, \dot{u}, \partial_i u). \quad (3.5)$$

The variation of $S[u]$ yields

$$\begin{aligned} \delta S[u] &= \int dt \int d^3x \left\{ \frac{\partial \mathcal{L}}{\partial u} \delta u + \frac{\partial \mathcal{L}}{\partial \dot{u}} \delta \dot{u} + \frac{\partial \mathcal{L}}{\partial \partial_i u} \delta \partial_i u \right\} \\ &= \int dt \int d^3x \left\{ \frac{\partial}{\partial t} \left(\frac{\partial \mathcal{L}}{\partial \dot{u}} \delta u \right) + \partial_i \left(\frac{\partial \mathcal{L}}{\partial \partial_i u} \delta u \right) \right. \\ &\quad \left. + \left(\frac{\partial \mathcal{L}}{\partial u} - \frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{u}} - \partial_i \frac{\partial \mathcal{L}}{\partial \partial_i u} \right) \delta u \right\}. \end{aligned}$$

With the help of Gauss's formula we can convert the first two terms in the integral to boundary terms. The requirement that the solution should be a stationary point of the action leads to the equation of motion:

$$\frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{u}} + \partial_i \frac{\partial \mathcal{L}}{\partial \partial_i u} - \frac{\partial \mathcal{L}}{\partial u} = 0. \quad (3.6)$$

Like in classical mechanics, we can derive the conservation of energy as long as $\mathcal{L}$ does not explicitly depend on time. Indeed, from

$$\frac{\partial}{\partial t} \mathcal{L}(u, \dot{u}, \partial_i u) = \frac{\partial \mathcal{L}}{\partial u} \dot{u} + \frac{\partial \mathcal{L}}{\partial \dot{u}} \ddot{u} + \frac{\partial \mathcal{L}}{\partial \partial_i u} \partial_i \dot{u},$$

and using the equation of motion (3.6), we obtain

$$\begin{aligned} \frac{\partial}{\partial t} \mathcal{L} &= \left(\frac{\partial}{\partial t} \frac{\partial \mathcal{L}}{\partial \dot{u}} \right) \dot{u} + \frac{\partial \mathcal{L}}{\partial \dot{u}} \ddot{u} + \left(\partial_i \frac{\partial \mathcal{L}}{\partial \partial_i u} \right) \dot{u} + \frac{\partial \mathcal{L}}{\partial \partial_i u} \partial_i \dot{u} \\ &= \frac{\partial}{\partial t} \left(\frac{\partial \mathcal{L}}{\partial \dot{u}} \dot{u} \right) + \partial_i \left(\frac{\partial \mathcal{L}}{\partial \partial_i u} \dot{u} \right). \end{aligned}$$

Hence

$$\frac{\partial}{\partial t} \left(\frac{\partial \mathcal{L}}{\partial \dot{u}} \dot{u} - \mathcal{L} \right) + \partial_i \left(\frac{\partial \mathcal{L}}{\partial \partial_i u} \dot{u} \right) = 0, \quad (3.7)$$

from which we can read off the energy density

$$\rho_E = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{u}}} \dot{\mathbf{u}} - \mathcal{L} \quad (3.8)$$

and the energy current density

$$S^i = \frac{\partial \mathcal{L}}{\partial \partial_i \mathbf{u}} \dot{\mathbf{u}}. \quad (3.9)$$

If $\mathcal{L} = \mathcal{T} - \mathcal{U}$ is given by eqs. (3.3) and (3.4), the equation of motion reads

$$M_{rs} \ddot{u}^s - \partial_i K_{rs}^{ij} \partial_j u^s = M_{rs} \ddot{u}^s - \partial_i \sigma_r^i = 0. \quad (3.10)$$

The components of the elastic stress tensor σ_r^i are equal to the r th component of the force onto a surface with a normal vector pointing in direction i :

$$\sigma_r^i = \frac{\partial \mathcal{U}}{\partial \partial_i u^r} = K_{rs}^{ij} \partial_j u^s. \quad (3.11)$$

For the energy density we find

$$\rho_E = \frac{1}{2} \dot{\mathbf{u}} \cdot M \dot{\mathbf{u}} + \frac{1}{2} \partial_i \mathbf{u} \cdot K^{ij} \partial_j \mathbf{u} = \mathcal{T} + \mathcal{U}, \quad (3.12)$$

and for the energy current density

$$S^i = \dot{\mathbf{u}} \cdot K^{ij} \partial_j \mathbf{u}. \quad (3.13)$$

Written in matrix notation, the equation of motion becomes

$$M \ddot{\mathbf{u}} - K^{ij} \partial_i \partial_j \mathbf{u} = 0. \quad (3.14)$$

This is a generalized wave equation, for which we find solutions in the form of plane waves by making the following ansatz:

$$\mathbf{u}(t, \mathbf{x}) = \mathbf{v} e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})} \quad (\mathbf{v} \in \mathbb{C}^3). \quad (3.15)$$

Inserting this ansatz into the equation of motion leads to an eigenvalue equation

$$\begin{aligned} (K^{ij} k_i k_j - M \omega^2) \mathbf{v} &= 0, \\ \det(K^{ij} k_i k_j - M \omega^2) &= 0. \end{aligned} \quad (3.16)$$

The following properties of the solutions can be read off immediately:

- (a) The eigenfrequencies $\omega_\alpha(\mathbf{k})$ ($\alpha = 1, 2, 3$) are, as a function of $\mathbf{k}$, homogeneous of first order, i.e. $\omega_\alpha(\lambda \mathbf{k}) = \lambda \omega_\alpha(\mathbf{k})$ for $\lambda > 0$.
- (b) We can choose the components of the corresponding eigenvectors $\mathbf{v}_\alpha(\mathbf{k})$ to be real numbers, i.e. for an eigen-oscillation, $\mathbf{u}(t, \mathbf{x})$ points into a fixed direction. For $\mathbf{v}_\alpha(\mathbf{k})$ and $\mathbf{v}_\beta(\mathbf{k})$ we find an orthogonality relation: $\mathbf{v}_\alpha(\mathbf{k}) \cdot M \mathbf{v}_\beta(\mathbf{k}) = 0$ for $\alpha \neq \beta$.

The most general solution to the wave equation can be expressed as a superposition of eigenfunctions:

$$\mathbf{u}(t, \mathbf{x}) = \sum_{\alpha} \int d^3k [A_{\alpha}(\mathbf{k}) e^{i\omega_{\alpha}(\mathbf{k})t} + B_{\alpha}(\mathbf{k}) e^{-i\omega_{\alpha}(\mathbf{k})t}] \mathbf{v}_{\alpha}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}}, \quad (3.17)$$

where $A_{\alpha}(\mathbf{k}), B_{\alpha}(\mathbf{k}) \in \mathbb{C}$.

If the functions $A_{\alpha}(\mathbf{k})$ and $B_{\alpha}(\mathbf{k})$ are distributed around some mean value $\mathbf{k}_0$ with variance $|\Delta\mathbf{k}|$, the solution separates into three wave packets whose centers move with (in general different) *group velocities*

$$\mathbf{w}_{\alpha} = \left. \nabla \omega_{\alpha}(\mathbf{k}) \right|_{\mathbf{k}=\mathbf{k}_0}. \quad (3.18)$$

Furthermore, $\mathbf{w}_{\alpha}$ is the average flux velocity $\langle \mathbf{S} \rangle / \langle \rho_E \rangle$ for the energy of a solution. Indeed, from

$$K^{ij} k_i k_j \mathbf{v}_{\alpha}(\mathbf{k}) = M \omega_{\alpha}^2 \mathbf{v}_{\alpha}(\mathbf{k}),$$

one obtains, after taking the derivative with respect to k_i and scalar multiplication with $\mathbf{v}_{\alpha}(\mathbf{k})$, the following expectation values:

$$\langle S^i \rangle = \frac{1}{2} \omega_{\alpha} \mathbf{v}_{\alpha} \cdot K^{ij} k_j \mathbf{v}_{\alpha} = \frac{1}{4} \omega_{\alpha} \frac{\partial \omega_{\alpha}^2}{\partial k_i} \mathbf{v}_{\alpha} \cdot M \mathbf{v}_{\alpha} = \frac{1}{2} \omega_{\alpha}^2 w_{\alpha i} \mathbf{v}_{\alpha} \cdot M \mathbf{v}_{\alpha} \quad (3.19)$$

and

$$\langle \rho_E \rangle = \frac{1}{4} \omega_{\alpha}^2 \mathbf{v}_{\alpha} \cdot M \mathbf{v}_{\alpha} + \frac{1}{4} \mathbf{v}_{\alpha} \cdot K^{ij} k_i k_j \mathbf{v}_{\alpha} = \frac{1}{2} \omega_{\alpha}^2 \mathbf{v}_{\alpha} \cdot M \mathbf{v}_{\alpha}. \quad (3.20)$$

3.2 Isotropic elastic media

Within Hooke's approximation, the density of the potential energy of a deformed solid is a quadratic function of the quantities

$$u_{ik} = \frac{1}{2} (\partial_i u_k + \partial_k u_i),$$

which implies that it has the form

$$\mathcal{U} = \frac{1}{2} \lambda_{irjs} u_{ir} u_{js}, \quad (3.21)$$

where the coefficients λ_{irjs} are subject to the symmetry relations:

$$\lambda_{irjs} = \lambda_{rij s} = \lambda_{irs j} = \lambda_{jsir}. \quad (3.22)$$

If we combine the symmetric pairs of indices, (i, r) and (j, s) , into indices A and B , we obtain a symmetric matrix λ_{AB} of elasticity coefficients, where A and B can assume six different values. A symmetric 6×6 matrix can have 21 independent elements, so, in general, there are 21 elasticity coefficients.

For an isotropic solid, $\mathcal{U}$ should be rotationally invariant and the number of independent elasticity coefficients reduces to two:

$$\mathcal{U} = \frac{\lambda}{2}(u_{ii})^2 + \mu u_{ik}u_{ik} = \frac{\lambda}{2} \left(\sum_i u_{ii} \right)^2 + \mu \sum_{i,k} u_{ik}u_{ik}. \quad (3.23)$$

If we distinguish volume-changing and volume-preserving deformations, we may also write

$$\begin{aligned} u_{ij} &= u_{ij} - \frac{1}{3}\delta_{ij}u_{kk} + \frac{1}{3}\delta_{ij}u_{kk} = \tilde{u}_{ij} + \frac{1}{3}\delta_{ij}u_{kk}, \\ \mathcal{U} &= \frac{1}{2}\kappa(u_{ii})^2 + \mu\tilde{u}_{ik}\tilde{u}_{ik}, \end{aligned} \quad (3.24)$$

with $\kappa = \lambda + \frac{2}{3}\mu$. The coefficients λ and μ are called *Lamé coefficients*. The condition

$$\mathcal{U} \geq 0 \quad \text{implies} \quad \mu \geq 0 \quad \text{and} \quad \kappa = \lambda + \frac{2}{3}\mu \geq 0. \quad (3.25)$$

The elastic stress tensor follows from $\delta\mathcal{U} = \sigma_{ik}\delta u_{ik}$:

$$\sigma_{ik} = \kappa u_{rr}\delta_{ik} + 2\mu(u_{ik} - \frac{1}{3}\delta_{ik}u_{rr}). \quad (3.26)$$

In particular, we find

$$\sigma_{ii} = 3\kappa u_{ii}.$$

Solving for the relative displacements we get

$$u_{ik} = \frac{1}{9\kappa}\delta_{ik}\sigma_{rr} + \frac{1}{2\mu} \left(\sigma_{ik} - \frac{1}{3}\delta_{ik}\sigma_{rr} \right). \quad (3.27)$$

This is Hooke's law for an isotropic medium.

We find the relation between λ , μ and the well-known elasticity moduli by considering special types of deformations, for example the following:

(a) Homogeneous compression ($\sigma_{ik} = p\delta_{ik}$):

$$u_{ik} = \frac{1}{3\kappa}\sigma_{rr}\delta_{ik}$$

and

$$u_{rr} = \frac{\delta V}{V} = \frac{1}{\kappa}\sigma_{rr}, \quad (3.28)$$

hence κ is the *bulk modulus*.

(b) Shearing ($\sigma_{12} = \sigma_{21} = p$, $\sigma_{ik} = 0$ otherwise):

$$u_{12} = u_{21} = \frac{1}{2}\sin\alpha = \frac{1}{2\mu}p, \quad u_{ik} = 0 \text{ const.} \quad (3.29)$$

where α is the shear angle and μ turns out to be the *shear modulus*.

(c) One-sided expansion ($\sigma_{11} = p$, $\sigma_{ik} = 0$ otherwise):

$$\begin{aligned} u_{11} &= \frac{\Delta l}{l} = \left(\frac{1}{9\kappa} + \frac{1}{3\mu} \right) p = \frac{1}{E} p, \\ u_{22} &= u_{33} = \left(\frac{1}{9\kappa} - \frac{1}{6\mu} \right) p. \end{aligned} \quad (3.30)$$

Hence, $E = [(1/9\kappa) + (1/3\mu)]^{-1}$ is *Young's modulus*, and the *shear contraction* $\sigma = -u_{22}/u_{11}$ is given by

$$\sigma = \frac{1}{2} \frac{3\kappa - 2\mu}{3\kappa + \mu}. \quad (3.31)$$

Hence, from $\kappa, \mu \geq 0$ it follows that: $-1 \leq \sigma \leq 1/2$. In practice, we even have $\sigma \geq 0$, which implies $3\lambda = 3\kappa - 2\mu \geq 0$.

A straightforward calculation leads from eq. (3.10) to the wave equation for an isotropic elastic medium. If we assume the following simple form for the kinetic energy density,

$$\mathcal{T} = \frac{1}{2} \rho \dot{\mathbf{u}}^2, \quad (3.32)$$

we obtain

$$\rho \ddot{u}_i - \partial_k \sigma_{ki} = 0,$$

hence

$$\rho \ddot{u}_i - \lambda \partial_i \partial_k u_k - 2\mu \partial_k u_{ki} = 0, \quad (3.33)$$

i.e.

$$\rho \ddot{\mathbf{u}} - (\lambda + \mu) \nabla (\nabla \cdot \mathbf{u}) - \mu \Delta \mathbf{u} = 0. \quad (3.34)$$

The exponential ansatz $\mathbf{u}(t, \mathbf{x}) = \mathbf{v} e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}$ leads to the eigenvalue equation

$$[\mu \mathbf{k}^2 \delta_{ij} + (\lambda + \mu) k_i k_j - \rho \omega^2 \delta_{ij}] v_j = 0, \quad (3.35)$$

from which we can determine the following eigenvalues and eigenvectors:

- Transverse oscillations ($\mathbf{v}_{1,2}$ orthogonal to $\mathbf{k}$)

$$\omega_{1,2} = \sqrt{\frac{\mu}{\rho}} k, \quad \text{with phase velocity} \quad c_t = \frac{\omega_{1,2}}{k} = \sqrt{\frac{\mu}{\rho}}. \quad (3.36)$$

- Longitudinal oscillations ($\mathbf{v}_3$ collinear to $\mathbf{k}$)

$$\omega_3 = \sqrt{\frac{\lambda + 2\mu}{\rho}} k, \quad \text{with velocity} \quad c_l = \sqrt{\frac{\lambda + 2\mu}{\rho}}. \quad (3.37)$$

From $\lambda, \mu \geq 0$, we obtain the following inequality for the propagation velocities of the longitudinal and transverse waves:

$$c_l \geq \sqrt{2} c_t. \quad (3.38)$$

Indeed, seismograph measurements of earthquakes detect longitudinal waves first.

3.3 Wave surfaces and ray surfaces

Now we return to the general wave equation (3.14). For fixed ω , the eigenvalue equation (3.16),

$$f(\omega, \mathbf{k}) = \det(K^{ij} k_i k_j - M\omega^2) = 0, \quad (3.39)$$

can be interpreted as an equation for the wave vector $\mathbf{k}$. This leads to a surface in $\mathbb{R}^3$ which, for elastic waves, consists generally of three shells, because for each $\mathbf{k}$ there exist three eigenfrequencies ω_α ($\alpha = 1, 2, 3$). From

$$\nabla f(\omega_\alpha(\mathbf{k}), \mathbf{k}) = \frac{\partial f}{\partial \omega} \nabla_k \omega_\alpha(\mathbf{k}) + (\nabla_k f)(\omega_\alpha(\mathbf{k}), \mathbf{k}) = 0, \quad (3.40)$$

we can deduce that the three group velocity vectors, $\mathbf{w}_\alpha = \nabla_k \omega_\alpha$, are orthogonal to this surface. Frequency $\omega(\mathbf{k})$ is a homogeneous function of first order in $\mathbf{k}$, i.e.

$$\begin{aligned} \mathbf{k} \cdot \nabla_k \omega(\mathbf{k}) &= \omega(\mathbf{k}) \quad \text{or} \\ \frac{\mathbf{k} c}{\omega} \cdot \frac{\mathbf{w}}{c} &= \mathbf{n} \cdot \mathbf{s} = 1, \end{aligned} \quad (3.41)$$

where we have introduced some reference velocity c such that the vectors $\mathbf{n}$ and $\mathbf{s}$ become dimensionless. The physical interpretation of $\mathbf{n}$ and $\mathbf{s}$ is the following:

- The vector $\mathbf{n} = \mathbf{k} c / \omega$ points in the direction of the wave vector $\mathbf{k}$, and its absolute value is given by $|\mathbf{n}| = kc / \omega = c / c_p$, where c_p is the velocity (to be more precise, the phase velocity) of the wave corresponding to $\mathbf{k}$ (and $\omega(\mathbf{k})$). This ratio of velocities is called the *refractive index* (or *index of refraction*) and $\mathbf{n}$ is a vector that tells us the refractive index for the corresponding directions of $\mathbf{k}$. The surface defined by

$$g(\omega, \mathbf{n}) = \det(K^{ij} n_i n_j - Mc^2) = 0 \quad (3.42)$$

is called the *wave surface*. As $\omega(\mathbf{k})$ is a homogeneous function of order 1 in $\mathbf{k}$, $g(\omega, \mathbf{n})$ does not depend on ω .

- The vector $\mathbf{s} = \mathbf{w} / c$ is called the *ray vector*. It is parallel to the normal vector of the wave surface, $(\nabla_n g)(\omega, \mathbf{n})$, and normalized by the condition $\mathbf{n} \cdot \mathbf{s} = 1$. We have seen that, for elastic media, $c\mathbf{s}$ is the group velocity and also the velocity of the energy flux, which in general is *not* collinear to the direction of $\mathbf{k}$.

If we choose some point in $\mathbf{0} \in \mathbb{R}^3$ and draw, for a fixed frequency ω , the corresponding ray vectors $\mathbf{s}$ for any direction, we obtain a multi-shelled surface,

$$h(\omega, \mathbf{s}) = 0,$$

the so-called *ray surface*. The ray surface describes the flux of energy for all directions, i.e. it represents the form of wave fronts that would result from a point-like disturbance at the origin $\mathbf{0}$. By definition, $\mathbf{s}$ is perpendicular to the wave surface and $\mathbf{n} \cdot \mathbf{s} = 1$. Therefore, we obtain the ray surface from the wave surface by a simple geometrical construction (see fig. 3.1). For each point $\mathbf{n}$ on the wave surface, we draw the tangent plane, construct the line perpendicular to this plane through the center of the wave surface, and normalize this line such that $\mathbf{n} \cdot \mathbf{s} = 1$.

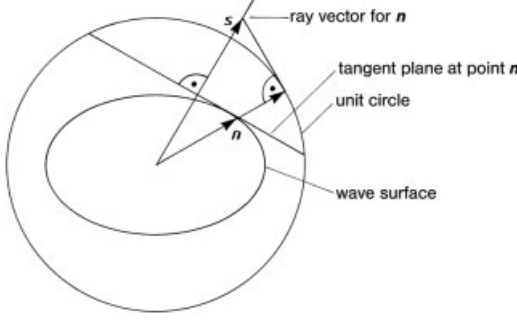


Figure 3.1: Construction of the ray surface from the wave surface.

Wave surface and ray surface are “dual” to each other: the above construction applied to the ray surface leads back to the wave surface. For a proof of this statement, we consider the variation $\delta \mathbf{n}$ of some $\mathbf{n}$ along the wave surface. Hence, $\delta \mathbf{n}$ is tangent to the wave surface and by definition of $\mathbf{s}$ we have $\mathbf{s} \cdot \delta \mathbf{n} = 0$. On the other hand, from $\mathbf{n} \cdot \mathbf{s} = 1$ we obtain

$$\mathbf{s} \cdot \delta \mathbf{n} + \mathbf{n} \cdot \delta \mathbf{s} = \mathbf{n} \cdot \delta \mathbf{s} = 0, \quad (3.43)$$

which means that $\mathbf{n}$ is orthogonal to the ray surface. The normalization condition, $\mathbf{n} \cdot \mathbf{s} = 1$, is symmetric in $\mathbf{n}$ and $\mathbf{s}$. Therefore, $\mathbf{n}$ may be constructed from the ray surface in the same way as $\mathbf{s}$ from the wave surface.

Examples

1. For the one-component wave equation

$$\frac{1}{c^2} \ddot{u} - A_{ij} \partial_i \partial_j u = 0 \quad (u(t, \mathbf{x}) \in \mathbb{R}), \quad (3.44)$$

where A_{ij} is some symmetric positive definite matrix, the wave surface has the shape of an ellipsoid:

$$f = A_{ij} n_i n_j - 1 = 0. \quad (3.45)$$

Because $\mathbf{n} \cdot \mathbf{s} = 1$, the ray vector $s_i = \alpha \nabla_i f = 2\alpha A_{ij} n_j$ is determined to be $s_i = A_{ij} n_j$, from which we get $n_i = A_{ij}^{-1} s_j$, where A_{ij}^{-1} is the inverse matrix of A_{ij} . Hence, the ray surface is given by

$$h = A_{ij}^{-1} s_i s_j - 1 = 0, \quad (3.46)$$

which is the ellipsoid inverse to the wave surface. We can easily verify the interpretation of the wave surfaces as an image of the shapes of the wave fronts, because the wave equation (3.44) admits solutions of the form

$$u(t, \mathbf{x}) = \frac{g(R \mp ct)}{R}, \quad (3.47)$$

where $R = \sqrt{A_{ij}^{-1}x^i x^j}$, and g is an arbitrary function of one variable. In particular,

$$G_\omega(\mathbf{x}) = \frac{1}{\sqrt{\det A}} \frac{e^{(i\omega/c)R}}{R} \quad (3.48)$$

satisfies the equation

$$(A_{ij}\partial_i\partial_j + \omega^2/c^2)G_\omega(\mathbf{x}) = -4\pi\delta(\mathbf{x}).$$

2. The wave surface for an isotropic elastic medium follows from eqs. (3.35)–(3.37):

$$f(\omega, \mathbf{n}) = \left(\mathbf{n}^2 - \frac{c^2}{c_1^2} \right) \left(\mathbf{n}^2 - \frac{c^2}{c_t^2} \right)^2 = 0. \quad (3.49)$$

This describes a sphere of radius $n_l = c/c_1$ and a second sphere (which has to be counted twice because the transverse waves are degenerate) of radius $n_t = c/c_t$. The ray surface consists of spheres with radii $s_l = c_1/c$ and $s_t = c_t/c$.

In the following chapters we will meet further examples of wave surfaces and ray surfaces.

4 Crystal optics

4.1 The normal ellipsoid

In this chapter we will discuss the propagation of light in transparent isotropic and anisotropic media, especially in crystals. Compared to electrical effects, we can neglect magnetic phenomena and will always assume $\mu_{ik} = \delta_{ik}$, i.e. $\mathbf{B} = \mu_0 \mathbf{H}$.

Under these conditions, the optical properties of a medium are determined by the permittivity tensor $\epsilon_{ij}(\omega) = \epsilon_{ji}(\omega)$. If absorption effects are negligible, $\epsilon_{ij}(\omega)$ is real and positive definite; the same holds for the inverse tensor $\epsilon_{ij}^{-1}(\omega)$, with $\epsilon_{ij}(\omega)\epsilon_{jk}^{-1}(\omega) = \delta_{ik}$.

Associated with the tensor $\epsilon_{ij}(\omega)$ is the so-called *Fresnel's ellipsoid*:

$$E_F(\omega) = \{\mathbf{x} \in \mathbb{R}^3 \mid \mathbf{x} \cdot \epsilon(\omega)\mathbf{x} = x_i \epsilon_{ij}(\omega) x_j = 1\}. \quad (4.1)$$

Even more important for a visualization of the optical properties of a medium is the *normal ellipsoid* or *index ellipsoid*, associated with the inverse tensor $\epsilon_{ij}^{-1}(\omega)$:

$$E_N(\omega) = \{\mathbf{x} \in \mathbb{R}^3 \mid \mathbf{x} \cdot \epsilon^{-1}(\omega)\mathbf{x} = 1\}. \quad (4.2)$$

The electric energy density satisfies

$$\rho_{\text{el}} = \frac{1}{2} \mathbf{E} \cdot \mathbf{D} = \frac{\epsilon_0}{2} \mathbf{E} \cdot \epsilon(\omega) \mathbf{E} = \frac{1}{2\epsilon_0} \mathbf{D} \cdot \epsilon^{-1}(\omega) \mathbf{D},$$

and therefore $E_F(\omega)$ is equal to the set of all vectors $\mathbf{E}$ for which $\rho_{\text{el}} = \epsilon_0/2$, and $E_N(\omega)$ is equal to the set of all vectors $\mathbf{D}$ for which $\rho_{\text{el}} = 1/(2\epsilon_0)$. The two ellipsoids are dual to each other in the sense of chapter 3: the normal vectors on E_F determine the direction of $\mathbf{D}$, and the normal vectors on E_N determine the directions of $\mathbf{E}$.

We can choose an orthogonal system of principal axes, for which $\epsilon_{ij}(\omega)$ and $\epsilon_{ij}^{-1}(\omega)$ become diagonal:

$$\epsilon_{ij}(\omega) = \epsilon_i(\omega) \delta_{ij}, \quad \epsilon_{ij}^{-1}(\omega) = \frac{1}{\epsilon_i(\omega)} \delta_{ij}. \quad (4.3)$$

The eigenvalues $\epsilon_i(\omega)$ are called *principal values of the electric permeability*. By convention we let their labels increase with their values:

$$0 \leq \epsilon_1(\omega) \leq \epsilon_2(\omega) \leq \epsilon_3(\omega). \quad (4.4)$$

Adopting the principal axes as a basis, the equations for Fresnel's ellipsoid and the normal ellipsoid simplify significantly:

$$\epsilon_1(\omega)x^2 + \epsilon_2(\omega)y^2 + \epsilon_3(\omega)z^2 = 1,$$

and

$$\frac{x^2}{\epsilon_1(\omega)} + \frac{y^2}{\epsilon_2(\omega)} + \frac{z^2}{\epsilon_3(\omega)} = 1. \quad (4.5)$$

The axes of these ellipsoids have lengths $1/\sqrt{\epsilon_1(\omega)}$, $1/\sqrt{\epsilon_2(\omega)}$, $1/\sqrt{\epsilon_3(\omega)}$, and $\sqrt{\epsilon_1(\omega)}$, $\sqrt{\epsilon_2(\omega)}$, $\sqrt{\epsilon_3(\omega)}$, respectively (see fig. 4.1).

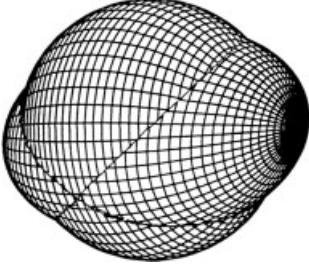


Figure 4.1: Circular cuts through an ellipsoid.

In general, not only the principal values $\epsilon_i(\omega)$ but also the directions of the principal axes depend on the frequency; in this case we speak of a *dispersion of the axes*.

However, if certain directions are distinguished due to the symmetry of the crystal lattice, the principal axes are fixed; and two of the principal values are equal if the crystal lattice is invariant under rotations around some axis with an angle $\Theta \neq 180^\circ$. According to increasing symmetry, one distinguishes seven crystal systems, for which the details are summarized in table 4.1.

The notations for the principal axes are as follows:

G *dispersion of the axis*: this is the general case; the direction of the axis depends on the frequency;

F *fixed axis*: the direction of the axis does not depend on the frequency;

D *degeneracy*: there are at least two axes of equal length; due to the degeneracy, the directions of these axes are to a certain degree arbitrary, but the space spanned by the axes of equal length is fixed.

In addition, Table 4.1 names the symmetry of the translational lattice and the shape of the normal ellipsoid.

We shall soon need to investigate intersections of the normal ellipsoid with planes through the center. The intersection of an ellipsoid E , given by

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} + \frac{z^2}{A_3^2} = 1, \quad A_1 \leq A_2 \leq A_3,$$

with a plane $H : m_1x + m_2y + m_3z = 0$ is an ellipse with principal axes $a_1 \leq a_2$. The lengths of the principal axes of this ellipse can be identified as the extremal points of the

Table 4.1: Crystal systems and their normal ellipsoids.

Crystal system	Symmetry of the lattice	Principal axes of ϵ_{ij}	Normal ellipsoid
Triclinic	General parallelepiped	G G G	Unequal axes
Monoclinic	Parallelepiped with one edge orthogonal to a parallelogram	F G G	Unequal axes
Orthogonal = orthorhombic	Rectangular parallelepiped	F F F	Unequal axes
Trigonal	Cube, stretched along a diagonal direction of a body	F D D	Rotational ellipsoid
Tetragonal	Rectangular parallelepiped with quadratic base	F D D	Rotational ellipsoid
Hexagonal	Hexagonal-prismatic	F D D	Rotational ellipsoid
Cubic	Cube	D D D	Sphere

central distance functional restricted to an i -dimensional subspace $\mathbb{R}^i$:

$$\begin{aligned}
 A_i &= \min_{\mathbb{R}^i} \max_{\mathbf{x} \in \mathbb{R}^i \cap E} |\mathbf{x}| = \max_{\mathbb{R}^{4-i}} \min_{\mathbf{x} \in \mathbb{R}^{4-i} \cap E} |\mathbf{x}| \\
 a_i &= \min_{\mathbb{R}^i \subset H} \max_{\mathbf{x} \in \mathbb{R}^i \cap E \cap H} |\mathbf{x}|.
 \end{aligned} \tag{4.6}$$

From this representation we can easily verify that

$$A_1 \leq a_1 \leq A_2 \leq a_2 \leq A_3. \tag{4.7}$$

Of special importance in optics are those intersection planes for which the intersection with the normal ellipsoid is a circle. The normal vectors for these planes are called *optic axes*.

From inequality (4.7) we deduce for circular intersections: $a_1 = a_2 = A_2$. The orientation of such planes can be obtained by calculating the intersection of the ellipsoid

$$\frac{x^2}{A_1^2} + \frac{y^2}{A_2^2} + \frac{z^2}{A_3^2} = 1$$

with the sphere

$$\frac{x^2 + y^2 + z^2}{A_2^2} = 1.$$

Taking the difference of these two equations gives

$$x^2 \left(\frac{1}{A_1^2} - \frac{1}{A_2^2} \right) = z^2 \left(\frac{1}{A_2^2} - \frac{1}{A_3^2} \right).$$

So the planes with a circular intersection contain the 2-direction, and the angle θ between their normal vectors and the 3-direction is given by

$$\tan \theta = \pm \sqrt{\frac{1/A_1^2 - 1/A_2^2}{1/A_2^2 - 1/A_3^2}}. \quad (4.8)$$

For the general case $A_1 < A_2 < A_3$, there are two circular intersections that are mirror images of each other. For this reason, crystals with $\epsilon_1 < \epsilon_2 < \epsilon_3$ are called *optically biaxial* (see fig. 4.1).

If at least two of the principal axes have equal length, the two circular intersections coincide. Crystals for which $\epsilon_1 = \epsilon_2 < \epsilon_3$ or $\epsilon_1 < \epsilon_2 = \epsilon_3$ are called *optically uniaxial*. Finally, for a sphere, all intersections are circular. According to our crystal classification in table 4.1, the crystals belonging to the first three systems are biaxial, those belonging to the fourth to sixth systems are uniaxial, and the crystals belonging to the cubic system are identical to completely isotropic substances with respect to their dielectric properties.

4.2 Plane waves in crystals

We consider Maxwell's equations (2.76) for $\rho \equiv 0$ and $\mathbf{j} \equiv 0$ and make an exponential ansatz for the quantities $\mathbf{E}$, $\mathbf{D}$, $\mathbf{B}$, and $\mathbf{H}$:

$$\mathbf{E} = \mathbf{E}_0 \exp[-i(\omega t - \mathbf{k} \cdot \mathbf{x})], \quad \dots$$

Furthermore, we assume that $\mu_{ik} = \delta_{ik}$, i.e. $\mathbf{B} = \mu_0 \mathbf{H}$, and obtain

$$\mathbf{n} \cdot \mathbf{B}_0 = 0, \quad \mathbf{n} \cdot \mathbf{D}_0 = 0, \quad (4.9)$$

$$\mathbf{n} \times \mathbf{E}_0 = c\kappa \mathbf{B}_0, \quad \mathbf{n} \times \mathbf{H}_0 = -c\kappa \mathbf{D}_0. \quad (4.10)$$

Here, $\mathbf{n} = (c/\omega)\mathbf{k}$ is the vector of the refractive index defined in chapter 3, and the first two equations follow from the last two. For convenience, we will suppress the lower index "0" in $\mathbf{E}_0$, $\mathbf{D}_0$, $\mathbf{B}_0$, and $\mathbf{H}_0$. Eliminating $\mathbf{B}$ yields

$$\mathbf{n} \times (\mathbf{n} \times \mathbf{E}) = c\kappa \mathbf{n} \times \mathbf{B} = -c^2 \kappa^2 \mu_0 \mathbf{D} = \mathbf{n}(\mathbf{n} \cdot \mathbf{E}) - \mathbf{n}^2 \mathbf{E}. \quad (4.11)$$

Making use of $D_i = \epsilon_{ik}(\omega)\epsilon_0 E_k$ and $c^2 \kappa^2 \epsilon_0 \mu_0 = 1$, we obtain the eigenvalue equation

$$[\mathbf{n}^2 \delta_{ik} - n_i n_k - \epsilon_{ik}(\omega)] E_k = 0. \quad (4.12)$$

The equation for the wave surface is now given by

$$f(\omega, \mathbf{n}) = \det[\mathbf{n}^2 \delta_{ik} - n_i n_k - \epsilon_{ik}(\omega)] = 0. \quad (4.13)$$

Due to the frequency dependence of ϵ_{ik} , ω will no longer be homogeneous in $\mathbf{k}$.

Although our arguments are easily generalized to the case of complex ϵ_{ik} , we shall restrict ourselves to the most important case of negligible absorption, i.e. real $\epsilon_{ik}(\omega) = \epsilon_{ki}(\omega)$.

Under these conditions, we may choose the eigenvector components $E_k^{(\alpha)}$ of the eigenvalue equation to be real-valued, which, remarkably, implies linearly polarized eigenwaves. Setting $E_k = (1/\epsilon_0)\epsilon_{kj}^{-1}D_j$, the eigenvalue equation assumes the equivalent form

$$[\mathbf{n}^2 \epsilon_{ik}^{-1}(\omega) - n_i n_r \epsilon_{rk}^{-1}(\omega) - \epsilon_{ir}^{-1}(\omega) n_r n_k - \delta_{ik}] D_k = 0. \quad (4.14)$$

We have added an extra term $\epsilon_{ir}^{-1} n_r n_k D_k$, which vanishes since $\mathbf{n} \cdot \mathbf{D} = 0$. In this way we may choose the matrix, for which we search the eigenvectors, to be symmetric, which implies that different eigenvectors $\mathbf{D}^{(\alpha)}$ and $\mathbf{D}^{(\beta)}$ are mutually orthogonal.

In order to determine the ray surface, we first prove that the average value of the Poynting vector,

$$\langle \mathbf{S} \rangle = \frac{1}{2\kappa} \mathbf{E} \times \mathbf{H}, \quad (4.15)$$

is perpendicular to the wave surface. Indeed, taking the variation of Maxwell's equations (4.10),

$$\begin{aligned} \kappa c \delta \mathbf{D} &= \delta \mathbf{H} \times \mathbf{n} + \mathbf{H} \times \delta \mathbf{n}, \\ \kappa c \delta \mathbf{B} &= \delta \mathbf{n} \times \mathbf{E} + \mathbf{n} \times \delta \mathbf{E}, \end{aligned}$$

and inserting Maxwell's equations once more, we obtain

$$\kappa c (\mathbf{E} \cdot \delta \mathbf{D} + \mathbf{H} \cdot \delta \mathbf{B}) = \kappa c (\mathbf{B} \cdot \delta \mathbf{H} + \mathbf{D} \cdot \delta \mathbf{E}) + 2\delta \mathbf{n} \cdot (\mathbf{E} \times \mathbf{H}).$$

From the symmetry of ϵ_{ik} we can conclude that $\mathbf{E} \cdot \delta \mathbf{D} = \mathbf{D} \cdot \delta \mathbf{E}$. Furthermore, because $\mathbf{B} = \mu_0 \mathbf{H}$, we have $\mathbf{H} \cdot \delta \mathbf{B} = \mathbf{B} \cdot \delta \mathbf{H}$, and, therefore,

$$\delta \mathbf{n} \cdot (\mathbf{E} \times \mathbf{H}) = 0. \quad (4.16)$$

Now we prove that the vector for the energy flux velocity divided by c ,

$$\mathbf{s} = \frac{\langle \mathbf{S} \rangle}{\langle \rho_E \rangle c}, \quad (4.17)$$

with

$$\langle \rho_E \rangle = \frac{1}{4} (\mathbf{E} \cdot \mathbf{D} + \mathbf{H} \cdot \mathbf{B}), \quad (4.18)$$

satisfies the normalization condition $\mathbf{s} \cdot \mathbf{n} = 1$. Therefore, it is equal to the ray vector. Indeed, from eqs. (4.9) and (4.10) we find

$$\begin{aligned} \kappa c \mathbf{E} \cdot \mathbf{D} &= \mathbf{E} \cdot (\mathbf{H} \times \mathbf{n}) = \mathbf{n} \cdot (\mathbf{E} \times \mathbf{H}), \\ \kappa c \mathbf{H} \cdot \mathbf{B} &= (\mathbf{n} \times \mathbf{E}) \cdot \mathbf{H} = \mathbf{n} \cdot (\mathbf{E} \times \mathbf{H}), \end{aligned}$$

from which we get $\mathbf{s} \cdot \mathbf{n} = 1$. As a side result we see that also for anisotropic media, in the average, the electric and magnetic energies of plane waves are equal. From the equality

$$\mathbf{s} = \frac{1}{\kappa c} \frac{1}{\mathbf{E} \cdot \mathbf{D}} \mathbf{E} \times \mathbf{H} = \frac{1}{\kappa c} \frac{1}{\mathbf{B} \cdot \mathbf{H}} \mathbf{E} \times \mathbf{H}, \quad (4.19)$$

we obtain

$$\begin{aligned} \mathbf{s} \cdot \mathbf{H} &= 0, & \mathbf{s} \cdot \mathbf{E} &= 0, \\ \mathbf{s} \times \mathbf{D} &= \frac{1}{\kappa c} \mathbf{H}, & \mathbf{s} \times \mathbf{B} &= -\frac{1}{\kappa c} \mathbf{E}. \end{aligned} \quad (4.20)$$

Comparison with Maxwell's equations reveals that the wave surface and the ray surface are related by the following exchanges:

$$\mathbf{n} \leftrightarrow \mathbf{s}, \quad \mathbf{B} \leftrightarrow \mathbf{H}, \quad \mathbf{E} \leftrightarrow \mathbf{D}, \quad \epsilon_{ik} \leftrightarrow \epsilon_{ik}^{-1}. \quad (4.21)$$

The relative orientations of the vectors $\mathbf{E}$, $\mathbf{D}$, $\mathbf{n}$, $\mathbf{s}$, and $\mathbf{B}$ for a plane wave may be illustrated in a simple diagram (see fig. 4.2). According to what we have just derived, $\mathbf{E}$, $\mathbf{D}$, $\mathbf{s}$, and $\mathbf{n}$ are perpendicular to $\mathbf{B}$ and, therefore, lie in the same plane. Furthermore, $\mathbf{n}$ and $\mathbf{D}$ are mutually orthogonal; the same holds for $\mathbf{s}$ and $\mathbf{E}$.

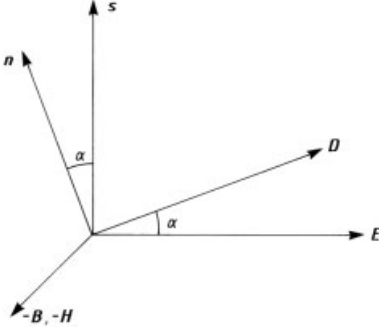


Figure 4.2: Relative orientations of the vectors $\mathbf{B}$, $\mathbf{H}$, $\mathbf{E}$, $\mathbf{D}$, $\mathbf{n}$, and $\mathbf{s}$.

Notice that, in general, in an anisotropic medium the vector perpendicular to the wave vector $\mathbf{k}$ is no longer $\mathbf{E}$ but $\mathbf{D}$.

Now we want to study the eigenvalue problem for plane waves in crystals in more detail. In this context it will turn out to be useful to choose the coordinate axes along the directions of the principal axes of $\epsilon_{ik}(\omega)$. The corresponding equation for the wave surface follows after a short calculation:

$$\begin{aligned} 0 &= f(\omega, \mathbf{n}) = \det[\mathbf{n}^2 \delta_{ij} - n_i n_j - \epsilon_{ij}(\omega)] \\ &= n^2 (\epsilon_1 n_1^2 + \epsilon_2 n_2^2 + \epsilon_3 n_3^2) \\ &\quad - \{n_1^2 \epsilon_1 (\epsilon_2 + \epsilon_3) + n_2^2 \epsilon_2 (\epsilon_1 + \epsilon_3) + n_3^2 \epsilon_3 (\epsilon_1 + \epsilon_2)\} + \epsilon_1 \epsilon_2 \epsilon_3. \end{aligned} \quad (4.22)$$

This equation is of *second* order and not, as one might suspect from a 3×3 determinant, of third order in the quantities n_i^2 . The technical reason for this behavior is the degeneracy of the matrix $n^2 \delta_{ij} - n_i n_j$. The wave surface contains only two shells due to the fact that an electromagnetic wave can only have two independent polarization directions.

For $\mathbf{n}$ fixed, there exist two, in general different, frequencies; and for fixed ω and given direction of the wave vector $\mathbf{k}$, there exist two values for the refractive index $|\mathbf{n}|$. An exception are the *double points* where the two shells of the wave surface intersect and the two values of the refractive index for a given direction of $\mathbf{k}$ coincide.

To determine the polarization directions of the eigenwaves, we start from the eigenvalue equation (4.14) for $\mathbf{D}$:

$$\mathbf{n}^2 \epsilon^{-1}(\omega) \mathbf{D} - \mathbf{n}[\mathbf{n} \cdot \epsilon^{-1}(\omega) \mathbf{D}] = \mathbf{D}. \quad (4.23)$$

We know that $\mathbf{n} \cdot \mathbf{D} = 0$, so the components of $\mathbf{D}$ lie in the plane perpendicular to $\mathbf{n}$. Let $P_\perp$ be the orthogonal projection onto this plane, i.e. $P_\perp(\mathbf{D}) = \mathbf{D}_\perp = \mathbf{D}$, and $P_\perp(\mathbf{n}) = 0$. We apply $P_\perp$ to the eigenvalue equation (4.23) and obtain the new eigenvalue equation

$$[P_\perp \epsilon^{-1}(\omega) P_\perp] \mathbf{D}_\perp = \frac{1}{n^2} \mathbf{D}_\perp. \quad (4.24)$$

Introducing a basis $(e_\alpha)_{\alpha=1,2}$ in the plane perpendicular to $\mathbf{n}$, we may write this equation in the form

$$\left(\epsilon_{\alpha\beta}^{-1}(\omega) - \frac{1}{n^2} \delta_{\alpha\beta} \right) D_\beta = 0. \quad (4.25)$$

The (mutually orthogonal) eigenvectors of this equation are the polarization vectors of the eigenwaves. Equation (4.24) admits a nice interpretation in terms of the normal ellipsoid,

$$\mathbf{D} \cdot \epsilon^{-1}(\omega) \mathbf{D} = 1.$$

If we restrict this equation to vectors $\mathbf{D}_\perp$ orthogonal to $\mathbf{n}$, i.e. if we intersect the normal ellipsoid with the plane $\mathbf{n} \cdot \mathbf{D} = 0$, we obtain the following equation for an ellipse:

$$\mathbf{D}_\perp \cdot P_\perp \epsilon^{-1}(\omega) P_\perp \mathbf{D}_\perp = 1.$$

The principal axes of this ellipse are determined by the eigenvalue equation (4.24),

$$[P_\perp \epsilon^{-1}(\omega) P_\perp] \mathbf{D}_\perp = \frac{1}{n^2} \mathbf{D}_\perp.$$

After scalar multiplication with $\mathbf{D}_\perp$, and using the equation for the ellipse, we see that the eigenvalues n are equal to the lengths of the principal axes of this ellipse.

We can now summarize the construction of the polarization vectors $\mathbf{D}$ and the corresponding refractive indices for an eigenwave propagating along a given direction $\mathbf{n}/|\mathbf{n}| = \mathbf{k}/|\mathbf{k}|$ with frequency ω (see fig. 4.3): Determine the intersection of the normal ellipsoid with the plane perpendicular to $\mathbf{n}$. The orientations of the principal axes of the resulting ellipse are equal to the directions of the polarizations of the eigenwaves, and the lengths of these axes yield the corresponding indices of refraction. From the normal ellipsoid we can also construct the vectors $\mathbf{E}$ and $\mathbf{s}$ corresponding to the polarization vector $\mathbf{D}$: $\mathbf{E}$ is perpendicular to the normal ellipsoid at point $\mathbf{D}$; $\mathbf{s}$ lies in the plane spanned by $\mathbf{n}$, $\mathbf{D}$, and $\mathbf{E}$, is orthogonal to $\mathbf{E}$, and is normalized by $\mathbf{s} \cdot \mathbf{n} = 1$.

If we replace the normal ellipsoid by Fresnel's ellipsoid, $\mathbf{E} \cdot \epsilon(\omega) \mathbf{E} = 1$, and $\mathbf{n}$ by $\mathbf{s}$, we first obtain the ray vectors $\mathbf{s}$ and the polarization vectors $\mathbf{E}$ of the eigenwaves, and in a second step the vectors $\mathbf{D}$ and $\mathbf{n}$.

Of special importance are those directions $\mathbf{n}/|\mathbf{n}|$ for which the sections with the normal ellipsoid are circles; these are the optic axes. In this case, the polarization vector $\mathbf{D}$ may be

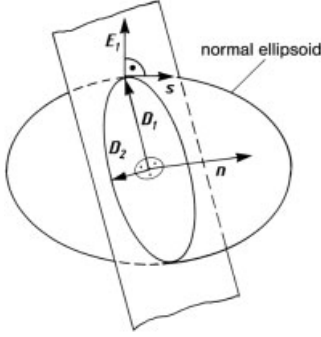


Figure 4.3: Construction of the eigensolutions using the normal ellipsoid.

chosen arbitrarily in the plane perpendicular to $\mathbf{n}$, and for a given orientation of $\mathbf{n}$, we only have one value for the refractive index, $n = \sqrt{\epsilon_2}$ ($\epsilon_1 < \epsilon_2 < \epsilon_3$). Obviously, these special vectors $\mathbf{n}$ correspond to the double point on the wave surface where the two shells of the surface intersect. According to section 4.1, there are no more than two optic axes for the non-cubic case. Therefore, these two shells of the wave surface intersect at one or two pairs of diametric points, and not, as one might suspect, along a one-dimensional curve.

Similar relations hold for the ray vector and the ray surface.

If we allow $\epsilon_{ij}(\omega)$ to be complex, the eigen-directions $\mathbf{D}^{(\alpha)}$ are also complex and the eigenwaves are elliptically polarized. One can show that in this case there are no self-intersections of the wave surface. However, if $\text{Im } \epsilon_{ij} \ll \text{Re } \epsilon_{ij}$, the polarization ellipses are stretched to a very oblong shape, almost like a line, and for certain directions the two shells of the wave surface approach each other closely.

In the following two sections we will discuss in more detail the propagation of light in isotropic crystals that are optically uniaxial or biaxial. The most important results that we will encounter in this context are summarized in table 4.2.

4.3 Optically uniaxial crystals

In the cubic case, $\epsilon_{ij} = \epsilon(\omega)\delta_{ij}$, the equation for the wave surface, eq. (4.22), degenerates to

$$\epsilon n^4 - 2\epsilon^2 n^2 + \epsilon^3 = \epsilon(n^2 - \epsilon)^2 = 0, \quad (4.26)$$

and we obtain a sphere that counts twice. The refractive index $n = \sqrt{\epsilon}$ is independent of the polarization and the direction of propagation, and the ray vector $\mathbf{s} = \mathbf{n}/n^2$ is always parallel to $\mathbf{n}$. There exist also eigenwaves with circular and elliptical polarization.

More complicated, but also more interesting, is the case of optically uniaxial crystals, $\epsilon_1 = \epsilon_2 = \epsilon_\perp$. Such a crystal is called *positive* if $\epsilon_3 > \epsilon_\perp$, and *negative* for $\epsilon_3 < \epsilon_\perp$ (see fig. 4.4). The normal ellipsoid is a ellipsoid of revolution, and the 3-axis coincides with the optic axis. The wave surface (4.22) simplifies to

$$\begin{aligned} n^2(\epsilon_\perp n_\perp^2 + \epsilon_3 n_3^2) - \{\epsilon_\perp(\epsilon_3 + \epsilon_\perp)n_\perp^2 + 2\epsilon_3\epsilon_\perp n_3^2\} + \epsilon_\perp^2 \epsilon_3 \\ = (n^2 - \epsilon_\perp)(\epsilon_\perp n_\perp^2 + \epsilon_3 n_3^2 - \epsilon_\perp \epsilon_3) = 0, \end{aligned} \quad (4.27)$$

where $n_\perp^2 = n_1^2 + n_2^2$. This corresponds to a sphere of radius $\sqrt{\epsilon_\perp}$ together with a circumscribed or inscribed ellipsoid of revolution with axis lengths $1/\sqrt{\epsilon_3}$, $1/\sqrt{\epsilon_3}$, and $1/\sqrt{\epsilon_\perp}$.

Table 4.2: Properties of isotropic and anisotropic crystal systems.

Isotropic		Anisotropic	
		Uniaxial	Biaxial
Crystal systems	Cubic	Trigonal, tetragonal, hexagonal	Orthorhombic, monoclinic, triclinic
Polarization of eigenwaves	Arbitrary elliptic polarization	Only linear polarization, except for $\mathbf{n}$ parallel to the optic axis	Only linear polarization, except for $\mathbf{n}$ parallel to one of the two optic axes
Ray vector $\mathbf{s}$	Normal propagation of light: $\mathbf{s}$ is always parallel to $\mathbf{n}$	Generically, one ordinary ray with $\mathbf{s}$ parallel to $\mathbf{n}$ and polarization orthogonal to the principal plane and one extraordinary ray: $\mathbf{s}$ is not parallel to $\mathbf{n}$, polarization in the principal plane	Generically, two extraordinary rays, $\mathbf{s}$ is not parallel to $\mathbf{n}$
Refraction	Snell's law of refraction holds	Generically, double refraction; one ordinary ray satisfying Snell's law of refraction, and one extraordinary ray, violating Snell's law of refraction	Generically, double refraction; two extraordinary rays both violating Snell's law; conical refraction if $\mathbf{n}$ is parallel to an optic axis

The waves corresponding to the sphere and the ellipsoid are called *ordinary* and *extraordinary* waves, respectively, and we also talk about ordinary and extraordinary rays. For the ordinary ray, $\mathbf{s}$ is always parallel to $\mathbf{n}$, like in the isotropic case, while for the extraordinary ray, $\mathbf{s}$ and $\mathbf{n}$ have different directions.

The plane spanned by the optic axis $\mathbf{e}_3$ and the vector of the refractive index $\mathbf{n}$ is called the *principal section*. The polarizations of the ordinary and extraordinary waves can be deduced from the construction of the normal ellipsoid, as described in the previous section. The normal ellipsoid is an ellipsoid of revolution with axis lengths $\sqrt{\epsilon_{\perp}}$, $\sqrt{\epsilon_{\perp}}$, and $\sqrt{\epsilon_3}$. The section of the normal ellipsoid with an arbitrary plane through the center yields an ellipse perpendicular to $\mathbf{n}$. The length of the axis perpendicular to the principal section is always $n = \sqrt{\epsilon_{\perp}}$; the other axis of the ellipse lies within the principal section.

As one might expect from symmetry considerations, the polarization of the ordinary wave is perpendicular to the principal section, while that of the extraordinary wave is parallel to the principal section. These statements may also be derived analytically from the following

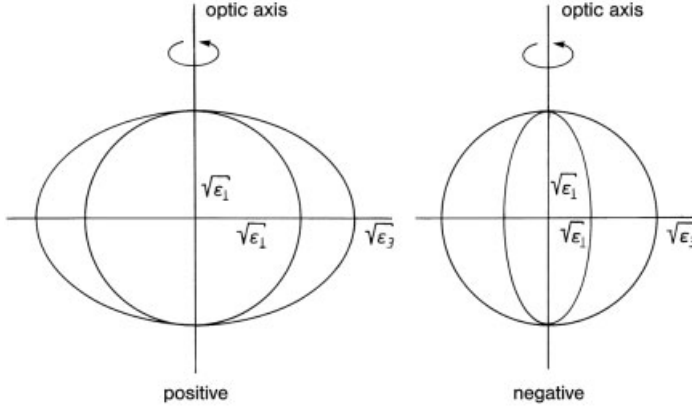


Figure 4.4: Wave surfaces for positive and negative optically uniaxial crystals.

representation of the eigenvalue problem (4.14) for D :

$$\begin{pmatrix} (n_1^2 - n^2)/\epsilon_{\perp} + 1 & 0 & 0 \\ 0 & (n_2^2 - n^2)/\epsilon_{\perp} + 1 & 0 \\ 0 & 0 & (n_3^2 - n^2)/\epsilon_3 + 1 \end{pmatrix} \begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} = 0.$$

For the ordinary ray we have $\mathbf{s} = \mathbf{n}/\epsilon_{\perp}$. Without loss of generality, we can place $\mathbf{n}$ into the 1–3 plane and obtain, from the equation for the wave surface, for the extraordinary ray

$$\frac{n_3^2}{\epsilon_{\perp}} + \frac{n_1^2}{\epsilon_3} = 1. \quad (4.28)$$

Because $\mathbf{s}$ is perpendicular to the wave surface and $\mathbf{s} \cdot \mathbf{n} = 1$, we obtain

$$s_1 = \frac{n_1}{\epsilon_3}, \quad s_3 = \frac{n_3}{\epsilon_{\perp}}, \quad (4.29)$$

such that

$$\tan \beta = \frac{n_1}{n_3} = \frac{\epsilon_3}{\epsilon_{\perp}} \tan \beta' = \frac{\epsilon_3}{\epsilon_{\perp}} \frac{s_1}{s_3}, \quad (4.30)$$

where β and β' are the angles between the optic axis $\mathbf{e}_3$ and $\mathbf{n}$ and $\mathbf{s}$, respectively. Furthermore,

$$\frac{1}{n^2} = \frac{\cos^2 \beta}{\epsilon_{\perp}} + \frac{\sin^2 \beta}{\epsilon_3}. \quad (4.31)$$

We should comment on the following special cases:

1. $\mathbf{n}$ parallel to the optic axis. The principal section degenerates to a straight line; the ordinary and extraordinary waves have the same phase velocity $c/\sqrt{\epsilon_{\perp}}$ and the same ray vector $\mathbf{s} = \mathbf{e}_3/\sqrt{\epsilon_{\perp}}$. There exist elliptically polarized eigenwaves and, for a linearly polarized wave, the orientation of the polarization remains unchanged.

2. *$\mathbf{n}$ orthogonal to the optic axis.* For the ordinary and the extraordinary rays we have: $n_o = \sqrt{\epsilon_\perp}$, $n_e = \sqrt{\epsilon_3}$, $\mathbf{s}_o = \mathbf{n}/\epsilon_\perp$, and $\mathbf{s}_e = \mathbf{n}/\epsilon_3$. The ordinary and the extraordinary rays have the same direction but differ with respect to their group and phase velocities. An initially linearly polarized wave will, in general, become elliptically polarized due to the different propagation velocities of the two polarization components within the crystal.

In general, $\mathbf{s}_o$ and $\mathbf{s}_e$ are not parallel. For a fixed orientation of $\mathbf{n}$ and fixed ω there exist two rays. When a ray with given wave vector $\mathbf{k}$ hits the surface of an anisotropic crystal, there will, in general, appear two refracted rays. This phenomenon is called *double refraction*.

4.4 Optically biaxial crystals

For the case of optically biaxial crystals with $\epsilon_1 < \epsilon_2 < \epsilon_3$, the wave surface is more complicated. We may visualize its form by calculating its intersection with the coordinate planes. So, in eq. (4.22),

$$n^2(\epsilon_1 n_1^2 + \epsilon_2 n_2^2 + \epsilon_3 n_3^2) - n_1^2 \epsilon_1 (\epsilon_2 + \epsilon_3) - n_2^2 \epsilon_2 (\epsilon_1 + \epsilon_3) - n_3^2 \epsilon_3 (\epsilon_1 + \epsilon_2) + \epsilon_1 \epsilon_2 \epsilon_3 = 0,$$

we set $n_1 = 0$, $n_2 = 0$, and $n_3 = 0$, one by one. For example, the intersection with the 1–2 plane is given by

$$n^2(\epsilon_1 n_1^2 + \epsilon_2 n_2^2) - n_1^2 \epsilon_1 (\epsilon_2 + \epsilon_3) - n_2^2 \epsilon_2 (\epsilon_1 + \epsilon_3) + \epsilon_1 \epsilon_2 \epsilon_3 = (n^2 - \epsilon_3)(\epsilon_1 n_1^2 + \epsilon_2 n_2^2 - \epsilon_1 \epsilon_2) = 0. \quad (4.32)$$

Hence, the intersection consists of a circle and an ellipse. Because of the ordering $\epsilon_1 < \epsilon_2 < \epsilon_3$, the ellipse lies inside the circle for $n_3 = 0$. For $n_1 = 0$ the circle lies inside the ellipse; and for $n_2 = 0$ the ellipse and the circle intersect at two opposite pairs of points (see fig. 4.5).

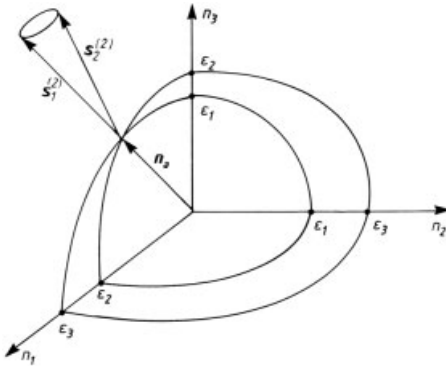


Figure 4.5: Wave surfaces for optically biaxial crystals.

These intersection points are just the four double points of the wave surface, and the corresponding directions (in the 1–2 plane) are just the optic axes of the crystal, for which the intersections with the normal ellipsoid $x^2/\epsilon_1 + y^2/\epsilon_2 + z^2/\epsilon_3 = 1$ consist of circles and the two eigenvalues are degenerate. Indeed, the intersection of

$$\frac{n_1^2}{\epsilon_2} + \frac{n_3^2}{\epsilon_2} = 1 \quad \text{and} \quad \frac{n_1^2}{\epsilon_3} + \frac{n_3^2}{\epsilon_1} = 1$$

yields

$$n_1^2 \left(\frac{1}{\epsilon_2} - \frac{1}{\epsilon_3} \right) = n_3^2 \left(\frac{1}{\epsilon_1} - \frac{1}{\epsilon_2} \right),$$

i.e.

$$\frac{n_1}{n_3} = \tan \theta = \pm \sqrt{\frac{1/\epsilon_1 - 1/\epsilon_2}{1/\epsilon_2 - 1/\epsilon_3}} = \pm \sqrt{\frac{\epsilon_3(\epsilon_2 - \epsilon_1)}{\epsilon_1(\epsilon_3 - \epsilon_2)}}, \quad (4.33)$$

which agrees with the result obtained in section 4.1.

The shells of the ray surfaces intersect transversely only at isolated points. Therefore, in the vicinity of the intersection points, they are cone-like such that the tips of the two cones coincide. The orientations of the ray vectors are given by the normal directions of the wave surface. As fig. 4.5 illustrates, there are infinitely many ray vectors that correspond to the singular vector $\mathbf{n}_A$ that sits at the double points and is directed towards the optic axis. All these ray vectors lie on the surface of the cone. The optic axis $\mathbf{n}_A$ itself lies on this surface, and, because of the degeneracy of the eigenmodes under rotations around the optic axis, the intersection curve of this ray cone with a plain perpendicular to the vector $\mathbf{n}_A$ has to be a circle. This agrees with the construction, described in section 4.2, of the vector $\mathbf{s}$ from the normal ellipsoid, if the intersection is circular. The projection of the ray cone onto the 1–3 plane is shown in fig. 4.5. In this plane the two ray vectors are

$$\mathbf{s}^{(1)} = \frac{1}{\epsilon_2}(n_1, n_3) \quad \text{and} \quad \mathbf{s}^{(2)} = \left(\frac{n_1}{\epsilon_3}, \frac{n_3}{\epsilon_1} \right). \quad (4.34)$$

From this representation we easily obtain the cone angle in the 1–3 plane. For an optically uniaxial crystal, the ray cone degenerates into a single ray.

The non-uniqueness of ray vectors for a given normal vector $\mathbf{n}_A$ leads to the remarkable phenomenon of *inner conical refraction*, which may be phrased as follows (see also fig. 4.6, left): Suppose we are given a biaxial crystal whose surface is cut perpendicular to its optic axis $\mathbf{n}_A$. A thin ray of unpolarized light hits the surface of this crystal parallel to $\mathbf{n}_A$. In this case the wave vector will be parallel to $\mathbf{n}_A$ also inside the crystal, and there will be a whole cone of corresponding ray vectors. The light leaving the crystal at the opposite side will be concentrated along the surface of a circular cylinder.

There also exists the phenomenon of *outer conical refraction*: The ray surface consists of two shells and a whole cone of wave vectors corresponds to the exceptional direction $\mathbf{s}_A$ (see fig. 4.6, right).

4.5 Reflection and refraction at interfaces

The phenomena of reflection and refraction, which may occur when an arbitrary, not necessarily electromagnetic, wave hits the interface between two different media, can be quite complicated, in particular for non-isotropic media. We shall first derive a general idea of the phenomena that we might expect.

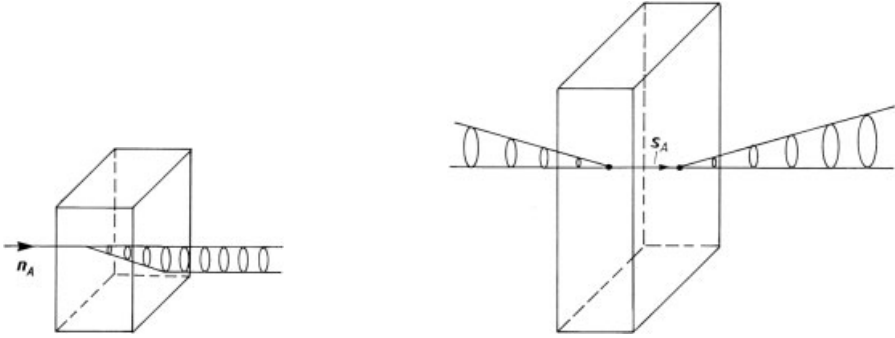


Figure 4.6: (Left) Principle of inner conical refraction. (Right) Principle of outer conical refraction.

As the incoming waves consist, in general, of a thin ray bundle which hits the interface only in the vicinity of a point, we may, for simplicity, assume the interface to be flat with normal vector $\mathbf{N}$. Furthermore, any ray bundle can be interpreted as a superposition of plane waves.

Therefore, without loss of generality, we represent the incoming wave as a plane eigen-wave $v_\alpha \exp[i(\mathbf{k} \cdot \mathbf{x} - \omega t)]$, which approaches the interface from medium 1.

The outgoing waves will then be linear combinations of eigenwaves. First, we want to determine the possible frequencies and wave vectors $(\omega'_\beta, \mathbf{k}'_\beta, \omega''_\gamma, \mathbf{k}''_\gamma)$ of the *reflected* waves, returning into medium 1, and the *refracted* waves, entering medium 2, respectively (see fig. 4.7).

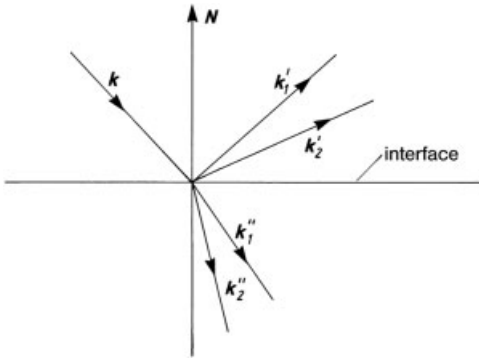


Figure 4.7: The different wave vectors involved in the reflection and refraction at the interface between media 1 and 2.

There are certain continuity conditions for the displacement of the waves that hold everywhere on the interface and for all times. As one might easily convince oneself, these conditions are only fulfilled if for all waves the frequencies as well as the longitudinal components, which are tangential to the interface, are equal:

$$\begin{aligned} \omega &= \omega'_\beta = \omega''_\gamma \\ \mathbf{k}_l &= \mathbf{k}'_{\beta l} = \mathbf{k}''_{\gamma l} \end{aligned} \quad \text{for all } \beta, \gamma. \quad (4.35)$$

We still have to determine the transverse components ($\mathbf{k}'_{\beta t}$ and $\mathbf{k}''_{\gamma t}$) of the wave vectors, which are perpendicular to the interface.

The tips of the wave vectors in both media, 1 and 2, have to lie on the wave surface corresponding to the frequency ω . If the wave surface has N shells, there are exactly N wave vectors in each medium, i.e. $\mathbf{k}'_{\beta}$ ($\alpha = 1, \dots, N$) and $\mathbf{k}''_{\gamma}$ ($\gamma = 1, \dots, N$). Their tangential components are given by $\mathbf{k}_l$ and their components parallel to the normal $\mathbf{N}$ have the same sign as the normal component $\mathbf{k}_t$ of $\mathbf{k}$ in medium 2 and the opposite sign in medium 1.

Denoting the corresponding eigen-amplitudes in medium 1 and 2 with v_{α} , v_{β} , and w_{γ} , we arrive at the following ansatz:

$$\begin{aligned} \text{medium 1} \quad u &= v_{\alpha} \exp[i(\mathbf{k} \cdot \mathbf{x} - \omega t)] + \sum_{\beta=1}^N A_{\alpha\beta} v_{\beta} \exp[i(\mathbf{k}'_{\beta} \cdot \mathbf{x} - \omega t)], \\ \text{medium 2} \quad u &= \sum_{\gamma=1}^N B_{\alpha\gamma} w_{\gamma} \exp[i(\mathbf{k}''_{\gamma} \cdot \mathbf{x} - \omega t)]. \end{aligned} \quad (4.36)$$

The coefficients $A_{\alpha,\beta}$ and $B_{\alpha,\gamma}$ can be determined from the continuity conditions along the interface.

For some values of $\mathbf{k}_l$, it may happen that there are less than N eigenmodes with real normal components $\mathbf{k}''_t$ or $\mathbf{k}'_t$. In this case the solution of the wave surface equation yields complex values for $\mathbf{k}'_t$ or $\mathbf{k}''_t$, which correspond to exponentially decreasing waves. A famous example of such a situation is the phenomenon of *total reflection*.

Because all the involved wave vectors share the same component parallel to the interface, they are coplanar. In general, however, the corresponding ray vectors $\mathbf{s}_{\alpha}$, $\mathbf{s}'_{\beta}$, and $\mathbf{s}''_{\gamma}$ are not coplanar.

After these preliminary remarks, we now discuss the reflection and refraction of electromagnetic waves at an interface between two not necessarily isotropic media.

In this case the wave surface consists of two shells, i.e. $N = 2$, and for the general case we would expect two reflected and two refracted rays. The continuity conditions at the interface require the continuity of the tangential components of $\mathbf{E}$ and $\mathbf{H}$:

$$\begin{aligned} \mathbf{N} \times (\mathbf{E}_1 - \mathbf{E}_2) &= 0, \\ \mathbf{N} \times (\mathbf{H}_1 - \mathbf{H}_2) &= 0, \end{aligned} \quad (4.37)$$

as well as the continuity of the normal components of $\mathbf{D}$ and $\mathbf{B}$:

$$\begin{aligned} \mathbf{N} \cdot (\mathbf{D}_1 - \mathbf{D}_2) &= 0, \\ \mathbf{N} \cdot (\mathbf{B}_1 - \mathbf{B}_2) &= 0. \end{aligned} \quad (4.38)$$

Taking into account Maxwell's equations, the continuity conditions (4.38) already follow from (4.37):

$$\mathbf{k} \times \mathbf{E} = \kappa\omega\mathbf{B}, \quad \mathbf{k} \times \mathbf{H} = -\kappa\omega\mathbf{D}. \quad (4.39)$$

Scalar multiplication of these equations with $\mathbf{N}$ yields the required result.

So, we are left with the continuity conditions of the tangential components of $\mathbf{E}$ and $\mathbf{H}$. They imply also that the normal component of the Poynting vector, $\mathbf{S} = (1/\kappa)\mathbf{E} \times \mathbf{H}$, is continuous. Indeed, this follows also from physical reasons, as otherwise energy would be accumulated in the interface.

Using Maxwell's equations (4.39) as well as the material equations $\mathbf{B} = \mu_0\mu\mathbf{H}$ and $\mathbf{D} = \epsilon_0\epsilon\mathbf{E}$, we may express all other fields in terms of $\mathbf{E}$ (or $\mathbf{H}$). Thus, we can try an ansatz of the following type to solve the problem:

$$\begin{aligned} \text{medium 1} \quad \mathbf{E} &= \mathbf{e}_\alpha \exp[i(\mathbf{k} \cdot \mathbf{x} - \omega t)] + \sum_{\beta=1}^2 A_{\alpha\beta} \mathbf{e}'_\beta \exp[i(\mathbf{k}'_\beta \cdot \mathbf{x} - \omega t)], \\ \text{medium 2} \quad \mathbf{E} &= \sum_{\gamma=1}^2 B_{\alpha\gamma} \mathbf{e}''_\gamma \exp[i(\mathbf{k}''_\gamma \cdot \mathbf{x} - \omega t)]. \end{aligned} \quad (4.40)$$

Here $\mathbf{e}_\alpha$, $\mathbf{e}'_\beta$, and $\mathbf{e}''_\gamma$ are polarization vectors.

For a fixed value of α , the four coefficients $A_{\alpha,\beta}$ and $B_{\alpha,\beta}$ can be calculated from the four independent continuity conditions (4.37).

In this way we can determine the directions as well as the intensities of the reflected and the refracted rays.

4.6 Fresnel's equations

In this section we will explicitly perform the calculation for the simplest case where both media are isotropic and free of absorption:

$$\epsilon_{ij}^1 = \epsilon_1 \delta_{ij}, \quad \epsilon_{ij}^2 = \epsilon_2 \delta_{ij}, \quad n_1 = \sqrt{\epsilon_1}, \quad n_2 = \sqrt{\epsilon_2}.$$

We also assume that $\mu_{ij}^1 = \mu_{ij}^2 = \delta_{ij}$.

For this case the wave surfaces degenerate to spheres (counting twice), both eigenmodes of oscillation are degenerate, and we have $\mathbf{k}'_1 = \mathbf{k}'_2 =: \mathbf{k}'$ and $\mathbf{k}''_1 = \mathbf{k}''_2 =: \mathbf{k}''$. Furthermore, the ray vectors corresponding to the wave vectors are parallel to $\mathbf{k}'$ or $\mathbf{k}''$ (see fig. 4.8).

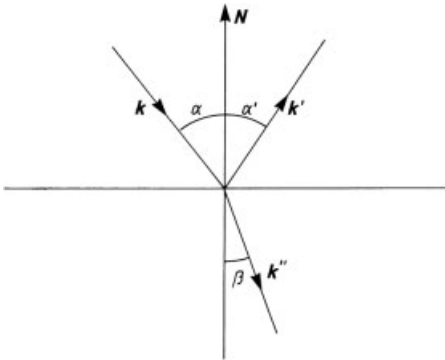


Figure 4.8: Reflection and refraction at an interface between two isotropic and absorption-free media 1 and 2.

Since

$$\mathbf{k}_l = \mathbf{k}'_l = \mathbf{k}''_l$$

and

$$|\mathbf{k}| = |\mathbf{k}'| = n_1 \frac{\omega}{c}, \quad |\mathbf{k}''| = n_2 \frac{\omega}{c}, \quad (4.41)$$

we obtain for

$$\sin \alpha = \frac{|\mathbf{k}_l|}{|\mathbf{k}|}, \quad \sin \alpha' = \frac{|\mathbf{k}'_l|}{|\mathbf{k}|}, \quad \sin \beta = \frac{|\mathbf{k}''_l|}{|\mathbf{k}|}$$

the *reflection law*

$$\alpha = \alpha' \quad (4.42)$$

as well as *Snell's law of refraction*

$$n_1 \sin \alpha = n_2 \sin \beta. \quad (4.43)$$

Furthermore, we obtain from (4.41)

$$\begin{aligned} \mathbf{k}'_t &= -\mathbf{k}_t, \\ \mathbf{k}''_t &= \sqrt{k'^2 - k_t^2} = \sqrt{(n_2^2/n_1^2)k^2 - k_t^2} = k\sqrt{(n_2^2/n_1^2) - \sin^2 \alpha}. \end{aligned} \quad (4.44)$$

For $n_1 > n_2$ there are no solutions of the condition $n_1 \sin \alpha > n_2$ with a real value of k''_t ; this describes the case of total reflection at the interface in front of an optically thinner medium.

Finally, we calculate the ratios of intensities. For this purpose, we distinguish two cases:

1. *Transverse polarization.* The vector $\mathbf{E}$ of the incoming wave oscillates perpendicularly to the so-called *plane of incidence* spanned by $\mathbf{k}$ and $\mathbf{N}$. Hence, it oscillates parallel to the interface. It turns out that in this case also the vectors $\mathbf{E}'$ and $\mathbf{E}''$ of the refracted and reflected waves are perpendicular to the plane of incidence. If we denote the components of the electric field vectors that are perpendicular to the plane of incidence by E , E' , and E'' , the continuity condition for the tangential component of the electric field vectors reads

$$E + E' = E''. \quad (4.45)$$

The vectors $\mathbf{H}$, $\mathbf{H}'$, and $\mathbf{H}''$ of the incoming, reflected, and refracted waves, respectively, oscillate in the plane of incidence and orthogonal to their corresponding wave vectors. Because of eq. (4.39), their lengths are given by

$$\begin{aligned} H &= \frac{1}{\mu_0 \kappa \omega} k E = \frac{n_1}{\mu_0 \kappa c} E, \\ H' &= \frac{n_1}{\mu_0 \kappa c} E', \quad H'' = \frac{n_2}{\mu_0 \kappa c} E'', \end{aligned}$$

and the continuity of the tangential components of the magnetic field vector leads to the condition

$$n_1 E \cos \alpha - n_1 E' \cos \alpha = n_2 E'' \cos \beta. \quad (4.46)$$

2. *Polarization in the plane of incidence.* The E vectors oscillate in the plane of incidence, and the H vectors are orthogonal to the E vectors. In this case one finds for H , H' , and H'' , as well as $E = (\mu_0 \kappa c / n_1) H$, $E' = (\mu_0 \kappa c / n_1) H'$, and $E'' = (\mu_0 \kappa c / n_1) H''$ the continuity conditions

$$H + H' = H'',$$

$$\left(\frac{1}{n_1} H - \frac{1}{n_2} H' \right) \cos \alpha = \frac{1}{n_2} H'' \cos \beta.$$

Solving for E' and E'' yields, for the case of transverse polarization,

$$\begin{aligned} \rho_t &= \frac{E'}{E} = \frac{n_1 \cos \alpha - n_2 \cos \beta}{n_1 \cos \alpha + n_2 \cos \beta}, \\ \tau_t &= \frac{E''}{E} = \frac{2n_1 \cos \alpha}{n_1 \cos \alpha + n_2 \cos \beta}, \end{aligned} \quad (4.47)$$

and, for the case of longitudinal polarization,

$$\begin{aligned} \rho_l &= \frac{E'}{E} = \frac{n_2 \cos \alpha - n_1 \cos \beta}{n_2 \cos \alpha + n_1 \cos \beta}, \\ \tau_l &= \frac{E''}{E} = \frac{2n_1 \cos \alpha}{n_2 \cos \alpha + n_1 \cos \beta}. \end{aligned} \quad (4.48)$$

Equations (4.47) and (4.48) are called *Fresnel's equations*.

For the Poynting vectors, we obtain the relations $\mathbf{N} \cdot \mathbf{S} \sim EH \cos \alpha \sim n_1 E^2 \cos \alpha$, $\mathbf{N} \cdot \mathbf{S}' \sim n_1 E'^2 \cos \alpha$, and $\mathbf{N} \cdot \mathbf{S}'' \sim n_2 E''^2 \cos \beta$, from which we can easily calculate the reflection and transmission coefficients:

- E perpendicular to the plane of incidence

$$\begin{aligned} r_t &= \left(\frac{n_1 \cos \alpha - n_2 \cos \beta}{n_1 \cos \alpha + n_2 \cos \beta} \right)^2 = \frac{\sin^2(\alpha - \beta)}{\sin^2(\alpha + \beta)} \\ t_t &= \frac{4n_1 n_2 \cos \alpha \cos \beta}{(n_1 \cos \alpha + n_2 \cos \beta)^2} = \frac{\sin 2\alpha \sin 2\beta}{\sin^2(\alpha + \beta)}. \end{aligned} \quad (4.49)$$

- E in the plane of incidence

$$\begin{aligned} r_l &= \left(\frac{n_2 \cos \alpha - n_1 \cos \beta}{n_2 \cos \alpha + n_1 \cos \beta} \right)^2 \\ &= \left(\frac{\sin 2\alpha - \sin 2\beta}{\sin 2\alpha + \sin 2\beta} \right)^2 = \frac{\tan^2(\alpha - \beta)}{\tan^2(\alpha + \beta)} \\ t_l &= \frac{4n_1 n_2 \cos \alpha \cos \beta}{(n_2 \cos \alpha + n_1 \cos \beta)^2} = \frac{4 \sin 2\alpha \sin 2\beta}{(\sin 2\alpha + \sin 2\beta)^2}. \end{aligned} \quad (4.50)$$

Obviously, $r_t + t_t = r_l + t_l = 1$. For $\alpha + \beta = \pi/2$, we have $\sin 2\alpha = \sin 2\beta$ and $r_l = 0$. The corresponding angle α_B is called the *Brewster angle*; it is determined by the relation $\tan \alpha_B = n_2/n_1$. The ray that is reflected at an angle α_B is polarized transversely with respect to the plane of incidence.

If the direction of the incident wave is orthogonal to the interface, the two cases degenerate into one, and we obtain

$$r = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2, \quad t = \frac{4n_1 n_2}{(n_1 + n_2)^2}. \quad (4.51)$$

If we have the situation of total reflection, the refracted wave is exponentially damped, $\cos \beta$ becomes purely imaginary, and $|E'/E| = 1$. In this case $r_l = r_t = 1$. The totally reflected wave is phase-shifted relative to the incoming wave, while the refracted wave, according to eqs. (4.47) and (4.48), is always in phase with the incoming wave. For an “ordinary” reflected wave, only the phase differences $\Delta\varphi = 0$ and $\Delta\varphi = \pi$ occur.

4.7 The Fabry–Perot interferometer

The *Fabry–Perot interferometer* is one of the most precise instruments in optics and has numerous applications. Its principle is based on the multiple reflection of plane waves. The simple arrangement of this interferometer consists of two highly planar and parallel interfaces which are partly transparent and partly reflective for incoming light (fig. 4.9). Let the optical distance between the two planes be d .

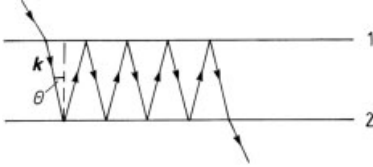


Figure 4.9: Principle of a Fabry–Perot interferometer.

We determine the transmissivity of the total apparatus for a plane wave that hits the first interface almost from above with an amplitude E .

We denote the transmission amplitude for this wave at the first interface by τ_1 , and the transmission amplitude for a wave that hits the second interface from above by τ_2 . Furthermore, ρ is the reflection amplitude of both interfaces for waves that hit the surfaces from between the two planes. For small incident angles we may neglect the influence of polarization.

Before the wave leaves the apparatus from below, it could have been reflected at the inner interfaces $2n$ times. The total amplitude leaving the apparatus is given by

$$E' = \sum_{n=0}^{\infty} E_n, \quad (4.52)$$

with $E_n = \tau_1 [\rho^2 \exp(2ikd/\cos \theta)]^n \tau_2 E$, where $2ikd/\cos \theta$ is just the phase difference after two transitions of the distance between the two plates. From

$$\rho^2 = r \exp[2i\delta(k, \theta)] \quad (4.53)$$

and

$$\varphi = \delta + ikd/\cos \theta, \quad (4.54)$$

we obtain

$$E' = \frac{\tau_1 \tau_2}{1 - r \exp(2i\varphi)} E, \quad (4.55)$$

and the total transmissivity of the apparatus is given by

$$T(k, \theta) = |\tau_1 \tau_2|^2 \frac{1}{|1 - r \exp(2i\varphi)|^2}. \quad (4.56)$$

If the reflection coefficient has a value close to $r = 1$, the transmissivity $T(k, \theta)$ exhibits a sharp peak at $\varphi(k, \theta) = m\pi$ (m integer).

The condition $\varphi = m\pi$ may be realized for fixed k either by changing the distance d , or by varying the incident angle θ . For a so-called *Fabry–Perot etalon*, separation d is fixed. Usually it consists of a highly plane-parallel glass plate with mirrored sides such that the reflection coefficient r is close to $r = 1$.

5 Electro-, magneto- and elastooptical phenomena

An electric or magnetic polarization of some medium may be due not only to an applied electromagnetic field but also to other external influences, like e.g. mechanical deformations. Furthermore, a material may exhibit spontaneous polarization without any external influence. In this chapter we will describe some of the more important polarization effects of this type, but we will restrict ourselves to electric polarization. First, we will discuss effects of zeroth and first order, i.e. electric polarization that is spontaneous or depends linearly on some external influence. Later we will also consider effects of higher order.

5.1 Polarization effects up to first order – optical activity

We use the following ansatz for the electric displacement density of the medium:

$$D_i(\omega, \mathbf{x}) = P_i^{(0)} + \epsilon_{ij}(\omega)\epsilon_0 E_j(\omega, \mathbf{x}) + \lambda_{ijk}(\omega)u_{jk}(\omega, \mathbf{x}) + \gamma_{ijk}(\omega)\epsilon_0 \partial_k E_j(\omega, \mathbf{x}) + \dots \quad (5.1)$$

This ansatz includes the following:

- A spontaneous polarization $P_i^{(0)}$. Materials for which $P_i^{(0)} \neq 0$ are called *pyroelectric*.
- A linear dependence on an electric field. Here ϵ_{ij} is the dielectric permeability, which we have already discussed in detail.
- A linear dependence on the distortion tensor $u_{jk} = \frac{1}{2}(\partial_j u_k + \partial_k u_j)$. Materials for which $\lambda_{ijk} = \lambda_{ikj} \neq 0$ are called *piezoelectric*. Piezoelectricity is known from quartz crystals, which become polarized under distortions. This phenomenon is used in technology to generate very stable oscillating circuits.
- A linear dependence on the gradient of the electric field. The so-called *natural optical activity* is an example of such an effect. When we derived the dielectric material equations, we made the assumption that the applied fields were homogeneous for a volume containing many molecular cores. The condition was $|\mathbf{k}|d \ll 1$, where d is the typical molecular distance. The contribution proportional to $\partial_k E_j(\omega, \mathbf{x})$ is a first correction term to this homogeneity assumption and is particularly important for rapidly fluctuating fields. On the other hand, for very large wavenumbers, $|\mathbf{k}|d \geq 1$, the phenomenological theory of material properties, obtained by averaging over many molecules, becomes meaningless. In this case we are dealing with the scattering of electromagnetic waves at spatially distributed single molecules.

The tensor quantities $P_i^{(0)}$, ϵ_{ij} , λ_{ijk} , and γ_{ijk} are strongly constrained by the spatial symmetry of the medium (which we take to be infinitely extended): the tensors must be invariant under all transformations, which corresponds to a spatial symmetry of the medium. For translations, this is a trivial condition. However, those combinations of rotations and reflections – i.e. elements of the improper rotation group $O(3)$ – which leave the medium invariant are more restrictive. For instance, if a point reflection $\mathbf{x} \rightarrow -\mathbf{x}$ belongs to the symmetries of the medium, the quantities $P_i^{(0)}$, λ_{ijk} , and γ_{ijk} vanish, because a tensor of odd order changes its sign under this operation.

There are a total of 32 different subgroups of $O(3)$, the so-called point groups, which transform a spatial translational lattice into itself. This corresponds to the classification of crystals according to the 32 *crystal classes*.

On the other hand, the seven *crystal systems* follow from the classification of crystals according to $\Gamma \subset O(3)$, the symmetry group of the translational lattice. Obviously, the symmetry group Γ' of a crystal is a subgroup of the symmetry group Γ of its translational lattice. If $\Gamma = \Gamma'$, the crystal is called *holohedral*. In general, a crystal with symmetry group Γ' belongs to the crystal system for which the lattice has the smallest possible symmetry group $\Gamma \supseteq \Gamma'$.

For $P_i^{(0)} \neq 0$, the symmetry group $\Gamma' \subset O(3)$ can only contain rotations around no more than one axis and it should not contain any reflections at planes orthogonal to this axis. Pyroelectricity is only possible for 10 of the 32 crystal classes.

The constraints for piezoelectric crystals are less restrictive. Piezoelectricity is possible for all pyroelectric crystals as well as for 10 more classes, making a total of 20 crystal classes.

Next we consider the polarization effect linear in $\partial_k E_j$ and look for the constraints that correspond to the symmetry principles of the kinetic coefficients and the passivity condition for the quantities $\gamma_{ijk}(\omega)$. In general, the symmetry condition may be written in the form

$$\int d^3x \mathbf{P}(\mathbf{x}) \cdot \delta \mathbf{E}(\mathbf{x}) = \int d^3x \delta \mathbf{P}(\mathbf{x}) \cdot \mathbf{E}(\mathbf{x}). \quad (5.2)$$

For our case it follows that

$$\int d^3x \delta E_i \gamma_{ijk}(\omega) \partial_k E_j = \int d^3x E_i \gamma_{ijk}(\omega) \partial_k \delta E_j,$$

and after partial integration we obtain

$$\gamma_{ijk}(\omega) = -\gamma_{jik}(\omega). \quad (5.3)$$

The passivity condition (2.94),

$$\int dt \int d^3x \mathbf{E} \cdot \frac{\partial \mathbf{P}}{\partial t} \geq 0,$$

implies for a harmonic time dependence:

$$\begin{aligned}
 \operatorname{Re} \int d^3x (-i\omega) E_i^* \gamma_{ikl}(\omega) \partial_l E_k &= \operatorname{Im} \int d^3x E_i^* \omega \gamma_{ikl}(\omega) \partial_l E_k \\
 &= \frac{\omega}{2i} \int d^3x [E_i^* \gamma_{ikl}(\omega) \partial_l E_k - E_i \gamma_{ikl}^*(\omega) \partial_l E_k^*] \\
 &= \frac{\omega}{2i} \int d^3x [E_i^* \gamma_{ikl}(\omega) \partial_l E_k - E_i^* \gamma_{ikl}^*(\omega) \partial_l E_k] \\
 &= \int d^3x E_i^* \operatorname{Im}\{\omega \gamma_{ikl}(\omega)\} \partial_l E_k \geq 0.
 \end{aligned} \tag{5.4}$$

Therefore, in the absence of absorption, we find $\gamma_{ikl}(\omega)$ to be real. In this case the relation between the fields $\mathbf{E}$ and $\mathbf{D}$ for plane waves, taking ordinary dielectric permeability into account, may be written as

$$\begin{aligned}
 D_i(\omega) &= [\epsilon_{ik}(\omega) + i\gamma_{ikl}(\omega)k_l] \epsilon_0 E_k \\
 &= [\epsilon_{ik}(\omega) + i(\omega/c)\gamma_{ikl}(\omega)n_l] \epsilon_0 E_k,
 \end{aligned} \tag{5.5}$$

where $\epsilon_{ik}(\omega)$ is real. In addition to the dielectric permeability, we find a $\mathbf{k}$ -dependent imaginary part. As $\gamma_{ikl}k_l$ is antisymmetric with respect to the indices i and k , this term can also be written in the form

$$\gamma_{ikl}k_l = \epsilon_{ikr}g_r. \tag{5.6}$$

Here g_r is the so-called *gyration tensor*, which depends on the wave vector k_l or the vector of the refractive index $n_l = (c/\omega)k_l$:

$$g_r = g_{rl}n_l. \tag{5.7}$$

The matrix g_{rl} is also constrained by the symmetry of the medium. For an isotropic medium or a cubic crystal we have

$$g_{rl} = \text{const.} \times \delta_{rl}. \tag{5.8}$$

For a detailed description of the eigenwave solutions it is helpful to introduce the tensor ζ_{ik} , which is the inverse of the Hermitian tensor $\epsilon_{ik} + i\epsilon_{ikr}g_r$. For most cases we have $|\mathbf{g}|, |\mathbf{G}| \ll |\epsilon_{ik}|, |\epsilon_{ik}^{-1}|$, which, in first approximation, leads to

$$\zeta_{ik} = \epsilon_{ik}^{-1} + i\epsilon_{ikr}G_r, \tag{5.9}$$

where

$$\epsilon_{ikr}G_r = -\epsilon_{ij}^{-1}\epsilon_{jsr}g_r\epsilon_{sk}^{-1}. \tag{5.10}$$

Projecting the eigenvalue equation (4.23), $n^2\zeta(\omega)\mathbf{D} - \mathbf{n}[\mathbf{n} \cdot \zeta(\omega)\mathbf{D}] = \mathbf{D}$, onto the plane orthogonal to $\mathbf{n}$, we get

$$\left(\zeta_{ab} - \frac{1}{n^2} \delta_{ab} \right) D_\beta = 0 \quad (\alpha, \beta = 1, 2)$$

or

$$\left(\epsilon_{\alpha\beta}^{-1} + i\epsilon_{\alpha\beta} G_r - \frac{1}{n^2} \delta_{\alpha\beta} \right) D_\beta = 0. \quad (5.11)$$

In the two-dimensional space orthogonal to $\mathbf{n}$, we choose a basis such that $\epsilon_{\alpha\beta}^{-1}$ becomes diagonal. Denoting the diagonal elements by $1/n_1^2$ and $1/n_2^2$, we obtain the following eigenvalue equation:

$$\begin{pmatrix} 1/n_1^2 - 1/n^2 & +iG_n \\ -iG_n & 1/n_2^2 - 1/n^2 \end{pmatrix} \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = 0.$$

Here, G_n is the component of $\mathbf{G}$ that is parallel to $\mathbf{n}$. The eigenvalues are

$$\frac{1}{n_\pm^2} = \frac{1}{2} \left(\frac{1}{n_1^2} + \frac{1}{n_2^2} \right) \pm \sqrt{\frac{1}{4} \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)^2 + G_n^2}, \quad (5.12)$$

and the eigenvectors satisfy

$$D_2^\pm = i\rho_\pm D_1^\pm,$$

where

$$\rho_\pm = \frac{1}{2G_n} \left[\frac{1}{n_1^2} - \frac{1}{n_2^2} \mp \sqrt{\left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)^2 + 4G_n^2} \right]. \quad (5.13)$$

From these results we deduce that $D_2^\pm/D_1^\pm$ is purely imaginary, the eigenwaves are elliptically polarized, and the axes of the polarization ellipses are collinear to the principal orientations of $\epsilon_{\alpha\beta}^{-1}$. Because of the relation $\rho_+\rho_- = -1$, the polarization ellipses of the two eigenwaves are related to each other by a rotation of 90° .

In general, the effect of the imaginary part, $i\epsilon_{ik}G_r$, is small, i.e. $\rho_+ \ll 1$, and the polarization ellipses are extremely prolonged and almost resemble linear polarizations.

However, this is no longer true if the eigenvalues of $\epsilon_{\alpha\beta}^{-1}$ coincide. If $\epsilon_{ik} = \epsilon\delta_{ik}$, this is the case for all $\mathbf{n}$. The term $i\epsilon_{ik}G_r$ leads to a qualitatively different behavior because it reduces the symmetry. Writing $1/n_1^2 = 1/n_2^2 = 1/n_0^2 = 1/\epsilon$ we now obtain

$$\frac{1}{n_\pm^2} = \frac{1}{n_0^2} \pm G_n \quad \text{and} \quad \rho_\pm = \mp 1. \quad (5.14)$$

The eigenvectors are circularly polarized and have different phase velocities. Furthermore, due to eq. (5.10), we have in first approximation

$$G_i = -\frac{1}{n_0^4} g_i, \quad (5.15)$$

and, because of the symmetry of the medium, $g_i = \alpha n_i$ and $G_n = \alpha/n_0^3$ (compare eq. (5.8)).

Plane-polarized waves can be written as a superposition of waves with a circular polarization. If a plane-polarized wave passes through such a medium, the different phase velocities of the right- and left-circular components lead to a phase shift between the two planar components and to a rotation of the plane of polarization.

Let us consider a wave propagating along the 3-axis and being plane-polarized along the 1-direction. Then we have

$$\begin{aligned} D_1 &= \frac{1}{2}A(e^{ik_+z} + e^{ik_-z}) = A e^{ikz} \cos(\kappa z), \\ D_2 &= \frac{1}{2}iA(-e^{ik_+z} + e^{ik_-z}) = A e^{ikz} \sin(\kappa z), \end{aligned} \quad (5.16)$$

where

$$\begin{aligned} k_{\pm} &= \frac{\omega}{c}n_{\pm} = \frac{\omega}{c}(n_0 \mp \frac{1}{2}G_n n_0^3) = \frac{\omega}{c}(n_0 \mp \frac{1}{2}\alpha), \\ k &= \frac{1}{2}(k_+ + k_-) = \frac{\omega}{c}n_0, \end{aligned} \quad (5.17)$$

$$\kappa = \frac{1}{2}(k_+ - k_-) = -\frac{\omega}{c}\frac{\alpha}{2}. \quad (5.18)$$

Hence, the angle φ between the plane of polarization and the 1-direction is given by

$$\varphi = \kappa z = -\frac{\omega}{c}\frac{\alpha z}{2}. \quad (5.19)$$

This rotation of the plane of polarization is called the *natural optical activity* of a medium. As we discussed at the beginning of this section, this phenomenon can only occur if the medium is mirror-asymmetric.

For instance, tartaric acid is a molecule existing in two different forms that are mirror images of each other. Living organisms only produce one of these “enantiomorphic” forms. A solution of natural tartaric acid is an isotropic medium that exhibits optical activity. The rotation of the plane of polarization per unit length is proportional to the concentration of the solution and can be used to measure this concentration. The sense of rotation of the plane of polarization around the direction of propagation $\mathbf{k}/k$ is independent of $\mathbf{k}$. Hence, if a plane-polarized wave passes through a medium in both directions, e.g. because it is reflected at the far side, there will be no net change of polarization.

5.2 Polarization effects of higher order

For many situations of theoretical and practical importance, terms of quadratic or even higher powers in the external influences on the electric polarization are crucial. In general, these terms may be incorporated by letting the coefficients ϵ_{ij} , λ_{ijk} , and γ_{ijk} , introduced in section 5.1, depend on these external influences in first or higher order. In the present context we restrict ourselves to a discussion of the particularly important dependence of the electric permeability ϵ_{ik} on various quantities.

First of all we look at the dependences on distortions and shear flows. Finally, ϵ_{ij} may also depend on the electromagnetic fields $\mathbf{E}$ and $\mathbf{B}$. For very intense radiation, such a dependence on the field strengths can lead to measurable effects, which we will partly discuss later in the context of nonlinear optics. Here we will only consider special nonlinear effects, where an applied homogeneous electromagnetic field changes the optical properties of a medium.

5.2.1 Dependence on distortions

The most general linear dependence of the quantity ϵ_{ij} on the distortion u_{rs} is of the form

$$\epsilon_{ij} = \epsilon_{ij}^{(0)} + \xi_{ijrs} u_{rs}. \quad (5.20)$$

The effects are particularly dramatic for a medium that is isotropic in its undistorted state. In this case the distortion may lead to an asymmetry and double refraction. On placing the distorted medium between two polarization filters with orthogonal polarization directions, the distortions become directly visible upon illumination. This effect is used in material sciences to study the strength and breaking of materials. For an isotropic medium we have

$$\epsilon_{ij}(\omega) = \epsilon^{(0)} \delta_{ij} + \lambda_1(\omega) u_{rr} \delta_{ij} + \lambda_2(\omega) (u_{ij} - \delta_{ij} u_{rr}).$$

The term with coefficient λ_1 leads to small changes of the isotropic refractive index, while the term with coefficient λ_2 gives rise to a qualitatively different anisotropic behavior of the medium. Coefficients λ_1 and λ_2 are called *elasto-optical constants*.

5.2.2 Dependence on shear flows

Anisotropy effects may also be produced by shear flows in a fluid. If $\mathbf{v}(\omega, \mathbf{x})$ is the flow field for an incompressible fluid, the most general linear ansatz is of the form

$$\epsilon_{ik} = \epsilon^{(0)} \delta_{ik} + \mu_1 \left(\frac{\partial v_i}{\partial x^k} + \frac{\partial v_k}{\partial x^i} \right) + \mu_2 \left(\frac{\partial v_i}{\partial x^k} - \frac{\partial v_k}{\partial x^i} \right). \quad (5.21)$$

The last term on the right-hand side (with coefficient μ_2) vanishes after a transformation to a suitable local reference system. Within the framework of classical mechanics, it can be shown that this implies $\mu_2 = 0$; however, quantum effects allow for $\mu_2 \neq 0$. In any case, we find, in general, that $\mu_2 \ll \mu_1$. The anisotropy effect related to $\mu_1 \neq 0$ is called *Maxwell's effect*.

5.2.3 Influence of electric fields

We start with a discussion of the influence of electric fields. A linear relation between $\epsilon_{ij}(\omega)$ and $\mathbf{E}$ is of the following form:

$$\epsilon_{ik}(\omega) = \epsilon_{ik}^{(0)}(\omega) + \alpha_{ikr} E_r. \quad (5.22)$$

The effect related to this type of dependence is called the *Pockels effect*. The tensor α_{ikr} is constrained by the same symmetry conditions as the piezoelectric tensor.

Effects *quadratic* in $\mathbf{E}$ are not subject to any symmetry constraints of the medium. They are most easily observed in isotropic and mirror-symmetric media, which for symmetry reasons do not exhibit a Pockels effect. In this case, a quadratic dependence is related to a qualitative change of the behavior of the medium. A second-order dependence on $\mathbf{E}$ is of the following type:

$$\epsilon_{ik} = (\epsilon^{(0)} + \beta \mathbf{E}^2) \delta_{ik} + \alpha E_i E_k. \quad (5.23)$$

The second term is crucial, because it leads to optical anisotropy, while the first term, $\beta \mathbf{E}^2 \delta_{ik}$, only induces an unimportant change of the isotropic dielectric permeability.

Due to the additional term $\alpha E_i E_k$, the electric permeability ϵ_{ik} is no longer invariant under arbitrary rotations but only under rotations around an axis parallel to the field vector $\mathbf{E}$. In the presence of a homogeneous field, a formerly isotropic medium shows the optical properties of a uniaxial crystal. This leads to the so-called *Kerr effect*: double refraction under the influence of an electric field. The optic axis is parallel to the direction of $\mathbf{E}$ and the principal values of the refractive index are given by $\epsilon_{1,2} = \epsilon^{(0)}$ and $\epsilon_3 = \epsilon^{(0)} + \alpha E^2$.

Both the Pockels effect and Kerr effect are utilized for rapid switching or modulation of laser beams, for instance. The principle of a Pockels cell is simple (see fig. 5.1). A medium is placed between two orthogonal polarization filters. The transparency of the system for a light ray can be influenced by an external field applied to the medium.

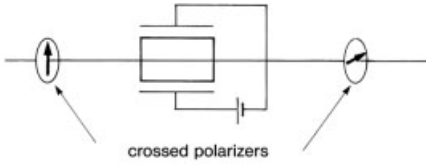


Figure 5.1: Principle of a Pockels cell.

5.2.4 Dependence on magnetic fields

Very interesting phenomena occur if the dielectric permeability ϵ_{ik} depends on a homogeneous magnetic field. General symmetry requirements for the kinetic coefficients imply that

$$\epsilon_{ik}(\omega, \mathbf{B}) = \epsilon_{ki}(\omega, -\mathbf{B}), \quad (5.24)$$

and therefore a term linear in $\mathbf{B}$ has to be antisymmetric in i and k . The passivity condition (2.94) now reads

$$\frac{\omega}{2i} [\epsilon_{ik}(\omega, \mathbf{B}) - \epsilon_{ki}^*(\omega, \mathbf{B})] \text{ is positive semidefinite.}$$

Hence, for a transparent medium, $\epsilon_{ik}(\omega, \mathbf{B})$ must be Hermitian but not necessarily real. Splitting into real and imaginary parts,

$$\epsilon_{ik}(\omega, \mathbf{B}) = \epsilon'_{ik}(\omega, \mathbf{B}) + i\epsilon''_{ik}(\omega, \mathbf{B}), \quad (5.25)$$

we see that the real part is symmetric in i and k , and even with respect to $\mathbf{B}$, while the imaginary part is antisymmetric and odd:

$$\begin{aligned} \epsilon'_{ik}(\omega, \mathbf{B}) &= \epsilon'_{ki}(\omega, \mathbf{B}) = \epsilon'_{ik}(\omega, -\mathbf{B}), \\ \epsilon''_{ik}(\omega, \mathbf{B}) &= -\epsilon''_{ki}(\omega, \mathbf{B}) = -\epsilon''_{ik}(\omega, -\mathbf{B}). \end{aligned} \quad (5.26)$$

Keeping only the linear approximation in $\mathbf{B}$ changes only the imaginary part:

$$\epsilon_{ik} = \epsilon_{ik}^{(0)} + i\epsilon_{ikr}g_r, \quad (5.27)$$

where

$$g_r = g_{rs}B_s. \quad (5.28)$$

For the isotropic case we have

$$\epsilon_{ik} = \epsilon^{(0)}\delta_{ik} + \alpha\epsilon_{ikr}B_r. \quad (5.29)$$

The discussion of optical properties follows closely that on natural optical activity. If we apply a magnetic field $\mathbf{B}$ to a medium, which in the absence of an external magnetic field is isotropic, then the plane of polarization of a plane-polarized light beam with a wave vector $\mathbf{k} = (\omega/c)\mathbf{n}$ that is not orthogonal to $\mathbf{B}$ will be rotated by an angle

$$\varphi = \frac{\omega}{2cn_0}aB_nz = RB_nz \quad \text{with} \quad B_n = \frac{\mathbf{n} \cdot \mathbf{B}}{n_0}. \quad (5.30)$$

This rotation of the plane of polarization, discovered by Michael Faraday in 1845, is called the *Faraday effect*. The Faraday effect constituted the first hint pointing to a relation between optics and electromagnetism. The constant R is called *Verdet's constant*. Despite the mathematical analogy between this effect and the treatment of natural optical activity, we should emphasize the following differences:

1. There are no symmetry constraints for the Faraday effect. In contrast to $\mathbf{n}$, the vector $\mathbf{B}$ does not change sign under point reflections. Therefore, the term $\alpha\epsilon_{ikr}B_r$ is compatible with reflections.
2. Changing the direction $\mathbf{n}$ of the beam also changes the sign of B_n and thereby also the orientation of the plane of polarization. If a beam passes through the medium in both directions (being reflected in between), the magnitude of the effect doubles.

On a microscopic level, the Faraday effect can be explained as follows. In the presence of a magnetic field, the charges in a medium not only oscillate but also rotate with frequency $\omega_L = eB/m$ (the Larmor frequency). Therefore, depending on the orientation of the circularly polarized light, the medium reacts differently.

For $\mathbf{n} \cdot \mathbf{B} = 0$, the Faraday effect is absent. In analogy to the Kerr effect, one can now observe an effect quadratic in $\mathbf{B}$, which is called the *Cotton-Mouton effect*.

6 Foundations of nonlinear optics

6.1 Nonlinear polarization – combination frequencies

In the previous chapter we mentioned certain effects that occur in homogeneous media, which we could attribute to the field dependence of the dielectric constant $\epsilon_{ij}(\mathbf{E}, \mathbf{B})$. However, up to now we have confined ourselves to the influence of homogeneous, time-independent (or slowly varying) fields on ϵ_{ij} . In this case the equations for the wave fields remained linear and could be solved easily.

Yet, if the field strength of the wave field itself reaches a scale where the dependence of the dielectric tensor ϵ_{ij} on these fields becomes noticeable, we enter the realm of *nonlinear optics*: the wave equations become nonlinear, thus allowing for a vast number of new phenomena, which we will discuss in this chapter.

The electric field strengths of light waves coming from ordinary sources reach values of about 100 V/m, and they induce charge displacements inside materials that are of the order of 10^{-14} cm and thus small compared to typical atomic distances of about 10^{-8} cm. For these values, the polarization depends linearly on the electric field strength.

However, for field strengths of about 10^6 V/m and charge displacements of the order of 10^{-10} cm, nonlinear effects may become important. Such values of the field strengths are common for laser beams and may also be found in glass-fiber cables where large radiation powers pass through very small cross-sections.

Nonlinear effects in the magnetic permeability μ_{ij} as well as nonlinear dependences on the magnetic field strength $\mathbf{B}$ turn out to be less important and will not be considered here.

We will restrict our discussion to the dependences of the dielectric tensor on the electric field strength $\mathbf{E}$ and on the displacement tensor u_{ij} . For nonlinear contributions that are not too large, it is helpful to make a Taylor expansion in $\mathbf{E}$ and u_{ij} . We write the electric polarization as the sum of a linear and a nonlinear term,

$$P_i = \chi_{ij}\epsilon_0 E_j + \lambda_{ijk} u_{jk} + P'_i, \quad (6.1)$$

and take into account the following contributions for the nonlinear quantity $P'(\mathbf{E}, u)$:

$$\begin{aligned} P'_i = & \alpha_{ijk} E_j E_k + \alpha_{ijkl} E_j E_k E_l + \dots \\ & + \xi_{ijrs} E_j u_{rs} + \xi_{ijkrs} E_j E_k u_{rs} + \xi_{ijklrs} E_j E_k E_l u_{rs} + \dots \end{aligned} \quad (6.2)$$

The case $\alpha_{ijk} \neq 0$ is not possible for isotropic media. Invariance under point reflections requires all material tensors of odd order to vanish.

The presence of nonlinear terms has far-reaching consequences, the most important one being the loss of the superposition principle for solutions of the wave equation. A further

unpleasant side-effect for practical calculations is that for a harmonic time dependence we have to work with the real fields $E = E_0 \cos(\omega t + \varphi)$ instead of the complex exponential function.

If the incoming field E_i, u_{ij} exhibits a harmonic time dependence of frequency ω , the nonlinear polarization $\mathbf{P}'$ also contains higher harmonics $2\omega, 3\omega$, due to identities like

$$\begin{aligned}\cos^2(\omega t) &= \frac{1}{2}[1 + \cos(2\omega t)], \\ \cos^3(\omega t) &= \frac{1}{4}[3 \cos(\omega t) + \cos(3\omega t)],\end{aligned}\tag{6.3}$$

and so on, as well as time-independent contributions. This leads to phenomena like *frequency multiplication* and *optical rectification*: a signal of frequency ω generates signals with frequencies that are multiples of ω as well as a time-independent signal.

More generally, let us consider the case in which the incoming fields consist of several frequency components $\omega_1, \omega_2, \omega_3, \dots$. Due to identities like

$$\cos(\omega_1 t) \cos(\omega_2 t) = \frac{1}{2} \{ \cos[(\omega_1 + \omega_2)t] + \cos[(\omega_1 - \omega_2)t] \},\tag{6.4}$$

$$\begin{aligned}\cos(\omega_1 t) \cos(\omega_2 t) \cos(\omega_3 t) &= \frac{1}{4} \{ \cos[(\omega_1 + \omega_2 + \omega_3)t] + \cos[(\omega_1 + \omega_2 - \omega_3)t] \\ &\quad + \cos[(\omega_1 - \omega_2 + \omega_3)t] + \cos[(\omega_1 - \omega_2 - \omega_3)t] \},\end{aligned}\tag{6.5}$$

and so on, the polarization $\mathbf{P}'$ also contains the *combination frequencies* of $\{\omega_i\}$:

$$\Omega_{\{\epsilon\}} = \sum_i \epsilon_i \omega_i,\tag{6.6}$$

where $\epsilon_i = \pm 1$.

Because of the nonlinear polarization $\mathbf{P}'$, an incoming wave with frequency ω generates higher harmonics with frequencies $2\omega, 3\omega$, as well as a rectified signal. Due to the frequency dependence of the refractive index, these signals will propagate with different velocities. Therefore, a spatially concentrated signal tends to “diverge” inside a nonlinear medium.

In eq. (6.2) a term of n th order in E_i and of m th order in u_{ij} corresponds to a nonlinear coupling of $n + 1$ electromagnetic waves and m sound waves.

After quantizing the wave motions, a wave with frequency ω and wave vector $\mathbf{k}$ will be associated to quanta with energy $E = \hbar\omega$ and momentum $\mathbf{p} = \hbar\mathbf{k}$. The quanta of the electromagnetic wave field are called *photons*, and those of the sound wave field are *phonons*. Hence, the nonlinearities describe interactions of $n + 1$ photons with m phonons.

If we associate positive frequencies with outgoing quanta and negative frequencies with incoming quanta, the combination law (6.6) translates into the conservation of energy for all quanta taking part in the interaction.

The typical frequencies of light and sound waves are in the range of 10^{15} Hz and 10^3 – 10^5 Hz, respectively, and thus differ by many orders of magnitude. The difference in wavelengths of 10^{-5} cm and 10–0.1 cm, respectively, is also large, but, due to the different propagation velocities, not quite as impressive.

Because of these differences in scale, the effects related to the term $\xi_{ijrs} E_j u_{rs}$ in eq. (6.2), which is linear in E_j and u_{rs} , are in general not treated as part of nonlinear optics, since u_{rs}

may be considered to be quasi-constant compared to E_i . Nevertheless, we will discuss these effects in this chapter.

We should also mention that the nonlinear polarization $\mathbf{P}'$ and the electromagnetic waves induced by the nonlinearities give rise to further combination waves, and we may expect all higher combination frequencies

$$\Omega_{\{n\}} = \sum n_i \omega_i \quad (n_i \text{ integer}) \quad (6.7)$$

of the original frequency ω to contribute. Only for the case of small nonlinearities may one hope that the higher combination frequencies are suppressed compared to the lower ones. These relations will be studied in the next section.

6.2 Nonlinear waves in a medium

In the absence of external charges and currents, Maxwell's equations inside a medium assume the form

$$\begin{aligned} \nabla \cdot \mathbf{D} &= 0, & \nabla \cdot \mathbf{B} &= 0, \\ \nabla \times \mathbf{H} &= \kappa \frac{\partial \mathbf{D}}{\partial t}, & \nabla \times \mathbf{E} &= -\kappa \frac{\partial \mathbf{B}}{\partial t}. \end{aligned} \quad (6.8)$$

We neglect magnetic effects and set $\mathbf{B} = \mu_0 \mathbf{H}$. Applying $\nabla \times$ to the fourth equation and using the third one yields

$$\nabla(\nabla \cdot \mathbf{E}) - \Delta \mathbf{E} = -\kappa^2 \mu_0 \frac{\partial^2 \mathbf{D}}{\partial t^2}. \quad (6.9)$$

Like in eq. (6.1), we separate linear and nonlinear contributions in $\mathbf{D}(\mathbf{E})$ but neglect distortion effects and write

$$D_i = \epsilon_{ij} \epsilon_0 E_j + P'_i(\mathbf{E}). \quad (6.10)$$

With $\kappa^2 \mu_0 \epsilon_0 = 1/c^2$ we obtain

$$\frac{1}{c^2} \epsilon_{ij} \frac{\partial^2 E_j}{\partial t^2} - \nabla_i(\nabla \cdot \mathbf{E}) - \Delta E_i = -\kappa^2 \mu_0 \frac{\partial^2 P'_i(\mathbf{E})}{\partial t^2}. \quad (6.11)$$

On the left-hand side of eq. (6.11) we find a wave operator for a linear medium; the nonlinearity is represented by the term on the right-hand side, which deprives the wave equation of its linearity. The nonlinear wave equation (6.11) is much too complicated to be solvable in general.

Next we will assume a series of specializations and approximations by which we hope to arrive at a closed solvable system of equations, which still describes the important aspects of the general problem. The first assumption is as follows:

- The wave propagates only along the 3-axis; there is rotational invariance around the 3-axis, and the fields $\mathbf{E}$ and $\mathbf{P}'$ are transverse to the 3-axis.

This assumption will be satisfied if, in a linear approximation, the medium is optically uniaxial and the optic axis is the 3-axis (compare section 4.3). With this assumption, eq. (6.11) simplifies to

$$\frac{\epsilon}{c^2} \frac{\partial^2 E_i}{\partial t^2} - \frac{\partial^2 E_i}{\partial z^2} = -\kappa^2 \mu_0 \frac{\partial^2 P'_i}{\partial t^2} \quad (i = 1, 2). \quad (6.12)$$

Furthermore, we make the following assumption:

- The nonlinearity is only quadratic, i.e.

$$P'_i = \alpha_{ijk} E_j E_k \quad (i, j, k = 1, 2). \quad (6.13)$$

The following assumption is more drastic:

- The nonlinearity is small enough that, to a good approximation, the fields E_i and P'_i contain only three frequencies: the fundamental frequencies ω_1 and ω_2 as well as their sum $\omega_3 = \omega_1 + \omega_2$.

This corresponds to the ansatz:

$$E_i = \sum_{\nu=1}^3 E_i^{\omega_\nu}, \quad P'_i = \sum_{\nu=1}^3 P'_i{}^{\omega_\nu}, \quad (6.14)$$

with

$$\begin{aligned} E_i^{\omega_\nu}(t, z) &= \text{Re}\{E_i^\nu(z) e^{i(k_\nu z - \omega_\nu t)}\}, \\ P'_i{}^{\omega_\nu}(t, z) &= \text{Re}\{P'_i{}^\nu(z) e^{i(k_\nu z - \omega_\nu t)}\}, \end{aligned} \quad (6.15)$$

for $\nu = 1, 2, 3$, where $k_\nu = [\sqrt{\epsilon(\omega_\nu)}/c]\omega_\nu = (\sqrt{\epsilon_\nu}/c)\omega_\nu$. Because of eq. (6.13), the frequency parts P'^{ω_ν} are given by

$$\begin{aligned} P'_i{}^{\omega_3} &= \text{Re}\{\alpha_{ijk} E_j^1 E_k^2 e^{i(k_1 z + k_2 z - \omega_3 t)}\}, \\ P'_i{}^{\omega_1} &= \text{Re}\{\alpha_{ijk} E_j^3 E_k^{2*} e^{i(k_3 z - k_2 z - \omega_1 t)}\}, \\ P'_i{}^{\omega_2} &= \text{Re}\{\alpha_{ijk} E_j^3 E_k^{1*} e^{i(k_3 z - k_1 z - \omega_2 t)}\}. \end{aligned} \quad (6.16)$$

For $\nu = 3$ the equations following from these assumptions for the frequency components in eq. (6.12) read

$$\begin{aligned} &\left(-k_3^2 - \frac{\partial^2}{\partial z^2}\right) \text{Re}\{E_i^3(z) e^{i(k_3 z - \omega_3 t)}\} \\ &= \kappa^2 \mu_0 \omega_3^2 \text{Re}\{\alpha_{ijk} E_j^1(z) E_k^2(z) e^{i(k_1 z + k_2 z - \omega_3 t)}\}. \end{aligned} \quad (6.17)$$

Similarly for $\nu = 1, 2$. This is a system of ordinary differential equations for $E_i^\nu(z)$. In addition, we assume that the following holds:

- The variations of the amplitudes $E_i^\nu(z)$ are small for distances of about a wavelength such that

$$\left| \frac{\partial^2}{\partial z^2} E_i^\nu(z) \right| \ll \left| k_\nu \frac{\partial}{\partial z} E_i^\nu(z) \right|,$$

and the following approximation is valid:

$$\frac{\partial^2}{\partial z^2} E_i^\nu(z) e^{i(k_\nu z - \omega_\nu t)} = \left(-k_\nu^2 + 2ik_\nu \frac{\partial}{\partial z} E_i^\nu \right) e^{i(k_\nu z - \omega_\nu t)}. \quad (6.18)$$

This leads to

$$\text{Re} \left\{ ik_3 \frac{\partial E_i^3}{\partial z} e^{i(k_3 z - \omega_3 t)} \right\} = -\frac{1}{2} \kappa^2 \mu_0 \omega_3^2 \text{Re} \{ \alpha_{ijk} E_j^1 E_k^2 e^{i(k_1 z + k_2 z - \omega_3 t)} \}, \quad (6.19)$$

and similarly for $\nu = 1, 2$.

Our final assumption is as follows:

- The phases of the quantities in brackets on both sides of eq. (6.19) differ approximately by $\delta\varphi = \pi$. This is known as the *rotating phase approximation* and corresponds to the assumption of a fixed phase relation between $E_i^{\omega_\nu}$ and the nonlinear polarization $P_i^{\omega_\nu}$.

This last condition permits us to change to complex quantities, since eq. (6.20) also holds for the imaginary part:

$$\begin{aligned} \frac{dE_i^3}{dz} &= \frac{i}{2} \frac{\kappa^2 \mu_0 \omega_3^2}{k_3} \alpha_{ijk} E_j^1 E_k^2 e^{iz\Delta k}, \\ \frac{dE_i^1}{dz} &= \frac{i}{2} \frac{\kappa^2 \mu_0 \omega_1^2}{k_1} \alpha_{ijk} E_j^3 E_k^{2*} e^{-iz\Delta k}, \\ \frac{dE_i^2}{dz} &= \frac{i}{2} \frac{\kappa^2 \mu_0 \omega_2^2}{k_2} \alpha_{ijk} E_j^3 E_k^{1*} e^{-iz\Delta k}. \end{aligned} \quad (6.20)$$

The quantity $\Delta k = k_1 + k_2 - k_3$ is called the *phase mismatch*.

The system of eqs. (6.20), describing the so-called *three-wave interaction*, is of fundamental importance for a model-based discussion of the influence of nonlinearities that are not too dominant. A phase mismatch $\Delta k \neq 0$ tends to “smear out” the oscillations on the right-hand side of eqs. (6.20) and to suppress larger values of the amplitudes E_i^ν . We will come back to this point in the next section. Complete phase matching, $\Delta k = k_1 + k_2 - k_3 = 0$, can be interpreted in a particle picture as momentum conservation of the photons participating in the interaction.

If the above-mentioned simplifications are not justified, e.g. if the nonlinearities are too dominant, one has to go back to the full system of nonlinear oscillation equations. Under certain circumstances all complexities of nonlinear dynamics may show up: extremely sensible dependence on the initial conditions, chaotic behavior, strange attractors, sudden qualitative changes of the behavior as a result of minor changes in external parameters, and much more. However, the subject of nonlinear dynamics is not the topic of this book, and for details the reader is referred to the literature given at the end.

Surprisingly, there is a class of nonlinear partial differential equations whose solutions exhibit no chaotic behavior and can be constructed explicitly. In most cases these are special differential equations in one space and one time coordinate. The simplest example is the *nonlinear Schrödinger equation* for a complex function $\psi(\tau, \zeta)$:

$$i \frac{\partial \psi}{\partial \tau} + \frac{\partial^2 \psi}{\partial \zeta^2} + 2|\psi|^2 \psi = 0. \quad (6.21)$$

We will study this equation in more detail in section 6.8. The nonlinear Schrödinger equation can be used to explain such phenomena as self-focussing and automatic phase matching (see sections 6.3 and 6.6). One can easily verify that eq. (6.21) has a certain class of special solutions:

$$\psi_0(\tau, \zeta) = \frac{2b \exp\{-2i[2(a^2 - b^2)\tau - a\zeta]\}}{\cosh[2b(\zeta - 4a\tau)]}. \quad (6.22)$$

These solutions correspond to so-called *solitary waves*: $|\psi_0(\tau, \zeta)|$ describes a spatially concentrated signal of width $1/2b$ propagating with a velocity $v = 4a$ without changing its shape. The width of the signal decreases with its height.

The possibility of non-decaying, spatially constant excitations is remarkable in itself. The effects of dispersion and nonlinearity, each supporting the tendency of a signal to diverge, just compensate in this case.

Even more remarkable is the fact that eq. (6.22) admits N -soliton solutions for arbitrary values of N . In the limit $t \rightarrow -\infty$, these solutions describe N spatially separated signals of the form (6.22) with, in general, different widths and velocities. Letting time evolve, the faster signals will overtake the slower ones. When two signals approach each other, they become distorted for a moment but regain their original shape after separation. In the limit $t \rightarrow \infty$, the solution corresponds to separated signals again.

This property of shape conservation already suggests that an infinite number of conservation laws underlie eq. (6.21). A detailed analysis confirms this suggestion. The nonlinear Schrödinger equation even has the property of *complete integrability*: like in the case of integrable systems with finitely many degrees of freedom, the conservation laws can be used to determine all solutions of the equation.

Another example of a completely integrable nonlinear wave equation is the *sine-Gordon equation*:

$$\left(\frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} \right) \varphi + \sin \varphi = 0. \quad (6.23)$$

The fundamental non-trivial solutions of this equation are the so-called *kink and anti-kink solutions*:

$$\varphi_{\pm}(t, x) = -4 \arctan \exp \left(\pm \frac{x - x_0 - vt}{\sqrt{1 - v^2}} \right). \quad (6.24)$$

Their asymptotic behavior is given by

$$\begin{aligned} \lim_{x \rightarrow \pm\infty} \varphi_{\pm}(t, x) &= -2\pi, \\ \lim_{x \rightarrow \mp\infty} \varphi_{\pm}(t, x) &= 0. \end{aligned} \quad (6.25)$$

These solutions resemble upward or downward steps, smeared out with a width $\sqrt{1-v^2}$. The transition from one asymptotic value to the other happens mainly in a region around $x = x_0 + vt$. There are solutions with N kinks and $\bar{N}$ anti-kinks; furthermore, there exist *breather solutions*: bound states of kinks and anti-kinks oscillating around their common center. The simplest breather solution is given by

$$\varphi_B = 4 \arctan \left(\tan \alpha \frac{\sin(t \sin \alpha)}{\cosh(x \sin \alpha)} \right). \quad (6.26)$$

An application of the sine–Gordon equation is the theoretical treatment of self-induced transparency (see sections 6.3 and 6.6).

As a final example for a completely integrable equation, we mention the general *three-wave interaction*:

$$\begin{aligned} \frac{\partial A_3}{\partial t} + v_3 \frac{\partial A_3}{\partial x} &= i A_1 A_2, \\ \frac{\partial A_1}{\partial t} + v_1 \frac{\partial A_1}{\partial x} &= i A_3 A_2^*, \\ \frac{\partial A_2}{\partial t} + v_2 \frac{\partial A_2}{\partial x} &= i A_3 A_1^*. \end{aligned} \quad (6.27)$$

Equation (6.20), describing a general three-wave interaction, can be brought into the above form if each of the waves $\mathbf{E}^i$ has fixed polarizations.

The interested reader will find details about the huge field of completely integrable systems in the references at the end of the book.

6.3 Survey of phenomena in nonlinear optics

We begin our discussion of special nonlinear optical effects with a survey of those phenomena which are to be expected if the nonlinearities are not too strong.

In table 6.1, the effects are ordered according to the powers of E_i and u_{ij} as they occur in the nonlinear polarization P'_i . The second column summarizes the “static” effects already discussed in section 5.2. In this case only the (quasi-)static external fields $E_i^{(0)}$ and $u_{ij}^{(0)}$ give rise to nonlinear contributions, while the wave field E_i still enters in a linear way. The third column contains the most important “dynamic” effects, which are nonlinear in the wave fields.

The phenomena related to the terms quadratic in E are also called *three-photon phenomena*, because this case describes the coupling of three electromagnetic fields; the phenomena related to terms cubic in E are so-called *four-photon phenomena*. The effects linear in E in the third row are usually not considered as part of nonlinear optics.

The phenomenon of *frequency doubling* occurs when light with some frequency ω enters a nonlinear medium. The nonlinearity of the medium causes the emergence of higher harmonics with frequency 2ω . This phenomenon is used to generate coherent light in frequency ranges that are not directly accessible for lasers, because of the lack of suitable materials.

Parametric amplification can occur when a *pumping wave* of frequency ω_3 and a *signal wave* of frequency ω_1 pass through the medium simultaneously. Under suitable circumstances the signal wave gets amplified within the nonlinear medium. Furthermore, a so-called *idler*

Table 6.1: Effects of nonlinear polarization.

	Static	Dynamic
$P' \sim E^2$	Pockels effect	Frequency doubling, optical rectification; two-photon absorption; parametric amplification
$P' \sim E^3$	Kerr effect	Frequency tripling, three-photon absorption; intensity dependence of the refractive index, self-focussing; phase matching; self-induced transparency, optical bistability
$P' \sim Eu$	Elastooptical effects, electrostriction	Brillouin scattering
$P' \sim E^3u$		Stimulated Brillouin scattering

wave of frequency $\omega_2 = \omega_3 - \omega_1$ is generated. What distinguishes parametric amplification compared to other amplification mechanisms is its extremely low noise. One can also imagine optical control elements, which, in analogy to transistors, are based on light instead of electric currents. At present a completely new area of applications seems to be emerging and *photonics* is about to become the optical analogue of electronics. In the following section we will discuss the phenomena of frequency doubling and parametric amplification in more detail.

Two-photon absorption is of special importance in spectroscopy. Two beams of light with frequency ω enter the medium from opposite directions and induce electromagnetic transitions with excitation energy $\Delta E = 2\hbar\omega$. The advantage of two-photon spectroscopy is that the otherwise annoying Doppler broadening related to inner movements inside the absorber are suppressed. A further possibility is multiple absorption.

Effects trilinear in E show up in an intensity dependence of the real and imaginary parts of the refractive index. We shall describe the corresponding phenomena later. The imaginary part of the refractive index, i.e. the absorption coefficient, can depend on the intensity due to a saturation effect: the levels that get excited by absorption processes are already occupied. In this case one observes a decrease of the absorption coefficient.

Brillouin scattering is an inelastic scattering effect where electromagnetic radiation and sonar vibrations are coupled. Classically, this corresponds to a distortion dependence of the refractive index. In a particle picture, a photon of frequency ω and a phonon of frequency ω_S merge to a photon of frequency $\omega + \omega_S$:

$$\hbar\omega + \hbar\omega_S \longrightarrow \hbar(\omega + \omega_S).$$

Spectroscopically, one observes an *anti-Stokes line* with increased frequency. On the other hand, the creation of a phonon may lead to a *Stokes line* with lower frequency:

$$\hbar\omega \longrightarrow \hbar\omega_S + \hbar(\omega - \omega_S).$$

If the light couples to molecular vibrations instead of lattice vibrations, one speaks of *Raman scattering*. *Stimulated Brillouin scattering* (and *stimulated Raman scattering*) consist of an interaction of several photons with oscillation quanta.

The reader may wonder why we have omitted, in table 6.1, a row with $P' \sim E^2 u$. This would correspond to processes of the following type:

$$\begin{aligned}\hbar\omega_1 + \hbar\omega_2 &\longrightarrow \hbar(\omega_1 + \omega_2 - \omega_S) + \hbar\omega_S, \\ 2\hbar\omega &\longrightarrow \hbar(2\omega - \omega_S) + \hbar\omega_S.\end{aligned}$$

Because of the huge scale differences between ω_S on the one side and ω_1 and ω_2 on the other side, these processes correspond to a parametric amplification with an additional coupling to lattice vibrations, or to an almost frequency doubling with the participation of an oscillation quantum. Such processes do not have any obvious applications.

6.4 Parametric amplification and frequency doubling

A quantitative discussion of the process of parametric amplification starts from the fundamental equations (6.20) for a three-wave interaction. No relevant information is lost if we assume a fixed polarization direction for each of the three waves.

A simple rescaling of the remaining components of $\mathbf{E}^1$, $\mathbf{E}^2$, and $\mathbf{E}^3$ leads to the equations:

$$\begin{aligned}\frac{dA_3}{dz} &= iA_1A_2 e^{iz\Delta k}, \\ \frac{dA_1}{dz} &= iA_2^*A_3 e^{-iz\Delta k}, \\ \frac{dA_2}{dz} &= iA_1^*A_3 e^{-iz\Delta k}.\end{aligned}\tag{6.28}$$

To relate these equations with the situation of parametric amplification, we identify A_3 with the intensive pump wave, A_1 with the signal wave to be amplified, and A_2 with the additional idler wave (see section 6.3).

We may assume $|A_3| \gg |A_1|, |A_2|$. In this case $\partial A_3 / \partial z$ is small of second order, i.e. A_3 is almost constant. Without loss of generality, we further assume A_3 to be real. For the resulting simplified system of equations,

$$\begin{aligned}\frac{dA_1}{dz} &= iA_3A_2^* e^{-iz\Delta k}, \\ \frac{dA_2^*}{dz} &= -iA_3A_1 e^{iz\Delta k},\end{aligned}\tag{6.29}$$

we make the substitutions

$$u = A_1 e^{iz\Delta k/2} \quad \text{and} \quad v = A_2^* e^{-iz\Delta k/2},\tag{6.30}$$

and obtain a system with constant coefficients that is easily solved:

$$\begin{aligned}\frac{du}{dz} &= \frac{1}{2}i\Delta k u + iA_3v, \\ \frac{dv}{dz} &= -iA_3u - \frac{1}{2}i\Delta k v.\end{aligned}\tag{6.31}$$

For the solution with initial conditions $A_2(0) = 0$ and $A_1(0) \neq 0$ we find:

$$\begin{aligned} A_1(z) &= A_1(0) \left(\cosh(\lambda z) - \frac{i\Delta k}{2\lambda} \sinh(\lambda z) \right) e^{iz\Delta k/2}, \\ A_2(z) &= iA_1(0) \frac{A_3}{\lambda} \sinh(\lambda z) e^{iz\Delta k/2}, \end{aligned} \quad (6.32)$$

where

$$\lambda = \sqrt{A_3^2 - \frac{1}{4}(\Delta k)^2}.$$

The exponential increase of $A_1(z)$ expresses the amplification. For non-vanishing phase mismatch $\Delta k = k_1 + k_2 - k_3$, we find additional oscillations.

We now come to the phenomenon of frequency doubling. In this case we have $\omega_1 = \omega_2 = \omega$ and $\omega_3 = 2\omega$, and in addition $A_1 = A_2$, such that eq. (6.28) simplifies to

$$\begin{aligned} \frac{dA_3}{dz} &= iA_1^2 e^{iz\Delta k}, \\ \frac{dA_1}{dz} &= iA_1^* A_3 e^{-iz\Delta k}. \end{aligned} \quad (6.33)$$

We are looking for a solution with initial condition $A_3(0) = 0$, because the higher mode wave is supposed to be produced inside the medium. From eq. (6.33) we immediately see that, initially, A_1 changes only very slowly while the amplitude A_3 gradually increases.

Let us first limit our discussion to this initial region where we can assume A_1 to be constant. There we find

$$\frac{dA_3}{dz} = iA_1^2(0) e^{iz\Delta k}, \quad (6.34)$$

i.e.

$$A_3(z) = A_1^2(0) \left(\frac{e^{iz\Delta k} - 1}{\Delta k} \right). \quad (6.35)$$

Hence, we obtain for the intensity $I_3 \sim |A_3(z)|^2$ of the higher harmonics:

$$I_3(z) \sim |A_1(0)|^4 \frac{\sin^2(z\Delta k/2)}{(\Delta k)^2}. \quad (6.36)$$

For $\Delta k \neq 0$ we find an oscillating dependence on z , while the height of the maxima increases as $(\Delta k)^{-2}$.

The oscillations caused by the *phase mismatch* $\Delta k \neq 0$ suppress a strong increase of the amplitude A_3 , so that the initial region just considered usually extends over the entire nonlinear medium. The oscillations of $I_3(z)$ can be observed directly, when we direct a high-intensity light beam of frequency ω onto a plate of variable thickness z and measure the signal leaving with frequency 2ω .

In the following we will consider the case $\Delta k = 0$. So, we solve

$$\frac{dA_3}{dz} = iA_1^2, \quad \frac{dA_1}{dz} = iA_1^* A_3, \quad (6.37)$$

with the initial condition $A_3(0) = 0$. We may assume $A_1(0)$ to be real, because if $A_1(z)$ and $A_2(z)$ are solutions of (6.37), so are $e^{i\alpha} A_1(z)$ and $e^{2i\alpha} A_3(z)$ for arbitrary values of α . Now $dA_3(0)/dz$ is purely imaginary. We set $B_1(z) = A_1(z)$ and $B_3(z) = -iA_3(z)$, i.e.

$$\frac{dB_3}{dz} = B_1^2, \quad \frac{dB_1}{dz} = -B_1^* B_3. \quad (6.38)$$

Solutions with initial conditions $B_3(0) = 0$ and $B_1(0) > 0$ are real for all values of z , because they are also the unique solutions of the real system $dB_3/dz = B_1^2$, $dB_1/dz = B_1 B_3$. Equation (6.38) can be solved by using the conservation law following from eq. (6.38):

$$\frac{d}{dz}(B_1 B_1^* + B_3 B_3^*) = 0. \quad (6.39)$$

Obviously, the conserved quantity is proportional to the total intensity. From

$$B_1^2(z) + B_3^2(z) = I = B_1^2(0),$$

we obtain

$$\frac{dB_3}{dz} = I - B_3^2, \quad (6.40)$$

and because $B_3(0) = 0$ we find that

$$B_3(z) = B_1(0) \tanh[z B_1(0)], \quad (6.41)$$

$$B_1(z) = \sqrt{B_1^2(0) - B_3^2(z)} = B_1(0) \frac{1}{\cosh[z B_1(0)]}.$$

We observe that $B_3(z)$ increases monotonically and, in the limit $z \rightarrow \infty$, approaches $B_1(0)$, while $B_1(z)$ decreases monotonically and approaches zero.

For a sufficiently thick medium, the total intensity is transformed into the second harmonic with double frequency. This transformation proceeds faster if the intensity $I = B_1^2(0)$ is larger.

6.5 Phase matching

After we have convinced ourselves of the importance of phase matching $\Delta k = 0$ in nonlinear optics in order to get good results for a combination signal, we now want to describe how to realize phase matching. We will discuss this problem in the concrete context of frequency doubling. For this example of a three-wave interaction we have $\omega_1 = \omega_2 = \omega$, $\mathbf{k}_1 = \mathbf{k}_2 = \mathbf{k}$, and the matching conditions for the frequencies and wave vectors read

$$\omega_3 = 2\omega \quad \text{and} \quad \mathbf{k}_3 = 2\mathbf{k}.$$

Hence, the wave vectors are parallel. As we did in chapter 3, we introduce the vectors for the refractive index, i.e. $\mathbf{n}_1 = \mathbf{n}(\omega_1) = \mathbf{k}_1 c / \omega_1$ and $\mathbf{n}_3 = \mathbf{n}(\omega_3) = \mathbf{k}_3 c / \omega_3$, and obtain the phase matching condition:

$$\mathbf{n}_1 = \mathbf{n}_3 \quad \text{or} \quad \mathbf{n}(\omega) = \mathbf{n}(2\omega). \quad (6.42)$$

The dispersion of the refractive index implies in general that $|\mathbf{n}(\omega)| \neq |\mathbf{n}(2\omega)|$, and eq. (6.42) is not easily realized. The equality of the vectors for the refractive index means that the wave surfaces defined in chapter 3 for the frequencies ω and 2ω must intersect at the point $\mathbf{n}(\omega) = \mathbf{n}(2\omega)$. For optically uniaxial crystals, the wave surfaces consist of the ordinary contribution – a sphere – and the extraordinary contribution – an ellipsoid of revolution that just touches the sphere from inside or outside. Figure 6.1 shows that indeed intersection points of the ordinary and extraordinary contributions of the wave surfaces can occur for different frequencies.

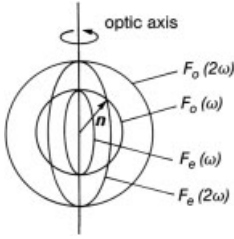


Figure 6.1: A cut through the wave surfaces for the frequencies ω and 2ω for an optically uniaxial crystal. Here $\mathbf{n}$ denotes the vector for the refractive index; and $F_o(\omega)$, $F_o(2\omega)$ are the ordinary and $F_e(\omega)$, $F_e(2\omega)$ the extraordinary shells of the wave surfaces for the frequencies ω and 2ω , respectively.

The ordinary surface for the frequency ω intersects the extraordinary surface for the frequency 2ω at point $\mathbf{n}$. Hence, for this choice of $\mathbf{n}$, we find phase matching. In general, however, the corresponding ray vectors $\mathbf{s}_o(\omega)$ and $\mathbf{s}_e(2\omega)$ do not coincide. As fig. 6.2 illustrates, an incoming regular beam with frequency ω has the effect that light with frequency 2ω will be emitted all over the plane spanned by $\mathbf{s}_o(\omega)$ and $\mathbf{s}_e(2\omega)$.

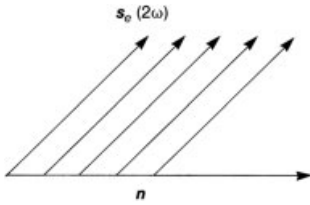


Figure 6.2: The ray vectors $\mathbf{n} = \mathbf{s}_o(\omega) \parallel \mathbf{n}$ and $\mathbf{s}_e(2\omega)$ for the ordinary and the extraordinary rays.

This flaw does not occur if the ordinary wave surface for the frequency ω and the extraordinary wave surface for the frequency 2ω meet by chance at just one point $\mathbf{n}$, because in this case $\mathbf{n}$, $\mathbf{s}_o(\omega)$, and $\mathbf{s}_e(2\omega)$ are parallel (see fig. 6.3). Fortunately, this degenerate case can be realized for some materials, e.g. in a KDP (potassium dihydrogen phosphate) crystal at a wavelength of $\lambda = 0.5145 \mu\text{m}$. Fine tuning makes use of the temperature dependence of the refractive index. According to section 4.3, the linear polarizations of the ordinary wave with frequency ω and the phase-matched extraordinary wave with frequency 2ω are orthogonal. (As linear polarized waves have vanishing angular momentum in the propagation direction, balance of angular momentum is also guaranteed.)

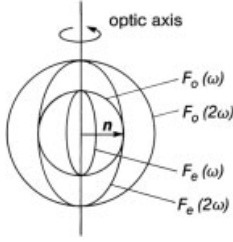


Figure 6.3: A cut through the wave surfaces with frequencies ω and 2ω for the degenerate case. The notations are as in fig. 6.1. The vectors $s_o(\omega)$ and $s_e(2\omega)$ are equal.

The phase matching condition $n(\omega) = n(2\omega)$ implies that the propagation velocities for the wave with frequency ω and the first harmonic with frequency 2ω must coincide. This requirement can be easily visualized. A wave with frequency ω generates permanently, by nonlinear polarization along its trajectory through the medium, a higher harmonic with frequency 2ω . If the fundamental wave and the higher harmonic have different propagation velocities, the higher harmonic generated at different points gets out of phase and they cancel each other due to interference.

6.6 Self-focussing, optical bistability, phase self-modulation

We now turn to a discussion of those phenomena which are trilinear in E_i ,

$$P'_i = \alpha_{ijkl} E_i E_k E_l, \quad (6.43)$$

i.e. the four-photon processes. In contrast to three-photon processes, the existence of such couplings does not depend on any symmetry properties of the medium. The relevant effects are already contained in the most symmetric case,

$$P'_i = \alpha E_i |\mathbf{E}|^2, \quad (6.44)$$

which, for an isotropic medium, implies an intensity dependence of the refractive index n of the form

$$n = \sqrt{\epsilon^{(0)} + \alpha |\mathbf{E}|^2} = \sqrt{\epsilon^{(0)}} + \alpha / [2\sqrt{\epsilon^{(0)}}] |\mathbf{E}|^2 + \dots,$$

i.e.

$$n = n_0 + n_1 I + \dots, \quad (6.45)$$

where n_1 can be real or complex; the real refractive index as well as the absorption coefficient can depend on the intensity.

Unlike in the previous sections, we will now restrict ourselves to a qualitative discussion of the phenomena and will not enter into detailed model studies (like, for example, the nonlinear Schrödinger equation or the sine-Gordon equation) but refer to the literature listed at the end of the book.

We first consider a real refractive index and assume $n_1 > 0$, i.e. the refractive index increases with increasing intensity. This case is more important and more frequent.

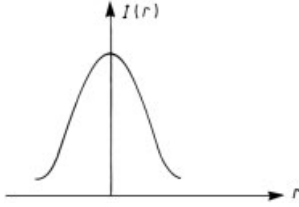


Figure 6.4: Intensity profile of a light ray perpendicular to the direction of propagation.

We now suppose that an intense bundle of parallel rays is directed onto such a medium. The intensity will be perpendicular to the direction of propagation and will decrease from inside to outside, as indicated in fig. 6.4.

Since the refractive index decreases with distance from the center of the beam and since, according to Snell's law of refraction, the direction of the beam is refracted into the direction ∇n , i.e. toward an increasing refractive index, the incoming bundle of rays will contract in the direction perpendicular to the ray (fig. 6.5). For obvious reasons, this effect is called *self-focussing*.

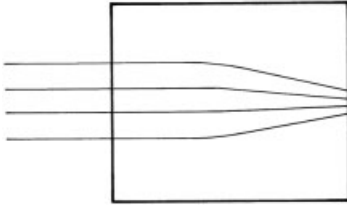


Figure 6.5: Self-focussing of a light beam in a nonlinear medium.

A large incoming intensity can lead to a vast contraction factor and an extremely high energy flux density. The field strength may reach values that are sufficient for ionization processes and for the destruction of the nonlinear material.

Instead of a longitudinally extended ray bundle, we now consider a light pulse that is concentrated in the longitudinal direction. At the front and rear ends of the pulse, the intensity and the refractive index are smaller and thus the propagation velocity is larger. As the front and rear parts of the pulse propagate faster than the middle part, there is a tendency for the pulse to change its shape, as indicated in fig. 6.6. While the pulse is passing through the medium, its front side will become flatter while its rear side will become steeper. This nonlinear effect is called *phase self-modulation*.

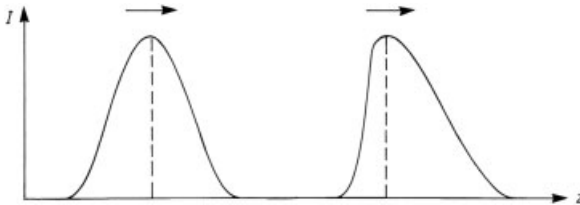


Figure 6.6: Changing pulse shape by phase self-modulation.

Let us now investigate how the spatial Fourier transform of the pulse changes inside a medium. The non-constant refractive index leads to a change of wavelengths such that the bandwidth Δk_z of the signal broadens as a function of k_z . We can even say more: due to the different slopes of the sides, the smaller wavenumbers k_z will contribute more to the front side and the larger wavenumbers to the rear side.

(Superimposed on the nonlinear effect is the normal linear dispersion effect, which leads to a diverging pulse, because the propagation velocity depends on the wavelength; however, this effect does not change the bandwidth Δk_z and, therefore, does not interfere with the nonlinear effect.)

After the pulse has passed through the nonlinear medium and acquired the shape just described, one can, in a second step, use a suitable device to give the pulse a narrow spatial concentration along the z axis by allowing the shorter wavelengths from behind to catch up with the longer wavelengths in front. This is possible, for example, by letting the distorted pulse traverse a medium with anomalous dispersion where the propagation velocity decreases as a function of the wavelength.

Practically, the second step is achieved by reflection at two refraction gratings. As one can see in fig. 6.7, the gratings reflect the longer wavelengths through a larger angle, thereby retarding them due to the longer distance.

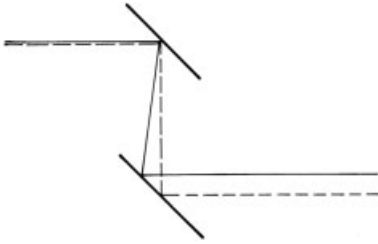


Figure 6.7: Reflection at two refraction gratings delays waves with larger wavelengths.

The total factor for the longitudinal contraction reached by such methods can be of the order of 100. For wavelengths of about $\lambda \approx 6 \times 10^{-4}$ mm, one can obtain in this way pulse durations of the order of femtoseconds (10^{-15} s).

In section 6.3 we have already mentioned that for very high light intensities the absorption coefficient can dramatically decrease as the result of saturation effects. We now describe briefly how this effect can be used for the construction of an optically bistable device.

A medium showing saturation behavior for absorption is placed between the two étalons of a Fabry–Perot interferometer and the setup is illuminated by light with intensity I_0 . We have already discussed the theory of the Fabry–Perot étalons in section 4.7, but now we have to take into account the absorption of the medium in order to determine the transmission of the device. From the derivation in section 4.7 it is obvious that this can easily be achieved by replacing the reflection coefficient r in eq. (4.56) by rt^2 , where t denotes the intensity decrease for one passage as well as a possible phase shift. From eq. (4.55) we now obtain for the emitted intensity I_1 :

$$I_1 = |\tau_1 \tau_2|^2 \frac{1}{|1 - rt^2 e^{+2i\varphi}|^2} I_0 =: T(t) I_0. \quad (6.46)$$

For the present discussion, t depends on the intensity I' inside the étalons immediately in front of the medium, and we learn from the calculation in section 4.7 that $I_1 = |\tau_2 t|^2 I'$. Therefore, we obtain the following relation between I_1 and I_0 :

$$I_0 = I_1 \frac{1}{T(t(I_1))}. \quad (6.47)$$

For small values of I_1 , T is almost independent of I_1 . The same is true for very large values of I_1 , when t has reached its saturation value $|t| \approx 1$. In these two regimes, the relation between I_0 and I_1 is almost linear.

In between, $t(I_1)$ and $T(t(I_1))$ increase as a function of I_1 . In general, the relation between I_1 and I_0 is no longer monotonic but follows the pattern sketched in fig. 6.8. If we start from small values and gradually increase the intensity of I_0 , the transmitted intensity I_1 initially also increases continuously. However, at a value $I_0 = A_0$ it will suddenly jump from A_1 to B_1 , and from there on increase again continuously as a function of I_1 . If now I_0 is decreased, also I_1 will decrease continuously first, but then make a jump at $I_0 = B_0 < A_0$ from B'_1 to A'_1 .

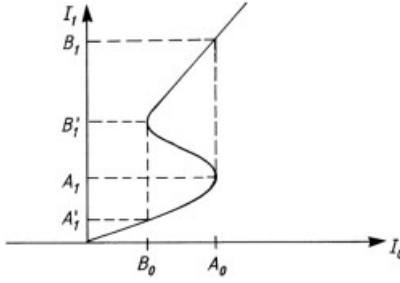


Figure 6.8: Optical bistability.

Hence, we find bistability and a hysteresis effect in the relation between I_0 and I_1 . The arrangement just described corresponds to an optical switch element for which the variation of I_0 induces jumps between two stable branches of $I_1(I_0)$.

6.7 Phase conjugation

A plane mirror reflects a plane wave in such a way that only the direction of propagation is reversed, while surfaces of constant phase remain planes. An incoming plane wave $u(t, \mathbf{x}) = e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}$ becomes a reflected wave $\bar{u}(t, \mathbf{x}) = R e^{-i(\omega t + \mathbf{k} \cdot \mathbf{x})}$, where R has a fixed amplitude and phase. In a similar way, a spherical mirror reflects a matching spherical wave into itself: $u(t, \mathbf{x}) = (1/r) e^{-i(\omega t - \mathbf{k} \cdot \mathbf{r})}$ is transformed into $\bar{u}(t, \mathbf{x}) = (R/r) e^{-i(\omega t + \mathbf{k} \cdot \mathbf{r})}$ (see fig. 6.9).

In general, given a wave $u_\omega(t, \mathbf{x}) = e^{-i\omega t} A(\mathbf{x}) e^{i\varphi(\mathbf{x})} = e^{-i\omega t} u_\omega(\mathbf{x})$ with fixed frequency ω , arbitrary amplitude $A(\mathbf{x})$, and phase $\varphi(\mathbf{x})$, we define the so-called *phase conjugated wave* to be:

$$u_\omega^c(t, \mathbf{x}) = e^{-i\omega t} A(\mathbf{x}) e^{-i\varphi(\mathbf{x})} := e^{-i\omega t} u_\omega^c(\mathbf{x}) = e^{-i\omega t} u_\omega^*(\mathbf{x}). \quad (6.48)$$

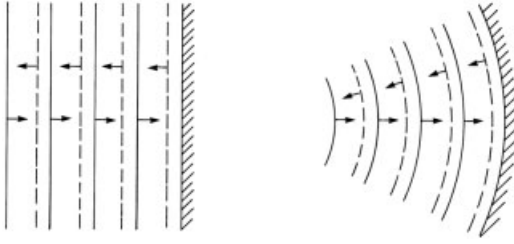


Figure 6.9: Inversion of wave fronts at a plane mirror and a spherical mirror.

The surfaces of constant phase of $u(t, \mathbf{x})$ and $u^c(t, \mathbf{x})$ have the same shape, but the directions of propagation of $u(t, \mathbf{x})$ and $u^c(t, \mathbf{x})$ are reversed. For the Fourier transform we use the identities

$$\begin{aligned} u_\omega(\mathbf{x}) &= (2\pi)^{-3/2} \int d^3k e^{i\mathbf{k}\cdot\mathbf{x}} \tilde{u}_\omega(\mathbf{k}), \\ u_\omega^c(\mathbf{x}) &= u_\omega^*(\mathbf{x}) = (2\pi)^{-3/2} \int d^3k e^{-i\mathbf{k}\cdot\mathbf{x}} \tilde{u}_\omega^*(\mathbf{k}) \\ &= (2\pi)^{-3/2} \int d^3k e^{i\mathbf{k}\cdot\mathbf{x}} \tilde{u}_\omega^*(-\mathbf{k}), \end{aligned}$$

and obtain

$$\tilde{u}_\omega^c(\mathbf{k}) = \tilde{u}_\omega^*(-\mathbf{k}). \quad (6.49)$$

One might ask if it is possible to construct a mirror that transforms *any* wave $u_\omega(t, \mathbf{x})$ into $Ru_\omega^c(t, \mathbf{x})$, i.e. each incoming wave is reflected into itself (fig. 6.10). In order to reflect a wave without changing its shape, such a phase conjugating mirror has to adapt itself automatically to the shape of any incoming wave surfaces. In this section, we will show how such a mirror can be realized with the help of nonlinear media. An eye looking into such a phase conjugating mirror would only see its own pupil.

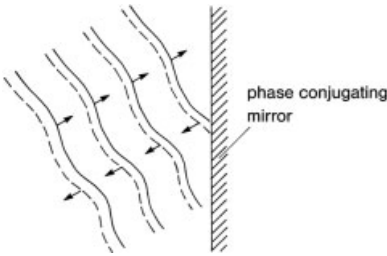


Figure 6.10: Reflection at a phase conjugating mirror.

Of course, the main reason for the broad interest in this construction is not based on this effect, but because it helps to undo the distortions that a light wave experiences during its passage through an irregular medium (fig. 6.11).

For instance, refraction at spatial density fluctuations will stretch a highly concentrated light bundle in the atmosphere. This expansion becomes broader if the light bundle runs back after being reflected at an ordinary mirror, but it is compensated if it has been reflected at a phase conjugating mirror.

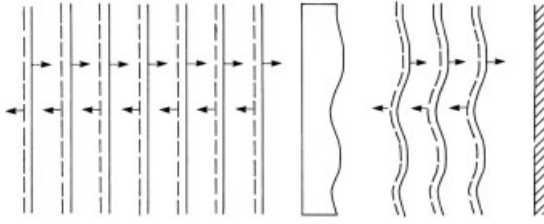


Figure 6.11: Compensation of the distortions of wave fronts after reflection at a phase conjugating mirror.

We consider a wave $u_\omega(t, \mathbf{x})$ entering a nonlinear medium. The relevant mechanism for the production of a phase conjugated wave $Ru_\omega^c(t, \mathbf{x})$ consists of stimulated Brillouin scattering, i.e. the interaction of a light wave with (ultra-)sound waves previously excited by the wave itself. The following explanations anticipate some knowledge about refraction and holography, which we will study in a later chapter and which may be skipped in a first reading.

The wave $u_\omega(t, \mathbf{x})$ is refracted at ultrasonic waves which are always present in the medium. Refraction occurs, because ultrasonic waves consist of density fluctuations and refractive index is density-dependent. Since the refraction takes place at a moving grating, the refracted wave shows a small frequency shift $|\Delta\omega| \leq \omega c_s/c$, where c_s is the velocity of sound and c is the velocity of light. The sign of the frequency shift depends on the propagation direction of the ultrasonic wave: $\Delta\omega$ is positive when the ultrasonic wave approaches the wave $u_\omega(t, \mathbf{x})$. For a refraction exactly in the backwards direction, a special effect occurs. First, in this case the frequency shift is maximal (while it vanishes in the forward direction); and, second, the interference of the refracted wave and the incoming wave produces “beats”, i.e. modulated amplitude patterns that travel through the medium with velocity $\pm c_s$.

By electrooptical interaction the pattern of the electromagnetic beats generates a correspondingly shaped ultrasonic wave, which again amplifies the backward refraction and thereby itself. This self-amplification eventually leads to a spatial phase lattice consisting of ultrasonic waves, which mimics the shape of the wave surfaces $u_\omega(t, \mathbf{x})$ and results in a conjugated wave by backward refraction.

A different and frequently used method is based on the working principle of holography. It utilizes the so-called *degenerate four-wave interaction*, a special case of stimulated Brillouin scattering, which is also mediated by sound waves. A planar reference wave $u_0(t, \mathbf{x})$ and a phase conjugated plane wave $u_0^c(t, \mathbf{x})$ are directed simultaneously and from opposite directions into the medium (see fig. 6.12). The wave pattern generated by the superposition of the reference wave $u_0(t, \mathbf{x})$ and the additional object wave $u(t, \mathbf{x})$ produces a phase lattice of analogous shape, which resembles a kind of volume hologram.

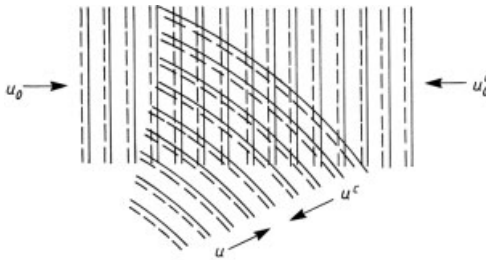


Figure 6.12: Phase conjugation by degenerate four-wave interaction.

The wave u_0^c serves as the reconstruction wave and produces holographically the conjugated wave u^c . The production mechanism is stabilized and amplified by the fact that the superposition of u_0^c and u^c generate the same phase lattice.

If we compare the above considerations with our discussion on holography in chapter 11, we notice that:

- the transition from an amplitude to a phase lattice does not fundamentally change the mechanism of holography, and,
- if the reconstruction wave is phase conjugated, the principal wave generated holographically is also phase conjugated to the object wave.

6.8 Fiber optics and optical solitons

Compared to radio waves, light waves have a frequency and bandwidth that are several orders of magnitude larger. Therefore, light waves are particularly suited for the construction of information transmission channels of very high capacity. Given a typical bandwidth of 5×10^{13} Hz, information transmission rates of 10^{12} bit/s become possible.

Optical information transmission uses *glass fibers*. Today's systems reach transmission rates of about 10^8 bit/s; systems with transmission rates of 10^{12} bit/s use *optical solitons* and are presently being developed.

The basic structure of a glass fiber is always the same (fig. 6.13): a core with a diameter of 3–50 μm is placed inside a cladding of 150–200 μm diameter. By suitable doping, e.g. with GeO_2 , one can arrange that the refractive index of the core glass is smaller than that of the cladding. Depending on whether the change of the refractive index between core and cladding occurs stepwise or gradually, one speaks of step profile (or step index) fibers or gradient profile fibers, respectively.

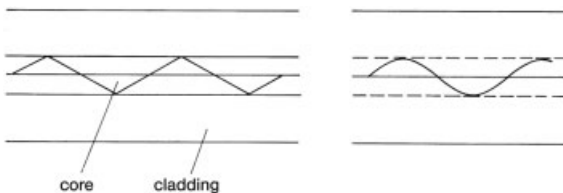


Figure 6.13: Beam trajectories for step profile and gradient profile fibers.

A light beam traveling through the core and having only a small inclination with respect to the fiber axis will remain in the core region of the fiber. In step profile fibers it will be totally reflected at the interface between core and cladding, while in gradient profile fibers it will be bent back toward the center of the core due to the decreasing refractive index.

The signal in the fiber is digitized, i.e. it consists of a series of short pulses. However, a disturbing influence may come from the so-called *mode dispersion*: beams that are inclined relative to the axis of the cable traverse a longer trajectory compared to beams parallel to the axis. Differences in the transition times can lead to a broadening of an initially short pulse.

In gradient profile fibers, by a suitable choice of the refractive index profile $n(r)$, one can arrange that this time delay along the longer trajectory is compensated due to the decrease in $n(r)$ and the corresponding increase in velocity for larger values of r . This method is generally used in today's glass-fiber cables, and yields a reduction of mode dispersion by a factor of 100; the temporal broadening of the pulses amounts to about 0.5 ns/km.

Two minima in the absorption coefficients of glass confine the practically used wavelength regimes to two windows at 1.3 μm and 1.55 μm .

If the diameter of the core is reduced to only a few micrometers, the just described propagation and mode dispersion in terms of ray optics is no longer valid. In this case, a wave theoretical treatment is necessary and one has to solve the following equation:

$$\left(\Delta + n^2(r, \omega) \frac{\omega^2}{c^2} \right) u = 0. \quad (6.50)$$

In cylindrical coordinates the ansatz

$$u(r, \varphi, z) = \chi_m(r) e^{im\varphi} e^{ikz} \quad (6.51)$$

leads to the equation

$$\left(\frac{d^2}{dr^2} + \frac{1}{r} \frac{d}{dr} - \frac{m^2}{r^2} + n^2(r, \omega) \frac{\omega^2}{c^2} - k^2 \right) \chi_m(r) =: [\Lambda(\omega^2) - k^2] \chi_m(r) = 0. \quad (6.52)$$

We look for eigenmodes $\chi_{\nu m}(r)$ of the differential operator $\Lambda(\omega^2)$ that are regular at $r = 0$ and decay sufficiently fast for $r \rightarrow \infty$. They correspond to the discrete eigenvalues $\kappa_{\nu m}^2(\omega^2)$:

$$\Lambda(\omega^2) \chi_{\nu m}(r) = \kappa_{\nu m}^2(\omega^2) \chi_{\nu m}(r) \quad (\nu = 1, 2, \dots; m = 0, \pm 1, \dots). \quad (6.53)$$

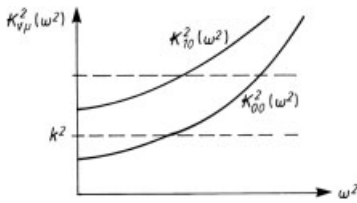


Figure 6.14: Dispersion relations for two different modes.

The dispersion relation, which gives us ω as a function of k , follows from the equation $\kappa_{\nu m}^2(\omega^2) = k^2$. Figure 6.14 shows the graphical solution of ω as a function of k . We see the following:

- In general, we find a different dispersion relation for each eigenvalue $\kappa_{\nu m}^2(\omega^2)$. This corresponds to the phenomenon of mode dispersion in the language of wave optics.
- For very small values of $k^2 < \kappa_{00}^2(0)$, no wave can propagate. For large values of k^2 , several or even many modes are possible.

- Mode dispersion can be avoided completely in so-called *single-mode fibers*. The dimensions of these fibers are such that only the fundamental mode with eigenvalue $\kappa_{00}^2(\omega^2)$ is possible. Single-mode fibers have very small core diameters, because the distance between eigenvalues increases as the core diameter becomes smaller and, eventually, only one mode is allowed in the wavenumber range for k .

Common single-mode fibers use a step profile for $n(r)$. Let us investigate the propagation behavior in single-mode fibers in more detail. The disturbing mode dispersion is completely absent, but, because of the dependence $k^2 = \kappa_{00}(\omega^2)$, there still exists a small dispersion effect, which is called *chromatic dispersion*, because it is the frequency dependence of the refractive index that contributes. We write the dispersion relation in the form:

$$k = f(\omega). \quad (6.54)$$

The derivative of f is given by $f'(\omega) = dk/d\omega = (d\omega/dk)^{-1}$, i.e.

$$f'(\omega) = \frac{1}{v_g(\omega)}, \quad (6.55)$$

where $v_g(\omega)$ denotes the group velocity, which also depends on the frequency, a phenomenon called *group velocity dispersion*. Depending on whether $dv_g(\omega)/d\omega$ is positive or negative, we speak of normal or anomalous group velocity dispersion. Obviously,

$$f''(\omega) = -\frac{1}{v_g^2} \frac{dv_g}{d\omega}. \quad (6.56)$$

As we have already mentioned in section 6.2, nonlinearities can compensate the chromatic dispersion if the nonlinear wave equation admits soliton solutions. We now show that the wave propagation in single-mode fibers with nonlinear behavior can be described by the nonlinear Schrödinger equation.

For k and ω fixed, a monochromatic wave may be written as

$$u = \chi_{00}(r) e^{-i(\omega t - f(\omega)z)}.$$

In general, a propagating wave is of the form $u = \chi_{00}(r)E(t, z)$. For $E(t, z)$ we can assume a one-dimensional nonlinear wave equation of the type (6.12), where now we replace the third and fifth assumptions of section 6.2 by this one:

- $E(t, z)$ can be written as

$$E(t, z) = \exp\{-i[\omega_0 t - f(\omega_0)z]\}e(t, z) \quad (6.57)$$

with slowly varying $e(t, z)$.

In particular, this implies that only small frequencies and wavenumbers contribute to the Fourier transform of $e(t, z)$. Using the Fourier transform $\tilde{E}(\omega, k)$, the nonlinear wave equation reads:

$$\text{Re} \left\{ \frac{1}{2\pi} \int d\omega \int dk e^{-i(\omega t - kz)} [k^2 - f^2(\omega)] \tilde{E}(\omega, k) \right\} = -\frac{\partial^2}{\partial t^2} P'(E). \quad (6.58)$$

The above assumption allows for the approximation

$$k^2 - f^2(\omega) \approx 2k_0[k - f(\omega)] \approx 2k_0[\Delta k - f'(\omega_0)\Delta\omega - \frac{1}{2}f''(\omega_0)(\Delta\omega)^2], \quad (6.59)$$

with $k_0 = f(\omega_0)$, such that eq. (6.58) now becomes

$$\begin{aligned} & \text{Re} \left\{ \frac{1}{2\pi} e^{-i(\omega_0 - k_0 z)} 2k_0 \int d\omega \int dk e^{-i(\Delta\omega t - \Delta k z)} \right. \\ & \quad \times [\Delta k - f'(\omega_0)\Delta\omega - \frac{1}{2}f''(\omega_0)(\Delta\omega)^2] \tilde{E}(\omega, k) \Big\} \\ & = \text{Re} \left\{ 2k_0 e^{-i(\omega_0 - k_0 z)} \left(-i \frac{\partial}{\partial z} - i f'(\omega_0) \frac{\partial}{\partial t} + \frac{1}{2} f''(\omega_0) \frac{\partial^2}{\partial t^2} \right) e(t, z) \right\} \\ & = -\frac{\partial^2}{\partial t^2} P'(E). \end{aligned}$$

Along similar lines as in section 6.2, we finally obtain the equation:

$$-i \left(\frac{\partial}{\partial z} + \frac{1}{v_g} \frac{\partial}{\partial t} \right) e(t, z) + \frac{1}{2} f''(\omega_0) \frac{\partial^2}{\partial t^2} e(t, z) = \alpha |e(t, z)|^2 e(t, z). \quad (6.60)$$

The following simple substitutions,

$$\begin{aligned} \zeta &= \frac{1}{v_g \tau_0} (z - v_g t), \quad \tau = -\frac{f''(\omega_0) z}{2\tau_0^2}, \\ \psi(\tau, \zeta) &= \sqrt{\frac{\alpha \tau_0^2}{|f''(\omega_0)|}} e(\tau, \zeta), \end{aligned} \quad (6.61)$$

lead to the nonlinear Schrödinger equation:

$$i \frac{\partial \psi}{\partial \tau} + \frac{\partial^2 \psi}{\partial \zeta^2} - 2\epsilon \psi |\psi|^2 = 0. \quad (6.62)$$

Here, $\epsilon = -1$ for anomalous group velocity dispersion (i.e. $f'' < 0$) and $\epsilon = +1$ for normal group velocity dispersion.

The case $f'' < 0$ is realized for a single-mode fiber. Written in terms of the new variables (6.61), the soliton solution (6.22) of the nonlinear Schrödinger equation reads:

$$\psi(t, z) = 2b \frac{\exp \left[-\frac{2ia}{v_g \tau_0} (z - v_g t) + 4i(a^2 - b^2) \frac{f''}{2\tau_0^2} z \right]}{\cosh \left[2b \left(\frac{z - v_g t}{v_g \tau_0} - \frac{2af''z}{\tau_0^2} \right) \right]}. \quad (6.63)$$

In practice, one tries to implement *frequency matching*: the phase of ψ is time-independent because $E(t, z)$ oscillates with frequency ω_0 . So we set $a = 0$ in eq. (6.63) and obtain for the form of the soliton:

$$\psi(t, z) = 2b \frac{\exp \left[\frac{-2ib^2 f'' z}{\tau_0^2} \right]}{\cosh \left[\frac{2b(z - v_g t)}{v_g \tau_0} \right]}. \quad (6.64)$$

This is a spatially concentrated pulse-like excitation that propagates through the glass fiber without any change of shape and with group velocity v_g . The height of the pulse is larger for smaller width.

Such solitons can be used in single-mode fibers as dispersion-free signal pulses. They allow for transmission rates of about 10^{12} bit/s.

The case $\epsilon = 1$ in the nonlinear Schrödinger equation is also of physical importance. It can be used for a detailed discussion of momentum contraction, which we described in section 6.5.

7 Short-wave asymptotics

7.1 Introductory remarks

Common to all optical systems is that light passes through an inhomogeneous medium where the inhomogeneities, like lenses and mirrors, are suitably shaped according to their purpose. Having so far studied the propagation of waves in homogeneous but possibly anisotropic media, we now will turn to the discussion of inhomogeneous media. Most of the time we will assume that the inhomogeneous media are pointwise isotropic. On the other hand, anisotropic media often exhibit a crystalline structure or at least a certain short-range ordering of their atoms and thus are piecewise homogeneous.

Common everyday optical phenomena are often determined by the fact that the wavelength of the light is small compared to the dimensions of the optical systems and, in many cases, even small compared to the typical length scale of its inhomogeneities. Indeed, the wave-like features of light are often hidden to such an extent that, for centuries, Newton's corpuscular theory, which treated light as a flux of small particles, lasted in mainstream scientific thinking. Only in sophisticated experiments is the wave-like character of light revealed by interference and diffraction phenomena. Quite a different matter is that, due to the wave-particle duality, light also exhibits particle-like properties, which dominate its behavior at very short wavelengths and low intensities.

Since the wavelengths of light are tiny with respect to the size of normal optical systems, the relevant solutions $u(t, \mathbf{x})$ of the wave equation are locally given by plane waves. This approximation holds for spatial regions that are large compared to the wavelength but small compared to the typical dimensions of the system. Therefore, the amplitudes and the propagation direction of a wave will not change considerably within a distance of several wavelengths. As a generalization of the expression for a plane wave with fixed frequency ω and a wave vector of absolute value $k = \omega/c$,

$$u_k(t, \mathbf{x}) = \varphi_0 e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})} = \varphi_0 e^{-ik(ct - \mathbf{n} \cdot \mathbf{x})},$$

we write $u(t, \mathbf{x})$ in the following way:

$$u_k(t, \mathbf{x}) = \varphi_k(\mathbf{x}) e^{ik[S(\mathbf{x}) - ct]}. \quad (7.1)$$

The (real) function $S(\mathbf{x})$ determines the phase of the wave at $\mathbf{x}$, and the function $\varphi_k(\mathbf{x})$ (also real) fixes the amplitude of the wave. We assume that the variations of $\varphi_k(\mathbf{x})$ as a function of $\mathbf{x}$ are much smaller than the variations of the phase $kS(\mathbf{x})$. To be more precise, we require that $S(\mathbf{x})$ is independent of k , and for large values of k the following expansion for $\varphi_k(\mathbf{x})$ is

valid:

$$\varphi_k(\mathbf{x}) = \varphi_0(\mathbf{x}) + \frac{1}{k}\varphi_1(\mathbf{x}) + \frac{1}{k^2}\varphi_2(\mathbf{x}) + \cdots. \quad (7.2)$$

This ansatz for $u_k(t, \mathbf{x})$, which we will insert into the wave equation, is fundamental for the short-wave approximation in optics. The second fundamental tool in this context is the method of stationary phases, which we will discuss in chapter 10.

Before we apply the short-wave expansion to Maxwell's equations, we want to list some important comments on the ansatz for $u_k(t, \mathbf{x})$.

1. The function $S(\mathbf{x})$ is sometimes called the *eikonal function*, because of its relevance for the theory of imaging in geometrical optics. The surfaces $S(\mathbf{x}) = \text{const.}$ are surfaces of constant phase. Their shape depends crucially on the chosen solution. Expanding around $\mathbf{x} = \mathbf{x}_0$ with $S(\mathbf{x}) = S(\mathbf{x}_0) + (\mathbf{x} - \mathbf{x}_0) \cdot \nabla S(\mathbf{x}_0) + \cdots$ and comparing with the expression for a plane wave, we observe that

$$\mathbf{k}(\mathbf{x}_0) = k\mathbf{n}(\mathbf{x}_0) = k\nabla S(\mathbf{x}_0) \quad (7.3)$$

is the local wave vector in $\mathbf{x}_0$, which is perpendicular to the surface of constant phase through $\mathbf{x}_0$. Furthermore, $\mathbf{n}(\mathbf{x}_0) = \nabla S(\mathbf{x}_0)$ is the local refractive index vector and $\mathbf{S}(\mathbf{x}_0) = \mathbf{n}/n^2$ is the local ray vector with $\mathbf{S} \cdot \mathbf{n} = 1$.

For a plane wave we have $S(\mathbf{x}) = \mathbf{n} \cdot \mathbf{x}$ with $\mathbf{n} \equiv \text{const.}$, and the surfaces of constant phase are planes. On the other hand, when $u_k(t, \mathbf{x})$ describes the propagation of a point-like disturbance starting at $\mathbf{x} = 0$, the surfaces $S(\mathbf{x}) = \text{const.}$ are, for not too large values of $|\mathbf{x}|$, spheres or ellipsoids, which may be more or less distorted and can be viewed as a generalization of the ray surface.

2. Taking superpositions of solutions of the type $u_k(t, \mathbf{x})$, where the frequency is concentrated in a small band around k , one can convince oneself that the surfaces $S(\mathbf{x}) = \text{const.}$ are also surfaces of shell-like concentrations of shock-wave solutions. The energy of such solutions is mainly concentrated within a small region around $S(\mathbf{x}) = \text{const.}$
3. Surfaces of constant phase, $S(\mathbf{x}) = \text{const.}$, also determine the directions of the energy flux. In locally isotropic media the energy flux is parallel to the local wave vector $\mathbf{k}(\mathbf{x}) = k\mathbf{n}(\mathbf{x})$ and, hence, always perpendicular to the surfaces of constant phase. In this way *light rays* result as the orthogonal trajectories to the phase surfaces. In optical experiments, these light rays are directly observable as stretched out wave packets. Indeed, the energy of a wave packet always follows the direction of the energy flux vector. The uncertainties of a wave packet in space and wavelength (Δx and Δk , respectively) satisfy the general uncertainty relations:

$$\Delta k \Delta x \geq \frac{1}{2} \quad \text{or} \quad \frac{\Delta k}{k} \frac{\Delta x}{x} \geq \frac{1}{2kx}. \quad (7.4)$$

Within the validity range of the short-wave approximation, one can construct wave packets for which the uncertainty orthogonal to the energy flux vector may be neglected. The trajectories of these packets coincide with the light rays.

The theory of the light-wave field that results from keeping only the leading order in $1/k$ is called *geometrical optics*.

7.2 Short-wave expansion of Maxwell's equations

In a locally isotropic medium, Maxwell's equations for fields with a fixed frequency ω and vanishing charge and current densities, $\rho \equiv 0$ and $j \equiv 0$, read

$$\begin{aligned}\epsilon_0 \nabla \cdot (\epsilon \mathbf{E}) &= 0, & \mu_0 \nabla \cdot (\mu \mathbf{H}) &= 0, \\ \nabla \times \mathbf{H} &= -i\omega \kappa \epsilon \epsilon_0 \mathbf{E}, & \nabla \times \mathbf{E} &= i\omega \kappa \mu \mu_0 \mathbf{H},\end{aligned}\tag{7.5}$$

with $\mathbf{D} = \epsilon_0 \epsilon \mathbf{E}$, $\mathbf{B} = \mu_0 \mu \mathbf{H}$, $\epsilon = \epsilon(\omega, \mathbf{x})$, and $\mu = \mu(\omega, \mathbf{x})$. We insert the ansatz described in section 7.1 into these equations,

$$\begin{aligned}\mathbf{E}(t, \mathbf{x}) &= \mathbf{e}(\mathbf{x}) e^{ik[S(\mathbf{x}) - ct]}, \\ \mathbf{H}(t, \mathbf{x}) &= \mathbf{h}(\mathbf{x}) e^{ik[S(\mathbf{x}) - ct]},\end{aligned}$$

with $k = \omega/c$. Taking into account that

$$\begin{aligned}\nabla \times \mathbf{E} &= (\nabla \times \mathbf{e} + ik \nabla S \times \mathbf{e}) e^{ik[S(\mathbf{x}) - ct]}, \\ \nabla \cdot \mathbf{E} &= (\nabla \cdot \mathbf{e} + ik \nabla S \cdot \mathbf{e}) e^{ik[S(\mathbf{x}) - ct]},\end{aligned}\tag{7.6}$$

we obtain

$$\begin{aligned}\mathbf{e} \cdot \nabla S &= -\frac{1}{ik} \left(\mathbf{e} \cdot \frac{\nabla \epsilon}{\epsilon} + \nabla \cdot \mathbf{e} \right), \\ \mathbf{h} \cdot \nabla S &= -\frac{1}{ik} \left(\mathbf{h} \cdot \frac{\nabla \mu}{\mu} + \nabla \cdot \mathbf{h} \right), \\ (\nabla S \times \mathbf{e} - \mu \mu_0 \kappa c \mathbf{h}) &= -\frac{1}{ik} \nabla \times \mathbf{e}, \\ (\nabla S \times \mathbf{h} + \epsilon \epsilon_0 \kappa c \mathbf{e}) &= -\frac{1}{ik} \nabla \times \mathbf{h}.\end{aligned}\tag{7.7}$$

The right-hand sides of these equations should be negligible in leading order for $k \rightarrow \infty$. “Dangerous” are only those situations for which the expressions following $-1/ik$ become unusually large. For normal optical systems this can happen

- in the immediate vicinity of edges, where the amplitudes $\mathbf{e}$ and $\mathbf{h}$ as well as ϵ and μ depend strongly on the space variable $\mathbf{x}$, and
- close to focal points or, more generally, caustics where the light rays get concentrated such that the amplitudes $\mathbf{e}$ and $\mathbf{h}$ become very large.

Apart from these two situations, which we will investigate later, we can trust the approximation of geometrical optics and assume for $k \rightarrow \infty$:

$$\begin{aligned}\mathbf{e} \cdot \nabla S &= 0, & \mathbf{h} \cdot \nabla S &= 0, \\ \nabla S \times \mathbf{e} &= \mu \mu_0 \kappa c \mathbf{h}, & \nabla S \times \mathbf{h} &= -\epsilon \epsilon_0 \kappa c \mathbf{e}.\end{aligned}\tag{7.8}$$

These equations resemble the result we obtained by making a plane-wave ansatz for the solution of Maxwell's equations, where now ∇S plays the role of $\mathbf{n}$. Eliminating $\mathbf{e}$ or $\mathbf{h}$ leads to

$$\nabla S \times (\nabla S \times \mathbf{e}) = -\epsilon\mu\mathbf{e}$$

or

$$(\nabla S)^2 = n^2 \quad \text{with} \quad n^2(\omega, \mathbf{x}) = \epsilon\mu. \quad (7.9)$$

This partial differential equation for the eikonal function $S(\mathbf{x})$ is called the *eikonal equation*. Different solutions for this equation correspond to different wave fields. The physical meaning of the eikonal equation is that the local propagation velocity is given by $c(\mathbf{x}) = c/n(\mathbf{x})$.

In leading order of $k \rightarrow \infty$, the following expressions for the time-averaged energy density and Poynting vector $\mathbf{S}$ are easily determined:

$$\begin{aligned} \langle \rho^{\text{el}} \rangle &= \frac{\epsilon\epsilon_0}{4} \mathbf{e} \cdot \mathbf{e}^* = \frac{1}{4\kappa c} (\mathbf{e} \times \mathbf{h}^*) \cdot \nabla S, \\ \langle \rho^{\text{mag}} \rangle &= \frac{\mu\mu_0}{4} \mathbf{h} \cdot \mathbf{h}^* = \frac{1}{4\kappa c} (\mathbf{e} \times \mathbf{h}^*) \cdot \nabla S = \langle \rho^{\text{el}} \rangle, \\ \langle \rho^{\text{E}} \rangle &= \langle \rho^{\text{el}} \rangle + \langle \rho^{\text{mag}} \rangle, \\ \langle \mathbf{S} \rangle &= \frac{1}{2\kappa} \text{Re}(\mathbf{e} \times \mathbf{h}^*) = \frac{1}{2\kappa^2 c \mu \mu_0} (\mathbf{e} \cdot \mathbf{e}^*) \nabla S \\ &= \frac{c}{n^2} \langle \rho^{\text{E}} \rangle \nabla S = \frac{c}{n} \langle \rho^{\text{E}} \rangle \frac{\nabla S}{n} = \frac{I}{n} \nabla S. \end{aligned} \quad (7.10)$$

We make the following observations:

- The average electric and magnetic energy densities are equal.
- Energy flows along the direction of the vector ∇S , i.e. perpendicular to the surfaces of constant phase and along the direction of the local wave vector.
- From the eikonal equation we see that $\boldsymbol{\sigma} = (1/n)\nabla S$ is a unit vector, and, therefore, $I := |\langle \mathbf{S} \rangle| = (c/n)\langle \rho^{\text{E}} \rangle$. Hence, the velocity of the energy flow is indeed equal to the local phase velocity c/n .

Notice that, in leading order of k , there is only the eikonal equation for the phase kS , and there are no restrictions for the amplitudes $\mathbf{e}$ and $\mathbf{h}$. By choosing $\mathbf{e}$ and $\mathbf{h}$ properly, we can generate solutions which, in leading order, correspond to sharply bounded rays that are everywhere parallel to ∇S . The requirement of energy conservation,

$$\nabla \langle \mathbf{S} \rangle = \nabla \left(\frac{I}{n} \nabla S \right) = 0,$$

yields

$$\frac{I}{n} \Delta S + \nabla S \cdot \nabla \left(\frac{I}{n} \right) = 0, \quad \text{or} \quad \Delta S + \nabla S \cdot \nabla \ln \left(\frac{I}{n} \right) = 0. \quad (7.11)$$

As we have seen, the light rays are the flux lines $\mathbf{x}(\tau)$ of the vector field ∇S . They satisfy the differential equation:

$$\frac{d\mathbf{x}}{d\tau} = \nabla S. \quad (7.12)$$

Along the light rays $\mathbf{x}(\tau)$, we obtain for eq. (7.11):

$$\Delta S(\mathbf{x}(\tau)) + \frac{d}{d\tau} \ln \left(\frac{I}{n} \right) = 0. \quad (7.13)$$

Notice that along the rays the intensity change is determined by the divergence of the flow field ∇S . Integrating both sides we get

$$\frac{I}{n}(\mathbf{x}(\tau)) = \frac{I}{n}(\mathbf{x}(0)) \exp \left(- \int_0^\tau d\tau' \Delta S(\mathbf{x}(\tau')) \right). \quad (7.14)$$

For the simplest case, $n \equiv 1$, we find that $S(\mathbf{x}) = |\mathbf{x}|$ is a rotationally symmetric solution of the eikonal equation. Now,

$$\nabla S = \frac{\mathbf{x}}{|\mathbf{x}|}, \quad \mathbf{x}(\tau) = \tau \frac{\mathbf{x}(0)}{|\mathbf{x}(0)|}, \quad |\mathbf{x}(\tau)| = \tau$$

and

$$\Delta S(\mathbf{x}(\tau)) = \frac{2}{|\mathbf{x}(\tau)|} = \frac{2}{\tau}.$$

Therefore,

$$I(\mathbf{x}) = I(\mathbf{x}_0) \frac{1}{|\mathbf{x}|^2}. \quad (7.15)$$

This describes the well-known quadratic decrease of the intensity as a function of distance.

If we had followed the same arguments for an anisotropic medium, the short-wave expansion would have been much more complicated. For each point, we would have obtained two different local eigensolutions with different propagation velocities, as we have seen in chapter 3. For this reason, the calculation would have involved two different eikonal functions corresponding to the two orientations of the principal polarizations.

7.3 The scalar wave equation

We now turn to the short-wave expansion for the scalar wave equation:

$$\left(\frac{n^2(\mathbf{x})}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) u(t, \mathbf{x}) = 0, \quad (7.16)$$

which follows as the equation of motion from the Lagrangian density

$$\mathcal{L} = \frac{1}{2} \left[\frac{n^2(\mathbf{x})}{c^2} \left(\frac{\partial u(t, \mathbf{x})}{\partial t} \right)^2 - (\nabla u(t, \mathbf{x}))^2 \right]. \quad (7.17)$$

As explained in chapter 3, this leads us to the following energy density and energy current density:

$$\begin{aligned}\rho^E &= \frac{1}{2} \left[\frac{n^2}{c^2} \left(\frac{\partial u(t, \mathbf{x})}{\partial t} \right)^2 + (\nabla u(t, \mathbf{x}))^2 \right], \\ \mathbf{S} &= - \frac{\partial u(t, \mathbf{x})}{\partial t} \nabla u(t, \mathbf{x}).\end{aligned}\tag{7.18}$$

The ansatz $u(t, \mathbf{x}) = u(\mathbf{x}) \exp(-i\omega t)$ yields Helmholtz's equation:

$$\left(\frac{1}{k^2} \Delta + n^2(\mathbf{x}) \right) u(\mathbf{x}) = 0 \quad \text{with} \quad k = \frac{\omega}{c}.\tag{7.19}$$

Writing $u(\mathbf{x}) = \varphi_k(\mathbf{x}) e^{ikS(\mathbf{x})}$, with $\varphi_k(\mathbf{x}) = \varphi_0(\mathbf{x}) + (1/k)\varphi_1(\mathbf{x}) + \dots$, we obtain

$$\begin{aligned}0 &= \left(\frac{1}{k^2} \Delta + n^2(\mathbf{x}) \right) u(\mathbf{x}) \\ &= \left(\frac{1}{k^2} \Delta \varphi_k + \frac{2i}{k} \nabla \varphi_k \nabla S + \frac{i\varphi_k}{k} \Delta S - \varphi_k (\nabla S)^2 + n^2 \varphi_k \right) e^{ikS}.\end{aligned}\tag{7.20}$$

Thus we find in zeroth order for $1/k \rightarrow 0$:

$$(\nabla S)^2 = n^2,\tag{7.21}$$

and in first order:

$$\begin{aligned}\varphi_0 \Delta S + 2 \nabla S \nabla \varphi_0 &= 0, \\ \Delta S + \nabla S \nabla \ln \varphi_0^2 &= 0, \\ \nabla(\varphi_0^2 \nabla S) &= 0.\end{aligned}\tag{7.22}$$

So we get the same eikonal equation as we did for Maxwell's equations. The equation of first order is called the *transport equation*.

Looking at the expressions for ρ^E and $\mathbf{S}$, we find in leading order of $1/k \rightarrow 0$:

$$\langle \rho^E \rangle = \frac{1}{2} n^2 k^2 \varphi_0^2, \quad \langle \mathbf{S} \rangle = \frac{1}{2} c k^2 \varphi_0^2 \nabla S = \frac{c}{n} \langle \rho^E \rangle \frac{\nabla S}{n}.\tag{7.23}$$

Hence, the transport equation describes the energy conservation and is identical to the energy balance condition for Maxwell's equations. So, as long as we do not have to take into account information about the polarization of light waves, which in any case depends strongly on the details of the system, the scalar wave equations provide, in first order of $1/k$, the same information about phases and intensities as Maxwell's equations. Therefore, the following discussion will be based on the scalar wave equation.

Along a light ray, where $d\mathbf{x}/d\tau = \nabla S$, we find in particular:

$$\varphi_0(\mathbf{x}(\tau)) = \varphi_0(\mathbf{x}(0)) \exp \left(-\frac{1}{2} \int_0^\tau d\tau' \Delta S(\mathbf{x}(\tau')) \right).\tag{7.24}$$

Let us consider a system of rays,

$$\mathbf{x}(\mathbf{x}_0, \tau) \quad \text{with} \quad \frac{d\mathbf{x}(\mathbf{x}_0, \tau)}{d\tau} = \nabla S(\mathbf{x}(\mathbf{x}_0, \tau)), \quad (7.25)$$

with initial condition

$$\mathbf{x}(\mathbf{x}_0, 0) = \mathbf{x}_0. \quad (7.26)$$

The mapping $\mathbf{x}_0 \mapsto g_\tau(\mathbf{x}_0) = \mathbf{x}(\mathbf{x}_0, \tau)$ defines a one-parameter family of (locally) invertible mappings from $\mathbb{R}^3$ into itself. The Jacobian $\mathcal{J}(\mathbf{x}_0, \tau) = \det(\partial \mathbf{x}^i(\mathbf{x}_0, \tau) / \partial x_0^j)$ is a measure for the distortions of volumes under these mappings. As is well known, $\mathcal{J}$ satisfies the following differential equation:

$$\frac{1}{\mathcal{J}} \frac{d\mathcal{J}}{d\tau} = \nabla \cdot \nabla S = \Delta S. \quad (7.27)$$

Therefore,

$$\mathcal{J}(\mathbf{x}_0, \tau) = \exp \left(\int_0^\tau d\tau' \Delta S(\mathbf{x}(\tau')) \right),$$

and instead of eq. (7.23) we can also write

$$\varphi_0(\mathbf{x}(\tau)) = \frac{\varphi_0(\mathbf{x}(0))}{\sqrt{\mathcal{J}(\mathbf{x}(\tau))}}. \quad (7.28)$$

We note that now the conservation of energy reveals itself as a τ independence of the following integral:

$$\int d^3x_0 \varphi_0(g_\tau(\mathbf{x}_0))^2.$$

7.4 Phase surfaces and rays

From a solution of the eikonal equation $(\nabla S)^2 - n^2 = 0$, a system of rays is obtained that are trajectories orthogonal to the level surfaces determined by the condition $S(\mathbf{x}) = \text{const.}$ (see fig. 7.1). On the other hand, we will see later that we can reconstruct a solution of the eikonal equation from a system of rays with suitable initial conditions on a single level surface. As $|(1/n)\nabla S| = 1$, the equation

$$n\boldsymbol{\sigma} = \nabla S \quad (7.29)$$

determines a field $\boldsymbol{\sigma}$ of unit vectors. Hence, if we choose the arc length s as a curve parameter for the light rays, they satisfy the following differential equation:

$$\frac{d\mathbf{x}}{ds} = \boldsymbol{\sigma}(\mathbf{x}(s)). \quad (7.30)$$

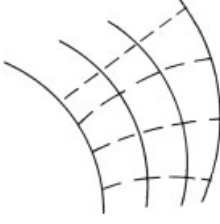


Figure 7.1: For a solution of the eikonal equation one obtains a system of rays that are orthogonal trajectories to the level surfaces of constant phase.

A simple calculation shows that

$$\begin{aligned} \frac{d}{ds} n \frac{d\mathbf{x}}{ds} &= \frac{d}{ds} (n\boldsymbol{\sigma}) = \frac{d}{ds} \nabla S = (\boldsymbol{\sigma} \cdot \nabla) \nabla S = \frac{1}{n} (\nabla S \cdot \nabla) \nabla S \\ &= \frac{1}{2n} \nabla (\nabla S)^2 = \frac{1}{2n} \nabla (n)^2 = \nabla n. \end{aligned} \quad (7.31)$$

Therefore, light rays are always curved towards the direction of increasing values of n . From

$$\nabla \times (n\boldsymbol{\sigma}) = \nabla \times \nabla S = 0, \quad (7.32)$$

we find that the integral

$$\int_{\gamma x_0}^{\mathbf{x}} n \boldsymbol{\sigma} \cdot d\mathbf{x}$$

is independent of the path γ that connects x_0 with $\mathbf{x}$. This implies that the phase $S(\mathbf{x})$ assumes a well-defined value at each point. From $\nabla \times (n\boldsymbol{\sigma}) = 0$ we can also derive Snell's law of refraction. We assume that n jumps from n_1 to n_2 at some interface, but is constant otherwise. By choosing a closed path γ (see fig. 7.2) which traverses the interface at two points and runs parallel to the interface in the regions where $n \equiv n_1$ and $n \equiv n_2$, we obtain

$$\int_{\gamma} d\mathbf{x} n \cdot \boldsymbol{\sigma} = 0,$$

and thus

$$n_1 \sigma_1'' = n_2 \sigma_2'',$$

where $\sigma_{1,2}''$ are the components of $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ parallel to the interface. Indeed, since $\sigma_{1,2}'' = \sin \theta_{1,2}$, where $\theta_{1,2}$ are the angles between the rays and the normal to the interface, we find

$$n_1 \sin \theta_1 = n_2 \sin \theta_2. \quad (7.33)$$

One may object to this derivation of the law of refraction, because the sudden change of n seems to contradict the applicability of the approximation underlying geometrical optics. Indeed, we know that refraction always comes together with reflection. Geometrical optics cannot predict the relative intensities of refracted and reflected waves.

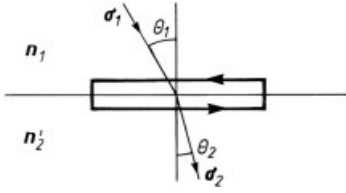


Figure 7.2: Geometrical illustration of Snell's law of refraction.

7.5 Fermat's principle

Evaluating the path-independent integral,

$$\int_{\gamma x_0}^x \nabla S \cdot d\mathbf{x} = \int_{\gamma x_0}^x n\boldsymbol{\sigma} \cdot d\mathbf{x} = \int_{x_0}^x n\boldsymbol{\sigma} \frac{d\mathbf{x}}{ds} ds = S(\mathbf{x}) - S(\mathbf{x}_0), \quad (7.34)$$

along some arbitrary path γ and making use of the inequality

$$\boldsymbol{\sigma} \cdot \frac{d\mathbf{x}}{ds} \leq |\boldsymbol{\sigma}| \left| \frac{d\mathbf{x}}{ds} \right| = 1,$$

we obtain

$$\int_{\gamma x_0}^x n\boldsymbol{\sigma} \cdot d\mathbf{x} \leq \int_{\gamma x_0}^x n ds. \quad (7.35)$$

Equality holds only if the path γ corresponds to a light ray, which, of course, always points into the direction of $\boldsymbol{\sigma}$ such that in this case $\boldsymbol{\sigma} \cdot (d\mathbf{x}/ds) = 1$. This leads us to *Fermat's principle*:

Among all trajectories γ running from $\mathbf{x}_0$ to $\mathbf{x}$, the trajectories that are realized by light rays yield the smallest possible value, $S(\mathbf{x}) - S(\mathbf{x}_0)$, for the integral $\int_{\gamma x_0}^x n ds$.

From Fermat's principle we easily derive the law of refraction as well as the law of reflection.

Fermat's principle has the following geometrical interpretation. Apart from the natural-length element

$$ds^2 = \dot{x}_i \dot{x}_i d\tau^2, \quad (7.36)$$

we define on $\mathbb{R}^3$ a new light-length element:

$$dl^2 = n^2(\mathbf{x}) \dot{x}_i \dot{x}_i d\tau^2 = n^2 ds^2. \quad (7.37)$$

With respect to the natural-length element, the arc length of a curve $\mathbf{x}(t)$ is given by

$$s_{21} = \int ds = \int_{t_1}^{t_2} \sqrt{\dot{\mathbf{x}}^2} d\tau. \quad (7.38)$$

However, expressed in terms of the light-length element, we find

$$l_{21} = \int dl = \int_{t_1}^{t_2} n(\mathbf{x}(\tau)) \sqrt{\dot{\mathbf{x}}^2} d\tau. \quad (7.39)$$

Geodesics are the shortest connections between two points. With respect to the natural-length element these are just straight lines. However, with respect to the light-length element, the geodesics are light rays. Due to the relation $n(x) = c/c(x)$, the length

$$l = c \int_{\gamma} \frac{ds}{c(x)} = cT \quad (7.40)$$

is proportional to the time that a light ray needs to travel along the path γ . Since the frequency ω is fixed, T is proportional to the number of oscillations as well as to the sum of the local wavelengths along γ . Therefore, *Fermat's principle* is sometimes stated in the following way:

A light ray runs from x_1 to x_2 along the path that is traversed *in the least time*.

Fermat's principle is a further example of the relevance of variational principles in physics. The analogy between geometrical optics and mechanics is based on the fact that both theories can be formulated in terms of variational principles.

7.6 Analogy between mechanics and geometrical optics

There exists a close analogy between the eikonal equation in optics and the Hamilton–Jacobi equation in mechanics, as well as between light rays and the trajectories in a mechanical system. Just as a reminder for the reader, we briefly describe the Hamilton–Jacobi theory for a mechanical system with the following Lagrange function:

$$L(\mathbf{x}, \dot{\mathbf{x}}) = \frac{1}{2} m_{ik}(\mathbf{x}) \dot{x}_i \dot{x}_k - V(\mathbf{x}). \quad (7.41)$$

The canonical momenta are

$$p_i = \frac{\partial L}{\partial \dot{x}_i} = m_{ik} \dot{x}_k, \quad (7.42)$$

and the Hamiltonian,

$$H(\mathbf{x}, \mathbf{p}) = p_i \dot{x}_i(\mathbf{x}, \mathbf{p}) - L(\mathbf{x}, \dot{\mathbf{x}}(\mathbf{x}, \mathbf{p})),$$

is given by

$$H(\mathbf{x}, \mathbf{p}) = \frac{1}{2} m_{ik}^{-1} p_i p_k + V. \quad (7.43)$$

Here, m_{ik}^{-1} denotes the elements of the inverse matrix of (m_{ij}) . *Hamilton's equations of motion*, namely

$$\frac{\partial H}{\partial p_i} = \dot{x}_i, \quad \frac{\partial H}{\partial x_i} = -\dot{p}_i, \quad (7.44)$$

are equivalent to Lagrange's equation of motion,

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{x}}} - \frac{\partial L}{\partial \mathbf{x}} = 0.$$

The Hamilton–Jacobi equation is the following partial differential equation:

$$H(\mathbf{x}, \nabla S) = \frac{1}{2} m_{ik}^{-1} \partial_i S \partial_k S + V = E. \quad (7.45)$$

If S is a solution of the Hamilton–Jacobi equation, we obtain the trajectories $\mathbf{x}(t)$ of the system as the orthogonal trajectories of the level surfaces of S in the following way. The momentum $p_i(t) = m_{ij} \dot{x}_j(t)$ is given by

$$p_i(t) = \partial_i S(\mathbf{x}(t)). \quad (7.46)$$

So in general it is not $\dot{x}_i$ but p_i that is orthogonal to the level surfaces of S . (If $m_{ij} = m\delta_{ij}$, as in most cases, $\dot{\mathbf{x}}$ and $\mathbf{p}$ are parallel.) As a proof of this statement we first verify

$$\dot{x}_i = \frac{\partial H}{\partial p_i} = m_{ij}^{-1} p_j, \quad \text{i.e. indeed} \quad p_i = m_{ij} \dot{x}_j, \quad (7.47)$$

and then calculate

$$\begin{aligned} \frac{dp_i}{dt} &= \frac{d}{dt} (m_{ij} \dot{x}_j) = \frac{d}{dt} \partial_i S = \dot{x}_k \partial_k \partial_i S = m_{kr}^{-1} \partial_r S \partial_k \partial_i S \\ &= \frac{1}{2} \partial_i (m_{kr}^{-1} \partial_k S \partial_r S) - \frac{1}{2} (\partial_i m_{kr}^{-1}) \partial_k S \partial_r S \\ &= -\partial_i V - \frac{1}{2} (\partial_i m_{kr}^{-1}) \partial_k S \partial_r S \\ &= -(\partial_i H)(\mathbf{x}, \mathbf{p}). \end{aligned} \quad (7.48)$$

So, the trajectories orthogonal to the level surfaces of S are, in the way described above, solution curves of the mechanical system. In chapter 9, we will see in a more general context that also the reverse statement is true: all solution curves are orthogonal trajectories to the level surfaces of S . The function S is closely related to the action $\int dt L(\mathbf{x}, \dot{\mathbf{x}})$, if one chooses for $\mathbf{x}(t)$ a solution of the equation of motion. Indeed,

$$\begin{aligned} \int_{t_0}^t dt' L(\mathbf{x}, \dot{\mathbf{x}}) &= \int_{t_0}^t dt' [\mathbf{p} \cdot \dot{\mathbf{x}} - H(\mathbf{x}, \mathbf{p})] \\ &= \int \mathbf{p} \cdot d\mathbf{x} - E(t - t_0) = S(\mathbf{x}) - S(\mathbf{x}_0) - E(t - t'). \end{aligned} \quad (7.49)$$

The fundamental idea of using the Hamilton–Jacobi method in order to find solutions of the equation of motion consists in determining sufficiently many solutions of the Hamilton–Jacobi equation and then calculating *all* orbital curves as orthogonal trajectories. We now write the eikonal equation $(\nabla S)^2 = n^2$ in one of the following equivalent forms:

$$\begin{aligned} \text{(A)} \quad & \frac{1}{2} (\nabla S)^2 - \frac{1}{2} n^2 = 0, \\ \text{(B)} \quad & \frac{1}{2n} (\nabla S)^2 - \frac{1}{2} n = 0, \\ \text{(C)} \quad & \frac{1}{2n^2} (\nabla S)^2 - \frac{1}{2} = 0. \end{aligned} \quad (7.50)$$

In all three cases we can identify the equation as the Hamilton–Jacobi equation of an appropriate mechanical system, and the three equations differ merely by their choice of time

coordinates. In addition, we find for all three cases $E = 0$ such that S is the action functional. (A different version of the analogy between optics and mechanics uses the so-called extended phase space where t and $(-E)$ are introduced as additional position and momentum coordinates.) Table 7.1 itemizes the analogy in more detail.

Table 7.1: Analogy of light rays and orbital curves of mechanical systems.

(A)	(B)	(C)
$m = 1, \quad V = -\frac{1}{2}n^2$	$m = n, \quad V = -\frac{1}{2}n$	$m = n^2, \quad V = -\frac{1}{2}$
$L = \frac{1}{2}\dot{\mathbf{x}}^2 + \frac{1}{2}n^2$	$L = \frac{1}{2}n\dot{\mathbf{x}}^2 + \frac{1}{2}n$	$L = \frac{1}{2}n^2\dot{\mathbf{x}}^2 + \frac{1}{2}$
$H = \frac{\mathbf{p}^2}{2} - \frac{n^2}{2}$	$H = \frac{\mathbf{p}^2}{2n} - \frac{n}{2}$	$H = \frac{\mathbf{p}^2}{2n^2} - \frac{1}{2}$
$\frac{d\mathbf{x}}{d\tau} = \dot{\mathbf{x}} = \mathbf{p} = \nabla S$	$\frac{d\mathbf{x}}{ds} = \frac{1}{n} \mathbf{p} = \dot{\mathbf{x}}$	$\frac{d\mathbf{x}}{dl} = \dot{\mathbf{x}} = \frac{\mathbf{p}}{n^2}$
$ \dot{\mathbf{x}} = n$	$ \dot{\mathbf{x}} = \left \frac{1}{n} \nabla S \right = 1$	$ \dot{\mathbf{x}} = \frac{1}{n}$
$E = \frac{1}{2}\dot{\mathbf{x}}^2 - \frac{1}{2}n^2 = 0$	$E = \frac{1}{2}n\dot{\mathbf{x}}^2 - \frac{1}{2}n = 0$	$E = \frac{1}{2}n^2\dot{\mathbf{x}}^2 - \frac{1}{2} = 0$
$\frac{d^2\mathbf{x}}{d\tau^2} = \frac{1}{2}\nabla n^2$	$\frac{d}{ds} \left(n \frac{d\mathbf{x}}{ds} \right) = \nabla n$	$\frac{d}{dl} \left(n^2 \frac{d\mathbf{x}}{dl} \right) = \frac{1}{2}\dot{\mathbf{x}}^2 \nabla n^2$

In all three cases, the value of $|\dot{\mathbf{x}}|$ follows either from $\mathbf{p} = \nabla S$ and the eikonal equation, or from $E = 0$, and the law of refraction follows from the conservation of the momentum component parallel to the interface between the media.

The action S and the momentum $\mathbf{p}$ are the same in all cases. Therefore, we obtain the same light rays with different parametrizations, and the relations between the parameters τ , s , and l are:

$$\frac{d\tau}{ds} = \frac{1}{n}, \quad \frac{dl}{ds} = n. \quad (7.51)$$

In addition, we should make the following comments on the systems (A), (B), and (C):

- (A) This system has the closest similarity to a normal mechanical system with a space-independent mass, and it was the basis for the particle theory of light. It turns out that the particle velocity is proportional to n , i.e. it is larger in a denser medium. However, the observed propagation velocity of light is proportional to $1/n$, which contradicts the concept of a particle.
- (B) In this case, the curve parameter is the natural arc length in ordinary space. The equation of motion is identical to the equation we derived for light rays in section 7.4.
- (C) Now, the curve parameter $l = \int n ds$ corresponds to the “light length”, i.e. the length with respect to the metric which occurs in Fermat’s principle and which is proportional

to the running time of light. This metric differs from the natural flat metric ds^2 by a conformal factor n^2 , which changes the distances but not the angles; for this reason it is called *conformally flat*. With l as the curve parameter, $d\mathbf{x}/dl$ is equal to the phase velocity of light. However, the price to pay is a position-dependent “mass” $m = n^2(\mathbf{x})$, which makes an interpretation as a normal mechanical system in the sense of a corpuscular theory difficult.

For none of the systems (A)–(C) is Fermat’s principle equivalent to the corresponding mechanical variational principle. However, the equation of motion for system (B),

$$\frac{d}{ds} \left(n \frac{d\mathbf{x}}{ds} \right) = \nabla n,$$

is just the equation for geodesics with respect to a light metric $n^2 ds^2 = dl^2$ and with s as curve parameter. Indeed, from

$$\delta \int n ds = \delta \int n \sqrt{\left(\frac{d\mathbf{x}}{dt} \right)^2} dt = 0$$

we obtain

$$\begin{aligned} 0 &= \int \left[\nabla n \sqrt{\left(\frac{d\mathbf{x}}{dt} \right)^2} \cdot \delta \mathbf{x} + n \frac{d\mathbf{x}/dt}{\sqrt{(d\mathbf{x}/dt)^2}} \cdot \frac{d \delta \mathbf{x}}{dt} \right] dt \\ &= \int \left[\nabla n \cdot \delta \mathbf{x} + n \frac{d\mathbf{x}}{ds} \cdot \frac{d \delta \mathbf{x}}{ds} \right] ds = \int ds \left[\nabla n - \frac{d}{ds} \left(n \frac{d\mathbf{x}}{ds} \right) \right] \delta \mathbf{x} = 0. \end{aligned}$$

Fermat’s principle,

$$\delta \int n \sqrt{\left(\frac{d\mathbf{x}}{dt} \right)^2} dt = 0,$$

has the advantage that it is completely independent of the parametrization t . Of course, for geodesics one finds again that

$$\int_{x_0}^x n ds = S(\mathbf{x}) - S(\mathbf{x}_0). \quad (7.52)$$

The mechanical analogy of Fermat’s principle in optics is the *principle of Maupertuis*: The orbits for fixed energy

$$E = \frac{1}{2} m_{ij}(\mathbf{x}) \dot{x}_i \dot{x}_j + V$$

follow from the variational problem:

$$\delta \int \sqrt{(E - V) m_{ik} dx^i dx^k} = 0. \quad (7.53)$$

For the determination of light rays we have to set $E = 0$, due to the eikonal equation.

8 Geometrical optics

8.1 Fermat's principle and focal points

We have seen that in the framework of Fermat's principle we may, independently of the parametrization, describe light rays connecting a point $\mathbf{x}_1$ with a point $\mathbf{x}_2$ as geodesics for the conformally flat light metric $dl^2 = n^2 ds^2$, i.e. as the curves $\mathbf{x}(t)$ with $\mathbf{x}(t_1) = \mathbf{x}_1$, $\mathbf{x}(t_2) = \mathbf{x}_2$, for which the length

$$l_{12} = \int_{\mathbf{x}_1}^{\mathbf{x}_2} n ds = \int_{s_1}^{s_2} n(\mathbf{x}(s)) ds = \int_{t_1}^{t_2} n(\mathbf{x}(t)) \sqrt{\left(\frac{d\mathbf{x}}{dt}\right)^2} dt \quad (8.1)$$

is stationary. If the two points $\mathbf{x}_1$ and $\mathbf{x}_2$ are sufficiently close, there exists exactly one shortest light ray connecting $\mathbf{x}_1$ and $\mathbf{x}_2$, and the stationary value

$$l_{21}^{(\min)} = S(\mathbf{x}_2, \mathbf{x}_1) \quad (8.2)$$

corresponding to this geodesic is a minimum. Up to a factor k , $S(\mathbf{x}_2, \mathbf{x}_1)$ is the phase difference between point $\mathbf{x}_1$ and point $\mathbf{x}_2$, and it may also be interpreted as the variational functional of a mechanical system. The minimal distance $S(\mathbf{x}_2, \mathbf{x}_1)$ between $\mathbf{x}_2$ and $\mathbf{x}_1$ with respect to the light metric is proportional to the minimal time for the light to get from $\mathbf{x}_1$ to $\mathbf{x}_2$.

For larger distances between $\mathbf{x}_1$ and $\mathbf{x}_2$, there may be several light rays of equal length $S(\mathbf{x}_2, \mathbf{x}_1)$ that connect the points $\mathbf{x}_1$ and $\mathbf{x}_2$. If there exists an r -parameter family of light rays of equal length connecting points $\mathbf{x}_1$ and $\mathbf{x}_2$, the points $\mathbf{x}_1$ and $\mathbf{x}_2$ are called *focal* to each other of r th order. In particular, $\mathbf{x}_1$ and $\mathbf{x}_2$ are called *conjugated* if they are focal of second order. In this case there exists a cone with tip point $\mathbf{x}_1$, which has a non-vanishing cone angle such that all rays through $\mathbf{x}_1$ and within this cone meet again in $\mathbf{x}_2$ (see fig. 8.1). Hence, conjugated points $\mathbf{x}_1$ and $\mathbf{x}_2$ are mapped onto each other by the light rays.

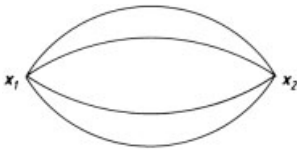


Figure 8.1: Two points are called conjugated if all the rays emitted from one point within a cone of non-vanishing cone angle meet again at the opposite point.

We obtain an *image* in an optical instrument when all rays starting from special points $\mathbf{x}_i$ meet again at the points conjugated to $\mathbf{x}_i$.

If a light ray on its way from a point x_1 to a point x_2 passes through a point x'_1 that is focal to x_1 , then this light ray is no longer a minimal geodesic from x_1 to x_2 . Indeed, as fig. 8.2 shows, there exist shorter connections between x_1 and x_2 , like, for example, the curve $x_1 \rightarrow x_3 \rightarrow x_4 \rightarrow x_2$. To construct such a shorter connection, one only has to choose points x_3 and x_4 close enough to each other, such that the light ray from x_3 to x_4 is the unique minimal geodesic from x_3 to x_4 .



Figure 8.2: A light ray passing through a focal point is not a minimal geodesic.

8.2 Perfect optical instruments

We now will investigate the possibility of constructing a perfect optical instrument, where, for *each* point x , all rays leaving x meet again at a point x' conjugate to x . We shall assume that x' is the first focal point of x along each ray on its way from x to x' ; this implies that the rays running from x to x' have minimal length. We will see that the construction of perfect optical instruments is subject to strict limitations.

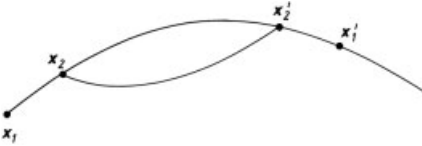


Figure 8.3: Imagery in a perfect optical instrument: x'_2 should not lie between x_2 and x'_1 .

Let us consider a minimal light ray running from x_1 to the first conjugated point x'_1 . Between x_1 and x'_1 there will be a second point x_2 on this light ray (see fig. 8.3). In a perfect optical instrument, the light rays leaving from x_2 meet at the first point x'_2 , which is conjugated to x_2 . Obviously, x'_2 also lies on the light ray passing through x_1 , x_2 , and x'_1 . However, this point x'_2 is located beyond the point x'_1 , and not between x_2 and x'_1 . Indeed, if x'_2 was between x_2 and x'_1 , the geodesic from x_1 to x'_1 would not be minimal, as we have just seen, which contradicts our assumption.

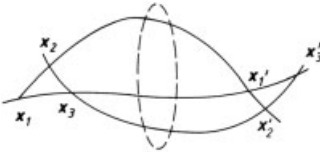


Figure 8.4: Imagery of three points in a perfect optical instrument.

Let us now look at the image of three points x_1 , x_2 , and x_3 in a perfect optical instrument. Figure 8.4 shows that

$$\begin{aligned} S(x'_2, x_2) + S(x_2, x_1) &= S(x'_2, x'_1) + S(x'_1, x_1), \\ S(x'_3, x_3) + S(x_3, x_2) &= S(x'_3, x'_2) + S(x'_2, x_2). \end{aligned}$$

By taking the sum we obtain

$$S(\mathbf{x}_3, \mathbf{x}_2) + S(\mathbf{x}_2, \mathbf{x}_1) = S(\mathbf{x}'_3, \mathbf{x}'_2) + S(\mathbf{x}'_2, \mathbf{x}'_1) + S(\mathbf{x}'_1, \mathbf{x}_1) - S(\mathbf{x}'_3, \mathbf{x}_3). \quad (8.3)$$

As a next step, we subdivide the triangle $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ into a polygon by inserting more and more points, and take the limit to a smooth curve. This leads us to the following statement: For a perfect optical instrument, let Φ be the mapping that associates to each point $\mathbf{x}$ its first conjugated point $\mathbf{x}' = \Phi(\mathbf{x})$. Furthermore, let $\gamma = \mathbf{x}(t)$ be a curve from $\mathbf{x}_1$ to $\mathbf{x}_2$ and $\gamma' = \Phi(\mathbf{x}(t))$ be its image under Φ . With this notation, the lengths $S[\gamma]$ and $S[\gamma']$ of the curves γ and γ' , respectively, as measured in the light metric, satisfy

$$S[\gamma'] = S[\gamma] + S(\mathbf{x}'_2, \mathbf{x}_2) - S(\mathbf{x}'_1, \mathbf{x}_1). \quad (8.4)$$

Hence, the difference $S[\gamma'] - S[\gamma]$ depends only on the initial and final points γ and γ' . Because of eq. (8.1) we have

$$\begin{aligned} S[\gamma'] - S[\gamma] &= \int_{t_1}^{t_2} dt \left[n(\Phi(\mathbf{x}(t))) \sqrt{\left(\frac{d\Phi(\mathbf{x}(t))}{dt} \right)^2} - n(\mathbf{x}(t)) \sqrt{\left(\frac{d\mathbf{x}(t)}{dt} \right)^2} \right] \\ &\equiv \int_{t_1}^{t_2} dt \Psi \left(\mathbf{x}(t), \frac{d\mathbf{x}}{dt} \right) \\ &= S(\Phi(\mathbf{x}(t_2)), \mathbf{x}(t_2)) - S(\Phi(\mathbf{x}(t_1)), \mathbf{x}(t_1)). \end{aligned} \quad (8.5)$$

Here, the integrand, $\Psi(\mathbf{x}, d\mathbf{x}/dt) = \Psi(\mathbf{x}, -d\mathbf{x}/dt)$, is a homogeneous function of degree one in the argument $d\mathbf{x}/dt$. Since the integral over Ψ depends only on the initial and final points of the curve, the integrand is an exact differential:

$$\Psi \left(\mathbf{x}(t), \frac{d\mathbf{x}}{dt} \right) = \frac{d\mathbf{x}}{dt} \cdot \nabla F(\mathbf{x}(t)).$$

Because $\Psi(\mathbf{x}, d\mathbf{x}/dt) = +\Psi(\mathbf{x}, -d\mathbf{x}/dt)$, we find $\nabla F \equiv 0$. We just proved the following statement:

For a perfect optical instrument, the optical length of each curve γ is equal to the length of its image $\Phi(\gamma)$. The optical distance $S(\mathbf{x}, \mathbf{x}')$ between a point $\mathbf{x}$ and its image point $\mathbf{x}' = \Phi(\mathbf{x})$ is independent of $\mathbf{x}$.

The proof outlined above is due to Carathéodory; it remains true even if we require the less stringent condition that the perfect optical instrument shall map all points inside a small open subset U onto an open subset $\Phi(U)$ by ray bundles with a non-vanishing opening angle.

If the refractive indices in U and $\Phi(U)$ are constant and equal, the mapping Φ can only be a congruence, i.e. a combination of rotations, mirror reflections, and translations, without any change of size. Such a mapping is easily realized by a simple system of flat mirrors.

8.3 Maxwell's fish-eye

The n -dimensional sphere $S^n \subset \mathbb{R}^{n+1}$ is defined to be the set

$$S^n = \left\{ (\xi_1, \dots, \xi_n, \zeta) \left| \sum_{i=1}^n \xi_i^2 + \zeta^2 = 1 \right. \right\}. \quad (8.6)$$

On this set we define the metric

$$ds^2 = \sum_{i=1}^n (d\xi_i)^2 + [d\zeta(\xi)]^2 = \sum_{i,j} \left(\delta_{ij} + \frac{\xi_i \xi_j}{1 - \xi^2} \right) d\xi_i d\xi_j, \quad (8.7)$$

where $\xi^2 = \sum_{i=1}^n \xi_i^2$. This metric is induced by the natural metric on $\mathbb{R}^{n+1}$, restricted to S^n . Geodesics on S^n are the great circles on S^n , i.e. the intersections of S^n with hyperplanes through the center of the sphere. All great circles passing through one point P of the sphere intersect again in P', the antipodal point of P. This fact can be utilized for the construction of a perfect optical instrument, as follows.

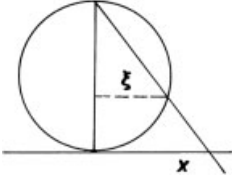


Figure 8.5: Stereographic projection of a sphere.

We define the stereographic projection of a sphere S^3 (excluding the north pole) onto $\mathbb{R}^3$ by the following mapping (see fig. 8.5):

$$\xi = \frac{x}{1 + \frac{1}{4}x^2}, \quad \zeta = \frac{1 - \frac{1}{4}x^2}{1 + \frac{1}{4}x^2}. \quad (8.8)$$

A simple calculation verifies that the stereographic projection is a conformal mapping. Indeed,

$$ds^2 = \sum_{i=1}^n (d\xi_i)^2 + [d\zeta(\xi)]^2 = \frac{(dx)^2}{(1 + \frac{1}{4}x^2)^2}. \quad (8.9)$$

The mapping (8.8) maps the sphere S^n with metric (8.7) (excluding the north pole) onto $\mathbb{R}^n$ with metric (8.9). Geodesics are mapped into geodesics. In particular, one property of the sphere is preserved under this mapping: all geodesics through one point meet in a conjugated point.

Therefore, if (by an additional trivial scale transformation) we choose the spatial dependence of the refractive index $n(x)$ to be of the form:

$$n(x) = \frac{n_0}{1 + x^2/a^2},$$

we obtain a perfect optical instrument called *Maxwell's fish-eye*. It was discovered by J. C. Maxwell in 1858 in a different context. The light rays in Maxwell's fish-eye are the images of great circles under a stereographic projection and, therefore, are also circular. (Only the great circles passing through the north and south poles correspond to straight lines in $\mathbb{R}^3$.) As for $|x| \rightarrow \infty$ the refractive index $n(x) = n_0/[1 + (x^2/a^2)]$ tends to zero, a physical realization of Maxwell's fish-eye is possible only in parts.

In general, a manifold with the property that for each point P all geodesics passing through P meet again at the first conjugated point P' is called a *Blaschke manifold* or a *Wiedersehen manifold* (literally a “meet again” manifold). According to a conjecture of Blaschke, any n -dimensional Wiedersehen manifold is isometric to S^n . A proof of this conjecture exists only for $n = 2$.

One can show that all conformal transformations of $\mathbb{R}^3$ can be generated from rotations, mirror reflections, translations, and inversions at the unit sphere. Therefore, Maxwell’s fish-eye is, apart from trivial systems consisting only of plane mirrors and assuming Blaschke’s conjecture to be right, the only perfect optical instrument.

8.4 Canonical transformations and eikonal functions

As is well known, the time evolution in a Hamiltonian system is given by a *canonical transformation*, which means the following. Suppose we are given a phase space

$$P = \{(\mathbf{x}, \mathbf{p}) \mid \mathbf{x}, \mathbf{p} \in \mathbb{R}^3\}, \quad (8.10)$$

on which we consider the 2-form:

$$\omega = d\mathbf{x} \wedge d\mathbf{p} = \sum_i dx^i \wedge dp_i = -d \sum_i p_i dx^i = d \sum_i x^i dp_i. \quad (8.11)$$

Then ω is closed: $d\omega = 0$. Solving Hamilton’s equations of motion for the initial data $(\mathbf{x}_0, \mathbf{p}_0)$, we obtain the following paths in phase space P :

$$\Phi_{(\mathbf{x}_0, \mathbf{p}_0)}(t) = (\mathbf{x}(t), \mathbf{p}(t)) \quad \text{with} \quad (\mathbf{x}_0, \mathbf{p}_0) = \Phi_{(\mathbf{x}_0, \mathbf{p}_0)}(0). \quad (8.12)$$

They define a family of mappings

$$\Phi_t : P \mapsto P, \quad (\mathbf{x}_0, \mathbf{p}_0) \mapsto \Phi_t(\mathbf{x}_0, \mathbf{p}_0) = \Phi_{(\mathbf{x}_0, \mathbf{p}_0)}(t) = (\mathbf{x}_t, \mathbf{p}_t), \quad (8.13)$$

and we have

$$d\mathbf{x}_0 \wedge d\mathbf{p}_0 = d\mathbf{x}_t \wedge d\mathbf{p}_t. \quad (8.14)$$

Equation (8.14) is just the defining property of a canonical transformation Φ_t . [In general, a manifold M with a non-degenerate 2-form ω with $d\omega = 0$ is called a *symplectic manifold*. According to a theorem of *Darboux*, we can define on any symplectic manifold M local coordinates (x_i, p_i) such that $\omega = \sum dx^i \wedge dp_i$. See chapter 9 for more details about symplectic manifolds.]

In section 7.6 we have seen that light rays can be identified as the orbits with energy $E = 0$ of suitable mechanical systems. The Lagrange function of the mechanical systems can be any of the three functions listed in table 7.1:

$$L_1 = \frac{1}{2}\dot{\mathbf{x}}^2 + \frac{1}{2}n^2, \quad L_2 = \frac{1}{2}n\dot{\mathbf{x}}^2 + \frac{1}{2}n, \quad L_3 = \frac{1}{2}n^2\dot{\mathbf{x}}^2 + \frac{1}{2}, \quad (8.15)$$

and, because of Fermat’s principle (eq. 8.1), we may also take

$$L_4 = n\sqrt{\dot{\mathbf{x}}^2}.$$

Therefore, to any system of geometrical optics we can assign a phase space $P = \{(\mathbf{x}, \mathbf{p})\}$. In all cases we have

$$\mathbf{p} = \frac{\partial L}{\partial \dot{\mathbf{x}}} = n\boldsymbol{\sigma}, \quad (8.16)$$

where $\boldsymbol{\sigma}$ is a unit vector along the direction of the light ray.

For an arbitrary parameter t along the light ray, the mapping

$$\Phi_t : (\mathbf{x}_0, \mathbf{p}_0) \longmapsto \Phi_t(\mathbf{x}_0, \mathbf{p}_0) = (\mathbf{x}_t, \mathbf{p}_t)$$

is a canonical transformation.

A canonical transformation can always be constructed from a generating function. In order to see how to get such a function, we consider the following 2-form on the Cartesian product $P \times P$ of the phase space with itself:

$$\Omega = d\mathbf{x}_0 \wedge d\mathbf{p}_0 - d\mathbf{x} \wedge d\mathbf{p}. \quad (8.17)$$

If $(\mathbf{x}, \mathbf{p})$ is obtained from $(\mathbf{x}_0, \mathbf{p}_0)$ by a canonical transformation Φ , we find $\Omega = 0$. Moreover, given a transformation $\Phi : P \rightarrow P$, we consider the *graph* of Φ :

$$\text{graph } \Phi = \{(\mathbf{x}_0, \mathbf{p}_0), \Phi(\mathbf{x}_0, \mathbf{p}_0)\} \subset P \times P. \quad (8.18)$$

Φ is a canonical transformation if and only if Ω vanishes on the submanifold $\text{graph } \Phi$ of $P \times P$.

We now assume that it is possible to introduce $(\mathbf{x}_0, \mathbf{x})$ as local coordinates on $\text{graph } \Phi$. This will always be the case for our geometrooptical system, when a light ray traveling from $\mathbf{x}_0$ to $\mathbf{x}$ is uniquely determined by its initial and final points such that, in particular, $\mathbf{p}_0$ and $\mathbf{p}$ are already determined by $\mathbf{x}_0$ and $\mathbf{x}$. This condition is fulfilled as long as $\mathbf{x}$ and $\mathbf{x}_0$ are not focal with respect to each other. Written in the coordinates $(\mathbf{x}_0, \mathbf{x})$, we find for $\text{graph } \Phi$:

$$\Omega = d\mathbf{x}_0 \wedge d\mathbf{p}_0 - d\mathbf{x} \wedge d\mathbf{p} = d(\mathbf{p}(\mathbf{x}, \mathbf{x}_0) \cdot d\mathbf{x} - \mathbf{p}_0(\mathbf{x}, \mathbf{x}_0) \cdot d\mathbf{x}_0) = 0. \quad (8.19)$$

Therefore, there exists, at least locally, a function $S(\mathbf{x}, \mathbf{x}_0)$ such that

$$\mathbf{p} \cdot d\mathbf{x} - \mathbf{p}_0 \cdot d\mathbf{x}_0 = dS(\mathbf{x}, \mathbf{x}_0), \quad (8.20)$$

i.e.

$$\mathbf{p} = \frac{\partial S}{\partial \mathbf{x}}, \quad \mathbf{p}_0 = -\frac{\partial S}{\partial \mathbf{x}_0}. \quad (8.21)$$

Indeed, for the cases of classical mechanics and geometrical optics, we already know such a function $S(\mathbf{x}, \mathbf{x}_0)$: it is the action functional or eikonal function $S(\mathbf{x}, \mathbf{x}_0)$ introduced in chapter 7, which gives us the optical distance between the points $\mathbf{x}_0$ and $\mathbf{x}$. This function $S(\mathbf{x}, \mathbf{x}_0)$ is also called the *point characteristic* or *point eikonal*, and it is a solution of the eikonal equation $(\nabla_{\mathbf{x}} S)^2 = n^2$.

When x_0 and x are conjugate points, it is more appropriate to introduce the initial position x_0 and the final momentum p as local coordinates on graph Φ . Indeed, in this case, a light ray should be uniquely determined by x_0 and p . Now,

$$0 = dx_0 \wedge dp_0 - dx \wedge dp = d(-x(p, x_0) \cdot dp - p_0(p, x_0) \cdot dx_0),$$

and there exists a function $M(p, x_0)$, the so-called *mixed characteristic*, for which

$$dM(p, x_0) = -x \cdot dp - p_0 \cdot dx_0, \quad \text{i.e.} \quad x = -\frac{\partial M}{\partial p}, \quad p_0 = -\frac{\partial M}{\partial x_0}. \quad (8.22)$$

The function $M(p, x_0)$ follows from the point characteristic $S(x, x_0)$ by a Legendre transformation from the variable x to the variable p :

$$\begin{aligned} M(p, x_0) &= S(x(p, x_0), x_0) - p \cdot x(p, x_0), \\ \frac{\partial M}{\partial x_0} &= \frac{\partial S}{\partial x_0} + \frac{\partial S}{\partial x} \cdot \frac{\partial x}{\partial x_0} - p \cdot \frac{\partial x}{\partial x_0} = \frac{\partial S}{\partial x_0} = -p_0, \\ \frac{\partial M}{\partial p} &= \frac{\partial S}{\partial x} \cdot \frac{\partial x}{\partial p} - x - p \frac{\partial x}{\partial p} = -x. \end{aligned} \quad (8.23)$$

If, in an environment of x and x_0 , the light rays are straight lines, one obtains the mixed characteristics from the point characteristics by a simple geometrical construction: $M(p, x_0) = S(x', x_0)$, where x' is constructed from x by going back along the ray with momentum p to the point where $p \cdot x' = 0$ (see fig. 8.6).



Figure 8.6: Construction of the mixed characteristics from the point characteristics.

Introducing p_0 and p as coordinates on graph Φ , we obtain the *angle characteristic* or the *angle eikonal*:

$$T(p, p_0) = S(x_0(p, p_0), x(p, p_0)) - p \cdot x(p, p_0) + p_0 \cdot x_0(p, p_0). \quad (8.24)$$

We find that

$$\frac{\partial T}{\partial p} = -x, \quad \frac{\partial T}{\partial p_0} = x_0. \quad (8.25)$$

One obtains additional eikonal functions by Legendre transformations with respect to arbitrary coordinates x_i, p_i or x_{0i}, p_{0i} .

On the other hand, from the procedure just described, it should be obvious that for a given eikonal function F one can reconstruct a canonical transformation Φ such that the variables of F are coordinates on graph Φ . For instance, from

$$\mathbf{p} = \frac{\partial S(\mathbf{x}, \mathbf{x}_0)}{\partial \mathbf{x}}, \quad \mathbf{p}_0 = -\frac{\partial S(\mathbf{x}, \mathbf{x}_0)}{\partial \mathbf{x}_0},$$

solving for $\mathbf{p}$ and $\mathbf{x}$, we obtain a canonical transformation

$$(\mathbf{x}, \mathbf{p}) = \Phi(\mathbf{x}_0, \mathbf{p}_0).$$

If the generating functions of the canonical transformations

$$\Phi : (\mathbf{x}_1, \mathbf{p}_1) \longmapsto (\mathbf{x}_2, \mathbf{p}_2),$$

$$\Psi : (\mathbf{x}_2, \mathbf{p}_2) \longmapsto (\mathbf{x}_3, \mathbf{p}_3)$$

are known, we can easily determine the generating functions of the combined transformation:

$$\Psi \circ \Phi : (\mathbf{x}_1, \mathbf{p}_1) \longmapsto (\mathbf{x}_3, \mathbf{p}_3).$$

For point characteristics one obtains

$$S_{31}(\mathbf{x}_3, \mathbf{x}_1) = S_{32}(\mathbf{x}_3, \mathbf{x}_2) + S_{21}(\mathbf{x}_2, \mathbf{x}_1), \quad (8.26)$$

where $\mathbf{x}_2$ has to be determined from the condition

$$\frac{\partial S_{32}}{\partial \mathbf{x}_2} + \frac{\partial S_{21}}{\partial \mathbf{x}_2} = 0. \quad (8.27)$$

This is a consequence of Fermat's principle. It means that a ray which, on its way from $\mathbf{x}_1$ to $\mathbf{x}_3$, passes through $\mathbf{x}_2$ will not have a kink at $\mathbf{x}_2$.

For mixed characteristics one finds

$$M_{31}(\mathbf{x}_3, \mathbf{p}_1) = M_{32}(\mathbf{x}_3, \mathbf{p}_2) + T_{21}(\mathbf{p}_2, \mathbf{p}_1), \quad (8.28)$$

with

$$\frac{\partial M_{32}}{\partial \mathbf{p}_2} + \frac{\partial T_{21}}{\partial \mathbf{p}_2} = 0. \quad (8.29)$$

This condition implies that a ray is continuous.

8.5 Imaging points close to the optic axis by wide spread ray bundles

We have seen that no “interesting” perfect optical instruments exist. Therefore, we now want to investigate the possibility of imaging at least certain curves or planes using optical instruments.

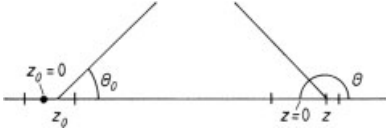


Figure 8.7: Imaging some segment I_0 of the optic axis near $z_0 = 0$ onto some other segment I of the axis near $z = 0$.

In this context we will mainly deal with the following situation, which is of greatest practical relevance: there should exist an *optic axis*, i.e. a straight light ray such that the optical arrangements are rotationally symmetric with respect to this axis as well as mirror-symmetric with respect to planes containing this axis.

We first discuss the problem of imaging a segment I_0 of the optic axis around the point $z_0 = 0$ onto some other segment I of the axis around $z = 0$ (see fig. 8.7). If each ray emanating from some point in I_0 arrives back at the optic axis at a point in I , we can easily show by restricting a suitable eikonal function to the submanifold $x_0 = y_0 = x = y = 0$ that the transformation

$$(z_0, p_0) \mapsto (f(z_0, p_0), g(z_0, p_0)) = \varphi(z_0, p_0) = (z, p) \quad (8.30)$$

is canonical, i.e.

$$dz_0 \wedge dp_0 = df \wedge dg = \left(\frac{\partial f}{\partial z_0} \frac{\partial g}{\partial p_0} - \frac{\partial f}{\partial p_0} \frac{\partial g}{\partial z_0} \right) dz_0 \wedge dp_0,$$

and thus

$$\frac{\partial f}{\partial z_0} \frac{\partial g}{\partial p_0} - \frac{\partial f}{\partial p_0} \frac{\partial g}{\partial z_0} = 1. \quad (8.31)$$

Here, $p_0 = n_0 \cos \theta_0$ and $p = n \cos \theta$ are the components along the optic axis of $\mathbf{p}_0$ and $\mathbf{p}$, respectively.

Now we require that, at least up to linear terms in z and z_0 , the interval I_0 is mapped onto I , i.e. in the linearized approximation,

$$\begin{aligned} z &= f(z_0, p_0) = a(p_0)z_0 + b(p_0), \\ p &= g(z_0, p_0) = c(p_0)z_0 + d(p_0), \end{aligned} \quad (8.32)$$

the following statements hold:

- $b(p_0) \equiv 0$, because $z_0 = 0$ is mapped onto $z = 0$;
- $a(p_0) \equiv a$ independently of p_0 , because all rays passing through z_0 will meet at the same point z .

Since the image mapping is a canonical transformation (eq. 8.31), we obtain

$$\begin{aligned} a(c'(p_0)z_0 + d'(p_0)) &\equiv 1, \\ ac'(p_0) &= c'(p_0) \equiv 0, \\ d' &\equiv 1/a. \end{aligned}$$

Therefore, $c(p_0) \equiv c = \text{const.}$ and $d(p_0) = (1/a)p_0$, i.e.

$$\begin{aligned} z &= a \cdot z_0, \\ p &= c \cdot z_0 + d \cdot p_0, \\ n \cos \theta &= c \cdot z_0 + d \cdot n_0 \cos \theta_0, \end{aligned} \quad (8.33)$$

where $ad = 1$. Hence, the canonical transformation is linear. Multiplication of the equations yields

$$zp = zn \cos \theta = acz_0^2 + adz_0n_0 \cos \theta_0 = acz_0^2 + z_0n_0 \cos \theta_0,$$

which holds for all θ_0 . Because $\theta_0 = 0$ implies $\theta = 0$, we find that

$$zn = acz_0^2 + z_0n_0.$$

Subtracting and using the identity $1 - \cos \theta = 2 \sin^2(\frac{1}{2}\theta)$ we finally obtain

$$zn \sin^2(\frac{1}{2}\theta) = z_0n_0 \sin^2(\frac{1}{2}\theta_0). \quad (8.34)$$

This so-called *Herschel's condition* is a necessary condition for an image mapping from I_0 onto I , mediated by ray bundles with a non-vanishing opening angle, to exist. Herschel's condition has many solutions.

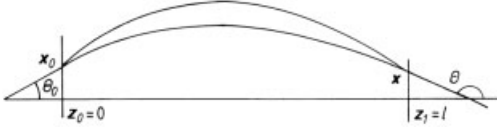


Figure 8.8: Imaging a small surface segment F_0 orthogonal to the optic axis at $z_0 = 0$ onto a surface segment F_1 at $z_1 = l$.

Next we consider image mappings from a small surface segment F_0 orthogonal to the optic axis at $z_0 = 0$ onto a surface segment F_1 at $z_1 = l$ (see fig. 8.8). Choosing z as the ray parameter we can rewrite Fermat's principle in the form:

$$\delta \int_0^l dz n \sqrt{1 + \left(\frac{dx_1}{dz}\right)^2 + \left(\frac{dx_2}{dz}\right)^2} = 0.$$

We easily see that, with the conjugated momentum

$$\frac{\partial L}{\partial \frac{dx_i}{dz}} = p_i = n \frac{dx_i/dz}{\sqrt{1 + \left(\frac{dx_1}{dz}\right)^2 + \left(\frac{dx_2}{dz}\right)^2}} = n \sin \theta_i,$$

the transformation

$$(\mathbf{x}_0, \mathbf{p}_0) \longmapsto (f(\mathbf{x}_0, \mathbf{p}_0), g(\mathbf{x}_0, \mathbf{p}_0)) \quad (8.35)$$

has to be canonical. Here $\mathbf{x}_0$ and $\mathbf{x}$ as well as $\mathbf{p}_0$ and $\mathbf{p}$ are vectors orthogonal to the optic axis. Taking into account the axial symmetry as well as the mirror symmetry, we obtain in linear approximation in $\mathbf{x}_0$:

$$\begin{aligned}\mathbf{x} &= A(|\mathbf{p}_0|)\mathbf{x}_0 + B(|\mathbf{p}_0|)\mathbf{p}_0, \\ \mathbf{p} &= C(|\mathbf{p}_0|)\mathbf{x}_0 + D(|\mathbf{p}_0|)\mathbf{p}_0,\end{aligned}$$

where $(\mathbf{x}, \mathbf{p})$ and $(\mathbf{x}_0, \mathbf{p}_0)$ are called *ray coordinates* at $z_1 = l$ and $z_0 = 0$. The requirement that, independently of $\mathbf{p}_0$, the point $\mathbf{x}_0$ is mapped onto $\mathbf{x}$ implies again that

$$A(|\mathbf{p}_0|) \equiv A = \text{const.}, \quad B(|\mathbf{p}_0|) \equiv 0,$$

and linearity follows from the canonical character of the mapping:

$$C \equiv \text{const.}, \quad D = 1/A.$$

So we find the following:

$$\begin{aligned}\mathbf{x} &= A\mathbf{x}_0, \\ \mathbf{p} &= C\mathbf{x}_0 + D\mathbf{p}_0.\end{aligned}\tag{8.36}$$

Due to the symmetry of the problem, we may choose all vectors to be in the xz plane such that

$$x = Ax_0, \quad n \sin \theta = Cx_0 + Dn_0 \sin \theta_0.$$

Multiplying the two equations together and neglecting ACx_0^2 , which is small compared to the other terms, we find *Abbe's sine condition* (discovered first in 1864 by R. Clausius and again in 1874 by H. von Helmholtz, and once more rediscovered and applied by E. Abbe in 1879),

$$xn \sin \theta = x_0 n_0 \sin \theta_0,\tag{8.37}$$

to be the necessary condition for the existence of an image mapping from F_0 onto F_1 . There are many ways for Abbe's sine condition to be satisfied. Two points z_0 and z_1 on the optic axis, for which Abbe's condition is satisfied for two surface segments F_0 in z_0 and F_1 in z_2 , are called *aplanatic points*. One can show that, apart from "trivial" perfect instruments, no more than one pair of surfaces can be mapped onto each other.

8.6 Linear geometrical optics and symplectic transformations

We consider two surfaces, F_0 and F , which are transverse to a reference ray (see fig. 8.9). A nearby second ray emanating from F_0 at $\mathbf{x}_0$ with momentum $\mathbf{p}_0$ will traverse F at point $\mathbf{x}$ with momentum $\mathbf{p}$. The transformation

$$(\mathbf{x}_0, \mathbf{p}_0) \longmapsto (\mathbf{x}, \mathbf{p}) = \Phi(\mathbf{x}_0, \mathbf{p}_0)$$



Figure 8.9: Linearization around a reference ray.

is canonical. As long as we only consider rays close to the reference ray, we can linearize the canonical transformation around x_0 and p_0 . Linear canonical transformations are called *symplectic transformations*.

It is worth while to study these transformations in general. For this purpose, we first introduce the fundamental concept of a symplectic vector space.

Definition 8.1. A symplectic vector space $(V, \mathcal{J})$ is a vector space V together with a non-degenerate antisymmetric bilinear form $\mathcal{J}$ defined on V .

In this section, V will be a symplectic vector space over $\mathbb{R}$. Notice, here and in the following, the close analogy with a Euclidean vector space on which a symmetric positive definite (and thus non-degenerate) bilinear form is defined.

Definition 8.2. Let $W \subset V$ be a subspace; then $W^\perp$ is defined by $W^\perp = \{x \mid \mathcal{J}(x, w) = 0 \text{ for all } w \in W\}$.

We find that $\dim W + \dim W^\perp = \dim V$ and $W^{\perp\perp} = W$. In general, however, we have $W \cap W^\perp \neq \{0\}$ and $W + W^\perp \neq V$, which is different from the Euclidean case.

Definition 8.3. A subspace $W \subset V$ is called symplectic if $\mathcal{J}$, restricted to W , is non-degenerate.

In this case, $W^\perp \cap W = \{0\}$ and $V = W \oplus W^\perp$, and from $W^{\perp\perp} = W$ we can deduce that also $W^\perp$ is symplectic.

Definition 8.4. A subspace $W \subset V$ is called isotropic if $\mathcal{J}$ vanishes identically on W .

This condition is equivalent to $W \subset W^\perp$.

Definition 8.5. A subspace $W \subset V$ is called co-isotropic if $W^\perp \subset W$.

Definition 8.6. A maximally isotropic subspace is called a Lagrangian subspace.

This means that $x \in W^\perp \Rightarrow x \in W$, i.e. $W^\perp \subset W$, and therefore $W = W^\perp$. In particular, we have $2 \dim W = \dim V$, and because there are always symplectic subspaces we find that symplectic vector spaces are always even-dimensional.

Definition 8.7. An isomorphism $S : (V_1, \mathcal{J}_1) \rightarrow (V_2, \mathcal{J}_2)$ is called symplectic if $\mathcal{J}_2(Sv, Sw) = \mathcal{J}_1(v, w)$ for all $v, w \in V_1$. A symplectic isomorphism $S : (V, \mathcal{J}) \rightarrow (V, \mathcal{J})$ is called a symplectic transformation. The set of symplectic transformations of V forms a group $\text{Sp}(V)$.

The linearization of a canonical transformation Φ mentioned at the beginning of this section will be the most important example of a symplectic isomorphism for us. In this case, V_1 and V_2 are the variations around (x_0, p_0) and $\Phi(x_0, p_0)$, and $\mathcal{J}_1$ and $\mathcal{J}_2$ are given by the 2-form ω in eq. (8.11) at the points (x_0, p_0) and $\Phi(x_0, p_0)$ of the phase space. Mathematically speaking, there is a symplectic isomorphism $T_m \Phi : T_m P \mapsto T_{\Phi(m)} P$ of the tangent spaces.

Theorem 8.1. *For any symplectic vector space of finite dimension, there exists a symplectic basis $(e_1, \dots, e_n, f_1, \dots, f_n)$ such that $\mathcal{J}(e_i, e_j) = \mathcal{J}(f_i, f_j) = 0$ and $\mathcal{J}(e_i, f_j) = \delta_{ij}$.*

Proof: Choose $e_1 \neq 0$ arbitrary and f_1 such that $\mathcal{J}(e_1, f_1) = 1$. Then f_1 exists because $\mathcal{J}$ is non-degenerate. Now proceed by induction. Let W be the symplectic subspace spanned by e_1 and f_1 , consider $W^\perp$, and construct e_2 and f_2 as above, etc. $\square$

Again we see that symplectic vector spaces are even-dimensional. We can easily convince ourselves that, for any subspace $W \subset V$, we can find an adapted symplectic basis $(e_1, \dots, e_n, f_1, \dots, f_n)$ such that

$$(e_1, \dots, e_r, f_1, \dots, f_s) \quad \text{where} \quad n \geq r \geq s \geq 0$$

is a basis of W . This leads to the following relations:

$$W \text{ symplectic} \iff r = s,$$

$$W \text{ isotropic} \iff s = 0,$$

$$W \text{ co-isotropic} \iff r = n,$$

$$W \text{ Lagrangian} \iff r = n, \quad s = 0.$$

With respect to a symplectic basis, $\mathcal{J}$ is represented by a matrix

$$\begin{pmatrix} 0 & \mathbf{1} \\ -\mathbf{1} & 0 \end{pmatrix},$$

where $\mathbf{1}$ is the $n \times n$ unit matrix. The theorem given above implies the isomorphy of all symplectic vector spaces of the same dimension. With respect to a basis, a symplectic transformation is represented by a matrix S that satisfies

$${}^t S \mathcal{J} S = \mathcal{J} \quad \text{or} \quad {}^t S = \mathcal{J} S^{-1} \mathcal{J}^{-1}.$$

The eigenvalues of a symplectic transformation satisfy:

$$\begin{aligned} \det(S - \lambda \mathbf{1}) &\iff \det({}^t S - \lambda \mathbf{1}) = 0, \\ &\iff \det(\mathcal{J} S^{-1} \mathcal{J}^{-1} - \lambda \mathbf{1}) = 0, \\ &\iff \det(\lambda S - \mathbf{1}) = 0, \\ &\iff \det[S - (1/\lambda) \mathbf{1}] = 0. \end{aligned}$$

So, if λ is an eigenvalue of a symplectic transformation, so are $1/\lambda$ and (because of the reality of S), $\bar{\lambda}$ and $1/\bar{\lambda}$.

The series of powers S^m of S is bounded only if all eigenvalues of S lie on the unit circle (see fig. 8.10). If all eigenvalues of S are simple and lie on the unit circle, this property also holds for all symplectic transformations S' in a sufficiently small neighborhood of S .

As is well known, orthogonal transformations R have determinant $\det R = \pm 1$. However, for a symplectic transformation S , we always find $\det S = 1$. This follows immediately from the defining property of a symplectic transformation:

$$\mathcal{J}' = {}^t S \mathcal{J} S = \mathcal{J}.$$

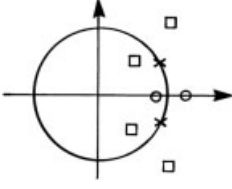


Figure 8.10: Only if all eigenvalues of S lie on the unit circle is the series of powers S^m of S bounded.

Indeed, using the $2n$ -dimensional totally antisymmetric ϵ -tensor we can define the quantity

$$D = \epsilon_{i_1 i_2 \dots i_{2n-1} i_{2n}} \mathcal{J}_{i_1 i_2} \cdots \mathcal{J}_{i_{2n-1} i_{2n}}$$

and we obtain

$$\begin{aligned} D &= D' = \epsilon_{i_1 i_2 \dots i_{2n-1} i_{2n}} S_{j_1 i_1} \cdots S_{j_{2n} i_{2n}} \mathcal{J}_{j_1 j_2} \cdots \mathcal{J}_{j_{2n-1} j_{2n}} \\ &= \epsilon_{j_1 j_2 \dots j_{2n-1} j_{2n}} \det S \mathcal{J}_{j_1 j_2} \cdots \mathcal{J}_{j_{2n-1} j_{2n}} \\ &= \det S \cdot D. \end{aligned}$$

Interpreted in a coordinate-free notation, this is just the invariance of $\mathcal{J} \wedge \cdots \wedge \mathcal{J}$. For $\dim V = 2$, the symplectic group $\text{Sp}(V)$ is equal to the group $\text{SL}(V)$ of linear mappings with determinant 1. The eigenvalues of S can be determined from the equation:

$$\lambda^2 + \lambda \text{tr } S + 1 = 0.$$

The series S^m is bounded if $\text{tr } S < 2$ and unbounded if $\text{tr } S > 2$.

8.7 Gaussian optics and image matrices

Now we consider systems with one optic axis, which are the most relevant applications of symplectic mappings. For a ray with a small inclination relative to the optic axis, we obtain a symplectic mapping that associates the coordinates of the penetration point and the inclination of the light ray for a plane F_0 perpendicular to the optic axis at the point z_0 to the coordinates of the penetration point and the inclination of this light ray at the plane F at the point z on the optic axis:

$$\begin{aligned} \mathbf{x} &= A\mathbf{x}_0 + B\mathbf{p}_0, \\ \mathbf{p} &= C\mathbf{x}_0 + D\mathbf{p}_0, \end{aligned} \tag{8.38}$$

where A , B , C , and D are linear mappings $\mathbb{R}^2 \rightarrow \mathbb{R}^2$.

Because of the rotational symmetry around the optic axis and the mirror symmetry at planes containing the optic axis, A , B , C , and D must be constants ($A = a\mathbf{1}$, $B = b\mathbf{1}, \dots$). The only constraint is $ad - bc = 1$. This approximation to geometrical optics is called *Gaussian optics*. If we assume the vectors $\mathbf{x}$, $\mathbf{x}_0$, $\mathbf{p}$, and $\mathbf{p}_0$ to be all in the xz plane, the transformation of the ray coordinates can be represented, in the linearized approximation, by a 2×2 matrix

$$T_{10} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

with $\det T_{10} = 1$:

$$\begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_0 \\ p_0 \end{pmatrix}. \quad (8.39)$$

For small inclination angles θ and θ_0 , we can write:

$$p_0 = n_0 \sin \theta_0 \approx n_0 \theta_0, \quad p = n \sin \theta \approx n \theta.$$

If we denote by T_{10} the mapping from plane F_0 to plane F_1 , and by T_{21} the mapping from F_1 to F_2 , the combined matrix $T_{20} = T_{21} \circ T_{10}$ represents the mapping from F_0 to F_2 .

If $b = 0$, we find $x = ax_0$ independently of p_0 . All rays through x_0 (with small inclination angle relative to the axis) pass through x . In this case the planes are mapped onto each other pointwise and are mutually *conjugated*. Then $a = 1/d$ is the *enlargement* of this mapping.

If $c = 0$, we have $p = dp_0$ independently of x_0 . Rays that are parallel in F_0 remain parallel in F . In this case, T_{10} is called *telescopic*, because a telescope adjusted to infinity has this property.

Two different basic elements serve as components of a general optical system of lenses and mirrors:

- Interspaces with constant refractive index n (fig. 8.11).

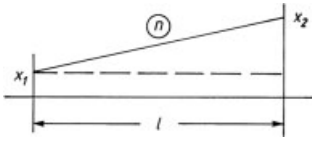


Figure 8.11: The interspace with a refractive index n as an optical unit element.

In this case we have

$$p_2 = p_1, \quad x_2 = x_1 + l\theta = x_1 + \frac{l}{n} p_1,$$

and the optical matrix for such an interspace is given by

$$S^L = \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \quad \text{with} \quad L = \frac{l}{n}, \quad (8.40)$$

where $L = l/n$ is the so-called *optical length*.

- Refracting surfaces (fig. 8.12). (A mirror can also be treated like a refracting surface if the optic axis is reversed.) We confine the discussion to surfaces that are rotationally symmetric and have the optic axis as their normal. Furthermore, we assume the rays to be close to the axis.

From fig. 8.12 we read off:

$$x_2 = x_1, \quad n_1(\alpha - \theta_1) = n_2(\alpha - \theta_2),$$

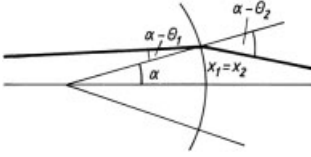


Figure 8.12: A refracting surface as an optical element. The same considerations apply for mirrors if the optic axis is reversed.

i.e.

$$x_2 = x_1, \quad p_2 = p_1 + (n_2 - n_1)\alpha = p_1 + \frac{n_2 - n_1}{R}x_1,$$

$$S^B = \begin{pmatrix} 1 & 0 \\ B & 1 \end{pmatrix}, \quad B = \frac{n_2 - n_1}{R}, \quad (8.41)$$

where $B = (n_2 - n_1)/R$ is called the *dioptric power* (refractive power) of the surface.

We now list some examples of combinations of these basic optical elements.

1. Combination of optical units:

$$S^{L_1} \circ S^{L_2} = S^{L_1+L_2}, \quad S^{B_1} \circ S^{B_2} = S^{B_1+B_2}. \quad (8.42)$$

In particular, if we consider a thin lens as a system with two refracting surfaces of negligible distance apart, we obtain

$$S^f = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \quad \text{with} \quad -\frac{1}{f} = (n_2 - n_1) \left(\frac{1}{R_2} - \frac{1}{R_1} \right). \quad (8.43)$$

We find

$$S^f \begin{pmatrix} x \\ p \end{pmatrix} = \begin{pmatrix} x' \\ p' \end{pmatrix} = \begin{pmatrix} x \\ -x/f + p \end{pmatrix},$$

i.e. $x' = x$; only the inclination changes.

2. Interspace–thin lens–interspace system:

$$S_{21} = S^{L_2} \cdot S^f \cdot S^{L_1} = \begin{pmatrix} 1 - L_2/f & L_1 + L_2 - (L_1 L_2/f) \\ -1/f & 1 - L_1/f \end{pmatrix}. \quad (8.44)$$

For $L_2 = f$ we find

$$S_{12} = \begin{pmatrix} 0 & f \\ -1/f & 1 - L_1/f \end{pmatrix},$$

i.e. $x_2 = fp_1$. Hence, all parallel rays with inclination p_1 will pass through the single point $x_2 = fp_1$ in the “back focal plane” behind the thin lens at a distance f . In particular, rays parallel to the axis (having inclination $p_1 = 0$) pass through the back focal point $x_2 = 0$.

For $L_1 = f$ we find

$$S_{12} = \begin{pmatrix} 1 - L_2/f & f \\ -1/f & 0 \end{pmatrix},$$

i.e. $p_2 = -x_1/f$. All rays passing through the point x_1 in the “front focal plane” will be mapped, independent of their inclination, into parallel rays with inclination $p_2 = -1/f x_1$. Those rays passing through the front focal point $x_1 = 0$ will be parallel to the axis behind the lens. These circumstances may be used for a geometrical construction of the image mapping by a thin lens (see fig. 8.13). The two planes F_1 and F_2 are conjugated if $L_1 + L_2 - (L_1 L_2)/f = 0$, or

$$\frac{1}{L_1} + \frac{1}{L_2} = \frac{1}{f}. \quad (8.45)$$

This is the famous *lens formula*.

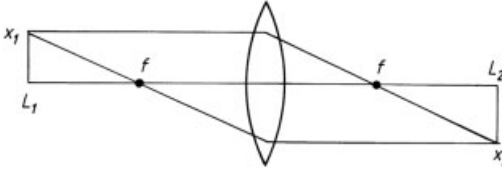


Figure 8.13: Imaging by a thin lens.

3. Lens–interspace–lens system:

$$S_{21} = S^{f_2} S^L S^{f_1} = \begin{pmatrix} 1 - L/f_1 & L \\ -1/f_1 - 1/f_2 + (L/f_1 f_2) & 1 - L/f_2 \end{pmatrix}. \quad (8.46)$$

The resulting dioptric power (refractive power) is given by

$$\frac{1}{f_{\text{res}}} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{L}{f_1 f_2} \quad \left(= \frac{L}{f^2} \quad \text{for} \quad f_1 = -f_2 = f \right). \quad (8.47)$$

A converging lens in combination with a diverging lens of opposite power results in a converging system.

4. Normal form of optical systems:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad ad - bc = 1$$

is the optical matrix of a Gaussian system that maps the ray coordinates (x_1, p_1) in plane F_1 into the ray coordinates

$$\begin{pmatrix} x_2 \\ p_2 \end{pmatrix} = \begin{pmatrix} ax_1 + bp_1 \\ cx_1 + dp_1 \end{pmatrix}$$

in plane F_2 .

We now will assume that $c \neq 0$, i.e. the system is non-telescopic. Telescopic systems can be reduced to non-telescopic systems by combining them with a fixed lens mapping.

A look at eq. (8.44) tells us that any non-telescopic image matrix can be written in the form $S_{21} = S^{L_2} S^f S^{L_1}$ with uniquely determined values for f , L_1 , and L_2 . On the other hand, there are uniquely determined values for h_1 and h_2 such that

$$\begin{pmatrix} 1 & -h_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & -h_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} = S^f,$$

with $f = -1/c$.

This observation admits the following interpretation. Consider the *front principal plane* H_1 , which is located behind F_1 with optical length h_1 and the *back principal plane* located in front of F_2 with optical length h_2 . For this case, the Gaussian mapping of rays from H_1 to H_2 is simply the mapping S^f of a thin lens with $f = -1/c$. Hence, knowing the position of the principal planes allows for a simple construction of the ray mapping. The front and back focal planes, B_1 and B_2 , are at a distance f in front of H_1 and behind H_2 , respectively. Rays parallel to the axis are transformed into rays through the back focal point, and rays passing through the front focal point are transformed into rays parallel to the axis (see fig. 8.14). The lens formula $1/L_1 + 1/L_2 = 1/f$ holds, if L_1 and L_2 are interpreted as the distances from the two principal planes.

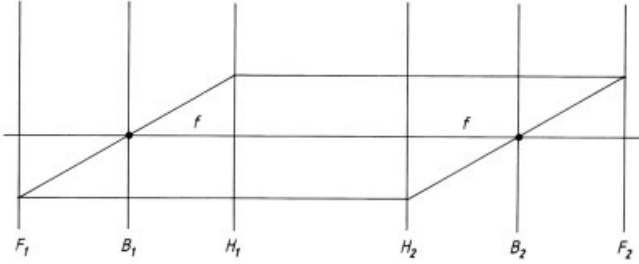


Figure 8.14: Location of the principal planes H_1 and H_2 as well as the focal planes B_1 and B_2 in the normal form of an optical system.

From the general form of a non-telescopic Gaussian optical matrix,

$$G = S^{L_1} S^f S^{L_2} = \begin{pmatrix} 1 - L_2/f & L_1 + L_2 - L_1 L_2/f \\ -1/f & 1 - L_1/f \end{pmatrix},$$

and considering the case in which F_1 and F_2 are conjugated ($L_1 + L_2 - L_1 L_2/f = 0$), one finds the *longitudinal magnification*

$$v_l = \frac{dl_2}{dl_1} = \frac{n_2}{n_1} \frac{dL_2}{dL_1}.$$

Indeed, from $dL_1 + dL_2 - (L_1/f)dL_2 - (L_2/f)dL_1 = 0$ we obtain

$$\frac{dL_2}{dL_1} = -\frac{1 - (L_2/f)}{1 - (L_1/f)},$$

$$v_l = \frac{n_2}{n_1} \frac{1 - (L_2/f)}{1 - (L_1/f)}.$$

On the other hand, the *transverse magnification* is given by

$$v_t = \frac{x_2}{x_1} = 1 - \frac{L_2}{f} = \frac{1}{1 - L_1/f},$$

and we find

$$v_l = \frac{n_2}{n_1} v_t^2.$$

This proves that Herschel's condition,

$$z_1 n_1 \sin^2(\theta_1/2) = z_2 n_2 \sin^2(\theta_2/2),$$

and Abbe's sine condition,

$$x_1 n_1 \sin \theta_1 = x_2 n_2 \sin \theta_2,$$

are incompatible for $\theta_1 \neq \theta_2$, because they yield

$$v_l = \frac{n_1}{n_2} \frac{\sin^2(\theta_1/2)}{\sin^2(\theta_2/2)},$$

$$v_t = \frac{n_1}{n_2} \frac{\sin \theta_1}{\sin \theta_2},$$

i.e. $v_l \neq (n_2/n_1)v_t^2$ for $\theta_1 \neq \theta_2$. This fact agrees with the non-existence of non-trivial perfect optical instruments, for which the two conditions should be fulfilled.

8.8 Lens defects and Seidel's theory of aberrations

As we have seen, in Gaussian optics the mapping of the ray coordinates from plane F_0 to F is given by

$$\begin{aligned} \mathbf{x} &= a\mathbf{x}_0 + b\mathbf{p}_0, \\ \mathbf{p} &= c\mathbf{x}_0 + d\mathbf{p}_0, \end{aligned} \quad \text{with } ad - bc = 1. \tag{8.48}$$

Within the Gaussian approximation the eikonal functions $S_G(\mathbf{x}, \mathbf{x}_0)$ and $M_G(\mathbf{p}, \mathbf{x}_0)$ can be determined up to irrelevant constants by solving for $\mathbf{x}, \mathbf{x}_0$ and $\mathbf{p}, \mathbf{x}_0$, respectively:

$$\mathbf{p}_0 = \frac{1}{b}(\mathbf{x} - a\mathbf{x}_0), \quad \mathbf{p} = \frac{1}{b}(d\mathbf{x} - \mathbf{x}_0),$$

and since $\partial S_G/\partial \mathbf{x} = \mathbf{p}$ and $\partial S_G/\partial \mathbf{x}_0 = -\mathbf{p}_0$, we find that

$$S_G(\mathbf{x}, \mathbf{x}_0) = \frac{1}{2b}(d\mathbf{x}^2 - 2\mathbf{x} \cdot \mathbf{x}_0 + a\mathbf{x}_0^2) + \text{const.} \quad (8.49)$$

Obviously, S is only defined for $b \neq 0$, i.e. when F and F_0 are not conjugate. Similarly, one finds

$$\mathbf{p}_0 = \frac{1}{d}(\mathbf{p} - c\mathbf{x}_0), \quad \mathbf{x} = \frac{1}{d}(\mathbf{x}_0 + b\mathbf{p}),$$

and since $\partial M_G/\partial \mathbf{p} = -\mathbf{x}$ and $\partial M_G/\partial \mathbf{x}_0 = -\mathbf{p}_0$, we obtain for the mixed characteristics within the Gaussian approximation:

$$M_G(\mathbf{p}, \mathbf{x}_0) = -\frac{1}{2d}(b\mathbf{p}^2 + 2\mathbf{p} \cdot \mathbf{x}_0 - c\mathbf{x}_0^2). \quad (8.50)$$

M_G is defined even for $b = 0$. In a similar way we get

$$T_G(\mathbf{p}, \mathbf{p}_0) = -\frac{1}{2c}(a\mathbf{p}^2 - 2\mathbf{p}\mathbf{p}_0 + d\mathbf{p}_0^2). \quad (8.51)$$

We now assume that, in a Gaussian approximation, the planes F_0 and F are conjugated and we want to determine systematically the defocussing effect of the next-to-leading terms. This is called *Seidel's theory of aberrations*.

A suitable quantity for this purpose is the mixed characteristic $M(\mathbf{p}, \mathbf{x}_0)$, which we write in the form

$$\begin{aligned} M(\mathbf{p}, \mathbf{x}_0) &= M_G(\mathbf{p}, \mathbf{x}_0) + M_1(\mathbf{p}, \mathbf{x}_0) \\ &= \frac{1}{2d}(c\mathbf{x}_0^2 - 2\mathbf{p} \cdot \mathbf{x}_0) + M_1(\mathbf{p}, \mathbf{x}_0), \end{aligned} \quad (8.52)$$

where M_1 contains terms in $\mathbf{p}$ and $\mathbf{x}_0$ which are of higher than second order. (Since F_0 and F_1 are conjugate, we have $b = 0$ in the Gaussian approximation.) Because of the rotational invariance of the system, M_1 can only depend on the invariant terms $u = \mathbf{x}_0^2$, $v = \mathbf{p}^2$, and $w = 2\mathbf{x}_0 \cdot \mathbf{p}$. The lowest non-trivial contributions in a Taylor expansion of M_1 are of the form

$$M_1(\mathbf{p}, \mathbf{x}_0) = -[\frac{1}{4}Fu^2 + \frac{1}{4}Av^2 + \frac{1}{8}(C - D)w^2 + \frac{1}{2}Duv + \frac{1}{2}Euw + \frac{1}{6}Bvw]. \quad (8.53)$$

Note that, due to the rotational invariance, there are no terms of third order. From this we get

$$\mathbf{x} = -\frac{\partial M}{\partial \mathbf{p}} = -\frac{\partial M_G}{\partial \mathbf{p}} - \frac{\partial M_1}{\partial \mathbf{p}} = \mathbf{x}_G + \Delta\mathbf{x} = \frac{1}{d}\mathbf{x}_0 + \Delta\mathbf{x}. \quad (8.54)$$

While $\mathbf{x}_G = (1/d)\mathbf{x}_0$ is independent of $\mathbf{p}$, the expression $\Delta\mathbf{x} = -\partial M_1/\partial \mathbf{p}$ will, in general, depend on $\mathbf{p}$ such that the system is no longer focussing. The deviations will increase with increasing $\mathbf{x}_0$ and $\mathbf{p}$. A short calculation gives

$$\Delta\mathbf{x} = -\frac{\partial M_1}{\partial \mathbf{p}} = [Eu + \frac{1}{3}Bv + \frac{1}{2}(C - D)w]\mathbf{x}_0 + [Av + Du + \frac{1}{3}Bw]\mathbf{p}. \quad (8.55)$$

The relevance of the different contributions becomes most obvious in polar coordinates for $\mathbf{p}$:

$$\mathbf{x}_0 = (x_0, 0), \quad \mathbf{p} = p(\cos \varphi, \sin \varphi),$$

i.e.

$$u = x_0^2, \quad v = p^2, \quad w = 2x_0p \cos \varphi,$$

and

$$\begin{aligned} \Delta x_1 &= Ap^3 \cos \varphi + \frac{1}{3}Bp^2x_0(2 \cos^2 \varphi + 1) + Cpx_0^2 \cos \varphi + Ex_0^3, \\ \Delta x_2 &= Ap^3 \sin \varphi + \frac{2}{3}Bp^2x_0 \sin \varphi \cos \varphi + Dpx_0^2 \sin \varphi. \end{aligned} \quad (8.56)$$

Here we have ordered the terms according to increasing powers of x_0 and decreasing powers of p .

- The first term,

$$\Delta \mathbf{x}^A = Ap^2 \mathbf{p}, \quad (8.57)$$

is called *spherical aberration*. It is independent of x_0 and can be suppressed by a small aperture such that only rays that are almost parallel to the axis can pass.

- The second term can be written in the form

$$(\Delta x_1^B, \Delta x_2^B) = \frac{1}{3}Bp^2x_0(2 + \cos 2\varphi, \sin 2\varphi). \quad (8.58)$$

Keeping x_0 and p fixed, which may be realized by a screen with a circular aperture, $(\Delta x_1^B, \Delta x_2^B)$ describes a circle of radius $\frac{1}{3}Bp^2x_0$ and central point $(\frac{2}{3}Bp^2x_0, 0)$. If p is varied within the interval $0 \leq p \leq \rho$, one obtains a picture like that in fig. 8.15. The corresponding image defect increases with x_0p^2 and is called *coma* (presumably because of its similarity with a comet).

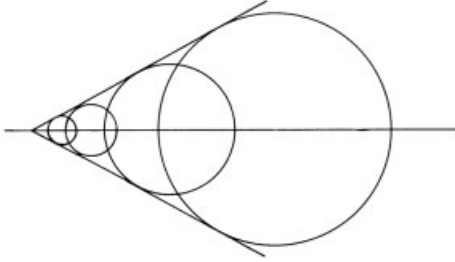


Figure 8.15: Coma is a type of aberration that deforms the image of a point in a characteristic way.

- The third term,

$$(\Delta x_1^C, \Delta x_2^C) = px_0^2(C \cos \varphi, D \sin \varphi), \quad (8.59)$$

is called *astigmatism* (“having no point”). Keeping x_0 and p fixed, $(\Delta x_1^C, \Delta x_2^C)$ describes an ellipse, which degenerates to a circle for $C = D$. If one tries to improve focusing by a small displacement L of the image plane, one finds

$$(\Delta x_1'^c, \Delta x_2'^c) = ((Cx_0^2 + L)p \cos \varphi, (Dx_0^2 + L)p \sin \varphi). \quad (8.60)$$

As long as $C \neq D$, the two terms $\Delta x_1'^c$ and $\Delta x_2'^c$ cannot vanish simultaneously. As a result of astigmatism, the image of $\mathbf{x}_0$ is not a point but consists of two mutually orthogonal image lines of length $2|C - D|x_0^2 p$. Their respective distances from $\mathbf{x}$ are $-Cx_0^2$ and $-Dx_0^2$.

- Finally, the last term in eq. (8.56) has the form

$$\Delta \mathbf{x} = E x_0^2 \mathbf{x}_0. \quad (8.61)$$

It is independent of $\mathbf{p}$ and thus does not influence the focussing properties of the system. However, it results in a *distortion* of the image. A rectangular net is mapped onto a distorted net like the ones shown in fig. 8.16. Depending on the sign of E , one speaks of a *barrel distortion* ($E > 0$) or a *pin-cushion distortion* ($E < 0$).

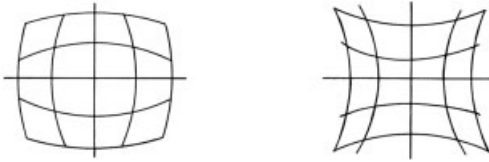


Figure 8.16: Barrel distortion (left) and pin-cushion distortion (right) of a rectangular lattice.

Any of the above-mentioned aberrations can be avoided or suppressed by a suitable construction of the optical system.

Apart from these geometrical lens defects, we also have *chromatic aberration*, i.e. a disturbance of the image by a wavelength dependence of the refractive index. It can be suppressed by combining different glass types. Furthermore, there are distortions due to diffractive effects, which are beyond the realm of geometrical optics.

9 Geometric theory of caustics

In this chapter, we will first generalize the discussion on short-wave asymptotics to more general partial differential equations. Then we will give a brief introduction to the geometric theory of characteristics, Lagrangian manifolds, and caustics.

9.1 Short-wave asymptotics for linear partial differential equations

Let

$$H\left(x, -\frac{i}{k}\partial\right) = A(x) + \sum_{r=1}^N \left(-\frac{i}{k}\right)^r A_{i_1 \dots i_r}(x) \frac{\partial}{\partial x^{i_1}} \cdots \frac{\partial}{\partial x^{i_r}} \quad (9.1)$$

be a general linear differential operator of order N acting on scalar functions

$$u : \mathbb{R}^n \longrightarrow \mathbb{C}.$$

If, in the differential operator, we replace all derivatives $-(i/k)(\partial/\partial x^j)$ by $p_j \in \mathbb{R}$, we obtain, for $x, p \in \mathbb{R}^n$, a function $H(x, p) : \mathbb{R}^{2n} \rightarrow \mathbb{C}$, which is called the *symbol* σ_H of the differential operator. We shall always assume that $H(x, p)$ has a zero in p for any $x \in \mathbb{R}^n$. For a linear differential operator, $H(x, p)$ is a polynomial in p . However, one may generalize the above considerations to so-called *pseudo-differential operators*, for which $H(x, p)$, as a function of x and p , should not increase too rapidly.

For commutators of differential operators we find the important identity:

$$\sigma_{[A, B]} = \frac{1}{ik} \{\sigma_A, \sigma_B\} + \mathcal{O}(k),$$

where $\{\sigma_A, \sigma_B\}$ is the Poisson bracket of the symbols. This relation is fundamental for the theory of the classical limit of quantum mechanics.

We are interested in the asymptotic behavior of solutions of the partial differential equation

$$H\left(x, -\frac{i}{k}\partial\right) u(x) = 0 \quad (9.2)$$

for large values of k . As explained in chapter 7, we insert the following ansatz into the differential equation:

$$u(x) = \varphi_k(x) e^{ikS(x)}, \quad \varphi_k(x) = \varphi_0(x) + \frac{1}{k} \varphi_1(x) + \cdots \quad (9.3)$$

In this way we find differential equations for the quantities S and φ_i . We will discuss the first two orders in $1/k$ explicitly.

The ansatz (9.3) is well known in the theory of quasi-classical approximations to quantum mechanics and goes by the name of the *WKB ansatz for the wave function*. In this context, the variable k is replaced by $1/\hbar$, the inverse of Planck's constant. The relation between classical mechanics and quantum mechanics is the same as the one between geometrical optics and wave optics.

If, for arbitrary $\xi \in \mathbb{R}^n$, we define the operator

$$D = \xi_j \frac{\partial}{\partial x_j}, \quad (9.4)$$

we obtain immediately that

$$\begin{aligned} -\frac{i}{k} D e^{ikS} &= e^{ikS} DS, \\ \left(-\frac{i}{k} D\right)^2 e^{ikS} &= e^{ikS} \left((DS)^2 - \frac{i}{k} D^2 S\right), \\ &\vdots \\ \left(-\frac{i}{k} D\right)^r e^{ikS} &= e^{ikS} \left\{ (DS)^r - \frac{i}{k} \binom{r}{2} (DS)^{r-2} D^2 S + \mathcal{O}\left(\frac{1}{k^2}\right) \right\}. \end{aligned} \quad (9.5)$$

Furthermore, Leibniz's rule tells us

$$\begin{aligned} \left(-\frac{i}{k} D\right)^r \varphi_k e^{ikS} &= e^{ikS} \left\{ (DS)^r \varphi_k - \frac{i}{k} \binom{r}{2} (DS)^{r-2} D^2 S \varphi_k - \frac{i}{k} r (DS)^{r-1} D \varphi_k + \mathcal{O}\left(\frac{1}{k^2}\right) \right\}. \end{aligned} \quad (9.6)$$

In the explicit calculation of $H(x, -(i/k)\partial) \varphi_k e^{ikS}$, we encounter terms of this form, and we finally obtain

$$\begin{aligned} H\left(x, -\frac{i}{k} \partial\right) \varphi_k e^{ikS} &= e^{ikS} \left\{ H(x, \partial S) \varphi_0 - \frac{i}{k} \left[\frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial p_j} \frac{\partial^2 S}{\partial x^i \partial x^j} \varphi_0 + \frac{\partial H}{\partial p_i} \frac{\partial \varphi_0}{\partial x^i} \right] \right. \\ &\quad \left. + \frac{1}{k} H(x, \partial S) \varphi_1 + \mathcal{O}\left(\frac{1}{k^2}\right) \right\}. \end{aligned} \quad (9.7)$$

In zeroth order we find the so-called *characteristic equation*

$$H(x, \partial S) = 0, \quad (9.8)$$

and in first order the *transport equation*

$$\frac{\partial H(x, \partial S)}{\partial p_i} \frac{\partial \varphi_0(x)}{\partial x^i} + \frac{1}{2} \frac{\partial^2 H(x, \partial S)}{\partial p_i \partial p_j} \frac{\partial^2 S(x)}{\partial x^i \partial x^j} \varphi_0(x) = 0. \quad (9.9)$$

The special case we discussed so far is Helmholtz's equation,

$$-\left(\frac{1}{2k^2}\Delta + \frac{1}{2}n^2\right)u = 0, \quad (9.10)$$

and it is of special importance for optics.

In this case, the symbol function is given by

$$H(x, p) = \frac{1}{2}p^2 - \frac{1}{2}n^2. \quad (9.11)$$

The characteristic equation,

$$\frac{1}{2}(\nabla S)^2 - \frac{1}{2}n^2 = 0, \quad (9.12)$$

is the eikonal equation, and the transport equation reads

$$\nabla S \cdot \nabla \varphi_0 + \frac{1}{2}\varphi_0 \Delta S = 0. \quad (9.13)$$

It agrees with eq. (7.22) given in chapter 7.

The solutions of the characteristic equation (9.8) and the transport equation (9.9) can be obtained from the Hamiltonian system of ordinary differential equations:

$$\dot{x}^i = \frac{\partial H(x, p)}{\partial p^i}, \quad \dot{p}_i = -\frac{\partial H(x, p)}{\partial x^i}.$$

On the other hand, a solution of the characteristic equation leads to solutions of the Hamiltonian system, which again shows the connection between wave theory and classical mechanics. In the next section we shall see that the following problems are intimately related.

A) Solution of the linear partial differential equation of n th order:

$$H\left(x, -\frac{i}{k}\partial\right)u = 0,$$

for large values of k .

B) Solution of the characteristic equation, i.e. the nonlinear partial differential equation of first order:

$$H(x, \partial S) = 0,$$

as well as the transport equation.

C) Solutions of the nonlinear system of ordinary differential equations:

$$\dot{x} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial x},$$

which are sometimes called the *characteristics* for the differential equation A.

These relations will be explained in the following sections. As an aside we note that this theory can be generalized to linear partial differential equations for m -component vector functions:

$$u : \mathbb{R}^n \longrightarrow \mathbb{C}^m.$$

However, for this complicated case, one has to introduce m eikonal functions S_i ($i = 1, \dots, m$).

9.2 Solution of the characteristic equation

In this section, we will describe the relation between the solutions of Hamilton's equations,

$$\dot{x}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x^i}, \quad (9.14)$$

and the characteristic equation

$$H(x, \partial S) = 0. \quad (9.15)$$

For Helmholtz's equation (9.10), the corresponding Hamiltonian system,

$$\dot{\mathbf{x}} = \mathbf{p}, \quad \dot{\mathbf{p}} = \frac{1}{2} \nabla n^2, \quad \text{i.e.} \quad \ddot{\mathbf{x}} = \frac{1}{2} \nabla n^2, \quad (9.16)$$

is just the "equation of motion" for light rays. In general, the following theorem holds.

Theorem 9.1. *Let $S(x)$ be a solution of the characteristic equation $H(x, \partial S) = 0$, and let $(x^i(\tau), p_i(\tau))$ be a solution of the Hamiltonian system $\dot{x}^i = \partial H / \partial p_i$, $\dot{p}_i = -\partial H / \partial x^i$ with initial conditions: $x^i(0) = x_0^i$, $p_i(0) = p_{0i} = (\partial S / \partial x^i)(x_0)$. Then, along the solution trajectory, we have $p_i(\tau) = (\partial S / \partial x^i)(x^i(\tau))$. Furthermore, $S(x(\tau)) = S(x(0)) + \int_0^\tau d\tau' p_i(\tau') \dot{x}^i(\tau')$.*

Proof: We consider the path $\tilde{x}(\tau)$, which follows from the differential equation

$$\dot{\tilde{x}}^i(\tau) = \frac{\partial H}{\partial p_i} \left(\tilde{x}(\tau), \frac{\partial S}{\partial x}(\tilde{x}(\tau)) \right)$$

with the initial condition $\tilde{x}(0) = x_0$. In addition we define

$$\tilde{p}_i(\tau) = \frac{\partial S}{\partial x^i}(\tilde{x}(\tau)) \quad \text{such that} \quad \tilde{p}_i(0) = p_{0i}.$$

We will show that $(\tilde{x}(\tau), \tilde{p}(\tau))$ is also a solution of the Hamiltonian system. Because the solutions of ordinary differential equations are determined uniquely by the initial conditions, we have $x^i(\tau) = \tilde{x}^i(\tau)$ and $p_i(\tau) = \tilde{p}_i(\tau) = (\partial S / \partial x^i)(x(\tau))$.

The first part of Hamilton's equations,

$$\dot{\tilde{x}}^i(\tau) = \frac{\partial H}{\partial p_i}(\tilde{x}(\tau), \tilde{p}(\tau)), \quad (9.17)$$

follows immediately from the definition of $\tilde{x}^i$ and $\tilde{p}_i$.

From the eikonal equation, $H(x, \partial S) = 0$, we obtain by differentiation

$$\frac{\partial H}{\partial x^i}(x, \partial S) + \frac{\partial H}{\partial p_j}(x, \partial S) \frac{\partial^2 S}{\partial x^i \partial x^j} = 0.$$

So, the second part of Hamilton's equation is also satisfied:

$$\dot{\tilde{p}}_i = \frac{d}{d\tau} \frac{\partial S}{\partial x^i}(\tilde{x}) = \frac{\partial^2 S}{\partial x^i \partial x^j} \dot{\tilde{x}}^j = \frac{\partial^2 S(\tilde{x})}{\partial x^i \partial x^j} \frac{\partial H}{\partial p_j} \left(\tilde{x}, \frac{\partial S}{\partial x}(\tilde{x}) \right) = -\frac{\partial H}{\partial x^i}(\tilde{x}, \tilde{p}). \quad (9.18)$$

For Helmholtz's equation we have $\dot{x} = p$, i.e.

$$\dot{x}(\tau) = \nabla S(x(\tau)); \quad (9.19)$$

once more we find the light rays to be the orthogonal trajectories to the surfaces of constant phase.

Conversely, we can use the solutions of the Hamiltonian system for solving the “initial value problem” of the characteristic equation.

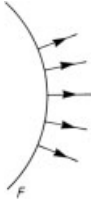


Figure 9.1: Construction of a solution of the eikonal equation from an $(n-1)$ -dimensional hypersurface.

Let F be an $(n-1)$ -dimensional hypersurface in $\mathbb{R}^n$ defined by the equation $S_0(x) = 0$ (see fig. 9.1). Introducing local parameters $y \in \mathbb{R}^{n-1}$, the surface can be written in the form $x = x(y)$ such that we have $S_0(x(y)) \equiv 0$ identically in y . For each point $x(y)$ on F we define $p(y) = (\partial S_0 / \partial x)(x(y))$. Furthermore, the Hamiltonian function should vanish on F : $H(x(y), p(y)) = 0$.

We now consider solutions of the Hamiltonian system that start from the point $x(y) \in F$ with initial momentum $p(y)$, i.e. functions $\xi(\tau, y)$ and $\pi(\tau, y)$ for which

$$\frac{\partial \xi}{\partial \tau}(\tau, y) = \frac{\partial H}{\partial p}(\xi(\tau, y), \pi(\tau, y)), \quad \frac{\partial \pi}{\partial \tau}(\tau, y) = -\frac{\partial H}{\partial x}(\xi(\tau, y), \pi(\tau, y)), \quad (9.20)$$

and

$$\xi(0, y) = x(y), \quad \pi(0, y) = p(y) = \frac{\partial S_0}{\partial x}(x(y)).$$

We now assume that $\partial \xi / \partial \tau(0, y)$ is nowhere tangential to F . In a certain neighborhood of the initial surface F , each point $x \in \mathbb{R}^n$ lies locally on exactly one trajectory $\xi(\tau, y)$. For not too large values of τ , the mapping

$$\xi : (\tau, y) \longmapsto \xi(\tau, y) \in \mathbb{R}^n \quad (9.21)$$

is invertible and smooth, i.e. it is a diffeomorphism. Under these conditions the Jacobian matrix

$$D\xi(\tau, y) = \left(\frac{\partial \xi}{\partial \tau}(\tau, y), \frac{\partial \xi}{\partial y^1}(\tau, y), \dots, \frac{\partial \xi}{\partial y^{n-1}}(\tau, y) \right) \quad (9.22)$$

is an invertible $n \times n$ matrix and the Jacobian determinant

$$\mathcal{J}(\tau, y) = \det D\xi(\tau, y) \quad (9.23)$$

satisfies $\mathcal{J}(\tau, y) \neq 0$.

If the values of τ become larger, the mapping $\xi(\tau, y)$ may no longer be a diffeomorphism; in particular, there may be “critical” points τ_0, y_0 where $\mathcal{J}(\tau_0, y_0) = 0$. This case, where paths $\xi(\tau, y)$ with infinitesimally neighboring parameter values y intersect at a point $\xi(\tau_0, y_0)$, will be of interest in our discussion of caustics.

Right now, we only consider a neighborhood of F where $\xi(\tau, y)$ is a diffeomorphism.

As an identity in τ and y we find

$$H(\xi(\tau, y), \pi(\tau, y)) = 0, \quad (9.24)$$

because for $\tau = 0$ this is a consequence of the initial conditions, and

$$\frac{\partial H}{\partial \tau} H(\xi(\tau, y), \pi(\tau, y)) = \frac{\partial H}{\partial x} \dot{\xi} + \frac{\partial H}{\partial p} \dot{\pi} = \frac{\partial H}{\partial x} \frac{\partial H}{\partial p} - \frac{\partial H}{\partial p} \frac{\partial H}{\partial x} = 0.$$

Differentiation with respect to y yields

$$\frac{\partial H}{\partial x}(\xi(\tau, y), \pi(\tau, y)) \frac{\partial \xi}{\partial y} + \frac{\partial H}{\partial p}(\xi(\tau, y), \pi(\tau, y)) \frac{\partial \pi}{\partial y} = 0. \quad (9.25)$$

The idea for solving the characteristic equation using the solutions $\xi(\tau, y), \pi(\tau, y)$ consists in constructing the solution $S(x)$ from the curves $\xi(\tau, y)$ which start at the surface F , such that the level surfaces of S have $\pi(\tau, y)$ as their normal. For this purpose, we make the ansatz

$$S \circ \xi(\tau, y) = S(\xi(\tau, y)) = \int_0^\tau d\tau' \pi(\tau', y) \frac{\partial \xi}{\partial \tau'}(\tau', y). \quad (9.26)$$

If we can prove that

$$\frac{\partial S}{\partial x}(\xi(\tau, y)) = \pi(\xi(\tau, y)), \quad (9.27)$$

we have also proven (because of eq. (9.24)) that S is a solution of the characteristic equation with the initial condition $S(x(y)) = S_0(x(y)) = 0$. The chain rule gives us

$$\frac{\partial}{\partial \tau} S \circ \xi = \frac{\partial S}{\partial x}(\xi) \frac{\partial \xi}{\partial \tau} \quad (9.28)$$

and

$$\frac{\partial}{\partial y} S \circ \xi = \frac{\partial S}{\partial x}(\xi) \frac{\partial \xi}{\partial y}. \quad (9.29)$$

Since $D\xi$ is invertible, we only have to prove that

$$\frac{\partial}{\partial \tau} S \circ \xi = \pi \frac{\partial \xi}{\partial \tau} \quad (9.30)$$

and

$$\frac{\partial}{\partial y} S \circ \xi = \pi \frac{\partial \xi}{\partial y}. \quad (9.31)$$

Identity (9.30) follows immediately from the definition of S (eq. (9.26)). Concerning eq. (9.31), we first find for $\tau = 0$:

$$\frac{\partial}{\partial y} S \circ \xi(0, y) = \frac{\partial}{\partial y} S(x(y)) = \frac{\partial}{\partial y} S_0(x(y)) = p(y) \frac{\partial x}{\partial y} = \pi(0, y) \frac{\partial \xi}{\partial y}(0, y),$$

due to the initial condition, and it remains to prove that

$$\frac{\partial}{\partial \tau} \frac{\partial}{\partial y} S \circ \xi = \frac{\partial}{\partial \tau} \left(\pi \frac{\partial \xi}{\partial y} \right).$$

However, we have that

$$\begin{aligned} \frac{\partial}{\partial \tau} \frac{\partial}{\partial y} S \circ \xi &= \frac{\partial}{\partial y} \frac{\partial}{\partial \tau} S \circ \xi \\ &= \frac{\partial}{\partial y} \left(\pi \frac{\partial \xi}{\partial \tau} \right) && \text{(because of eq. (9.30))} \\ &= \frac{\partial \pi}{\partial y} \frac{\partial H}{\partial p} + \pi \frac{\partial}{\partial \tau} \frac{\partial \xi}{\partial y} && \text{(because of Hamilton's equations)} \\ &= -\frac{\partial H}{\partial x} \frac{\partial \xi}{\partial y} + \pi \frac{\partial}{\partial \tau} \frac{\partial \xi}{\partial y} && \text{(because of eq. (9.25))} \\ &= \frac{\partial \pi}{\partial \tau} \frac{\partial \xi}{\partial y} + \pi \frac{\partial}{\partial \tau} \frac{\partial \xi}{\partial y} && \text{(because of Hamilton's equations)} \\ &= \frac{\partial}{\partial \tau} \left(\pi \frac{\partial \xi}{\partial y} \right), \end{aligned}$$

which completes the proof. $\square$

We find a second useful geometrical interpretation of our solution of the initial value problem for the characteristic equation by reformulating the problem not in position space but in the $2n$ -dimensional phase space:

$$P = \{(x, p) \mid x, p \in \mathbb{R}^n\}.$$

A submanifold $\Lambda \subset P$ is called *isotropic* if the 2-form

$$\omega = \sum_{i=1}^n dx^i \wedge dp_i = -d \sum p_i dx^i$$

vanishes on this submanifold. This implies that, with respect to the bilinear form $\omega(p)$, the tangent space $T_p \Lambda$ to Λ is an isotropic subspace (in the sense of section 8.6) of the tangent space $T_p P$ of P for all $p \in \Lambda$. An isotropic submanifold Λ of P with (maximal) dimension n is called a *Lagrangian submanifold*.

To the $(n-1)$ -dimensional initial manifold $F = \{x \mid S_0(x) = 0\} = \{x(y)\}$ we associate the following $(n-1)$ -dimensional submanifold of the phase space:

$$\tilde{F} = \left\{ \left(x(y), \frac{\partial S_0}{\partial x}(x(y)) \right) \right\} \subset P. \quad (9.32)$$

$\tilde{F}$ is isotropic, because on $\tilde{F}$ we have

$$p_i dx^i = \frac{\partial S_0}{\partial x^i}(x(y)) dx^i(y) = dS_0(x(y)),$$

and therefore

$$\omega|_{\tilde{F}} = -d dS_0|_{\tilde{F}} = 0.$$

For the same reason, the manifold

$$\Lambda_F = \{(\xi(\tau, y), \pi(\tau, y))\} \subset P, \quad (9.33)$$

which one obtains by combining all solution curves of the Hamiltonian system starting on $\tilde{F}$, is a Lagrangian submanifold of P :

$$p_i dx^i|_{\Lambda_F} = \pi_i d\xi^i = \frac{\partial S}{\partial x^i} d\xi^i = dS|_{\Lambda_F}.$$

For any fixed value of $\tau = \tau_0$, the $(n-1)$ -dimensional submanifolds

$$\tilde{F}_{\tau_0} = \{(\xi(\tau_0, y), \pi(\tau_0, y))\} \quad (9.34)$$

are isotropic and we have $\tilde{F}_0 = \tilde{F}$.

This formulation of the initial value problem of the characteristic equation suggests the following important generalization: Not every $(n-1)$ -dimensional isotropic manifold can be written as

$$\tilde{F} = \left\{ (x, p) \left| S_0(x) = 0, p = \frac{\partial S_0}{\partial x} \right. \right\}. \quad (9.35)$$

A counter-example is the following $(n-1)$ -dimensional manifold associated to a given point $x_0 \in \mathbb{R}^n$:

$$\tilde{F}_0 = \{(x, p) \mid x = x_0, H(x_0, p) = 0\}.$$

This manifold is also isotropic, because $dx^i|_{\tilde{F}_0} = 0$.

Again we can use the solution trajectories starting on $\tilde{F}_{x_0}$,

$$(\xi(\tau, y), \pi(\tau, y)), \quad \xi(0, y) = x_0, \quad \pi(0, y) = p(y), \quad H(x_0, p(y)) = 0,$$

for the construction of a Lagrangian submanifold Λ_{x_0} and for the determination of a solution $S(x, x_0)$ of the characteristic equation. However, $S(x, x_0)$ will no longer be differentiable at the point $x = x_0$.

In this case, one considers all curves $\xi(\tau, y)$ emanating from a point x_0 instead of a manifold F and constructs the solutions of the eikonal equation as the level surfaces of S orthogonal to $\pi(\tau, y)$ (see fig. 9.2).

For Helmholtz's equation, this solution $S(x, x_0)$ is just the point characteristic. Because $(\nabla S)^2 = n^2$, we find in a vicinity of $x = x_0$ the solution $S(x, x_0)$ to be of the form

$$S(x, x_0) = n(x_0)|x - x_0| + R(x, x_0), \quad (9.36)$$

where $R(x, x_0)$ is regular at $x = x_0$ and $R(x_0, x_0) = 0$.

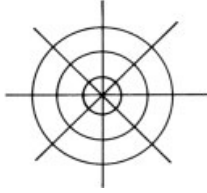


Figure 9.2: A solution of the eikonal equation, which is orthogonal to the momenta of all characteristics emanating from a point x_0 .

9.3 Solution of the transport equation

A solution of the transport equation,

$$\frac{\partial H}{\partial p_i}(x, \partial S) \frac{\partial \varphi_0}{\partial x^i} + \frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial p_j}(x, \partial S) \frac{\partial^2 S}{\partial x^i \partial x^j} \varphi_0(x) = 0, \quad (9.37)$$

with the initial value problem

$$\varphi_0(x(y)) = \psi(y), \quad (9.38)$$

can also be constructed (at least locally) from the solutions $(\xi(\tau, y), \pi(\tau, y))$ of the corresponding Hamiltonian system.

Obviously, the first term in eq. (9.37) can be written as

$$\frac{d}{d\tau} \varphi_0(\xi(\tau, y)) = \frac{\partial \varphi_0}{\partial x^i} \frac{\partial \xi^i}{\partial \tau}(\tau, y) = \frac{\partial H}{\partial p_i}(\xi, \partial S(\xi)) \frac{\partial \varphi_0}{\partial x^i}(\xi). \quad (9.39)$$

The second term in the transport equation is closely related to the distortion of the volume element $\mathcal{J}(\tau, y) = \det D\xi(\tau, y)$ due to the flux

$$\frac{\partial \xi}{\partial \tau} = \frac{\partial H}{\partial p}(\xi, \partial S(\xi)). \quad (9.40)$$

As is well known, the relative volume distortion is given by the divergence of the vector field on the right-hand side, i.e.

$$\frac{1}{\mathcal{J}} \frac{d\mathcal{J}}{d\tau} = \frac{\partial}{\partial \xi^i} \frac{\partial H}{\partial p_i}(\xi, \partial S(\xi)) = \frac{\partial^2 H}{\partial p_i \partial x^i}(\xi, \partial S(\xi)) + \frac{\partial^2 H}{\partial p_i \partial p_j}(\xi, \partial S(\xi)) \frac{\partial^2 S}{\partial x^i \partial x^j}.$$

Therefore,

$$\frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial p_j} \frac{\partial^2 S}{\partial x^i \partial x^j} = \frac{1}{\sqrt{\mathcal{J}}} \frac{d\sqrt{\mathcal{J}}}{d\tau} - \frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial x^i}, \quad (9.41)$$

and for the transport equation along the paths $\xi(\tau, y)$ we can write

$$\frac{d}{d\tau} (\varphi_0 \circ \xi) + \frac{1}{\sqrt{\mathcal{J}}} \frac{d\sqrt{\mathcal{J}}}{d\tau} \varphi_0 \circ \xi = \frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial x^i}(\xi, \partial S(\xi)) \varphi_0 \circ \xi$$

or

$$\frac{d}{d\tau} \frac{(\sqrt{\mathcal{J}} \varphi_0 \circ \xi)}{(\sqrt{\mathcal{J}} \varphi_0 \circ \xi)} = \frac{1}{2} \frac{\partial^2 H}{\partial p_i \partial x^i}(\xi, \partial S(\xi)).$$

By integration we find

$$\varphi_0(\xi(\tau, y)) = \frac{\sqrt{\mathcal{J}(0, y)}}{\sqrt{\mathcal{J}(\tau, y)}} \varphi_0(\xi(0, y)) \exp \left\{ \int_0^\tau d\tau' \frac{\partial^2 H}{\partial p_i \partial x^i}(\xi(\tau', y), \partial S(\xi(\tau', y))) \right\}. \quad (9.42)$$

For the eikonal equation and, more generally, whenever the operator $H(x, -(i/k)\partial)$ is formally self-adjoint, the term $\frac{1}{2}(\partial^2 H / \partial p_i \partial x^i)$ in eqs. (9.41) and (9.42) vanishes.

In this case, up to terms of order $1/k^2$, we obtain for the equation $H(x, -(i/k)\partial)u = 0$ the solution

$$u(x) = \frac{\sqrt{\mathcal{J}(0, y(x))}}{\sqrt{\mathcal{J}(\tau(x), y(x))}} \varphi_0(\xi(0, y(x))) e^{ikS(x)}, \quad (9.43)$$

which is valid as long as $\xi(\tau, y)$ is a diffeomorphism. In this formula, $\tau(x)$ and $y(x)$ are values for τ and y that are uniquely determined by the condition $\xi(\tau(x), y(x)) = x$. Notice that the amplitude tends to infinity at the critical points where $\mathcal{J} = 0$.

Looking again at Helmholtz's equation, the equation for $(d/d\tau)\sqrt{\mathcal{J}}$ simplifies to

$$\frac{1}{\sqrt{\mathcal{J}}} \frac{d\sqrt{\mathcal{J}}}{d\tau} = \frac{1}{2} \Delta S(\xi(\tau, y)), \quad (9.44)$$

which is in agreement with the result we obtained in section 7.3. Energy conservation makes the factor $1/\sqrt{\mathcal{J}}$ in the solution plausible.

As we explained in the previous section, we can construct the solution from the isotropic submanifold

$$\tilde{F}_{x_0} = \{(x, p) \mid x = x_0, H(x_0, p) = 0\}.$$

We obtain a function

$$G(x, x_0) = A(x, x_0) e^{ikS(x, x_0)}, \quad (9.45)$$

which for $x \neq x_0$ solves the equation

$$H\left(x, -\frac{i}{k}\partial\right) G(x, x_0) = 0 \quad (9.46)$$

up to terms of order $1/k^2$ and which has a well-defined singularity at $x = x_0$.

These relations will now be studied in more detail for the case of Helmholtz's equation:

$$-\left(\frac{1}{2k^2}\Delta + \frac{n^2}{2}\right)u = 0.$$

For a constant refractive index, $n(\mathbf{x}) \equiv 1$, the light rays passing through $\mathbf{x}_0$ are given by

$$\xi(\tau, \mathbf{p}) = \mathbf{x}_0 + \tau \mathbf{p}, \quad \pi(\tau, \mathbf{p}) = \mathbf{p}, \quad \mathbf{p}^2 = 1, \quad (9.47)$$

that is, $S(\mathbf{x}, \mathbf{x}_0) = |\mathbf{x} - \mathbf{x}_0|$. For $\tau \neq 0$ the mapping $\xi(\tau, \mathbf{p})$ is a diffeomorphism and $\mathcal{J}(\tau(\mathbf{x}), \mathbf{p}(\mathbf{x})) = \text{const.} \times (\mathbf{x} - \mathbf{x}_0)^2$. So we find

$$G_0(\mathbf{x}, \mathbf{x}_0) = \frac{1}{|\mathbf{x} - \mathbf{x}_0|} e^{ik|\mathbf{x} - \mathbf{x}_0|}, \quad (9.48)$$

which is even defined globally as long as $\mathbf{x} \neq \mathbf{x}_0$ and which, for $\mathbf{x} \neq \mathbf{x}_0$, is an exact solution of Helmholtz's equation (not only up to terms of order $\mathcal{O}(1/k^2)$). Taking into account the singularity at $\mathbf{x} = \mathbf{x}_0$, one finds

$$(\Delta + k^2)G_0(\mathbf{x}, \mathbf{x}_0) = -4\pi\delta(\mathbf{x} - \mathbf{x}_0). \quad (9.49)$$

Hence, $G(\mathbf{x}, \mathbf{x}_0)$ is the Green's function for Helmholtz's equation.

If $n(\mathbf{x}) \neq \text{const.}$, we find similar relations. Close to $\mathbf{x} = \mathbf{x}_0$ one obtains $S(\mathbf{x}, \mathbf{x}_0) = n(\mathbf{x}_0)|\mathbf{x} - \mathbf{x}_0| + \text{regular terms}$, and $\sqrt{\mathcal{J}(\tau(\mathbf{x}), \mathbf{p}(\mathbf{x}))} = \text{const.} \times |\mathbf{x} - \mathbf{x}_0| + \text{regular terms}$. Therefore, at $\mathbf{x} \approx \mathbf{x}_0$, the function

$$G(\mathbf{x}, \mathbf{x}_0) = A(\mathbf{x}, \mathbf{x}_0) e^{ikS(\mathbf{x}, \mathbf{x}_0)} \quad (9.50)$$

agrees with the “free Green's function” $G_0(\mathbf{x}, \mathbf{x}_0)$ up to regular terms and is (at least locally and up to non-leading terms in k) the Green's function for Helmholtz's equation, which in the vicinity of $\mathbf{x} = \mathbf{x}_0$ resembles an outgoing spherical wave.

At this point we will not proceed with a discussion of the equations for the quadratic and higher orders in $1/k$, which one obtains by inserting the ansatz (9.3) into the wave equation (9.2). The initial value problem for these equations can also be solved by integration along the characteristics $\xi(\tau, y)$, $\pi(\tau, y)$.

9.4 Focal points and caustics

We now turn to a discussion of the situation in which the mapping

$$\xi : (\tau, y) \longmapsto \xi(\tau, y) \in \mathbb{R}^n, \quad (9.51)$$

constructed from the solutions of the Hamiltonian system, is no longer a diffeomorphism. In this case, the solution curves $(\xi(\tau, y), \pi(\tau, y))$ were defined by the condition that their initial points at $\tau = 0$ are located in an $(n - 1)$ -dimensional isotropic submanifold of phase space, $\tilde{F} \subset P$. We have already mentioned two special cases:

- (i) We start from an $(n - 1)$ -dimensional surface of constant phase in space, $F = \{x \mid S_0(x) = 0\}$, and construct the following $(n - 1)$ -dimensional surface in phase space:

$$\tilde{F} = \left\{ (x(y), p(y)) \left| S_0(x(y)) = 0, p(y) = \frac{\partial S_0}{\partial x}(x(y)) \right. \right\}, \quad (9.52)$$

where we require $H(x(y), p(y)) = 0$. Here $x(y)$ is a point of the $(n - 1)$ -dimensional surface $F = \{x \mid S_0(x) = 0\}$, and $p(y)$ points into the normal direction of F .

(ii) In this second special case:

$$\tilde{F} = \tilde{F}_{x_0} = \{(x(y), p(y)) \mid x(y) = x_0, H(x(y), p(y)) = 0\}. \quad (9.53)$$

The initial point in space is fixed.

Those points $x = \xi(\tau, y)$ for which

$$\mathcal{J}(\tau, y) = \det D\xi(\tau, y) = 0 \quad (9.54)$$

are called *focal points* of $\tilde{F}$, and the set of all focal points, i.e. the set Σ of all critical values of ξ , is called the *caustic* of $\tilde{F}$.

If $x = \xi(\tau, y)$ is not a focal point, we know from the implicit function theorem that the mapping ξ is a diffeomorphism from at least a small neighborhood of the parameter point (τ, y) into a small neighborhood of x .

On the other hand, a focal point is characterized as the intersection point of neighboring trajectories $\xi(\tau, y)$ with infinitesimally different values of y . Neighboring initial points are mapped onto the same point. However, the mapping $(\xi(0, y), \pi(0, y)) \mapsto (\xi(\tau, y), \pi(\tau, y))$ is a canonical transformation, because it describes the time evolution of a Hamiltonian system, and according to Liouville's theorem such a mapping is volume-preserving in phase space. Therefore, a contraction in position space must be accompanied by a stretching in momentum space. In particular, if the Jacobian $\mathcal{J}(\tau, y)$ vanishes at a critical point, the directions of the different intersecting trajectories have to diverge. But as $\pi(\tau, y)$ is perpendicular to the wave fronts, the different wave fronts will exhibit an irregular behavior at a focal point.

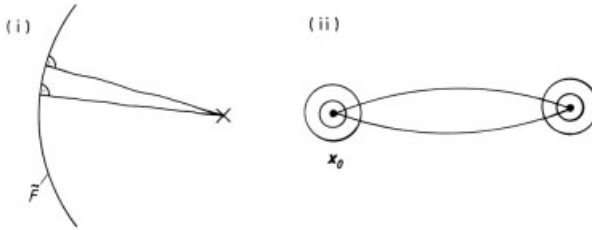


Figure 9.3: A focal point for a surface F and for a point x_0 .

Fig. 9.3 sketches what happens in the above-mentioned cases (i) and (ii). In case (i), the focal points of $\tilde{F}$ are also referred to as the focal points of F , and in case (ii), one speaks of the focal points of point x_0 , which coincides with our convention of section 8.1.

The *order* of a focal point $x = \xi(\tau, y)$ is defined to be $n - r$, where $r = \text{rank } D\xi(\tau, y)$ is the rank of the Jacobian matrix. Finally, for the curve segment for fixed y ,

$$\mathcal{C} = \{\xi(\tau, y) \mid 0 \leq \tau \leq \tau_0\},$$

we define the *Morse index* $\mu(\mathcal{C})$ to be the number of focal points on $\mathcal{C}$ multiplied by their rank. Here we assume that the end-point $\xi(\tau_0, y)$ is not focal.

Now we want to discuss the role of focal points and caustics in geometrical optics in more detail. For a fixed value of y , the paths $\xi(\tau, y)$ are geodesics with respect to the *light metric*

$dl^2 = n^2 ds^2$. Furthermore, we have $\dot{\xi} = \pi$, and according to eq. (9.19) the surfaces of constant phase are perpendicular to the rays. Focal points are convergence points of neighboring light rays. *Focal points of lenses* are focal points of order two.

Let $\mathcal{C}$ be a geodesic from x_0 to x , and such that x is not a focal point for x_0 . If $\mu(\mathcal{C}) = 0$, the path $\mathcal{C}$ is the shortest path from x_0 to x . This is no longer the case for $\mu(\mathcal{C}) > 0$. For $\mu(\mathcal{C}) = 0$, the second variation $l''(\mathcal{C})$ of the arc length is a positive definite quadratic functional; in general, $\mu(\mathcal{C})$ is just the dimension of the space where $l''(\mathcal{C})$ is negative definite. In a way, one may interpret $\mu(\mathcal{C})$ as the number of independent short-cuts for the geodesic $\mathcal{C}$. These relations form the basics of the so-called *Morse theory*, an interesting mathematical area combining geometry and variational calculus.

Now, let $\bar{F}$ be of the form (9.52). The fundamental principles of geometrical optics are most obvious when the refractive index, $n(x)$, is constant: $n(x) \equiv 1$. In this case, Hamilton's equations read

$$\dot{\xi} = \pi, \quad \dot{\pi} = 0, \quad \text{i.e.} \quad \ddot{\xi} = 0 \quad (9.55)$$

(see fig. 9.4). The light rays are straight lines perpendicular to the surface F and can be written as

$$\xi(\tau, y) = x(y) + N(y)\tau, \quad (9.56)$$

where $x(y) \in F$, and $N(y)$ is the unit normal vector of F at point $x(y)$.

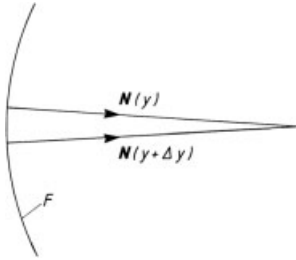


Figure 9.4: Focal points for a surface F for $n = 1$.

The Jacobian matrix $D\xi(\tau, y)$ has the form

$$D\xi(\tau, y) = \left(N, \frac{\partial x}{\partial y^1} + \tau \frac{\partial N}{\partial y^1}, \frac{\partial x}{\partial y^2} + \tau \frac{\partial N}{\partial y^2} \right). \quad (9.57)$$

The zeros of the Jacobian determinant are most easily determined by multiplying $D\xi(\tau, y)$ by the matrix

$$A = \left(N, \frac{\partial x}{\partial y^1}, \frac{\partial x}{\partial y^2} \right). \quad (9.58)$$

Here $\partial x / \partial y^1$ and $\partial x / \partial y^2$ are tangent vectors for F and orthogonal to N . So we know that $\det A \neq 0$ whenever y^1 and y^2 are a parametrization of F . We may even assume A to be

positive definite. Since $\mathbf{N} \cdot (\partial/\partial y^i)\mathbf{x} = 0$ and $\mathbf{N} \cdot (\partial/\partial y_i)\mathbf{N} = \frac{1}{2}(\partial/\partial y^i)\mathbf{N}^2 = 0$, the matrix $AD\xi$ has the form

$$\left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & \frac{\partial \mathbf{x}}{\partial y^i} \cdot \frac{\partial \mathbf{x}}{\partial y^k} + \tau \frac{\partial \mathbf{x}}{\partial y^i} \cdot \frac{\partial \mathbf{N}}{\partial y^k} \\ 0 & \end{array} \right). \quad (9.59)$$

Hence, we have reduced the problem to an investigation of the (2×2) matrix:

$$D_{ik} = \frac{\partial \mathbf{x}}{\partial y^i} \cdot \frac{\partial \mathbf{x}}{\partial y^k} - \tau \mathbf{N} \cdot \frac{\partial^2 \mathbf{x}}{\partial y^i \partial y^k} = g_{ik} - \tau L_{ik}, \quad (9.60)$$

where we have used $0 = (\partial/\partial y^k)[\mathbf{N} \cdot (\partial/\partial y^i)\mathbf{x}]$. In eq. (9.60), g_{ik} is positive definite and determines the metric on F ; it is called the *first fundamental form* of F ; and L_{ik} is called the *second fundamental form* of F .

By a suitable choice of parameters y^i , we can arrange that, for any given point, $g_{ik}(y) = \delta_{ik}$; an additional orthogonal transformation will make L_{ik} diagonal. The two diagonal elements of L_{ik} are just the two *principal curvatures*, $1/R_1$ and $1/R_2$, of the surface:

$$D_{ik} = \begin{pmatrix} 1 - (\tau/R_1) & 0 \\ 0 & 1 - (\tau/R_2) \end{pmatrix}. \quad (9.61)$$

Note that the focal points of F are just the centers of curvature of F . Therefore, the Morse index of a light ray $\mathcal{C}$ emanating from a point $\mathbf{x}(y) \in F$ into a direction perpendicular to F is just equal to the number of centers of curvature of the area segment around $\mathbf{x}(y)$ on $\mathcal{C}$: $0 \leq \mu(\mathcal{C}) \leq 2$.

Qualitatively, the situation does not change much if $n(\mathbf{x})$ is not constant and the light rays are no longer straight lines. The “centers of curvature” are merely determined by the geodesics with respect to a different metric. In general, however, a light ray may pass through several caustics, i.e. the case $\mu(\mathcal{C}) > 2$ is also possible.

9.5 Behavior of phases in the vicinity of caustics

Looking at the local solution,

$$u(\mathbf{x}) = \frac{\sqrt{\mathcal{J}(0, y(x))}}{\sqrt{\mathcal{J}(\tau(x), y(x))}} \varphi_0(\xi(0, y(x))) e^{ikS(x)} + \mathcal{O}\left(\frac{1}{k^2}\right), \quad (9.62)$$

it is obvious that the amplitude gets very large close to a caustic, because $\mathcal{J}(\tau(x), y(x))$ vanishes there. From the equation derived in section 9.3,

$$\frac{1}{\mathcal{J}} \frac{d\mathcal{J}}{d\tau} = \frac{\partial^2 H}{\partial p_i \partial x^i}(\xi, \partial S(\xi)) + \frac{\partial^2 H}{\partial p_i \partial p_j}(\xi, \partial S(\xi)) \frac{\partial^2 S(\xi)}{\partial x^i \partial x^j}, \quad (9.63)$$

which simplifies within the framework of geometrical optics to

$$\frac{1}{\mathcal{J}} \frac{d\mathcal{J}}{d\tau} = \frac{\partial^2 S}{\partial x^i \partial x^i} = \Delta S, \quad (9.64)$$

we see that because of $\pi(\tau, y) = (\partial S / \partial x)(\xi(\tau, y))$, the term $(\partial \pi_i / \partial x^i)(\xi(\tau, y))$ approaches infinity at the caustics Σ : At non-focal points x , the direction of the ray π is determined by x . At focal points x_0 , rays converge from different directions such that close to x_0 the derivative $(\partial \pi / \partial x)$ becomes very large and approaches infinity at the focal point. This leads to a diverging amplitude.

The expression for $u(x)$, given in eq. (9.62), is no longer valid on a caustic. However, we can learn from this expression that there will be a rapid change of the phase – in the limit $k \rightarrow \infty$ even a jump – due to the change of sign of $\mathcal{J}$ when a caustic is crossed. At a focal point of order r , the determinant $\mathcal{J}$ has an r -fold zero so that one expects for $\mathcal{J}^{-1/2}$ a phase factor $(-i)^r = \exp(-i\pi r/2)$. This phase factor has to be taken into account when the point x on a ray $\mathcal{C}$ lies beyond the first focal point.



Figure 9.5: There may be several light rays from F to x .

Furthermore, in general, there may be more than one ray emanating from the initial manifold F to the point x (fig. 9.5). In this case, the local solution $u(x)$ generalizes to

$$u(x) = \sum_j \frac{\sqrt{|\mathcal{J}(0, y_j(x))|}}{\sqrt{|\mathcal{J}(\tau_j(x), y_j(x))|}} e^{(-i\pi/2)\mu_j} \varphi_0(\xi(0, y_j(x))) e^{ikS_j(x)} + \mathcal{O}\left(\frac{1}{k^2}\right). \quad (9.65)$$

Here the sum runs over all solutions $(\xi(\tau, y_j), \pi(\tau, y_j))$ of the Hamiltonian system, for which

$$(\xi(0, y_j), \pi(0, y_j)) \in \tilde{F}$$

are located on the isotropic initial manifold and for which

$$\xi(\tau_j, y_j) = x.$$

In eq. (9.65), μ_j denotes the Morse index of the curve segment

$$\mathcal{C}_j = \{\xi(\tau, y_j) \mid 0 \leq \tau \leq \tau_j\},$$

and

$$S_j(x) = \int_0^{\tau_j} d\tau' \pi(\tau', y_j) \dot{\xi}(\tau', y_j)$$

is just the action along $\mathcal{C}_j$. This expression for $u(x)$ is valid whenever x does not lie exactly in the caustic for $\tilde{F}$.

The change of phase when a light ray traverses through a caustic was first observed by L. G. Gouy in 1890. In Fresnel's mirror experiment, a point-like source P illuminates two

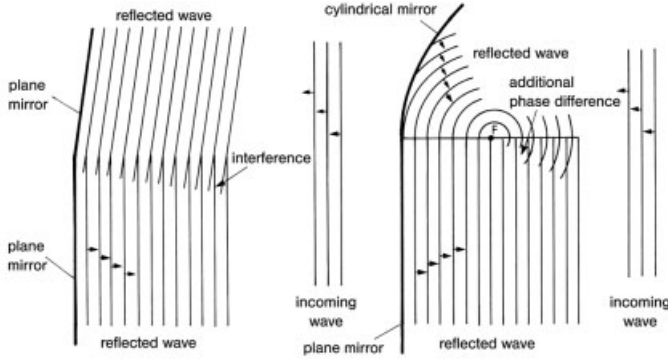


Figure 9.6: Sketch of the experiment of Gouy.

mirrors, which are inclined to each other at a very small angle (see fig. 9.6). One observes two close-by mirror images, P_1 and P_2 , of the point source, together with a system of interference fringes resembling the pattern of interfering spherical waves emanating from P_1 and P_2 . If one of the mirrors is replaced by a cylindrical mirror whose axis is parallel to the cut line between the two mirrors, the interference pattern is shifted in exactly the way one would expect from a phase shift of a plane wave of $\pi/2$, depending on whether one looks at the pattern in front of or behind the focal line of the mirror. If the second mirror is a spherical mirror, the change in the interference pattern corresponds to a phase shift of π .

9.6 Caustics, Lagrangian submanifolds and Maslov index

We will see that the geometrical interpretation of the wave equation involving the phase space P , the isotropic $(n-1)$ -dimensional initial manifold

$$\tilde{F} = \{x(y), p(y)\} = \{(\xi(0, y), \pi(0, y))\} \subset P,$$

the isotropic $(n-1)$ -dimensional manifolds

$$\tilde{F}_{\tau_0} = \{(\xi(\tau, y), \pi(\tau, y)) \mid \tau = \tau_0\} \subset P,$$

and the Lagrangian manifold

$$\Lambda_{\tilde{F}} = \{(\xi(\tau, y), \pi(\tau, y))\} \subset P$$

is particularly suited for the description of the caustics of $\tilde{F}$.

If $\mathcal{J}(\tau, y) = \det D\xi(\tau, y) \neq 0$, i.e. if $x = \xi(\tau, y)$ does not lie on the caustic of $\tilde{F}$, the transformation $(\tau, y) \mapsto \xi(\tau, y)$ is a local diffeomorphism, and instead of (τ, y) we may as well choose $x = \xi(\tau, y)$ to be the local coordinates on the Lagrangian manifold $\Lambda_{\tilde{F}}$. In other words, if $x = \xi(\tau, y)$ does not lie on the caustic, the projection

$$\text{pr} : \Lambda_{\tilde{F}} \longrightarrow \mathbb{R}^n, \quad (\xi(\tau, y), \pi(\tau, y)) \longmapsto \xi(\tau, y) \quad (9.66)$$

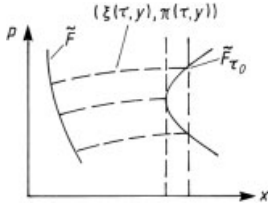


Figure 9.7: Phase space description of caustics.

of $\Lambda_{\tilde{F}}$ onto the position space $\mathbb{R}^n$ is a local diffeomorphism. Looking at the problem this way, the caustic $\Sigma \subset \mathbb{R}^n$ turns out to be the set of critical values of the projection pr .

The set $\tilde{\Sigma} \subset \Lambda_{\tilde{F}}$ of points $(\xi(\tau, y), \pi(\tau, y))$ for which $\mathcal{J}(\tau, y) = 0$, and for which ξ is not a local diffeomorphism, is exactly the set of critical points of pr , and the caustic Σ results as a projection from $\tilde{\Sigma} : \text{pr}(\tilde{\Sigma}) = \Sigma$ (see fig. 9.7). If

$$(x_0(\tau, y), \pi_0(\tau, y)) = (x_0, p_0) = (x_0^1, \dots, x_0^n, p_1^0, \dots, p_n^0) \in \tilde{\Sigma} \quad (9.67)$$

is a critical point of pr , the components of x are no longer permissible as coordinates on $\Lambda_{\tilde{F}}$. For x_0 a focal point of order one, there will exist local coordinates for a neighborhood of $(x_0, p_0) \in \Lambda_{\tilde{F}}$ that are of the form $(p_1, x^2, \dots, x^n)$, and x_1 will be a function of these coordinates, $x^1 = x^1(p_1, x^2, \dots, x^n)$, such that $(\partial x^1 / \partial p_1)(p_1^0, \dots, p_n^0) = 0$. Depending on the sign of $(\partial x^1 / \partial p_1)(p_1^0, x^2, \dots, x_n^0)$, it is, at least locally, possible to say when a point $(x, p) \in \Lambda_{\tilde{F}} \setminus \tilde{\Sigma}$ is on the “positive” or “negative” side of $\Lambda_{\tilde{F}}$.

For generic $\Lambda_{\tilde{F}}$, the following general result due to Maslov holds:

- The set of critical points $\tilde{\Sigma} \in \Lambda_{\tilde{F}}$ consists of $(n - 1)$ -dimensional submanifolds and boundary components of dimension $(n - 3)$, $(n - 4)$, $\dots$. There is no component of dimension $(n - 2)$.
- $\tilde{\Sigma}$ divides $\Lambda_{\tilde{F}}$ into a positive and a negative part.

Let $\tilde{\gamma}$ now be a path in $\Lambda_{\tilde{F}}$ with initial point (x_1, p_1) and final point (x_2, p_2) , which either does not intersect $\tilde{\Sigma}$ at all or which intersects it transversely, i.e. with a non-vanishing angle (see fig. 9.8). For each point (x^α, y^α) in the intersection of $\tilde{\gamma}$ with $\tilde{\Sigma}$, we define $n_\alpha(\tilde{\gamma}) = \pm 1$, depending on whether $\tilde{\gamma}$ traverses the critical subset $\tilde{\Sigma}$ in (x^α, p^α) from the negative to the positive side of $\Lambda_{\tilde{F}}$ or vice versa. The *Maslov index* $\tilde{\mu}(\tilde{\gamma})$ of the path $\tilde{\gamma}$ is now defined by

$$\tilde{\mu}(\tilde{\gamma}) = \sum_{\alpha} n_{\alpha}(\tilde{\gamma}), \quad (9.68)$$

where the sum runs over all intersection points of $\tilde{\gamma}$ with $\tilde{\Sigma}$.

It seems plausible, and can indeed be proven, that $n(\tilde{\gamma})$ does not change under continuous deformations of $\tilde{\gamma}$ as long as the initial and final points remain fixed. In defining the Maslov index, one can even get rid of the transversality requirement for the intersections of $\tilde{\gamma}$ with $\tilde{\Sigma}$.

In particular, let $\tilde{\gamma}$ be of the form

$$\tilde{\gamma} = \{(\xi(\tau, y), \pi(\tau, y)) \mid 0 \leq \tau \leq \tau_0, y = y_0\},$$

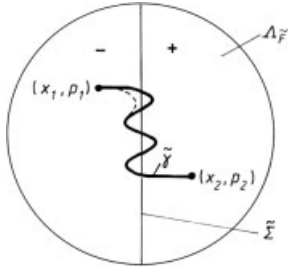


Figure 9.8: Maslov index of a path $\tilde{\gamma}$. For the depicted case we have $\mu(\tilde{\gamma}) = 1$.

then the projection

$$\gamma = \text{pr } \tilde{\gamma} = \{\xi(\tau, y) \mid 0 \leq \tau \leq \tau_0, y = y_0\}$$

is a “light ray” in $\mathbb{R}^n$, and for positive definite $(\partial^2 H / \partial p_i \partial p_j)$ we have

$$\tilde{\mu}(\tilde{\gamma}) = \mu(\gamma).$$

In this case, the Maslov index of $\tilde{\gamma}$ is equal to the Morse index of γ . However, a general path in Λ_F cannot be projected onto light rays in $\mathbb{R}^n$, which makes the Maslov index an important generalization of the Morse index.

The expression for the solution $u(x)$, which we presented in section 9.5, becomes ill-defined for $x \in \Sigma$. The geometrical theory of caustics, which we have just discussed, helps in the explicit construction of the solutions close to caustics. The strategy is briefly described below:

1. One chooses a suitable Fourier transform such that one can represent the wave equation $H(x, -(i/k)\nabla)u(x)$ partly in position and partly in momentum variables, i.e. using the variables

$$(p_1, x^2, \dots, x^n).$$

2. In the new variables, one makes a short-wave expansion,

$$u(p_1, x^2, \dots, x^n) = \varphi(p_1, x^2, \dots, x^n) e^{ik\hat{S}(p_1, x^2, \dots, x^n)},$$

and calculates $\hat{S}$ and φ up to order $1/k$.

3. By performing the inverse Fourier transformation, one can represent the solution in position space.

For Lagrangian manifolds, Arnol'd has determined normal forms for $\tilde{\Sigma}$ and Σ when they can be represented, up to equivalence, as polynomials of order ≤ 6 . For these cases, the strategy just outlined yields standard integral expressions for $u(x)$ in the vicinity of a caustic.

9.7 Supplementary remarks on geometrical short-wave asymptotics

Up to now we have kept the presentation of the geometrical optics of short-wave asymptotics and the geometry of caustics as elementary as possible and have deliberately made no use of elaborate concepts from differential geometry. For the interested reader, in this section we want at least to sketch the differential geometric formulation and generalization of the results obtained up to now. Some supplementary remarks concerning the geometric short-wave asymptotics will follow.

Concerning the mathematical tools, the interested reader is referred to textbooks on differential geometry. However, we should emphasize that this section as well as sections 10.8 and 10.9 of the following chapter will not be needed for the rest of this book.

In the preceding sections we proceeded from the phase space

$$P = \{(x, p) \mid x, p \in \mathbb{R}^n\} = \mathbb{R}^{2n}$$

on which the symplectic form

$$\omega = \sum_{i=1}^n dx^i \wedge dp_i \quad (9.69)$$

is defined. The generalizations of these “flat” phase spaces are symplectic manifolds:

Definition 9.1. A symplectic manifold (P, ω) is a $(C^\infty\text{-})$ manifold P , on which a closed and nowhere degenerate 2-form ω is defined. That is, $d\omega = 0$, and the antisymmetric bilinear form $\omega(m)$ on the tangent space $T_m P$ of a point m is non-degenerate for all $m \in P$.

Symplectic manifolds always have even dimensions.

Definition 9.2. A smooth bijection $\Phi : P \mapsto P$ is called a symplectic transformation or canonical transformation if ω is kept invariant (i.e. if $\Phi^* \omega = \omega$).

Already in chapter 8 we mentioned the following theorem:

Theorem 9.2 (Darboux’s theorem). Every point of a $2n$ -dimensional symplectic manifold (P, ω) has a neighborhood U on which one can choose coordinates x^i, p_i such that the symplectic 2-form ω assumes the form (9.69).

Because of Darboux’s theorem, the results obtained for $P = \mathbb{R}^{2n}$ hold locally for any symplectic manifold. The notion of a Lagrangian submanifold (see section 9.2) also applies to arbitrary symplectic manifolds.

If we want to generalize the short-wave asymptotics for linear differential operators from flat space $\mathbb{R}^n$ to an arbitrary manifold Q , or if we want to consider mechanics on an arbitrary configuration space Q , the corresponding phase space is just the *cotangent bundle* T^*Q of the manifold Q .

The cotangent space T_q^*Q of Q at a point $q \in Q$ is the dual space of the tangent space $T_q Q$. If we denote by x^i the local coordinates on Q , the differentials dx^i provide a basis for all cotangent spaces in the validity range of the coordinates.

So, we obtain coordinates on T^*Q by complementing the coordinates $x^i(q)$ of a point q by the components p^i of the cotangent vectors $\alpha = \sum_{i=1}^n p^i dx^i(q)$ in q .

The projection

$$\pi : T^*Q \longrightarrow Q \quad (9.70)$$

is defined in such a way that each cotangent vector at a point q is associated to its base point q . Using coordinates x and p , we simply have

$$\pi(x, p) = x. \quad (9.71)$$

On T^*Q is defined a canonical 1-form θ which in coordinates x and p reads

$$\theta = \sum_i p_i dx^i. \quad (9.72)$$

Independent of coordinates, we can define the 1-form θ by the following identity, which should hold for every $(q, \alpha) \in T^*Q$ and every $v \in TT^*Q$:

$$\theta_{(q, \alpha)}(v) = \alpha(\pi_* v). \quad (9.73)$$

In a natural way we can make the cotangent bundle T^*Q into a symplectic manifold by introducing the closed and nowhere degenerate 2-form

$$\omega = -d\theta. \quad (9.74)$$

The coordinates x and p turn out to be Darboux coordinates on T^*Q because

$$\omega = -d\theta = -d \sum_{i=1}^n p_i dx^i = \sum_{i=1}^n dx^i \wedge dp_i. \quad (9.75)$$

In general, (P, ω) is called *exact symplectic* if ω is not only closed but even exact, i.e. if a 1-form θ exists such that $\omega = -d\theta$. Hence, in a natural way cotangent bundles are exact symplectic manifolds. The symplectic manifolds we will encounter in this book will always be cotangent bundles.

The asymptotic expansion of solutions of linear partial differential equations on manifolds Q can always be formulated on $P = T^*Q$. The symbol function $H(x, p)$, defined in section 7.1, now becomes a function on T^*Q . If $S : Q \rightarrow \mathbb{R}$ is a function on Q , the 1-form

$$dS : Q \longrightarrow T^*Q \quad (9.76)$$

is a function with values in T^*Q . Writing the characteristic equation (9.8) in the form

$$H \circ dS = 0 \quad (9.77)$$

also leads to an interpretation on T^*Q .

Hamilton's equations,

$$\dot{x}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x^i}, \quad (9.78)$$

are the equations of motion on T^*Q . As we described in section 7.2, one obtains a Lagrangian submanifold $\Lambda \subset T^*Q$ by starting from an isotropic $(n-1)$ -dimensional submanifold $\tilde{F} \subset H^{-1}(0)$ and solving Hamilton's equations (9.78). The only requirement needed is that the solution curves of eq. (9.78) leave $\tilde{F}$, i.e. that the vector field

$$X_H = \left(\frac{\partial H}{\partial p}, -\frac{\partial H}{\partial q} \right) \quad (9.79)$$

is transverse with respect to $\tilde{F}$. As in section 9.6, $\tilde{\Sigma}$ is the set of critical points of the projection $\pi : \Lambda \rightarrow Q$, and the caustic $\Sigma = \pi(\tilde{\Sigma})$ is the set of critical values of π . Away from the critical points, Λ has the form

$$\Lambda = \left\{ (x, p) \left| p = \frac{\partial S}{\partial x} \right. \right\} \subset T^*Q, \quad (9.80)$$

where S is a solution of eq. (9.77). S can also be generalized to a function $\tilde{S}$ on Λ where

$$S \circ \pi = \tilde{S}. \quad (9.81)$$

Even the amplitude φ_0 in the short-wave expansion of the solutions of the partial differential equation may be reinterpreted as a function $\tilde{\varphi}_0$ on the Lagrangian manifold $\Lambda \subset T^*Q$.

Because of eq. (9.39), the transport equation (9.37) appears in this view as the projection of an equation on Λ . For given initial values for $\tilde{\varphi}_0$ on the isotropic $(n-1)$ -dimensional submanifold $\tilde{F} \subset H^{-1}(0) \cap T^*Q$, one obtains $\tilde{\varphi}_0$ on Λ along the solution of Hamilton's equations (9.78).

Therefore, the entire short-wave expansion “lives” on Λ rather than on Q . In the sequel of eq. (9.41), it turned out to be convenient to extract from φ_0 the square root of the Jacobian determinant $\mathcal{J}$. Using a coordinate-free geometrical language, this suggests the use of so-called *half-densities* on Λ . This means the following: We first define an r -density on an n -dimensional vector space V as a complex-valued function $\rho : \Lambda^n V \mapsto \mathbb{C}$ on the n th exterior power of V , having the homogeneity property

$$\rho(tv) = |t|^r \rho(v) \quad (9.82)$$

for $v \in \Lambda^n V$. Because $\Lambda^n V$ is one-dimensional, the r -densities form a one-dimensional vector space $\Omega_r(V)$. By definition, half-densities on V have $r = 1/2$. Let M be an n -dimensional manifold, then the half-density bundle

$$\Omega_{n/2}(M) = \bigcup_{m \in M} \Omega_{n/2}(T_m M) \quad (9.83)$$

is a complex line bundle on M with obvious topology. Half-densities on M are sections of the bundle $\Omega_{n/2}(M)$; they may be interpreted as square roots of n -forms on M .

These considerations suggest the interpretation of the asymptotic expansion of the solution as a half-density on Λ . However, this is not sufficient, because the phases that occur in the

expansion (9.65) and are due to crossings of the critical set $\tilde{\Sigma}$ or the caustics Σ have not yet been taken into account.

It turns out that we have to introduce an additional line bundle on Λ , the so-called Maslov bundle $M(\Lambda)$. A Maslov bundle is a vector bundle over Λ with a complex one-dimensional fiber $\mathbb{C}$. However, this time the structure group is $\mathbb{Z}_4$ and it contains the information about the additional phases.

So, the asymptotic solution is a section in the tensor product $\Omega_{n/2}(\Lambda) \otimes M(\Lambda)$ of the bundles of half-densities and the Maslov bundle over Λ . The asymptotically leading part of the so-defined solution is called the *symbol* of the solution $u(x)$. The relation of this definition with the notion of a symbol of differential operators will be explained in section 10.9.

We want to sketch briefly the definition of a Maslov bundle $M(\Lambda)$ and, in this context, comment on a second way to introduce the Maslov index $\tilde{\mu}(\tilde{\gamma})$ of a path $\tilde{\gamma}$ in Λ . Additional information may be found in the literature for this chapter, in particular in the book and article of Guillemin and Sternberg and in the thesis of Meinrenken. First, however, we have to explain briefly the invariants of Lagrangian subspaces $L \subset V$ of a symplectic vector space $(V, \mathcal{J})$.

Any two Lagrangian subspaces $L, L' \subset V$ can be transformed into each other by a symplectic transformation. A pair of Lagrangian subspaces $L_1, L_2 \subset V$ can be transformed into a second pair $L'_1, L'_2 \subset V$ only if $\dim(L_1 \cap L_2) = \dim(L'_1 \cap L'_2)$. For a triple of Lagrangian subspaces L_1, L_2, L_3 , there exists exactly one more invariant apart from the four dimensions of the intersections $L_1 \cap L_2$, $L_1 \cap L_3$, $L_2 \cap L_3$, and $L_1 \cap L_2 \cap L_3$. This invariant is the signature $s(L_1, L_2, L_3)$ of the quadratic form B defined on the exterior direct sum $L_1 \oplus L_2 \oplus L_3$ by

$$B(x_1, x_2, x_3) = \mathcal{J}(x_1, x_2) + \mathcal{J}(x_2, x_3) + \mathcal{J}(x_3, x_1). \quad (9.84)$$

The signature $s(L_1, L_2, L_3)$ is locally constant and depends piecewise continuously on L_1 , L_2 , and L_3 . Now, let $\Lambda \subset T^*Q$ be a Lagrangian submanifold. At each point $m \in \Lambda$ we consider three Lagrangian subspaces of the corresponding tangent space $T_m T^*Q$ of T^*Q :

1. L_{1m} is the space of vertical tangent vectors at T^*Q in $m \in T^*Q$. By a simple identification we find $L_{1m} = T_{\pi(m)}^*Q$, where $\pi : T^*Q \rightarrow Q$ is the projection of the bundle T^*Q .
2. $L_{2m} = T_m \Lambda$ is the tangent space at Λ in m .
3. L_{3m} is an auxiliary space transverse to L_{1m} and L_{2m} : $L_{3m} \cap L_{1m} = L_{3m} \cap L_{2m} = \{0\}$.

Let $\mathcal{L}_{3m}$ be the set of suitable auxiliary spaces at point $m \in T^*Q$, i.e. the set of Lagrangian subspaces of $T_m T^*Q$ that are transverse to L_{1m} and L_{2m} .

For a sufficiently small open neighborhood $\mathcal{U} \subset \Lambda$, we can choose the mapping $m \mapsto L_{3m}$ to be continuous. Furthermore, there exists a covering $\mathcal{A}$ of Λ of such neighborhoods. The fiber $M(m)$ of the Maslov bundle $M(\Lambda)$ in $m \in \Lambda$ is defined as follows: $M(m)$ is the complex one-dimensional vector space of all mappings

$$\psi : \mathcal{L}_{3m} \longrightarrow \mathbb{C}$$

with the property

$$\psi(L'_{3m}) = \exp\left\{\frac{1}{4}i\pi[s(L_{1m}, L_{2m}, L_{3m}) - s(L_{1m}, L_{2m}, L'_{3m})]\right\}\psi(L_{3m}) \quad (9.85)$$

for all $L'_{3m}, L_{3m} \in \mathcal{L}_{3m}$. Indeed, the properties of s guarantee that this defines a complex line bundle over Λ with structure group $\mathbb{Z}_4$.

In order to define the Maslov index for a path $\tilde{\gamma}$ in Λ ,

$$\tilde{\gamma} : [0, 1] \longrightarrow \Lambda,$$

we proceed from the above-mentioned covering $\mathcal{A}$ of Λ . By choosing sufficiently many intermediate points,

$$0 = t_0 < t_1 < \cdots < t_k = 1,$$

we can always arrange that

$$\tilde{\gamma}([t_{i-1}, t_i]) \subset U_i$$

lies within the covering $\mathcal{A}$ for a suitable U_i .

The Maslov index $\tilde{\mu}(\tilde{\gamma})$ is now given by

$$\mu(\tilde{\gamma}) = -\frac{1}{2} \sum_{i=1}^k [s(L_{1m}, L_{2m}, L_{3m}^{U_i})|_{m=\tilde{\gamma}(t_i)} - s(L_{1m}, L_{2m}, L_{3m}^{U_i})|_{m=\tilde{\gamma}(t_{i-1})}]. \quad (9.86)$$

One can easily show that this construction is well defined, i.e. that it is independent of the choice of covering and the choice of intermediate points. In section 10.8, we will describe an alternative way to define the Maslov bundle $M(\Lambda)$ and the Maslov index $\tilde{\mu}(\tilde{\gamma})$.

10 Diffraction theory

10.1 Survey

Deviations from geometrical optics related to diffraction phenomena are due to the wave-like nature of light; they are the subject of classical *diffraction theory* that we want to discuss in the present chapter.

In section 10.2, we will start with some considerations concerning the principles of Huygens and Huygens–Fresnel. The starting point will be the concept of a *wavelet* from which any wave field can be reconstructed as long as at least one single surface of constant phase is known. Huygens’ principle leads to the reconstruction of all surfaces of constant phase from a given initial surface. As we shall see, this is a consequence of the eikonal equation for the phase of a wave field. In addition, the principle of Huygens–Fresnel allows the reconstruction of the total wave field, not only the surfaces of constant phase or wave front.

In section 10.3, we will gradually develop the method of stationary phases, which is one of the fundamental tools in the treatment of wave fields with small wavelengths.

Kirchhoff’s identities for the solutions of the scalar wave equation as well as the justification of the Huygens–Fresnel principle will be given in section 10.4. In applying the method of stationary phases to the integral representation of Kirchhoff, we will recover the short-wave expansion of the wave field derived in the previous chapter.

The theory of diffraction resulting from Kirchhoff’s identities making an approximate ansatz also due to Kirchhoff will be the subject of section 10.5. The geometrical boundaries of shadows follow from Kirchhoff’s diffraction integral using the method of stationary phases. We will also describe the approximations to Kirchhoff’s theory derived by Fresnel and Fraunhofer.

Sections 10.6 to 10.8 contain applications of the diffraction theories of Fresnel and Fraunhofer. In the two final sections, 10.9 and 10.10, we will give more details about the symplectic geometrical theory of wave fields. In particular, we will discuss the fundamental notions of Morse families and Fourier integral operators.

10.2 The principles of Huygens and Fresnel

A quite illustrative and helpful concept for the description of wave fields turns out to be the notion of *wavelets*, a form of elementary waves that are solutions corresponding to a point-like

disturbance of the wave field. For a constant refractive index they are given by

$$G_0(\mathbf{x}, \mathbf{x}') = \frac{e^{ik|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}, \quad (10.1)$$

while in more general cases the Green's function

$$G(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')} \quad (10.2)$$

is to be considered as the wavelet.

Huygens' principle allows a geometrically intuitive construction of the surfaces of constant phase (see fig. 10.1). Starting from a surface F_0 of constant phase, we consider each point of F_0 as the origin of a new wavelet. All other surfaces of constant phase are now obtained as the enveloping surface of the family of phase surfaces of these wavelets. We shall show that this form of Huygens' principle indeed follows from the eikonal equation for S .

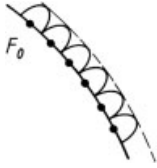


Figure 10.1: Construction of wave fronts as the envelope of wavelets according to Huygens' principle.

Consider a family of real-valued functions on $\mathbb{R}^n$,

$$f_\alpha(x) = f(x, \alpha), \quad (10.3)$$

which depend on $r < n$ parameters α . We define the envelope $F(x)$ as the unique function which, for all $x \in \mathbb{R}^n$, coincides in zeroth and first order with one of the functions in our family. Hence, for all x , the level surfaces of $F(x)$ are tangential to one of the level surfaces from our family.

The envelope function $F(x)$ has the general form

$$F(x) = f(x, \alpha(x)), \quad (10.4)$$

where $\alpha(x)$ tells us which of the functions from our family are tangential to $F(x)$. The condition of being tangential reads

$$\frac{\partial F}{\partial x}(x) = \frac{\partial f}{\partial x}(x, \alpha) \Big|_{\alpha=\alpha(x)}.$$

Because

$$\frac{\partial F}{\partial x}(x) = \frac{\partial f}{\partial x}(x, \alpha) \Big|_{\alpha=\alpha(x)} + \frac{\partial f}{\partial \alpha}(x, \alpha(x)) \frac{\partial \alpha(x)}{\partial x},$$

we find

$$\frac{\partial f}{\partial \alpha}(x, \alpha(x)) \frac{\partial \alpha}{\partial x} = 0,$$

which for the non-degenerate case implies that

$$\frac{\partial f}{\partial \alpha}(x, \alpha(x)) = 0. \quad (10.5)$$

Therefore, we can determine $\alpha(x)$ by inverting this equation.

For our purpose, the wave front F_0 will be given by $F_0 = \{\mathbf{x}_0(y)\}$ such that $S(\mathbf{x}_0(y)) \equiv S_0$. The family of wavelets with center on F_0 is now

$$S_y(\mathbf{x}) = S_0 + S(\mathbf{x}, \mathbf{x}_0(y)), \quad (10.6)$$

where y assumes the role of the family parameter. The enveloping condition reads

$$\frac{\partial S}{\partial \mathbf{x}_0}(\mathbf{x}, \mathbf{x}_0(y)) \cdot \frac{\partial \mathbf{x}_0}{\partial y} = 0, \quad (10.7)$$

i.e. $\mathbf{p}_0 = -\partial S/\partial \mathbf{x}_0$ has to be perpendicular to the surface F_0 . However, this is just the initial condition for the characteristics that we used in section 9.2 for the solution of the initial value problem of the eikonal equation:

$$\begin{aligned} S(\mathbf{x}) &= S(\mathbf{x}_0(y(\mathbf{x}))) + \int_0^{\tau(\mathbf{x})} d\tau' \boldsymbol{\pi}(\tau', y(\mathbf{x})) \cdot \frac{d}{d\tau'} \boldsymbol{\xi}(\tau', y(\mathbf{x})) \\ &= S_0 + S(\mathbf{x}, \mathbf{x}_0(y(\mathbf{x}))). \end{aligned} \quad (10.8)$$

Huygens' principle can also be justified starting from Fermat's principle. Later, Fresnel generalized Huygens' principle of wavelets by showing that it is possible not only to reconstruct the surfaces of constant phases but also to find a wave solution by superimposing all wavelets emanating from some fixed surface of constant phase. Such a construction helped Fresnel to understand why light in a homogeneous medium propagates along straight lines despite its wave-like nature. This problem had been the reason why Newton eventually rejected the wave theory of light in favor of a corpuscular theory (see chapter 1).

Let us consider a spherical wave originating from $\mathbf{x}_0$ (see fig. 10.2). The wave surfaces are spherical shells. The problem consists in explaining why an observer in $\mathbf{x}$ sees only light rays coming along the straight line connecting $\mathbf{x}$ and $\mathbf{x}_0$ but not light coming from, say, $\mathbf{y}$, although this point also emits a wavelet.

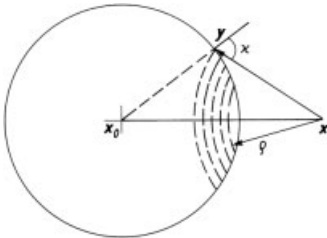


Figure 10.2: Construction of Fresnel's zones.

Fresnel solved this problem by imagining spherical shells centered around $\mathbf{x}$ with radius $r_n = n\lambda + \rho$ and looking at the intersection lines with the spherical wave surface, which in this way are decomposed into so-called *Fresnel zones* ($\lambda = 2\pi/k$ is the wavelength). All

points in a given zone Z_n have the same distance $\rho + (n-1)\lambda \leq r \leq \rho + n\lambda$ from $\mathbf{x}$ and can be considered as the origin of wavelets of the form

$$C\kappa(\chi) \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|} \quad (\mathbf{y} \in Z_n).$$

Here, C denotes an as yet unknown normalization constant and $\kappa(\chi)$ an *inclination factor* that takes into account that the strength of the spherical wave in $\mathbf{y}$ may depend on the angle χ between the normal direction of the wave surface and the direction $\mathbf{x} - \mathbf{y}$. From the inner to the outer boundary of a Fresnel zone, the phases of the wavelets change by just 2π . For small wavelengths, the zones are also small, and it becomes plausible that, by interference, the wavelets from the outer half of each zone just cancel with the wavelets from the inner half of the following zone. Just the contributions from the inner half of the first zone and the outer half of the last zone remain. Both are very small for small λ , which explains the almost straight propagation of light.

However, we do not observe the backward wave emanating from the last zone. Fresnel took this into account by assuming the inclination factor to vanish, $\kappa(\chi) = 0$, for $\chi \geq \pi/2$.

Hence, the principle of Huygens and Fresnel corresponds to the assumption of an identity of the form

$$u(\mathbf{x}) = C \int_{F_0} df(\mathbf{y}) \kappa(\chi) u(\mathbf{x}(\mathbf{y})) \frac{e^{ik|\mathbf{x}-\mathbf{x}(\mathbf{y})|}}{|\mathbf{x}-\mathbf{x}(\mathbf{y})|}, \quad (10.9)$$

where we have to integrate over some wave front F_0 .

For a test of this assumption, we calculate the integral for a spherical wave and the surface $|\mathbf{y}| = \text{const.}$, $|\mathbf{x}| > |\mathbf{y}|$. For $\kappa(\chi) \equiv 1$, the integral can be calculated explicitly by elementary methods:

$$\begin{aligned} & C \int_{|\mathbf{y}|=R} df(\mathbf{y}) \frac{e^{ik|\mathbf{y}|}}{|\mathbf{y}|} \cdot \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{|\mathbf{x}-\mathbf{y}|} \\ &= 2\pi C \frac{e^{ik|\mathbf{y}|}}{|\mathbf{y}|} |\mathbf{y}|^2 \int_{-1}^1 d\cos\chi \frac{\exp[ik(x^2 + y^2 - 2xy\cos\chi)^{1/2}]}{(x^2 + y^2 - 2xy\cos\chi)^{1/2}} \\ &= \frac{2\pi i}{k} C \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|} (1 - e^{2ik|\mathbf{y}|}). \end{aligned} \quad (10.10)$$

The best way to reproduce $u(\mathbf{x})$ is by setting

$$C = -\frac{ik}{2\pi} = -\frac{i}{\lambda}. \quad (10.11)$$

This coincides with an observation already made by Fresnel, according to which the phases of the wavelets must lag behind the phases of the primary wave by $\pi/2$. Furthermore, the intensity is weighted by a factor $1/\lambda$. (From the ansatz, it is obvious that C must have the dimension of an inverse length.) However, the expression we have just calculated contains an unwanted “backward” wave coming from the rear side of the sphere, the phase difference with respect to the forward wave being of the diameter of the phase surface. It must be eliminated by

a suitably chosen inclination factor. Now the phase surfaces are given by spheres $|\mathbf{x}| = \text{const.}$, like one would expect from Huygens' principle.

In the coming sections we will attack the problem of the range of validity of the Huygens–Fresnel principle.

10.3 The method of stationary phases

The principle of Huygens and Fresnel as expressed in eq. (10.9) leads one to consider integrals of the general form

$$I(k) = \int d^n y f(y) e^{ikS(y)}. \quad (10.12)$$

In particular, we are interested in the behavior of these integrals in the limit $k \rightarrow \infty$. Intuitively, one would expect that for larger values of k the expression $e^{ikS(y)}$, considered as a function of y , becomes increasingly oscillatory such that the integrand will vanish in the average for most regions in y space. An exception to this argument will be neighborhoods of stationary points y_0 of the function $S(y)$. In the limit $k \rightarrow \infty$ they should yield the dominant contributions to $I(k)$. Using three lemmas as intermediate steps, we will now confirm this intuitive argument and derive a leading-order expression for $I(k)$. In order to avoid convergence problems for the integral, we will assume $f(y)$ to have compact support.

We shall prove first the following corollary:

Corollary 10.1. *If $S(y)$ has no stationary point in the support of f , i.e. in the closure of the set of points where $f(y) \neq 0$, then $I(k)$ decreases faster than any power of k :*

$$\lim_{k \rightarrow \infty} k^N I(k) = 0 \quad \text{for all } N > 0. \quad (10.13)$$

Proof. We consider the differential operator

$$D = \frac{1}{ik} \frac{1}{\|\partial S(y)\|^2} \partial_j S \frac{\partial}{\partial y^j}. \quad (10.14)$$

By assumption, S has no stationary point in the integration region and we find everywhere

$$\|\partial S(y)\|^2 = \partial_i S \partial_i S \neq 0,$$

so D is indeed well defined. D is chosen in such a way that $D e^{ikS} = e^{ikS}$. The proof now follows after partial integration:

$$\begin{aligned} I(k) &= \int d^n y f(y) D^{N+1} e^{ikS(y)} \\ &= \int d^n y (D^{\dagger(N+1)} f)(y) e^{ikS(y)} = \mathcal{O}\left(\frac{1}{k^{N+1}}\right), \end{aligned} \quad (10.15)$$

where the “adjoint” operator $D^\dagger$ is defined by

$$D^\dagger f = -\frac{1}{ik} \partial_j \left(\frac{\partial_j S}{\|\partial S\|^2} f \right). \quad (10.16)$$

□

According to Corollary 10.1 it is sufficient to restrict the integration to small regions, each of which contains only one stationary point of $S(y)$. The error thus made vanishes faster than any inverse power of k . The integral is additive in the contributions from the stationary points up to terms vanishing faster than any power of k^{-N} . As a warm-up for the general case, we first consider one-dimensional integrals of the form

$$\int_{-\infty}^{\infty} dt g(t) e^{-\lambda t^2/2}.$$

We have the following corollary:

Corollary 10.2. *If $g(t)$, $g'(t)$, and $g''(t)$ are bounded, the integral*

$$I(\lambda) = \int_{-\infty}^{\infty} dt g(t) e^{-\lambda t^2/2}$$

exists for $\lambda \neq 0$, $\operatorname{Re} \lambda \geq 0$, and is holomorphic for $\operatorname{Re} \lambda > 0$.

Proof: The idea of the proof consists of a two-fold partial integration such that the integral

$$\int_{t_1}^{t_2} dt g(t) e^{-\lambda t^2/2}$$

assumes a form where we can easily estimate its asymptotic behavior for large values of $|t_1|$ and $|t_2|$. Because of

$$e^{-\lambda t^2/2} = -\frac{1}{\lambda t} \frac{d}{dt} e^{-\lambda t^2/2}, \quad (10.17)$$

we find

$$\begin{aligned} \int_{t_1}^{t_2} dt g(t) e^{-\lambda t^2/2} &= \int_{t_1}^{t_2} dt \left(-\frac{1}{\lambda t} g \right) \frac{d}{dt} e^{-\lambda t^2/2} \\ &= -\frac{1}{\lambda t} g(t) e^{-\lambda t^2/2} \Big|_{t_1}^{t_2} + \frac{1}{\lambda} \int_{t_1}^{t_2} dt \left(\frac{g}{t} \right)' e^{-\lambda t^2/2} \\ &= -\frac{1}{\lambda t} g(t) e^{-\lambda t^2/2} \Big|_{t_1}^{t_2} - \frac{1}{\lambda^2 t} \left(\frac{g}{t} \right)' e^{-\lambda t^2/2} \Big|_{t_1}^{t_2} \\ &\quad + \frac{1}{\lambda^2} \int_{t_1}^{t_2} dt \left[\frac{1}{t} \left(\frac{g}{t} \right)' \right]' e^{-\lambda t^2/2}. \end{aligned} \quad (10.18)$$

The uniform convergence of all terms for $\lambda \neq 0$, $\operatorname{Re} \lambda \geq 0$ and the holomorphy for $\operatorname{Re} \lambda > 0$ are obvious. □

Furthermore, we write

$$g(t) = g(0) + [g(t) - g(0)] = g(0) + th(t) \quad (10.19)$$

and show the following corollary:

Corollary 10.3. *For $\lambda \neq 0$, $\operatorname{Re} \lambda \geq 0$ we have*

$$\int_{-\infty}^{\infty} dt g(t) e^{-\lambda t^2/2} = \sqrt{\frac{2\pi}{\lambda}} g(0) \left[1 + \mathcal{O}\left(\frac{1}{\lambda}\right) \right], \quad (10.20)$$

where, by definition, $1/\sqrt{\lambda} = 1/\sqrt{|\lambda|} e^{-i\varphi/2}$ when λ is written in the form $\lambda = |\lambda| e^{i\varphi}$.

Proof: The proof is straightforward. First, for real $\lambda > 0$:

$$\int_{t_1}^{t_2} dt \exp(-\lambda t^2/2) = \sqrt{2\pi/\lambda}.$$

According to Corollary 10.2 we may continue to complex values of $\lambda \neq 0$, $\operatorname{Re} \lambda \geq 0$. Furthermore, for $\operatorname{Re} \lambda > 0$ we have:

$$\int_{-\infty}^{\infty} dt th(t) e^{-\lambda t^2/2} = -\frac{1}{\lambda} \int_{-\infty}^{\infty} dt h(t) \frac{d}{dt} e^{-\lambda t^2/2} = \frac{1}{\lambda} \int_{-\infty}^{\infty} dt h' e^{-\lambda t^2/2},$$

and the final integral is of the same type. □

We now come back to our integral

$$\int d^n y f(y) e^{ikS(y)}.$$

We determine the contribution of one integration domain U that contains only one stationary point y_0 of S : $(\partial S/\partial y^i)(y_0) = 0$. In addition, we assume that the stationary point y_0 is regular, i.e. the matrix of second derivatives,

$$S''_{ik}(y_0) = \frac{\partial^2 S}{\partial y_i \partial y_k}(y_0),$$

should be invertible: $\det S''(y_0) \neq 0$. In a sufficiently small neighborhood $U' \in U$, we can introduce coordinates z^i with $y^i(0) = y_0^i$ such that in U' the function $S(y(z))$ assumes the form

$$S(y(z)) = S(y_0) + \frac{1}{2}Q(z) = S(y_0) + \frac{1}{2}(-z_1^2 - \cdots - z_l^2 + z_{l+1}^2 + \cdots + z_n^2).$$

We now have to determine

$$\int_{y(z) \in U'} dz |J(z)| f(y(z)) \exp\left[\frac{1}{2}ikQ(z)\right],$$

where $J = \det \partial y(z)/\partial z$ is the Jacobian matrix of the transformation. Because $Q(z)$ has only one stationary point $z = 0$, the error from extending the integration domain to the whole of $\mathbb{R}^n$ vanishes faster than any inverse power of k . We only have to continue $|J|f(y(z))$ in some way to $\mathbb{R}^n$.

Furthermore,

$$\frac{\partial}{\partial z^r} S(y(z)) = \frac{\partial S}{\partial y^i}(y(z)) \frac{\partial y^i}{\partial z^r},$$

and thus

$$\begin{aligned} \pm \delta_{rs} &= \frac{1}{2} \frac{\partial^2}{\partial z^r \partial z^s} Q(z) = \frac{\partial^2}{\partial z^r \partial z^s} S(y(z)) \\ &= \frac{\partial^2 S}{\partial y^i \partial y^k}(y(z)) \frac{\partial y^i}{\partial z^r} \frac{\partial y^k}{\partial z^s} + \frac{\partial S}{\partial y^i}(y(z)) \frac{\partial^2 y^i}{\partial z^r \partial z^s}. \end{aligned}$$

At the stationary point $y_0 = y(0)$ we have $\partial S/\partial y^i = 0$ and, therefore,

$$|\det S''(y_0)| |J(0)|^2 = 1,$$

i.e.

$$|J(0)| = |\det S''(y_0)|^{-1/2}. \quad (10.21)$$

Putting all the results together we finally find the following theorem:

Theorem 10.1. *Let f be a function with compact support and let S have only finitely many regular stationary points y_i within the domain of f . Then*

$$\begin{aligned} &\int d^n y f(y) e^{ikS(y)} \\ &= \left(\frac{2\pi}{k}\right)^{n/2} \sum_{y_j} \frac{\exp[i(\pi/4) \text{sign } S''(y_j)]}{\sqrt{|\det S''(y_j)|}} f(y_j) e^{ikS(y_j)} + \mathcal{O}(k^{-n/2-1}). \end{aligned} \quad (10.22)$$

Here, $\text{sign } S''(y_i)$ equals the number of positive eigenvalues minus the number of negative eigenvalues of $S''(y_i)$.

As an example for the method of stationary phases, we consider the integral

$$\int_{|y|=y} d\mathbf{f}(\mathbf{y}) \frac{e^{ik(|y|+|\mathbf{x}-\mathbf{y}|)}}{|\mathbf{y}||\mathbf{x}-\mathbf{y}|},$$

which we already calculated in the last section. The stationary points of $S(\mathbf{y}) = |\mathbf{y}| + |\mathbf{x} - \mathbf{y}|$ can be read off fig. 10.2: $S(\mathbf{y})$ is

$$\begin{aligned} \text{minimal at} \quad \mathbf{y}_{01} &= \frac{|\mathbf{y}|}{|\mathbf{x}|} \mathbf{x}, & \text{such that } |\mathbf{y}_{01}| + |\mathbf{x} - \mathbf{y}_{01}| &= |\mathbf{x}|, \\ \text{maximal at} \quad \mathbf{y}_{02} &= -\frac{|\mathbf{y}|}{|\mathbf{x}|} \mathbf{x}, & \text{such that } |\mathbf{y}_{02}| + |\mathbf{x} - \mathbf{y}_{02}| &= 2|\mathbf{y}| + |\mathbf{x}|. \end{aligned}$$

If we define the 3-axis to be along $\mathbf{x}$ and introduce y_1 and y_2 as new coordinates, the matrices of second derivatives are given by

$$\begin{aligned} S''(\mathbf{y}_{01}) &= \frac{|\mathbf{x}|}{|\mathbf{y}|(|\mathbf{x}| - |\mathbf{y}|)} \mathbf{1}, \\ S''(\mathbf{y}_{02}) &= -\frac{|\mathbf{x}|}{|\mathbf{y}|(|\mathbf{x}| + |\mathbf{y}|)} \mathbf{1}, \end{aligned} \quad (10.23)$$

such that, finally,

$$\int_{|\mathbf{y}|=y} d\mathbf{f}(\mathbf{y}) \frac{e^{ik(|\mathbf{y}|+|\mathbf{x}-\mathbf{y}|)}}{|\mathbf{y}||\mathbf{x}-\mathbf{y}|} = \frac{2\pi i}{k} \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|} (1 - e^{2ik|\mathbf{y}|}) + \mathcal{O}\left(\frac{1}{k^2}\right). \quad (10.24)$$

In this case the method of stationary phases even yields the exact result (10.10). The two stationary points correspond to the forward and backward waves.

10.4 Kirchhoff's representation of the wave amplitude

Let $u(\mathbf{x}')$ be a solution of Helmholtz's equation:

$$[\Delta' + k^2 n^2(\mathbf{x}')]u(\mathbf{x}') = 0;$$

and let $G(\mathbf{x}, \mathbf{x}')$ be the Green's function describing an outgoing wave for $\mathbf{x} \approx \mathbf{x}'$:

$$[\Delta' + k^2 n^2(\mathbf{x}')]G(\mathbf{x}, \mathbf{x}') = -4\pi\delta(\mathbf{x}-\mathbf{x}').$$

The following identity holds:

$$\begin{aligned} &\nabla' \{u(\mathbf{x}') \nabla' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \nabla' u(\mathbf{x}')\} \\ &= u(\mathbf{x}') \Delta' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \Delta' u(\mathbf{x}') \\ &= u(\mathbf{x}') [\Delta' + k^2 n^2(\mathbf{x}')] G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') [\Delta' + k^2 n^2(\mathbf{x}')] u(\mathbf{x}') \\ &= -4\pi u(\mathbf{x}') \delta(\mathbf{x}-\mathbf{x}'). \end{aligned} \quad (10.25)$$

Integration over some compact domain W with boundary ∂W yields

$$\begin{aligned} u(\mathbf{x}) &= 0 && \text{for } \mathbf{x} \notin W, \\ u(\mathbf{x}) &= \frac{1}{4\pi} \int_{\partial W} d\mathbf{f}' \{u(\mathbf{x}') \nabla' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \nabla' u(\mathbf{x}')\} && \text{for } \mathbf{x} \in W. \end{aligned} \quad (10.26)$$

Here, the orientation of $d\mathbf{f}'$ is such that it points *into* W .

This identity for u holds for any Green's function of the operator $\Delta + n^2 k^2$. In particular, if we use the Green's function $G_D(\mathbf{x}, \mathbf{x}')$ for Dirichlet boundary conditions, defined by $G_D(\mathbf{x}, \mathbf{x}') = 0$ for $\mathbf{x}' \in \partial W$, the second term in the integral vanishes and we see that $u(\mathbf{x})$ is already determined by its value on ∂W .

For physical applications there are more interesting domains than compact domains, e.g. half-spaces or the regions outside closed surfaces (fig. 10.3), which we can think of as limits of compact domains. We first add a boundary in the form of a half-sphere or sphere of radius R

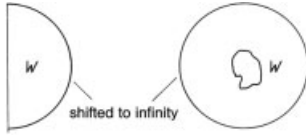


Figure 10.3: Two different non-compact domains where Kirchhoff's representation of wave fields is applied.

and then take the limit $R \rightarrow \infty$. The contribution of this boundary tends to zero for $R \rightarrow \infty$, if $u(\mathbf{x})$ satisfies the so-called *radiation condition*. This condition means that $u(\mathbf{x})$ describes only outgoing waves, which is guaranteed if, for large values of $|\mathbf{x}|$, the function $u(\mathbf{x})$ can be represented as a superposition of outgoing spherical waves whose centers are all located inside a certain compact domain. For physically relevant situations, the radiation condition is always satisfied. In this case, the representation of $u(\mathbf{x})$ in terms of the values of u and $\mathbf{n} \cdot \nabla u$ on the boundary ∂W , which is due to Kirchhoff, also holds for non-compact W .

Even apart from the fact that ∂W is an arbitrary surface and not necessarily a surface of constant phase for $u(\mathbf{x})$, Kirchhoff's representation of $u(\mathbf{x})$ differs, at first sight, essentially from the ansatz

$$u(\mathbf{x}) = \int_{\partial W} df' C\kappa(\chi) u(\mathbf{x}') G(\mathbf{x}, \mathbf{x}'), \quad (10.27)$$

which forms the basis of the Huygens–Fresnel principle.

The similarity reveals itself when we insert into Kirchhoff's integral the ansätze of the short-wave expansion:

$$u(\mathbf{x}) = a(\mathbf{x}) e^{ik\phi(\mathbf{x})}, \quad G(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')}. \quad (10.28)$$

For both cases, the derivative of the exponential yields a factor k and thus the leading terms.

The other terms may be neglected as long as the amplitude varies more slowly than the phases, e.g. if $|\mathbf{x}|$, $|\mathbf{x}'|$, and $|\mathbf{x} - \mathbf{x}'|$ are large compared to the wavelength λ and if caustics are avoided.

Under these conditions we obtain from eq. (10.26):

$$\begin{aligned} u(\mathbf{x}) &= \frac{ik}{4\pi} \int_{\partial W} df' a(\mathbf{x}') A(\mathbf{x}, \mathbf{x}') [\nabla' S(\mathbf{x}, \mathbf{x}') - \nabla' \phi(\mathbf{x}')] e^{ik[\phi(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')] } \\ &= \frac{ik}{4\pi} \int_{\partial W} df' [\nabla' S(\mathbf{x}, \mathbf{x}') - \nabla' \phi(\mathbf{x}')] u(\mathbf{x}') G(\mathbf{x}, \mathbf{x}'). \end{aligned} \quad (10.29)$$

Because $\phi(\mathbf{x}')$ and $S(\mathbf{x}, \mathbf{x}')$ are solutions of the eikonal equation, we have

$$|\nabla' S(\mathbf{x}, \mathbf{x}')| = |\nabla' \phi(\mathbf{x}')| = n(\mathbf{x}'). \quad (10.30)$$

In order to compare with the Huygens–Fresnel principle, we now assume that the phase $\phi(\mathbf{x}')$ of u is constant on ∂W . In this case we have $d\mathbf{f}' \cdot \nabla' \phi = df' n$ and $d\mathbf{f}' \cdot \nabla' S(\mathbf{x}, \mathbf{x}') = -df' n \cos \chi(\mathbf{x}')$, where $\chi(\mathbf{x}')$ is the angle between the normal of the surface and the ray direction of $G(\mathbf{x}, \mathbf{x}')$ at point $\mathbf{x}' \in \partial W$. [The relations become particularly simple for the free case, $G(\mathbf{x}, \mathbf{x}') = e^{ik|\mathbf{x} - \mathbf{x}'|}/|\mathbf{x} - \mathbf{x}'|$, where the rays are straight lines.] We obtain

$$u(\mathbf{x}) = -\frac{ik}{2\pi} \int_{\partial W} df' n(\mathbf{x}') \frac{1 + \cos \chi(\mathbf{x}')}{2} u(\mathbf{x}') G(\mathbf{x}, \mathbf{x}'). \quad (10.31)$$

We now see that the Huygens–Fresnel principle holds for small wavelengths where the short-wave expressions (10.28) hold. Each wavelet contributes with a weight $-(ik/2\pi)n(\mathbf{x}') = -i/\lambda$, and, in particular, there exists a phase shift of $-\pi/2$ with respect to the primary wave. The inclination factor is $\kappa(\chi) = \frac{1}{2}(1 + \cos \chi)$. It differs from the inclination factor used by Fresnel but it also leads to a suppression of the backward wave.

Applying the method of stationary phases to Kirchhoff's integral (10.29) leads to interesting insights. Let $y \in \mathbb{R}^2$ be a parameter for the surface. Then the condition for a stationary point of $\phi(\mathbf{x}'(y)) + S(\mathbf{x}, \mathbf{x}'(y))$ reads

$$[\nabla' \phi(\mathbf{x}'(y)) + \nabla' S(\mathbf{x}, \mathbf{x}'(y))] \cdot \frac{\partial \mathbf{x}'}{\partial y} = 0. \quad (10.32)$$

Therefore, the vector

$$\begin{aligned} \nabla' \phi(\mathbf{x}'(y)) + \nabla' S(\mathbf{x}, \mathbf{x}'(y)) &= \mathbf{p}_u(\mathbf{x}'(y)) - \mathbf{p}_G^x(\mathbf{x}, \mathbf{x}'(y)) \\ &= n(\mathbf{x}')[\boldsymbol{\sigma}_u(\mathbf{x}'(y)) - \boldsymbol{\sigma}_G^x(\mathbf{x}'(y))] \end{aligned}$$

has to be perpendicular to ∂W at point $(\mathbf{x}'(y))$. Here, $\boldsymbol{\sigma}_u$ is the ray unit vector of the function $u(\mathbf{x}')$, and $\boldsymbol{\sigma}_G^x$ is the ray unit vector for the Green's function. There are two ways to fulfill this stationarity condition.

(a) *Forward wave:* $\mathbf{p}_u^x = \mathbf{p}_G$.

(b) *Backward wave:* $\mathbf{p}_u$ and $\mathbf{p}_G^x$ are symmetric with respect to the surface normal N (fig. 10.4), and in this case we find

$$\begin{aligned} N \cdot (\mathbf{p}_u + \mathbf{p}_G^x) &= 0, \\ df' N \cdot (\mathbf{p}_u + \mathbf{p}_G^x) &= df' \cdot (\nabla' \phi - \nabla' S) = 0. \end{aligned} \quad (10.33)$$

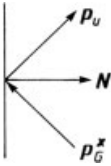


Figure 10.4: Orientation of the wave normal for the backward wave.

Thus, the special form of the integrand in Kirchhoff's integral (10.29) guarantees that the contributions of the stationary points leading to backward waves vanish within the approximation of stationary phases. Only contributions of type (a) remain.

The condition of stationarity, $\mathbf{p}_u = \mathbf{p}_G$, implies that for $k \rightarrow \infty$ only those light rays which traverse ∂W continuously and without kinks contribute to the wave $u(\mathbf{x})$ at point $\mathbf{x}$. If ∂W is a surface of constant phase for u , i.e. if $\phi(\mathbf{x}')$ is constant on ∂W , then $\mathbf{p}_G = -\nabla' S(\mathbf{x}, \mathbf{x}')$ is perpendicular to ∂W , and the only contributions to the amplitude at $\mathbf{x}$ are from those rays which we employed in chapter 9 in order to find the solution of the initial value problem of the wave equation. For these rays we have $\chi = 0$ and the inclination factor equals unity, $\kappa(\chi) = 1$. They traverse the phase surface ∂W at exactly those points $\mathbf{x}'(y_i) \in \partial W$ whose distance $S(\mathbf{x}, \mathbf{x}'(y_i))$ from $\mathbf{x}$, measured in the light metric, is stationary.

We can indeed convince ourselves that the method of stationary phases applied to Kirchhoff's integral yields the expression for $u(\mathbf{x})$ that we obtained in section 9.5. In particular, the phase factor $\exp(-i\pi\mu_j/2)$ in eq. (9.65) is identical to the phase $(-i) \exp[(\frac{1}{4}i\pi) \text{sign } S''(y_i)]$, where

$$(S''(y))_{ik} = \frac{\partial^2 S(\mathbf{x}, \mathbf{x}'(y))}{\partial y^i \partial y^k}.$$

Here μ_j is just the number of focal points on the light ray between $\mathbf{x}'(y_i)$ and $\mathbf{x}$. According to our discussion in section 9.4, the focal points of order r are the “centers of curvature” of ∂W , and these, in turn, are the points where the eigenvalues of S'' change sign, which leads to a jump of $2r$ in the signature of S'' . We reconfirm this last statement once more for the case where the light rays are straight lines, i.e. where $S(\mathbf{x}, \mathbf{x}'(y)) = |\mathbf{x} - \mathbf{x}'(y)|$. The relation

$$\frac{\partial}{\partial y^i} |\mathbf{x} - \mathbf{x}'| = \frac{\mathbf{x}'(y) - \mathbf{x}}{|\mathbf{x} - \mathbf{x}'(y)|} \cdot \frac{\partial \mathbf{x}'}{\partial y^i} = 0 \quad (10.34)$$

implies that $\mathbf{N} = (\mathbf{x} - \mathbf{x}')/|\mathbf{x} - \mathbf{x}'|$ points into the normal direction and thus

$$\begin{aligned} \frac{\partial^2}{\partial y^i \partial y^k} |\mathbf{x} - \mathbf{x}'(y)| &= -\mathbf{N} \cdot \frac{\partial^2 \mathbf{x}'}{\partial y^i \partial y^k} + \frac{1}{|\mathbf{x} - \mathbf{x}'(y)|} \frac{\partial \mathbf{x}'}{\partial y^i} \cdot \frac{\partial \mathbf{x}'}{\partial y^k} \\ &= \frac{1}{|\mathbf{x} - \mathbf{x}'|} g_{ik} - L_{ik} \end{aligned} \quad (10.35)$$

is again proportional to the matrix we already encountered in section 9.4.

The value $|\mu_j| > 2$ is possible because the action S is primarily defined on the Lagrangian manifold $\Lambda_{\partial W}$ and may be multivalued in position space. Therefore, the light ray may pass through several focal points.

Finally, we want to consider the important and illustrative special case of a system that is symmetric around its optic axis and where we may confine our discussion to rays close to the axis and with small inclinations relative to the optic axis (fig. 10.5).

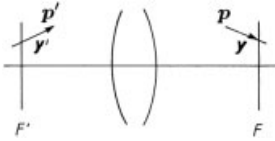


Figure 10.5: Kirchhoff's integral for a Gaussian optical system.

Let F' and F be two non-conjugated planes with an optical separation z measured along the optic axis. As we did in chapter 8, we use the following ray coordinates: $\mathbf{y}'$ and $\mathbf{y}$ are vectors in F' and F , respectively, and $\mathbf{p}'$ and $\mathbf{p}$ are the two components of the ray momenta perpendicular to the optic axis.

The image matrix from F' to F will be given by

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \mathbf{y}' \\ \mathbf{p}' \end{pmatrix} \quad \text{with} \quad ab - cd = 1, \quad b \neq 0. \quad (10.36)$$

According to eq. (8.8.2) the point characteristic reads

$$S(\mathbf{y}, \mathbf{y}') = \frac{1}{2b}(d\mathbf{y}^2 - 2\mathbf{y} \cdot \mathbf{y}' + a\mathbf{y}'^2) + z, \quad (10.37)$$

and the Green's function

$$G(\mathbf{y}, \mathbf{y}') = a(\mathbf{y}, \mathbf{y}') e^{ikS(\mathbf{y}, \mathbf{y}')}$$

is given by

$$G(\mathbf{y}, \mathbf{y}') = (-1)^n \frac{1}{b} \exp\left(\frac{ik}{2b}(d\mathbf{y}^2 - 2\mathbf{y} \cdot \mathbf{y}' + a\mathbf{y}'^2) + ikz\right). \quad (10.38)$$

The pre-factor $a(\mathbf{y}, \mathbf{y}') = (-1)^n/b = [\det D\xi(\tau, \mathbf{p}')]^{-1/2}$ follows from the conservation of the energy flux as well as from the form of the light rays, which close to F can be written as

$$\xi(\tau, \mathbf{p}') = (a\mathbf{y}' + b\mathbf{p}' + (\tau - z)\mathbf{p}', \tau).$$

In eq. (10.38), n denotes the number of planes that are conjugated to F' and are located between F' and F ; because $\mu = 2n$, we find $e^{-i\pi\mu/2} = (-1)^n$.

Eventually, Kirchhoff's integral assumes the form

$$u_F(\mathbf{y}) = -\frac{i}{\lambda b} (-1)^n e^{ikz} \int d^2y' u_{F'}(\mathbf{y}') \exp\left(\frac{ik}{2b}(d\mathbf{y}^2 - 2\mathbf{y} \cdot \mathbf{y}' + a\mathbf{y}'^2)\right). \quad (10.39)$$

Note that for coinciding F and F' (i.e. for $z = 0$, $a = d = 1$, and $b \rightarrow 0$) one indeed obtains $u_F = u_{F'}$.

10.5 Kirchhoff's theory of diffraction

The typical arrangement of a diffraction experiment is depicted in fig. 10.6. The light from some source hits a screen S , through which it can pass only at certain apertures. The light wave is observed on a surface F behind the screen (as seen from the source). There may be optical components like lenses or mirrors in front of and behind the screen, i.e. the refractive index need not be constant.

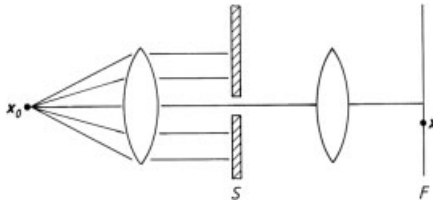


Figure 10.6: Principle arrangement of a diffraction experiment.

Within the framework of Kirchhoff's diffraction theory, the screen together with its apertures is identified with the boundary ∂W of some domain W . Kirchhoff's identity,

$$u(x) = \frac{1}{4\pi} \int_{\partial W} d\mathbf{f}' [u(\mathbf{x}') \nabla' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \nabla' u(\mathbf{x}')], \quad (10.40)$$

provides a relation between the value of u at some arbitrary point $\mathbf{x} \in W$ and the values of u as well as the normal derivatives of u on the boundary ∂W . Since the screen might have a reactive influence on the light wave due to diffraction, the values of u and $\mathbf{N} \cdot \nabla u$ on ∂W are not known from the beginning. The light is diffracted at the screen and will be deflected. Immediately behind the screen these effects may be assumed to be small enough such that the wave $u(\mathbf{x})$ there is given by geometrical optics. In this sense, Kirchhoff's diffraction theory makes the following assumption about the behavior of u and $\mathbf{N} \cdot \nabla u$ on the dark side of ∂W :

On the screen we have $u(\mathbf{x}') = \mathbf{N} \cdot \nabla u(\mathbf{x}') \equiv 0$. Within the apertures of the screen, $u(\mathbf{x}')$ and $\mathbf{N} \cdot \nabla u(\mathbf{x}')$ coincide with the light wave that would be there in the absence of the screen.

According to Kirchhoff's theory, in W we have

$$u(\mathbf{x}) = \frac{1}{4\pi} \int_A d\mathbf{f}' [w(\mathbf{x}') \nabla' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \nabla' w(\mathbf{x}')], \quad (10.41)$$

where the integral is to be taken over the apertures A of the screen and $w(\mathbf{x}')$ is given by the undisturbed wave of the light source.

Strictly speaking, this ansatz is not consistent. If one calculates $u(\mathbf{x})$ according to formula (10.41), one finds a non-vanishing amplitude on the screen, which, however, is small. Furthermore, we are only dealing with a scalar amplitude and, hence, dismiss all diffractive phenomena related to the polarization of light. There exist extensions of Kirchhoff's theory for vector fields (which we will not elaborate further in this context), but the intensities are reproduced quite well by the scalar theory. For some rare cases the boundary value problem corresponding to a diffractive arrangement is exactly solvable. It turns out that Kirchhoff's theory describes the situation close to the geometrical boundaries of the shadows, where the most interesting effects are to be expected, quite successfully.

If we replace the screen S by a complementary screen S' , where the apertures and the non-transparent parts just exchange places, the sum of the corresponding amplitudes u and u' are given by an integral over the total boundary ∂W , and from Kirchhoff's identity we obtain

$$u(\mathbf{x}) + u'(\mathbf{x}) = u_0(\mathbf{x}), \quad (10.42)$$

where $u_0(\mathbf{x})$ is the undisturbed amplitude in the absence of the screen. For each point $\mathbf{x}$ where $u_0(\mathbf{x})$ vanishes we find equal intensities:

$$|u(\mathbf{x})|^2 = |u'(\mathbf{x})|^2. \quad (10.43)$$

If, for instance, the light coming from the source is sharply focussed at a point $\mathbf{x}'_0$ on the observer's surface F , the intensities for the two arrangements are equal everywhere on F except at the image point $\mathbf{x}'_0$. This equality of diffraction phenomena for complementary screens is known as *Babinet's principle*.

We now proceed with a more detailed discussion of Kirchhoff's formula (10.41). Once again we employ the ansätze

$$w(\mathbf{x}) = a(\mathbf{x}) e^{ik\phi(\mathbf{x})} \quad \text{and} \quad G(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')}.$$

and disregard the derivatives of $a(\mathbf{x}')$ and $A(\mathbf{x}, \mathbf{x}')$, and we obtain:

$$u(\mathbf{x}) = \frac{ik}{4\pi} \int_A d\mathbf{f}' a(\mathbf{x}') A(\mathbf{x}, \mathbf{x}') [\nabla' S(\mathbf{x}, \mathbf{x}') - \nabla' \phi(\mathbf{x}')] e^{ik[\phi(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')]} \quad (10.44)$$

In leading order of k this disregard is well justified, but it also holds if the amplitudes $a(\mathbf{x}')$ and $A(\mathbf{x}, \mathbf{x}')$ within the apertures do not vary much within the distance of a wavelength. This should be true if there are no focal points of $w(\mathbf{x}')$ or $G(\mathbf{x}, \mathbf{x}')$ within the apertures, which in any case would contradict the idea of a diffraction experiment. Therefore, the expression above should not be identified with the leading term of a short-wave expansion of u ; in particular, the wavelength does not necessarily have to be small compared to the dimensions of the aperture. It is just this case for which we expect formula (10.44) to describe the deviations from geometrical optics.

We obtain the behavior for $k \rightarrow \infty$ by applying to eq. (10.44) the method of stationary phases. If there exists a light ray from the source to the point $\mathbf{x}$ on F which intersects the boundary ∂W at a point $\mathbf{x}'(\mathbf{x}) \in A$ of the aperture, then, for $k \rightarrow \infty$, $u(\mathbf{x})$ is given by the undisturbed amplitude in the absence of a screen. For all other cases, $u(\mathbf{x})$ decreases for $k \rightarrow \infty$ faster than any power of k . These domains on F just correspond to the regions of geometrical light and geometrical shadow. From the discussion in section 10.3 we also know that the error introduced by changing the functions in front of the exponential function in the integral outside a certain small neighborhood of the stationary point $\mathbf{x}'(\mathbf{x})$ decreases rapidly for increasing values of k . Therefore, the $\mathbf{x}'$ dependence of those pre-exponential factors is negligible even if the dimensions of the apertures are not so small that an essential change of these quantities is impossible anyhow.

The $\mathbf{x}'$ dependence of $\nabla' S(\mathbf{x}, \mathbf{x}') = -\mathbf{p}(\mathbf{x}, \mathbf{x}')$ and $\nabla' \phi(\mathbf{x}') = \mathbf{p}_0(\mathbf{x}')$ can be estimated even more precisely. The derivatives of these quantities with respect to $\mathbf{x}'$ are of the order of the curvature of the wave fronts of w and $G(\mathbf{x}, \mathbf{x}')$ at point $\mathbf{x}'$. They are small if there are no focal points of w or G inside the aperture, and if the source and the observation point $\mathbf{x}$ are sufficiently far away from the screen.

Furthermore, we can neglect the $\mathbf{x}'$ dependence of $A(\mathbf{x}, \mathbf{x}')$ and $\nabla' S(\mathbf{x}, \mathbf{x}')$ in the pre-exponential terms. The quantities $a(\mathbf{x}')$ and $\nabla' \phi(\mathbf{x}')$ are allowed to have stronger variations, because we want to include the description of an important generalization of diffraction where the apertures may have a varying transparency, like, for example, a slide.

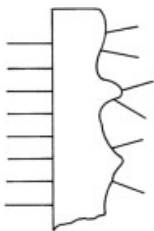


Figure 10.7: Diffraction at an amplitude or phase object.

Furthermore, variations of the optical density along the aperture may introduce additional changes in phase (see fig. 10.7). However, these changes are with respect to the tangential direction and they drop out in $d\mathbf{f}' \cdot \nabla' \phi$. Depending on whether the variation in amplitude or

phase is dominant, one speaks of an *amplitude object* or a *phase object*. Our expression for the diffracted wave is now of the form

$$u(\mathbf{x}) = -\frac{ik}{4\pi} A(\mathbf{x}, \mathbf{x}'(\mathbf{x})) (\cos \chi_0 + \cos \chi) \int_A df' a(\mathbf{x}') e^{ik\phi(\mathbf{x}')} e^{ikS(\mathbf{x}, \mathbf{x}')}. \quad (10.45)$$

Here, χ_0 and χ are the angles between the wave normals of w and G , respectively, and the normal of the surface A . For a better understanding of the influence of the aperture, we introduce quantities that describe the modification of the incoming wave by the aperture. Immediately in front of the aperture w will be given by

$$w(\mathbf{x}') = a_0(\mathbf{x}') e^{ik\phi_0(\mathbf{x}')} \quad (10.46)$$

and immediately behind we write w in the form

$$w(\mathbf{x}') = a(\mathbf{x}') e^{ik\phi(\mathbf{x}')} = a_0(\mathbf{x}') e^{ik\phi_0(\mathbf{x}')} t(\mathbf{x}') e^{i\varphi(\mathbf{x}')}. \quad (10.47)$$

Here $t(\mathbf{x}')$ is the *transmission* and $\varphi(\mathbf{x}')$ is the *phase modulation*.

With these quantities we can write for u :

$$\begin{aligned} u(\mathbf{x}) &= -\frac{ik}{4\pi} A(\mathbf{x}, \mathbf{x}'(\mathbf{x})) a_0(\mathbf{x}, \mathbf{x}'(\mathbf{x})) (\cos \chi_0 + \cos \chi) \\ &\quad \times \int_A df' e^{ik\varphi(\mathbf{x}')} t(\mathbf{x}') e^{ik[\phi_0(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')].} \end{aligned} \quad (10.48)$$

In general, the factors in front of the integral vary only weakly as a function of $\mathbf{x}$, and the interesting quantity that we now want to study is

$$\begin{aligned} U(\mathbf{x}) &= \int_A df' t(\mathbf{x}') e^{i\varphi(\mathbf{x}')} e^{ik[\phi_0(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')]} \\ &= \int_{\partial W} df' P(\mathbf{x}') e^{ik[\phi_0(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')].} \end{aligned} \quad (10.49)$$

Here we have introduced the so-called *pupil function* on ∂W :

$$P(\mathbf{x}') = \begin{cases} t(\mathbf{x}') e^{i\varphi(\mathbf{x}')} & \text{for } \mathbf{x}' \in A, \\ 0 & \text{for } \mathbf{x}' \notin A. \end{cases} \quad (10.50)$$

If A is a simple hole, we have $P(\mathbf{x}') = 1$ for $\mathbf{x}' \in A$ and $P(\mathbf{x}') = 0$ for $\mathbf{x}' \notin A$.

The *optical transmission function* $K(\mathbf{x}, \mathbf{x}_0)$ is defined to be the wave $u(\mathbf{x})$ that would emerge behind the diffracting screen if the source in front of the screen were a point source located at $\mathbf{x}_0$:

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}_0) &= \frac{1}{4\pi} \int_A df' [G(\mathbf{x}', \mathbf{x}_0) \nabla' G(\mathbf{x}, \mathbf{x}') - G(\mathbf{x}, \mathbf{x}') \nabla' G(\mathbf{x}', \mathbf{x}_0)] \\ &= -\frac{ik}{4\pi} A(\mathbf{x}, \mathbf{x}'(\mathbf{x})) A(\mathbf{x}'(\mathbf{x}), \mathbf{x}_0) (\cos \chi_0 + \cos \chi) \\ &\quad \times \int_{\partial W} df' P(\mathbf{x}') e^{ik[S(\mathbf{x}', \mathbf{x}_0) + S(\mathbf{x}, \mathbf{x}')].} \end{aligned} \quad (10.51)$$

Obviously, K is the Green's function which takes into account the influence of the screen. In particular, we find for a constant refractive index in front of and behind the screen:

$$G(\mathbf{x}, \mathbf{x}') = \frac{e^{ik|\mathbf{x}-\mathbf{x}'|}}{|\mathbf{x}-\mathbf{x}'|}$$

and

$$K(\mathbf{x}, \mathbf{x}_0) = \frac{-ik}{|\mathbf{x}-\mathbf{x}'(\mathbf{x})||\mathbf{x}_0-\mathbf{x}'(\mathbf{x})|} \left(\frac{\cos \chi_0 + \cos \chi}{4\pi} \right) \times \int_{\partial W} df' P(\mathbf{x}') e^{ik(|\mathbf{x}'-\mathbf{x}_0|+|\mathbf{x}-\mathbf{x}'|)}. \quad (10.52)$$

The transmission function $K(\mathbf{x}, \mathbf{x}_0)$ characterizes the imaging properties of an optical instrument by describing the image of a point source. In an optical instrument, the inclination angles are usually restricted to small values by circular apertures. Effectively, the diffracting aperture is given by the *pupil*, which may be a really existing or merely imaginary aperture providing the lateral limits for the rays.

In order to determine the function

$$U(\mathbf{x}) = \int_A df' P(\mathbf{x}') e^{ik[\phi_0(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}')]}, \quad (10.53)$$

we choose a reference point $\mathbf{x}'_0$ in the aperture and make a Taylor expansion in $\boldsymbol{\xi}$:

$$\mathbf{x}' = \mathbf{x}'_0 + \boldsymbol{\xi}.$$

This Taylor expansion exists if there are no focal points within the aperture. We find

$$\begin{aligned} \phi_0(\mathbf{x}') + S(\mathbf{x}, \mathbf{x}') &=: \Psi(\mathbf{x}') \\ &= \Psi(\mathbf{x}'_0) + [\mathbf{p}_0(\mathbf{x}'_0) - \mathbf{p}(\mathbf{x}'_0, \mathbf{x})] \boldsymbol{\xi} + \frac{1}{2} \frac{\partial^2 \Psi(\mathbf{x}_0)}{\partial x'^i \partial x'^k} \xi^i \xi^k + \dots \end{aligned} \quad (10.54)$$

The quadratic term is negligible if

$$k \frac{\partial^2 \Psi}{\partial x'^i \partial x'^k} \xi^i \xi^k \ll 2\pi.$$

As we have seen, the second derivatives of Ψ are of the order of $1/R_0$ or $1/R$, the curvatures of the wave fronts. If we characterize the dimension of the aperture by d , the above condition reads

$$k \left(\frac{1}{R_0} + \frac{1}{R} \right) d^2 \ll 2\pi \quad \text{or} \quad \lambda R \gg d^2, \quad \lambda R_0 \gg d^2. \quad (10.55)$$

The case for which a linearization of the phase Ψ is justified is called *Fraunhofer diffraction*; otherwise we speak of *Fresnel diffraction*.

For instance, for a constant refractive index we find

$$\begin{aligned} |\mathbf{x}' - \mathbf{x}_0| + |\mathbf{x} - \mathbf{x}'| &= R + R_0 + (\mathbf{N}_0 - \mathbf{N}) \cdot \boldsymbol{\xi} + \frac{1}{2} \left(\frac{1}{R_0} + \frac{1}{R} \right) \xi^2 \\ &\quad - \frac{(\mathbf{N} \cdot \boldsymbol{\xi})^2}{2R} - \frac{(\mathbf{N}_0 \cdot \boldsymbol{\xi})^2}{2R_0}, \end{aligned}$$

where

$$R_0 = |\mathbf{x}'_0 - \mathbf{x}_0|, \quad R = |\mathbf{x} - \mathbf{x}'_0|, \quad \mathbf{N}_0 = \frac{\mathbf{x}'_0 - \mathbf{x}_0}{R_0}, \quad \mathbf{N} = \frac{\mathbf{x} - \mathbf{x}'_0}{R}. \quad (10.56)$$

The radii of curvature are just the distances of the source and the observer from the aperture.

The second derivatives of $S(\mathbf{x}, \mathbf{x}')$ with respect to $\mathbf{x}'$ vanish if $\mathbf{x}$ lies in the focal plane or, more generally, in the focal surface of the optical instrument behind the screen. Indeed, for this case we know from our discussion in chapter 8 that

$$\mathbf{p}'(\mathbf{x}, \mathbf{x}') = -\frac{\partial S}{\partial \mathbf{x}'} = \mathbf{p}'(\mathbf{x}) \quad (10.57)$$

is independent of the point $\mathbf{x}' \in A$ (see fig. 10.8).

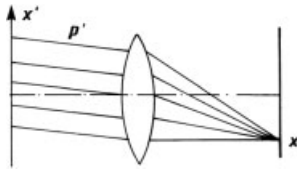


Figure 10.8: If F is focal with respect to A , then $\mathbf{p}'(\mathbf{x}, \mathbf{x}')$ does not depend on $\mathbf{x}'$.

So the conditions for Fraunhofer diffraction are fulfilled if the observation point $\mathbf{x}$ and the source $\mathbf{x}_0$ are both located in a focal plane with respect to A . In this case, the image of $\mathbf{x}_0$ lies in the image surface F. These conditions are also fulfilled if F is a focal surface and A is a surface of constant phase of $w(\mathbf{x}')$.

Fraunhofer diffraction is the simplest and most important form of diffraction phenomenon. If the aperture A is located in a plane, the function

$$U(\mathbf{x}) = \int d^2\xi P(\xi) e^{ik[p_0 - p(\mathbf{x})] \cdot \xi} \quad (10.58)$$

is now simply the Fourier transform of the pupil function. Up to weakly varying factors, the same also holds for the wave solution $u(\mathbf{x})$ and the optical transmission function $K(\mathbf{x}, \mathbf{x}_0)$. The Fourier transform of the pupil function manifests itself in the spatial diffraction pattern in the focal plane F. We shall see soon that this feature gives rise to attractive methods to manipulate the image of an optical instrument.

10.6 Diffraction at an edge

Before we turn to the phenomenon of Fraunhofer diffraction, we first want to discuss an example where Fraunhofer's approximation is not applicable. Let the screen S have the shape of a half-plane, i.e. the aperture A is also a half-plane. Assume that the screen is illuminated by a point source located at $\mathbf{x}_0$. For convenience we place $\mathbf{x}_0$, S, and F as shown in fig. 10.9: S will be in the 1–2 plane and the edge is oriented along the 2-axis. For an everywhere constant

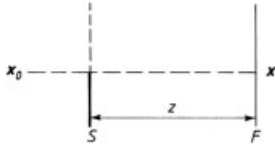


Figure 10.9: Diffraction at an edge.

refractive index, the wave is, up to almost constant factors, given by

$$U(x) = \int_0^\infty d\xi_1 \int_{-\infty}^\infty d\xi_2 \quad (10.59)$$

$$\times \exp \left\{ ik \left[\sqrt{x_0^2 + \xi_1^2 + \xi_2^2} + \sqrt{z^2 + (x_1 - \xi_1)^2 + (x_2 - \xi_2)^2} \right] \right\}.$$

If x is close to the boundary of the geometrical shadow, the value of ξ for which the exponent becomes stationary is also small; and for $x_0, z \gg |\xi|$, we expand the exponent in terms of ξ , because the main contribution to the integral stems from small values of ξ . Up to second order this expansion yields

$$\begin{aligned} & \sqrt{x_0^2 + \xi_1^2 + \xi_2^2} + \sqrt{z^2 + (x_1 - \xi_1)^2 + (x_2 - \xi_2)^2} \\ &= x_0 + z + \frac{\xi_1^2 + \xi_2^2}{2x_0} + \frac{(x_1 - \xi_1)^2 + (x_2 - \xi_2)^2}{2z}. \end{aligned} \quad (10.60)$$

Therefore, if we were to omit all terms quadratic in ξ , the integral would be divergent. Including the quadratic terms we obtain

$$\begin{aligned} U(x_1, x_2) &= e^{ik[x_0 + z + (x_1^2 + x_2^2)/2z]} \int_{-\infty}^\infty d\xi_2 e^{ik[-p_2 \xi_2 + (\xi_2^2/2R)]} \\ &\times \int_0^\infty d\xi_1 e^{ik[-p_1 \xi_1 + (\xi_1^2/2R)]}, \end{aligned}$$

where $p_1 = x_1/z$, $p_2 = x_2/z$, and $1/R = 1/x_0 + 1/z$. From

$$ik \left(\frac{\xi^2}{2R} - p\xi \right) = \frac{ik}{2R} (\xi - Rp)^2 - \frac{ikRp^2}{2},$$

we find

$$\begin{aligned} U(x_1, x_2) &= \exp \left[ik \left(x_0 + z + \frac{(x_1^2 + x_2^2)}{2(x_0 + z)} \right) \right] \\ &\times \int_{-\infty}^\infty d\xi_2 e^{(ik/2R)(\xi_2 - Rp_2)^2} \int_0^\infty d\xi_1 e^{(ik/2R)(\xi_1 - Rp_1)^2}. \end{aligned} \quad (10.61)$$

The first integral yields the constant $\sqrt{2\pi R/k} e^{i\pi/4}$ and is independent of $p_2 = x_2/z$. For the second integral we obtain

$$\int_{-Rp_1}^\infty d\xi e^{(ik/2R)\xi^2} = \sqrt{\frac{R}{k}} \int_{-\sqrt{kR}p_1}^\infty d\eta e^{i\eta^2/2}.$$

The function

$$F(v) = \mathcal{C}(v) + i\mathcal{S}(v) = \frac{1}{\sqrt{2\pi}} \int_v^\infty d\eta e^{i\eta^2/2} \quad (10.62)$$

is called *Fresnel's integral*.

For the final result we get

$$U(x_1, x_2) = 2\pi \frac{R}{k} e^{i\pi/4} \exp \left[ik \left(x_0 + z \frac{x_1^2 + x_2^2}{2(x_0 + 2z)} \right) \right] F \left(-\sqrt{kR} \frac{x_1}{z} \right). \quad (10.63)$$

The variation of the intensity, $|U|^2$, is solely due to the factor $|F(-\sqrt{kR}(x_1/z))|^2$. Special values for Fresnel's integral F are $F(-\infty) = e^{i\pi/4}$ and $F(0) = \frac{1}{2} e^{i\pi/4}$. The function

$$\phi(v) = F(0) - F(v) = \int_0^v d\eta e^{i\eta^2/2}$$

satisfies

$$\phi(-v) = -\phi(v), \quad \phi(\pm\infty) = \pm \frac{1}{2} e^{i\pi/4}.$$

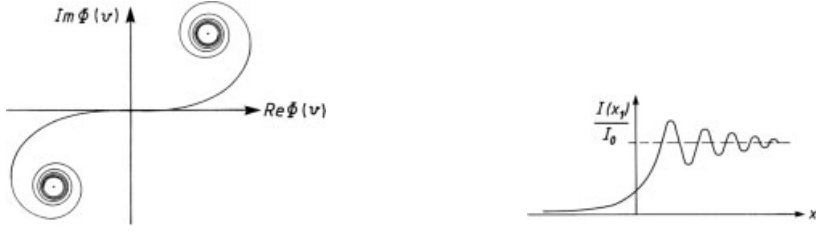


Figure 10.10: (Left) Cornu spiral. (Right) Diffraction image of an edge.

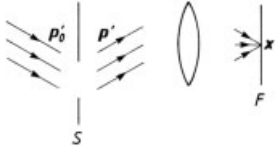
Making a plot of the curve $(\text{Re } \phi(v), \text{Im } \phi(v))$ with curve parameter v we obtain a graph that is called a *Cornu spiral* (fig. 10.10, left). From this plot we can immediately read off the behavior of the intensity $I(x_1) = |U|^2$ for the diffraction pattern at a half-plane depicted in fig. 10.10 (right). The distance between the geometrical border of the shadow and the first maximum of this curve is of the order $z/\sqrt{kR}$.

10.7 Examples of Fraunhofer diffraction

For Fraunhofer diffraction we have seen that the diffracted wave is given, up to insignificant factors, by the Fourier transform of the pupil function:

$$U(\mathbf{x}) = \int d^2\xi P(\boldsymbol{\xi}) e^{ik(\mathbf{p}_0 - \mathbf{p}') \cdot \boldsymbol{\xi}}. \quad (10.64)$$

Here, $\mathbf{p}_0'$ and $\mathbf{p}'$ are the momentum vectors of the incoming and outgoing waves, respectively. The $\mathbf{x}$ dependence enters because of the $\mathbf{x}$ dependence of $\mathbf{p}'(\mathbf{x})$ (see fig. 10.11).

**Figure 10.11:** Fraunhofer diffraction.

If $\mathbf{x}$ is located in the focal plane F of the optical system behind the screen, which in linear approximation may be represented by an optical matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we can write

$$\mathbf{x} = a\boldsymbol{\xi} + b\mathbf{p}'.$$

In addition, because F is supposed to be the focal plane, we have $a = 0$ and the optical matrix is of the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & f \\ -1/f & d \end{pmatrix},$$

i.e.

$$\mathbf{x} = f\mathbf{p}', \quad \mathbf{p}' = \frac{\mathbf{x}}{f}.$$

In the following we will assume that $\mathbf{p}'_0 = 0$ (the case $\mathbf{p}'_0 \neq 0$ only leads to a trivial displacement of the diffraction pattern), and we consider diffraction at apertures of various shapes.

10.7.1 Diffraction by a rectangle

In this case the following hold:

$$-\frac{a_1}{2} \leq \xi_1 \leq \frac{a_1}{2}, \quad -\frac{a_2}{2} \leq \xi_2 \leq \frac{a_2}{2}.$$

From

$$\begin{aligned} \int_{-a/2}^{a/2} d\xi e^{-ikp'\xi} &= \frac{-1}{ikp'} (e^{-ikp'a/2} - e^{+ikp'a/2}) = \frac{\sin(kp'a/2)}{(kp'a/2)} a \\ &= \frac{\sin(p'a\pi/\lambda)}{(p'a\pi/\lambda)} a \end{aligned}$$

we obtain

$$U = \frac{\sin(p'_1 a_1 \pi / \lambda)}{(p'_1 a_1 \pi / \lambda)} \cdot \frac{\sin(p'_2 a_2 \pi / \lambda)}{(p'_2 a_2 \pi / \lambda)} a_1 a_2. \quad (10.65)$$

The intensity is proportional to $|U|^2$.

For $a \rightarrow \infty$ one obtains diffraction by a slit. The diffraction pattern is homogeneous along the 2-axis. Along the 1-axis we find a distribution like the one shown in fig. 10.12. The minima are equally separated by $\Delta p = \lambda/a_1$, and the first minimum is found at $p_m = \lambda/a_1$, in accordance with the elementary theory of diffraction by a slit.

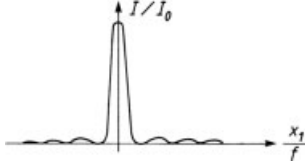


Figure 10.12: Diffraction pattern for a slit.

10.7.2 Diffraction by a circular aperture

In this case we have:

$$|\xi| \leq a.$$

We make a transformation to polar coordinates,

$$\begin{aligned} \xi_1 &= \rho \cos \varphi, & p'_1 &= p \cos \psi, \\ \xi_2 &= \rho \sin \varphi, & p'_2 &= p \sin \psi, \end{aligned}$$

where $\mathbf{p}' \cdot \xi = p\rho \cos(\varphi - \psi)$, and obtain

$$U = \int_0^a d\rho \rho \int_0^{2\pi} d\varphi e^{ikp\rho \cos(\varphi - \psi)}.$$

The integral does not depend on ψ and can be expressed by a Bessel function:

$$U = 2\pi \int_0^a d\rho \rho J_0(kp\rho) = \pi a^2 \frac{2J_1(kpa)}{kpa} = \pi a^2 \frac{2J_1(2\pi pa/\lambda)}{(2\pi pa/\lambda)}. \quad (10.66)$$

For large values of v , the zeros of $J_1(v)$ become equally spaced with separation π . The first zeros are at

$$v_1 = 1.22\pi, \quad v_2 = 2.33\pi, \quad v_3 = 3.24\pi. \quad (10.67)$$

The diffraction pattern consists of a bright circular area surrounded by a concentric system of alternating dark and bright rings (see 10.13). The first minimum is found at

$$p_m = \frac{r_m}{f} = 0.61 \frac{\lambda}{a}. \quad (10.68)$$

This diffraction pattern was first derived by G. B. Airy and is of fundamental importance in the theory of optical instruments. For such a system, the pattern is determined by the diffraction by circular apertures. Therefore, in the plane where, according to the laws of geometrical

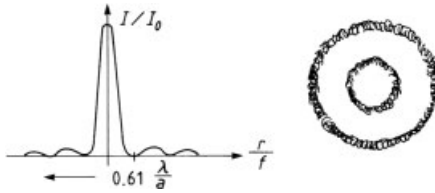


Figure 10.13: Diffraction pattern for a circular hole: the Airy function.

optics, a sharp image would appear, the transmission function K is always an Airy function, $2J_1(v)/v$. Additional geometrical aberration effects should be convoluted with the Airy function.

For the image of a point-like source, the resolution of optical instruments is limited by the finite radius of the central maximum of the Airy function. We can estimate the smallest resolvable separation of point sources by requiring that the maximum of one image just coincides with the minimum of the second.



Figure 10.14: (Left) Principle of a telescope. (Right) Resolution of a microscope.

In a telescope (the principle is shown in fig. 10.14, left), it is the front aperture that leads to diffraction. Therefore, the smallest resolvable angle between two stars at infinity is given by $p_m = 0.61 \lambda/a$.

For a microscope, it is the front lens with focal length f that yields the main diffractive effect (fig. 10.14, right). The distance between primary object and front lens is roughly equal to f . The wave fronts of two neighboring points with separation r are inclined towards each other at an angle of the order of $p = r/f$. Therefore, a rough estimate for the smallest resolvable distance yields

$$r_m = 0.61 \lambda \frac{f}{a}. \quad (10.69)$$

10.7.3 Arrangements of several identical structures

For several diffracting apertures or, more generally, diffracting structures, which are located at different positions, the pupil function is given by

$$P(\xi) = \sum_{j=1}^N P_0(\xi - r_j),$$

where P_0 is the pupil function of a single structure. In this case

$$\begin{aligned}
 U &= \int d^2\xi P(\xi) e^{-ik\mathbf{p}'\cdot\xi} = \sum_{j=1}^N \int d^2\xi P_0(\xi - \mathbf{r}_j) e^{-ik\mathbf{p}'\cdot\xi} \\
 &= \left(\sum_{j=1}^N e^{-ik\mathbf{p}'\cdot\mathbf{r}_j} \right) \int d^2\xi P_0(\xi) e^{-ik\mathbf{p}'\cdot\xi} \\
 &= U_0 \cdot \sum_{j=1}^N e^{-ik\mathbf{p}'\cdot\mathbf{r}_j}.
 \end{aligned} \tag{10.70}$$

The diffracted wave is the product of a factor U_0 describing the diffraction at a single structure and a factor $A = \sum_{j=1}^N e^{-ik\mathbf{p}'\cdot\mathbf{r}_j}$ that depends on the positions of the various structures. We discuss two particularly important cases for the spatial distribution of diffracting structures, namely a random distribution of many structures and a regular lattice arrangement (see fig. 10.15), and then consider the case of diffraction at two circular holes.

(a) Random distribution (fig. 10.15a)

The average intensity is given by $I = \langle |A|^2 \rangle_{\text{ave}} |U_0|^2$, where

$$\langle |A|^2 \rangle_{\text{ave}} = \left\langle \sum_{i,j=1}^N e^{-ik\mathbf{p}'\cdot(\mathbf{r}_i - \mathbf{r}_j)} \right\rangle_{\text{ave}}. \tag{10.71}$$

For a random spatial distribution, all the terms in this sum with $i \neq j$ will mutually cancel each other and we obtain $\langle |A|^2 \rangle_{\text{ave}} = N$; thus

$$I = N |U_0|^2 = NI_0. \tag{10.72}$$

The result is the diffraction pattern of a single structure, but with N -fold intensity. We can observe this effect when we look at a light source through a misted window, where the small water droplets act as diffracting structures. Another example is the halo around the Moon on not completely clear nights. The diffracting structures do not need to be located in a plane, because our theory also holds for light diffraction at centers that are randomly distributed in space.

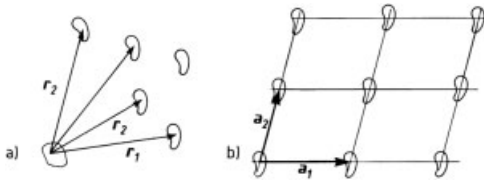


Figure 10.15: Diffraction by (a) randomly distributed structures and (b) by a regular lattice.

(b) Regular lattice (fig. 10.15b)

In this case, $\mathbf{r}_{n_1, n_2} = n_1 \mathbf{a}_1 + n_2 \mathbf{a}_2$ ($0 \leq n_i \leq N_i - 1$), where $\mathbf{a}_1$ and $\mathbf{a}_2$ are fixed lattice vectors. The sum of the geometric series yields

$$\begin{aligned} A &= \sum_{n_1=0}^{N_1-1} e^{-ik\mathbf{p}' \cdot \mathbf{a}_1 n_1} \sum_{n_2=0}^{N_2-1} e^{-ik\mathbf{p}' \cdot \mathbf{a}_2 n_2} \\ &= \frac{e^{-ik\mathbf{p}' \cdot \mathbf{a}_1 N_1} - 1}{e^{-ik\mathbf{p}' \cdot \mathbf{a}_1} - 1} \frac{e^{-ik\mathbf{p}' \cdot \mathbf{a}_2 N_2} - 1}{e^{-ik\mathbf{p}' \cdot \mathbf{a}_2} - 1} \end{aligned} \quad (10.73)$$

and

$$|A|^2 = \frac{\sin^2(N_1 \pi \mathbf{p}' \cdot \mathbf{a}_1 / \lambda)}{\sin^2(\pi \mathbf{p}' \cdot \mathbf{a}_1 / \lambda)} \frac{\sin^2(N_2 \pi \mathbf{p}' \cdot \mathbf{a}_2 / \lambda)}{\sin^2(\pi \mathbf{p}' \cdot \mathbf{a}_2 / \lambda)}. \quad (10.74)$$

For a linear arrangement, $N_2 = 1$ and the second factor drops out. For a spatial distribution, a so-called *spatial lattice*, we obtain a third factor of the same type.

The minima of $|A|^2$ are located at integral values of $m_i = (N_i \mathbf{p}' \cdot \mathbf{a}_i) / \lambda$, where $0 < m_i < N_i$. The values $m_i = 0$ and $m_i = N_i$ correspond to maxima. For increasing values of N_i , these so-called *principal maxima* become increasingly steeper and smaller (the width is of the order $1/N_i$), and for large N_i they are considerably higher than the so-called *secondary* or *side maxima* between the principal minima. In this case, the total intensity

$$I = |U_0|^2 |A|^2 \quad (10.75)$$

is sharply concentrated at the principal maxima of $|A|^2$. This property is used for the construction of diffraction gratings ($N_2 = 1$ and $N_1 = N$ large), where the resolution $\lambda / \Delta \lambda$ for different wavelengths is given by $\lambda / \Delta \lambda = N$.

For a spatial lattice, the principal maxima are determined by Bragg's condition: $\mathbf{p}' \cdot \mathbf{a}_i / \lambda = M_i$, where $i = 1, 2, 3$.

(c) Two circular holes

Finally, we discuss diffraction at two circular holes of radius a separated by a distance $r > a$. In this case $N_1 = 2$ and $N_2 = 1$, and U is given by

$$|U|^2 = |U_0|^2 \left(\frac{\sin(2\pi p' r / \lambda)}{\sin(\pi p' r / \lambda)} \right)^2 = 4 |U_0|^2 \cos^2 \left(\frac{\pi p' r}{\lambda} \right), \quad (10.76)$$

where $|U_0|^2$ is Airy's diffraction pattern. The resulting diffraction pattern is shown in fig. 10.16: the circles of Airy's pattern are traversed by parallel stripes of maxima and minima.

10.8 Optical image processing in Fourier space

We now consider an optical system that maps the object plane S, where the pupil function $P(\xi)$ is defined, onto the image plane B (fig. 10.17). This mapping associates to $P(\xi)$ an amplitude



Figure 10.16: Diffraction pattern of two circular holes.

and phase distribution $P'(\xi) = e^{i\alpha}vP(v\xi)$ in the image plane B, which differs from $P(\xi)$ only by a (positive or negative) scale factor v and a net phase α . This follows from the fact that the mapping from S to B is stigmatic, and the optical distance between conjugated points is always equal. The pre-factor v is a consequence of the conservation of the total intensity. Located between S and B is the focal plane F, where (up to a scale transformation) we find the Fourier transform $\tilde{P}$ of P as the amplitude and phase distribution of the wave function. Placing a screen at F will allow us to see directly $|\tilde{P}|^2$ as the intensity distribution. Hence, the part of the optical system that is located between F and B effectively induces a back-transformation from the Fourier transform $\tilde{P}$ to the original pupil function (up to scale factors).

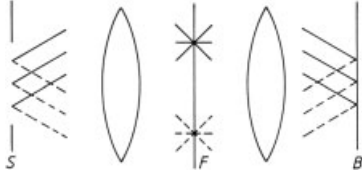


Figure 10.17: Imaging the object plane S onto the image plane B. When S is described by the pupil function P , we find in the focal plane F, up to a scale factor, the Fourier transform $\tilde{P}$.

Strictly speaking, because the plane F can never have an infinite extension, the components of $\tilde{P}(\mathbf{p})$ are cut off for large values of $\mathbf{p}$. This is just another way to understand the restrictions on the resolution of a system due to diffraction effects.

Denoting by $\tilde{P}(\mathbf{p})$ the Fourier transformation of $P(\xi)$, we find the following well-known relations between a function and its Fourier transform:

$$\begin{aligned}
 \text{(i)} \quad & P(\xi + \mathbf{a}) \quad \text{and} \quad e^{i\mathbf{p} \cdot \mathbf{a}} \tilde{P}(\mathbf{p}), \\
 \text{(ii)} \quad & e^{i\mathbf{b} \cdot \xi} P(\xi) \quad \text{and} \quad \tilde{P}(\mathbf{p} - \mathbf{b}), \\
 \text{(iii)} \quad & P(\alpha\xi) \quad \text{and} \quad (1/\alpha^2) \tilde{P}((1/\alpha)\mathbf{p}), \\
 \text{(iv)} \quad & P^*(\xi) \quad \text{and} \quad \tilde{P}^*(-\mathbf{p}).
 \end{aligned} \tag{10.77}$$

These relations may be interpreted as follows:

- (i) A translation of P corresponds to a simple phase shift of the image at F, while the intensity distribution $|\tilde{P}|^2$ remains unchanged.
- (ii) A linear phase shift in ξ of the pupil function P corresponds to a translation of the image at F.
- (iii) A dilation of P corresponds to a contraction of $\tilde{P}$: the smaller the structure at S, the more extended it appears at F.

- (iv) If P is real at S (or has a fixed phase), like e.g. for an amplitude object that is illuminated from a direction perpendicular to S , we have $|\tilde{P}(\mathbf{p})|^2 = |\tilde{P}(-\mathbf{p})|^2$, i.e. the diffraction pattern is mirror symmetric around the origin.

The image at B can be influenced by placing suitable filters in plane F , which either induce specific phase shifts or suppress certain Fourier components with a template.

For a better understanding, we consider $\tilde{P}(\mathbf{p})$ for various pupil functions. If S consists of a simple large hole without any primary object, we have $P \equiv 1$ and $\tilde{P}(\mathbf{p}) = \text{const.} \times \delta(\mathbf{p})$. In the plane F , we only observe a central maximum at the origin. Figure 10.18 shows the Fourier transforms for simple line and net structures as well as for a lattice structure with additional imperfections.

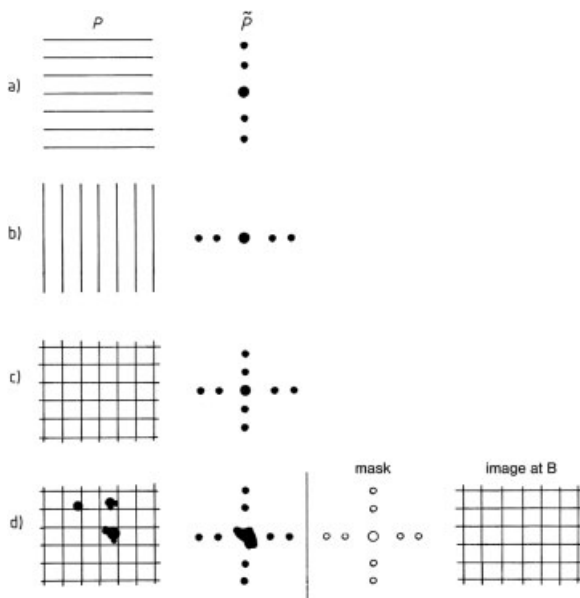


Figure 10.18: The Fourier transforms for various pupil functions. One can eliminate the interfering structures in (d) by placing a special mask at F that allows only the Fourier transform of the lattice to pass through it.

If we mask the Fourier components corresponding to $\tilde{P}$ at plane F , we can eliminate the image of P at plane B . In principle, any unwanted structure can be suppressed by a suitable mask. We obtain such a mask by placing the structure as a pupil function at plane S and making a photograph of its image on plane F . In this way, we can suppress the irritating raster in a newspaper picture. We can eliminate the disturbing spots in fig. 10.18(d) with the help of a mask that has small holes where the maxima of the Fourier transform of the lattice net are located. Placing a circular aperture at F around the center and making the radius of the hole smaller, the lattice pattern in the image of the pupil function becomes gradually diffused and eventually vanishes.

We can learn from (iii) that, in general, a filter whose transparency decreases with increasing distance from the center will suppress the higher Fourier components and thus the smaller structures. The result is a less pronounced contrast. On the other hand, a filter at F with increasing transparency away from the center enlarges the contrast of the image at B.

An interesting effect is achieved by a mask that suppresses the central maximum at $\mathbf{p} = 0$. In the absence of a mask, uniformly dark regions at B result from interference of the homogeneous background field stemming from the central maximum with the other structures. If the central maximum is masked, these previously dark regions become light. Sharp transitions between dark and light areas at S will be visible as linear structures at B.

Many microorganisms are phase objects when we consider them as optical structures. They are uniformly transparent but according to their structure their inner refractive index exhibits modulations. Such phase objects are described by pupil functions of the form

$$P(\xi) = A e^{ik\varphi(\xi)}, \quad A = \text{const.} \quad (10.78)$$

The Fourier transform $|\tilde{P}|^2$ usually corresponds at F to a structured intensity distribution, which, after Fourier back-transformation, returns to a uniform intensity distribution at B. In order to make phase objects observable at B, one first has to transform them into amplitude objects. For biological objects, this is often achieved by coloring, which, however, usually kills the organism. A better solution consists in processing the image at F.

Since the phase differences are small, we may assume that $k\varphi(\xi) \ll 1$. The general case can be treated in a similar way with only a manageable increase in effort. Up to first order in $k\varphi$ we obtain

$$P(\xi) = A[1 + ik\varphi(\xi)], \quad \text{i.e.} \quad |P|^2 = |A|^2[1 + \mathcal{O}(k^2\varphi^2)],$$

and the intensity variations are of the order of $k^2\varphi^2$. The Fourier transform of P is given by

$$\tilde{P}(\mathbf{p}) = A[4\pi^2\delta(\mathbf{p}) + ik\tilde{\varphi}(\mathbf{p})]. \quad (10.79)$$

There are various methods to manipulate $\tilde{P}$, three of which are as follows.

(a) The dark-ground method

An opaque disk placed in the center of F removes the central maximum. The resulting distribution on F is given by

$$\tilde{P}'(\mathbf{p}) = iAk\tilde{\varphi}(\mathbf{p}),$$

from which we obtain after a back-transformation to the plane B:

$$P'(\xi) = iAk\varphi(\xi).$$

If no phase object is present, we have $\varphi(\xi) \equiv 0$, and B appears dark. In front of this dark background the phase object is visible as the intensity distribution

$$|P'(\xi)|^2 = |A|^2 k^2 \varphi(\xi)^2.$$

(b) The phase contrast method

This was developed by F. Zernicke (Nobel prize, 1953). A transparent filter in the center shifts the phase of the central maximum by $+\pi/2$ or $-\pi/2$. Now,

$$\tilde{P}'(\mathbf{p}) = A[\pm 4\pi^2 i \delta(\mathbf{p}) + ik\tilde{\varphi}(\mathbf{p})],$$

and the back-transformation leads to

$$P'(\boldsymbol{\xi}) = iA[\pm 1 + k\varphi(\boldsymbol{\xi})].$$

The intensity,

$$|P'|^2 = |A|^2\{1 \pm 2k\varphi(\boldsymbol{\xi}) + [k\varphi(\boldsymbol{\xi})]^2\},$$

exhibits a modulation of $\varphi(\boldsymbol{\xi})$ in first order. The phase object becomes visible as an amplitude object at B.

(c) The Schlieren method

In practice, a higher intensity for dark-ground or phase contrast microscopes is achieved not by using plane waves for the primary illumination but by using the congruent light from ring-shaped sources. In this case, the zeroth-order maximum at F is also ring-shaped and can be manipulated by a ring-shaped mask.

A variant of this method, already discovered in 1864 by August Toepler, is the so-called *Schlieren method* (the word “Schlieren” refers to the German word for streaks or striae), which is of special importance in fluid dynamics. With this method one can make visible the streaks – small variants of the refractive index – in a fluid.

In this arrangement a small fluid-filled cell is placed at S. A knife-edge is introduced in the focal plane and covers about one-half of the image at F. When the edge approaches the focal point, the streaks become perceptible in the form of cloudy structures on the screen B, and their spatial and temporal fluctuations are easily observed.

10.9 Morse families

We have already mentioned at the end of section 9.6 that in the vicinity of caustics it can be helpful to generalize the simple ansatz

$$u(x) = e^{ikS(x)}\varphi_k(x)$$

of the short-wave asymptotics into an integral ansatz of the form

$$u(x) = \left(\frac{k}{2\pi}\right)^{r/2} \int_{\mathbb{R}^r} d^r a e^{ikS(x;a)} \varphi_k(x, a). \quad (10.80)$$

In the preceding sections, integrals of this type have occurred on several occasions. The integration runs over r parameters $(a_1, \dots, a_r)$, and in order to avoid convergence problems

we assume that the integrand has compact support as a function of a . The method of stationary phases reveals that for $k \rightarrow \infty$ only the points in

$$M_S = \left\{ (x, a) \left| \frac{\partial S}{\partial a}(x; a) = 0 \right. \right\} \subset Q \times \mathbb{R}^r \quad (10.81)$$

contribute. Furthermore, we shall always assume that the $r \times (r+n)$ matrix of second derivatives has rank r ,

$$\text{rank} \left(\frac{\partial^2 S}{\partial a^i \partial a^j} \left| \frac{\partial^2 S}{\partial a^i \partial x^k} \right. \right) = r \quad (i, j = 1, \dots, r; k = 1, \dots, n). \quad (10.82)$$

If this regularity condition is fulfilled, M_S is an n -dimensional submanifold of $Q \times \mathbb{R}^r$, and the set

$$\Lambda_S = \left\{ (x, p) \in P \left| \frac{\partial S}{\partial a}(x; a) = 0, p = \frac{\partial S}{\partial x}(x; a) \right. \right\} \subset P \quad (10.83)$$

is a Lagrangian submanifold of the phase space $P = (T^*Q)$.

The 2-form $\omega = \sum dx^i \wedge dp_i = -d(\sum p_i dx^i)$ vanishes on Λ_S , which is an immediate consequence of the fact that on Λ_S we have

$$\sum p_i dx^i = \sum \frac{\partial S}{\partial x^i} dx^i = dS.$$

Furthermore, the regularity condition (10.82) guarantees that Λ_S has maximal dimension n .

If a Lagrangian manifold Λ_S and a family of functions $S(x; a)$ are related in the way described by eq. (10.83), $S(x; a)$ is called a *Morse family* for Λ_S . For a given Lagrangian submanifold Λ , there exist many Morse families S that locally parametrize Λ , i.e. for which locally $\Lambda = \Lambda_S$. It turns out that in the vicinity of a caustic of order r one needs at least r auxiliary parameters a_i and that there always exists a local parametrization with r parameters. In many cases one chooses certain momentum coordinates p as auxiliary parameters.

A Morse family $S(x; a)$ is a generalization of the generating function $S(x)$ of a Lagrangian manifold Λ_S . Indeed, away from the caustics, Lagrangian manifolds are always of the form

$$\Lambda_S = \left\{ (x, p) \left| p = \frac{\partial S(x)}{\partial x} \right. \right\}. \quad (10.84)$$

In chapter 8 we have seen that the graph of a canonical transformation $\Phi : P \rightarrow P$ is a Lagrangian submanifold of $P \times P$ if we equip $P \times P$ with the symplectic form

$$\Omega = \sum dx^i \wedge dp_i - \sum dx_0^i \wedge dp_{0i}. \quad (10.85)$$

Also, $S(x, x_0; a)$ is a Morse family for a canonical transformation Φ if S is a Morse family for graph Φ . This is a generalization of the generating function of a canonical transformation.

The following composition rules hold:

- If $S_1(x_0; a)$ is a Morse family for $\Lambda \subset P$ and $S_2(x, x_0; b)$ is a Morse family for a canonical transformation $\Phi : P \rightarrow P$, then

$$S(x; x_0, a, b) = S_2(x, x_0; b) + S_1(x_0; a) \quad (10.86)$$

is a Morse family (with auxiliary parameters x_0 , a , and b) for the Lagrangian manifold $\Phi(\Lambda) \subset P$, which is the image of Λ under the mapping Φ .

- If $S_2(x, x_1; a)$ is a Morse family for $\Phi_2 : P \rightarrow P$ and $S_1(x, x_1; b)$ is a Morse family for $\Phi_1 : P \rightarrow P$, then

$$S(x, x_0; x_1, a, b) = S_2(x, x_1; a) + S_1(x_1, x_0; b) \quad (10.87)$$

is a Morse family (with parameters x_1 , a , and b) for the combined canonical transformation $\Phi_2 \circ \Phi_1 : P \rightarrow P$.

The proofs of these statements follow from the conditions of stationarity:

$$\begin{aligned} 0 &= \frac{\partial S(x; x_0, a, b)}{\partial x_0} \\ &= \frac{\partial S_2(x, x_0; a)}{\partial x_0} + \frac{\partial S_1(x_0; a)}{\partial x_0} = -p_0^{(2)} + p_0^{(1)}; \\ 0 &= \frac{\partial S(x, x_0; x_1, a, b)}{\partial x_1} \\ &= \frac{\partial S_2(x, x_1; a)}{\partial x_1} + \frac{\partial S_1(x_1, x_0; b)}{\partial x_1} = -p_1^{(2)} + p_1^{(1)}. \end{aligned} \quad (10.88)$$

Finally, the Maslov bundle $M(\Lambda)$ and the Maslov index of a path $\tilde{\gamma}$ in Λ can be defined using Morse families. We have seen that both are related to the phases associated with a crossing of caustics.

The contribution of a stationary point a_0 to the integral in eq. (10.80) is of the form

$$u^{a_0}(x) = \frac{\exp[(i\pi/4) \operatorname{sign} S''_{aa}(x; a_0)]}{\sqrt{|\det S''_{aa}(x; a_0)|}} e^{ikS(x; a_0)} \varphi(x, a_0). \quad (10.89)$$

The phases of the asymptotic expansion are given by the signature of the matrix

$$(S_{aa}(x, a_0))_{ij} = \frac{\partial^2 S}{\partial a_i \partial a_j}(x, a_0). \quad (10.90)$$

Now, let $n \in \Lambda$ and denote by $\mathcal{S}(n)$ the set of all Morse families parametrizing Λ in some neighborhood of n . The fiber of the Maslov bundle $M(\Lambda)$ at point $m \in \Lambda$ is defined to be the set of all mappings

$$\psi : \mathcal{S}(n) \longrightarrow \mathbb{C}$$

with the property that, for all $S_1(x; a), S_2(x; a') \in \mathcal{S}(n)$, we have

$$\psi(S_1) = \exp \left\{ \frac{i\pi}{4} \left[\text{sign} \left(\frac{\partial^2 S_1}{\partial a^2} \right) - \text{sign} \left(\frac{\partial^2 S_2}{\partial a'^2} \right) \right] \right\} \psi(S_2). \quad (10.91)$$

We can now define the Maslov index $\tilde{\mu}(\tilde{\gamma})$ for some path $\tilde{\gamma}$ in Λ as follows. Let $\tilde{\gamma} : [0, 1] \rightarrow \Lambda$ be a path in the Lagrangian submanifold Λ . We choose points $0 = t_0 < t_1 < \dots < t_k = 1$ such that, for some suitable Morse families S_i , we have $\tilde{\gamma}([t_{i-1}, t_i]) \subset \Lambda_{S_i}$. We now define

$$\tilde{\mu}(\tilde{\gamma}) = -\frac{1}{2} \sum_{i=1}^k \left[\text{sign} \left(\frac{\partial^2 S_i}{\partial a_i^2}(\tilde{\gamma}(t_i)) \right) - \text{sign} \left(\frac{\partial^2 S_i}{\partial a_i^2}(\tilde{\gamma}(t_{i-1})) \right) \right]. \quad (10.92)$$

The proof of the equivalence of this definition with the definition given in chapter 9 is quite elaborate.

10.10 Oscillatory functions and Fourier integral operators

In this section we will introduce the concept of oscillatory functions and Fourier integral operators. This will be done in a slightly simplified form and without claiming mathematical rigor. These notions fit exactly the needs of theoretical optics.

Let Q be a manifold. If a function $f : Q \rightarrow \mathbb{C}$ can be written as a locally finite sum of integrals of the form

$$I(x, k) = \left(\frac{k}{2\pi} \right)^{r/2} \int d^r a e^{ikS(x; a)} \varphi_k(x, a), \quad (10.93)$$

it is called an *oscillatory function* on Q .

As in the previous section, we assume that $\varphi_k(x, a)$ has compact support with respect to a and that $S(x; a)$ is a Morse family that locally parametrizes a Lagrangian submanifold $\Lambda \subset T^*Q$. In this case one speaks of f as an oscillatory function associated to Λ .

Local finiteness means that for each point $x \in Q$ only finitely many terms of the form (10.93) contribute to the sum. In section 9.7 we have seen that it may be useful to generalize this definition slightly in such a way that f assumes values in the product $\Omega_{n/2}(\Lambda) \otimes M(\Lambda)$ of half-density bundles and Maslov bundles. Expanding with respect to stationary phases, one notes that in the limit $k \rightarrow \infty$ the integral (10.93) really “lives” on the manifold

$$M_S = \left\{ (x, a) \left| \frac{\partial S(x; a)}{\partial a} = 0 \right. \right\} \subset Q \times \mathbb{R}^r,$$

or, similarly, on the Lagrangian submanifold Λ_S that one obtains from M_S by the following injection (“Lagrange immersion”):

$$\begin{aligned} i_S : M_S &\longrightarrow \Lambda_S \subset T^*Q \\ (x, a) &\longmapsto \left(x, \frac{\partial S(x; a)}{\partial x} \right) \in T^*Q. \end{aligned} \quad (10.94)$$

More precisely, in the limit $k \rightarrow \infty$ it is only the intersection of M_S with the support of $\varphi_k(x, a)$ and its image under i_S that matters:

$$i_S(M_S \cap \text{supp } \varphi_k) = F(I). \quad (10.95)$$

This Lagrangian submanifold $F(I) \subset \Lambda_S$ is called the *frequency set* of the oscillatory integral I . If f is a locally finite sum of integrals I_i of the form (10.93), the frequency set on which f “lives” is given by

$$F(f) \subseteq \bigcup_i F(I_i). \quad (10.96)$$

If, for some function

$$g = e^{-ik\psi(x)}\chi(x), \quad (10.97)$$

the intersection of the set

$$M_g = \left\{ (x, p) \mid x \in \text{supp } \chi, p = \frac{\partial \psi}{\partial x} \right\} \subset T^*Q \quad (10.98)$$

with the set $F(I)$ is empty, Theorem 10.1 on stationary phases tells us that the function

$$\begin{aligned} A(k) = \langle I, g \rangle &= \left(\frac{k}{2\pi} \right)^{n/2} \int d^n x I(x, k) g(x) \\ &= \left(\frac{k}{2\pi} \right)^{(n+r)/2} \int d^n x d^r a e^{ikS(x; a) - ik\psi(x)} \psi_k(x, a) \chi(x) \end{aligned} \quad (10.99)$$

will vanish in the limit $k \rightarrow \infty$ faster than any power of k^{-N} . In this sense I is concentrated on $F(I)$. For an arbitrary oscillatory function f , the frequency set $F(f)$ is defined in such a way that for $M_g \cap F(f) = \emptyset$ we have

$$\lim_{k \rightarrow \infty} k^{-N} \langle f, g \rangle = 0 \quad \text{for all } N \geq 0. \quad (10.100)$$

The requirement that the integrand in eq. (10.93) has compact support with respect to a is still too limiting. We can get rid of it by considering locally finite sums of integrals (10.93) also with respect to a . This idea is behind the following definition of an *oscillatory distribution*.

Definition 10.1. Let $\mathcal{A} = \{U_\alpha\}$ be a locally finite covering of T^*Q with open and relatively compact sets U_α . Furthermore, let $\Lambda \subset T^*Q$ be a Lagrangian submanifold. The locally finite sum

$$f = \sum_\alpha f_\alpha \quad (10.101)$$

is called an oscillatory distribution associated to Λ if f_α are oscillatory functions associated to Λ with compact support and $F(f_\alpha) \subset U_\alpha$.

Also for oscillatory distributions the concept of a frequency set remains meaningful. The asymptotically leading part $\sigma(f)$ of f , which we obtain by the method of stationary phases, is called the *symbol* of f or sometimes the *principal symbol*.

Let $OF_0(Q)$ be the space of oscillatory functions with compact support in Q . A linear operator

$$K : OF_0(Q) \longrightarrow OF_0(Q),$$

$$(Ku)(x) = \int d^n x' K(x, x') u(x') \quad (10.102)$$

is called a *Fourier integral operator* if the integral kernel $K(x, x')$ is an oscillatory distribution on $Q \times Q$. The integral kernel K is associated to a canonical transformation $\Phi : T^*Q \rightarrow T^*Q$ if $K(x, x')$ is associated to the Lagrangian submanifold graph $\Phi \subset T^*Q \times T^*Q$. In general, the integral kernel $K(x, x')$ of a Fourier integral operator is the sum of integrals of the form

$$I(x, x') = \left(\frac{k}{2\pi}\right)^r \int d^n p d^r a e^{ikp(x-x')} \varphi_k(x, x'; a).$$

We now discuss some particularly important examples of Fourier integral operators for the case $Q = \mathbb{R}^n$.

(a) Unit operator

Its kernel is

$$K_\delta(x, x') = \left(\frac{k}{2\pi}\right)^n \int d^n p e^{ikp \cdot (x-x')} = \delta(x - x'). \quad (10.103)$$

The Morse family $S(x, x'; p) = p \cdot (x - x')$ parametrizes the Lagrangian manifold

$$F(K_\delta) = \Lambda_\delta = \{((x, p), (x, -p))\} \subset T^*Q \times T^*Q. \quad (10.104)$$

As a graph Λ corresponds to the identity transformation. The asymptotic expansion is trivial. At the stationary point we have $(\partial S / \partial p) = 0$, i.e. $S(x, x'; p) = 0$, and we find

$$\sigma(K_\delta) = \left(\frac{k}{2\pi}\right)^n. \quad (10.105)$$

(b) Invertible point transformation $x' \rightarrow h(x')$

Now, the kernel is given by

$$K_h(x, x') = \left(\frac{k}{2\pi}\right)^{n/2} \int d^n p e^{ikp \cdot [x - h(x')]} = \delta(x - h(x')). \quad (10.106)$$

For the frequency set we find

$$F(K_h) = \Lambda_h = \{(h(x'), p), (x', -(\partial h / \partial x') p)\}, \quad (10.107)$$

which corresponds to the graph of the following canonical transformation:

$$\Phi : (x, p) \mapsto \left(h(x), \left(\frac{\partial h}{\partial x} \right)^{-1} p \right). \quad (10.108)$$

Again

$$\sigma(K_h) = \left(\frac{k}{2\pi} \right)^n. \quad (10.109)$$

(c) Differential operator $H(x, -(\mathbf{i}/k)\partial)$

This is like that used in section 9.1. We have

$$\begin{aligned} (Hu)(x) &= H\left(x, -\frac{\mathbf{i}}{k}\partial\right) \left(\frac{k}{2\pi}\right)^{n/2} \int d^n p e^{ikp \cdot x} \tilde{u}(p) \\ &= \left(\frac{k}{2\pi}\right)^{n/2} \int d^n p e^{ikp \cdot x} H(x, p) \tilde{u}(p) \\ &= \left(\frac{k}{2\pi}\right)^n \int d^n x' \int d^n p e^{ikp \cdot (x-x')} H(x, p) u(x'), \end{aligned}$$

from which we obtain for the kernel of the operator:

$$K_H(x, x') = \left(\frac{k}{2\pi}\right)^n \int d^n p e^{ikp \cdot (x-x')} H(x, p). \quad (10.110)$$

In this sense, differential operators are also Fourier integral operators. The Morse family $S(x, x'; p) = p \cdot (x - x')$ is the same as for the unit operator $\delta(x - x')$.

Hence, all differential operators are associated to the canonical unit transformation; furthermore, $S(x, x'; p)$ vanishes on Λ_S . These two properties even characterize the set of pseudo-differential operators among all Fourier integral operators.

Obviously, the symbol of H is

$$\sigma_H = \left(\frac{k}{2\pi}\right)^n H(x, p), \quad (10.111)$$

and it agrees, up to a trivial factor, with the symbol of H that we defined in section 9.1. This justifies the name “symbol” for the asymptotically leading term.

The following elementary operations involving oscillatory functions and Fourier integral operators seem close at hand:

1. Application of a Fourier integral operator K to an oscillatory function u with compact support.
2. Composition of two Fourier integral operators K_1 and K_2 , for which the composition $K_1 \circ K_2$ is also a Fourier integral operator.

A third operation might be less obvious:

3. Transition to the adjoint $K^\dagger$ of a Fourier integral operator.

It is possible to formulate a so-called *symbol calculus* which, for given symbols or, more generally, for asymptotic expansions of K_1 , K_2 , and u , allows one to determine the symbols of the asymptotic expansions of $K_1 u$, $K_1 \circ K_2$, K_1^+ , and K_2^+ . The calculations we performed in chapter 9 in the context of asymptotic expansions are examples of such operations. The symbol calculus is closely related to the notion of star products, which will be explained in chapter 13.

Recalling what we have said in the previous section 10.9 about Morse families makes the following statements obvious. Let the Fourier integral operators K_1 and K_2 be associated to the canonical transformations Φ_1 and Φ_2 , respectively, and let u be an oscillatory function with compact support associated to $\Lambda \subset T^*Q$. Then

- $K_1 u$ is associated to the Lagrangian submanifold $\Phi_1(\Lambda)$, the image of Λ under Φ_1 ;
- $K_1 \circ K_2$ is associated to the canonical transformation $\Phi_1 \circ \Phi_2$.

Furthermore, the frequency sets satisfy the following locality property:

$$F(K_1 u) \subset \Phi_1(F(u)).$$

A brief look at the content of the previous and the following chapters reveals that the subject of theoretical optics as a theory of wave fields is packed with oscillatory functions and Fourier integral operators.

11 Holography

11.1 The principle of holography

In the previous chapter, notably in section 10.3, we have seen that for fixed frequency $\omega = kc$ a solution $u(\mathbf{x})$ of the wave equation

$$(\Delta + k^2 n^2)u(\mathbf{x}) = 0 \quad (11.1)$$

is already determined by its values on a surface S , at least under the condition that the solution satisfies the radiation condition at infinity.

Now let $u(\mathbf{x})$ be the (monochromatic) wave radiating from some object G . If one could generate a wave on S that coincides completely, in amplitude and phase, with $u(\mathbf{x})$ on S , an observer B behind S would not be able to distinguish this wave from the original wave $u(\mathbf{x})$. In particular, the observer can see the image of G from different perspectives. If the amplitude and phase of $u(\mathbf{x})$ are only produced on a segment A of S , the wave of G can only be reconstructed for a certain range of volume angles for the observer B (fig. 11.1).

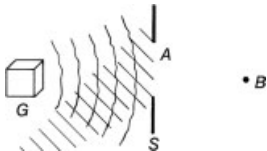


Figure 11.1: Superposition of an object wave and a reference wave during the taking of a hologram.

In principle, the reconstruction of a wave for such a range of volume angles could be realized by illuminating a special type of slide, placed in the aperture A of the screen, with a planar wave. The transparency of this slide should reproduce the amplitude of the wave, and the optical thickness of the slide should reproduce the phase of the wave on S . However, technically such a slide is almost impossible to realize, because it is difficult to measure the phase of u at a point.

Using a special trick, however, holography comes quite close to this ideal. The idea is to superimpose the primary wave $u(\mathbf{x})$ emanating from the object with a *reference wave* $u_0(\mathbf{x}) = a_0 e^{i\mathbf{k}_0 \cdot \mathbf{x}}$ (with $\mathbf{k}_0^2 = k^2$). Denoting by $\boldsymbol{\xi}$ some coordinates on S , the intensity of the superposition $u_0(\mathbf{x}) + u(\mathbf{x})$ on S is given by

$$\begin{aligned} I(\boldsymbol{\xi}) &= |u_0 + u|^2 \\ &= |a_0|^2 + a_0^* u(\boldsymbol{\xi}) \exp(-i\mathbf{k}_0 \cdot \boldsymbol{\xi}) + a_0 u^*(\boldsymbol{\xi}) \exp(+i\mathbf{k}_0 \cdot \boldsymbol{\xi}) + |u(\boldsymbol{\xi})|^2. \end{aligned} \quad (11.2)$$

We assume that the intensity of the reference wave on S is much higher than the intensity of $u(\boldsymbol{\xi})$, so that the term $|u(\boldsymbol{\xi})|^2$ in $I(\boldsymbol{\xi})$ can be neglected.

If we place a photographic film at S (or A), the blackening of the film will be proportional to the intensity I and the developed film will exhibit a complicated pattern of bright and dark structures, whose transparency is given by

$$t(\xi) = 1 - \gamma I(\xi), \quad (11.3)$$

and which for the naked eye has no obvious similarity with the object G.

This slide is now placed at A and illuminated from behind by a *reconstruction wave* $u_1(x) = a_1 e^{ik_1 \cdot x}$. Immediately behind the film, which is often called a *hologram*, the wave will be

$$\begin{aligned} w(\xi) &= t(\xi)u_1(\xi) \\ &= a_1(1 - \gamma|a_0|^2) \exp(ik_1 \cdot \xi) - \gamma a_1 a_0^* u(\xi) \exp[i(k_1 - k_0) \cdot \xi] \\ &\quad - \gamma a_1 a_0 u^*(\xi) \exp[i(k_1 + k_0) \cdot \xi]. \end{aligned} \quad (11.4)$$

We first set $k_1 = k_0$. Within a certain region for the volume angle set by the aperture A, an observer B will observe a diffracted wave that is the superposition of three amplitudes:

1. a wave $e^{ik_0 \cdot x}$;
2. a wave that coincides up to a constant factor with the wave $u(x)$ coming from the object G; and
3. a wave that differs by a factor $-\gamma a_1 a_0 e^{2ik_0 \cdot \xi}$ from the complex-conjugate wave u^* on A.

Next we assume that at A the direction of the reference wave is inclined with respect to the preferred direction k_G/k of the wave vector for the object wave u . In this case, the three waves are concentrated around different wave vectors in Fourier space, which are

1. $k = k_0$,
2. $k = k_G$, and
3. $k = k(k_G + 2k_0)/|k_G + 2k_0|$.

For an observer B the waves arrive from different directions. Within a certain volume angle around k_G/k the observer only perceives one wave that is (almost) not superimposed by the other waves and is indistinguishable from the object wave $u(x)$. Thereby one can reconstruct, at least in certain regions of the volume angle, the total wave field u and thus a three-dimensional picture of G.

In other directions one can observe the reference wave or the complex-conjugate wave. This conjugate wave gives the observer a picture of G that is inverted in a bizarre form. First, because of the factor $e^{2ik \cdot \xi}$, this picture seems to come from a different direction. Second, because of complex conjugation, there is an inversion of phase shifts. This has the effect that the front and back parts of G seem to be perspectively inverted, while at the same time the image of the intersections of the different parts of G seem to have the correct spatial relations. Because of these irritating effects, the image of the complex-conjugate wave is rather annoying; however, since $t(\xi)$ is real it cannot be avoided.

There are other situations for which the construction of phase conjugated mirrors producing the complex-conjugate wave $u^*(\mathbf{x})$ for a wave $u(\mathbf{x})$ are of great interest (see chapter 6).

It is also interesting to observe that in a certain way any part of the hologram contains the full information about the object G. Indeed, using only a part of the hologram corresponds to a smaller window A and thus to a smaller region of volume angles where G is visible. One can think of a hologram as a planar diffraction grating whose diffractive image reproduces a given object G.

There are no principle changes for the case in which the reconstruction wave and the reference wave are different. For $\mathbf{k}_1 = \alpha \mathbf{k}_0$, the image of G appears in a different color and seems to be reduced by a factor α . If the directions of $\mathbf{k}_0$ and $\mathbf{k}_1$ are different, the image of G also appears in a different direction.

From the arguments given above it should be obvious that the reference wave and the reconstruction wave do not have to be planar. Also suitable are waves u_0 and u_1 that on A have the forms $u_0(\boldsymbol{\xi}) = a_0(\boldsymbol{\xi}) \exp[ik\psi(\boldsymbol{\xi})]$ and $u_1(\boldsymbol{\xi}) = \exp(i\mathbf{q} \cdot \boldsymbol{\xi})u_0(\boldsymbol{\xi})$, respectively, with only weak variations of $a_0(\boldsymbol{\xi})$.

11.2 Modifications and applications

Ever since its discovery by Dennis Gabor (Nobel prize, 1971) in the late 1940s, the importance of holography has increased substantially due to steady technological progress. The biggest technical problem consists in keeping the phase differences of the wave field stable among the different points of the aperture A during the exposure of the hologram, because otherwise there will be no sharp interference image. The necessary “coherence” of the wave field can only be achieved by laser light, and the brief exposure times can be realized with pulsed high-power lasers and highly sensitive photographic emulsions. The progress of this development is revealed by the fact that holographic portraits already exist.

The basic arrangement described in the last section can be modified in many ways:

1. Apart from the *amplitude holograms* described above, *phase holograms* are also in use. In this case it is not the transparency that is influenced by the exposure of a film but the refractive index, such that the hologram produces a spatially modulated phase shift of the reconstruction wave. For a phase hologram, the transmission $t(\boldsymbol{\xi})$ in eq. (11.3) has to be replaced by $t(\boldsymbol{\xi}) = \exp[i\gamma I(\boldsymbol{\xi})] \approx 1 + i\gamma I(\boldsymbol{\xi})$, with $\gamma I(\boldsymbol{\xi}) \ll 1$. Up to an irrelevant factor of $-i$, one obtains the same conditions as for an amplitude hologram.
2. The reference wave can also consist of spherical waves whose radius of curvature at A is adjusted to the radius of curvature of the wavefronts of the object wave. This reduces fluctuations in the distances between the maxima and minima of the intensity distribution $I(\boldsymbol{\xi})$ for different parts of the hologram and thereby diminishes the demands for the resolution of the emulsions.
3. Colored spatial pictures are produced by the superposition of holograms for different colors.
4. Using the same method, any surface, not only a screen S, in the line of the radiation of some optical instrument can be used for the reconstruction of the wave field in amplitude

and phase. Of special importance is the reconstruction in the focal plane F behind the screen (see fig. 10.17). In this case it is not the pupil function P that is reconstructed on S but its Fourier transform $\tilde{P}$. In this way the given resolution of the photographic emulsion can often be used more efficiently. Such holograms are called *Fourier transform holograms*.

5. Instead of transparent holograms, reflecting holograms are also possible.
6. Holograms can also be generated by ultrasound waves. In the arrangement shown in fig. 11.2, a sound wave is reflected at some submerged object G and interferes with a reference wave of a second correlated sound source. This produces a hologram of ripples at the surface, which can be observed by the reflection of laser light and provides an “X-ray” image of G .

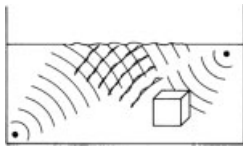


Figure 11.2: Principle of acoustic holography.

We now want to describe briefly a few special applications of holography.

11.2.1 Observing small object deformations

The idea is to compare the holograms of a deformed and an undeformed object, which can be done in several ways:

- Make a double exposure of a film with the holograms of the undeformed and the deformed object.
- Produce a hologram of the undeformed object first and place it at the aperture A ; the object remains in its position and the reference wave also remains unchanged. Now the hologram of the undeformed object is illuminated by the wave field of the deformed object.

In both cases the deformations will show up in the interference patterns of the image seen by the observer B .

- If the object oscillates elastically with small amplitudes during the exposure time, the holographic images reveals nodes and anti-nodes of the vibration.

11.2.2 Holographic optical instruments

When we illuminate a film from one side with a spherical wave and from the other side with a plane reference wave and record the resulting interference pattern, we obtain a hologram that focuses the reconstruction wave into one point, i.e. which acts like a converging lens.

In this way or, more recently, using computer-generated holograms, one can manufacture optical elements that are light and flat and are able to generate any given wave field.

11.2.3 Pattern recognition

If a substructure $g(\xi)$ is part of some pattern $f(\xi)$ at some position ξ_0 (see fig. 11.3), the convolution

$$(f * Ig)(\xi) = \frac{1}{2\pi} \int d^2\eta f(\eta)g(\eta - \xi) =: h(\xi) \quad (11.5)$$

exhibits a steep maximum at $\xi = \xi_0$, where $Ig(\xi) := g(-\xi)$. Due to the convolution theorem, a Fourier transforms leads to

$$\tilde{f}(\mathbf{k})\tilde{Ig}(\mathbf{k}) = \tilde{h}(\mathbf{k}).$$

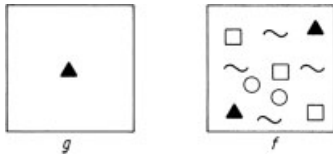


Figure 11.3: Searching for a substructure g in some pattern f .

When we place a Fourier hologram of Ig in the focal plane F and use this as a filter for the interference pattern of f , the function $\tilde{h}(\mathbf{k})$ shows up as the irradiation distribution (see fig. 10.17). After a Fourier back-transformation one observes bright spots in the focal plane B at all places where the structure g occurs in f .

11.3 Volume holograms

There is no need to restrict the recording of the intensity $I(\mathbf{x}) = |u(\mathbf{x}) + u_0(\mathbf{x})|^2$ of the interference field of the reference wave u_0 and the object wave u to the plane of a single film; one can also produce images within a certain volume. Suitable media are thick and light-sensitive emulsion layers or crystals of potassium bromide (KBr), where the exposed parts produce colored centers with variable transparency. Also in use are ferroelectric crystals like lithium niobate and potassium tantalate niobate. They react to exposure with a change of the refractive index and one obtains a spatial phase object.

Such a *volume hologram* can be visualized as a stack of planar holograms. In contrast to the case of planar holograms, it is now important that the reference wave and the reconstruction wave are equal up to a constant factor, otherwise the holographic images of the different planar holograms in the stack would be “off beat” and would in the average almost everywhere cancel each other by interference.

Alternatively, a volume hologram can be described as a spatial lattice that is constructed in such a way that the Bragg diffraction field represents a given object.

Like for planar holograms, the calculations for volume holograms are also easily performed. A monochromatic reconstruction wave $u_1(\mathbf{x}) = a_1 e^{ik_1 \cdot \mathbf{x}}$ with frequency $\omega_1 = c|\mathbf{k}_1|$ illuminates a volume with a certain distribution $t(\mathbf{x})$ of scattering centers. A scattered wave $G_{\omega_1}(\mathbf{x} - \mathbf{x}')$ emanates from a center at point $\mathbf{x}'$ with the same frequency ω_1 , which is proportional to both $t(\mathbf{x}')$ and the value $u_1(\mathbf{x}')$ of the reconstruction wave at point $\mathbf{x}'$. The total

scattered wave is proportional to

$$w(\mathbf{x}) = (2\pi)^{-3/2} \int d^3x' t(\mathbf{x}') e^{i\mathbf{k}_1 \cdot \mathbf{x}'} G_{\omega_1}(\mathbf{x} - \mathbf{x}'). \quad (11.6)$$

(In addition there exists the unscattered primary wave $a'_1 e^{i\mathbf{k}_1 \cdot \mathbf{x}}$.)

The Fourier transform of the scattered wave follows from the convolution theorem:

$$\tilde{w}(\mathbf{k}) = \tilde{t}(\mathbf{k} - \mathbf{k}_1) \tilde{G}_{\omega_1}(\mathbf{k}). \quad (11.7)$$

Since $w(\mathbf{x})$ is (almost) monochromatic, we have

$$\tilde{w}(\mathbf{k}) = 0 \quad \text{for } |\mathbf{k}| \neq \frac{\omega_1}{c}. \quad (11.8)$$

In a volume hologram the distribution $t(\mathbf{x})$ results from illuminating the light-sensitive volume with the interference pattern of the reference wave $u_0(\mathbf{x}) = a e^{i\mathbf{k}_0 \cdot \mathbf{x}}$ and the object wave $u(\mathbf{x})$, both monochromatic with frequency $\omega_0 = c |\mathbf{k}_0|$:

$$\begin{aligned} t(\mathbf{x}) &= 1 - \gamma I(\mathbf{x}), \\ I(\mathbf{x}) &= |u_0(\mathbf{x}) + u(\mathbf{x})|^2 \\ &= |a_0|^2 + a_0^* u(\mathbf{x}) e^{-i\mathbf{k}_0 \cdot \mathbf{x}} + a_0 u^*(\mathbf{x}) e^{i\mathbf{k}_0 \cdot \mathbf{x}} + |u(\mathbf{x})|^2. \end{aligned} \quad (11.9)$$

As we did before, we neglect the term $|u(\mathbf{x})|^2$ and find

$$\tilde{t}(\mathbf{k}) = \text{const.} \times \delta(\mathbf{k}) - \gamma a_0^* \tilde{u}(\mathbf{k} + \mathbf{k}_0) - \gamma a_0 \tilde{u}^*(\mathbf{k}_0 - \mathbf{k}).$$

According to eq. (11.7), the Fourier transform of the scattered wave $w(\mathbf{x})$ generated from this function has the form

$$\tilde{w}(\mathbf{k}) = [\text{const.} \times \delta(\mathbf{k} - \mathbf{k}_1) - \gamma a_0^* \tilde{u}(\mathbf{k} + \mathbf{k}_0 - \mathbf{k}_1) - \gamma a_0 \tilde{u}^*(\mathbf{k}_0 + \mathbf{k}_1 - \mathbf{k})] \tilde{G}_{\omega_1}(\mathbf{k}).$$

The first term is a contribution to the unscattered direct part of the reconstruction wave, and as such of no further interest. If $\mathbf{k}_1 \neq \pm \mathbf{k}_0$, the other two contributions vanish for almost all values of $\mathbf{k}$ and only produce reflections in certain directions, which have little to do with the original wave $u(\mathbf{x})$. They are easily identified and separated. Indeed, $\tilde{G}_{\omega_1}(\mathbf{k})$ is sharply concentrated around $|\mathbf{k}| = \omega_1/c$ such that for $k \neq \omega_1/c$ we have in good approximation $\tilde{G}_{\omega_1}(\mathbf{k}) = 0$. Furthermore, we also have that $\tilde{u}(\mathbf{k} + \mathbf{k}_0 - \mathbf{k}_1) = 0$ for $|\mathbf{k} + \mathbf{k}_0 - \mathbf{k}_1| \neq \omega_0/c$ and $\tilde{u}^*(\mathbf{k}_0 + \mathbf{k}_1 - \mathbf{k}) = 0$ for $|\mathbf{k}_0 + \mathbf{k}_1 - \mathbf{k}| \neq \omega_0/c$.

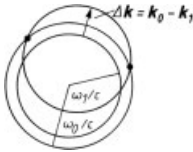


Figure 11.4: The support of $\tilde{w}(\mathbf{k})$ as the intersection of two spheres.

The relationships become clear when we take a look at fig. 11.4. For $\mathbf{k}_0 \neq \pm \mathbf{k}_1$, the scattered light is concentrated on the surfaces of cones that result from the intersection of two non-concentric spheres. On the other hand, if $\mathbf{k}_0 = \mathbf{k}_1$, and thus also $\omega_0 = \omega_1$, we have

$$\tilde{w}(\mathbf{k}) = [\text{const.} \times \delta(\mathbf{k} - \mathbf{k}_1) - \gamma a_0^* \tilde{u}(\mathbf{k}) - \gamma a_0 \tilde{u}^*(2\mathbf{k}_0 - \mathbf{k})] \tilde{G}_{\omega_1}(\mathbf{k}).$$

Because $\tilde{G}_{\omega_1}(\mathbf{k})$ is rotationally invariant, it only contributes a constant factor. Apart from the direct unscattered wave, there are two scattered waves, the first of which differs from the object wave only by a constant, and the second of which is concentrated on the surface of a cone. For $\mathbf{k}_1 = -\mathbf{k}_0$ it is the phase conjugated wave $\tilde{u}^*(-\mathbf{k})$ that dominates.

The fact that reference and reconstruction waves must have the same wave vector $\mathbf{k}_0 = \mathbf{k}_1$ is more of an advantage than a disadvantage. In this way, several volume holograms can be recorded with different values of $\mathbf{k}_0$ on one and the same light-sensitive volume. In the reconstruction, only that hologram becomes visible for which the reference vector $\mathbf{k}_0$ coincides with the one offered by the reconstruction vector.

This leads to attractive possibilities for optical data storage with extremely high storage density and convenient decoding methods. By recording a volume hologram simultaneously with three reference waves whose wave vectors coincide with respect to their direction but not with respect to their value, one obtains a colored three-dimensional picture. For the reconstruction one can illuminate this hologram even with white light, as the correct wave vectors are filtered out automatically. Simple volume holograms can be found almost for free.

12 Coherence theory

12.1 Coherent and incoherent light

In all our previous discussions on optics, we have not been concerned with the origin of the light wave fields whose behavior in different optical systems we wanted to study. The wave fields we considered were always monochromatic; sometimes they consisted of plane waves and sometimes the wave fields emanated from a point-like center of disturbance. In any case they possessed the property of complete temporal and spatial *coherence*: the phase of a field at point $\mathbf{x}_0$ at time t_0 determined the phase for all other times t at the same point; likewise, the phase at t_0 at point $\mathbf{x}_0$ determined the phases at all other points $\mathbf{x}$ for t_0 .

However, not even a single free atom radiates ideal monochromatic light. On the contrary, the excited states responsible for the emission of light have a lifetime τ_c of about 10^{-8} s, which corresponds to a wave train of length $l_c = c\tau_c = 3$ m. Associated with the finite lifetime τ_c is a frequency uncertainty $\Delta\omega \approx 1/\tau_c$ of the emitted radiation. For an atom interacting with its environment, the so-called *coherence time* τ_c and the *coherence length* $l_c = c\tau_c$ may even be much shorter.

In addition, a common light source consists of a very large number of atoms and molecules, which radiate almost independently. Under these circumstances, for a fixed time t_0 and close to the light source, there should be no perceptible correlations of phases for the wave field at different locations; and for a given point $\mathbf{x}_0$ the information about the phase should be lost after the coherence time τ_c .

On the other hand, interference phenomena like diffraction with a visible pattern of maxima and minima are only possible if the phase differences exhibit a certain temporal and spatial stability. These interference phenomena were observed long before the development of lasers, where atoms and molecules produce correlated radiation. These effects are due to the fact that, far away from a radiating source, a (slight) spatial coherence emerges all by itself. In order to understand this phenomenon, let us consider a source with spatial extension d' , whose molecules radiate uncorrelatedly in a small frequency band of width $\Delta\omega$ around $\bar{\omega}$.

An observer separated from the source by a distance $R \gg d'$ performs measurements within a volume of linear dimension $d \ll R$ (see fig. 12.1). Let ξ'_1 and ξ'_2 be two points within the source and let $\mathbf{R} + \xi_1$ and $\mathbf{R} + \xi_2$ be two points close to the observer. For each of these two points, the observer measures the path length difference for the radiation emanated at points

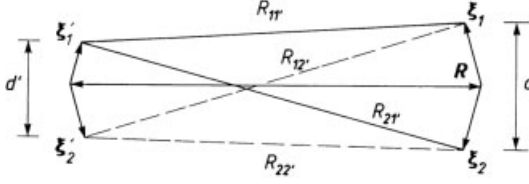


Figure 12.1: Coherence of light from distant sources.

ξ'_1 and ξ'_2 :

$$\begin{aligned}\Delta R_{1'} &= R_{11'} - R_{21'} = \sqrt{(R + \xi_1 - \xi'_1)^2} - \sqrt{(R + \xi_2 - \xi'_1)^2}, \\ \Delta R_{2'} &= R_{12'} - R_{22'} = \sqrt{(R + \xi_1 - \xi'_2)^2} - \sqrt{(R + \xi_2 - \xi'_2)^2}.\end{aligned}\quad (12.1)$$

The observer will find spatial coherence if the path length differences for the radiation from different points of the source are small compared with the wavelength of the radiation:

$$\Delta R_{1'} - \Delta R_{2'} < \bar{\lambda}, \quad \text{where} \quad \bar{\lambda} = \frac{2\pi}{k} = \frac{2\pi c}{\omega}. \quad (12.2)$$

In lowest non-vanishing order a Taylor expansion in ξ_1 , ξ_2 , ξ'_1 , and ξ'_2 yields

$$\Delta R_{1'} - \Delta R_{2'} = -\frac{(\xi_2 - \xi_1)^\perp \cdot (\xi'_2 - \xi'_1)^\perp}{R} = \mathcal{O}\left(\frac{dd'}{R}\right), \quad (12.3)$$

where $\xi^\perp$ denotes the component of ξ perpendicular to $\mathbf{R}$. So the coherence condition reads

$$dd' < \bar{\lambda}R \quad (12.4)$$

or, equivalently,

$$d < \frac{\bar{\lambda}}{\alpha}, \quad d' < \frac{\bar{\lambda}}{\alpha'}. \quad (12.5)$$

Here,

$$\alpha = \frac{d}{R} \quad \text{and} \quad \alpha' = \frac{d'}{R}$$

are the angles at which the source is seen by the observer and at which the observer is seen by the source, respectively.

The typical angle at which we see a star is about $0.005''$ (arc-seconds), which for $\bar{\lambda} = 5 \times 10^{-7}$ m corresponds to an extension of $d \approx 20$ m. Table 12.1 lists the *coherence radius* d for three typical distant light sources. Light sources are often radiating surfaces. In this case the *coherence volume* has the dimensions $l_c = c\tau_c$ in the ray direction and $d = \bar{\lambda}/\alpha'$ perpendicular to the ray direction.

In the coming sections, we will develop a detailed quantitative formulation of the concept of coherence, which, in particular, incorporates the random character of a wave field of a natural light source.

Table 12.1: Coherence lengths for distant sources.

Largest fixed star (Betelgeuse)	$\alpha' = 0.047''$	$d \approx 2 \text{ m}$
Source with	$\alpha' = 1''$	$d \approx 10 \text{ cm}$
Light source of 5 mm as seen from a distance of 100 m	$\alpha' \approx 10''$	$d \approx 1 \text{ cm}$

12.2 Real and analytical signals

This section will provide some tools for the later treatment of non-monochromatic wave fields. Up to now, we have mainly considered monochromatic waves of the form

$$v_\omega(t, \mathbf{x}) = e^{-i\omega t} u_k(\mathbf{x}) \quad \text{with} \quad k = \frac{\omega}{c}, \quad (12.6)$$

where $u_k(\mathbf{x})$ is a solution of Helmholtz's equation,

$$[\Delta + k^2 n^2(\mathbf{x})] u_k(\mathbf{x}) = 0. \quad (12.7)$$

Linear superposition gives us the most general solution of the wave equation,

$$\left(\frac{1}{c^2} n^2(\mathbf{x}) \frac{\partial^2}{\partial t^2} - \Delta \right) v(t, \mathbf{x}) = 0, \quad (12.8)$$

in terms of a temporal Fourier representation

$$v(t, \mathbf{x}) = \frac{1}{\sqrt{2\pi}} \int d\omega e^{-i\omega t} u_k(\mathbf{x}). \quad (12.9)$$

In a similar way, from the Green's function

$$G_k(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')} \quad (12.10)$$

for Helmholtz's equation, we obtain the Green's function

$$\begin{aligned} G(t, \mathbf{x}, \mathbf{x}') &= \frac{1}{2\pi} \int d\omega e^{-i\omega t} G_k(\mathbf{x}, \mathbf{x}') \\ &= \frac{A(\mathbf{x}, \mathbf{x}')}{2\pi} \int d\omega e^{-i\omega[t - (1/c)S(\mathbf{x}, \mathbf{x}')] } \\ &= A(\mathbf{x}, \mathbf{x}') \delta\left(t - \frac{1}{c}S(\mathbf{x}, \mathbf{x}')\right), \end{aligned} \quad (12.11)$$

which satisfies the equation

$$\left(\frac{1}{c^2} n^2(\mathbf{x}) \frac{\partial^2}{\partial t^2} - \Delta \right) G(t, \mathbf{x}, \mathbf{x}') = 4\pi \delta(t) \delta^{(3)}(\mathbf{x} - \mathbf{x}'). \quad (12.12)$$

Here $G(t, \mathbf{x}, \mathbf{x}')$ is just the retarded Green's function of the wave equation: the time at position $\mathbf{x}$ is retarded with respect to a source at $\mathbf{x}'$ by a running time $(1/c)S(\mathbf{x}, \mathbf{x}')$. In particular, we obtain for the vacuum, $n(\mathbf{x}) \equiv 1$:

$$G_0(t, \mathbf{x}, \mathbf{x}') = \frac{1}{|\mathbf{x} - \mathbf{x}'|} \delta\left(t - \frac{1}{c}|\mathbf{x} - \mathbf{x}'|\right). \quad (12.13)$$

Of course, for physical solutions of the wave equation, $v(t, \mathbf{x})$ is real, which we will indicate in the following by an index “r”. We slightly change the notation such that the Fourier representation reads

$$v^r(t, \mathbf{x}) = \frac{1}{\sqrt{2\pi}} \int d\omega e^{-i\omega t} \tilde{v}^r(\omega, \mathbf{x}). \quad (12.14)$$

Since v^r is real, the Fourier transform satisfies

$$\tilde{v}^r(\omega, \mathbf{x}) = \tilde{v}^r(-\omega, \mathbf{x})^*. \quad (12.15)$$

Hence, the total information about v^r is already contained in the values of the Fourier transform $\tilde{v}^r$ for positive frequencies. It is convenient to introduce complex fields whose Fourier components vanish for negative frequencies. We define

$$v(t, \mathbf{x}) = \frac{1}{\sqrt{2\pi}} \int_0^\infty d\omega e^{-i\omega t} \tilde{v}^r(\omega, \mathbf{x}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty d\omega e^{-i\omega t} \tilde{v}(\omega, \mathbf{x}), \quad (12.16)$$

where

$$\tilde{v}(\omega, \mathbf{x}) = \begin{cases} \tilde{v}^r(\omega, \mathbf{x}) & \text{for } \omega \geq 0, \\ 0 & \text{for } \omega < 0. \end{cases} \quad (12.17)$$

Obviously

$$v^r(t, \mathbf{x}) = v(t, \mathbf{x}) + v(t, \mathbf{x})^*, \quad (12.18)$$

and we immediately recover v^r from v . The complex function $v(t, \mathbf{x})$ is called the *analytical signal*, which is justified by the fact that, due to the absence of negative frequency components, $v(t, \mathbf{x})$ as a function of t can be analytically continued to a holomorphic function on the lower t half-plane. According to our discussion in section 2.4, applied to analytical signals, we observe that the real signals

$$\begin{aligned} v^r(t, \mathbf{x}) &= v(t, \mathbf{x}) + v^*(t, \mathbf{x}), \\ v^i(t, \mathbf{x}) &= -i[v(t, \mathbf{x}) - v^*(t, \mathbf{x})] \end{aligned} \quad (12.19)$$

are related by dispersion relations, i.e. they are Hilbert transforms of each other, and v can be written in terms of v^r or v^i .

From now on we shall always assume that complex signals $v(t, \mathbf{x})$ do not contain any negative frequencies. If the integral

$$\int dt |v(t, \mathbf{x})|^2$$

exists, unitarity of the Fourier transformation implies that

$$\int dt |v(t, \mathbf{x})|^2 = \int d\omega |\tilde{v}(\omega, \mathbf{x})|^2 \quad (12.20)$$

or, more generally,

$$\int dt v(t+\tau, \mathbf{x}_1) v^*(t, \mathbf{x}_2) = \int d\omega e^{-i\omega\tau} \tilde{v}(\omega, \mathbf{x}_1) \tilde{v}^*(\omega, \mathbf{x}_2). \quad (12.21)$$

However, we are mainly interested in signals like $v(t, \mathbf{x}) = e^{-i\omega t} u_k(\mathbf{x})$, for which the integrals do not exist. Therefore, we are only allowed to perform calculations with temporally averaged intensities of the form

$$I(\mathbf{x}) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T dt |v(t, \mathbf{x})|^2. \quad (12.22)$$

For a more rigorous treatment we will use signals with a “cut-off”,

$$v_T^r(t, \mathbf{x}) = \begin{cases} v^r(t, \mathbf{x}) & \text{for } |t| \leq T, \\ 0 & \text{for } |t| > T, \end{cases} \quad (12.23)$$

which are confined to the interval $[-T, T]$. Indeed, a stationary signal is an idealization, because no signal acts for an infinite time. For each finite value of T , we have

$$\frac{1}{2T} \int_{-\infty}^{\infty} dt v_T^r(t+\tau, \mathbf{x}_1) v_T^r(t, \mathbf{x}_2) = \frac{1}{2T} \int_{-\infty}^{\infty} d\omega e^{-i\omega\tau} \tilde{v}_T^r(\omega, \mathbf{x}_1) \tilde{v}_T^{r*}(\omega, \mathbf{x}_2), \quad (12.24)$$

and we can expect that the limit $T \rightarrow \infty$ exists. We introduce double angle brackets to denote the temporal average of a quantity:

$$\langle\langle A(t) \rangle\rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T dt A(t). \quad (12.25)$$

With this notation the equation above can be written in the form:

$$\langle\langle v^r(t+\tau, \mathbf{x}_1) v^r(t, \mathbf{x}_2) \rangle\rangle = \int d\omega e^{-i\omega\tau} \lim_{t \rightarrow \infty} \frac{1}{2T} \tilde{v}_T^r(\omega, \mathbf{x}_1) \tilde{v}_T^{r*}(\omega, \mathbf{x}_2). \quad (12.26)$$

Using the simplified notation $v_1(t) = v(t, \mathbf{x}_1)$, which we shall adopt frequently in this chapter, we obtain for the analytic signal $v(t, \mathbf{x})$:

$$\begin{aligned} & \langle\langle v_1(t+\tau) v_2^*(t) \rangle\rangle \\ &= \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T dt \int \frac{d\omega}{\sqrt{2\pi}} \int \frac{d\omega'}{\sqrt{2\pi}} e^{-i\omega(t+\tau)} e^{+i\omega't} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}^*(\omega') \\ &= \lim_{T \rightarrow \infty} \int \frac{d\omega}{\sqrt{2\pi}} \int \frac{d\omega'}{\sqrt{2\pi}} \int_{-T}^T dt e^{-i(\omega-\omega')t} e^{-i\omega\tau} \frac{1}{2T} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}^*(\omega'). \end{aligned}$$

From the identity

$$\lim_{T \rightarrow \infty} \int_{-T}^T dt e^{-i(\omega - \omega')t} = 2\pi \delta(\omega - \omega'),$$

we find

$$\langle\langle v_1(t + \tau) v_2^*(t) \rangle\rangle = \int d\omega e^{-i\omega\tau} \tilde{\Gamma}_{12}(\omega), \quad (12.27)$$

where

$$\tilde{\Gamma}_{12}(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}^*(\omega). \quad (12.28)$$

Here $\tilde{\Gamma}_{12}(\omega)$ satisfies the conditions $\tilde{\Gamma}_{12}(\omega) = \tilde{\Gamma}_{21}^*(\omega)$ and $\tilde{\Gamma}_{12}(\omega) = 0$ for $\omega < 0$. This rather formal calculation can be justified rigorously. For a light wave field that does not contain any additional static field, we have $\langle\langle v_1(t) \rangle\rangle = 0$. We shall always assume that the temporal average of v vanishes. From

$$\frac{1}{2T} \int_{-T}^T dt v_1(t) = \frac{1}{2T} \tilde{v}_{1T}(0), \quad (12.29)$$

we can deduce that $\tilde{v}(\omega)$ can only be weakly singular for $\omega \rightarrow 0$, which then implies

$$\begin{aligned} \langle\langle v_1(t + \tau) v_2(t) \rangle\rangle &= \lim_{T \rightarrow \infty} \int dt \int \frac{d\omega}{\sqrt{2\pi}} \int \frac{d\omega'}{\sqrt{2\pi}} e^{-i(\omega + \omega')t} e^{-i\omega\tau} \frac{1}{2T} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}(\omega') \\ &= \int d\omega \int d\omega' \delta(\omega + \omega') e^{-i\omega\tau} \lim_{T \rightarrow \infty} \frac{1}{2T} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}(\omega') \\ &= \int d\omega e^{-i\omega\tau} \lim_{T \rightarrow \infty} \frac{1}{2T} \tilde{v}_{1T}(\omega) \tilde{v}_{2T}(-\omega) \\ &= 0, \end{aligned}$$

because $\tilde{v}_{1T}$ and $\tilde{v}_{2T}$ do not contain any negative frequencies. From $v^r = v + v^*$ we get

$$\langle\langle v_1^r(t + \tau) v_2^r(t) \rangle\rangle = \langle\langle v_1(t + \tau) v_2^*(t) \rangle\rangle + \langle\langle v_1^*(t + \tau) v_2(t) \rangle\rangle. \quad (12.30)$$

In particular, for $\mathbf{x}_1 = \mathbf{x}_2$ and $\tau = 0$ the mean intensity at $\mathbf{x}_1$ is given by

$$I(\mathbf{x}_1) = \langle\langle [v_1^r(t)]^2 \rangle\rangle = 2 \langle\langle v_1(t) v_1^*(t) \rangle\rangle = 2 \int d\omega \tilde{\Gamma}_{11}(\omega). \quad (12.31)$$

So $\tilde{\Gamma}_{11}(\omega)$ has the meaning of a spectral density for the intensity. For $\tau \neq 0$ we obtain

$$\begin{aligned} \frac{1}{2} \langle\langle v_1^r(t + \tau) v_1^r(t) \rangle\rangle &= \text{Re} \langle\langle v_1(t + \tau) v_1^*(t) \rangle\rangle \\ &= \text{Re} \int d\omega e^{-i\omega\tau} \tilde{\Gamma}_{11}(\omega). \end{aligned} \quad (12.32)$$

12.3 The light wave field as a stochastic process

A natural light source consists of many atoms and molecules whose microstates cannot be controlled by observation. Therefore, light signals $v(t, \mathbf{x})$ at given space-time points will be more or less random. In other words, for each space-time point $(t, \mathbf{x})$ there exists a random variable $V(t, \mathbf{x})$ whose distribution depends on the characteristic properties of the light source. Hence, in our description of the light field we are dealing with a collection of uncountably many random variables or, more precisely, a *random function* $V(t, \mathbf{x})$. Expressed in a mathematical language, the wave field should be described by a *stochastic process*. Each actually existing wave field $v(t, \mathbf{x})$ is a realization of the stochastic process, and the probability distribution characterizing the stochastic process determines the relative probabilities of different realizations.

As for all random variables A , we can consider the expectation value $\langle A \rangle$. Of special interest are the expectation values

$$\begin{aligned} & \Gamma(t_1, \mathbf{x}_1, \dots, t_r, \mathbf{x}_r; t_{r+1}, \mathbf{x}_{r+1}, \dots, t_n, \mathbf{x}_n) \\ &= \langle V(t_1, \mathbf{x}_1) \cdots V(t_r, \mathbf{x}_r) V^*(t_{r+1}, \mathbf{x}_{r+1}) V^*(t_n, \mathbf{x}_n) \rangle, \end{aligned} \quad (12.33)$$

which are called *generalized correlation functions*. We will often use the simplified notation

$$\Gamma_{1, \dots, n}(t_1, \dots, t_n) = \langle V_1(t_1) \cdots V_r(t_r) V_{r+1}^*(t_{r+1}) V_n^*(t_n) \rangle. \quad (12.34)$$

The following is a list of a few particularly important special cases:

- (i) $\Gamma_{12}(t+\tau, t) = \langle V_1(t+\tau) V_2^*(t) \rangle$ the *covariance function*,
- (ii) $\Gamma_{11}(t, t) = \langle V_1(t) V_1^*(t) \rangle \geq 0$ the *intensity function*,
- (iii) $\gamma_{12}(t+\tau, t) = \frac{\Gamma_{12}(t+\tau, t)}{[\Gamma_{11}(t+\tau)]^{1/2} [\Gamma_{22}(t)]^{1/2}}$ the *correlation function*.

Independently of further specifications, every random variable C satisfies the inequality

$$\langle C^* C \rangle \geq 0,$$

and writing $C = \sqrt{\langle B^* B \rangle} A + e^{i\varphi} \sqrt{\langle A^* A \rangle} B$ we find

$$\langle C^* C \rangle = 2\langle A^* A \rangle \langle B^* B \rangle + \sqrt{\langle A^* A \rangle} \sqrt{\langle B^* B \rangle} \langle e^{i\varphi} A^* B + e^{-i\varphi} B^* A \rangle \geq 0.$$

From this we obtain the fundamental inequality

$$|\langle AB^* \rangle| \leq \sqrt{\langle A^* A \rangle} \sqrt{\langle B^* B \rangle}. \quad (12.35)$$

The equality sign holds for the case $C_0 = A/\sqrt{\langle A^* A \rangle} + e^{i\varphi_0} B/\sqrt{\langle B^* B \rangle} = 0$. Setting $A = V_1(t+\tau)$ and $B = V_2(t)$ this inequality implies

$$|\Gamma_{12}(t+\tau, t)|^2 \leq \Gamma_{11}(t+\tau, t+\tau) \Gamma_{22}(t, t)$$

or

$$0 \leq |\gamma_{12}(t+\tau, t)| \leq 1. \quad (12.36)$$

We now summarize some additional properties of stochastic processes that describe the light wave fields in which we are interested.

- (A) Because each realization $v(t, \mathbf{x})$ satisfies a linear wave equation and because expectation values are linear, the generalized correlation functions satisfy the same wave equation with respect to all of their arguments, e.g.

$$\left(\frac{n^2(\mathbf{x})}{c^2} \frac{\partial^2}{\partial t^2} - \Delta \right) \langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle = 0. \quad (12.37)$$

- (B) Each realization contains only non-negative frequencies; in particular, we have that $\langle \tilde{V}(\omega, \mathbf{x}) \rangle = 0$ for $\omega < 0$. As we have seen in the last section, we may require this property without loss of generality.
- (C) $\langle V(t, \mathbf{x}) \rangle = 0$. This only implies that we deal exclusively with oscillating fields without any static contributions. From (B) and (C) we also find

$$\langle V(t_1, \mathbf{x}_1) V(t_2, \mathbf{x}_2) \rangle = \langle V^*(t_1, \mathbf{x}_1) V^*(t_2, \mathbf{x}_2) \rangle = 0. \quad (12.38)$$

- (D) The stochastic process is *stationary*:

$$\begin{aligned} & \langle V_1(t + \tau_1) \cdots V_r(t + \tau_r) V_{r+1}^*(t + \tau_{r+1}) \cdots V_n^*(t + \tau_n) \rangle \\ &= \langle V_1(\tau_1) \cdots V_r(\tau_r) V_{r+1}^*(\tau_{r+1}) \cdots V_n^*(\tau_n) \rangle. \end{aligned} \quad (12.39)$$

So, the generalized correlation functions depend only on time differences. In particular, we will use the following notation:

$$\begin{aligned} \Gamma_{12}(t + \tau, t) &= \Gamma_{12}(\tau, 0) =: \Gamma_{12}(\tau), \\ \Gamma_{11}(t, t) &= \Gamma_{11}(0), \end{aligned} \quad (12.40)$$

and

$$\begin{aligned} |\Gamma_{12}(\tau)| &\leq \sqrt{\Gamma_{11}(0)} \sqrt{\Gamma_{22}(0)}, \\ 0 &\leq \gamma_{12}(\tau) \leq 1. \end{aligned} \quad (12.41)$$

For a real signal we find

$$\begin{aligned} \Gamma_{12}^r(\tau) &= \langle V_1^r(t + \tau) V_2^r(t) \rangle = 2 \operatorname{Re} \Gamma_{12}(\tau), \\ \Gamma_{11}^r(0) &= \langle V_1^r(t)^2 \rangle = 2 \Gamma_{11}(0), \end{aligned} \quad (12.42)$$

and

$$0 \leq \gamma_{12}^r = \operatorname{Re} \gamma_{12} = \frac{\Gamma_{12}^r(\tau)}{\sqrt{\Gamma_{11}^r} \sqrt{\Gamma_{22}^r}} \leq 1. \quad (12.43)$$

- (E) The stochastic process is *ergodic*, which means that for (almost) all realizations we may replace the expectation value $\langle \cdot \rangle$ by the temporal average $\langle\langle \cdot \rangle\rangle$:

$$\langle\langle v_1(t + \tau_1) \cdots v_n^*(t + \tau_n) \rangle\rangle = \langle V_1(\tau_1) \cdots V_n^*(\tau_n) \rangle. \quad (12.44)$$

In other words, we may think of an ergodic process as a motion for which the state of a system approaches, for sufficiently large time-scales, any given state arbitrarily closely and arbitrarily often. Ergodic systems contain incommensurable frequencies.

Because of $\langle\langle A(t) \rangle\rangle = \langle\langle A(t + t_0) \rangle\rangle$, an ergodic process is always stationary, which makes (D) a consequence of (E).

One more important consequence of ergodicity will be emphasized in particular. A similar calculation as in the previous section, which takes into account the stationarity of the process, yields

$$\begin{aligned} \Gamma_{12}(\tau) &= \langle V_1(t + \tau) V_2^*(t) \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T dt \langle V_1(t + \tau) V_2^*(t) \rangle \\ &= \int d\omega e^{-i\omega\tau} \tilde{\Gamma}_{12}(\omega), \end{aligned} \quad (12.45)$$

where

$$\tilde{\Gamma}_{12}(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} \langle \tilde{V}_{1T}(\omega) \tilde{V}_{2T}^*(\omega) \rangle.$$

In particular,

$$\tilde{\Gamma}_{11}(\tau) = \langle V(t + \tau, \mathbf{x}) V^*(t, \mathbf{x}) \rangle = \int d\omega e^{-i\omega\tau} \tilde{\Gamma}_{11}(\omega),$$

where

$$\tilde{\Gamma}_{11}(\omega) = \lim_{T \rightarrow \infty} \frac{1}{2T} \langle \tilde{V}_{1T}(\omega) \tilde{V}_{1T}^*(\omega) \rangle \geq 0. \quad (12.46)$$

For a stochastic process we may replace the expectation values by time averages. In this case we have already seen that $\tilde{\Gamma}_{11}(\omega)$ is the spectral density of the time average of the intensity. This is the content of the *Wiener–Khinchin theorem*:

Theorem 12.1 (Wiener–Khinchin theorem). *The covariance function*

$$\Gamma_{11}(\tau) = \langle\langle v_1(t + \tau) v_1^*(t) \rangle\rangle = \langle V_1(t + \tau) V_1^*(t) \rangle \quad (12.47)$$

and the spectral intensity distribution $\tilde{\Gamma}_{11}(\omega)$ are Fourier transforms of each other.

The Wiener–Khinchin theorem may be used to determine the spectral distribution $\tilde{\Gamma}_{11}(\omega)$ by measuring $\Gamma_{11}(\tau)$. So, with the help of time measurements, we can measure the widths of spectral lines even when the resolution of an ordinary spectrometer turns out to be insufficient.

Because $\tilde{\Gamma}_{12}(\omega) = 0$ for $\omega < 0$, the real and imaginary parts of $\Gamma_{12}(\tau)$ are related by dispersion relations.

In a quantum theoretical treatment of the radiation field, the quantities $V(t, \mathbf{x})$ and $V^*(t, \mathbf{x})$ correspond to the annihilation and creation operators of light quanta. The expectation value

$$\langle V(t_1, \mathbf{x}_1) \dots V^*(t_n, \mathbf{x}_n) \rangle = \text{Tr}\{\rho V(t_1, \mathbf{x}_1) \dots V^*(t_n, \mathbf{x}_n)\}$$

becomes the quantum-mechanical expectation value of the quantum field in a (mixed) state characterized by a density matrix ρ . Also in a quantum-mechanical treatment, the properties (A) to (D) remain valid. Quantum expectation values will be dealt with in chapter 14.

12.4 Gaussian stochastic processes

In this section, we will discuss a stochastic property that can be attributed to many light wave fields. In many cases, we are allowed to treat the field $V(t, \mathbf{x})$ as a Gaussian stochastic process, at least within a good approximation. Before we can explain the consequences of this property, we have to define Gaussian random variables.

Let Z_i ($i = 1, \dots, n$) be a set of stochastic variables. Without loss of generality, we may assume these variables to have vanishing mean values: $\langle Z_i \rangle = 0$.

The variables Z_i are simultaneously *Gaussian distributed* if their probability density is of the form

$$w(z) = \frac{1}{N} e^{-\frac{1}{2} z \cdot A z}. \quad (12.48)$$

Here, $z = (z_1, \dots, z_n)$, $z \cdot A z = \sum_{i,j=1}^n z_i A_{ij} z_j$, (A_{ij}) is a positive definite symmetric matrix, and N is a normalization factor that guarantees that $\int d^n z w(z) = 1$. We can diagonalize (A_{ij}) by a suitable coordinate transformation, in which case the normalization integral becomes a product of one-dimensional Gaussian integrals:

$$\int dx e^{-\frac{1}{2} a x^2} = \sqrt{\frac{2\pi}{a}}. \quad (12.49)$$

In this way we can determine the normalization factor to be $N = (2\pi)^{n/2} (\det A)^{-1/2}$, so that we finally obtain

$$w(z) = \left(\det \frac{A}{2\pi} \right)^{1/2} e^{-\frac{1}{2} z \cdot A z}. \quad (12.50)$$

By a simple translation $z_i \mapsto z_i + c_i$, we can define Gaussian distributed random variables with non-vanishing expectation value. According to the central limit theorem, we can expect the variables Z_i to be Gaussian distributed if the randomness of their distribution is the consequence of many independent additive influences. For instance, elementary error calculus is based on the assumption of a Gaussian distribution.

The generalized covariances,

$$\langle Z_{i_1} \dots Z_{i_r} \rangle = \int d^n z z_{i_1} \dots z_{i_r} w(z),$$

are most easily obtained from the generating function

$$F(j) = \langle e^{j \cdot Z} \rangle = \frac{1}{N} \int d^n z e^{j \cdot z - \frac{1}{2} z \cdot A z}.$$

Indeed, taking derivatives with respect to j yields

$$\langle Z_{i_1} \cdots Z_{i_r} \rangle = \frac{\partial}{\partial j_1} \cdots \frac{\partial}{\partial j_r} F(j) \Big|_{j=0}.$$

We calculate $F(j)$ by making a quadratic completion and obtain

$$F(j) = e^{\frac{1}{2} j \cdot A^{-1} j},$$

where A^{-1} is the inverse matrix of A . This immediately leads to

$$\langle Z_i Z_k \rangle = A_{ik}^{-1},$$

i.e. the covariances of second order are simply given by the components of the inverse matrix $(A_{ik})^{-1}$. For a Gaussian distribution the higher covariances are already determined by the covariances of second order, e.g.

$$\langle Z_i Z_k Z_r Z_s \rangle = A_{ik}^{-1} A_{rs}^{-1} + A_{ir}^{-1} A_{ks}^{-1} + A_{is}^{-1} A_{kr}^{-1}.$$

In general, all odd covariances vanish and all even covariances can be represented as sums of products of second covariances.

For us, this property – factorization of the higher covariances into second-order covariances – will be the most important feature of Gaussian stochastic processes. As a side remark we mention that, if Z_i ($i = 1, \dots, n$) are Gaussian variables, then also the stochastic variables obtained by a linear transformation,

$$W_k = \sum_i C_{ki} Z_i,$$

are simultaneously Gaussian distributed, and

$$\langle W_i W_k \rangle = \sum_{r,k} C_{ir} C_{ks} \langle Z_r Z_s \rangle.$$

The stochastic process $\{V^r(t, \mathbf{x})\}$ is called *Gaussian* if for any value of n and any n space-time points $(t_1, \mathbf{x}_1), \dots, (t_n, \mathbf{x}_n)$ the random variables

$$V^r(t_1, \mathbf{x}_1), \dots, V^r(t_n, \mathbf{x}_n)$$

are simultaneously Gaussian distributed. In a similar way we define an analytic Gaussian process $V(t, \mathbf{x})$.

Due to the factorization property of generalized covariances and the additional relation $\langle V_1 V_2 \rangle = \langle V_1^* V_2^* \rangle = 0$, the covariances for which the numbers of V and V^* differ all vanish:

$$\langle V_1 \cdots V_r V_{r+1}^* \cdots V_n^* \rangle = 0.$$

If $V(t, \mathbf{x})$ is a Gaussian stochastic process, the linearly transformed process

$$W(t, \mathbf{x}) = \int dt' \int d^3x' K(t-t', \mathbf{x}, \mathbf{x}') V(t', \mathbf{x}')$$

is also Gaussian and we find

$$\begin{aligned} & \langle W(t_1, \mathbf{x}_1) W^*(t_2, \mathbf{x}_2) \rangle \\ &= \int dt'_1 \int dt'_2 \int d\mathbf{x}'_1 \int d\mathbf{x}'_2 \\ & \quad \times K(t_1-t'_1, \mathbf{x}_1, \mathbf{x}'_1) K^*(t_2-t'_2, \mathbf{x}_2, \mathbf{x}'_2) \cdot \langle V(t'_1, \mathbf{x}'_1) V^*(t'_2, \mathbf{x}'_2) \rangle. \end{aligned}$$

From these relations it should be obvious that properties (A) to (D) from section 12.3 and, if K is not too singular, also property (E) apply for W if they apply for V .

These invariances under linear transformations are of great importance for the light wave fields $V(t, \mathbf{x})$ that we consider. Indeed, because $V(t, \mathbf{x})$ satisfies a linear wave equation and contains no negative frequencies, $V(t, \mathbf{x})$ depends, for arbitrary values of t , linearly on the initial values $V(t_0, \mathbf{x})$. Therefore, $V(t, \mathbf{x})$ is Gaussian if the process $V(t_0, \mathbf{x})$ is Gaussian for any time t_0 . Since $V(t, \mathbf{x})$ is already determined linearly by the values $V(t, \mathbf{x}(y))$ on a two-dimensional surface F , it suffices for V to be Gaussian on F . If the source of the light wave field is a glowing surface F , e.g. the filament of a bulb or the surface of a star, the different parts of this surface radiate in a completely uncorrelated way, and for arbitrary times and points $\mathbf{x}_1, \dots, \mathbf{x}_n$ on F the variables $V(t_1, \mathbf{x}_1) \dots V(t_n, \mathbf{x}_n)$ are simultaneously Gaussian distributed. On F we even have $\langle V(t_1, \mathbf{x}_1) V^*(t_1, \mathbf{x}_2) \rangle = \Gamma(0, \mathbf{x}_1, \mathbf{x}_2) = a(\mathbf{x}_1) \delta(\mathbf{x}_1 - \mathbf{x}_2)$. A Gaussian process on F , for which the covariance function is proportional to a δ -function, is called *white noise*.

If the radiation from the source on F corresponds to white noise, the process $V(t, \mathbf{x})$ will not necessarily remain a white noise process for arbitrary values of t . However, it will always remain a Gaussian process. In particular, the total information about the process is already contained in the covariance function $\Gamma_{12}(\tau)$. A Gaussian stochastic process can always be transformed into a white noise process by a suitable linear transformation of the variables.

Laser light is an example of a non-Gaussian stochastic light field. In this case, the higher generalized covariance functions contain new information that is independent of the second covariance functions. We shall come to these questions in chapter 14.

12.5 The quasi-monochromatic approximation

Some of the considerations concerning the coherence of a light wave field simplify considerably when, due to the nature of the light source or a suitable filter, only frequencies within a small region around a mean frequency $\bar{\omega}$ are relevant. In this case, the coherence time τ_c is large and we can derive an approximate expression for $\Gamma_{12}(\tau)$ that is valid as long as $\tau \ll \tau_c$. We start with an exact definition of the coherence time τ_c and the spectral width σ_ω .

If $V(t, \mathbf{x})$ is stationary, we cannot define τ_c as an average decay time of the signal, but only as the decay time of the correlation $\gamma_{11}(\tau)$ at a given space point.

Due to the relation $\gamma_{11}(\tau) = \gamma_{11}^*(-\tau)$, then $|\gamma_{11}^*(\tau)|^2$ is an even function. We normalize this function such that the integral is unity:

$$N = \int d\tau |\gamma_{11}(\tau)|^2 = \int d\omega |\tilde{\gamma}_{11}(\omega)|^2.$$

Furthermore,

$$\bar{\tau} = \frac{1}{N} \int d\tau \tau |\gamma_{11}(\tau)|^2 = 0.$$

We define

$$\begin{aligned} \tau_c^2 &= \frac{1}{N} \int d\tau \tau^2 |\gamma_{11}(\tau)|^2, \\ \bar{\omega} &= \frac{1}{N} \int d\omega \omega |\tilde{\gamma}_{11}(\omega)|^2, \\ \sigma_\omega^2 &= \frac{1}{N} \int d\omega (\omega - \bar{\omega})^2 |\tilde{\gamma}_{11}(\omega)|^2. \end{aligned} \quad (12.51)$$

Indeed, for almost monochromatic light, not only $\langle |\tilde{V}_1(\omega)| \rangle$ but also

$$\tilde{\gamma}_{11}(\omega) \sim \lim_{T \rightarrow \infty} \frac{1}{2T} \langle |\tilde{V}_{T1}(\omega)|^2 \rangle$$

are sharply peaked around $\bar{\omega}$. One can show that

$$\tau_c \sigma_\omega \gtrsim \frac{1}{2}. \quad (12.52)$$

In quantum mechanics this relation is known as the uncertainty relation. Equality holds when $|\gamma_{11}(\tau)|^2$ is a Gauss function:

$$|\gamma_{11}(\tau)|^2 = \text{const.} \times e^{-\tau^2/2\tau_c^2}. \quad (12.53)$$

The covariance function $\Gamma_{12}(\tau)$ is given by

$$\Gamma_{12}(\tau) = \int d\omega \tilde{\Gamma}_{12}(\omega) e^{-i\omega\tau}. \quad (12.54)$$

Since we assume that $\tilde{\Gamma}_{12}(\omega)$ is sharply concentrated around $\bar{\omega}$, it is helpful to consider the quantities

$$e^{i\bar{\omega}\tau} \Gamma_{12}(\tau) = \int d\omega \tilde{\Gamma}_{12}(\omega) e^{-i(\omega - \bar{\omega})\tau}. \quad (12.55)$$

The value of $|\tilde{\Gamma}_{12}(\omega)|$ is markedly different from zero only within a region of width σ_ω . For $\sigma_\omega \tau \ll 1$, i.e. $\tau \ll \tau_c$, the τ dependence on the right-hand side is negligible and can be replaced by its value at $\tau = 0$. Thus, for $\tau \ll \tau_c$, we get

$$\begin{aligned} e^{i\bar{\omega}\tau} \Gamma_{12}(\tau) &= \Gamma_{12}(0), \\ \Gamma_{12}(\tau) &\approx e^{-i\bar{\omega}\tau} \Gamma_{12}(0). \end{aligned}$$

This expression is called the *quasi-monochromatic approximation* for $\Gamma_{12}(\tau)$. The τ dependence of $\exp(i\bar{\omega}\tau)\Gamma_{12}(\tau)$ is much slower than the time dependence of the wave field. In interference experiments, the coherence lengths are often of the order of centimeters while the differences in the path length of interfering rays are of the order of a wavelength. In these cases the quasi-monochromatic approximation is legitimate.

Introducing the notation

$$e^{i\bar{\omega}\tau}\Gamma_{12}(\tau) = |\Gamma_{12}(\tau)| e^{i\alpha_{12}(\tau)} = \mu_{12}(\tau) e^{i\alpha_{12}(\tau)},$$

where $\alpha_{12}(\tau)$ and $\mu_{12}(\tau)$ are slowly varying functions, the monochromatic approximation may also be written in the form

$$\Gamma_{12}(\tau) \approx \mu_{12}(0) e^{i[\alpha_{12}(0) - \bar{\omega}\tau]}. \quad (12.56)$$

Quasi-monochromatic approximations exist also for the higher generalized covariance functions.

12.6 Coherence and correlation functions

We have seen that the value of the correlation function

$$\gamma_{12}(\tau) = \gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) = \frac{\langle V(t+\tau, \mathbf{x}_1) V^*(t, \mathbf{x}_2) \rangle}{\langle |V(t, \mathbf{x}_1)|^2 \rangle^{1/2} \langle |V(t, \mathbf{x}_2)|^2 \rangle^{1/2}} \quad (12.57)$$

is bounded by

$$0 \leq |\gamma_{12}(\tau)| \leq 1, \quad (12.58)$$

and $\gamma_{11}(0) = 1$. The interpretation of $\gamma_{12}(\tau)$ becomes obvious when we investigate the limiting cases $\gamma_{12}(\tau) = 0$ and $|\gamma_{12}(\tau)| = 1$.

If, for space-time points $(t+\tau, \mathbf{x}_1)$ and $(t, \mathbf{x}_2)$, we find $\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) = 0$, there are no correlations whatsoever between the values of the wave field at these two points. If, even more, we have $\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) = 0$ for all points $\mathbf{x}_1 \neq \mathbf{x}_2$, the wave field is totally incoherent and Γ assumes the form of white noise:

$$\Gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) = a\delta(\mathbf{x}_1 - \mathbf{x}_2). \quad (12.59)$$

However, as we have already mentioned and as we shall see again soon, there is a certain amount of coherence even for a surface source producing white noise.

Next we assume that for two space-time points $(t+\tau, \mathbf{x}_1)$ and $(t, \mathbf{x}_2)$, the correlation function is unity:

$$|\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)| = 1.$$

In this case, the equality sign holds in the correlation inequality (12.35). This implies, as we have shown in section 12.3, that for the random variables $V(t+\tau, \mathbf{x}_1)$ and $V(t, \mathbf{x}_2)$ a relation of the following type holds:

$$\frac{V(t+\tau, \mathbf{x}_1)}{\sigma_1} = e^{i\alpha(\tau, \mathbf{x}_1, \mathbf{x}_2)} \frac{V(t, \mathbf{x}_2)}{\sigma_2}.$$

Here $\sigma_{1,2}^2 = \langle |V(t, \mathbf{x}_{1,2})|^2 \rangle$, and $\alpha_{12}(\tau) = \alpha(\tau, \mathbf{x}_1, \mathbf{x}_2)$ is a fixed phase (i.e. not random). Indeed, in this case

$$\gamma_{12}(\tau) = \frac{\langle V_1(t+\tau) V_2^*(t) \rangle}{\sigma_1 \sigma_2} = e^{i\alpha(\tau, \mathbf{x}_1, \mathbf{x}_2)} \frac{\langle |V_2(t)|^2 \rangle}{\sigma_2^2} = e^{i\alpha_{12}(\tau)},$$

and $\alpha_{12}(\tau)$ is just the phase of $\gamma_{12}(\tau)$ (which is why we adopted the notation of the previous section). For the two space-time points, the two phases have a fixed difference and, therefore, are strictly correlated. In the extremal case, $|\gamma_{12}(\tau)| \equiv 1$ for *all* values of τ , $\mathbf{x}_1$, and $\mathbf{x}_2$:

$$\frac{V_1(t_1)}{\sigma_1} \equiv \frac{V_2(t_2)}{\sigma_2} e^{i\alpha_{12}}.$$

Evaluating this equation for other pairs of space-time points, we see that $\alpha_{13} = \alpha_{12} + \alpha_{23}$, and when we choose a reference point $(t_0, \mathbf{x}_0)$, we find $\alpha_{12} = \alpha_{10} - \alpha_{20}$, i.e.

$$\frac{V_1(t_1)}{\sigma_1} e^{-i\alpha_{10}} \equiv \frac{V_2(t_2)}{\sigma_2} e^{-i\alpha_{20}}.$$

On both sides of this equation we find the same random variable A with $\langle |A|^2 \rangle = 1$ and which no longer depends on the space-time point. Therefore,

$$V(t, \mathbf{x}) = \sqrt{\langle |V(t, \mathbf{x})|^2 \rangle} e^{i\alpha(t, \mathbf{x})} A, \quad (12.60)$$

and the only stochastic element is given by the space-time-independent stochastic variable A . In particular, the phase differences between the field values at any two space-time points are fixed. So, $|\gamma_{12}(\tau)| \equiv 1$ corresponds to the case of complete coherence of the wave field. This suggests defining $|\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)|$ as the *degree of coherence*, where $|\gamma(0, \mathbf{x}_1, \mathbf{x}_2)|$ characterizes the spatial coherence and $|\gamma(\tau, \mathbf{x}_1, \mathbf{x}_1)|$ the temporal coherence at a point $\mathbf{x}_1$. This interpretation is confirmed by the following experiment, which also provides a method for measuring $\gamma_{12}(\tau)$.

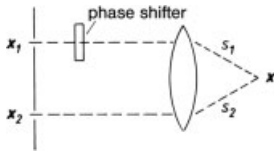


Figure 12.2: Principle of Young's interference experiment.

We are talking about *Young's interference experiment* (fig. 12.2). A screen with two tiny apertures at points $\mathbf{x}_1$ and $\mathbf{x}_2$ is placed into the wave field. We assume that the screen does not change the field in any essential way at the apertures. The interference of the two excitations emanating from $\mathbf{x}_1$ and $\mathbf{x}_2$ is observed at point $\mathbf{x}$ behind the screen, and the optical path lengths $S(\mathbf{x}, \mathbf{x}_1)$ and $S(\mathbf{x}, \mathbf{x}_2)$ will be denoted by s_1 and s_2 , respectively. The excitation at point $\mathbf{x}$ is now given by the stochastic variable

$$W(t, \mathbf{x}) = a_1 V(t - s_1/c, \mathbf{x}_1) + a_2 V(t - s_2/c, \mathbf{x}_2). \quad (12.61)$$

The coefficients a_1 and a_2 only depend on the size of the apertures and are, as explained in section 10.3, purely imaginary due to the Huygens–Fresnel principle. The average intensity at point $\mathbf{x}$ is proportional to

$$\begin{aligned} \langle |W(t, \mathbf{x})|^2 \rangle &= |a_1|^2 \langle |V(t, \mathbf{x}_1)|^2 \rangle + |a_2|^2 \langle |V(t, \mathbf{x}_2)|^2 \rangle \\ &\quad + |a_1| |a_2| \{ \langle V(t-s_1/c, \mathbf{x}_1) V^*(t-s_2/c, \mathbf{x}_2) \rangle \\ &\quad + \langle V(t-s_2/c, \mathbf{x}_2) V^*(t-s_1/c, \mathbf{x}_1) \rangle \}. \end{aligned}$$

The first two terms on the right-hand side are proportional to the intensities $I^{(1)}(\mathbf{x})$ and $I^{(2)}(\mathbf{x})$ which we would observe if only one of the apertures were present. The third term describes the interference. Expressed in terms of the intensities, we obtain

$$I(\mathbf{x}) = I^{(1)}(\mathbf{x}) + I^{(2)}(\mathbf{x}) + 2\sqrt{I^{(1)}(\mathbf{x})}\sqrt{I^{(2)}(\mathbf{x})}\operatorname{Re}\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2), \quad (12.62)$$

with $\tau = (s_2 - s_1)/c$. We see that $\operatorname{Re}\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)$ is directly observable: $\mathbf{x}_1$ and $\mathbf{x}_2$ can be varied by changing the position of the apertures, and τ can be changed either by shifting the observation point $\mathbf{x}$ or by placing a phase shifter behind one of the apertures.

Now $\operatorname{Im}\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)$ can be determined from dispersion relations or simply by shifting the phase of the light fields at one of the holes by $\pi/2$. In the quasi-monochromatic limit the τ dependence is particularly simple. Using the notations of section 12.5, we have $\gamma_{12}(\tau) = \mu_{12}(\tau) e^{-i\bar{\omega}\tau + i\alpha_{12}(\tau)}$ and thus

$$\begin{aligned} \operatorname{Re}\gamma_{12}(\tau) &\approx \mu_{12}(0) \cos[\alpha_{12}(0) - \bar{\omega}\tau], \\ \operatorname{Im}\gamma_{12}(\tau) &\approx \mu_{12}(0) \sin[\alpha_{12}(0) - \bar{\omega}\tau]. \end{aligned} \quad (12.63)$$

For $\mathbf{x}_1 = \mathbf{x}_2$, the quantity $\gamma_{11}(\tau) = \gamma(\tau, \mathbf{x}_1, \mathbf{x}_1)$ can also be measured directly in an interference experiment, as indicated in fig. 12.3. In this case, the excitation at point $\mathbf{x}_1$ is split into two rays (e.g. by a semitransparent mirror), which are later, after both rays have traversed paths of different lengths, reunited for interference at point $\mathbf{x}$. According to the Wiener–Khinchin theorem, we can determine from such a measurement of $\gamma_{11}(\tau)$ the quantity $\tilde{\gamma}_{11}(\omega)$, which allows a precise measurement of a spectral intensity distribution with an extremely sharp peak. The smaller this peak, the smaller are the variations of $\gamma_{11}(\tau)$ as a function of τ , which makes it easier to measure.

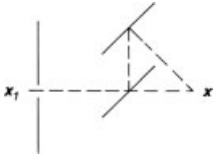


Figure 12.3: Measuring $\gamma_{11}(\tau)$.

The result (12.62) for the intensity can be rewritten in the form

$$\begin{aligned} I(\mathbf{x}) &= [1 - |\gamma_{12}(\tau)|] (I^{(1)} + I^{(2)}) \\ &\quad + |\gamma_{12}(\tau)| \{ I^{(1)} + I^{(2)} + 2\sqrt{I^{(1)}}\sqrt{I^{(2)}} \cos[\alpha_{12}(\tau) - \bar{\omega}\tau] \}. \end{aligned} \quad (12.64)$$

This remains valid even when the aperture is not tiny, as long as $\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)$ does not essentially vary along the diameter of the aperture. One can easily show that in this case $I^{(1)}$

and $I^{(2)}$ are the diffraction intensities for each of the two apertures. In general, the spatial dependence of the correlation $\gamma_{12}(\tau)$ is small compared to the spatial dependence of the signal $V(t, \mathbf{x})$.

Equation (12.64) confirms our interpretation of $|\gamma_{12}(\tau)|$ as the degree of coherence. There are two contributions for the intensity $I(\mathbf{x})$:

- An incoherent part, which is simply the sum of the two input intensities without further interference. This contribution is weighted by $1 - |\gamma_{12}|$.
- A coherent part, for which the excitations are added, and the intensity equals the square of the total signal. For instance, for $\alpha_{12} - \omega\tau = 0$ we find

$$I_{\text{coh}} = \left(\sqrt{I^{(1)}} + \sqrt{I^{(2)}} \right)^2.$$

The presence of the second part reveals itself in an interference pattern varying with $\mathbf{x}$. The *visibility* of this pattern is defined by

$$\mathcal{V} = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}}, \quad (12.65)$$

where I_{max} and I_{min} are the extremal values of I corresponding to $\cos(\alpha_{12} - \omega\tau) = \pm 1$. For this quantity we obtain

$$\mathcal{V} = \frac{2\sqrt{I^{(1)}}\sqrt{I^{(2)}}}{I^{(1)} + I^{(2)}} |\gamma_{12}|, \quad (12.66)$$

and, in particular, for $I^{(1)} = I^{(2)}$:

$$\mathcal{V} = |\gamma_{12}|. \quad (12.67)$$

12.7 The propagation of the correlation function

We now assume that on a surface F the covariance function

$$\Gamma_{12}(\tau) = \langle V_1(t + \tau) V_2^*(\tau) \rangle$$

is known and we pose the problem to determine $\Gamma_{12}(\tau)$ for all points in space. This problem is solvable because the covariance function satisfies the wave equation and we can apply the Huygens–Fresnel principle. Taking the Fourier transform of the signal with respect to time, $\tilde{V}(\omega, \mathbf{x})$, we can write Kirchhoff's identity (10.41) in the form

$$\tilde{V}(\omega, \mathbf{x}) = \frac{1}{4\pi} \int_F d\mathbf{f}' \{ \tilde{V}(\omega, \mathbf{x}') \nabla' G_\omega(\mathbf{x}, \mathbf{x}') - G_\omega(\mathbf{x}, \mathbf{x}') \nabla' \tilde{V}(\omega, \mathbf{x}') \}. \quad (12.68)$$

For $|\mathbf{x} - \mathbf{x}'| \gg \lambda$, from eq. (10.44) we obtain

$$\tilde{V}(\omega, \mathbf{x}) = \frac{-ik}{4\pi} \int_F df' (\cos \chi + \cos \chi_0) \tilde{V}(\omega, \mathbf{x}') G_\omega(\mathbf{x}, \mathbf{x}'), \quad (12.69)$$

where χ_0 and χ are the inclination angles of the normal wave vectors of $\tilde{V}(\omega, \mathbf{x})$ and G_ω with respect to the normal directions of F . In our case, χ_0 is a random variable, and in order to avoid a random inclination factor we replace the Green's function $G_\omega(\mathbf{x}, \mathbf{x}')$ by the Dirichlet Green's function $G_{D\omega}(\mathbf{x}, \mathbf{x}')$, which satisfies

$$G_{D\omega}(\mathbf{x}, \mathbf{x}') = 0 \quad \text{for} \quad \mathbf{x}' \in F. \quad (12.70)$$

The second term in eq. (12.69) now vanishes and we find

$$\tilde{V}(\omega, \mathbf{x}) = \frac{-ik}{4\pi} \int df' \tilde{V}(\omega, \mathbf{x}') \nabla' G_{D\omega}(\mathbf{x}, \mathbf{x}'). \quad (12.71)$$

If F is a plane, we obtain $G_{D\omega}$ from G_ω by applying the reflection principle:

$$G_{D\omega}(\mathbf{x}, \mathbf{x}') = G_\omega(\mathbf{x}, \mathbf{x}') - G_\omega(\mathbf{x}, \mathbf{x}''), \quad (12.72)$$

where $\mathbf{x}''$ follows from $\mathbf{x}'$ by reflection on F . Given

$$G_\omega(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')},$$

we get

$$G_{D\omega}(\mathbf{x}, \mathbf{x}') = A(\mathbf{x}, \mathbf{x}') e^{ikS(\mathbf{x}, \mathbf{x}')} - A(\mathbf{x}, \mathbf{x}'') e^{ikS(\mathbf{x}, \mathbf{x}'')}. \quad (12.73)$$

On F we have $\mathbf{x}' = \mathbf{x}''$, and for $|\mathbf{x} - \mathbf{x}'| \gg \lambda$ we obtain

$$(\mathbf{n} \cdot \nabla') G_D(\mathbf{x}, \mathbf{x}') = 2(\mathbf{n} \cdot \nabla') G_\omega(\mathbf{x}, \mathbf{x}'),$$

i.e.

$$\tilde{V}(\omega, \mathbf{x}) = -\frac{2ik}{4\pi} \int_F df' \tilde{V}(\omega, \mathbf{x}') \cos \chi G_\omega(\mathbf{x}, \mathbf{x}'). \quad (12.74)$$

This expression differs from the one we obtained before merely by the replacement $(\cos \chi + \cos \chi_0) \mapsto 2 \cos \chi$. In leading order of k , both expressions are identical. So we assume that

$$\tilde{V}(\omega, \mathbf{x}) = \int_F df' \Lambda \tilde{V}(\omega, \mathbf{x}') G_\omega(\mathbf{x}, \mathbf{x}'), \quad (12.75)$$

with an inclination factor $\Lambda = -(i/\lambda) \cos \chi$. Inserting the covariance function yields

$$\tilde{\Gamma}(\omega, \mathbf{x}_1, \mathbf{x}_2) = \int_F df'_1 df'_2 \Lambda_1 \Lambda_2^* \tilde{\Gamma}(\omega, \mathbf{x}'_1, \mathbf{x}'_2) G_\omega(\mathbf{x}_1, \mathbf{x}'_2) G_\omega^*(\mathbf{x}_2, \mathbf{x}'_2), \quad (12.76)$$

and after a Fourier transformation

$$\begin{aligned} \Gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) &= \frac{1}{2\pi} \int d\omega \tilde{\Gamma}(\omega, \mathbf{x}_1, \mathbf{x}_2) e^{-i\omega\tau} \\ &= \int_F df'_1 df'_2 \Lambda_1 \Lambda_2^* A(\mathbf{x}_1, \mathbf{x}'_1) A^*(\mathbf{x}_2, \mathbf{x}'_2) \\ &\quad \times \Gamma(\tau - S(\mathbf{x}_1, \mathbf{x}'_1)/c + S(\mathbf{x}_2, \mathbf{x}'_2)/c, \mathbf{x}'_1, \mathbf{x}'_2). \end{aligned} \quad (12.77)$$

This expression simplifies further when we make the following assumptions:

- The quasi-monochromatic approximation holds:

$$\tau - \frac{1}{c}S_{11'} + \frac{1}{c}S_{22'} \ll \tau_c. \quad (12.78)$$

- On F the function $\Gamma(\tau, \mathbf{x}'_1, \mathbf{x}'_2)$ corresponds to white noise:

$$\Gamma(\tau, \mathbf{x}'_1, \mathbf{x}'_2) = I(\tau, \mathbf{x}'_1) \delta^{(2)}(\mathbf{x}'_1 - \mathbf{x}'_2). \quad (12.79)$$

For a distant filament lamp or a fixed star, both observed through a filter, these assumptions are satisfied. We then find

$$\begin{aligned} \Gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) &= \int_F d\mathbf{f}'_1 \Lambda_1 \Lambda_2^* A(\mathbf{x}_1, \mathbf{x}'_1) A^*(\mathbf{x}_2, \mathbf{x}'_1) \\ &\quad \times I(\omega, \mathbf{x}'_1) \cdot e^{-i\bar{\omega}[\tau - (1/c)S_{11'} + (1/c)S_{22'}]}. \end{aligned} \quad (12.80)$$

For a very distant object we have $\Lambda_1 = \Lambda_2$, and we may put slowly varying quantities in front of the integral:

$$\begin{aligned} \Gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) &= A(\mathbf{x}_1, \mathbf{x}'_0) A^*(\mathbf{x}_2, \mathbf{x}'_0) \\ &\quad \times \int_F d\mathbf{f}'_1 |\Lambda|^2 I(\omega, \mathbf{x}'_1) \cdot e^{-i\bar{\omega}[\tau - (1/c)S_{11'} + (1/c)S_{22'}]}. \end{aligned} \quad (12.81)$$

The average intensity at point $\mathbf{x}_1$ is proportional to

$$\Gamma(0, \mathbf{x}_1, \mathbf{x}_1) = |A(\mathbf{x}_1, \mathbf{x}'_0)|^2 \int_F d\mathbf{f}' |\Lambda|^2 I(\omega, \mathbf{x}'). \quad (12.82)$$

So, we obtain as the final result for the normalized correlation function:

$$\gamma(0, \mathbf{x}_1, \mathbf{x}_2) = \frac{\int_F d\mathbf{f}' |\Lambda|^2 I(\bar{\omega}, \mathbf{x}') e^{(i\bar{\omega}/c)[S(\mathbf{x}_1, \mathbf{x}') - S(\mathbf{x}_2, \mathbf{x}')]}}{\int_F d\mathbf{f}' |\Lambda|^2 I(\omega, \mathbf{x}')}. \quad (12.83)$$

This expression resembles a diffraction integral in which the pupil function has been replaced by the intensity distribution $|\Lambda|^2 I(\bar{\omega}, \mathbf{x}')$. This fact is known as the *van Cittert–Zernicke theorem*:

Theorem 12.2 (van Cittert–Zernicke theorem). *In the quasi-monochromatic approximation, the correlation function $\gamma(0, \mathbf{x}_1, \mathbf{x}_2)$ for the radiation of a white planar source, considered as a function of $\mathbf{x}_1$, resembles the diffraction field of an aperture centered around $\mathbf{x}_2$ and with pupil function $|\Lambda|^2 I(\omega, \mathbf{x})$.*

For a very distant source we may in addition apply Fraunhofer's approximation, i.e. we may expand the exponent up to linear terms in $\mathbf{x}'_0$:

$$\frac{i\omega}{c}S(\mathbf{x}_1, \mathbf{x}') - \frac{i\omega}{c}S(\mathbf{x}_2, \mathbf{x}') = i\psi_{12} + ik\mathbf{p}_{12} \cdot \boldsymbol{\xi}, \quad (12.84)$$

and we get

$$\gamma(0, \mathbf{x}_1, \mathbf{x}_2) = e^{i\psi_{12}} \frac{\int_F d^2\xi |\Lambda|^2 I(\bar{\omega}, \xi) e^{ik\mathbf{p}_{12} \cdot \xi}}{\int d^2\xi |\Lambda|^2 I(\bar{\omega}, \xi)}. \quad (12.85)$$

If, for instance, the source consists of a homogeneous radiating circular disc of radius ρ , the function $\gamma(0, \mathbf{x}_1, \mathbf{x}_2)$ assumes the form of an Airy function:

$$\begin{aligned} \gamma(0, \mathbf{x}_1, \mathbf{x}_2) &= e^{i\psi_{12}} 2J_1(v)/v, \\ \text{with } v &= k\rho p = k\rho d/R. \end{aligned} \quad (12.86)$$

Here, $p = d/R$ is the angle at which the two points $\mathbf{x}_1$ and $\mathbf{x}_2$ (with $|\mathbf{x}_1 - \mathbf{x}_2| = d$) are seen from a source at a distance $R = |\mathbf{x}_{1,2} - \mathbf{x}'_0|$.

12.8 Amplitude and intensity interferometry

12.8.1 Amplitude interferometry: Michelson interferometer

As we have seen in section 10.7, the resolution of a telescope is limited by the diameter of the frontal aperture d . The smallest angle α that can just be resolved is of the order

$$\alpha = \lambda/d.$$

The *Michelson interferometer* is an arrangement by which one can achieve a virtual effective aperture diameter of up to 10 m. The principal features are shown in fig. 12.4. The light from the observed astronomical object falls on two mirrors, which are separated by a distance d . Two additional mirrors direct the rays into a telescope and map them into the focal plane. The distance d can be changed.

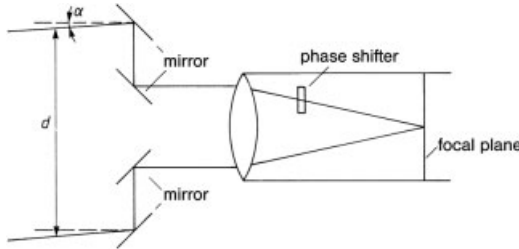


Figure 12.4: Principle of a Michelson interferometer.

We first consider the case of a single point-like object, such as a distant star with a small radius. If the mirrors are circular, we will see in the focal plane the diffraction image of two circular apertures of distance d , which is an Airy pattern, traversed by straight interference fringes (see section 10.6). The distance between the straight interference fringes is of the order $f\lambda/d$.

Suppose now that we observe two point-like objects, say, a binary system, having a very small angular distance α . Due to the different incidence angles there will be a relative phase shift at the two mirrors, which in turn results in a relative displacement of the diffraction

patterns transverse to the straight fringes by a distance $f\alpha$. If we now vary d , the distance between the straight fringes changes, and for $\lambda/2d = \alpha$ the maximum of the image of one component lies on top of the minimum of the image of the other component, which results in the minimal visibility of the straight fringes part of the interference pattern. In this way, even very narrow binary systems can be resolved.

The Michelson arrangement is useful not only for the resolution of binary systems but also for measuring the diameters of stars. In this case we should think of the object as a small disk-like luminosity distribution with a small angular diameter α . At first sight one might expect that for such an object the straight interference fringes would be completely smeared out; however, this is not the case, and the best way to understand this phenomenon is via the van Cittert–Zernicke theorem.

Both mirrors receive the same intensity of light from the star having angular diameter α . According to section 12.6, this leads to an intensity distribution in the focal plane that is given by

$$I(\mathbf{x}) = 2I^{(1)}(\mathbf{x})[1 + \text{Re} \gamma(\tau, \mathbf{x}_1, \mathbf{x}_2)], \quad (12.87)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the positions of the two mirrors and $|\mathbf{x}_1 - \mathbf{x}_2| = d$ is their distance apart. $I^{(1)}$ resembles an Airy distribution. But according to the van Cittert–Zernicke theorem we have

$$\gamma(\tau, \mathbf{x}_1, \mathbf{x}_2) = e^{i\psi_{12} - i\bar{\omega}\tau} \frac{\int d^2\xi |\Lambda|^2 I(\bar{\omega}, \xi) e^{ik\mathbf{p}_{12} \cdot \xi}}{\int d^2\xi |\Lambda|^2 I(\bar{\omega}, \xi)}. \quad (12.88)$$

Thus, the correlation function $\gamma_{12}(\tau)$ directly measures the Fourier transform of the intensity distribution. On the other hand, according to eq. (12.87), it is directly observable when we vary $\mathbf{x}_1$ and $\mathbf{x}_2$ and record the changes in the interference pattern. In this way we can determine not only the apparent diameter of the source but in principle even the total intensity distribution.

For a disk-like source like a fixed star, γ_{12} is described by an Airy distribution:

$$\gamma_{12}(0) \sim \frac{2J_1(\pi d\alpha/\lambda)}{\pi d\alpha/\lambda}, \quad (12.89)$$

where d is the separation of the mirrors and α is the apparent diameter of the star. The interference pattern has minimal visibility if the following condition is fulfilled:

$$\pi d\alpha/\lambda = 1.22\pi, \quad \text{i.e.} \quad \alpha = 1.22\lambda/d. \quad (12.90)$$

By varying d , one can determine the diameters of large fixed stars. For instance, already in the 1920s this method yielded the value $\alpha = 0.047''$ for Betelgeuse.

The problem in Michelson's method lies in the extremely high demands for the stability of the arrangement. The distance between the two mirrors must be kept stable within fractions of the wavelength of light, which is difficult to achieve, in particular, when it should be possible to direct the instrument towards different stars.

12.8.2 Photon correlation spectroscopy

This stability problem is avoided by the so-called *intensity correlation spectroscopy*, first proposed in 1957 by R. Hanbury Brown and R. Q. Twiss (fig. 12.5). What is measured in this

method are the correlations in the intensity fluctuations of the object signal at two distant points x_1 and x_2 :

$$\Delta I_{1,2} \sim (|V_{1,2}|^2 - \langle |V_{1,2}|^2 \rangle). \quad (12.91)$$

Cables conduct the two intensity signals to a correlator, which measures the correlations. The demands for mechanical stability are low. Furthermore, the intensity varies much slower than the signal itself, making the measurement of $\langle \Delta I_1 \Delta I_2 \rangle$ really manageable.

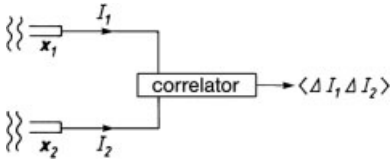


Figure 12.5: Principle of intensity correlation spectroscopy.

Using the tools derived up to now, we can easily understand the principles of the method of Hanbury Brown and Twiss. Because the wave field of an extended object can be considered (almost) as a Gaussian stochastic process, we can determine the correlation function $\langle \Delta I_1 \Delta I_2 \rangle$ using the factorization properties mentioned in sections 12.3 and 12.4:

$$\begin{aligned} \langle \Delta I_1 \Delta I_2 \rangle &\sim \langle [V_1 V_1^*(t + \tau) - \langle V_1 V_1^* \rangle] [V_2 V_2^*(t) - \langle V_2 V_2^* \rangle] \rangle \\ &= \langle V_1(t + \tau) V_1^*(t + \tau) V_2(t) V_2^*(t) \rangle - \langle V_1 V_1^* \rangle \langle V_2 V_2^* \rangle \\ &= \langle V_1(t + \tau) V_2^*(t) \rangle \langle V_1^*(t + \tau) V_2(t) \rangle \\ &\quad + \langle V_1 V_1^* \rangle \langle V_2 V_2^* \rangle - \langle V_1 V_1^* \rangle \langle V_2 V_2^* \rangle \\ &= |\Gamma_{12}(\tau)|^2. \end{aligned} \quad (12.92)$$

Hence, the correlation of the intensity fluctuations is a direct measure for the absolute value of the covariance function $\Gamma_{12}(\tau)$. The phase can easily be calculated within the quasi-monochromatic approximation, whereas, in general, it is determined by dispersion relations. This leads us to the surprising result that, via the van Cittert–Zernicke theorem, we can determine the shape of the source of a wave field by correlation measurements of intensity fluctuations. At least in principle it should be possible to determine the shape of an egg frying in a pan by measuring the intensity correlations of the sound.

Due to the extremely low intensities, the practical realization of this method has to rely on correlation measurements of photon counting rates in photomultipliers. This will be described in chapter 14.

Intensity correlation spectroscopy is mainly used in radioastronomy using radio waves. In this case, several radio telescopes, often thousands of kilometers apart, are connected for the measurement of the intensity correlations.

12.9 Dynamical light scattering

Stochastic wave fields do not always have their origin in some stochastic source – they can also be produced by reflection, refraction, diffraction, and scattering of coherent waves at stochastic

structures. An example of such a situation is the speckle effect, which we will describe in the next section.

Dynamical light scattering refers to the scattering of light at stochastically distributed scattering centers. Here, we want to describe in a simple way how we can obtain information about the shape and the dynamics of the distribution of these scattering centers by measuring correlation functions.

Let us first consider a single scattering center located at a point $\mathbf{x}_i(t)$ at time t . To simplify matters, we restrict the discussion to scalar wave theory. An incoming plane wave $e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})}$ will produce a scattered wave emanating from $\mathbf{x}_i(t)$. Like we did in section 11.3, we make the following ansatz for the scattered wave under the condition that $|\mathbf{v}_i(t)| = |\dot{\mathbf{x}}_i(t)| \ll c$:

$$v(t, \mathbf{x}) = A e^{-i\omega t} e^{i\mathbf{k} \cdot \mathbf{x}_i(t)} \frac{e^{ik|\mathbf{x} - \mathbf{x}_i(t)|}}{|\mathbf{x} - \mathbf{x}_i(t)|}. \quad (12.93)$$

For large distances from the scattering center we find approximately

$$v(t, \mathbf{x}) = A \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|} e^{-i\omega t} e^{-i\mathbf{q} \cdot \mathbf{x}_i(t)}, \quad (12.94)$$

where

$$\mathbf{q} = \mathbf{k}' - \mathbf{k} = k \frac{\mathbf{x}}{|\mathbf{x}|} - \mathbf{k}. \quad (12.95)$$

[Strictly speaking, the factor k in eq. (12.95) has to be replaced by the quantity $k' = k + \Delta k$, taking into account the Doppler shift. However, for $|\mathbf{v}_i|/c \ll 1$ and $\mathbf{q} \neq 0$, the difference is negligible.] For several stochastically distributed scattering centers, the locations $\mathbf{X}_i(t)$ are stochastic variables and the stochastic scattered wave field is given by

$$V(t, \mathbf{x}) = A \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|} e^{-i\omega t} \sum_i e^{-i\mathbf{q} \cdot \mathbf{X}_i(t)}. \quad (12.96)$$

From this we find for the covariance function:

$$\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle = |A|^2 \sum_{i,j} e^{-i\omega(t-t')} \langle e^{-i\mathbf{q} \cdot [\mathbf{X}_i(t) - \mathbf{X}_j(t')]} \rangle. \quad (12.97)$$

For the simplest and most common cases, the quantity $|A|^2$ is simply a constant. In general, $|A|^2$ is proportional to the differential cross-section for the scattering at a single scattering center.

Let $w_{ij}(t, t'; \mathbf{x}, \mathbf{x}')$ be the simultaneous position distribution on $\mathbb{R}^3 \times \mathbb{R}^3$, which corresponds to the joint probability that we find center i at time t close to $\mathbf{x}$ and center j at time t' close to $\mathbf{x}'$. Then

$$\langle e^{-i\mathbf{q} \cdot [\mathbf{X}_i(t) - \mathbf{X}_j(t')]} \rangle = \int d^3x \int d^3x' w_{ij}(t, t'; \mathbf{x}, \mathbf{x}') e^{-i\mathbf{q} \cdot (\mathbf{x} - \mathbf{x}')}. \quad (12.98)$$

Setting

$$w_{ij}(t, t'; \mathbf{x}, \mathbf{x}') = C_{ij}(t - t', \mathbf{x} - \mathbf{x}'), \quad (12.99)$$

we find that the expectation value (12.98) is proportional to the spatial Fourier transform of the so-called *pair distribution function* $C_{ij}(t-t', \mathbf{x}-\mathbf{x}')$:

$$\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle \sim |A|^2 \sum_{i,j} e^{-i\omega(t-t')} \tilde{C}_{ij}(t-t', \mathbf{q}). \quad (12.100)$$

(The $\mathbf{x}$ dependence on the right-hand side of eq. (12.100) is hidden in the momentum transfer $\mathbf{q}$.)

If, instead of discrete point-like scattering centers, we are considering a random scattering density distribution $\rho(t, \mathbf{x})$, eqs. (12.96) and (12.100) are replaced by

$$V(t, \mathbf{x}) = A \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|} e^{-i\omega t} \int d^3x' \rho(t, \mathbf{x}') e^{-i\mathbf{q} \cdot \mathbf{x}} \quad (12.101)$$

and

$$\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle \sim |A|^2 e^{-i\omega(t-t')} \tilde{C}(t-t', \mathbf{q}), \quad (12.102)$$

where now $\tilde{C}(t-t', \mathbf{q})$ is the spatial Fourier transform of the covariance function:

$$C(t-t', \mathbf{x}-\mathbf{x}') = \langle \rho(t, \mathbf{x}) \rho^*(t', \mathbf{x}') \rangle. \quad (12.103)$$

The *structure function* $\tilde{C}(\tau, \mathbf{q})$ contains information about the space-time correlations of the scattering centers. The quantity $\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle$, which is proportional to $\tilde{C}(\tau, \mathbf{q})$, can be measured directly with the methods described in section 12.6. However, in many cases, a more accessible quantity to measure is the Fourier transform:

$$\begin{aligned} I(\omega', \mathbf{q}) &= \int d\tau e^{i\omega'\tau} \langle V(t, \mathbf{x}) V^*(t+\tau, \mathbf{x}) \rangle \\ &\sim |A|^2 \int d\tau e^{i(\omega'-\omega)\tau} \tilde{C}(\tau, \mathbf{q}). \end{aligned} \quad (12.104)$$

According to the Wiener–Khinchin theorem (12.47), $I(\omega', \mathbf{q})$ is proportional to the spectral density of the scattered radiation that is observed at a momentum transfer $\mathbf{q}$, i.e. the differential cross-section of the inelastic scattering with frequency change $\Delta\omega = \omega' - \omega$.

Instead of measuring $\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle$, it is often more convenient to measure the correlations of intensity fluctuations, which are proportional to $\langle \Delta|V(t, \mathbf{x})|^2 \Delta|V(t', \mathbf{x}')|^2 \rangle$ and which, for a Gaussian stochastic process, contain the same information as $\langle V(t, \mathbf{x}) V^*(t', \mathbf{x}') \rangle$ (see section 12.8).

This general discussion about the scattering of waves at stochastically distributed centers can be applied not only to the scattering of light, but also, for example, to the scattering of neutrons in matter. Here we mention just a few of the numerous applications:

1. *Scattering of light by a dilute solution of polymer molecules.* The scattering centers are the segments of long chain molecules. Because the different segments within the chain have different positions along the chain, the scattering centers are not equivalent. The structure function $\tilde{C}_{ij}(\tau, \mathbf{q})$ depends on the dynamical behavior of the polymer chain, like, for example, the degree of stretching or entanglement. We can treat the scattering of light by macromolecules in dilute solutions or by colloidal solutions of small particles in a similar way.

2. *Dynamical pair correlation functions of material media.* In this case the scattering centers are the molecules of a gaseous, liquid, or solid medium. In general, the molecules may be considered as equivalent such that the indices i and j in eq. (12.100) can be omitted. The pair correlation function $C(0, \mathbf{x}-\mathbf{x}')$ determines the thermodynamic properties of the medium, while $C(\tau, \mathbf{x}-\mathbf{x}')$ contains information about the relaxation behavior following a disturbance or about transport coefficients such as the thermal conductivity, the diffusion constant, or the viscosity.
3. *Light scattering at density fluctuations in a transparent fluid.* This is an example of scattering by a continuous distribution. Basically, the scattering occurs at fluctuations of the refractive index, but these are mainly due to density fluctuations. $\tilde{C}(0, \mathbf{0})$ is related to the isothermal compressibility; the τ dependence of $\tilde{C}(\tau, \mathbf{q})$ is damped exponentially and characterizes the relaxation after a disturbance of the density distribution. Close to the critical point the density fluctuations are long-ranged and long-lived.

There are two types of density fluctuations in stationary media:

- thermal fluctuations due to temperature variations and thermal expansion – they are damped by thermal conduction; and
- mechanical fluctuations, i.e. sound waves, which are damped by inner friction and thermal conduction.

Associated with these two density fluctuations one observes two different kinds of maxima in the temporal Fourier transform $I(\omega', \mathbf{q})$ of $\tilde{C}(\tau, \mathbf{q})$:

- a so-called *Rayleigh peak* at $\omega' = \omega$, whose width is determined by the thermal conductivity; and
 - two so-called *Brillouin peaks* at the Doppler-shifted positions $\omega' = \omega \pm c_s |\mathbf{q}|$, where c_s is the velocity of sound – the width of the Brillouin peaks depends on the viscosity of the medium.
4. *X-ray spectroscopy of crystals.* With respect to the metric in a crystal lattice, the equal-time pair distribution $C(0, \mathbf{x}-\mathbf{x}')$ is spatially periodic, i.e. $\tilde{C}(0, \mathbf{q})$ is concentrated in the form of a δ -function on a reciprocal lattice with discrete values of $\mathbf{q}$ (see section 10.7). From the structure function $\tilde{C}(\tau, \mathbf{q})$ one can determine the form and the dynamics of the crystal lattice. In many cases, however, one is less interested in the structure function itself than in the factor $|A|^2$ in eq. (12.102), which characterizes the shape of a single scattering center. The complicated molecules to be investigated are crystallized such they are aligned and periodically distributed. The structure function $\tilde{C}(0, \mathbf{q})$ serves the purpose to concentrate the intensity of the scattered light along the directions of the discrete maxima of $\tilde{C}(0, \mathbf{q})$.

12.10 Granulation

When laser light is reflected from a diffuse surface F' , the produced radiation field exhibits a speckled structure that can be seen by the naked eye. This phenomenon, called the *speckle effect*, is a nice application of the results we obtained in this chapter.

Let the laser light have a fixed frequency ω . Due to the random irregularities on F' , a stochastic wave field $V(\omega, \mathbf{x}')$ ($\mathbf{x}' \in F'$) is generated right in front of F' . On some other surface F there will be, according to eq. (12.68) or (12.75), a linearly transformed field $V(\omega, \mathbf{x})$ with $\mathbf{x} \in F$. In particular, if F is the focal plane of the imaging system, the linear transformation is a Fourier transformation. In general, the wave field on F' will be close to Gaussian or even white noise. Therefore, there will be a Gaussian wave field also on F , which, however, for most cases is not white noise.

The wave field on F is called the *objective speckle pattern* and has to be distinguished from the *subjective speckle pattern*, which refers to the speckle pattern on the retina of our eye. Diffraction effects at the pupil of the eye contribute to the subjective speckle pattern. Only if F is the focal plane of some optical system may the objective and subjective speckle patterns be identified, but even in this case diffraction effects in the eye can make things more complicated: in some cases one perceives finer granules within each coarse grain.

We now consider the intensity distribution of the objective speckle pattern on F . For a fixed point $\mathbf{x} \in F$, the distribution of the complex amplitude $V(\omega, \mathbf{x})$ will be a two-dimensional Gaussian distribution of the form

$$w(v) = \frac{1}{N} e^{-v^* v / 2\sigma^2}. \quad (12.105)$$

Writing $v = v_1 + iv_2$ we get

$$\begin{aligned} w(v_1, v_2) dv_1 dv_2 &= \frac{1}{N} e^{-(v_1^2 + v_2^2)/2\sigma^2} |v| d|v| d\varphi \\ &= \frac{1}{2N} e^{-|v|^2/2\sigma^2} d|v|^2 d\varphi, \end{aligned} \quad (12.106)$$

where $|v|^2 = v_1^2 + v_2^2$.

So the distribution of the intensity $I = |v|^2/2$ at point $\mathbf{x} \in F$ is simply given by an exponential function:

$$w(I) = \frac{1}{\langle I \rangle} e^{-I/\langle I \rangle}, \quad (12.107)$$

where the average intensity is $\langle I \rangle = \sigma^2$, and for the variance of I one obtains

$$\sigma_I^2 = \langle I \rangle^2. \quad (12.108)$$

Notice that $I = 0$ has largest probability. The grain size of the speckle pattern is given by the correlation length of the intensity distribution.

According to eq. (12.92) and the van Cittert–Zernicke theorem (12.88), it is the correlation function $\gamma(0, \mathbf{x}_1, \mathbf{x}_2)$ on F that is relevant. If the illuminated surface on F' is circular with

diameter d , there will be an Airy distribution and the grain size is given by

$$D = 1.22 \frac{\lambda}{d} L, \quad (12.109)$$

where L is the distance between F' and F .

Apart from being a disturbing side-effect, the speckle effect also has many useful applications in measurement technology. It is used, for instance, to make visible tiny displacements or to observe the flow fields in fluids that have been dotted with small scattering particles. We will now explain the principle of such experiments for the particularly nice example of speckle interferometry of stars.

Irregularities in the density and humidity of the atmosphere lead to the scintillation of starlight. Due to these effects, the observed light of a star seems to have spatial and temporal fluctuations. The spatial extension of these inhomogeneities is of the order of 0.1 cm to 10 cm, and the lifetime is about 0.1 s. Starlight is coherent enough to produce a speckle pattern, which can be photographed if the exposure time is sufficiently short. The film then exhibits a speckle distribution $I_1(\mathbf{x})$.

Two close-by stars produce an intensity distribution of the form

$$I(\mathbf{x}) = I_1(\mathbf{x}) + \rho I_1(\mathbf{x} + \mathbf{d}), \quad (12.110)$$

which consists of two similar distributions that are slightly shifted with respect to each other. The factor ρ accounts for the different luminosities of the stars.

For the Fourier transform, which can also be realized optically by the methods described in chapter 10, we find

$$\tilde{I}(\mathbf{k}) = \tilde{I}_1(\mathbf{k})(1 + \rho e^{i\mathbf{k} \cdot \mathbf{d}}), \quad (12.111)$$

i.e.

$$|\tilde{I}(\mathbf{k})|^2 = |\tilde{I}_1(\mathbf{k})|^2 [1 + \rho^2 + 2\rho \cos(\mathbf{k} \cdot \mathbf{d})]. \quad (12.112)$$

From the width of the pattern of stripes traversing the distribution $|\tilde{I}(\mathbf{k})|^2$, one can determine the size of d and thus the angular distance of the two stars.

12.11 Image processing by filtering

Often a signal suffers interference by random influences. This leads to the problem of undoing the disturbances as far as possible and recovering the undisturbed signal. The so-called *Wiener filter* is one tool for this purpose.

As a concrete example, we think of an image that can be described by a two-dimensional intensity distribution $I(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^2$. We assume $I(\mathbf{x})$ to be a stochastic process (see section 12.3).

A stochastic disturbance $\Sigma(\mathbf{x})$ will be superimposed onto the distribution $I(\mathbf{x})$ and what we observe is the sum of the two:

$$Z(\mathbf{x}) = I(\mathbf{x}) + \Sigma(\mathbf{x}). \quad (12.113)$$

We can measure the covariance function,

$$C_{ZZ}(\mathbf{x} - \mathbf{y}) = \langle \Delta Z(\mathbf{x}) \Delta Z(\mathbf{y}) \rangle \quad (12.114)$$

with $\Delta Z(\mathbf{x}) = Z(\mathbf{x}) - \langle Z(\mathbf{x}) \rangle$, as well as $\langle Z(\mathbf{x}) \rangle$. Furthermore, we assume that we know the characteristic features of the disturbance Σ ,

$$C_{\Sigma\Sigma}(\mathbf{x} - \mathbf{y}) = \langle \Delta \Sigma(\mathbf{x}) \Delta \Sigma(\mathbf{y}) \rangle \quad (12.115)$$

and $\langle \Sigma(\mathbf{x}) \rangle$. For instance, Σ may be white noise.

In eqs. (12.114) and (12.115) we have assumed I and Σ to be translation-invariant. Furthermore,

$$C_{I\Sigma}(\mathbf{x} - \mathbf{y}) = \langle \Delta I(\mathbf{x}) \Delta \Sigma(\mathbf{y}) \rangle = 0, \quad (12.116)$$

i.e. the signal I and the disturbance Σ are uncorrelated.

As an estimator $\hat{I}$ for an optimal reconstruction of the signal I , we take a stochastic process that results from the observed signal Z by a linear transformation:

$$\hat{I}(\mathbf{x}) = (hZ)(\mathbf{x}) + w(\mathbf{x}) = \frac{1}{2\pi} \int d^2y h(\mathbf{x} - \mathbf{y}) Z(\mathbf{y}) + w(\mathbf{x}), \quad (12.117)$$

where again we have assumed translation invariance.

The non-stochastic quantities h and w are not yet known and will be determined from the condition that the expectation value

$$\begin{aligned} \sigma^2 = \langle (I - \hat{I})^2 \rangle &= \left\langle \int d^2x [I(\mathbf{x}) - \hat{I}(\mathbf{x})]^2 \right\rangle \\ &= \left\langle \int d^2x \left[I(\mathbf{x}) - \frac{1}{2\pi} \int d^2y h(\mathbf{x} - \mathbf{y}) Z(\mathbf{y}) - w(\mathbf{x}) \right]^2 \right\rangle \end{aligned} \quad (12.118)$$

becomes minimal. This expresses the idea that, in the average, $\hat{I}$ is an optimal reconstruction of I . Variation with respect to w yields

$$\langle I(\mathbf{x}) - (hZ)(\mathbf{x}) \rangle - w(\mathbf{x}) = \left\langle I(\mathbf{x}) - \int d^2y h(\mathbf{x} - \mathbf{y}) Z(\mathbf{y}) \right\rangle - w(\mathbf{x}) = 0, \quad (12.119)$$

i.e.

$$w(\mathbf{x}) = \langle I(\mathbf{x}) \rangle - (h\langle Z \rangle)(\mathbf{x}) = \langle I(\mathbf{x}) \rangle - \frac{1}{2\pi} \int d^2y h(\mathbf{x} - \mathbf{y}) \langle Z(\mathbf{y}) \rangle. \quad (12.120)$$

Variation with respect to h leads to the condition

$$0 = \langle [I(\mathbf{x}) - (hZ)(\mathbf{x}) - w(\mathbf{x})] Z(\mathbf{y}) \rangle. \quad (12.121)$$

Inserting eq. (12.120) we find

$$\begin{aligned} 0 &= \langle [\Delta I(\mathbf{x}) - (h\Delta Z)(\mathbf{x})] \Delta Z(\mathbf{y}) \rangle \\ &= C_{IZ}(\mathbf{x} - \mathbf{y}) - \frac{1}{2\pi} \int d^2u h(\mathbf{x} - \mathbf{u}) \langle \Delta Z(\mathbf{u}) \Delta Z(\mathbf{y}) \rangle \\ &= C_{IZ}(\mathbf{x} - \mathbf{y}) - \frac{1}{2\pi} \int d^2u h(\mathbf{x} - \mathbf{u}) C_{ZZ}(\mathbf{u} - \mathbf{y}). \end{aligned} \quad (12.122)$$

This is a linear integral equation for h . In new variables, it may be written in the form

$$C_{IZ}(\mathbf{x}) = \frac{1}{2\pi} \int d^2y h(\mathbf{x} - \mathbf{y}) C_{ZZ}(\mathbf{y}). \quad (12.123)$$

This equation is referred to as the *Yule–Walker equation*. Taking into account eqs. (12.113) and (12.116) we obtain

$$C_{IZ}(\mathbf{x}) = C_{II}(\mathbf{x}) = C_{ZZ}(\mathbf{x}) - C_{\Sigma\Sigma}(\mathbf{x}), \quad (12.124)$$

which contains only known and measurable quantities.

Due to the assumed translational invariance, the Yule–Walker equation for h can be solved easily by Fourier transformation and applying the convolution theorem:

$$\tilde{h}(\mathbf{k}) = \frac{\tilde{C}_{ZZ}(\mathbf{k}) - \tilde{C}_{\Sigma\Sigma}(\mathbf{k})}{\tilde{C}_{ZZ}(\mathbf{k})}. \quad (12.125)$$

Hence, we have found a “filter” for an optimal linear estimator for the reconstruction of the signal. This is called the Wiener filter.

12.12 Polarization of partially coherent light

We cannot describe the polarization effects of a stochastic light wave field using a scalar analytical signal $V(t, \mathbf{x})$. If we want to include these effects, we have to describe the three components of the electric field strength by three analytical stochastic signals $E_i(t, \mathbf{x})$. Now, the covariance function, $\langle E_i(t+\tau, \mathbf{x}_1) E_j^*(t, \mathbf{x}_2) \rangle$, assumes values in a matrix.

We restrict ourselves to the important case in which the wave field is observed only at one point $\mathbf{x}$; we set $\tau = 0$ and assume that the propagation of light occurs along the 3-axis. In this case $E_3 \equiv 0$, and the polarization can be described by a 2×2 matrix:

$$S = \begin{pmatrix} \langle E_1 E_1^* \rangle & \langle E_1 E_2^* \rangle \\ \langle E_2 E_1^* \rangle & \langle E_2 E_2^* \rangle \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}. \quad (12.126)$$

Obviously, $S_{11} \geq 0$, $S_{22} \geq 0$, and $S_{12} = S_{21}^*$, i.e. the matrix S is Hermitian: $S = S^+$. Any Hermitian 2×2 matrix may be written in the form

$$S = \begin{pmatrix} s_0 + s_1 & s_2 + is_3 \\ s_2 - is_3 & s_0 - s_1 \end{pmatrix} = s_0 \mathbf{1} + \mathbf{s} \cdot \boldsymbol{\sigma}, \quad (12.127)$$

where

$$\sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix}, \quad (12.128)$$

are the well-known σ matrices. (Compared to the standard notation used in quantum mechanics, the labeling is different in order to be consistent with the convention used in optics.) The

parameters s_0 , s_1 , s_2 , and s_3 characterize the polarization and are called the *Stokes parameters*. Our first aim will be to reveal their meaning. The allowed values for s_0 and $\mathbf{s}$ are not completely arbitrary but are subject to the conditions $s_0 \pm s_1 \geq 0$, i.e. $s_0 \geq 0$, and, in addition, we can deduce from the correlation inequality that

$$\begin{aligned} |\langle E_1 E_2^* \rangle|^2 &\leq \langle E_1 E_1^* \rangle \langle E_2 E_2^* \rangle, \\ S_{12} S_{21} &\leq S_{11} S_{22}, \\ \det S &= S_{11} S_{22} - S_{12} S_{21} = s_0^2 - \mathbf{s}^2 \geq 0. \end{aligned} \quad (12.129)$$

The determinant of S is non-negative, which leads to an additional constraint for the Stokes parameters:

$$s_0 \geq |\mathbf{s}| := s. \quad (12.130)$$

The trace of S ,

$$\text{Tr } S = \langle E_1 E_1^* \rangle + \langle E_2 E_2^* \rangle = 2s_0, \quad (12.131)$$

is proportional to the average intensity. A rotation around the 3-axis,

$$R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

transforms S into

$$S^R = R(\theta) S R^+(\theta). \quad (12.132)$$

Since

$$\text{Tr } S^R = \text{Tr}(R(\theta) S R^+(\theta)) = \text{Tr}(S R^+ R) = \text{Tr } S, \quad (12.133)$$

the intensity remains unchanged, as was to be expected.

Some more operations that one can perform on the wave field using matrices are as follows:

- A *phase shift* by a phase ψ of E_1 relative to E_2 by a so-called *compensator*, which consists of small plates cut from an optically uniaxial crystal. Such a compensator can be represented by a matrix

$$K = \begin{pmatrix} e^{i\psi/2} & 0 \\ 0 & e^{-i\psi/2} \end{pmatrix}.$$

We find

$$S^K = K(\psi) S K^+(\psi) = \begin{pmatrix} S_{11} & e^{i\psi} S_{12} \\ e^{-i\psi} S_{21} & S_{22} \end{pmatrix}, \quad (12.134)$$

and, of course, $\text{Tr } S = \text{Tr } S^K$.

- Propagation through a *polarization filter*. In this case the E vector is projected onto the polarization direction $\mathbf{n} = (\cos \vartheta, \sin \vartheta, 0)$ of the filter. A polarization filter is represented by a projection operator

$$P(\vartheta) = \begin{pmatrix} \cos^2 \vartheta & \sin \vartheta \cos \vartheta \\ \sin \vartheta \cos \vartheta & \sin^2 \vartheta \end{pmatrix}, \quad (12.135)$$

which satisfies $P^+(\vartheta) = P(\vartheta) = P^2(\vartheta)$. We find

$$S^P = PSP$$

and in this case

$$\text{Tr } S^P = \text{Tr}(PSP) = \text{Tr}(SP^2) = \text{Tr } SP \leq \text{Tr } S, \quad (12.136)$$

because part of the intensity is absorbed by the filter.

We can directly determine all the matrix elements of S by intensity measurements using combinations of polarization filters and compensators. For instance,

$$\begin{aligned} \text{Tr } SP(0) &= S_{11} = s_0 + s_1, \\ \text{Tr } SP(\pi/2) &= S_{22} = s_0 - s_1, \\ \text{Tr } SP(\pi/4) &= \frac{1}{2}(S_{11} + S_{22} + S_{12} + S_{21}) = s_0 + s_2, \\ \text{Tr}(K(\pi/2)SK^+(\pi/2)P(\pi/4)) &= \frac{1}{2}(S_{11} + S_{22} + iS_{12} - iS_{21}) = s_0 - s_3. \end{aligned} \quad (12.137)$$

For a better understanding of how the information about the polarization is hidden in the matrix S , we first consider some special cases:

1. *Completely unpolarized light*. There is no preferred direction and there is no correlation between the different components of E . So

$$S = \begin{pmatrix} \langle E^2 \rangle & 0 \\ 0 & \langle E^2 \rangle \end{pmatrix}, \quad \text{i.e. } S = s_0 \mathbf{1}, \quad \mathbf{s} = 0, \quad \det S = s_0^2. \quad (12.138)$$

2. *Strictly planar polarization*. Now, E_2 is already completely determined by E_1 and shares the same phase as E_1 :

$$\begin{aligned} E_1 &= AE_2 \quad (A \text{ real}), \\ S &= \langle E_1 E_1^* \rangle \begin{pmatrix} 1 & A \\ A & A^2 \end{pmatrix}. \end{aligned} \quad (12.139)$$

For this case we find $\det S = 0$ and $s_3 = 0$.

3. *Strictly elliptical polarization with the axes of the polarization ellipsoid in the 1–2 direction*. For this case also, E_2 is completely determined by E_1 and there is a phase difference between the two components of $\pm\pi/2$:

$$\begin{aligned} E_1 &= \pm i A E_2 \quad (\text{with } A > 0), \\ S &= \langle E_1 E_1^* \rangle \begin{pmatrix} 1 & \pm i A \\ \mp i A & A^2 \end{pmatrix}, \end{aligned} \quad (12.140)$$

Again, we find $\det S = 0$.

4. *Strictly elliptical polarization with arbitrary direction of the axes.* This follows by applying a rotation.

For all cases of strict polarizations we find $\det S = 0$. It turns out that any polarization matrix S can be written uniquely as the sum of a completely unpolarized and a completely polarized part:

$$S = S_u + S_p. \quad (12.141)$$

In order to find this splitting, we observe that S has the eigenvalues $\lambda_{1,2} = s_0 \pm |s| = s_0 \pm s$, and that $\det(S - \lambda_{1,2}\mathbf{1}) = 0$. Therefore, $S = \lambda_{1,2}\mathbf{1} + (S - \lambda_{1,2}\mathbf{1})$. The requirement that the diagonal elements of $S - \lambda_{1,2}\mathbf{1}$ shall not be negative leads to the unique splitting

$$S = \begin{pmatrix} s_0 - s & 0 \\ 0 & s_0 - s \end{pmatrix} + \begin{pmatrix} s + s_1 & s_2 + is_3 \\ s_2 - is_3 & s - s_1 \end{pmatrix}. \quad (12.142)$$

The intensity of the polarized part is

$$I_p = \text{Tr } S_p = 2s \quad (12.143)$$

and for the unpolarized part we find

$$I_u = \text{Tr } S_u = 2s_0 - 2s. \quad (12.144)$$

So, the degree of polarization is given by

$$p = \frac{I_p}{I_p + I_u} = \frac{s}{s_0}. \quad (12.145)$$

Indeed, we obtain $0 \leq p \leq 1$, where the extremal values correspond to the completely unpolarized and the completely polarized cases. Having found the degree of polarization, we now want to extract the type of polarization. For this purpose we consider the transformation behaviour of the Stokes parameters under a rotation $R(\theta)$. We find the following replacements:

$$\begin{aligned} s_0 &\mapsto s'_0 = s_0, \\ s_1 &\mapsto s'_1 = s_1 \cos(2\theta) - s_2 \sin(2\theta), \\ s_2 &\mapsto s'_2 = s_1 \sin(2\theta) + s_2 \cos(2\theta), \\ s_3 &\mapsto s_3, \\ s &\mapsto s. \end{aligned} \quad (12.146)$$

Making a rotation by an angle θ_0 , where

$$\tan(2\theta_0) = -s_2/s_1, \quad (12.147)$$

we can arrange $s'_2 = 0$ such that the polarized part assumes the form of eq. (12.140):

$$S_p = \begin{pmatrix} s + s'_1 & is_3 \\ -is_3 & s - s'_1 \end{pmatrix} = \begin{pmatrix} s \pm \sqrt{s_1^2 + s_2^2} & is_3 \\ -is_3 & s \mp \sqrt{s_1^2 + s_2^2} \end{pmatrix}. \quad (12.148)$$

Therefore, the polarized part describes

- a linearly polarized wave for $s_3 = 0$, and
- an elliptically polarized wave for $s_3 \neq 0$, where the sign determines the rotational direction along the polarization ellipsoid.

The directions of the axes of the polarization ellipsoid are given by

$$\tan \theta_0 = -s_2/s_1,$$

and the ratio of the absolute values of the axes is

$$\alpha^2 = \frac{s - \sqrt{s_1^2 + s_2^2}}{s + \sqrt{s_1^2 + s_2^2}}. \quad (12.149)$$

The values $s_1 = s_2 = 0$ imply a circular polarization and $s = \sqrt{s_1^2 + s_2^2}$, i.e. $s_3 = 0$, a linear polarization.

13 Quantum states of the electromagnetic field

13.1 Quantization of the electromagnetic field and harmonic oscillators

The quantization of the free electromagnetic field can be reduced to the quantization of an infinite system of uncoupled harmonic oscillators. So, we first have to recapitulate the quantization of a simple harmonic oscillator.

A simple classical harmonic oscillator is defined by the equation of motion

$$m\ddot{q} + kq = 0 \quad (13.1)$$

with solution

$$q(t) = \alpha \cos(\omega t) + \beta \sin(\omega t) \quad (13.2)$$

and the frequency ω given by $\omega = \sqrt{k/m}$. Replacing k by ω , we obtain for the conserved energy

$$E = \frac{m}{2} \dot{q}^2 + \frac{m\omega^2}{2} q^2 \quad (13.3)$$

and for the classical Hamiltonian

$$H_{cl} = \frac{p^2}{2m} + \frac{m\omega^2}{2} q^2. \quad (13.4)$$

For quantization, the momentum $p = m\dot{q}$ and the position q are replaced by operators P and Q with

$$[P, Q] = \frac{\hbar}{i} \mathbf{1}, \quad (13.5)$$

giving the Hamilton operator:

$$H = \frac{P^2}{2m} + \frac{m\omega^2}{2} Q^2. \quad (13.6)$$

Rescaling P and Q to dimensionless quantities, we obtain

$$H = \hbar\omega \left[\frac{1}{2} \left(\frac{P}{\sqrt{m\hbar\omega}} \right)^2 + \frac{1}{2} \left(\sqrt{\frac{m\omega}{\hbar}} Q \right)^2 \right] \quad (13.7)$$

$$= \hbar\omega \left(\frac{1}{2} X_2^2 + \frac{1}{2} X_1^2 \right). \quad (13.8)$$

The eigenvalues and eigenstates of H are obtained in a purely algebraic way. We introduce the operator

$$a = \frac{1}{\sqrt{2}}(X_1 + iX_2) \quad (13.9)$$

and its adjoint

$$a^+ = \frac{1}{\sqrt{2}}(X_1 - iX_2). \quad (13.10)$$

From

$$[X_1, X_2] = i\mathbf{1}, \quad (13.11)$$

we immediately obtain

$$[a, a^+] = \mathbf{1}. \quad (13.12)$$

Here a and a^+ are called the *annihilation operator* and the *creation operator*, respectively, and

$$X_1 = \frac{1}{\sqrt{2}}(a + a^+), \quad X_2 = \frac{1}{\sqrt{2}i}(a - a^+) \quad (13.13)$$

are called the *quadratures* associated with a and a^+ . In terms of a and a^+ , the Hamilton operator H becomes

$$H = \hbar\omega(a^+a + \frac{1}{2}). \quad (13.14)$$

The eigenvalues and eigenstates of $N = a^+a$ are easily obtained in the following way. Let $|\nu\rangle$ be a state with

$$N|\nu\rangle = \nu|\nu\rangle, \quad \langle\nu|\nu\rangle = 1,$$

then

$$\begin{aligned} Na|\nu\rangle &= aN|\nu\rangle + [N, a]|\nu\rangle = (\nu - 1)a|\nu\rangle, \\ \langle a\nu|a\nu\rangle &= \langle\nu|a^+a\nu\rangle = \nu, \end{aligned} \quad (13.15)$$

and

$$\begin{aligned} Na^+|\nu\rangle &= [N, a^+]|\nu\rangle + a^+N|\nu\rangle = (\nu + 1)a^+|\nu\rangle, \\ \langle a^+\nu|a^+\nu\rangle &= \langle\nu|aa^+|\nu\rangle = \nu + 1. \end{aligned} \quad (13.16)$$

So, a^+ and a raise and lower the eigenvalue of N by one unit. Because N is a non-negative operator, ν has to be a non-negative integer n .

Starting from a ground state $|0\rangle$ with $a|0\rangle = N|0\rangle = 0$, $\langle 0|0\rangle = 1$, we can obtain the excited states

$$|n\rangle = \frac{(a^+)^n}{\sqrt{n!}}|0\rangle \quad (13.17)$$

with $N|n\rangle = n|n\rangle$, which are also eigenstates of H :

$$H|n\rangle = \hbar\omega(n + \frac{1}{2})|n\rangle. \quad (13.18)$$

Harmonic oscillators in f -dimensional space can, by linear transformation, always be reduced to f decoupled harmonic oscillators with

$$H = \sum_{l=1}^f \hbar\omega_l(a_l^+ a_l + \frac{1}{2}), \quad (13.19)$$

where

$$[a_l, a_m] = [a_l^+, a_m^+] = 0, \quad [a_l, a_m^+] = \delta_{lm} \mathbf{1}. \quad (13.20)$$

The normalized eigenstates are

$$\begin{aligned} |n_1, \dots, n_f\rangle &= \frac{(a_1^+)^{n_1}}{\sqrt{n_1!}} \cdots \frac{(a_f^+)^{n_f}}{\sqrt{n_f!}} |0\rangle, \\ \langle n_1, \dots, n_f | n'_1, \dots, n'_f \rangle &= \delta_{n_1 n'_1} \cdots \delta_{n_f n'_f}, \\ H|n_1, \dots, n_f\rangle &= \sum_{l=1}^f \hbar\omega_l(n_l + \frac{1}{2})|n_1, \dots, n_f\rangle. \end{aligned} \quad (13.21)$$

From eq. (13.7) we see the following. Different harmonic oscillators with different values of $m\omega$ give rise to different quadratures X'_1 and X'_2 of the form:

$$X'_1 = e^r X_1, \quad X'_2 = e^{-r} X_2, \quad (13.22)$$

and different annihilation and creation operators

$$\begin{aligned} a' &= a \cosh r + a^+ \sinh r, \\ a'^+ &= a \sinh r + a^+ \cosh r. \end{aligned} \quad (13.23)$$

These X'_1 , X'_2 , a' , and a'^+ also fulfill the commutation relations (13.11) and (13.12). Another linear transformation conserving the commutation laws is a rotation of the quadratures:

$$\begin{aligned} X'_1 &= X_1 \cos \theta + X_2 \sin \theta, \\ X'_2 &= -X_1 \sin \theta + X_2 \cos \theta, \end{aligned} \quad (13.24)$$

or

$$a' = a e^{-i\theta}, \quad a'^+ = a^+ e^{i\theta}. \quad (13.25)$$

The transformations (13.22) and (13.23) are called *Bogolyubov transformations*, and every linear inhomogeneous transformation preserving Hermiticity properties and the commutation laws is a product of rotations, Bogolyubov transformations, and translations $a' = a + \alpha \mathbf{1}$,

$a'^+ = a^+ + \alpha^* \mathbf{1}$. For r degrees of freedom, the linear transformations preserving relations (13.20) are symplectic transformations.

An irreducible representation of the commutation relations (13.20) with Hermitian operators X_{1l} and X_{2l} is uniquely defined up to a unitary equivalence

$$a' = U^+ a U, \quad a'^+ = U^+ a^+ U \quad (13.26)$$

with unitary U . The exact form of U will be determined later.

Let us now turn to the quantization of the free electromagnetic field. Its energy is given by eq. (2.15) (making use of eq. (2.8)):

$$E = \int_V d^3x \left(\frac{\epsilon_0}{2} \mathbf{E}^2 + \frac{1}{2\mu_0} \mathbf{B}^2 \right) = \int_V d^3x \left(\frac{\epsilon_0}{2} \mathbf{E}^2 + \frac{c^2 \kappa^2 \epsilon_0}{2} \mathbf{B}^2 \right). \quad (13.27)$$

Here, we assume the fields to be restricted to a (possibly large) rectangular volume $V = ABC$ of edge lengths A , B , and C . Expressing $\mathbf{B}$ and $\mathbf{E}$ by the potentials $\mathbf{A}$ and ϕ according to eq. (2.18) and inserting the Coulomb gauge condition

$$\nabla \cdot \mathbf{A} = 0, \quad \phi = 0 \quad (13.28)$$

on the potentials, we obtain

$$E = \int_V d^3x \left[\frac{1}{2} \epsilon_0 \kappa^2 \dot{\mathbf{A}}^2 + \frac{1}{2} c^2 \kappa^2 \epsilon_0 (\nabla \times \mathbf{A})^2 \right]. \quad (13.29)$$

Now

$$\begin{aligned} (\nabla \times \mathbf{A})^2 &= \nabla_i A_j \nabla_i A_j - \nabla_i A_j \nabla_j A_i \\ &= \nabla_i (A_j \nabla_i A_j - A_j \nabla_j A_i) - A_j \Delta A_j + A_j \nabla_j \nabla_i A_i \\ &= -A_j \Delta A_j + \nabla_i (A_j \nabla_i A_j - A_j \nabla_j A_i). \end{aligned}$$

Hence, by partial integration,

$$E = \int_V d^3x \left(\frac{1}{2} \epsilon_0 \kappa^2 \dot{\mathbf{A}}^2 - \frac{1}{2} c^2 \kappa^2 \epsilon_0 \mathbf{A} \cdot \Delta \mathbf{A} \right). \quad (13.30)$$

Choosing a complete orthogonal system of eigenfunctions $\mathbf{u}_l(\mathbf{x})$ of the non-negative operator $-\Delta$ with

$$\Delta \mathbf{u}_l = -k_l^2 \mathbf{u}_l, \quad \int_V \mathbf{u}_l^* \cdot \mathbf{u}_m d^3x = \delta_{lm}$$

(depending on the boundary conditions on ∂V) we have

$$\begin{aligned} \mathbf{A}(t, \mathbf{x}) &= \sum_l q_l(t) \mathbf{u}_l(\mathbf{x}), \\ \dot{\mathbf{A}}(t, \mathbf{x}) &= \sum_l \dot{q}_l(t) \mathbf{u}_l(\mathbf{x}), \end{aligned} \quad (13.31)$$

and

$$E = \sum_l \left(\frac{1}{2} \epsilon_0 \kappa^2 \dot{q}_l^2 + \frac{1}{2} \epsilon_0 \kappa^2 \omega_l^2 q_l^2 \right) \quad (13.32)$$

with the frequency $\omega_l = k_l c$.

The quantization is now straightforward. The Hamiltonian is

$$\hat{H} = \sum_l \left(\frac{1}{2\epsilon_0 \kappa^2} P_l^2 + \frac{\epsilon_0 \kappa^2 \omega_l^2}{2} Q_l^2 \right). \quad (13.33)$$

Notice that the electric field contributes to the kinetic energy and the magnetic field to the potential energy.

Comparing with eq. (13.7) we immediately read off the annihilation and creation operators:

$$\begin{aligned} a_l &= \frac{1}{\sqrt{2}} \left(\sqrt{\frac{\epsilon_0 \kappa^2 \omega_l}{\hbar}} Q_l + i \sqrt{\frac{1}{\epsilon_0 \kappa^2 \omega_l \hbar}} P_l \right), \\ a_l^+ &= \frac{1}{\sqrt{2}} \left(\sqrt{\frac{\epsilon_0 \kappa^2 \omega_l}{\hbar}} Q_l - i \sqrt{\frac{1}{\epsilon_0 \kappa^2 \omega_l \hbar}} P_l \right), \end{aligned} \quad (13.34)$$

in terms of which

$$\hat{H} = \sum_l \hbar \omega_l \left(a_l^+ a_l + \frac{1}{2} \right). \quad (13.35)$$

Application of a_l^+ to an energy eigenstate $|0\rangle$ raises the energy by $\hbar \omega_l$ and corresponds to the creation of an electromagnetic or light quantum of energy $\hbar \omega_l$. It is natural to assume that the ground state $|0\rangle$ with no light quanta has energy zero. Hence, we replace $\hat{H}$ by the final Hamiltonian

$$H = \sum_l \hbar \omega_l a_l^+ a_l \quad (13.36)$$

of the electromagnetic field.

The operators of eq. (13.34) refer to the Schrödinger picture in quantum mechanics, where the time dependence resides in the states. In field theory, it is more common to work in the Heisenberg pictures, where states are time-independent and operators time-dependent. The relationship between Heisenberg and Schrödinger picture operators is given by

$$A_H(t) = e^{(i/\hbar)Ht} A e^{-(i/\hbar)Ht}, \quad (13.37)$$

and with eq. (13.36) this gives

$$a_l(t) = a_l e^{-i\omega_l t}, \quad a_l^+(t) = a_l^+ e^{i\omega_l t}. \quad (13.38)$$

The quantum vector potential is given by the operator

$$\begin{aligned} \mathbf{A}_H(t, \mathbf{x}) &= \sum_l \sqrt{\frac{\hbar}{\epsilon_0 \kappa^2 \omega_l}} \frac{[a_l e^{-i\omega_l t} \mathbf{u}_l(\mathbf{x}) + a_l^\dagger e^{+i\omega_l t} \mathbf{u}_l^*(\mathbf{x})]}{\sqrt{2}} \\ &= \mathbf{A}^{(+)} + \mathbf{A}^{(-)}. \end{aligned} \quad (13.39)$$

The annihilation and creation parts $\mathbf{A}^{(\pm)}$ are Hermitian conjugates of each other.

One easily derives the field operators for the electric and magnetic fields. For example,

$$\begin{aligned} \mathbf{E}_H(t, \mathbf{x}) &= \sum_l i \sqrt{\frac{\hbar \omega_l}{\epsilon_0}} \frac{[a_l \mathbf{u}_l(\mathbf{x}) e^{-i\omega_l t} - a_l^\dagger \mathbf{u}_l^*(\mathbf{x}) e^{+i\omega_l t}]}{\sqrt{2}} \\ &= \mathbf{E}^{(+)} + \mathbf{E}^{(-)}. \end{aligned} \quad (13.40)$$

The operators $\mathbf{A}^{(+)}(t, \mathbf{x})$ and $\mathbf{E}^{(+)}(t, \mathbf{x})$ annihilate a photon localized at the point $\mathbf{x}$, whereas $\mathbf{A}^{(-)}(t, \mathbf{x})$ and $\mathbf{E}^{(-)}(t, \mathbf{x})$ create a photon at the point $\mathbf{x}$.

The quadratures

$$X_{1l} = \frac{1}{\sqrt{2}}(a_l + a_l^\dagger), \quad X_{2l} = \frac{1}{i\sqrt{2}}(a_l - a_l^\dagger)$$

play an important role in quantum optics.

To finish this section, we give examples of the precise form of the basis functions $\mathbf{u}_l(\mathbf{x})$. For periodic boundary conditions we have

$$\mathbf{u}_l(\mathbf{x}) = \boldsymbol{\epsilon}_l \frac{e^{i\mathbf{k}_l \cdot \mathbf{x}}}{\sqrt{ABC}} \quad (13.41)$$

with

$$\mathbf{k}_l = \left(\frac{2\pi l_1}{A}, \frac{2\pi l_2}{B}, \frac{2\pi l_3}{C} \right), \quad l_1, l_2, l_3 \in \mathbb{Z}$$

and $\boldsymbol{\epsilon}_l \cdot \mathbf{k}_l = 0$. The $\boldsymbol{\epsilon}_l$ contains polarization information. For a basis we need two orthogonal vectors $\boldsymbol{\epsilon}_l$. Periodic boundary conditions correspond to running waves: a photon has energy $\hbar\omega_l$ and momentum $\hbar\mathbf{k}_l$.

For standing waves, we can take

$$\mathbf{u}_l(\mathbf{x}) = \sqrt{\frac{8}{ABC}} \begin{pmatrix} \epsilon_1^{(l)} \sin(k_1^{(l)} x_1) \cos(k_2^{(l)} x_2) \cos(k_3^{(l)} x_3) \\ \epsilon_2^{(l)} \cos(k_1^{(l)} x_1) \sin(k_2^{(l)} x_2) \cos(k_3^{(l)} x_3) \\ \epsilon_3^{(l)} \cos(k_1^{(l)} x_1) \cos(k_2^{(l)} x_2) \sin(k_3^{(l)} x_3) \end{pmatrix} \quad (13.42)$$

with

$$\mathbf{k}_l = (k_1^{(l)}, k_2^{(l)}, k_3^{(l)}) = \left(\frac{\pi l_1}{A}, \frac{\pi l_2}{B}, \frac{\pi l_3}{C} \right)$$

and $\mathbf{k}_l \cdot \boldsymbol{\epsilon}_l = 0$. This time, the sign of the momentum of one light quantum is undetermined.

In quantum mechanics, there are no privileged harmonic oscillators, and one pair of creation and annihilation operators is as good as any other.

For a quantized electromagnetic field, on the other hand, the above set of creation and annihilation operators is physically distinguished. Nevertheless, Bogolyubov transformations play an important role in connection with so-called squeezed states, which will be defined in the next section.

In quantum optics, quite often a considerable simplification applies for the electromagnetic field, because only a finite, small number of modes contributes.

13.2 Coherent and squeezed states

In the previous section, we mentioned that the inhomogeneous linear transformations, which preserve the Hermiticity properties and the commutation laws of the creation and annihilation operators and of the quadratures, can be realized by unitary transformations. We now give explicit expressions for these unitary transformations.

- We start with *translations* (*Galilean transformations*). For $\alpha \in \mathbb{C}$ we define a unitary operator

$$D(\alpha) = e^{\alpha a^\dagger - \alpha^* a}. \quad (13.43)$$

We readily see that

$$D^\dagger(\alpha) a D(\alpha) = a + \alpha, \quad D^\dagger(\alpha) a^\dagger D(\alpha) = a^\dagger + \alpha^*. \quad (13.44)$$

(For a proof, we replace α by αt , differentiate with respect to t , and observe that both sides of eq. (13.44) solve the same linear differential equation and coincide for $t = 0$.)

For the quadratures we find

$$D^\dagger(\alpha) X_1 D(\alpha) = X_1 + \frac{\alpha + \alpha^*}{\sqrt{2}}, \quad D^\dagger(\alpha) X_2 D(\alpha) = X_2 + \frac{\alpha - \alpha^*}{i\sqrt{2}}. \quad (13.45)$$

Using the well-known identity

$$e^A e^B = e^{A+B} e^{\frac{1}{2}[A,B]}, \quad (13.46)$$

which holds whenever the commutator $[A, B]$ commutes with A and B , we can write $D(\alpha)$ also as

$$D(\alpha) = e^{-\frac{1}{2}|\alpha|^2} e^{\alpha a^\dagger} e^{-\alpha^* a}. \quad (13.47)$$

- We now turn to *rotations*. For the unitary operator

$$R(\theta) = e^{-i\theta a^\dagger a} \quad (\theta \in \mathbb{R}) \quad (13.48)$$

we find

$$R^\dagger(\theta) a R(\theta) = a e^{-i\theta}, \quad R^\dagger(\theta) a^\dagger R(\theta) = a^\dagger e^{i\theta}, \quad (13.49)$$

and, hence,

$$\begin{aligned} R^+(\theta)X_1R(\theta) &= X_1 \cos \theta + X_2 \sin \theta = \frac{1}{\sqrt{2}}(a e^{-i\theta} + a^+ e^{i\theta}), \\ R^+(\theta)X_2R(\theta) &= -X_1 \sin \theta + X_2 \cos \theta = \frac{1}{i\sqrt{2}}(a e^{-i\theta} - a^+ e^{i\theta}). \end{aligned} \quad (13.50)$$

The time evolution operator

$$U(t) = e^{-i\omega a^+ a t} = e^{-(i/\hbar)Ht} \quad (13.51)$$

is, of course, a phase space rotation operator with $\theta = \omega t$.

- Finally, we consider *Bogolyubov transformations*. For $r \in \mathbb{R}$ we define the unitary operator

$$B(r) = e^{\frac{1}{2}ir(X_1X_2+X_2X_1)} \quad (13.52)$$

and obtain

$$B^+(r)X_1B(r) = X_1 e^{-r}, \quad B^+(r)X_2B(r) = X_2 e^r \quad (13.53)$$

or

$$\begin{aligned} B^+(r)aB(r) &= a \cosh r - a^+ \sinh r, \\ B^+(r)a^+B(r) &= a^+ \cosh r - a \sinh r. \end{aligned} \quad (13.54)$$

Written in terms of a and a^+ , the operator $B(r)$ assumes the form

$$B(r) = e^{\frac{1}{2}r[a^2 - (a^+)^2]}. \quad (13.55)$$

A slight generalization of eq. (13.55) is the so-called *squeezing operator*:

$$S(\xi) = e^{\frac{1}{2}[\xi^* a^2 - \xi (a^+)^2]} \quad \text{with } \xi \in \mathbb{C}. \quad (13.56)$$

For $\xi = r e^{i\theta}$, we infer from eq. (13.49) that

$$S(\xi) = R^+(\theta/2)B(r)R(\theta/2) \quad (13.57)$$

and

$$\begin{aligned} S^+(\xi)aS(\xi) &= a \cosh r - a^+ e^{i\theta} \sinh r, \\ S^+(\xi)a^+S(\xi) &= a^+ \cosh r - a e^{-i\theta} \sinh r. \end{aligned} \quad (13.58)$$

Coherent states are of utmost importance in quantum optics. For $\alpha \in \mathbb{C}$ we define the normalized coherent state (with respect to the pair a and a^+) by

$$|\alpha\rangle = D(\alpha)|0\rangle, \quad (13.59)$$

with $D(\alpha)$ from eq. (13.43). This means that $|\alpha\rangle$ is just a Galilean-transformed ground state.

Of course, $|\alpha\rangle = D(\alpha)|0\rangle$ is annihilated by the transformed annihilation operator a' :

$$\begin{aligned} a' &= D(\alpha)aD^\dagger(\alpha) = a - \alpha, \\ (a - \alpha)|\alpha\rangle &= D(\alpha)aD^\dagger(\alpha)D(\alpha)|0\rangle = D(\alpha)a|0\rangle = 0. \end{aligned}$$

In other words,

$$a|\alpha\rangle = \alpha|\alpha\rangle, \quad (13.60)$$

and $|\alpha\rangle$ is an eigenstate of the annihilation operator a with eigenvalue α . (We notice that the spectrum of a is the whole complex plane, whereas $a^\dagger$ has no eigenstates at all.)

With eq. (13.47) we can write $|\alpha\rangle$ in the equivalent form:

$$\begin{aligned} |\alpha\rangle &= e^{-\frac{1}{2}|\alpha|^2} e^{\alpha a^\dagger} |0\rangle \\ &= e^{-\frac{1}{2}|\alpha|^2} \sum_{n=0}^{\infty} \frac{\alpha^n (a^\dagger)^n}{n!} |0\rangle \\ &= e^{-\frac{1}{2}|\alpha|^2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \end{aligned} \quad (13.61)$$

(see eq. (13.17)). Hence,

$$\langle n|\alpha\rangle = e^{-\frac{1}{2}|\alpha|^2} \frac{\alpha^n}{\sqrt{n!}}, \quad (13.62)$$

and the probability of finding n photons in a given state $|\alpha\rangle$ equals

$$w_n = \frac{|\alpha|^{2n}}{n!} e^{-|\alpha|^2}. \quad (13.63)$$

This means that in a coherent state $|\alpha\rangle$ the photon number is Poisson distributed with mean value and variance, respectively,

$$\langle n\rangle = |\alpha|^2, \quad \sigma_n = |\alpha|. \quad (13.64)$$

For $\alpha \neq \beta$ the coherent states $|\alpha\rangle$ and $|\beta\rangle$ are not orthogonal. Instead, with eq. (13.61), we find

$$\langle\beta|\alpha\rangle = e^{-\frac{1}{2}(|\alpha|^2 + |\beta|^2)} \langle 0|e^{\beta^* a} e^{\alpha a^\dagger}|0\rangle \quad (13.65)$$

and, using eq. (13.60),

$$\langle\beta|\alpha\rangle = e^{-\frac{1}{2}(|\alpha|^2 + |\beta|^2) + \beta^* \alpha} = e^{-\frac{1}{2}(|\alpha|^2 + |\beta|^2 - \beta^* \alpha - \alpha^* \beta - \beta^* \alpha + \alpha^* \beta)}. \quad (13.66)$$

Thus

$$\langle\beta|\alpha\rangle = e^{-\frac{1}{2}|\alpha - \beta|^2 + \frac{1}{2}(\beta^* \alpha - \alpha^* \beta)} \quad (13.67)$$

and

$$|\langle\beta|\alpha\rangle|^2 = e^{-|\alpha-\beta|^2}. \quad (13.68)$$

The overlap of $|\alpha\rangle$ and $|\beta\rangle$ decreases exponentially with $|\alpha - \beta|^2$.

Further useful consequences of eq. (13.61) are the following identities:

$$a^+|\alpha\rangle = e^{-\frac{1}{2}|\alpha|^2} \frac{\partial}{\partial\alpha} e^{+\frac{1}{2}|\alpha|^2} |\alpha\rangle = \left(\frac{\partial}{\partial\alpha} + \frac{\alpha^*}{2} \right) |\alpha\rangle, \quad (13.69)$$

$$|n\rangle = \frac{1}{\sqrt{n!}} \left(\frac{\partial}{\partial\alpha} + \frac{\alpha^*}{2} \right)^n |\alpha\rangle \Big|_{\alpha=0}, \quad (13.70)$$

and

$$\langle m|C|n\rangle = \frac{1}{\sqrt{m!}\sqrt{n!}} e^{-|\alpha|^2} \frac{\partial^m}{\partial\alpha^{*m}} \frac{\partial^n}{\partial\alpha^n} e^{|\alpha|^2} \langle\alpha|C|\alpha\rangle \Big|_{\alpha=0}. \quad (13.71)$$

The set of $\{|\alpha\rangle\}$ of all coherent states is over-complete in the Hilbert space $\mathcal{H}$. One finds

$$\begin{aligned} \frac{1}{\pi} \int d^2\alpha |\alpha\rangle\langle\alpha| &= \mathbf{1}, \\ |\psi\rangle &= \frac{1}{\pi} \int d^2\alpha |\alpha\rangle\langle\alpha|\psi\rangle, \\ \langle\phi|\psi\rangle &= \frac{1}{\pi} \int d^2\alpha \langle\phi|\alpha\rangle\langle\alpha|\psi\rangle, \end{aligned} \quad (13.72)$$

where $d^2\alpha = d\alpha_R d\alpha_I$, with α_R and α_I being the real and imaginary parts of α , respectively. The identity

$$|\beta\rangle = \frac{1}{\pi} \int d^2\alpha |\alpha\rangle\langle\alpha|\beta\rangle = \frac{1}{\pi} \int d^2\alpha |\alpha\rangle e^{-\frac{1}{2}|\alpha-\beta|^2 + \frac{1}{2}(\alpha^*\beta - \alpha\beta^*)} \quad (13.73)$$

reveals that the coherent states are not linearly independent.

For a proof of the important identity (13.72), we show that for normalized eigenstates $|m\rangle$ and $|n\rangle$ of a^+a we have

$$\frac{1}{\pi} \int d^2\alpha \langle m|\alpha\rangle\langle\alpha|n\rangle = \delta_{mn}. \quad (13.74)$$

Indeed, according to eq. (13.62), we have

$$\frac{1}{\pi} \int d^2\alpha \langle m|\alpha\rangle\langle\alpha|n\rangle = \frac{1}{\pi} \int d^2\alpha \frac{(\alpha^*)^n \alpha^m}{\sqrt{m!}\sqrt{n!}} e^{-|\alpha|^2}, \quad (13.75)$$

and this integral is easily evaluated, e.g. using polar coordinates in the α_R vs. α_I plane, to yield δ_{mn} .

We also mention a useful formula for the trace of an operator A :

$$\text{Tr } A = \frac{1}{\pi} \int d^2\alpha \langle\alpha|A|\alpha\rangle. \quad (13.76)$$

For a proof we calculate

$$\begin{aligned}\text{Tr } A &= \sum_n \langle n|A|n \rangle = \sum_n \frac{1}{\pi} \int d^2\alpha \langle n|A|\alpha \rangle \langle \alpha|n \rangle \\ &= \frac{1}{\pi} \int d^2\alpha \sum_n \langle \alpha|n \rangle \langle n|A|\alpha \rangle = \frac{1}{\pi} \int d^2\alpha \langle \alpha|A|\alpha \rangle.\end{aligned}$$

Coherent states remain coherent under time evolution with the operator $U(t)$. From eq. (13.51) we get

$$\begin{aligned}U(t)|\alpha \rangle &= e^{-|\alpha|^2} U(t) e^{\alpha a^\dagger} U^\dagger(t) |0 \rangle \\ &= e^{-|\alpha|^2} e^{\alpha a^\dagger e^{-i\omega t}} |0 \rangle.\end{aligned}$$

Thus,

$$e^{-i\omega a^\dagger a t} |\alpha \rangle = |\alpha e^{-i\omega t} \rangle. \quad (13.77)$$

Coherent states are the ground states of a translated oscillator. A translation of the oscillator potential can be achieved by adding a linear term:

$$V = \frac{D}{2} q^2 - Eq = \frac{D}{2} \left(q - \frac{E}{D} \right)^2 - \frac{E^2}{2D}.$$

Thus, a coherent state of a harmonic oscillator can be prepared by starting from the ground state of an untranslated harmonic oscillator, then adiabatically switching on a linear term such that one always stays in the ground state of the varying Hamiltonian and then suddenly switching off the linear term.

From eq. (13.59) it is also clear that coherent states should be generated by any interaction Hamiltonian linear in a and $a^\dagger$ or X_1 and X_2 . Indeed, a time-dependent classical current distribution $\mathbf{J}(t, \mathbf{x})$ interacting with the electromagnetic field via the interaction Hamiltonian

$$H_I = \int d^3x \mathbf{J}(t, \mathbf{x}) \cdot \mathbf{A}(t, \mathbf{x})$$

will radiate off an electromagnetic field in a coherent state. In practice, coherent states of the electromagnetic field are produced by a laser working well above threshold.

Squeezed states are just coherent states constructed with the squeezed annihilation and creation operators:

$$a_\xi = S(\xi) a S^\dagger(\xi), \quad a_\xi^\dagger = S(\xi) a^\dagger S^\dagger(\xi), \quad (13.78)$$

with the squeezing operator $S(\xi)$ as defined in eq. (13.56). The squeezed vacuum state

$$|0, \xi \rangle = S(\xi) |0 \rangle \quad (13.79)$$

is annihilated by a_ξ :

$$a_\xi |0, \xi \rangle = 0. \quad (13.80)$$

With

$$D_\xi(\alpha) = e^{\alpha a_\xi^+ - \alpha^* a_\xi} = S(\xi)D(\alpha)S^+(\xi), \quad (13.81)$$

the normalized squeezed state $|\alpha, \xi\rangle$ is defined by

$$|\alpha, \xi\rangle = D_\xi(\alpha)|0, \xi\rangle. \quad (13.82)$$

A simple calculation yields

$$|\alpha, \xi\rangle = S(\xi)D(\alpha)S^+(\xi)S(\xi)|0\rangle = S(\xi)D(\alpha)|0\rangle = S(\xi)|\alpha\rangle.$$

Hence,

$$|\alpha, \xi\rangle = S(\xi)|\alpha\rangle, \quad (13.83)$$

and squeezed states are obtained from coherent states simply by applying the squeezing operator $S(\xi)$.

Obviously, the overlap relations (13.67) and the completeness condition (13.72) are also valid for squeezed states for a fixed squeezing parameter ξ . The time evolution as described by $U = e^{-i\omega a^+ a}$ for the squeezing operator $S(\xi)$ yields

$$U(t)S(\xi)U^+(t) = S(\xi e^{-2i\omega t}). \quad (13.84)$$

Thus, squeezed states remain squeezed states under time evolution:

$$U(t)|\alpha, \xi\rangle = |\alpha e^{-i\omega t}, \xi e^{-2i\omega t}\rangle. \quad (13.85)$$

The squeezing parameters will change with time.

The photon number distribution

$$w_{n,\alpha,\xi} = |\langle n|\alpha, \xi\rangle|^2 \quad (13.86)$$

is a complicated and rapidly varying function of n . For instance, it is immediately clear from eq. (13.79) that

$$w_{n,0,\xi} = |\langle n|0, \xi\rangle|^2 \quad (13.87)$$

vanishes for odd n .

Squeezed states are oscillator eigenstates for translated oscillators with modified stiffness. Starting from the ground state of an unperturbed harmonic oscillator, they can be created by first adiabatically switching on and then suddenly switching off a perturbation term in the Hamiltonian, which is quadratic in a and a^+ or X_1 and X_2 . Quadratic interaction Hamiltonians should generally give rise to squeezed states. In practice, squeezed states are manufactured in a routine way by nonlinear optical devices that produce photon pairs.

We conclude this section with the important discussion of uncertainty relations for the quadratures X_1 and X_2 . Let us define

$$\langle X_i \rangle_\rho = \text{Tr } X_i \rho \quad (i = 1, 2) \quad (13.88)$$

to be the expectation values of the quadratures in any (pure or mixed) quantum state ρ . With

$$\Delta X_i = X_i - \langle X_i \rangle_\rho, \quad (13.89)$$

we can define the uncertainties of the quadratures by the variances

$$\sigma_{X_i\rho}^2 = \langle (\Delta X_i)^2 \rangle_\rho. \quad (13.90)$$

We have the uncertainty relation

$$\sigma_{X_1\rho}\sigma_{X_2\rho} \geq \frac{1}{2}. \quad (13.91)$$

For instance, for the oscillator states $|n\rangle$ we calculate

$$\langle X_1 \rangle_n = \left\langle n \left| \frac{a + a^\dagger}{\sqrt{2}} \right| n \right\rangle = \frac{1}{\sqrt{2}} \langle n | n - 1 \rangle \sqrt{n} + \frac{1}{\sqrt{2}} \langle n | n + 1 \rangle \sqrt{n + 1} = 0$$

and likewise for $\langle X_2 \rangle_n = 0$. Then

$$\begin{aligned} \sigma_{X_1 n}^2 &= \langle n | X_1^2 | n \rangle = \langle n | \frac{1}{2} (a + a^\dagger)^2 | n \rangle \\ &= \frac{1}{2} \langle n | [a^2 + (a^\dagger)^2 + aa^\dagger + a^\dagger a] | n \rangle = \frac{1}{2} \langle n | (1 + 2a^\dagger a) | n \rangle = n + \frac{1}{2}. \end{aligned}$$

Likewise

$$\begin{aligned} \sigma_{X_2 n} &= \sigma_{X_1 n} = \sqrt{n + \frac{1}{2}} \\ \sigma_{X_1 n} \sigma_{X_2 n} &= n + \frac{1}{2}. \end{aligned} \quad (13.92)$$

We see that the uncertainties are equal and that the uncertainty relation is saturated with an equality sign for the ground state $n = 0$.

This symmetric saturation also holds for coherent states, which, after all, are only translated ground states:

$$\begin{aligned} \sigma_{X_i \alpha}^2 &= \langle \alpha | (\Delta X_i)^2 | \alpha \rangle = \langle 0 | D^\dagger(\alpha) (\Delta X_i)^2 D(\alpha) | 0 \rangle \\ &= \langle 0 | X_i^2 | 0 \rangle = \sigma_{X_i 0} = 1/\sqrt{2}. \end{aligned} \quad (13.93)$$

Conversely, let us investigate under what circumstances the uncertainty relation (13.91) is saturated. Starting from any state with density matrix

$$\rho = \sum_i w_i |\psi_i\rangle \langle \psi_i|, \quad (13.94)$$

we form

$$\begin{aligned} 0 &\leq \left\langle \left(\frac{\Delta X_1}{\sigma_{X_1\rho}} - i \frac{\Delta X_2}{\sigma_{X_2\rho}} \right) \left(\frac{\Delta X_1}{\sigma_{X_1\rho}} + i \frac{\Delta X_2}{\sigma_{X_2\rho}} \right) \right\rangle_\rho \\ &= \sum_i w_i \left\langle \psi_i \left| \left(\frac{\Delta X_1}{\sigma_{X_1\rho}} - i \frac{\Delta X_2}{\sigma_{X_2\rho}} \right) \left(\frac{\Delta X_1}{\sigma_{X_1\rho}} + i \frac{\Delta X_2}{\sigma_{X_2\rho}} \right) \right| \psi_i \right\rangle \\ &= \left\langle \frac{(\Delta X_1)^2}{\sigma_{X_1\rho}^2} + \frac{(\Delta X_2)^2}{\sigma_{X_2\rho}^2} + i \left[\frac{\Delta X_1}{\sigma_{X_1\rho}}, \frac{\Delta X_2}{\sigma_{X_2\rho}} \right] \right\rangle_\rho \\ &= 2 - \frac{1}{\sigma_{X_1\rho} \sigma_{X_2\rho}}. \end{aligned}$$

Equality $\sigma_{X_1\rho}\sigma_{X_2\rho} = \frac{1}{2}$ can only hold if $\rho = |\psi\rangle\langle\psi|$ is pure and if $|\psi\rangle$ is the unique state satisfying

$$\left(\frac{\Delta X_1}{\sigma_{X_1\rho}} + 2i\Delta X_2\sigma_{X_1\rho} \right) |\psi\rangle = 0. \quad (13.95)$$

This means that $|\psi\rangle$ has to be the ground state with respect to the pair

$$a'' = \frac{1}{\sqrt{2}} \left(\frac{\Delta X_1}{\sigma_{X_1\rho}\sqrt{2}} + i\Delta X_2\sigma_{X_1\rho}\sqrt{2} \right),$$

$$a''^+ = \frac{1}{\sqrt{2}} \left(\frac{\Delta X_1}{\sigma_{X_1\rho}\sqrt{2}} - i\Delta X_2\sigma_{X_1\rho}\sqrt{2} \right),$$

or, equivalently, a coherent state with respect to the pair

$$a' = \frac{1}{\sqrt{2}} \left(\frac{X_1}{\sigma_{X_1\rho}\sqrt{2}} + iX_2\sigma_{X_1\rho}\sqrt{2} \right),$$

$$a'^+ = \frac{1}{\sqrt{2}} \left(\frac{X_1}{\sigma_{X_1\rho}\sqrt{2}} - iX_2\sigma_{X_1\rho}\sqrt{2} \right).$$

Summarizing our results, we see that the uncertainty relation for X_1 and X_2 is saturated if and only if ρ is a pure state which is either a coherent state or a squeezed state generated by a pure Bogolyubov transformation. One readily evaluates

$$\begin{aligned} \sigma_{X_{1,2}\alpha}^2 &= \langle \alpha, r | [X_{1,2} - \langle \alpha, r | X_{1,2} | \alpha, r \rangle]^2 | \alpha, r \rangle \\ &= \langle \alpha | B^+(r) [X_{1,2} - \langle \alpha, r | X_{1,2} | \alpha, r \rangle]^2 B(r) | \alpha \rangle \\ &= \langle \alpha | [X_{1,2} e^{\mp r} - \langle \alpha, r | X_{1,2} | \alpha, r \rangle]^2 | \alpha \rangle \\ &= \langle \alpha | [X_{1,2} e^{\mp r} - \langle \alpha | B^+(r) X_{1,2} B(r) | \alpha \rangle]^2 | \alpha \rangle \\ &= e^{\mp 2r} \langle \alpha | [X_{1,2} - \langle \alpha | X_{1,2} | \alpha \rangle]^2 | \alpha \rangle \\ &= e^{\mp 2r} \sigma_{X_{1,2}\alpha}^2 = \frac{1}{2} e^{\mp 2r}. \end{aligned}$$

Hence,

$$\sigma_{X_1\alpha r} = \frac{1}{\sqrt{2}} e^{-r} \quad \text{and} \quad \sigma_{X_2\alpha r} = \frac{1}{\sqrt{2}} e^r. \quad (13.96)$$

The saturation of the uncertainty relation for X_1 and X_2 is symmetric, with $\sigma_{X_1} = \sigma_{X_2} = 1/\sqrt{2}$, exactly for coherent states. Equation (13.96) justifies the name “squeezed states”.

For more general squeezed states with complex squeezing parameters, an additional rotation is involved. In this case, the uncertainty relation is not saturated for X_1 and X_2 , but for the rotated quadratures

$$X_1(\phi) = X_1 \cos \phi - X_2 \sin \phi \quad \text{and} \quad X_2(\phi) = X_1 \sin \phi + X_2 \cos \phi$$

with suitable rotation angle ϕ .

In summary, coherent and squeezed states are exactly those states which are as close as possible to classical states, because they have minimal uncertainties.

One can visualize the situation by a plot (fig. 13.1), in which for a given state ρ the expectation values $\langle X_{1,2} \rangle_\rho$ are plotted and the uncertainty is indicated by an uncertainty area of minimal measure $\geq \pi/4$. For coherent states the area is minimal and circular, for Bogolyubov-transformed states the area is minimal and elliptic with axes parallel to the main axes, and for a general squeezed state one has minimal rotated elliptical areas.

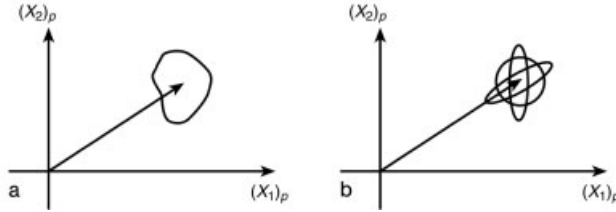


Figure 13.1: (a) Uncertainties for the quadratures in an arbitrary state. (b) Uncertainties for the quadratures in coherent and squeezed states.

One notices that the uncertainty relation for X_1 and X_2 gives rise to an uncertainty relation for the amplitude and the phase of $a = (1/\sqrt{2})(X_1 + iX_2)$ and, hence, of a light wave. By choosing a suitably squeezed state one can reduce the uncertainty of either amplitude or phase, taking into account an increased uncertainty of the conjugated quantity.

13.3 Operators, ordering procedures and star products

All operators acting on states of the electromagnetic field are functions of the creation and annihilation operators a_l^+ and a_l or, alternatively, of the quadrature operators X_{l1} and X_{l2} . Quantization is a procedure that associates operators $f(a, a^+)$ to classical functions $f(\alpha, \alpha^+)$. For functions like $f(\alpha, \alpha^*) = \alpha^2 \alpha^{*2}$ ordering problems arise. Should one define $f(a, a^+)$ by $a^2 a^{+2}$ or aa^+aa^+ or $\frac{1}{2}(a^2 a^{+2} + a^{+2} a^2)$ or otherwise? It is desirable to adopt an ordering procedure that associates to every function $f(\alpha, \alpha^*)$ a unique operator $(\text{Op } f)(a, a^+)$ such that

$$\text{Op}(c_1 f_1 + c_2 f_2) = c_1 \text{Op } f_1 + c_2 \text{Op } f_2 \quad \text{and} \quad \text{Op } 1 = 1. \quad (13.97)$$

Using a Fourier representation

$$f(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2 \beta e^{\beta \alpha^* - \beta^* \alpha} \tilde{f}(\beta, \beta^*), \quad (13.98)$$

or, with $\alpha = (1/\sqrt{2})(x_1 + ix_2)$ and $\beta = (1/\sqrt{2})(y_1 + iy_2)$,

$$f(x_1, x_2) = \frac{1}{2\pi} \int d^2 y e^{i(x_1 y_2 - x_2 y_1)} \tilde{f}(y_1, y_2), \quad (13.99)$$

we see that it suffices to define $\text{Op } f$ for the exponential functions

$$f_{\beta\beta^*}(\alpha, \alpha^*) := e^{(\beta\alpha^* - \beta^*\alpha)}, \quad f_{y_1 y_2}(x_1, x_2) := e^{i(x_1 y_2 - y_1 x_2)}. \quad (13.100)$$

For polynomial functions p , $\text{Op } p$ is then obtained by differentiating $\text{Op } f_{\beta\beta^*}$ with respect to β and β^* , and for more general functions f one can use eq. (13.97) to define $\text{Op } f$.

Let us now list five frequently used ordering procedures:

1. Normal ordering

$$\text{Op}_N f_{\beta\beta^*} = e^{\beta a^+} e^{-\beta^* a}. \quad (13.101)$$

All creation operators are placed to the left of all annihilation operators. For instance, $\text{Op}_N(\alpha^2 \alpha^{*2}) = a^{+2} a^2$.

The expectation value between coherent states is easily found to be

$$\langle \alpha | \text{Op}_N f | \alpha \rangle = f(\alpha, \alpha^*). \quad (13.102)$$

In particular,

$$\langle 0 | \text{Op}_N f | 0 \rangle = f(0, 0).$$

2. Anti-normal ordering

$$\text{Op}_A f_{\beta\beta^*} = e^{-\beta^* a} e^{\beta a^+}. \quad (13.103)$$

This time, all creation operators are placed to the right.

For this ordering, one finds from eq. (13.97) and the completeness relation (eq. (13.72)) for coherent states

$$\begin{aligned} \text{Op}_A f &= \frac{1}{\pi^2} \int d^2 \beta d^2 \alpha e^{-\beta^* a} |\alpha\rangle \langle \alpha| e^{\beta a^+} \tilde{f}(\beta, \beta^*) \\ &= \frac{1}{\pi^2} \int d^2 \beta d^2 \alpha e^{-\beta^* \alpha + \beta \alpha^*} \tilde{f}(\beta, \beta^*) |\alpha\rangle \langle \alpha| \\ &= \frac{1}{\pi} \int d^2 \alpha |\alpha\rangle f(\alpha, \alpha^*) \langle \alpha|. \end{aligned} \quad (13.104)$$

3. Standard ordering

$$\text{Op}_S f_{y_1 y_2} = e^{i y_2 X_1} e^{-i y_1 X_2}. \quad (13.105)$$

All operators X_2 are placed to the right of all operators X_1 . This procedure is most common if X_2 is the differential operator $X_2 = -i d/dx$ and X_1 is multiplication by x_1 . Then $\text{Op}_S f$ is a differential operator with all derivative operators placed to the right.

4. Anti-standard ordering

$$\text{Op}_{AS} f_{y_1 y_2} = e^{-i y_1 X_2} e^{+i y_2 X_1}. \quad (13.106)$$

5. Weyl ordering

$$\text{Op}_W f_{\beta\beta^*} = e^{\beta a^+ - \beta^* a} = D(\beta), \quad (13.107)$$

This is a very important ordering procedure, where $D(\beta)$ is the translation operator of eq. (13.43). Equivalently, we have

$$\text{Op}_W f_{y_1 y_2} = e^{i(y_2 X_1 - y_1 X_2)}. \quad (13.108)$$

The Weyl ordering corresponds to a complete symmetrization over all orderings in which polynomials in a and a^+ can be written. For example

$$\text{Op}_W(\alpha\alpha^*) = \frac{1}{2}(aa^+ + a^+a)$$

and

$$\text{Op}_W(\alpha^2\alpha^{*2}) = \frac{1}{6}(a^2a^{+2} + aa^+aa^+ + aa^{+2}a + a^+a^2a^+ + a^+aa^+a + a^{+2}a^2).$$

For real functions f , $\text{Op}_N f$, $\text{Op}_A f$, and $\text{Op}_W f$ are Hermitian. This is in general not true for $\text{Op}_S f$ and $\text{Op}_{AS} f$.

For $f \neq g$ we have $\text{Op}_X f \neq \text{Op}_X g$; hence Op_X has an inverse, and a non-commutative but associative so-called *star product* $*_X$ can be defined by

$$f *_X g = \text{Op}_X^{-1}(\text{Op}_X f \text{Op}_X g). \quad (13.109)$$

From eq. (13.97) we see that the star product is linear in its factors and that $1 *_X f = f *_X 1 = f$. Different ordering procedures give different but related star products. Indeed,

$$f *_X g = S_{XY}^{-1}((S_{XY} f) *_Y (S_{XY} g)), \quad (13.110)$$

where

$$S_{XY} f = \text{Op}_Y^{-1} \text{Op}_X(f).$$

Due to eq. (13.98), star products are already completely determined once we know $f_{\beta\beta^*} *_X f_{\gamma\gamma^*}$, with $f_{\beta\beta^*}$ and $f_{\gamma\gamma^*}$ given by eq. (13.100). For instance,

$$\begin{aligned} f_{\beta\beta^*} *_N f_{\gamma\gamma^*} &= \text{Op}_N^{-1}(e^{\beta a^+} e^{-\beta^* a} e^{\gamma a^+} e^{-\gamma^* a}) \\ &= \text{Op}_N^{-1}(e^{(\beta+\gamma)a^+} e^{-(\beta^*+\gamma^*)a} e^{-\beta^* \gamma}) \\ &= e^{-\beta^* \gamma} e^{(\beta+\gamma)\alpha^* - (\beta^*+\gamma^*)\alpha} \\ &= f_{\beta\beta^*} \exp\left(\left(\overleftarrow{\frac{\partial}{\partial\alpha}} \overrightarrow{\frac{\partial}{\partial\alpha^*}}\right) f_{\gamma\gamma^*}\right). \end{aligned} \quad (13.111)$$

Here the arrows over $\partial/\partial\alpha$ and $\partial/\partial\alpha^*$ indicate that $\partial/\partial\alpha^*$ acts to the right on $f_{\gamma\gamma^*}$ and $\partial/\partial\alpha$ acts to the left on $f_{\beta\beta^*}$. So, quite generally, we obtain

$$f *_N g = f \exp\left(\left(\overleftarrow{\frac{\partial}{\partial\alpha}} \overrightarrow{\frac{\partial}{\partial\alpha^*}}\right) g\right)$$

and, analogously,

$$\begin{aligned} f *_A g &= f \exp \left(-\frac{\overleftarrow{\partial}}{\partial \alpha^*} \frac{\overrightarrow{\partial}}{\partial \alpha} \right) g, \\ f *_S g &= f \exp \left(-i \frac{\overleftarrow{\partial}}{\partial x_2} \frac{\overrightarrow{\partial}}{\partial x_1} \right) g, \end{aligned} \quad (13.112)$$

$$\begin{aligned} f *_{AS} g &= f \exp \left(i \frac{\overleftarrow{\partial}}{\partial x_1} \frac{\overrightarrow{\partial}}{\partial x_2} \right) g, \\ f *_W g &= f \exp \left[\frac{1}{2} \left(\frac{\overleftarrow{\partial}}{\partial \alpha} \frac{\overrightarrow{\partial}}{\partial \alpha^*} - \frac{\overleftarrow{\partial}}{\partial \alpha^*} \frac{\overrightarrow{\partial}}{\partial \alpha} \right) \right] g \\ &= f \exp \left[\frac{i}{2} \left(\frac{\overleftarrow{\partial}}{\partial x_1} \frac{\overrightarrow{\partial}}{\partial x_2} - \frac{\overleftarrow{\partial}}{\partial x_2} \frac{\overrightarrow{\partial}}{\partial x_1} \right) \right] g. \end{aligned} \quad (13.113)$$

The Hermiticity property

$$(f *_X g)^* = g^* *_X f^* \quad (13.114)$$

holds for $X = N, A, W$, but not for $X = S, AS$.

Op_X and $*_X$ are invariant under phase space rotations (eq. (13.24)), for $X = N, A, W$, and invariant under Bogolyubov transformations for $X = S, AS, W$. Hence, also in this respect Op_W and $*_W$ represent the most symmetric case.

The non-commutativity of the star product comes from the non-commutativity of $\text{Op } f$ and $\text{Op } g$, which is regulated by Planck's constant $\hbar$. In our treatment we have scaled out $\hbar$ in order to have simple commutators $[a, a^+] = 1, [X_1, X_2] = i1$.

The $\hbar$ dependence can be visualized by replacing everywhere

$$a \rightarrow \frac{a}{\sqrt{\hbar}}, \quad a^+ \rightarrow \frac{a^+}{\sqrt{\hbar}}, \quad X_{1,2} \rightarrow \frac{X_{1,2}}{\sqrt{\hbar}}, \quad \alpha \rightarrow \frac{\alpha}{\sqrt{\hbar}}, \quad \alpha^* \rightarrow \frac{\alpha^*}{\sqrt{\hbar}}. \quad (13.115)$$

The $\hbar$ will reappear in our formulas. The star product $*_W$ will look like

$$f *_W g = f \exp \left[\frac{i\hbar}{2} \left(\frac{\overleftarrow{\partial}}{\partial x_1} \frac{\overrightarrow{\partial}}{\partial x_2} - \frac{\overleftarrow{\partial}}{\partial x_2} \frac{\overrightarrow{\partial}}{\partial x_1} \right) \right] g. \quad (13.116)$$

In lowest non-trivial order in $\hbar$, the commutator will be given by the Poisson bracket

$$f *_W g - g *_W f = i\hbar \{f, g\} + O(\hbar^2), \quad (13.117)$$

where $O(\hbar^2)$ indicates terms of higher order in $\hbar$. The same property is true for the other star products:

$$f *_X g - g *_X f = i\hbar \{f, g\} + O(\hbar^2). \quad (13.118)$$

For an operator A that is already ordered according to the ordering procedure X , one immediately reads off the inverse $f_A = \text{Op}_X^{-1} A$. A problem arises if A is not properly ordered.

In general, $\text{Op}_X^{-1} A$ can be obtained in the following way. The starting point is the fundamental identity:

$$\frac{1}{\pi} \text{Tr}(D^+(\beta)D(\alpha)) = \delta^{(2)}(\alpha - \beta) \quad (13.119)$$

or

$$\frac{1}{2\pi} \text{Tr}(e^{-i(X_1 y_2 - X_2 y_1)} e^{i(X_1 x_2 - X_2 x_1)}) = \delta(x_1 - y_1)\delta(x_2 - y_2), \quad (13.120)$$

which can be verified by direct calculation with eq. (13.76). Then, every operator A can be represented in the form

$$\begin{aligned} A &= \frac{1}{\pi} \int d^2\beta D(\beta) \text{Tr}(D^+(\beta)A) \\ &= \frac{1}{\pi} \int d^2\beta e^{\beta a^+ - \beta^* a} \text{Tr}(e^{-\beta a^+ + \beta^* a} A) \end{aligned} \quad (\text{Weyl}). \quad (13.121)$$

This gives A in Weyl ordered form. Reordering $e^{\beta a^+ - \beta^* a}$ we also find

$$\begin{aligned} A &= \frac{1}{\pi} \int d^2\beta e^{+\beta a^+} e^{-\beta^* a} \text{Tr}(e^{+\beta^* a} e^{-\beta a^+} A) \quad (\text{normal}) \\ &= \frac{1}{\pi} \int d^2\beta e^{-\beta^* a} e^{\beta a^+} \text{Tr}(e^{-\beta a^+} e^{\beta^* a} A) \quad (\text{anti-normal}) \\ &= \frac{1}{2\pi} \int d^2y e^{i(X_1 y_2 - X_2 y_1)} \text{Tr}(e^{-i(X_1 y_2 - X_2 y_1)} A) \\ &= \frac{1}{2\pi} \int d^2y e^{iX_1 y_2} e^{-iX_2 y_1} \text{Tr}(e^{+iX_2 y_1} e^{-iX_1 y_2} A) \quad (\text{standard}) \\ &= \frac{1}{2\pi} \int d^2y e^{-iX_2 y_1} e^{iX_1 y_2} \text{Tr}(e^{-iX_1 y_2} e^{iX_2 y_1} A) \quad (\text{anti-standard}). \end{aligned} \quad (13.122)$$

From eqs. (13.121) and (13.122) we immediately read off the inverses $\text{Op}_X^{-1}(A)$:

$$\begin{aligned} \text{Op}_W^{-1}(A) &= W_A(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2\beta e^{\beta \alpha^* - \beta^* \alpha} \text{Tr}(e^{-\beta a^+ + \beta^* a} A) \\ &= W_A(x_1, x_2) = \frac{1}{2\pi} \int d^2y e^{i(x_1 y_2 - x_2 y_1)} \text{Tr}(e^{-i(X_1 y_2 - X_2 y_1)} A) \\ \text{Op}_N^{-1}(A) &= Q_A(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2\beta e^{\beta \alpha^* - \beta^* \alpha} \text{Tr}(e^{\beta^* a} e^{-\beta a^+} A) \\ \text{Op}_A^{-1}(A) &= P_A(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2\beta e^{\beta \alpha^* - \beta^* \alpha} \text{Tr}(e^{-\beta a^+} e^{\beta^* a} A) \\ \text{Op}_S^{-1}(A) &= q_A(x_1, x_2) = \frac{1}{2\pi} \int d^2y e^{i(x_1 y_2 - x_2 y_1)} \text{Tr}(e^{iX_2 y_1} e^{-iX_1 y_2} A) \\ \text{Op}_{AS}^{-1}(A) &= p_A(x_1, x_2) = \frac{1}{2\pi} \int d^2y e^{i(x_1 y_2 - x_2 y_1)} \text{Tr}(e^{-iX_1 y_2} e^{iX_2 y_1} A). \end{aligned} \quad (13.123)$$

In the above equations, W_A is called the *Wigner function* of the operator A . It plays a fundamental role in quantum optics and quantum mechanics. Almost equally important are the

so-called P and Q functions P_A and Q_A . From eqs. (13.102), (13.104), and (13.122) we see

$$Q_A(\alpha, \alpha^*) = \langle \alpha | A | \alpha \rangle \quad (13.124)$$

and

$$A = \frac{1}{\pi} \int d^2\alpha |\alpha\rangle P_A \langle \alpha|. \quad (13.125)$$

For positive A , Q_A is also positive. This is in general not true for the other functions in eq. (13.123). Equation (13.124) reflects the remarkable fact that, due to the over-completeness of coherent states, an operator A is already determined by its diagonal elements $\langle \alpha | A | \alpha \rangle$.

By definition of the star product, eq. (13.109), $\text{Op}_X^{-1}(AB)$ is given by star products:

$$\begin{aligned} W_{AB} &= W_A *_{\text{W}} W_B, & Q_{AB} &= Q_A *_{\text{N}} Q_B, \\ P_{AB} &= P_A *_{\text{A}} P_B, & p_{AB} &= p_A *_{\text{AS}} p_B, & q_{AB} &= q_A *_{\text{A}} q_B. \end{aligned} \quad (13.126)$$

[The symbol σ_H of a differential operator H introduced at the beginning of section 9.1 and in section 10.10 should just be identified with $q_S = Q_S^{-1}(H) = \sigma_H$, and we have $\sigma_{AB} = \sigma_A *_{\text{S}} \sigma_B$, and $\sigma_{[A,B]} = (1/i\hbar)\{\sigma_A, \sigma_B\}$ corresponds to eq. (13.118).]

Reordering the exponentials under the traces in eq. (13.123), we can observe how the different functions W_A , P_A , Q_A , p_A , and q_A are related to each other:

$$\widetilde{W}(\beta, \beta^*) = e^{\frac{1}{2}|\beta|^2} \widetilde{Q}(\beta, \beta^*), \quad \widetilde{P}_A(\beta, \beta^*) = e^{|\beta|^2} \widetilde{Q}(\beta, \beta^*), \quad (13.127)$$

and, hence,

$$Q_A = e^{\frac{1}{2}S} W_A, \quad P_A = e^{-\frac{1}{2}S} W_A, \quad (13.128)$$

with the operator $S = \partial^2 / \partial \alpha \partial \alpha^*$. Similarly,

$$q_A = \exp\left(-\frac{i}{2} \frac{\partial^2}{\partial x_1 \partial x_2}\right) W_A, \quad p_A = \exp\left(\frac{i}{2} \frac{\partial^2}{\partial x_1 \partial x_2}\right) W_A. \quad (13.129)$$

From this we derive interesting and important identities for the trace of operators:

$$\begin{aligned} \text{Tr } A &= \frac{1}{\pi} \int d^2\alpha \langle \alpha | A | \alpha \rangle = \frac{1}{\pi} \int d^2\alpha Q_A(\alpha, \alpha^*) \\ &= \frac{1}{\pi} \int d^2\alpha e^{\frac{1}{2}S} W_A(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2\alpha e^S P_A(\alpha, \alpha^*) \\ &= \frac{1}{\pi} \int d^2\alpha W_A(\alpha, \alpha^*) = \frac{1}{\pi} \int d^2\alpha P_A(\alpha, \alpha^*), \end{aligned} \quad (13.130)$$

where the last line follows because all the derivatives in S integrate out. Likewise,

$$\begin{aligned} \text{Tr } A &= \frac{1}{2\pi} \int d^2x W_A(x_1, x_2) \\ &= \frac{1}{2\pi} \int d^2x p_A(x_1, x_2) = \frac{1}{2\pi} \int d^2x q_A(x_1, x_2). \end{aligned} \quad (13.131)$$

Moreover,

$$\mathrm{Tr} AB = \frac{1}{\pi} \int d^2\alpha W_{AB} = \frac{1}{\pi} \int d^2\alpha W_A *_W W_B = \frac{1}{\pi} \int d^2\alpha W_A W_B, \quad (13.132)$$

because the derivatives in the star product integrate out, and

$$\begin{aligned} \mathrm{Tr} AB &= \frac{1}{\pi} \int d^2\alpha W_A W_B = \frac{1}{\pi} \int d^2\alpha e^{-\frac{1}{2}S} Q_A e^{-\frac{1}{2}S} Q_B \\ &= \frac{1}{\pi} \int d^2\alpha (e^{-S} Q_A) Q_B = \frac{1}{\pi} \int d^2\alpha P_A Q_B = \frac{1}{\pi} \int d^2\alpha Q_A P_B. \end{aligned} \quad (13.133)$$

Similarly,

$$\mathrm{Tr} AB = \frac{1}{2\pi} \int d^2x W_A W_B = \frac{1}{2\pi} \int d^2x q_A p_B = \frac{1}{2\pi} \int d^2x p_A q_B. \quad (13.134)$$

Equations (13.133) and (13.134) also follow from eqs. (13.121) and (13.122).

Far beyond quantum optics, the theory of star products is a rapidly developing branch of mathematical physics and differential geometry. We defined star products on the simplest phase spaces $P = \mathbb{R}^2$ and $P = \mathbb{R}^{2n}$. More generally, star products can be defined on so-called *Poisson manifolds* P , whose defining property is the existence of a bilinear Poisson bracket $\{f, g\}$ for any smooth complex-valued functions $f, g : P \rightarrow \mathbb{C}$ on P : $f, g \in C^\infty(P)$, with the usual properties of Poisson brackets:

$$\begin{aligned} \text{(i)} \quad & \{f, g\} + \{g, f\} = 0, \\ \text{(ii)} \quad & \{f, gh\} = \{f, g\}h + g\{f, h\}, \\ \text{(iii)} \quad & \{f, \{g, h\}\} + \{g, \{h, f\}\} + \{h, \{f, g\}\} = 0. \end{aligned} \quad (13.135)$$

Symplectic manifolds as defined in chapter 9 are special Poisson manifolds.

In order to avoid convergence problems in expressions like eq. (13.116), one goes over to formal power series in an indeterminate λ (eventually meant to be replaced by $\hbar$). Let

$$f(p) = \sum_{n=0}^{\infty} f_n(p) \lambda^n \quad (13.136)$$

be a formal power series in λ with coefficients in $C^\infty(P)$. We denote the set of all such formal power series by $C^\infty(P)[[\lambda]]$. Then a star product on a Poisson manifold P is a (bilinear) associative product

$$\begin{aligned} * : C^\infty(P)[[\lambda]] \times C^\infty(P)[[\lambda]] &\rightarrow C^\infty(P)[[\lambda]] \\ (f, g) &\rightarrow f * g = \sum_{n=0}^{\infty} \lambda^n M_n(f, g), \end{aligned} \quad (13.137)$$

with the following properties:

$$\begin{aligned} \text{(i)} \quad & M_0(f, g) = fg, \quad \text{i.e.} \quad f * g = fg[1 + O(\lambda)], \\ \text{(ii)} \quad & M_1(f, g) - M_1(g, f) = i\{f, g\}, \quad \text{i.e.} \quad f * g - g * f = i\lambda\{f, g\}[1 + O(\lambda)], \\ \text{(iii)} \quad & 1 * f = f * 1 = f. \end{aligned} \quad (13.138)$$

Normally, also the following locality property for the supports of f , g and $f * g$ is postulated:

$$(iv) \quad \text{supp } f * g \subset \text{supp } f \cap \text{supp } g. \quad (13.139)$$

A star product is called *Hermitian* if

$$(v) \quad \overline{f * g} = \bar{g} * \bar{f}. \quad (13.140)$$

The star products defined in this section are all local, and $*_{\text{W}}$, $*_{\text{N}}$ and $*_{\text{AS}}$ are Hermitian.

13.4 The Q , P , and Wigner functions of a density operator

Let the state of the system be described by a density operator

$$\rho = \sum_i w_i |\psi_i\rangle \langle \psi_i|, \quad \rho \geq 0, \quad \text{Tr } \rho = 1. \quad (13.141)$$

The functions $Q_\rho(\alpha, \alpha^*)$, $P_\rho(\alpha, \alpha^*)$, and $W_\rho(\alpha, \alpha^*)$ are often simply called the Q , P , and Wigner functions of the system.

For a suitably ordered operator $A = \text{Op}_X f$, the expectation value

$$\langle A \rangle_\rho = \text{Tr } \rho A \quad (13.142)$$

can be simply expressed by

$$\begin{aligned} \langle \text{Op}_N f \rangle_\rho &= \frac{1}{\pi} \int d^2 \alpha f(\alpha, \alpha^*) P_\rho(\alpha, \alpha^*), \\ \langle \text{Op}_A f \rangle_\rho &= \frac{1}{\pi} \int d^2 \alpha f(\alpha, \alpha^*) Q_\rho(\alpha, \alpha^*), \\ \langle \text{Op}_W f \rangle_\rho &= \frac{1}{\pi} \int d^2 \alpha f(\alpha, \alpha^*) W_\rho(\alpha, \alpha^*). \end{aligned} \quad (13.143)$$

These equations follow immediately from eqs. (13.132) and (13.133). For $Q_\rho(\alpha, \alpha^*)$ we have (eq. 13.124)

$$\begin{aligned} Q_\rho(\alpha, \alpha^*) &= \langle \alpha | \rho | \alpha \rangle \geq 0, \\ \frac{1}{\pi} \int d^2 \alpha Q_\rho(\alpha, \alpha^*) &= 1, \\ Q_\rho(\alpha, \alpha^*) &= \sum_i w_i \langle \alpha | \psi_i \rangle \langle \psi_i | \alpha \rangle \leq 1. \end{aligned} \quad (13.144)$$

We have already mentioned that positivity will in general only hold for Q_ρ , but not for P_ρ and W_ρ . We shall soon see that $|W_\rho|$ is bounded but not $|P_\rho|$. Let us calculate the functions Q_ρ , P_ρ , and W_ρ for some typical and important density operators. We start with evaluating Q_ρ .

- *Pure coherent state.* From eq. (13.68) we immediately get (compare eq. (13.124))

$$Q_{|\gamma\rangle}(\alpha, \alpha^*) = \langle \alpha | \gamma \rangle \langle \gamma | \alpha \rangle = e^{-|\alpha - \gamma|^2}. \quad (13.145)$$

The result is a Gaussian function.

Also for a squeezed state we would have found a Gaussian, but this time with different variances for α_1 and α_2 .

- *Pure number state.* From eq. (13.71) or directly from eq. (13.70) we obtain for a pure number state

$$\begin{aligned} Q_{|n\rangle}(\alpha, \alpha^*) &= \frac{1}{n!} e^{-|\gamma|^2} \left(\frac{\partial^2}{\partial \gamma \partial \gamma^*} \right)^n e^{|\gamma|^2} Q_{|\gamma\rangle}(\alpha, \alpha^*) \Big|_{\gamma=\gamma^*=0} \\ &= \frac{|\alpha|^{2n}}{n!} e^{-|\alpha|^2}. \end{aligned} \quad (13.146)$$

- *Thermal state.* A thermal state is described by the density matrix

$$\rho_{\text{th}} = \frac{1}{Z} e^{-\tau a^\dagger a} = \sum_n w_n |n\rangle \langle n| \quad (13.147)$$

with

$$\tau = \frac{\hbar\omega}{kT}, \quad Z = \text{Tr} e^{-\tau a^\dagger a} = \frac{1}{1 - e^{-\tau}}, \quad w_n = \frac{e^{-\tau n}}{Z}.$$

The expectation value $\langle N \rangle = \text{Tr} a^\dagger a \rho_{\text{th}}$ is given by

$$\langle N \rangle = -\frac{\partial}{\partial \tau} \ln Z = \frac{e^{-\tau}}{1 - e^{-\tau}}. \quad (13.148)$$

Expressing $e^{-\tau}$ by $\langle N \rangle$ we can cast ρ_{th} into the form

$$\rho_{\text{th}} = \frac{1}{\langle N \rangle + 1} \sum_{n=0}^{\infty} \left(\frac{\langle N \rangle}{\langle N \rangle + 1} \right)^n |n\rangle \langle n|.$$

Now,

$$\begin{aligned} Q_{\text{th}}(\alpha, \alpha^*) &= \langle \alpha | \rho_{\text{th}} | \alpha \rangle = \sum_n w_n Q_{|n\rangle}(\alpha, \alpha^*) \\ &= \frac{1}{\langle N \rangle + 1} \exp \left(-\frac{|\alpha|^2}{\langle N \rangle + 1} \right). \end{aligned} \quad (13.149)$$

It is now an easy task to compute also W_ρ and P_ρ for the same set of density states. From eq. (13.127) we see that we can start out from the Fourier transform $\tilde{Q}_\rho(\beta, \beta^*)$, multiply by $e^{|\beta|^2/2}$ or $e^{|\beta|^2}$ and then calculate the inverse Fourier transform. The result is as follows:

- *Coherent state.*

$$W_{|\gamma\rangle}(\alpha, \alpha^*) = 2 e^{-2|\alpha - \gamma|^2}, \quad P_{|\gamma\rangle}(\alpha, \alpha^*) = \pi \delta^{(2)}(\alpha - \gamma). \quad (13.150)$$

- *Number state.*

$$\begin{aligned} W_{|n\rangle}(\alpha, \alpha^*) &= 2 \frac{e^{-|\gamma|^2}}{n!} \left(\frac{\partial^2}{\partial \gamma \partial \gamma^*} \right)^n e^{+|\gamma|^2} e^{-2|\alpha-\gamma|^2} \Big|_{\gamma=\gamma^*=0}, \\ P_{|n\rangle}(\alpha, \alpha^*) &= \pi \frac{e^{-|\gamma|^2}}{n!} \left(\frac{\partial^2}{\partial \gamma \partial \gamma^*} \right)^n e^{+|\gamma|^2} \delta^{(2)}(\alpha - \gamma) \Big|_{\gamma=\gamma^*=0}. \end{aligned} \quad (13.151)$$

- *Thermal state.*

$$\begin{aligned} W_{\text{th}}(\alpha, \alpha^*) &= \frac{1}{\langle N \rangle + \frac{1}{2}} \exp\left(-\frac{|\alpha|^2}{\langle N \rangle + \frac{1}{2}}\right), \\ P_{\text{th}}(\alpha, \alpha^*) &= \frac{1}{\langle N \rangle} \exp\left(-\frac{|\alpha|^2}{\langle N \rangle}\right). \end{aligned} \quad (13.152)$$

One notices that the distribution W_ρ is narrower than Q_ρ , and that P_ρ is narrower than W_ρ .

Let us visualize the dependence on $\hbar$ by introducing dimensional quantities α , α^* , and $X_{1,2}$ according to eq. (13.115). We define dimensional Q , P , and W functions by

$$\begin{aligned} Q_\rho^{(\hbar)}(\alpha, \alpha^*) &= \frac{1}{\hbar} Q\left(\frac{\alpha}{\sqrt{\hbar}}, \frac{\alpha^*}{\sqrt{\hbar}}\right), \\ W_\rho^{(\hbar)}(\alpha, \alpha^*) &= \frac{1}{\hbar} W\left(\frac{\alpha}{\sqrt{\hbar}}, \frac{\alpha^*}{\sqrt{\hbar}}\right), \\ P_\rho^{(\hbar)}(\alpha, \alpha^*) &= \frac{1}{\hbar} P\left(\frac{\alpha}{\sqrt{\hbar}}, \frac{\alpha^*}{\sqrt{\hbar}}\right). \end{aligned} \quad (13.153)$$

These functions are normalized to

$$\frac{1}{\pi} \int d^2\alpha Q_\rho^{(\hbar)} = \frac{1}{\pi} \int d^2\alpha W_\rho^{(\hbar)} = \frac{1}{\pi} \int d^2\alpha P_\rho^{(\hbar)} = 1, \quad (13.154)$$

and eq. (13.144) turns into

$$0 \leq Q_\rho^{(\hbar)} \leq 1/\hbar. \quad (13.155)$$

For the Wigner function, we can use its invariance under Bogolyubov transformations and employ also non-symmetric rescalings:

$$x_1 \rightarrow \kappa x_1, \quad x_2 \rightarrow x_2/\kappa\hbar. \quad (13.156)$$

The state ρ is said to have a classical limit ρ_{cl} if

$$\lim_{\hbar \rightarrow 0} W_\rho^{(\hbar)} = \lim_{\hbar \rightarrow 0} Q_\rho^{(\hbar)} = \lim_{\hbar \rightarrow 0} P_\rho^{(\hbar)} = \pi \rho_{\text{cl}} \quad (13.157)$$

exists and if ρ_{cl} is a classical state, i.e. a non-negative distribution on phase space.

A glance at eqs. (13.145)–(13.152) reveals that $\rho_\gamma = |\gamma\rangle\langle\gamma|$ has the classical pure state

$$\rho_{\gamma \text{cl}}(\alpha, \alpha^*) = \delta^{(2)}(\alpha - \gamma) \quad (13.158)$$

as a limit. Also ρ_{th} converges to the classical canonical distribution,

$$\rho_{\text{th cl}} = \frac{\omega}{\pi kT} \exp\left(-\frac{\omega|\alpha|^2}{kT}\right), \quad (13.159)$$

whereas, not surprisingly, $\rho_n = |n\rangle\langle n|$ has no classical limit.

If the classical limit exists, the limit of the expectation values is given by the classical expectation value:

$$\lim_{\hbar \rightarrow 0} \{\text{Tr}(\rho \text{Op}_X f)\} = \int d^2\alpha \rho_{\text{cl}}(\alpha, \alpha^*) f(\alpha, \alpha^*). \quad (13.160)$$

Negative values of W_ρ and P_ρ are a quantum feature of these functions and will disappear in the classical limit.

We now discuss the properties of the Wigner function $W_\rho(x_1, x_2)$ in some more detail, this time in terms of the quadrature coordinates x_1 and x_2 . Integrating

$$W_\rho(x_1, x_2) = \frac{1}{2\pi} \int d^2y e^{i(x_1 y_2 - x_2 y_1)} \text{Tr}(e^{-i(X_1 y_2 - X_2 y_1)} \rho) \quad (13.161)$$

over x_2 we obtain a δ -function:

$$\begin{aligned} \int dx_2 W_\rho(x_1, x_2) &= \int d^2y e^{ix_1 y_2} \delta(y_1) \text{Tr}(e^{-i(X_1 y_2 - X_2 y_1)} \rho) \\ &= \int dy_2 e^{ix_1 y_2} \text{Tr}(e^{-iX_1 y_2} \rho). \end{aligned} \quad (13.162)$$

Using eigenstates $|x'_1\rangle$ of X_1 we can write

$$\text{Tr}(e^{-iX_1 y_2} \rho) = \int dx'_1 \langle x'_1 | e^{-iX_1 y_2} \rho | x'_1 \rangle \quad (13.163)$$

and rewrite eq. (13.162) as

$$\begin{aligned} \int dx_2 W_\rho(x_1, x_2) &= \int dy_2 \int dx'_1 e^{i(x_1 - x'_1)y_2} \langle x'_1 | \rho | x'_1 \rangle \\ &= 2\pi \int dx'_1 \delta(x_1 - x'_1) \langle x'_1 | \rho | x'_1 \rangle \\ &= 2\pi \langle x_1 | \rho | x_1 \rangle \geq 0. \end{aligned}$$

The result is just the probability distribution for the measured values of the quadrature X_1 , up to a factor 2π . In the same way,

$$\int dx_1 W_\rho(x_1, x_2) = 2\pi \langle x_2 | \rho | x_2 \rangle \geq 0. \quad (13.164)$$

Although $W_\rho(x_1, x_2)$ is not a proper positive distribution function, its marginals are so.

The Wigner function is invariant under all squeezing transformations. For instance, we could have based all our calculations on rotated quadratures. Thus, the probability distribution for the rotated quadrature X_θ is obtained by integrating out the complementary variable:

$$2\pi \langle x_\theta | \rho | x_\theta \rangle = \int d^2x \delta(x_\theta - x_1 \cos \theta + x_2 \sin \theta) W(x_1, x_2). \quad (13.165)$$

Reordering $e^{-i(X_1 y_2 - X_2 y_1)}$ in eq. (13.161), we derive another frequently used form of the Wigner function:

$$\begin{aligned} W(x_1, x_2) &= \frac{1}{2\pi} \int d^2 y e^{i(x_1 y_2 - x_2 y_1 - \frac{1}{2} y_1 y_2)} \text{Tr}(e^{-iX_1 y_2} e^{iX_2 y_1} \rho) \\ &= \frac{1}{2\pi} \int d^2 y \int dx'_1 e^{i(x_1 y_2 - x_2 y_1 - \frac{1}{2} y_1 y_2)} \langle x'_1 | e^{-iX_1 y_2} e^{iX_2 y_1} \rho | x'_1 \rangle \\ &= \frac{1}{2\pi} \int d^2 y \int dx'_1 e^{i(x_1 y_2 - x_2 y_1 - \frac{1}{2} y_1 y_2 - x'_1 y_2)} \langle x'_1 | e^{iX_2 y_1} \rho | x'_1 \rangle. \end{aligned} \quad (13.166)$$

Observing that $e^{iX_2 y_1}$ is a translation operator for X_1 , we can continue:

$$\begin{aligned} W(x_1, x_2) &= \frac{1}{2\pi} \int d^2 y \int dx'_1 e^{i(x_1 y_2 - x_2 y_1 - \frac{1}{2} y_1 y_2 - x'_1 y_2)} \langle x'_1 + y_1 | \rho | x'_1 \rangle \\ &= \int dy_1 \int dx'_1 \delta(x_1 - x'_1 - \frac{1}{2} y_1) e^{-ix_2 y_1} \langle x'_1 + y_1 | \rho | x'_1 \rangle \\ &= \int dy_1 e^{-ix_2 y_1} \langle x_1 + \frac{1}{2} y_1 | \rho | x_1 - \frac{1}{2} y_1 \rangle. \end{aligned} \quad (13.167)$$

Notice how the quantum uncertainty relation between X_1 and X_2 reflects itself in a slight smearing of $\langle x_1 | \rho | x_1 \rangle$ around the diagonal. From eq. (13.167) we can derive a bound for $W(x_1, x_2)$ using Schwartz's inequality:

$$\begin{aligned} |W(x_1, x_2)| &= \left| \sum_i \int dy_1 e^{-ix_2 y_1} w_i \langle x_1 + \frac{1}{2} y_1 | \psi_i \rangle \langle \psi_i | x_1 - \frac{1}{2} y_1 \rangle \right| \\ &\leq \sum_i w_i \left(2 \int \frac{dy_1}{2} \left| \left\langle x_1 + \frac{y_1}{2} \middle| \psi_i \right\rangle \right|^2 \right)^{1/2} \\ &\quad \times \left(2 \int \frac{dy_1}{2} \left| \left\langle \psi_i \middle| x_1 - \frac{y_1}{2} \right\rangle \right|^2 \right)^{1/2} \\ &\leq 2. \end{aligned} \quad (13.168)$$

So,

$$|W(x_1, x_2)| \leq 2 \quad \text{and} \quad |W^{(\hbar)}(x_1, x_2)| \leq 2/\hbar. \quad (13.169)$$

The equation of motion for the density operator is the quantum Liouville equation,

$$i\hbar \frac{d}{dt} \rho = [\hat{H}, \rho], \quad (13.170)$$

where

$$\hat{H} = \text{Op}_W H \quad (13.171)$$

is the Hamilton operator. For the Wigner function $W_\rho^{(\hbar)}$, eq. (13.170) means simply

$$\frac{d}{dt} W_\rho^{(\hbar)} = \frac{1}{i\hbar} (H *_W W_\rho^{(\hbar)} - W_\rho^{(\hbar)} *_W H). \quad (13.172)$$

In lowest order in $\hbar$ this goes over into the classical Liouville equation,

$$\frac{d}{dt}\rho_{\text{cl}} = \{H, \rho_{\text{cl}}\},$$

with the Poisson bracket on the right-hand side (compare eq. (13.118)).

For a stationary state ρ we have

$$[\hat{H}, \rho] = 0, \quad H *_{\text{W}} W_{\rho}^{(\hbar)} - W_{\rho}^{(\hbar)} *_{\text{W}} H = 0, \quad \{H, \rho_{\text{cl}}\} = 0. \quad (13.173)$$

This equation can be solved order by order in $\hbar$ using the short-wave asymptotic methods of chapters 7–10, where the parameter k controlling the short-wave limit has to be replaced by $1/\hbar$.

The star product formalism yields considerable simplifications. For an energy eigenstate $\rho = |E\rangle\langle E|$, with $\hat{H}|E\rangle = E|E\rangle$, eq. (13.173) simplifies to

$$\hat{H}\rho = E\rho, \quad H *_{\text{W}} W_{\rho}^{(\hbar)} = EW_{\rho}^{(\hbar)}. \quad (13.174)$$

For a pure stationary state, the short-wave ansatz eq. (7.1) corresponds to a WKB ansatz for the Wigner function:

$$W_{\rho}^{(\hbar)} = E^{(i/\hbar)S} *_{\text{W}} C *_{\text{W}} E^{-(i/\hbar)S}, \quad (13.175)$$

where

$$\begin{aligned} E^{(i/\hbar)S} &= 1 + \frac{i}{\hbar}S + \frac{1}{2}\left(\frac{i}{\hbar}\right)^2 S *_{\text{W}} S + \frac{1}{6}\left(\frac{i}{\hbar}\right)^3 S *_{\text{W}} S *_{\text{W}} S + \dots \\ &= \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{i}{\hbar}\right)^n S^{*_{\text{W}} n} \end{aligned} \quad (13.176)$$

is the star-exponential function, which coincides with the normal exponential function if S depends only on x_1 .

In lowest order, $\{H; \rho_{\text{cl}}\} = 0$ means that ρ_{cl} has to be constant along the classical solutions of the equation of motion. If, in addition, the energy has a sharp value, ρ_{cl} is concentrated on the orbits of energy E . For higher orders in $\hbar$, this concentration will be somewhat softened, and W_{ρ} will have appreciable values only in some small neighborhood of the region $\{(x_1, x_2), H(x_1, x_2) = E\}$ in classical phase space, where the possible values of E are given by the Bohr–Sommerfeld quantization condition.

If the energy does not have a sharp value but is concentrated in a neighborhood of E , then, again, W_{ρ} will show the same concentration about the energy level surface corresponding to the energy E .

For the canonical thermal density operator $\rho = (1/Z) \exp(-\hat{H}/kT)$, we have, of course,

$$W_{\rho} = \frac{1}{Z} E^{-H/kT} \quad \text{with} \quad Z = \frac{1}{2\pi} \int d^2x E^{-H/kT}. \quad (13.177)$$

The scalar product of two states $|\varphi\rangle$ and $|\psi\rangle$ is given by

$$|\langle\varphi|\psi\rangle|^2 = \text{Tr}(\rho_{\varphi}\rho_{\psi}) = \int d^2x W_{|\varphi\rangle} W_{|\psi\rangle}. \quad (13.178)$$

The integrand is often concentrated on a small overlap region of the supports of $W_{|\varphi\rangle}^{(\hbar)}$ and $W_{|\psi\rangle}^{(\hbar)}$ and the integral can be evaluated by the method of stationary phases described in chapter 8. For details, see the comprehensive and lucid presentation of Schleich, mentioned in the literature list.

14 Detection of radiation fields

14.1 Beam splitters and homodyne detection

A beam splitter is a linear optical device by which an incident beam of light is partially transmitted and partially reflected (see fig. 14.1).

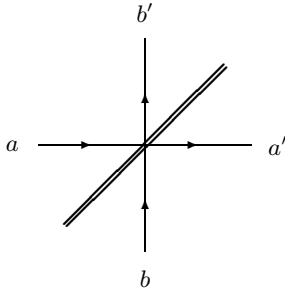


Figure 14.1: Principle of a beam splitter.

Usually, one considers two incident beams directed in such a way that their reflected and transmitted rays can interfere. In the simplest and most important case, all four beams involved are unimodal, i.e. each of them contains only one mode of the radiation field. (In the following we shall assume that all of them have the same frequency.)

The beam splitter will now realize a linear transformation between the annihilation operators a and b of the incident modes and the corresponding operators a' and b' of the outgoing modes:

$$a' = t_1 a + r_1 b, \quad b' = t_2 b + r_2 a, \quad (14.1)$$

and, hence,

$$a'^+ = t_1^* a^+ + r_1^* b^+, \quad b'^+ = t_2^* b^+ + r_2^* a. \quad (14.2)$$

There are no Bogolyubov transformations involved, because we have assumed that the beam splitter contains no nonlinear optical components.

More generally, we can consider linear transformations

$$\begin{aligned} a'_i &= \sum_{r=1}^{M_1} a_r C_{ri}, \\ a_j^{'+} &= \sum_{s=1}^{M_1} a_s^+ C_{sj}^* = \sum_{s=1}^{M_1} C_{js}^+ a_s^+ \quad (i, j = 1, \dots, M_2). \end{aligned} \quad (14.3)$$

The commutation relation $[a'_i, a_j^{'+}] = \delta_{ij}$ leads to

$$(C^+ C)_{ij} = \delta_{ij}, \quad (14.4)$$

whereas energy conservation or, equivalently, photon number conservation,

$$\sum_{r=1}^{M_1} a_r^+ a_r = \sum_{i=1}^{M_2} a_i^{'+} a'_i,$$

gives

$$(C C^+)_{rs} = \delta_{rs}. \quad (14.5)$$

Altogether this implies that C has to be a unitary square matrix and $M_1 = M_2$. Matrix C can be implemented by a unitary transformation U of the Hilbert space $U^+ a_r U = \sum_i a_i C_{ir}$, and U leaves the vacuum invariant.

For the beam splitter, we must have $M_1 = M_2 = 2$ and we see that, even if there is only one incident beam, we cannot neglect the second mode b in quantum optics. If only one beam is incident, the second incoming state is the vacuum state of the mode b .

For the beam splitter, “unitarity” means

$$\begin{aligned} |t_1|^2 + |r_1|^2 &= |t_2|^2 + |r_2|^2 = 1, \\ t_1 r_2^* + r_1 t_2^* &= t_1 r_1^* + r_2 t_2^* = 0. \end{aligned} \quad (14.6)$$

The precise values for the reflection and transmission coefficients depend on the physical details of the beam splitter.

Redefining the phases of the four modes a , b , a' , and b' changes the phases of the coefficients:

$$\begin{aligned} t_1 &\rightarrow t_1 e^{i(\delta_a - \delta_{a'})}, & t_2 &\rightarrow t_2 e^{i(\delta_b - \delta_{b'})}, \\ r_1 &\rightarrow r_1 e^{i(\delta_a - \delta_{b'})}, & r_2 &\rightarrow r_2 e^{i(\delta_b - \delta_{a'})}. \end{aligned}$$

In this way, the matrix

$$\begin{pmatrix} t_1 & r_1 \\ r_2 & t_2 \end{pmatrix}$$

can be cast into the symmetric form:

$$\begin{pmatrix} t_1 & r_1 \\ r_2 & t_2 \end{pmatrix} = \begin{pmatrix} \sqrt{T} & i\sqrt{1-T} \\ i\sqrt{1-T} & \sqrt{T} \end{pmatrix} \quad \text{with } 0 \leq T \leq 1. \quad (14.7)$$

For simplicity, we shall restrict ourselves to the discussion of the symmetric beam splitter with the symmetric matrix:

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & +i \\ +i & 1 \end{pmatrix} \quad (14.8)$$

and

$$\begin{aligned} a' &= \frac{1}{\sqrt{2}}(a + ib), & b' &= \frac{1}{\sqrt{2}}(b + ia), \\ a'^+ &= \frac{1}{\sqrt{2}}(a^+ - ib^+), & b'^+ &= \frac{1}{\sqrt{2}}(b^+ - ia^+). \end{aligned} \quad (14.9)$$

From this, it is easy to derive the corresponding transformation of the quantum states caused by the beam splitter.

The total vacuum state $|0\rangle_a|0\rangle_b$ remains a total vacuum state: $|0\rangle_a|0\rangle_b = |0\rangle_{a'}|0\rangle_{b'}$. For one-photon states we have, for instance,

$$|1, 0\rangle_{ab} \rightarrow \frac{1}{\sqrt{2}}|1, 0\rangle_{ab} - \frac{i}{\sqrt{2}}|0, 1\rangle_{ab} = |1, 0\rangle_{a'b'}. \quad (14.10)$$

The transformation of $|n_1, n_2\rangle_{ab}$ is easily evaluated. Particularly simple laws apply for coherent (or squeezed) states, e.g.

$$|\alpha\rangle_a|\beta\rangle_b = |\alpha, \beta\rangle_{ab} \rightarrow \left| \frac{1}{\sqrt{2}}(\alpha + i\beta), \frac{1}{\sqrt{2}}(\beta + i\alpha) \right\rangle_{ab} = |\alpha, \beta\rangle_{a'b'}. \quad (14.11)$$

Coherent states remain coherent but with transformed coherence parameters.

Even if the second ray is a vacuum state, we have

$$|\alpha, 0\rangle_{ab} \rightarrow \left| \frac{1}{\sqrt{2}}\alpha, \frac{i}{\sqrt{2}}\alpha \right\rangle_{ab} \quad (14.12)$$

with both components present.

A beam splitter can be used for measuring unknown states. Suppose that the state of the light field in the Heisenberg picture is given by the product of a coherent state $|\alpha\rangle_a$ and an unknown state ρ_b with respect to the operators a and b :

$$\rho_{ab} = |\alpha\rangle_{aa}\langle\alpha| \otimes \rho_b. \quad (14.13)$$

The photon number operators $N_{a'}$ and $N_{b'}$ on the outgoing side are given by

$$\begin{aligned} N_{a'} &= a'^+a' = \frac{1}{2}[a^+a + b^+b + i(a^+b - b^+a)], \\ N_{b'} &= b'^+b' = \frac{1}{2}[a^+a + b^+b - i(a^+b - b^+a)]. \end{aligned} \quad (14.14)$$

In the difference

$$\Delta N' = N_{a'} - N_{b'} = i(a^+b - b^+a), \quad (14.15)$$

half of the terms cancel out. It is convenient to measure $N_{a'}$, $N_{b'}$, and $\Delta N'$ by counting outgoing photons.

The expectation value in the state ρ_{ab} is simply

$$\langle \Delta N' \rangle_{\rho_{ab}} = \langle i(\alpha^* b - \alpha b^+) \rangle_{\rho_b} \quad (14.16)$$

or, with $\alpha = |\alpha| e^{i\theta}$,

$$\begin{aligned} \langle \Delta N' \rangle_{\rho_{ab}} &= |\alpha| \langle b e^{-i(\theta - \frac{1}{2}\pi)} + b^+ e^{i(\theta - \frac{1}{2}\pi)} \rangle_{\rho_b} \\ &= \sqrt{2} |\alpha| \langle X_{\theta - \frac{1}{2}\pi} \rangle_{\rho_b}. \end{aligned} \quad (14.17)$$

Remarkably enough, we obtain the expectation value of a rotated quadrature with respect to the operators b and b^+ . By means of a phase shifter for the oscillator b , we can vary θ in a controlled way and thus measure the expectation value of all rotated quadratures.

The detection scheme just described is called *homodyne detection*. The name refers to the fact that a and b belong to equal frequencies, otherwise one would talk about *heterodyne detection*. Homodyne detection only works if the operators of the incoming states are phase locked, i.e. they must have a stable time-independent phase difference.

When we try to calculate the expectation value $\langle (\Delta N')^2 \rangle_{\rho_{ab}}$, we first obtain terms like aa^+b^+b , where a^+ stands to the right of a such that the expectation value in the coherent state $|\alpha\rangle$ has to be calculated by reordering:

$$\begin{aligned} \langle aa^+b^+b \rangle_{\rho_{ab}} &= \langle a^+abb^+ \rangle_{\rho_{ab}} + \langle [a, a^+]bb^+ \rangle_{\rho_{ab}} \\ &= (1 + |\alpha|^2) \langle bb^+ \rangle_{\rho_b}. \end{aligned}$$

In the strong oscillator limit, $|\alpha| \gg 1$, the commutator terms can be neglected and we have

$$\langle (\Delta N')^2 \rangle_{\rho_{ab}} = 2|\alpha|^2 \langle X_{\theta - \frac{1}{2}\pi}^2 \rangle_{\rho_b}$$

and

$$\sigma_{\Delta N'}^2 = 2|\alpha|^2 \sigma_{X_{\theta - \frac{1}{2}\pi}}^2. \quad (14.18)$$

In particular, the probability distribution for the measured values of $\Delta N'$ directly gives the probability distribution for the measured values of $X_{\theta - \frac{1}{2}\pi}$:

$$w_{\rho_b}(x_{\theta - \frac{1}{2}\pi}) = \langle x_{\theta - \frac{1}{2}\pi} | \rho_b | x_{\theta - \frac{1}{2}\pi} \rangle = w_{\rho_b} \left(\frac{\Delta N'}{\sqrt{2}|\alpha|} \right). \quad (14.19)$$

In terms of the Weyl function $W(x_1, x_2)$, we have

$$w_{\rho_b}(x_\theta) = \frac{1}{2\pi} \langle x_\theta | \rho_b | x_\theta \rangle = \int d^2x \delta(x_1 \cos \theta + x_2 \sin \theta - x_\theta) W_{\rho_b}(x_1, x_2). \quad (14.20)$$

If $w(x_\theta)$ is known for all θ and for all values of x_θ , then the Weyl function $W_{\rho_b}(x_1, x_2)$ is completely known.

In eq. (14.20), $w_{\rho_b}(x_\theta)$ is just the *Radon transform* of the Weyl function $W(x_1, x_2)$. In general, the Radon transform of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$(Rf)(\mathbf{n}, k) = \int d^n x \delta(\mathbf{n} \cdot \mathbf{x} - k) f(\mathbf{x}). \quad (14.21)$$

The Radon transform $(Rf)(\mathbf{n}, k)$ is just the value of the integral over f along the hyperplane $\{\mathbf{x} \mid \mathbf{n} \cdot \mathbf{x} = k\}$. Inversion of the Radon transform is achieved by Fourier transformation with respect to k :

$$\begin{aligned} \int dk e^{ikA} (Rf)(\mathbf{n}, k) &= \int d^n x \int dk e^{ikA} \delta(\mathbf{n} \cdot \mathbf{x} - k) f(\mathbf{x}) \\ &= \int d^n x e^{iA\mathbf{n} \cdot \mathbf{x}} f(\mathbf{x}) = \tilde{f}(A\mathbf{n}) (2\pi)^{n/2}. \end{aligned} \quad (14.22)$$

The inverse Fourier transform of $\tilde{f}(A\mathbf{n})$ gives $f(\mathbf{x})$. Now,

$$w_{\rho_b}(x_\theta) = (RW_{\rho_b})(\mathbf{n}, k) \quad \text{with} \quad \mathbf{n} = (\cos \theta, \sin \theta), \quad k = x_\theta, \quad (14.23)$$

and W_{ρ_b} can be obtained by inverting the Radon transform.

In analogy to computer tomography imaging of the human body, which also makes essential use of the Radon transformation, one talks about tomographic reconstruction of the Wigner function.

With the so-called *eight-port interferometer*, it is even possible to measure two complementary quadratures. The arrangement consists of a combination of two homodyne detectors and is depicted in fig. 14.2.

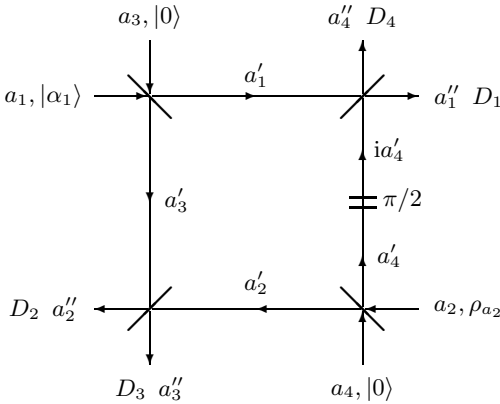


Figure 14.2: Principle of an eight-port interferometer.

There are four incoming rays corresponding to oscillators a_1 , a_2 , a_3 , and a_4 , and four outgoing rays with oscillators a'_1 , a'_2 , a'_3 , and a'_4 , whose photons are measured in detectors D_1 , D_2 , D_3 , and D_4 . The incoming state is prepared as

$$\rho = |\alpha_1\rangle e^{i\theta} \rangle_{a_1} \langle \alpha_1| e^{i\theta} \rangle \otimes \rho_{a_2} \otimes |0\rangle_{a_3} \langle 0| \otimes |0\rangle_{a_4} \langle 0|, \quad (14.24)$$

a coherent state with adjustable phase for a_1 , a state ρ_{a_2} for a_2 which is to be analyzed, and vacuum states for a_3 and a_4 . In one arm there is a device that shifts the phase by $\pi/2$.

We have:

$$\begin{aligned} [a'_i, a'_j] &= [a_i'^+, a_j'^+] = 0, & [a'_i, a_j'^+] &= \delta_{ij}, \\ [a''_i, a''_j] &= [a_i''^+, a_j''^+] = 0, & [a''_i, a_j''^+] &= \delta_{ij}. \end{aligned} \quad (14.25)$$

We readily calculate:

$$\begin{aligned} a''_2 &= \frac{1}{\sqrt{2}}(a'_2 + ia'_3), & a''_3 &= \frac{1}{\sqrt{2}}(a'_3 + ia'_2), \\ a''_1 &= \frac{1}{\sqrt{2}}(a'_1 - a'_4), & a''_4 &= \frac{i}{\sqrt{2}}(a'_4 + a'_1). \end{aligned} \quad (14.26)$$

The photon number operators $N_i'' = a_i''^+ a_i''$ all commute, and we obtain

$$\begin{aligned} \Delta N_{23} &= N_2'' - N_3'' = i(a_2'^+ a'_3 - a_3'^+ a'_2), \\ \Delta N_{14} &= N_1'' - N_4'' = -(a_1'^+ a'_4 + a_4'^+ a'_1). \end{aligned} \quad (14.27)$$

In the strong oscillator limit, $|\alpha_1| \gg 1$, we may replace a_1 by α_1 and a_1^+ by α_1^* , and replace $\alpha_1 \pm a_3$ by α_1 , and obtain

$$\begin{aligned} \Delta N_{23} &\rightarrow -\frac{1}{\sqrt{2}}(\alpha_1 a_2'^+ + \alpha_1^* a_2') \\ &= -\frac{1}{2}(\alpha_1 a_2^+ + \alpha_1^* a_2) + \frac{i}{2}(\alpha_1 a_4^+ - \alpha_1^* a_4) \\ &= \frac{|\alpha_1|}{\sqrt{2}}[-X_\theta(a_2) + X_{\theta+\frac{1}{2}\pi}(a_4)], \\ \Delta N_{14} &\rightarrow -\frac{1}{\sqrt{2}}(\alpha_1 a_4'^+ + \alpha_1^* a_4') \\ &= +\frac{i}{2}(\alpha_1 a_2^+ - \alpha_1^* a_2) - \frac{1}{2}(\alpha_1 a_4^+ + \alpha_1^* a_4) \\ &= \frac{|\alpha_1|}{\sqrt{2}}[X_{\theta+\frac{1}{2}\pi}(a_2) - X_\theta(a_4)]. \end{aligned}$$

In this approximation, ΔN_{23} and ΔN_{14} still commute. Furthermore, in expectation values with vacuum states for a_3 and a_4 , we may omit the parts with a_4 and a_4^+ .

Finally, for the expectation values with respect to a state and in the limit $|\alpha_1| \gg 1$ we may always replace

$$\Delta N_{23} \rightarrow -\frac{|\alpha_1|}{\sqrt{2}}X_\theta(a_2), \quad \Delta N_{14} \rightarrow \frac{|\alpha_1|}{\sqrt{2}}X_{\theta+\frac{1}{2}\pi}(a_2).$$

So, in this limit, we obtain two complementary rotated quadrature operators associated to a_2 , which do not commute but in the strong oscillator limit are equivalent with respect to expectation values for states of the form of eq. (14.24).

Measuring the joint count distribution of ΔN_{23} and ΔN_{14} , we can readily measure

$$Q_{\rho_{a_2}} = \langle \xi | \rho_{a_2} | \xi \rangle,$$

where $\xi = (1/\sqrt{2})(x_1 + ix_2)$.

14.2 Correlation functions and quantum coherence

In chapter 12, we have already discussed in some detail the importance and properties of expectation values of products of stochastic fields. Most of the content of chapter 12 can be literally taken over to expectation values of products of quantum fields with the caveat that this time the order of the non-commuting quantum fields has to be respected.

Keeping as close as possible to the notations of chapter 12, we denote by $V(x) = V(t, \mathbf{x})$ and $V^+(x) = V^+(t, \mathbf{x})$ the annihilation and creation parts of the operator for the electric or magnetic field (eq. (13.40)) which, just like the stochastic fields V and V^+ of chapter 12, contain only positive and negative frequencies, respectively.

Indices related to polarization information will be suppressed. The operator $I(x) = I(t, \mathbf{x}) = V^+(x)V(x)$ will be proportional to the energy density, the photon number density, and the intensity of the field.

Quite generally, we shall consider expectation values

$$\begin{aligned} \Gamma^{(n)}(x_1, \dots, x_{2n}) &= \langle V^+(x_1) \cdots V^+(x_n) V(x_{n+1}) \cdots V(x_{2n}) \rangle_\rho \\ &= \text{Tr}\{\rho V^+(x_1) \cdots V^+(x_n) V(x_{n+1}) \cdots V(x_{2n})\} \end{aligned} \quad (14.28)$$

associated to a given density matrix ρ . Expectation values with a different number of creation and annihilation operators are less important and will not be considered. The reason why we prefer normal ordering for the expectation values will become clear in the next section.

In addition, we shall need normalized expectation values:

$$\begin{aligned} \gamma^{(n)}(x_1, \dots, x_{2n}) &= \frac{\Gamma^{(n)}(x_1, \dots, x_{2n})}{[\Gamma^{(1)}(x_1, x_1)]^{1/2} \cdots [\Gamma^{(1)}(x_{2n}, x_{2n})]^{1/2}} \\ &= \frac{\Gamma^{(n)}(x_1, \dots, x_{2n})}{\langle I(x_1) \rangle^{1/2} \cdots \langle I(x_{2n}) \rangle^{1/2}}. \end{aligned} \quad (14.29)$$

Here $\gamma^{(n)}$ is called the *coherence function of order n*. The functions

$$\Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1) = \langle V^+(x_1) \cdots V^+(x_n) V(x_n) \cdots V(x_1) \rangle_\rho \quad (14.30)$$

and

$$\gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1) = \frac{\Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1)}{\langle I(x_1) \rangle \cdots \langle I(x_n) \rangle} \quad (14.31)$$

will turn out to be of particular importance. We have

$$\Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1) \geq 0, \quad (14.32)$$

because this quantity is the expectation value of a positive operator of the form $A^\dagger A$.

From Schwarz's inequality

$$|\langle A^+ B \rangle_\rho|^2 \leq \langle A^+ A \rangle_\rho \langle B^+ B \rangle_\rho, \quad (14.33)$$

we obtain the inequalities

$$|\Gamma^{(n)}(x_1, \dots, x_{2n})|^2 \leq \Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1) \\ \times \Gamma^{(n)}(x_{2n}, \dots, x_{n+1}, x_{n+1}, \dots, x_{2n}), \quad (14.34)$$

and, analogously, for the normalized expectation values $\gamma^{(n)}$. In particular,

$$|\Gamma^{(1)}(x_1, x_2)|^2 \leq \Gamma^{(1)}(x_1, x_1) \Gamma^{(1)}(x_2, x_2), \\ |\gamma^{(1)}(x_1, x_2)| \leq 1. \quad (14.35)$$

[There is no corresponding bound for the higher coherence functions $\gamma^{(n)}$.] If

$$|\gamma^{(1)}(x_1, x_2)| = 1, \quad (14.36)$$

the state ρ is said to be *first-order coherent*. This condition was analyzed in detail in chapter 12. For the quantum fields it means that in all expectation values in the state ρ , we can replace

$$V(x_2) = c(x_2, x_1)V(x_1), \quad V^+(x_2) = c^*(x_2, x_1)V^+(x_1), \quad (14.37)$$

with $c(x_2, x_1) \in \mathbb{C}$. Just as in the classical case, this is a very close relationship of the phases of the field at different points.

More generally, we can define coherence of higher order n_0 if

$$|\gamma^{(n)}(x_1, \dots, x_{2n})| = 1 \quad \text{for } n \leq n_0, \quad (14.38)$$

and perfect coherence if

$$|\gamma^{(n)}(x_1, \dots, x_{2n})| = 1 \quad \text{for all } n. \quad (14.39)$$

[If the state ρ contains at most n_0 photons, then, evidently, $\gamma^{(n)}(x_1, \dots, x_{2n}) = 0$ for $n > n_0$.]

If ρ is a multi-mode coherent state

$$\rho = |v\rangle\langle v| \quad \text{with} \quad V(x)|v\rangle = v(x)|v\rangle, \quad (14.40)$$

then we have

$$\Gamma^{(n)}(x_1, \dots, x_{2n}) = \langle v|V^+(x_1) \cdots V^+(x_n)V(x_{n+1}) \cdots V(x_{2n})|v\rangle \\ = v^*(x_1) \cdots v^*(x_n)v(x_{n+1}) \cdots v(x_{2n}). \quad (14.41)$$

This complete factorization of the expectation values immediately implies perfect coherence to all orders and explains the name “coherent state”.

The quantum field $V(x)$ obeys the same equations of motion as the classical field. Everything that was said in chapter 12 about the propagation of the classical field, including the van Cittert–Zernicke theorem, remains true in the quantum case.

Again, in most cases of importance, the state ρ will be stationary and the functions $\Gamma^{(n)}(x_1, \dots, x_{2n})$ will depend only on time differences. In particular, $\langle I(x) \rangle$ will be independent of time. The Wiener–Khinchin theorem remains valid.

14.3 Measurement of correlation functions

The expectation values $\Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1)$ can be measured directly by counting photons near the space-time points $x_1, \dots, x_n$. Suppose that a photon detector is placed at the point x_1 in space. Detection of a photon means its absorption by the detector, followed by a cascade of ionization processes. Thus, the interaction Hamiltonian of the detector and the electromagnetic field will be proportional to the annihilation operator $V(x_1)$. The transition probability $w(x_1)$ per unit time from an initial state $|i\rangle$ to a final state $|f\rangle$ will be proportional to

$$w^{fi}(x_1) \sim |\langle f|V(x_1)|i\rangle|^2 = \langle i|V^+(x_1)|f\rangle\langle f|V(x_1)|i\rangle. \quad (14.42)$$

Now, the final state $|f\rangle$ will not be registered, and the total probability per unit time of an absorption of a photon from the initial state $|i\rangle$ is

$$w^i(x_1) \sim \sum_f w^{fi}(x_1) \sim \langle i|V^+(x_1)V(x_1)|i\rangle. \quad (14.43)$$

In many cases, the initial state $|i\rangle$ is not precisely known and has to be described by a density matrix $\rho = \sum_i p_i |i\rangle\langle i|$. Then the total detection probability per unit time is

$$w_\rho(x_1) = \text{Tr}\{\rho V^+(x_1)V(x_1)\} = \langle V^+(x_1)V(x_1)\rangle_\rho = \Gamma^{(1)}(x_1, x_1). \quad (14.44)$$

Quite analogously, for n photon detectors, n annihilation operators will be relevant, and the joint probability for detecting photons at $x_1, \dots, x_n$ in time intervals $\Delta t_1, \dots, \Delta t_n$ will be $w(x_1, \dots, x_n)\Delta t_1 \cdots \Delta t_n$ with

$$\begin{aligned} w_\rho(x_1, \dots, x_n) &\sim \langle V^+(x_1) \cdots V^+(x_n)V(x_n) \cdots V(x_1)\rangle_\rho \\ &= \Gamma^{(n)}(x_1, \dots, x_n, x_n, \dots, x_1). \end{aligned} \quad (14.45)$$

If coherence up to order n_0 holds, eq. (14.38) will tell us

$$w_\rho(x_1, \dots, x_n) = w_\rho(x_1) \cdots w_\rho(x_n) \quad \text{for } n \leq n_0, \quad (14.46)$$

and the photon counts at different space-time points will be statistically uncorrelated up to order n_0 .

The correlation function $\Gamma^{(1)}(x_1, x_2)$ for different arguments x_1 and x_2 can be measured by interference as explained in connection with Young's interference experiment in chapter 12. If an interference device guarantees that at some point x we have $V(x) = a_1 V(x_1) + a_2 V(x_2)$, then $\Gamma^{(1)}(x_1, x_2)$ can be obtained by measuring $\langle V^*(x)V(x)\rangle_\rho$. Generalizing Young's interference scheme, one can, at least in principle, also measure $\Gamma(x_1, \dots, x_{2n})$ with all arguments different.

Let us treat a simple example in somewhat more detail, which will illustrate the description and understanding of the Michelson stellar interferometer and the Hanbury Brown and Twiss intensity interferometer in the framework of quantum optics. Suppose that the field $V(x)$ consists of only two modes of equal frequency ω and wave vectors $\mathbf{k}$ and $\mathbf{k}'$, with $|\mathbf{k} - \mathbf{k}'| \ll |\mathbf{k} + \mathbf{k}'|$. This is an idealization for a light source consisting of two point sources

and observed from a very large distance. The field $V(t, \mathbf{x})$ will then be given by (compare eq. (13.39)):

$$V(t, \mathbf{x}) = E_\omega (a_{\mathbf{k}} e^{-i(\omega t - \mathbf{k} \cdot \mathbf{x})} + a_{\mathbf{k}'} e^{-i(\omega t - \mathbf{k}' \cdot \mathbf{x})}). \quad (14.47)$$

An idealized Michelson stellar interferometer will measure

$$\begin{aligned} & \langle [V^+(x_1) + V^+(x_2)][V(x_1) + V(x_2)] \rangle_\rho \\ &= \langle V^+(x_1)V(x_1) \rangle_\rho + \langle V^+(x_2)V(x_2) \rangle_\rho + 2 \operatorname{Re} \langle V^+(x_1)V(x_2) \rangle_\rho \\ &= \langle I(x_1) \rangle + \langle I(x_2) \rangle + 2 \operatorname{Re} \Gamma^{(1)}(x_1, x_2), \end{aligned} \quad (14.48)$$

the same quantity as also measured by Young's interference experiment. With eq. (14.47) the interesting part of eq. (14.48) will be

$$\begin{aligned} & 2 \operatorname{Re} \langle V^+(x_1)V(x_2) \rangle_\rho \\ &= 2|E_\omega|^2 \operatorname{Re} \{ \langle a_{\mathbf{k}}^+ a_{\mathbf{k}} \rangle_\rho e^{+i\omega(t_1 - t_2) - i\mathbf{k} \cdot (\mathbf{x}_1 - \mathbf{x}_2)} + \langle a_{\mathbf{k}'}^+ a_{\mathbf{k}'} \rangle_\rho e^{i\omega(t_1 - t_2) - i\mathbf{k}' \cdot (\mathbf{x}_1 - \mathbf{x}_2)} \\ & \quad + \langle a_{\mathbf{k}}^+ a_{\mathbf{k}'} \rangle_\rho e^{i\omega(t_1 - t_2) - i(\mathbf{k} \cdot \mathbf{x}_1 - \mathbf{k}' \cdot \mathbf{x}_2)} \langle a_{\mathbf{k}'}^+ a_{\mathbf{k}} \rangle_\rho e^{i\omega(t_1 - t_2) - i(\mathbf{k}' \cdot \mathbf{x}_1 - \mathbf{k} \cdot \mathbf{x}_2)} \}. \end{aligned} \quad (14.49)$$

If ρ is a coherent state, all four terms in eq. (14.49) will contribute. But in most cases of interest non-vanishing expectation values have equal numbers of creation and annihilation operators for each mode separately. This is seen to be the case for a thermal state ρ_{th} or if ρ is a photon number eigenstate $|n_{\mathbf{k}}, n_{\mathbf{k}'}\rangle$ and also for phase diffused laser light. Assuming in addition that

$$\langle a_{\mathbf{k}}^+ a_{\mathbf{k}} \rangle_\rho = \langle a_{\mathbf{k}'}^+ a_{\mathbf{k}'} \rangle_\rho = \langle N \rangle_\rho, \quad (14.50)$$

and taking $t_1 = t_2$, we obtain

$$\begin{aligned} & 2 \operatorname{Re} \langle V^+(x_1)V(x_2) \rangle_\rho \\ &= 2|E_\omega|^2 \langle N \rangle_\rho \{ \cos[\mathbf{k} \cdot (\mathbf{x}_1 - \mathbf{x}_2)] + \cos[\mathbf{k}' \cdot (\mathbf{x}_1 - \mathbf{x}_2)] \} \\ &= 4|E_\omega|^2 \langle N \rangle_\rho \cos\left\{\frac{1}{2}[(\mathbf{k} + \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\right\} \cos\left\{\frac{1}{2}[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\right\}. \end{aligned} \quad (14.51)$$

The factor $\cos\left\{\frac{1}{2}[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\right\}$ allows for a precise measurement of a small angle between $\mathbf{k}$ and $\mathbf{k}'$. The precision is limited by the first factor $\cos\left\{\frac{1}{2}[(\mathbf{k} + \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\right\}$, which varies rapidly with $\mathbf{x}_1 - \mathbf{x}_2$ and imposes strong requirements on the mechanical stability of the interferometer.

In the arrangement of Hanbury Brown and Twiss, one measures correlations in joint count rates $\langle V^+(x_1)V^+(x_2)V(x_2)V(x_1) \rangle_\rho$, making use of the lack of second-order coherence. Again, keeping only contributions with paired annihilation and creation operators and assuming $t_1 = t_2$, we obtain with eq. (14.47)

$$\begin{aligned} & \langle V^+(x_1)V^+(x_2)V(x_2)V(x_1) \rangle_\rho \\ &= |E_\omega|^4 \{ \langle a_{\mathbf{k}}^{+2} a_{\mathbf{k}}^2 \rangle_\rho + \langle a_{\mathbf{k}'}^{+2} a_{\mathbf{k}'}^2 \rangle_\rho + 2 \langle a_{\mathbf{k}}^+ a_{\mathbf{k}'}^+ a_{\mathbf{k}} a_{\mathbf{k}'} \rangle_\rho \operatorname{Re}(1 + e^{-i(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)}) \} \end{aligned} \quad (14.52)$$

and, assuming

$$\begin{aligned} & \langle a_{\mathbf{k}}^+ a_{\mathbf{k}} \rangle_\rho = \langle a_{\mathbf{k}'}^+ a_{\mathbf{k}'} \rangle_\rho = \langle N \rangle_\rho, \\ & \langle (a_{\mathbf{k}}^+ a_{\mathbf{k}})^2 \rangle_\rho = \langle (a_{\mathbf{k}'}^+ a_{\mathbf{k}'}^2) \rangle_\rho = \langle N^2 \rangle_\rho, \\ & \langle a_{\mathbf{k}}^+ a_{\mathbf{k}} a_{\mathbf{k}'}^+ a_{\mathbf{k}'} \rangle_\rho = \langle N \rangle_\rho^2, \end{aligned} \quad (14.53)$$

we find

$$\begin{aligned} & \langle V^+(x_1)V^+(x_2)V(x_2)V(x_1) \rangle_\rho \\ &= 2|E_\omega|^4 (\langle N^2 \rangle_\rho - \langle N \rangle_\rho^2 + \langle N \rangle_\rho^2 \{1 + \cos[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\}). \end{aligned} \quad (14.54)$$

Notice the absence of the cumbersome rapidly varying term in eq. (14.51). If ρ corresponds to the state $|1_{\mathbf{k}}, 1_{\mathbf{k}'}\rangle$ with exactly one photon in each mode, eq. (14.54) reduces to

$$\Gamma_{11}^{(2)} = 2|E_\omega|^4 \{1 + \cos[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\}. \quad (14.55)$$

For Poissonian light and phase diffused laser light we have

$$\langle N^2 \rangle_\rho - \langle N \rangle_\rho^2 = \langle N \rangle_\rho$$

and

$$\Gamma_{\text{Poisson}}^{(2)}(x_1, x_2, x_2, x_1) = 2|E_\omega|^4 \langle N \rangle_\rho^2 \{2 + \cos[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\}, \quad (14.56)$$

and for thermal light, for instance, starlight, we find that

$$\langle N_\rho^2 \rangle - \langle N \rangle_\rho^2 = \langle N \rangle_\rho^2 + \langle N \rangle_\rho$$

holds and

$$\Gamma_{\text{th}}^{(2)}(x_1, x_2, x_2, x_1) = 2|E_\omega|^4 \langle N \rangle_\rho^2 \{3 + \cos[(\mathbf{k} - \mathbf{k}') \cdot (\mathbf{x}_1 - \mathbf{x}_2)]\}. \quad (14.57)$$

The discussion in chapter 12 shows how the shape of the light source can be determined from a measurement of $\Gamma^{(1)}(x_1, x_2)$ or $\Gamma^{(2)}(x_1, x_2, x_2, x_1)$ even if the source contains more than just two modes $a_{\mathbf{k}}$ and $a_{\mathbf{k}'}$.

Interferometers can be used not only for detecting the properties of incoming light fields. It is also possible to assume the light field to be known and to use interference effects to detect distortions of the interferometer device. In this case a sensitive dependence on $\mathbf{x}_1 - \mathbf{x}_2$ as visible in eq. (14.51) is the highly desirable effect responsible for the sensibility of the instrument. A first example for this use of interferometry was Michelson's interferometer, originally devised to detect contractions due to motion through the ether.

A more recent and very attractive application of interferometry is the construction of gravitational wave detectors. Gravitational waves propagate with the speed of light and are emitted by any arrangement of masses with a time-dependent quadrupole moment. Unfortunately, under normal circumstances, the radiated power is so low that direct detection of gravitational radiation is extremely difficult. Indirect evidence for the existence of gravitational radiation was, however, obtained from observing the energy loss in the double pulsar PSR 1913+16. Strong gravitational radiation is to be expected from collapsing gravitating objects like supernovas or coalescing black holes. Detecting gravitational waves will open a new window for astronomy.

Gravitational waves are wave-like perturbations $h_{\mu\nu}$ of the space-time metric $g_{\mu\nu}$:

$$g_{\mu\nu} = g_{\mu\nu}^0 + h_{\mu\nu}. \quad (14.58)$$

Their effect is a relative length change of the order of magnitude $|h_{\mu\nu}|$. More precisely, transverse to the direction of propagation of a gravitational wave, one has a relative dilatation of order $|h|$ in one perpendicular direction and a contraction of the same size in the other.

This effect can in principle be detected with a Michelson interferometer (fig. 14.3). The challenge lies in the smallness of the effect. A theoretical estimate yields $|h| \sim 10^{-21}$. With $\Delta L = Lh$, for $L = 100$ km this corresponds to $\Delta L = 10^{-14}$ cm, less than the diameter of an atomic nucleus. One may ask how the surface of a mirror in an interferometer can possibly be defined to such precision, but, fortunately, the surface is composed of many atoms, leading to an averaging out of the effects of individual atoms and, more importantly, only displacements of the surface matter rather than the precise definition of its position.

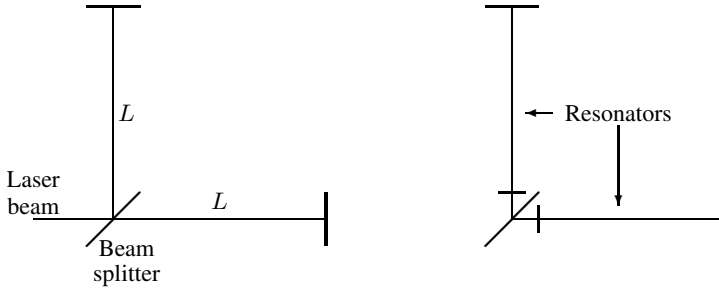


Figure 14.3: Michelson's interferometer.

The minimal detectable $\Delta L = hL$ is approximately given by

$$\Delta L \approx \frac{\lambda}{2\pi} \Delta\theta, \quad (14.59)$$

where λ is the wavelength of the light in the interferometer and $\Delta\theta$ its phase uncertainty.

The frequency of the gravitational wave to be detected is of the order of magnitude of 10–10 000 Hz, which corresponds to wavelengths of 30 000 to 3 km. Optimal resolution would require the dimension of the instrument to be about half a wavelength, but dimensions exceeding a few kilometers are not feasible for Earth-based interferometers, yet are planned for space-based instruments. The largest instruments that have been built, like the German–English project GEO 600 near Hannover, the American LIGO, and the French–Italian VIRGO, using all the tricks of interferometry, have reduced the phase uncertainty to the limit of the quantum theoretical phase uncertainty σ_θ . As we saw in section 13.2 (compare fig. 13.1), for unsqueezed light, this is given by

$$\sigma_\theta = \frac{1}{\sqrt{\langle N \rangle}}. \quad (14.60)$$

The average photon number $\langle N \rangle = \langle a^\dagger a \rangle$ is determined by the power P of the laser light and the averaging time t , which has to be smaller than the period of the gravitational wave:

$$\langle N \rangle = \frac{Pt}{\hbar\omega}. \quad (14.61)$$

Thus, σ_θ can be made very small by increasing the power. The main limiting effect for phase resolution will then be the quantum phase uncertainty of the vacuum mode entering into the unused port of the beam splitter in the interferometer. By illuminating the unused port with phase squeezed laser light, phase locked to the main beam, this phase uncertainty can be further reduced.

The sensitivity of the interferometer can be increased enormously by placing resonators in its arms. This way, the arm length L is replaced by the effective arm length $L_{\text{eff}} = RL$, where R is the number of reflections in the resonators. The setting with two arms may be replaced by a ring-like device (a Sagnac interferometer).

14.4 Anti-bunching and sub-Poissonian light

In this section, we shall discuss the properties of the count statistics of photons detected at a fixed given point $\mathbf{x}$. We shall consider stationary states ρ such that

$$\langle I(t, \mathbf{x}) \rangle_\rho = \langle V^+(t, \mathbf{x})V(t, \mathbf{x}) \rangle_\rho$$

is independent of t , and that

$$\gamma^{(1)}((t+\tau, \mathbf{x}), (t, \mathbf{x})) = \frac{\langle V^+(t+\tau, \mathbf{x})V(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho} =: \gamma^{(1)}(\tau) \quad (14.62)$$

and

$$\frac{\langle V^+(t, \mathbf{x})V^+(t+\tau, \mathbf{x})V(t+\tau, \mathbf{x})V(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} =: \gamma^{(2)}(\tau) \quad (14.63)$$

depend on τ but not on t .

In quantum field theory, the photon count correlation function (14.63) differs from the intensity correlation function

$$\begin{aligned} \frac{\langle I(t+\tau, \mathbf{x})I(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} &= \frac{\langle V^+(t+\tau, \mathbf{x})V(t+\tau, \mathbf{x})V^+(t, \mathbf{x})V(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} \\ &=: g^{(2)}(\tau), \end{aligned} \quad (14.64)$$

whereas for commuting classical stochastic fields, as discussed in chapter 12, $\gamma^{(2)}(\tau)$ and $g^{(2)}(\tau)$ coincide.

Schwarz's inequality for $g^{(2)}(\tau)$ together with the stationarity of ρ yield

$$|g^{(2)}(\tau)| \leq \frac{\langle I^2(t+\tau, \mathbf{x}) \rangle_\rho^{1/2} \langle I^2(t, \mathbf{x}) \rangle_\rho^{1/2}}{I(t, \mathbf{x})^2} = \frac{\langle I^2(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} = g^{(2)}(0). \quad (14.65)$$

Hence, for commuting fields or if the non-commutativity is negligible, we also have

$$|\gamma^{(2)}(\tau)| \leq \gamma^{(2)}(0). \quad (14.66)$$

Equation (14.66) is called the *bunching condition*. It means that the probability for two consecutive counts at times t and $t + \tau$ is maximal for $\tau = 0$; in other words, counts have a

tendency to aggregate. (By our derivation we see that eq. (14.66) is not a reflection of the boson character of the photons. Bunching may also occur for fermions – for instance, for neutrons.) If the non-commutativity of the fields is not negligible, we may have the classically forbidden inequality:

$$|\gamma^{(2)}(\tau)| > \gamma^{(2)}(0). \quad (14.67)$$

This is the *anti-bunching* condition. It means that for small E photons show a tendency to segregate: the probability for detecting one photon immediately after another one is suppressed. Anti-bunching may occur for fluorescence light. Fluorescence light can be emitted by an atom only after excitation from the ground state to an excited state. The emission of two immediately subsequent photons is suppressed because, before the next emission, the atom has to be excited again.

Our next task is the investigation of intensity and photon count variances at one and the same space-time point $(t, \mathbf{x})$. From eq. (14.65) we see that

$$\begin{aligned} g^{(2)}(0) &= \frac{\langle I^2(t, \mathbf{x}) \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} = 1 + \frac{\langle I^2(t, \mathbf{x}) \rangle_\rho - \langle I(t, \mathbf{x}) \rangle_\rho^2}{\langle I(t, \mathbf{x}) \rangle_\rho^2} \\ &= 1 + \frac{\langle [I(t, \mathbf{x}) - \langle I(t, \mathbf{x}) \rangle_\rho]^2 \rangle_\rho}{\langle I(t, \mathbf{x}) \rangle_\rho^2} \geq 1. \end{aligned} \quad (14.68)$$

If the non-commutativity of the field is negligible, the classical inequality

$$\gamma^{(2)}(0) \geq 1 \quad (14.69)$$

will hold. For genuine quantum states, also

$$\gamma^{(2)}(0) < 1 \quad (14.70)$$

will be possible.

To see the significance of eqs. (14.69) and (14.70), it is convenient, though not necessary, to assume that the light is sufficiently monochromatic and well collimated such that only one mode contributes. In this situation, $g^{(2)}(0)$ and $\gamma^{(2)}(0)$ simplify to

$$g^{(2)}(0) = \frac{\langle a^+ a a^+ a \rangle_\rho}{\langle a^+ a \rangle_\rho^2} \quad (14.71)$$

and

$$\gamma^{(2)}(0) = \frac{\langle a^+ a^+ a a \rangle_\rho}{\langle a^+ a \rangle_\rho^2}. \quad (14.72)$$

Now,

$$\begin{aligned} \gamma^{(2)}(0) &= \frac{\langle a^+ ([a^+, a] + a a^+) a \rangle_\rho}{\langle a^+ a \rangle_\rho^2} = \frac{\langle a^+ a a^+ a \rangle_\rho - \langle a^+ a \rangle_\rho}{\langle a^+ a \rangle_\rho^2} \\ &= \frac{\langle N^2 \rangle_\rho - \langle N \rangle_\rho}{\langle N \rangle_\rho^2} = \frac{\langle N^2 \rangle_\rho - \langle N^2 \rangle_\rho^2 + \langle N \rangle_\rho^2 - \langle N \rangle_\rho}{\langle N \rangle_\rho^2} \\ &= \frac{\langle (\Delta N)^2 \rangle_\rho - \langle N \rangle_\rho}{\langle N \rangle_\rho^2} + 1. \end{aligned} \quad (14.73)$$

Hence, the sign of $\langle(\Delta N)^2\rangle_\rho - \langle N\rangle_\rho$ determines whether $\gamma^{(2)}(0) \geq 1$ or $\gamma^{(2)}(0) < 1$. If the count statistic of the photons is Poissonian, then $\langle(\Delta N)^2\rangle_\rho = \langle N\rangle_\rho$. If the count distribution is broader, with larger variance, we are in the *super-Poissonian* case with

$$\langle(\Delta N)^2\rangle_\rho > \langle N\rangle_\rho \quad \text{and} \quad \gamma^{(2)}(0) > 1; \quad (14.74)$$

and if the distribution is narrower, we have the *sub-Poissonian* case with

$$\langle(\Delta N)^2\rangle_\rho < \langle N\rangle_\rho \quad \text{and} \quad \gamma^{(2)}(0) < 1. \quad (14.75)$$

Sub-Poissonian light is a genuine quantum effect and classically impossible. We already know and can immediately read off from eq. (14.72) that coherent light is Poissonian.

If $\rho = |n\rangle\langle n|$ is a pure number state, then we are in the extreme sub-Poissonian situation:

$$\langle\Delta N\rangle^2 = 0 \quad \text{and} \quad \gamma^{(2)}(0) = 1 - \frac{1}{\langle N\rangle_\rho}. \quad (14.76)$$

From this and also from eq. (14.73) it is evident that strongly sub-Poissonian light must have a small mean photon number $\langle N\rangle_\rho$. For sub-Poissonian light, the distribution function $P_\rho(\alpha, \alpha^*)$ of section 13.4 cannot be everywhere positive. Indeed, from eq. (13.143) we learn that

$$\begin{aligned} \gamma^{(2)}(0) &= \frac{1}{\langle a^\dagger a \rangle_\rho^2} \left[\frac{1}{\pi} \int d^2\alpha (\alpha^* \alpha)^2 P(\alpha, \alpha^*) - \left(\frac{1}{\pi} \int d^2\alpha \alpha^* \alpha P(\alpha, \alpha^*) \right)^2 \right] + 1 \\ &= \frac{1}{\langle a^\dagger a \rangle_\rho^2} \left[\frac{1}{\pi} \int d^2\alpha (\alpha^* \alpha)^2 P(\alpha, \alpha^*) - \langle a^\dagger a \rangle_\rho^2 \right] + 1 \\ &= \frac{1}{\langle a^\dagger a \rangle_\rho^2} \frac{1}{\pi} \int d^2\alpha (\alpha^* \alpha - \langle a^\dagger a \rangle_\rho)^2 P(\alpha, \alpha^*) + 1. \end{aligned}$$

If $P(\alpha, \alpha^*) \geq 0$ everywhere, we have $\gamma^{(2)}(0) \geq 1$.

For amplitude squeezed light we find that

$$\langle(\Delta N)^2\rangle_\rho < \langle N\rangle_\rho,$$

because the variance of N is smaller than for coherent light. Hence, also amplitude squeezed light is sub-Poissonian.

The ratio

$$f_\rho = \frac{\langle(\Delta N)^2\rangle_\rho}{\langle N\rangle_\rho} \quad (14.77)$$

is called the *Fano factor* of the state ρ . Sub-Poissonian light corresponds to $f_\rho < 1$.

The Fano factor changes when light falls on a beam splitter. If the beam splitter is described by eq. (14.1), and if the light corresponding to the operator a is in the state ρ , and the light corresponding to the operator b is in the vacuum state, then it is an easy exercise to show that the outgoing light rays with the operators a' and b' are in the states $\rho_{a'}$ and $\rho_{b'}$ with Fano factors

$$\begin{aligned} 1 - f_{\rho_{a'}} &= (1 - f_\rho) |t_1|^2, \\ 1 - f_{\rho_{b'}} &= (1 - f_\rho) |r_1|^2. \end{aligned}$$

In other words, the beam splitter pushes the light closer to the Poissonian case.

15 Interaction of radiation and matter

15.1 The electric dipole interaction

The main subject of this chapter is the detailed description of a system consisting of an atom in a light field. The Hamiltonian

$$H = H_{\text{light}} + H_{\text{atom}} + H_{\text{int}} = H_0 + H_{\text{int}} \quad (15.1)$$

will consist of two parts pertaining to the light field and the atom plus an additional interaction part. For each part we shall make simplifying assumptions, which could be lifted stepwise in later refinements.

- (a) We assume that only a finite number M of modes contribute to the field, hence

$$H_{\text{light}} = \hbar \sum_{\gamma=1}^M \Omega_{\gamma} a_{\gamma}^{\dagger} a_{\gamma}. \quad (15.2)$$

- (b) Only a finite number of discrete energy eigenstates of H_{atom} is effectively accessible. This, in particular, means that we neglect the center-of-mass motion of the atom and disregard ionization processes. Let $\{|\varphi_i\rangle\}$ ($i = 1, \dots, N$) be an orthogonal system of eigenstates,

$$H_{\text{atom}}|\varphi_i\rangle = E_i|\varphi_i\rangle = \hbar\omega_i|\varphi_i\rangle, \quad (15.3)$$

ordered such that

$$\omega_i \geq \omega_j \quad \text{for } i < j. \quad (15.4)$$

In the subspace spanned by the states $|\varphi_i\rangle$, an operator C can be written in the form

$$C = \sum_{i,j} |\varphi_i\rangle\langle\varphi_i| C |\varphi_j\rangle\langle\varphi_j| = \sum_{i,j} C_{ij} \sigma_{ij}. \quad (15.5)$$

The operators

$$\sigma_{ij} = |\varphi_i\rangle\langle\varphi_j| \quad (15.6)$$

satisfy

$$\sigma_{ij} \sigma_{rs} = \delta_{jr} \sigma_{is}.$$

In a matrix representation with respect to the basis $\{|\varphi_i\rangle\}$, σ_{ij} is a matrix with the entry 1 at position (i, j) and 0 entries elsewhere. Now H_{atom} will be of the form

$$H_{\text{atom}} = \sum_{i=1}^M \hbar\omega_i \sigma_{ii}. \quad (15.7)$$

- (c) The interaction between the atom and the field will be the usual Schrödinger picture dipole interaction Hamiltonian

$$H_{\text{int}} = \mathbf{D} \cdot \mathbf{E}(\mathbf{x}_0), \quad (15.8)$$

where $\mathbf{D}$ is the dipole operator $\mathbf{D} = \sum_A e_A \mathbf{Q}_A$ of the atom composed of the position operators of its constituents, and $\mathbf{E}(\mathbf{x}_0)$ is the electric field operator from eq. (13.40). This, in particular, implies that the field does not change appreciably over the diameter of the atom. This is true if the wavelengths of the radiation field are much larger than the atomic diameters.

If the center-of-mass motion of the atom were taken into account, the field $\mathbf{E}(\mathbf{x}_0)$ would also enter at points $\mathbf{x} \neq \mathbf{x}_0$. Under these hypotheses, H_{int} will be

$$H_{\text{int}} = \sum_{i,j,\gamma} \hbar(g_{ij}^\gamma \sigma_{ij} a_\gamma + g_{ij}^{\gamma*} \sigma_{ji} a_\gamma^\dagger), \quad (15.9)$$

with

$$\hbar g_{ij}^\gamma = i\sqrt{\frac{\hbar\omega_\gamma}{2\epsilon_0}} \langle \varphi_i | \mathbf{D} | \varphi_j \rangle \mathbf{u}_\gamma(\mathbf{x}_0). \quad (15.10)$$

So, altogether, our Hamiltonian will be given by

$$\begin{aligned} H &= \sum_{\gamma=1}^M \hbar\Omega_\gamma a_\gamma^\dagger a_\gamma + \sum_{i=1}^N \hbar\omega_i \sigma_{ii} + \sum_{i,j,\gamma} \hbar(g_{ij}^\gamma \sigma_{ij} a_\gamma + g_{ij}^{\gamma*} \sigma_{ji} a_\gamma^\dagger) \\ &= H_0 + H_{\text{int}}. \end{aligned} \quad (15.11)$$

Starting from a different field–atom interaction, for instance from an electric quadrupole or magnetic dipole interaction, we would have arrived at a Hamiltonian of the same form. What really matters is the locality of the interaction and the feature that the field enters only linearly.

The time evolution of the system is described by the unitary time development operator $U(t)$, which solves the differential equation plus initial condition:

$$i\hbar \frac{d}{dt} U(t) = H U(t), \quad U(0) = \mathbf{1}. \quad (15.12)$$

We also consider the free evolution operator $U_0(t)$ with

$$i\hbar \frac{d}{dt} U_0(t) = H_0 U_0(t), \quad U_0(0) = \mathbf{1}. \quad (15.13)$$

In the interaction picture, the important object will be the evolution operator

$$U_I(t) = U_0^+(t)U(t), \quad U_I(0) = \mathbf{1}, \quad (15.14)$$

which satisfies

$$i\hbar \frac{d}{dt} U_I(t) = H_I U_I(t) \quad (15.15)$$

with the interaction operator

$$H_I(t) = U_0^+(t)H_{\text{int}}U_0(t). \quad (15.16)$$

In our case we obtain

$$H_I(t) = \sum_{i,j,\gamma} \hbar (g_{ij}^\gamma \sigma_{ij} a_\gamma e^{i(\omega_i - \omega_j - \Omega_\gamma)t} + g_{ij}^{\gamma*} \sigma_{ji} a_\gamma^+ e^{-i(\omega_i - \omega_j - \Omega_\gamma)t}). \quad (15.17)$$

The various terms in H_I will give rise to particularly large transition probabilities between different atomic levels if the *resonance conditions* are fulfilled:

$$\Delta_{ij}^\gamma := (\omega_i - \omega_j - \Omega_\gamma) = 0. \quad (15.18)$$

The quantities Δ_{ij}^γ are called *detunings*.

In the *rotating phase approximation*, which will be adopted from here onwards, only interaction terms are kept in which an excitation of the atom is associated with the annihilation of a photon and a de-excitation with the creation of a photon. This corresponds to neglecting faster oscillating terms with larger detunings in eq. (15.17). The final Hamiltonian with which we shall work will then be

$$H = \sum_{\gamma=1}^M \hbar \Omega_\gamma a_\gamma^+ a_\gamma + \sum_{i=1}^N \hbar \omega_i \sigma_{ii} + \sum_{i<j,\gamma} \hbar (g_{ij}^\gamma \sigma_{ij} a_\gamma + g_{ij}^{\gamma*} \sigma_{ji} a_\gamma^+) \quad (15.19)$$

with

$$H_I = \sum_{i<j,\gamma} \hbar (g_{ij}^\gamma \sigma_{ij} a_\gamma e^{i\Delta_{ij}^\gamma t} + g_{ij}^{\gamma*} \sigma_{ji} a_\gamma^+ e^{-i\Delta_{ij}^\gamma t}). \quad (15.20)$$

The rotating phase approximation has already been employed in chapter 6 and has turned out to be a good and valuable approximation in many circumstances.

Strong fields interacting with the atom can sometimes be treated as classical fields. This corresponds to replacing the operators a_γ and a_γ^+ in eq. (15.20) by complex numbers α_γ and α_γ^* . Thus, in the semiclassical approximation, we have

$$H_{I,\text{sc}} = \sum_{i<j,\gamma} \hbar (g_{ij}^\gamma \sigma_{ij} \alpha_\gamma e^{i\Delta_{ij}^\gamma t} + g_{ij}^{\gamma*} \sigma_{ji} \alpha_\gamma^* e^{-i\Delta_{ij}^\gamma t}). \quad (15.21)$$

Now, the total Schrödinger picture Hamiltonian for the atom in the field will be time-dependent:

$$H_{\text{sc}} = \sum_i \hbar \omega_i \sigma_{ii} + \sum_{i<j,\gamma} \hbar (g_{ij}^\gamma \sigma_{ij} \alpha_\gamma e^{-i\Omega_\gamma t} + g_{ij}^{\gamma*} \sigma_{ji} \alpha_\gamma^* e^{i\Omega_\gamma t}). \quad (15.22)$$

The quantities $\Omega_{Rij}^\gamma = g_{ij}^\gamma \alpha_\gamma$ are called *complex Rabi frequencies*. Quite frequently, the experimental arrangement is such that only resonant terms with vanishing detuning are relevant. In this case, H_I and $H_{I,sc}$ are time-independent. In particular, the evolution equation (15.15) with $H_{I,sc}$ is just a system of linear differential equations with constant coefficients, which can be solved explicitly with standard methods. We shall return to this case in a later section.

The minimal case $M = 1$, $N = 2$ of a one-mode field interacting with a two-level system is of particular importance as a theoretical laboratory and also approximately realized in many experiments. With a slight change to the more standard notation

$$\sigma_3 = \sigma_{11} - \sigma_{22} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \sigma_+ = \sigma_{12} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \sigma_- = \sigma_{21} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad (15.23)$$

$$\sigma_+ \sigma_- = \sigma_{11}, \quad \sigma_- \sigma_+ = \sigma_{22}, \quad [\sigma_3, \sigma_\pm] = \pm 2\sigma_\pm, \quad \sigma_3 \sigma_\pm + \sigma_\pm \sigma_3 = 0, \quad (15.24)$$

the Hamiltonian of eq. (15.19) assumes the form

$$H = \hbar \Omega a^+ a + \frac{1}{2} \hbar \omega \sigma_3 + \hbar g \sigma_+ a + \hbar g^* \sigma_- a^+, \quad (15.25)$$

with $\omega = \omega_1 - \omega_2$.

To arrive at eq. (15.25), we have subtracted a constant $\frac{1}{2}(\omega_1 + \omega_2)\mathbf{1}$ from H . Equation (15.25) is called the *Jaynes–Cummings Hamiltonian*. Because of its great importance, we shall devote to it a section of its own (section 15.4).

The interaction Hamiltonian of the Jaynes–Cummings model is

$$H_I = \hbar(g\sigma_+ a e^{i\Delta t} + g^* \sigma_- a^+ e^{-i\Delta t}), \quad (15.26)$$

with

$$\Delta = \omega - \Omega. \quad (15.27)$$

In the semiclassical case, the evolution equation for

$$\begin{aligned} H_{I,sc} &= \hbar(g\alpha\sigma_+ e^{i\Delta t} + g^* \alpha^* \sigma_- e^{-i\Delta t}) \\ &= \hbar \begin{pmatrix} 0 & g\alpha e^{i\Delta t} \\ (g\alpha)^* e^{-i\Delta t} & 0 \end{pmatrix} = \hbar e^{\frac{1}{2}i\Delta\sigma_3 t} \begin{pmatrix} 0 & g\alpha \\ (g\alpha)^* & 0 \end{pmatrix} e^{-\frac{1}{2}i\Delta\sigma_3 t} \end{aligned} \quad (15.28)$$

can be solved in a closed form. Indeed, with $[\sigma_3, \sigma_\pm] = \pm 2\sigma_\pm$, it is easy to see that $H_{I,sc}$ in eq. (15.28) is also the interaction Hamiltonian of the simple time-independent Hamiltonian:

$$\hat{H} = \frac{1}{2} \hbar \Delta \sigma_3 + \hbar[g\alpha\sigma_+ + (g\alpha)^* \sigma_-]. \quad (15.29)$$

Here $\hat{H}$ can be rewritten as

$$\hat{H} = \hbar \begin{pmatrix} \frac{1}{2}\Delta & g\alpha \\ (g\alpha)^* & -\frac{1}{2}\Delta \end{pmatrix} = \hbar \Lambda \begin{pmatrix} \Delta/2\Lambda & g\alpha/\Lambda \\ (g\alpha)^*/\Lambda & -\Delta/2\Lambda \end{pmatrix} = \hbar \Lambda \Sigma, \quad (15.30)$$

with

$$\Lambda = \sqrt{\frac{1}{4}\Delta^2 + |g\alpha|^2} \quad \text{and} \quad \Sigma^2 = 1.$$

Now the exponential function $\exp[-(i/\hbar)\hat{H}t]$ can be readily evaluated to

$$e^{-(i/\hbar)\hat{H}t} = \cos(\Lambda t) \cdot \mathbf{1} - i \sin(\Lambda t) \cdot \Sigma. \quad (15.31)$$

Furthermore,

$$U_I(t) = e^{\frac{1}{2}i\Delta\sigma_3 t} e^{-(i/\hbar)\hat{H}t} \quad (15.32)$$

and

$$U(t) = e^{-\frac{1}{2}i\omega\sigma_3 t} U_I(t) = e^{-\frac{1}{2}i\Omega\sigma_3 t} e^{-(i/\hbar)\hat{H}t}. \quad (15.33)$$

So, finally, we obtain for the full time evolution operator of the semiclassical Jaynes–Cummings model:

$$U(t) = \begin{pmatrix} \left[\cos(\Lambda t) - i \frac{\Delta}{2\Lambda} \sin(\Lambda t) \right] e^{-\frac{1}{2}i\Omega t} & -i \frac{g\alpha}{\Lambda} \sin(\Lambda t) e^{-\frac{1}{2}i\Omega t} \\ -i \frac{g^*\alpha^*}{\Lambda} \sin(\Lambda t) e^{\frac{1}{2}i\Omega t} & \left[\cos(\Lambda t) + i \frac{\Delta}{2\Lambda} \sin(\Lambda t) \right] e^{\frac{1}{2}i\Omega t} \end{pmatrix}. \quad (15.34)$$

The occupation probabilities for the two levels are given by $w_{1,2} = |c_{1,2}(t)|^2$ with

$$\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix} = U(t) \begin{pmatrix} c_1(0) \\ c_2(0) \end{pmatrix}, \quad (15.35)$$

and $w_{1,2}(t)$ will oscillate periodically with period 2Λ . This result differs radically from the one we shall obtain for the full Jaynes–Cummings model.

Quite generally, the time evolution of a mixed state ρ of our system will be determined by the quantum Liouville equation:

$$\frac{d}{dt}\rho = \frac{1}{i\hbar}[H, \rho], \quad (15.36)$$

where H can be any one of the Hamiltonians described in this section. In realistic situations, the system will not be closed, and there will be losses by evading photons or occupation of states not explicitly included into the model. These effects can be dealt with phenomenologically by introducing a Hermitian non-negative damping operator Γ and replacing eq. (15.36) by

$$\frac{d}{dt}\rho = \frac{1}{i\hbar}[H, \rho] - \frac{1}{2}(\Gamma\rho + \rho\Gamma). \quad (15.37)$$

This equation guarantees that ρ remains Hermitian and positive. The losses are visible from the identity

$$\frac{d}{dt}\text{Tr}(\rho) = -\text{Tr}(\rho\Gamma) \leq 0. \quad (15.38)$$

This decrease of $\text{Tr } \rho$ may be canceled by adding another term to eq. (15.37):

$$\frac{d}{dt}\rho = \frac{1}{i\hbar}[H, \rho] - \frac{1}{2}(\Gamma\rho + \rho\Gamma) + \Gamma^{1/2}\rho\Gamma^{1/2}. \quad (15.39)$$

Equation (15.39) is a simple example of a rich class of equations of motion of so-called *Lindblad* form:

$$\frac{d}{dt}\rho = \frac{1}{i\hbar}[H, \rho] - \frac{1}{2}\sum_i \{V_i^+ V_i, \rho\} + \sum_i V_i \rho V_i^+, \quad (15.40)$$

which are known to be the most general linear equations of motion for ρ preserving Hermiticity and positivity and keeping $\rho = 1$ fixed.

15.2 Simple laser theory

Lasers are a crucial part of any experiment in quantum optics. It was only with the advent of the laser that quantum optics could start to develop. The functional principle of the laser is easily described.

Light is emitted by the transition from an excited level with energy E_3 to another level with energy E_2 . The light-emitting medium is placed in an optical resonator such that the emitted light triggers further stimulated transitions from E_3 to E_2 . If the resonator is correctly tuned, there will be feedback, and the medium will emit light of strong intensity in a coherent way. In order to maintain stationary conditions, the upper level E_3 has to be continuously repopulated by a process of *optical pumping*: by absorption of pumping energy from a third level E_1 , direct or indirect transitions are induced, which stabilize the occupation number of the level E_3 .

The details of a quantitative description are complicated and strongly dependent on the actual type of laser. Indeed, laser theory and technology have developed into a wide and important field of its own. Here we shall content ourselves with the description by simple rate equations, which allow the understanding of the laser principle under very general conditions. In our presentation, we follow H. A. Bachor. For details we refer the reader to the rich literature on lasers – for instance, the comprehensive monograph of A. E. Siegman.

Consider M atoms with three levels. The numbers of atoms in the states E_1 , E_2 , and E_3 are denoted by M_1 , M_2 , and M_3 , respectively. We refer to $J_1 = M_1/M$, $J_2 = M_2/M$, and $J_3 = M_3/M$ as the relative occupation numbers. Let $n = n_\gamma/M$ be the number of photons with energy $E_3 - E_2$ per atom. The strength of the pumping process and the rate of induced transitions between E_3 and E_2 are characterized by positive constants Π and G , respectively. Finally, we introduce positive constants γ_{32} and γ_{21} describing the rate of spontaneous transitions from E_3 to E_2 and from E_2 to E_1 , respectively, and a positive constant κ taking account of photon losses from absorption and from leaving the resonator.

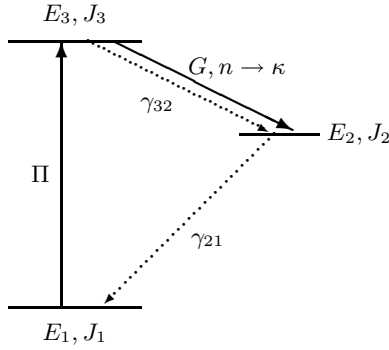


Figure 15.1: The electronic levels taking part in laser action.

The validity of the following rate equations is immediately clear from fig. (15.1):

$$\begin{aligned}
 \frac{dJ_1}{dt} &= -\Pi J_1 + \gamma_{21} J_2, \\
 \frac{dJ_2}{dt} &= G(J_3 - J_2)n + \gamma_{32} J_3 - \gamma_{21} J_2, \\
 \frac{dJ_3}{dt} &= -G(J_3 - J_2)n - \gamma_{32} J_3 + \Pi J_1, \\
 \frac{dn}{dt} &= G(J_3 - J_2)n - \kappa n.
 \end{aligned}
 \tag{15.41}$$

Equations (15.41) describe a nonlinear dynamical system for the quantities $J_1(t)$, $J_2(t)$, $J_3(t)$, and $n(t)$. The conservation law

$$\frac{d}{dt}(J_1 + J_2 + J_3) = 0
 \tag{15.42}$$

is a consequence of eqs. (15.41), and the constraint

$$J_1 + J_2 + J_3 = 1
 \tag{15.43}$$

can be consistently imposed. Stationary solutions are found by equating the left-hand sides of eqs. (15.41) to zero and solving for J_1 , J_2 , J_3 , and n . Two solutions are found, one with $n \equiv 0$ and one with

$$n = \frac{\Pi(\gamma_{21} - \gamma_{32})}{\kappa(2\Pi + \gamma_{21})} - \frac{1}{G} \frac{\Pi(\gamma_{21} + \gamma_{32}) + \gamma_{21}\gamma_{32}}{(2\Pi + \gamma_{21})}.
 \tag{15.44}$$

The second solution, which corresponds to continuous operation of the laser, is only physical if $n > 0$, which requires $\gamma_{21} > \gamma_{32}$.

In fact, $\gamma_{21} > \gamma_{32}$ is necessary for spontaneous depletion of the state E_2 , without which stimulated transitions from E_3 to E_2 would rapidly come to an end. In addition, for a positive stationary value of n , the pumping strength Π must be larger than a critical value Π_0 . If the pumping rate were too low, the occupation number of E_3 would be too small for a steady operation of the laser. This threshold behavior of Π is evident from eq. (15.44), because for very small Π the negative term on the right-hand side will dominate.

The last equation of (15.41) tells us that stationary light emission $n > 0$ requires an occupation number inversion $J_3 > J_1$. Contrary to the thermal equilibrium distribution, the occupation number of the higher state E_3 must exceed the occupation number of E_1 .

Depending on the relative values for the spontaneous lifetime γ_{32} and the loss rate κ , one can distinguish three types of lasers:

- Type 1: $\gamma_{32} \gg \kappa$, e.g. dye lasers;
- Type 2: $\gamma_{32} \approx \kappa$, e.g. visible gas lasers;
- Type 3: $\gamma_{32} \ll \kappa$, e.g. solid-state and semiconductor lasers.

A complete quantum-mechanical treatment of the laser is complicated. Two main approaches are possible.

In the first alternative, one considers a generalized quantum Liouville equation of the type eq. (15.37), where the density operator ρ may also depend on space and describes a quantum system consisting of a radiation field plus atoms or molecules with at least three levels. Suitable stochastic extra terms in the Hamiltonian may be introduced to describe the influence of external noise perturbations.

Alternatively, one may consider a Heisenberg equation of motion for the occupation number operators of the various levels and for the photon field number operators. If additional external noise terms are introduced, the dynamical equation to be solved will be a Heisenberg–Langevin equation.

A quantum-mechanical treatment will yield the fluctuations for the observables of the laser and will allow for a precise determination of the quantum state of the emitted light field. Not surprisingly, sufficiently far above the pumping threshold Π_0 , this state turns out to be very close to a coherent state, because the stimulated emission of the atoms in the laser is well correlated under these conditions.

15.3 Three-level systems and atomic interference

In this section, we shall concentrate on the interesting effects of the coherent superposition of atomic eigenstates on transitions mediated by the electromagnetic field. The minimal arrangement to study such phenomena is an atomic three-level system with two possible electromagnetic transitions. There are three different configurations, which are depicted in fig. 15.2: a Λ configuration with one upper level and two nearly degenerate lower levels; a V configuration with one lower level and two nearly degenerate upper levels; and a cascade configuration.

Consider the spontaneous decay of a system in the V configuration prepared in a superposition state

$$|\psi\rangle = c_1 e^{-i\omega_1 t} |\varphi_1\rangle + c_2 e^{-i\omega_2 t} |\varphi_2\rangle. \quad (15.45)$$

The transition amplitude to $|\varphi_3\rangle$ will be a superposition of two terms with different frequencies, and the occupation probability $w_3(t)$ of the state $|\varphi_3\rangle$ will contain a beat mode term with an oscillation in the difference frequency $\omega_1 - \omega_2$. The same oscillations can be observed in the counting rate of photons emerging from the transition. This typically quantum interference

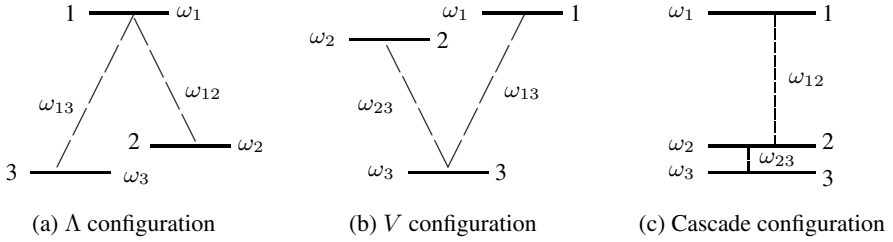


Figure 15.2: The level configurations for three-level systems.

phenomenon is completely analogous to the interference pattern in the quantum-mechanical double-slit experiment: two possible pathways through two different holes interfere to give the typical pattern. The interference disappears if “which path” information is available telling us through which hole the particle reached the screen. In the same way, the beat mode in the transition will disappear if there is information about the excited state from which the ground state was reached, for instance by a precision measurement of the energy of the outgoing photons.

Quantum beats were first observed in the so-called *Hanle effect*, where a coherent superposition of two states with different magnetic quantum numbers is produced by interaction with an electromagnetic field transverse to the original quantization axis. Interference phenomena of transition amplitudes are also visible in induced transitions. By forcing the system into a suitable superposition state, it is even possible to suppress transitions completely by destructive interference of transition amplitudes. In this way, absorption can be blocked for Λ configurations and emission for V configurations. The corresponding superposition state is also called a *dark state* or a *trapped state*.

Such phenomena can be discussed in a simple exactly solvable model. Consider the semi-classical interaction picture Hamiltonian $H_{I,sc}$ of eq. (15.22) with a three-level system and two interaction terms with an external classical light field that is tuned to resonance such that the detunings vanish. Then $H_{I,sc}$ is of the form

$$\begin{aligned}
 H_{I,sc} &= \hbar(\Omega_2\sigma_{12} + \Omega_3\sigma_{13} + \Omega_2^*\sigma_{21} + \Omega_3^*\sigma_{31}) \\
 &= \hbar \begin{pmatrix} 0 & \Omega_2 & \Omega_3 \\ \Omega_2^* & 0 & 0 \\ \Omega_3^* & 0 & 0 \end{pmatrix}, \tag{15.46}
 \end{aligned}$$

where, this time, we have not insisted on a special ordering of the atomic levels and

$$\Omega_2 = (g_{12}\alpha_{12})^{(*)}, \quad \Omega_3 = (g_{13}\alpha_{13})^{(*)}, \tag{15.47}$$

are the complex Rabi frequencies of their complex conjugates. They can be adjusted by changing the phases and intensities of the external radiation fields.

The matrix (15.46) is the matrix representation with respect to the energy eigenstates

$$|\varphi_1\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad |\varphi_2\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad |\varphi_3\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{15.48}$$

of the unperturbed atomic Hamiltonian

$$H_0 = \hbar \begin{pmatrix} \omega_1 & 0 & 0 \\ 0 & \omega_2 & 0 \\ 0 & 0 & \omega_3 \end{pmatrix}.$$

The eigenvalues λ_i and eigenstates $|\psi_i\rangle$ of eq. (15.46) are readily determined:

$$\begin{aligned} \lambda_1 &= 0, \quad |\psi_1\rangle = \frac{1}{\Omega} \begin{pmatrix} 0 \\ \Omega_3 \\ -\Omega_2 \end{pmatrix}, \\ \lambda_2 &= \Omega, \quad |\psi_2\rangle = \frac{1}{\Omega\sqrt{2}} \begin{pmatrix} \Omega \\ \Omega_2^* \\ \Omega_3^* \end{pmatrix}, \\ \lambda_3 &= -\Omega, \quad |\psi_3\rangle = \frac{1}{\Omega\sqrt{2}} \begin{pmatrix} -\Omega \\ \Omega_2^* \\ \Omega_3^* \end{pmatrix}, \end{aligned} \quad (15.49)$$

with

$$\Omega = \sqrt{|\Omega_2|^2 + |\Omega_3|^2}. \quad (15.50)$$

Then, with respect to the basis $|\psi_1\rangle, |\psi_2\rangle, |\psi_3\rangle$, the time evolution operator

$$U_I(t) = e^{-(i/\hbar)H_I t}$$

is given by the matrix:

$$U_I(t) \simeq \begin{pmatrix} 1 & 0 & 0 \\ 0 & e^{-i\Omega t} & 0 \\ 0 & 0 & e^{i\Omega t} \end{pmatrix}, \quad (15.51)$$

and with respect to the basis $|\varphi_1\rangle, |\varphi_2\rangle, |\varphi_3\rangle$, we have

$$U_I(t) \simeq \frac{1}{\Omega^2} \begin{pmatrix} \Omega^2 \cos(\Omega t) & -i\Omega\Omega_2 \sin(\Omega t) & -i\Omega\Omega_3 \sin(\Omega t) \\ -i\Omega\Omega_2^* \sin(\Omega t) & |\Omega_3|^2 + |\Omega_2|^2 \cos(\Omega t) & \Omega_2^* \Omega_3 [\cos(\Omega t) - 1] \\ -i\Omega\Omega_3^* \sin(\Omega t) & \Omega_2 \Omega_3^* [\cos(\Omega t) - 1] & |\Omega_2|^2 + |\Omega_3|^2 \cos(\Omega t) \end{pmatrix}. \quad (15.52)$$

In our model, the trapped state without transitions to other states is just given by

$$|\psi_1\rangle = \frac{1}{\Omega} \begin{pmatrix} 0 \\ \Omega_3 \\ -\Omega_2 \end{pmatrix}.$$

Let us mention two different methods to prepare the trapped state $|\psi_1\rangle$.

1. *Adiabatic population transfer.* One starts with $\Omega_3 = 0$; then the trapped state is given by $|\psi_1(\Omega_3=0)\rangle = -|\varphi_3\rangle$. The system is prepared in the eigenstate $|\varphi_3\rangle$ of H_0 , which can be easily achieved. Now one slowly switches on Ω_3 by increasing the resonant external field with frequency ω_{13} . According to the adiabatic theorem, the system will remain in the eigenstate $|\psi_1(\Omega_3)\rangle$ of $H_I(\Omega_3)$ with eigenvalue $\lambda_1 = 0$, which is always a trapped state corresponding to the instantaneous value of Ω_3 . (If one subsequently slowly switches off Ω_2 , one will have achieved an adiabatic transition of the state from $|\varphi_3\rangle$ to $|\varphi_2\rangle$.)
2. *Combination of pumping and decay.* A pair of atomic levels can also be forced into a trapped state by optical pumping. The difference frequency of the two levels is irradiated with high intensity. The emerging coherent superposition of the two states can also be described as a superposition of the trapped state and a state orthogonal to it. The trapped component will survive because of its stability, whereas the orthogonal component will decay.

An ensemble of atoms in a trapped state is called *phasonium*, a substance with many surprising properties.

In what follows, we shall describe several phenomena originating from the coherent superposition of atomic states in as simple a way as possible. A more detailed description based on an equation of motion for the density operator ρ similar to eq. (15.37) can be rather complicated. For details we refer the interested reader to more comprehensive monographs like Scully and Zubairy.

15.3.1 Electromagnetically induced transparency

This is the effect of preventing absorption by forcing the system into a trapped state from which or into which transitions are blocked. Then the so prepared phasonium will be transparent for the frequency that belongs to the blocked transition. In our simple model of eq. (15.46), this effect can be understood in the following way.

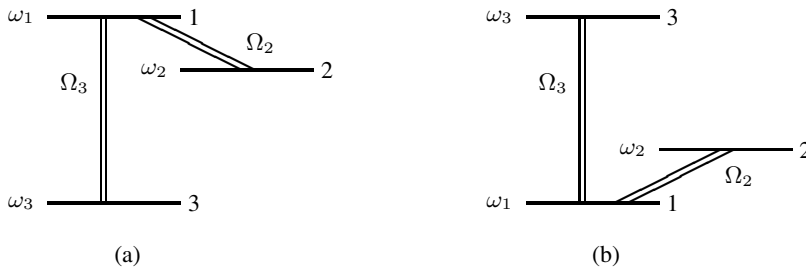


Figure 15.3: Configurations for electromagnetically induced transparency.

We assume the system to be in one of the configurations depicted in fig. 15.3. We apply strong pumping between the levels ω_1 and ω_2 , i.e. irradiation of high-intensity light with frequency $|\omega_1 - \omega_2|$. In our model, this corresponds to the situation $|\Omega_2| \gg |\Omega_3|$.

From eqs. (15.49) and (15.50) we see that in the limit $\Omega_3/\Omega_2 \rightarrow 0$ the eigenstates of H_I are given by

$$|\psi_1\rangle = \begin{pmatrix} 0 \\ 0 \\ -\Omega_2/|\Omega_2| \end{pmatrix}, \quad |\psi_2\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ \Omega_2^*/|\Omega_2| \\ 0 \end{pmatrix}, \quad |\psi_3\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ \Omega_2^*/|\Omega_2| \\ 0 \end{pmatrix}. \quad (15.53)$$

This means that there are no transitions between $|\varphi_3\rangle$ on the one hand and any linear combination of $|\varphi_1\rangle$ and $|\varphi_2\rangle$ on the other. Indeed, in the same limit we have

$$U_I(t) = \begin{pmatrix} \cos(|\Omega_2|t) & -i(\Omega_2/|\Omega_2|) \sin(|\Omega_2|t) & 0 \\ -i(\Omega_2^*/|\Omega_2|) \sin(|\Omega_2|t) & \cos(|\Omega_2|t) & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (15.54)$$

For small but finite values of Ω_3/Ω_2 , we can easily evaluate the transition rates:

- Consider first the situation of fig. 15.3(a). Assume the system to be in the state $|\varphi_3\rangle$. Then the occupation probabilities of the levels $|\varphi_1\rangle$ and $|\varphi_2\rangle$ are given by eq. (15.52)

$$w_{13}(t) = |\langle\varphi_1|U_I(t)|\varphi_3\rangle|^2 = \left|\frac{\Omega_3}{\Omega}\right|^2 \sin^2(\Omega t) \ll 1, \quad (15.55)$$

$$w_{23}(t) = |\langle\varphi_2|U_I(t)|\varphi_3\rangle|^2 = \frac{|\Omega_2\Omega_3|^2}{|\Omega|^4} [\cos(\Omega t) - 1]^2 \ll 1.$$

- For the configuration of fig. 15.3(b) we assume that the system is initially in the state $|\varphi_1\rangle$ or $|\varphi_2\rangle$ (or any superposition of the two). Then we readily find

$$w_{31}(t) = w_{13}(t), \quad w_{32}(t) = w_{23}(t),$$

with $w_{13}(t)$ and $w_{23}(t)$ as in eq. (15.55).

Hence, in any case, absorption from the lower state(s) is suppressed. To get an intuitive feeling for this result, we notice that, for small t , transitions between $|\varphi_1\rangle$ and $|\varphi_3\rangle$, irrespective of the values of Ω_2 and Ω_3 , will have a small probability

$$w_{13}(t) = w_{31}(t) \sim t^2.$$

Direct transitions between $|\varphi_2\rangle$ and $|\varphi_3\rangle$ are impossible in our model. They have to go via the state $|\varphi_1\rangle$. Correspondingly, eq. (15.52) tells us that, for small t , the transition probability is even smaller:

$$w_{23}(t) = w_{32}(t) \sim t^4.$$

Now, due to strong pumping, there are fast oscillations of the system between the states $|\varphi_1\rangle$ and $|\varphi_2\rangle$. As a consequence, the state will have already switched between $|\varphi_1\rangle$ and $|\varphi_2\rangle$ before any transition of either of these two states to the state $|\varphi_3\rangle$ can have occurred with sizable probability.

15.3.2 Refractive index enhancement

In chapter 2, we saw that absorption is determined by the imaginary part of the refractive index and that at resonance the refractive index is purely imaginary. A little bit off resonance, both the real and the imaginary parts are large. It is now possible to suppress absorption by means of trapped states and at the same time to maintain a large real part of the refractive index. This can be achieved, for example, by a weak additional occupation of the excited level.

In this way, phasonium may be transparent and show very large refraction at the same time.

15.3.3 Lasing without inversion

In section 15.2, we saw that a lasing transition with positive gain requires population inversion: the occupation of the higher level must dominate the lower level. Otherwise, the resonant light field would induce more absorption from the lower level than emission from the upper level.

Now, if the lower level is replaced by a doublet of states forced into a trapped state, absorption may be blocked and lasing will be possible without inversion. In our simple model, we consider the system to be in the Λ configuration of fig. 15.2(a). We assume the initial state to be given by

$$|\psi(0)\rangle = c_1|\psi_1\rangle + c_2|\varphi_1\rangle = c_1|\psi_1\rangle + \frac{c_2}{\sqrt{2}}(|\psi_2\rangle - |\psi_3\rangle) \quad (15.56)$$

(compare eq. (15.49)). For later time t , the state will be

$$|\psi(t)\rangle = c_1|\psi_1\rangle + \frac{c_2}{\sqrt{2}}(|\psi_2\rangle e^{-i\Omega t} - |\psi_3\rangle e^{i\Omega t}). \quad (15.57)$$

The occupation probability for the upper state $|\varphi_1\rangle$ will be

$$w_1(t) = |\langle\varphi_1|\psi(t)\rangle|^2 = \frac{1}{2}|c_2|^2|e^{-i\Omega t} + e^{i\Omega t}|^2 = 2|c_2|^2 \cos^2(\Omega t). \quad (15.58)$$

For small t , this is a decreasing function of t , which means lasing with a positive gain irrespective of the values of c_1 and c_2 . Now, for $|c_1| \gg |c_2|$, the lower levels are much more populated than the upper one, and we have lasing without inversion.

15.3.4 Correlated emission laser

This is another application of coherent superposition that we would just like to mention. Spontaneous transitions from the upper to the lower of the two lasing levels usually occur randomly and uncorrelated. They are a source of noise for the emerging laser light, which is present even if other sources, for instance due to noisy pumping, have been suppressed. If now the upper level of a laser is replaced by a doublet of states prepared in a coherent superposition, spontaneous emission will be correlated just like induced emission, and noise will be further reduced. This device is, for instance, applied in gravitational detectors.

15.4 The Jaynes–Cummings model

The Jaynes–Cummings Hamiltonian (eq. (15.25)),

$$\begin{aligned} H &= \hbar\Omega a^\dagger a + \frac{1}{2}\hbar\omega\sigma_3 + \hbar(g\sigma_+ a + g^*\sigma_- a^\dagger) \\ &= H_0 + \hbar(g\sigma_+ a + g^*\sigma_- a^\dagger), \end{aligned} \quad (15.59)$$

describes an extremely instructive model system. On the one hand, it captures many characteristic physical features of systems with interacting matter and radiation; on the other, the associated quantum-mechanical problem can be solved in closed form. With the notation of section 15.1, we can write the eigenstates of the free Hamiltonian H_0 as

$$|n, i\rangle = |n\rangle \otimes |\varphi_i\rangle \quad (i = 1, 2) \quad (15.60)$$

with

$$\begin{aligned} H_0|n, 2\rangle &= (\hbar\Omega n - \frac{1}{2}\hbar\omega)|n, 2\rangle, \\ H_0|n, 1\rangle &= (\hbar\Omega n + \frac{1}{2}\hbar\omega)|n, 1\rangle. \end{aligned} \quad (15.61)$$

The quantity

$$C_1 := a^\dagger a + \frac{1}{2}\sigma_3 \quad (15.62)$$

is conserved because excitation is associated with absorption of a photon and de-excitation with emission of a photon. C_1 can be interpreted as the total excitation degree of atom plus photons. One readily verifies that

$$[H, C_1] = 0. \quad (15.63)$$

The states $|n, i\rangle$ are eigenstates of C_1 :

$$\begin{aligned} C_1|n, 1\rangle &= (n + \frac{1}{2})|n, 1\rangle, \\ C_1|n, 2\rangle &= (n - \frac{1}{2})|n, 2\rangle. \end{aligned} \quad (15.64)$$

The eigenvalue $(n - \frac{1}{2})$ is two-fold degenerate for $n \geq 1$:

$$\begin{aligned} C_1|n, 2\rangle &= (n - \frac{1}{2})|n, 2\rangle, \\ C_1|n-1, 1\rangle &= (n - \frac{1}{2})|n-1, 1\rangle, \end{aligned} \quad (15.65)$$

whereas for $n = 0$ the only eigenstate $|0, 2\rangle$ with

$$C_1|0, 2\rangle = -\frac{1}{2}|0, 2\rangle \quad (15.66)$$

is the unique ground state of H :

$$H|0, 2\rangle = -\frac{1}{2}\hbar\omega|0, 2\rangle. \quad (15.67)$$

Another conserved quantity is

$$C_2 := \frac{1}{2}\Delta\sigma_3 + g\sigma_+a + g^*\sigma_-a^+, \quad \Delta = \omega - \Omega, \quad (15.68)$$

but there is a relation between H , C_1 , and C_2 :

$$H = \hbar\Omega C_1 + \hbar C_2. \quad (15.69)$$

The conservation of C_1 facilitates the diagonalization of H . Eigenstates of H can be assumed to be eigenstates of C_1 . In the two-dimensional eigenspace of C_1 that belongs to the eigenvalue $n - \frac{1}{2}$ ($n > 0$), the eigenstates of H are of the form

$$|\psi_n\rangle = b_{n-1,1}|n-1, 1\rangle + b_{n,2}|n, 2\rangle. \quad (15.70)$$

The Hamiltonian (15.59), restricted to this eigenspace of C_1 , is given by the matrix

$$\hbar \begin{pmatrix} (n-1)\Omega + \frac{1}{2}\omega & \sqrt{n}g \\ \sqrt{n}g^* & n\Omega - \frac{1}{2}\omega \end{pmatrix} = \hbar(n - \frac{1}{2})\Omega \mathbf{1} + \hbar \begin{pmatrix} \frac{1}{2}\Delta & \sqrt{n}g \\ \sqrt{n}g^* & -\frac{1}{2}\Delta \end{pmatrix}, \quad (15.71)$$

which can easily be diagonalized to give the eigenvalues and eigenvectors of H . We find

$$|\psi_{n\pm}\rangle = b_{n-1,1}^{\pm}|n-1, 1\rangle + b_{n,2}^{\pm}|n, 2\rangle, \quad (15.72)$$

with

$$\begin{aligned} b_{n,2}^{\pm} &= \frac{-1}{N_{\pm}}(\frac{1}{2}\Delta \mp \Lambda_n), & b_{n-1,1}^{\pm} &= \frac{1}{N_{\pm}}\sqrt{n}g, \\ \Lambda_n &:= \sqrt{\frac{1}{4}\Delta^2 + n|g|^2}, \\ N_{\pm}^2 &:= 2\Lambda_n^2 \mp \Delta\Lambda_n, \end{aligned} \quad (15.73)$$

and

$$H|\psi_{n\pm}\rangle = \hbar\Omega_{n\pm}|\psi_{n\pm}\rangle \quad (15.74)$$

with

$$\Omega_{n\pm} = (n - \frac{1}{2})\Omega \pm \Lambda_n. \quad (15.75)$$

For $g \neq 0$, this spectrum differs from the spectrum in eq. (15.61) of H_0 , a phenomenon known as the *dynamical Stark effect*. From these results, one immediately calculates the time translation operator $U = \exp[-(i/\hbar)Ht]$.

In view of the importance of the Jaynes–Cummings model, we derive $U(t)$ also in another way, which gives us additional insight. Just as in the semiclassical case of section 15.1, the interaction Hamiltonian H_I of eq. (15.26) can be written as

$$\begin{aligned} H_I &= \hbar(g\sigma_+a e^{i\Delta t} + g^*\sigma_-a^+ e^{-i\Delta t}) \\ &= \hbar e^{+\frac{1}{2}i\Delta\sigma_3 t}(g\sigma_+a + g^*\sigma_-a^+) e^{-\frac{1}{2}i\Delta\sigma_3 t} \end{aligned} \quad (15.76)$$

(compare eq. (15.24)). In view of eq. (15.16), eq. (15.76) means that H_I is also the interaction Hamiltonian for the Hamiltonian

$$\hat{H} = \frac{1}{2}\hbar\Delta\sigma_3 + \hbar(g\sigma_+a + g^*\sigma_-a^+) = \hbar C_2 = \hbar \begin{pmatrix} \frac{1}{2}\Delta & ga \\ g^*a^+ & -\frac{1}{2}\Delta \end{pmatrix}, \quad (15.77)$$

such that

$$U_I(t) = e^{\frac{1}{2}i\Delta\sigma_3 t} e^{-(i/\hbar)\hat{H}t} \quad (15.78)$$

and

$$U(t) = U_0(t)U_I(t) = e^{-i\Omega(a^+a + \frac{1}{2}\sigma_3)t} e^{-(i/\hbar)\hat{H}t} = e^{-i\Omega C_1 t} e^{-(i/\hbar)\hat{H}t}. \quad (15.79)$$

Here $\hat{H}$ is time-independent and $e^{-(i/\hbar)\hat{H}t}$ can be evaluated in the following way. First we have, using eq. (15.24):

$$\begin{aligned} \left(\frac{1}{\hbar}\hat{H}\right)^2 &= \frac{1}{4}\Delta^2\mathbf{1} + |g|^2(\sigma_+\sigma_-aa^+ + \sigma_-\sigma_+a^+a) \\ &= \begin{pmatrix} \frac{1}{4}\Delta^2 + |g|^2aa^+ & 0 \\ 0 & \frac{1}{4}\Delta^2 + |g|^2a^+a \end{pmatrix} =: \begin{pmatrix} \hat{\Lambda}_+^2 & 0 \\ 0 & \hat{\Lambda}_-^2 \end{pmatrix} =: \hat{\Lambda}^2, \end{aligned} \quad (15.80)$$

where the operators

$$\begin{aligned} \hat{\Lambda}_+ &= \sqrt{\frac{1}{4}\Delta^2 + |g|^2aa^+}, \\ \hat{\Lambda}_- &= \sqrt{\frac{1}{4}\Delta^2 + |g|^2a^+a} \end{aligned} \quad (15.81)$$

satisfy

$$\begin{aligned} \hat{\Lambda}_+|n\rangle &= \sqrt{\frac{1}{4}\Delta^2 + |g|^2(n+1)}|n\rangle =: \Lambda_{n+1}|n\rangle, \\ \hat{\Lambda}_-|n\rangle &= \sqrt{\frac{1}{4}\Delta^2 + |g|^2n}|n\rangle =: \Lambda_n|n\rangle. \end{aligned} \quad (15.82)$$

Then

$$\begin{aligned} \left(\frac{1}{\hbar}\hat{H}\right)^{2n} &= \hat{\Lambda}^{2n} = \begin{pmatrix} \hat{\Lambda}_+^{2n} & 0 \\ 0 & \hat{\Lambda}_-^{2n} \end{pmatrix}, \\ \left(\frac{1}{\hbar}\hat{H}\right)^{2n+1} &= \frac{1}{\hbar}\hat{H}\hat{\Lambda}^{2n}, \end{aligned}$$

and with eq. (15.77)

$$U_I(t) = \begin{pmatrix} \left[\cos(\hat{\Lambda}_+t) - i\frac{\Delta}{2}\frac{1}{\hat{\Lambda}_+}\sin(\hat{\Lambda}_+t) \right] e^{\frac{1}{2}i\Delta t} & -iga\frac{1}{\hat{\Lambda}_-}\sin(\hat{\Lambda}_-t)e^{\frac{1}{2}i\Delta t} \\ -ig^*a^+\frac{1}{\hat{\Lambda}_+}\sin(\hat{\Lambda}_+t)e^{-\frac{1}{2}i\Delta t} & \left[\cos(\hat{\Lambda}_-t) + i\frac{\Delta}{2}\frac{1}{\hat{\Lambda}_-}\sin(\hat{\Lambda}_-t) \right] e^{-\frac{1}{2}i\Delta t} \end{pmatrix}. \quad (15.83)$$

As a first check, we verify the stability of the ground state

$$|0, 2\rangle = |0\rangle \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

i.e.

$$U_I(t)|0, 2\rangle = [\cos(\tfrac{1}{2}\Delta t) + i\sin(\tfrac{1}{2}\Delta t)]e^{-\frac{1}{2}i\Delta t}|0, 2\rangle = |0, 2\rangle. \quad (15.84)$$

Writing the state at time t as

$$|\psi(t)\rangle = \sum_{n=0}^{\infty} [c_{n-1,1}(t)|n-1, 1\rangle + c_{n,2}(t)|n, 2\rangle], \quad (15.85)$$

with $c_{-1,1} = 0$ by definition, we find that eq. (15.83) is equivalent to

$$\begin{pmatrix} c_{n-1,1}(t) \\ c_{n,2}(t) \end{pmatrix} = \begin{pmatrix} a & b \\ -b^* & a^* \end{pmatrix} \begin{pmatrix} c_{n-1,1}(0) \\ c_{n,2}(0) \end{pmatrix}, \quad (15.86)$$

with

$$\begin{aligned} a &= [\cos(\Lambda_n t) - i(\Delta/2\Lambda_n) \sin(\Lambda_n t)] e^{\frac{1}{2}i\Delta t}, \\ b &= -ig\sqrt{n} (1/\Lambda_n) \sin(\Lambda_n t) e^{\frac{1}{2}i\Delta t}. \end{aligned}$$

and where Λ_n is defined in eq. (15.73).

One immediate consequence of eq. (15.84) or (15.86) is *spontaneous emission*. The state $|0, 1\rangle$ consisting of no photon and the atom in the excited state is not stable. Rather, for $c_{0,1}(0) = 1$ and $c_{1,2}(0) = 0$,

$$\begin{aligned} c_{0,1}(t) &= [\cos(\Lambda_1 t) - i(\Delta/2\Lambda_1) \sin(\Lambda_1 t)] e^{\frac{1}{2}i\Delta t}, \\ c_{1,2}(t) &= -ig^*(1/\Lambda_1) \sin(\Lambda_1 t) e^{-\frac{1}{2}i\Delta t}. \end{aligned} \quad (15.87)$$

With eq. (15.86), it is an easy task to calculate the occupation probabilities from the initial values:

$$w_{n,1}(t) = |c_{n,1}(t)|^2 \quad \text{and} \quad w_{n,2}(t) = |c_{n,2}(t)|^2. \quad (15.88)$$

The probabilities for the occupation of the atomic states are then

$$w_1(t) = \sum_{n=0}^{\infty} w_{n,1}(t) \quad \text{and} \quad w_2(t) = \sum_{n=0}^{\infty} w_{n,2}(t), \quad (15.89)$$

whereas the probability for the presence of n photons is

$$w_n^p(t) = w_{n,1}(t) + w_{n,2}(t). \quad (15.90)$$

To simplify our expressions, we assume that at time $t = 0$ the atom is in the excited state; hence $w_{n,2}(0) = 0$. Then

$$\begin{aligned} w_{n,1}(t) &= \left(\cos^2(\Lambda_{n+1}t) + \frac{\Delta^2}{4\Lambda_{n+1}^2} \sin^2(\Lambda_{n+1}t) \right) w_{n,1}(0), \\ w_{n,2}(t) &= \frac{|g|^2 n}{\Lambda_n^2} \sin^2(\Lambda_n t) w_{n-1,1}(0). \end{aligned} \quad (15.91)$$

One checks that in fact

$$w_1(t) + w_2(t) = 1, \quad (15.92)$$

and after a short calculation one obtains

$$\begin{aligned} W(t) &:= w_1(t) - w_2(t) = 1 - 2w_2(t) \\ &= 1 - 2 \sum_{n=0}^{\infty} \frac{|g|^2(n+1)}{\Lambda_{n+1}^2} \sin^2(\Lambda_{n+1}t) w_{n,1}(0) \\ &= \sum_{n=0}^{\infty} \frac{1}{\Lambda_{n+1}^2} \left[\frac{1}{4}\Delta^2 + |g|^2(n+1) \cos(2\Lambda_{n+1}t) \right] w_{n,1}(0). \end{aligned} \quad (15.93)$$

The quantity $W(t) = w_1(t) - w_2(t)$ is called *occupation inversion* (see fig. 15.4).

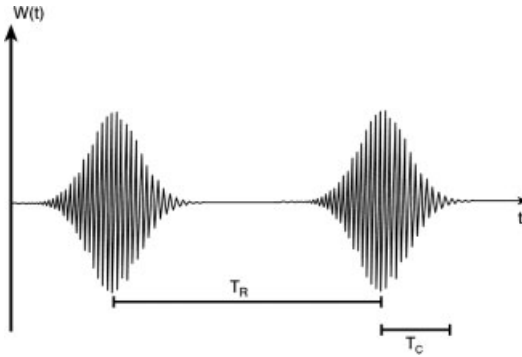


Figure 15.4: Occupation inversion as a function of time.

In the quasi-classical treatment of section 15.1, we found a periodic behavior for $W(t)$. In the Jaynes–Cummings model, $W(t)$ is only periodic if the initial distribution $w_{n,1}$ has a sharp photon number n_0 . If, on the other hand, $w_{n,1}$ is a distribution with mean value $\langle N \rangle$ and variance σ_N , the time behavior of $W(t)$ will be totally different, and will coincide with what is actually observed.

First, $W(0) = 1$. Now, $W(t)$ contains many periodic components with different frequencies $2\Lambda_{n+1}$. The fast oscillations with typical frequency $2\Lambda_{\langle N \rangle + 1}$ will fall out of phase after some time has elapsed. After a *collapse time* T_C , the inversion will be very small due to destructive interference of the different periodic modes. The uncertainty relation between time

and frequency allows one to estimate the collapse time T_C as

$$T_C = \frac{1}{2(\Lambda_{\langle N \rangle + 1 + \sigma_N} - \Lambda_{\langle N \rangle + 1 - \sigma_N})} \approx \frac{1}{4\sigma_N (\partial \Lambda_n / \partial n)|_{\langle N \rangle + 1}}. \quad (15.94)$$

After an even longer time T_R , the periodic oscillations, which after all have discrete frequencies, will be in phase again, and the inversion will again assume a larger value. This phenomenon is called *inversion revival*. Collapses and revivals will alternate periodically. Revival is a purely quantum-mechanical effect of the discreteness of the photon number. A good estimate of the revival time is

$$2T_R(\Lambda_{\langle N \rangle + 1} - \Lambda_{\langle N \rangle}) = 2T_R \left. \frac{\partial \Lambda_n}{\partial n} \right|_{\langle N \rangle} = 2\pi. \quad (15.95)$$

With Λ_n from eq. (15.73)

$$\frac{\partial \Lambda_n}{\partial n} = \frac{|g|^2}{2\Lambda_n} \quad (15.96)$$

and

$$T_R = 2\pi \frac{\Lambda_{\langle N \rangle}}{|g|^2} = \frac{2\pi}{|g|} \sqrt{\frac{\Delta^2}{4|g|^2} + \langle N \rangle}, \quad (15.97)$$

$$T_C = \frac{\Lambda_{\langle N \rangle + 1}}{2|g|^2 \sigma_N} = \frac{\sqrt{(\Delta^2/4|g|^2) + \langle N \rangle} + 1}{2\sigma_N |g|}.$$

If the initial distribution is Poissonian with $\sigma_N = \langle N \rangle^{1/2}$, for $\langle N \rangle \gg 1$ one obtains

$$T_C = \sqrt{1 + \frac{\Delta^2}{4|g|^2 \langle N \rangle}} \frac{1}{2|g|}. \quad (15.98)$$

We conclude this section with a short discussion on the interesting limit of strong detuning:

$$\Delta \gg |g| \sqrt{\langle N \rangle}. \quad (15.99)$$

Then

$$\Lambda_n = \frac{\Delta}{2} + \frac{|g|^2 n}{\Delta} + O\left(\frac{1}{\Delta^2}\right).$$

A glance at eq. (15.83) or (15.86) shows that to leading order

$$\lim_{\Delta \rightarrow \infty} U_I(t) = \mathbf{1}. \quad (15.100)$$

Not surprisingly, transitions are suppressed for very strong detuning. In next-to-leading order, we can approximately insert into eq. (15.83):

$$\begin{aligned} \left(\cos(\hat{\Lambda}_+ t) - i \frac{\Delta}{2\hat{\Lambda}_+} \sin(\hat{\Lambda}_+ t) \right) e^{\frac{1}{2}i\Delta t} &\approx \exp\left(-i \frac{|g|^2}{\Delta} (a^+ a + 1)t\right), \\ \left(\cos(\hat{\Lambda}_- t) + i \frac{\Delta}{2\hat{\Lambda}_-} \sin(\hat{\Lambda}_- t) \right) e^{-\frac{1}{2}i\Delta t} &\approx \exp\left(i \frac{|g|^2}{\Delta} a^+ a t\right), \end{aligned} \quad (15.101)$$

whereas the off-diagonal elements are still negligible.

With these approximations, $U_I(t)$ becomes the time evolution operator for the effective Hamiltonian

$$H_{\text{eff}} = \hbar \frac{|g|^2}{\Delta} (\sigma_3 a^\dagger a + \sigma_{11}), \quad (15.102)$$

and $U_I(t)$ will produce slow variations of the relative phases of the excited states. This effect is exploited in a variety of experimental setups.

15.5 The micromaser

Micromasers and microlasers owe their existence to the spectacular progress in the construction of electromagnetic resonators. The decay time of the electromagnetic field is much longer than the interaction time with atoms inside the resonator, which means that photons can be absorbed and re-emitted many times before they get lost by dissipation.

The micromaser is the first and most advanced example of such a device. A microwave field in a cavity is tuned to be in resonance with a transition between two Rydberg states of the hydrogen atom. Rydberg states are excited states of the hydrogen atom with large principal quantum numbers. They are ideally suited for experimentation in this field for at least three reasons:

1. The transition frequencies between Rydberg states lie in the experimentally convenient microwave range. For instance, the transition frequency between the frequently employed levels $63 \text{ p}_{3/2}$ and $61 \text{ d}_{3/2}$ is 21.5 GHz.
2. The strength of the interaction with the electromagnetic field is proportional to the transition dipole moments, which are excessively large for Rydberg states. The diameter of an excited hydrogen atom and its transition dipole elements are proportional to n^2 , which leads to transition probabilities of the order of n^4 . A few photons will already suffice to saturate a transition.
3. Rydberg atoms can conveniently be prepared and detected by electric field ionization. If a homogeneous electric field $\mathbf{E}$ is applied to a hydrogen atom, then the Coulomb field, in which the electron of the atom is bound, is deformed and lowered in the direction $-\mathbf{E}$. As a result, atomic levels that are bound states without the field $\mathbf{E}$ become levels in the continuum, and formerly bound electrons are pulled out of the atom. If the field $\mathbf{E}$ is increased, weakly bound electrons with higher principal quantum numbers n are liberated first and more strongly bound electrons later. Alternatively, by properly adjusting the strength of $\mathbf{E}$, one can liberate electrons in states with $n > n_0$ for any given n_0 , whereas electrons with $n \leq n_0$ remain bound.

Working with a cavity in the microwave regime also brings about a technical difficulty: disturbing thermal photons have to be suppressed by cooling the device to very low temperatures. Inserting numbers into Planck's formula

$$\bar{n}_{\text{th}} = \frac{1}{e^{\hbar\omega/kT} - 1}, \quad (15.103)$$

one finds $\bar{n}_{\text{th}} = 1.5$ at $T = 2$ K and $\omega = 21.5$ GHz, still too high for ambitious purposes. Nowadays, $T = 0.3$ K and $\bar{n} = 0.054$ have been achieved.

The principle of a micromaser is simple. A ray of hydrogen atoms prepared in one of two Rydberg states is velocity-selected and injected into a cooled microwave cavity tuned to the transition frequency between the two atomic levels. Inside the cavity, the atoms can emit or absorb a photon by interacting with the electromagnetic field. The atoms pass through the cavity and, after they leave it, their state of excitation is measured by two electrostatic field ionization detectors, the first one tuned to the upper level, and the second one tuned to the lower level.

Long decay times of several seconds can be achieved for the field in the resonator, and the intensity for the atomic beam can be so low that on average less than one atom is present in the cavity.

If the atoms enter the cavity in the excited state and if the cavity field is initially in the ground state, then the atoms traversing the cavity will deposit their energy as photons in the cavity until, due to the inevitable losses, some saturation point is reached.

The device is a minimal realization of a laser (or rather maser) arrangement. An electromagnetic field is built up by stimulated emission of photons by a small number of excited atoms. As opposed to an ordinary laser, the radiation in the cavity cannot be coupled out, because this would spoil the quality of the resonator. The atoms traversing the cavity play a dual role, both producing and detecting the field in the cavity. The photon statistics in the cavity is very interesting and shows many non-classical features.

The theory of the micromaser uses the Jaynes–Cummings model described in the previous section. The velocity-selected atoms spend a controllable time τ in the cavity. During this time-span, the dynamics of the two-level system in interaction with the cavity field is well described by the Jaynes–Cummings Hamiltonian of eq. (15.25).

The equation of motion for the density operator is of Lindblad type and given by eq. (15.40), where H is the Jaynes–Cummings Hamiltonian, and dissipation by photon losses is accounted for by an additional term

$$L_\rho = -\frac{1}{2}c(\bar{n}_{\text{th}} + 1)(a^+a\rho + \rho a^+a - 2a\rho a^+) - \frac{1}{2}c\bar{n}_{\text{th}}(aa^+\rho + \rho aa^+ - 2a^+\rho a). \quad (15.104)$$

Phenomena like collapse and revival of the occupation inversion can be observed in the micromaser. In the next section, we shall describe how the micromaser may be used for “quantum state engineering”, the preparation of a given state of the electromagnetic field. Other applications for experiments on the foundations of quantum theory will be described in chapter 16.

Recently, also microlasers working in the regime of visible light have become feasible. Two different arrangements are currently in use. One of them takes profit from the fact that mirrors of extremely high reflectivity, formerly classified and reserved for military purposes, have become available. This allows the construction of optical resonators of sufficiently high quality for a microlaser. The difficulty of weaker coupling due to smaller transition dipole elements in the optical regime is overcome by making the spatial dimensions of the resonator very small. A glance at eqs. (13.42) and (15.10) will reveal that this increases the coupling strength. Fortunately, thermal photons are less of a problem in the optical regime and the requirements for cooling are very mild.

Another arrangement uses modes of the light field propagating along the inside of a glass sphere, which can be excited by a laser. They are called “whispering gallery modes” in analogy with the famous acoustic phenomenon in the gallery of St. Paul’s Cathedral in London.

15.6 Quantum state engineering

In this section, we shall describe techniques for the measurement and preparation of atoms and electromagnetic states. We concentrate on the state space of the Jaynes–Cummings model: a two-state atomic system and a one-mode electromagnetic field. Measurement and preparation are intimately related in quantum theory, because every measurement implies a projection on a state possessing the measured value with certainty.

Let us first turn to two-state atomic systems. The eigenstates $|\varphi_1\rangle$ and $|\varphi_2\rangle$ of the free Hamiltonian H_0 are easily prepared by an energy measurement. The problem is the preparation of an arbitrary superposition:

$$|\varphi\rangle = c_1|\varphi_1\rangle + c_2|\varphi_2\rangle, \quad |c_1|^2 + |c_2|^2 = 1. \quad (15.105)$$

This can be achieved by first preparing the system in the state $|\varphi_1\rangle$ and then applying an appropriate unitary transformation. Such a transformation can, for instance, be realized by the time evolution operator of eq. (15.34) of the quasi-classical Jaynes–Cummings model. In practice, the system initially in the state $|\varphi_1\rangle$ is brought into interaction with a classical external wave field for a controlled time interval t_0 . By adjusting t_0 and the parameters g , α and Δ properly, arbitrary values of c_1 and c_2 can be obtained. The relative phase between c_1 and c_2 can also be changed by allowing the system to evolve under the free Hamiltonian operator H_0 . Preparation and measurement of states of the electromagnetic field are more difficult because it is harder to exert controlled influences on them. In chapter 14, we saw how homodyne detection can be applied for state measurements.

The micromaser may be used for both measuring and preparing the field in the cavity. A given cavity field can be analyzed by injecting atoms into it, allowing them to interact with the field for a controlled time, and then measuring the state of the atoms. In the time dependence of the inversion, collapses and revivals can be seen, and eq. (15.93) teaches us that the photon number statistics can be read off from the Fourier transform of the inversion function $W(t)$.

Electric field states of the cavity with given photon number n can be prepared in the micromaser in the following way. Individual excited atoms are passed through the resonator, which is initially in its zero-photon ground state. The excitation level of the atoms leaving the resonator is measured, and, every time an atom leaves the cavity in a de-excited state, an additional photon has been deposited in the resonator. If photon losses by dissipation can be neglected, states of given photon number have been prepared.

The deposition of photons can be discussed quantitatively in the Jaynes–Cummings model. For zero detuning $\Delta = 0$, eq. (15.91) gives

$$w_{n,2}(t_0) = \sin^2(g\sqrt{n}t_0) \quad (15.106)$$

for the probability $w_{n,2}(t_0)$ that an excited atom raises the photon number in the oscillator from $n - 1$ to n within an interaction time t_0 .

Equation (15.106) shows that an n -photon state can be prepared even without measuring the excitation state of the outgoing atoms. One just has to fix t_0 such that

$$\sin[g\sqrt{(n+1)}t_0] = 0. \quad (15.107)$$

Then it is not possible to raise the photon number from n to $n+1$. If eq. (15.107) is fulfilled, one says that the maser is in an upwards trapping state. In this case, one just injects excited atoms into the micromaser with the cavity field initially in the ground state. Some atoms will increase the photon number, some of them will not, but the increase will stop when the photon number n is reached.

There are also downwards trapped states of the laser, in which a de-excited atom cannot pick up a photon. They provide another mechanism for number state preparation.

A more ambitious aim is the preparation of a state

$$|\psi\rangle = \sum_{n=0}^N d_n^{(N)} |n\rangle, \quad \sum_{n=0}^N |d_n^{(N)}|^2 = 1, \quad (15.108)$$

of the resonator field, which is an arbitrary superposition of number states up to photon number N .

The mechanism we are going to describe makes essential use of the fact that the eigenstates of the Jaynes–Cummings model as given by the equations following eq. (15.72) do not have sharp values of photon and atomic excitation number but show entanglement between atomic and electromagnetic quantum numbers. Moreover, a product state $|n, i\rangle$ ($n = 0, 1, \dots; i = 1, 2$) will evolve into an entangled state.

Consider now a general state

$$|\Psi\rangle = \sum_{i=1}^2 \sum_{n=0}^{\infty} c_{n,i} |n, i\rangle \quad (15.109)$$

in the Jaynes–Cummings model. If a measurement of the atomic state gives the result $|\varphi\rangle = \alpha_1|\varphi_1\rangle + \alpha_2|\varphi_2\rangle$, the state $|\Psi\rangle$ of the field–atom system is reduced to a state

$$|\psi\rangle = \sum_{n=0}^{\infty} (\alpha_1^* c_{n1} + \alpha_2^* c_{n2}) |n\rangle \quad (15.110)$$

of the cavity field. Now, $|\psi\rangle$ is a coherent superposition of photon number states. The strategy for photonic state preparation is now the following.

Let the cavity be initially in its ground state. An atom is injected in a superposition state $|\varphi\rangle^{(1)} = \alpha_1^{(1)}|\varphi_1\rangle + \alpha_2^{(1)}|\varphi_2\rangle$ and interacts with the cavity for a time t_1 . When the atom leaves the cavity, its excitation state is measured. When the atom is in the upper state, the cavity will still be in its ground state; when it is in the lower state, the cavity will be in a superposition of its ground state $|0\rangle$ and its first excited state $|1\rangle$. In the latter case, a second atom is injected in a different superposition state, interacts for a time t_2 , and is detected after leaving the cavity. If this second atom is found in the lower state, one may proceed; otherwise the run will be rejected and one starts again from the beginning.

Proceeding in this way, after N steps the state of the cavity will be a superposition of the first N excited states like eq. (15.108). By choosing the values t_i , $\alpha_1^{(i)}$, and $\alpha_2^{(i)}$ properly, one can give the coefficients $d_n^{(N)}$ in eq. (15.108) the desired values.

Let us follow these steps analytically by using eq. (15.86) for the case $\Delta = 0$ and $g > 0$. After time t , the states $|n, 1\rangle$ and $|n, 2\rangle$ will have evolved into

$$\begin{aligned} |n, 1\rangle &\rightarrow \cos [g\sqrt{(n+1)t}] |n, 1\rangle - i \sin [g\sqrt{(n+1)t}] |n+1, 2\rangle \\ &= C_{n+1}(t) |n, 1\rangle + S_{n+1}(t) |n+1, 2\rangle, \\ |n, 2\rangle &\rightarrow -i \sin [g\sqrt{n}t] |n-1, 1\rangle + \cos [g\sqrt{n}t] |n, 2\rangle \\ &= S_n(t) |n-1, 1\rangle + C_n(t) |n, 2\rangle. \end{aligned} \quad (15.111)$$

Now, in the first step we have

$$\alpha_1^{(1)}|0, 1\rangle + \alpha_2^{(1)}|0, 2\rangle \rightarrow \alpha_1^{(1)}C_1(t_1)|0, 1\rangle + \alpha_1^{(1)}S_1(t_1)|1, 2\rangle + \alpha_2^{(1)}C_0(t_1)|0, 2\rangle. \quad (15.112)$$

After measurement and projection onto $|\varphi_2\rangle$, this reduces to

$$|\psi\rangle^{(1)} = \alpha_1^{(1)}S_1(t_1)|1\rangle + \alpha_2^{(1)}|0\rangle = d_0^{(1)}|0\rangle + d_1^{(1)}|1\rangle. \quad (15.113)$$

One notes that, after normalization, arbitrary superpositions of $|0\rangle$ and $|1\rangle$ can be produced.

Let the cavity state after the $(N-1)$ th step be given by

$$|\psi\rangle^{(N-1)} = \sum_{n=0}^{N-1} d_n^{(N-1)} |n\rangle. \quad (15.114)$$

Time evolution yields

$$\begin{aligned} |\psi\rangle^{(N-1)} &\otimes (\alpha_1^{(N)}|\varphi_1\rangle + \alpha_2^{(N)}|\varphi_2\rangle) \\ &\rightarrow \sum_{n=0}^{N-1} d_n^{(N-1)} \{ \alpha_1^{(N)} [C_{n+1}(t_N)|n, 1\rangle + S_{n+1}(t_N)|n+1, 2\rangle] \\ &\quad + \alpha_2^{(N)} [S_n(t_N)|n-1, 1\rangle + C_n(t_N)|n, 2\rangle] \}, \end{aligned} \quad (15.115)$$

and, after projection onto $|\varphi_2\rangle$,

$$\begin{aligned} |\psi\rangle^{(N)} &= \sum_{n=0}^{N-1} d_n^{(N-1)} \{ \alpha_1^{(N)} S_{n+1}(t_N) |n+1\rangle + \alpha_2^{(N)} C_n(t_N) |n\rangle \} \\ &= \sum_{n=0}^N d_n^{(N)} |n\rangle. \end{aligned} \quad (15.116)$$

It is not difficult to see that, for suitable $d_n^{(N-1)}$, $\alpha_{1,2}^{(N)}$, and t_N , arbitrary superpositions (eq. (15.108)) can be produced.

We conclude this section with an example of state preparation, in which the Jaynes–Cummings model with strong detuning is employed. We start out from a state

$$|\Psi\rangle = \frac{1}{\sqrt{2}} \sum_{n=0}^{\infty} c_n |n\rangle \otimes (|\varphi_1\rangle + |\varphi_2\rangle) = \frac{1}{\sqrt{2}} \sum_{n=0}^{\infty} c_n (|n, 1\rangle + |n, 2\rangle). \quad (15.117)$$

Applying the time evolution of eq. (15.102) of the Jaynes–Cummings model with large detuning yields

$$|\Psi\rangle \rightarrow \frac{1}{\sqrt{2}} \sum_{n=0}^{\infty} c_n \left[\exp\left(-i \frac{|g|^2}{\Delta} (n+1)t\right) |n, 1\rangle + \exp\left(i \frac{|g|^2}{\Delta} nt\right) |n, 2\rangle \right]. \quad (15.118)$$

Projection of the atomic state onto

$$|\varphi\rangle = \frac{1}{\sqrt{2}} \left[\exp\left(-i \frac{|g|^2}{\Delta} t\right) |\varphi_1\rangle + |\varphi_2\rangle \right] \quad (15.119)$$

yields the photonic state

$$|\psi\rangle = \frac{1}{2} \sum_n c_n \left[\exp\left(-i \frac{|g|^2}{\Delta} nt\right) |n\rangle + \exp\left(i \frac{|g|^2}{\Delta} nt\right) |n\rangle \right]. \quad (15.120)$$

Choosing

$$c_n = \frac{\beta^n}{\sqrt{n!}} e^{-\frac{1}{2}|\beta|^2}$$

corresponds to an initial coherent state of the electromagnetic field:

$$|\Psi\rangle = \frac{1}{\sqrt{2}} |\beta\rangle \otimes (|\varphi_1\rangle + |\varphi_2\rangle). \quad (15.121)$$

Then the emerging state $|\psi\rangle$ of eq. (15.120) will be

$$|\psi\rangle = \frac{1}{2} (|\beta e^{i\delta t}\rangle + |\beta e^{-i\delta t}\rangle), \quad (15.122)$$

with $\delta = |g|^2/\Delta$. This is a superposition of two coherent states. Now, coherent states are as close as possible to classical states, and the superposition comes close to the superposition of classical states. In analogy to the famous “Schrödinger’s cat” paradox, where a superposition of two states of a cat, one dead and one alive, is considered, the states of eq. (15.122) are called *Schrödinger cat states*.

15.7 The Paul trap

Nowadays, experiments with individual atoms have become feasible. This progress was made possible by the construction of traps that confine single atoms to a small volume where they can stay and be observed for weeks and months while their interaction with radiation can be

studied and quantum transitions can be observed in detail. Ions are usually stored by means of electric fields acting on their charge, whereas neutral atoms are stored in magnetic fields, which act on their magnetic moment.

For trapping and observing an atom or ion, it is desirable to reduce its kinetic energy. Major progress has been made in this direction with *laser cooling*, a technique that allows the temperature in the trap to be reduced to a few microkelvins and even provides another mechanism to confine neutral particles. The principle of laser cooling is quite simple. Electromagnetic radiation is directed into the trap at a frequency slightly below the resonance frequency of the confined particles at rest. An outgoing particle experiences Doppler-shifted radiation and can be excited by absorbing a photon from the radiation. Thereby, it obtains an inwards directed momentum. Subsequently, the particle will be de-excited by radiating off a photon into a random direction. On average, the net effect of absorption and emission will be an inwards directed kick applied to the particle. In this way, a dramatic reduction of the momenta of the particles in the trap and temperatures in the microkelvin range can be achieved. If, in addition, the faster particles are allowed to escape from the trap by “evaporation”, temperatures in the nanokelvin range are reached. For a large number of confined Bose particles, Bose–Einstein condensation was observed.

The theoretical description of the confinement of a two-level atom or ion in a trap is closely analogous to the Jaynes–Cummings model of sections 15.1 and 15.4.

We start out with a simple model and will apply refinements later. So, we assume that the particle is confined by an isotropic oscillator potential with frequency ω_R and treat the one-mode electromagnetic field interacting with it as a classical field. The interaction with this field is assumed to be a dipole interaction. The model Hamiltonian for this situation is given by

$$\begin{aligned} H &= \frac{\mathbf{P}^2}{2M} + \frac{M\omega_R^2}{2}\mathbf{Q}^2 + \frac{1}{2}\hbar\omega\sigma_3 + H_{\text{int}} \\ &= H_{\text{CM}} + H_{\text{AT}} + H_{\text{int}} = H_0 + H_{\text{int}}, \end{aligned} \quad (15.123)$$

where H_{CM} describes the motion in the oscillator potential, and H_{AT} the two levels of the particle. For a homogeneous running classical light wave, the interaction Hamiltonian with the classical field has the form

$$H_{\text{int}}^r = \hbar(\alpha g\sigma_+ e^{-i(\omega_L t - \mathbf{k} \cdot \mathbf{Q})} + \alpha^* g^* \sigma_- e^{i(\omega_L t - \mathbf{k} \cdot \mathbf{Q})}), \quad (15.124)$$

and for a standing classical light wave we have

$$H_{\text{int}}^s = 2\hbar g(\alpha\sigma_+ + \alpha^*\sigma_-) \cos(\omega_L t) \sin(\mathbf{k} \cdot \mathbf{Q}), \quad (15.125)$$

where ω_L is the frequency of the classical (laser) radiation.

Introducing the creation and annihilation operators

$$\begin{aligned} b^+ &= \frac{1}{\sqrt{2}} \left(\sqrt{\frac{M\omega_R}{\hbar}} \mathbf{Q} - i \frac{1}{\sqrt{M\omega_R\hbar}} \mathbf{P} \right), \\ b &= \frac{1}{\sqrt{2}} \left(\sqrt{\frac{M\omega_R}{\hbar}} \mathbf{Q} + i \frac{1}{\sqrt{M\omega_R\hbar}} \mathbf{P} \right), \end{aligned} \quad (15.126)$$

as in section 13.1, the Hamiltonian assumes the form

$$H = \hbar\omega_R(\mathbf{b}^+\mathbf{b} + \frac{1}{2}) + \frac{1}{2}\hbar\omega\sigma_3 + H_{\text{int}}, \quad (15.127)$$

with

$$H_{\text{int}}^r = \hbar \left\{ \alpha g \sigma_+ \exp(-i\omega_L t) \exp \left[i \sqrt{\frac{\hbar}{2M\omega_R}} \mathbf{k} \cdot (\mathbf{b} + \mathbf{b}^+) \right] \right. \\ \left. + \alpha^* g^* \sigma_- \exp(i\omega_L t) \exp \left[-i \sqrt{\frac{\hbar}{2M\omega_R}} \mathbf{k} \cdot (\mathbf{b} + \mathbf{b}^+) \right] \right\} \quad (15.128)$$

or

$$H_{\text{int}}^s = 2\hbar g(\alpha\sigma_+ + \alpha^*\sigma_-) \cos(\omega_L t) \sin \left(\sqrt{\frac{\hbar}{2M\omega_R}} \mathbf{k} \cdot (\mathbf{b} + \mathbf{b}^+) \right). \quad (15.129)$$

The quantity

$$\boldsymbol{\eta} = \sqrt{\frac{\hbar}{2M\omega_R}} \mathbf{k} \quad (15.130)$$

is called the *Lamb–Dicke parameter*.

The typical amplitude d of the harmonic oscillator in low states of excitation is of the order of $d \approx \sqrt{\hbar/(2M\omega_R)}$, and the Lamb–Dicke parameter is the ratio d/λ of the amplitude and the wavelength of the electromagnetic radiation.

In the interaction picture, the interaction Hamiltonian $H_I = U_0^\dagger H_{\text{int}} U_0$ can be calculated just like in section 15.1:

$$H_I^r = \hbar \{ \alpha g \sigma_+ \exp[i(\omega - \omega_L)t] \exp[i\boldsymbol{\eta}(\mathbf{b} e^{-i\omega_R t} + \mathbf{b}^+ e^{i\omega_R t})] + \text{h.c.} \} \\ = \hbar \alpha g \sigma_+ e^{i(\omega - \omega_L)t} D(i\boldsymbol{\eta} e^{i\omega_R t}) + \text{h.c.}, \quad (15.131)$$

where $D(i\boldsymbol{\eta} e^{i\omega_R t})$ is just the displacement operator of eq. (13.43). For the standing wave we obtain

$$H_I^s = -i\hbar g \cos(\omega_L t) (\alpha\sigma_+ e^{i\omega t} + \alpha^*\sigma_- e^{-i\omega t}) [D(i\boldsymbol{\eta} e^{i\omega_R t}) - D(-i\boldsymbol{\eta} e^{i\omega_R t})]. \quad (15.132)$$

The form of $H_I^{r,s}$ shows that atomic transitions are in general associated with both multiple emission and absorption of vibration quanta, called phonons, which here play the role of the photons in section 15.1.

At this level of generality, there is no place for a rotating wave approximation, yet. Sometimes, special approximation apply, which we spell out for the running wave case, leaving the completely analogous case of the standing wave for the reader.

1. For small $|\boldsymbol{\eta}| \ll 1$, the so-called *Lamb–Dicke approximation* applies. Only the first two terms of the exponential in eq. (15.131) are kept:

$$H_I^r \approx \hbar \alpha g \sigma_+ e^{i(\omega - \omega_L)t} (1 + i\boldsymbol{\eta} \cdot \mathbf{b} e^{-i\omega_R t} + i\boldsymbol{\eta} \cdot \mathbf{b}^+ e^{i\omega_R t}) + \text{h.c.} \quad (15.133)$$

Sometimes, a rotating wave approximation will apply, which consists in keeping only the interaction term with the slowest oscillation. If $|\omega - \omega_L - \omega_R|$ is minimal, the rotating wave approximation gives the Jaynes–Cummings model (with phonons rather than photons):

$$H_I^r \approx \hbar \alpha g \sigma_+ \boldsymbol{\eta} \cdot \mathbf{b} e^{i(\omega - \omega_L - \omega_R)t} + \text{h.c.} \quad (15.134)$$

If, on the other hand, $|\omega - \omega_L + \omega_R|$ is minimal, the resulting interaction Hamiltonian is

$$H_I^r \approx \hbar \alpha g \sigma_+ \boldsymbol{\eta} \cdot \mathbf{b}^+ e^{i(\omega - \omega_L + \omega_R)t} + \text{h.c.} \quad (15.135)$$

This is the *anti Jaynes–Cummings model*, which is also exactly solvable. The conserved quantity that guarantees solvability is the difference

$$C_{-1} := -\frac{1}{2}\sigma_3 + b^+b \quad (15.136)$$

between the atomic and phononic excitation numbers. An excitation of the atom is associated with the emission of a phonon.

2. In the interaction Hamiltonian (15.131) it may also be a good approximation to keep the term with the slowest variation. If, for some integer s , $|\omega - \omega_L - s\omega_R|$ is minimal, we obtain

$$H_I^r \approx \hbar \alpha g \sigma_+ \frac{(i\boldsymbol{\eta} \cdot \mathbf{b}^{(+)})^{|s|}}{|s|!} e^{i(\omega - \omega_L - s\omega_R)t} + \text{h.c.}, \quad (15.137)$$

where for positive s the annihilation operator $\mathbf{b}$ enters and for negative s the creation operator $\mathbf{b}^+$ enters. This is the exactly solvable *s-phonon Jaynes–Cummings model* or the *s-photon anti Jaynes–Cummings model*, where an atomic excitation is associated with the absorption or emission of s phonons, respectively. The conserved quantity is

$$C_s = \frac{1}{2}s\sigma_3 + b^+b. \quad (15.138)$$

The Hamiltonian H_{CM} of eq. (15.123) for the center-of-mass motion is unrealistic, because an electrostatic or magnetostatic oscillator potential cannot exist. In fact, the potential V has to obey Laplace's equation

$$\Delta V = \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \right) V = 0 \quad (15.139)$$

and hence cannot have a minimum. Not all of its second derivatives can be positive. The most general quadratic potential V with rotation symmetry about the 3-axis and fulfilling eq. (15.139) is of the form

$$V = \frac{1}{2}Mf(x_1^2 + x_2^2 - 2x_3^2). \quad (15.140)$$

For at least one direction, this is repulsive. The way to achieve confinement nevertheless is to replace the constant f by a periodic function $f(t)$ with $f(t) = f(t+T)$, which can change sign

and will periodically interchange the attractive and repulsive directions. For the discussion of this mechanism, we can treat the three directions separately and consider the equations:

$$\ddot{x}_{1,2} + f(t)x_{1,2} = 0, \quad \ddot{x}_3 - 2f(t)x_3 = 0. \quad (15.141)$$

The function $f(t)$ is normally chosen to be of the form

$$f(t) = A + B \cos(\omega_{rf}t). \quad (15.142)$$

The resulting equation is called *Mathieu's equation*. The equation for x_3 is obtained from the equation for x_1 by the replacements $A \rightarrow -2A$ and $B \rightarrow -2B$.

Now, let $\epsilon_1(t)$ and $\epsilon_2(t)$ be any two solutions of the equation

$$\ddot{x} + f(t)x = 0 \quad (15.143)$$

with periodic f . Then $\epsilon_1(t+T)$ and $\epsilon_2(t+T)$ are also solutions and must be linear combinations of $\epsilon_1(t)$ and $\epsilon_2(t)$:

$$\begin{pmatrix} \epsilon_1(t+T) \\ \epsilon_2(t+T) \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \end{pmatrix}. \quad (15.144)$$

By a suitable choice of a basis of solutions, one can (generically) diagonalize the matrix in eq. (15.144) and construct multiplicative solutions $\epsilon_{\mu_1}(t)$ and $\epsilon_{\mu_2}(t)$ with

$$\epsilon_{\mu_{1,2}}(0) = 1 \quad \text{and} \quad \epsilon_{\mu_{1,2}}(t+T) = e^{i\mu_{1,2}T} \epsilon_{\mu_{1,2}}(t). \quad (15.145)$$

Such multiplicative solutions are called *Floquet solutions*, and the quantities $\mu_{1,2}$ are the *Floquet exponents*. If both Floquet exponents are real, the solution is stable; if one of them has an imaginary part, an instability for $t \rightarrow \infty$ will occur. For eq. (15.143) we have $\mu_1 = -\mu_2$, because for any two solutions of eq. (15.143) we have

$$\frac{d}{dt}[\epsilon_1(t)\dot{\epsilon}_2(t) - \epsilon_2(t)\dot{\epsilon}_1(t)] = 0, \quad (15.146)$$

and inserting ϵ_{μ_1} and ϵ_{μ_2} gives

$$\begin{aligned} & \epsilon_{\mu_1}(t+T)\dot{\epsilon}_{\mu_2}(t+T) - \epsilon_{\mu_2}(t+T)\dot{\epsilon}_{\mu_1}(t+T) \\ &= e^{i(\mu_1+\mu_2)T}[\epsilon_{\mu_1}(t)\dot{\epsilon}_{\mu_2}(t) - \epsilon_{\mu_2}(t)\dot{\epsilon}_{\mu_1}(t)] \\ &= \epsilon_{\mu_1}(t)\dot{\epsilon}_{\mu_2}(t) - \epsilon_{\mu_2}(t)\dot{\epsilon}_{\mu_1}(t). \end{aligned} \quad (15.147)$$

Moreover, if $f(t) = f(-t)$, as for the choice in eq. (15.142), then, with $\epsilon_{\mu}(0) = 1$, we find

$$\epsilon_{-\mu}(t) = \epsilon_{\mu}(-t). \quad (15.148)$$

For real μ , which is the stable case of interest, we also have

$$\epsilon_{-\mu}(t) = \epsilon_{\mu}(t)^*. \quad (15.149)$$

Every Floquet solution can be written in the form

$$\epsilon_{\mu}(t) = e^{i\mu t} e_{\mu}(t), \quad (15.150)$$

with periodic $e_\mu(t)$. Hence,

$$\epsilon_\mu(t) = \sum_{n=-\infty}^{\infty} c_n e^{i(\mu+n)t}. \quad (15.151)$$

If eqs. (15.148) and (15.149) are fulfilled, the coefficients c_n are real and $\dot{\epsilon}_\mu(0)$ is purely imaginary.

The parameters A and B in eq. (15.142) can be chosen such that the solution of eq. (15.141) is stable in all directions. In fact, eq. (15.143) with f from eq. (15.142) can be written in the form

$$-\ddot{x} - [|B| + B \cos(\omega_{rf}t)]x = (A - |B|)x. \quad (15.152)$$

One recognizes the Schrödinger equation in a periodic nowhere-positive potential, and the stable regions correspond to the energy bands.

To perform the transition to the interaction picture for the Hamiltonian (15.123), with

$$H_{\text{CM}} = \frac{P}{2M} + M \frac{f(t)}{2} (Q_1^2 + Q_2^2) - M f(t) Q_3^2 \quad (15.153)$$

instead of the simple oscillator Hamiltonian, we have to solve the equations of motion:

$$\dot{Q}_i(t) = \frac{1}{i\hbar} [Q_i(t), H_{\text{CM}}], \quad \dot{P}_i(t) = \frac{1}{i\hbar} [P_i(t), H_{\text{CM}}]. \quad (15.154)$$

It will suffice to treat only one component:

$$\dot{Q}(t) = \frac{P(t)}{M}, \quad \dot{P}(t) + M f(t) Q(t) = 0 \quad (15.155)$$

or

$$P(t) = M \dot{Q}(t), \quad \ddot{Q}(t) + f(t) Q(t) = 0. \quad (15.156)$$

These are the same equations as the classical equations of motion. The solution of eq. (15.156) is best presented in terms of quadratures,

$$X_1(t) = \sqrt{\frac{M\omega_R}{\hbar}} Q(t), \quad X_2(t) = \frac{1}{\sqrt{M\omega_R\hbar}} P(t), \quad (15.157)$$

and annihilation and creation operators,

$$b(t) = X_1(t) + iX_2(t), \quad b^+(t) = X_1(t) - iX_2(t). \quad (15.158)$$

As opposed to eq. (15.123), the frequency ω_R is not given by the Hamiltonian but rather an arbitrary reference frequency that can be chosen for convenience. Changing ω_R corresponds to a Bogolyubov transformation. In terms of the quadratures, the equations of motion are

$$\ddot{X}_1(t) + f(t) X_1(t) = 0, \quad X_2(t) = \frac{1}{\omega_R} \dot{X}_1(t). \quad (15.159)$$

It is now easy to write down the solution of eq. (15.159) with the initial values $X_1(0) = X_1$ and $X_2(0) = X_2$. Let $\epsilon(t)$ be the solution of eq. (15.143) with the initial values

$$\epsilon(0) = 1, \quad \dot{\epsilon}(0) = i\omega_R. \quad (15.160)$$

Then

$$\begin{aligned} X_1(t) &= X_1 \operatorname{Re} \epsilon(t) + X_2 \operatorname{Im} \epsilon(t), \\ X_2(t) &= X_1 \frac{\operatorname{Re} \dot{\epsilon}(t)}{\omega_R} + X_2 \frac{\operatorname{Im} \dot{\epsilon}(t)}{\omega_R}, \end{aligned} \quad (15.161)$$

and

$$\begin{aligned} b(t) &= \frac{1}{2}b \left[\epsilon^*(t) + \frac{i}{\omega_R} \dot{\epsilon}^*(t) \right] + \frac{1}{2}b^+ \left[\epsilon(t) + \frac{i}{\omega_R} \dot{\epsilon}(t) \right], \\ b^+(t) &= \frac{1}{2}b^+ \left[\epsilon(t) - \frac{i}{\omega_R} \dot{\epsilon}(t) \right] + \frac{1}{2}b \left[\epsilon^*(t) - \frac{i}{\omega_R} \dot{\epsilon}^*(t) \right]. \end{aligned} \quad (15.162)$$

In particular,

$$b(t) + b^+(t) = b\epsilon^*(t) + b^+\epsilon(t). \quad (15.163)$$

By choosing $i\omega_R = \dot{\epsilon}_\mu(0)$, one can achieve that the solution ϵ is a Floquet solution ϵ_μ .

The commutation relations

$$[X_1(t), X_2(t)] = i, \quad [b(t), b^+(t)] = 1 \quad (15.164)$$

are a direct consequence of eq. (15.146).

Using the results of section 13.2 we see that the time evolution is given by a combination,

$$R(\theta_1(t))B(r(t))R(\theta_2(t)), \quad (15.165)$$

of phase space rotations and Bogolyubov transformations.

The interaction Hamiltonian can now be immediately written down as

$$\begin{aligned} H_I^r &= \hbar\alpha g \sigma_+ e^{i(\omega - \omega_L)t} D(i\eta\epsilon(t)), \\ H_I^s &= \hbar\alpha g \cos(\omega_L t) (\sigma_+ e^{i\omega t} + \sigma_- e^{-i\omega t}) [D(i\eta\epsilon(t)) - D(-i\eta\epsilon(t))]. \end{aligned} \quad (15.166)$$

Comparing with eqs. (15.131) and (15.132), we see that the only change from the simple to the time-dependent oscillator is the substitution $e^{i\omega_R t} \rightarrow \epsilon(t)$. All the subsequent approximations can be performed in complete analogy. One only has to keep in mind that, in searching for the lowest frequency terms, the contribution of the Floquet exponent μ also has to be considered.

The simple time-independent oscillator potential in eq. (15.123) may be considered as an effective time-averaged approximation of the more realistic time-dependent confining potential.

15.8 Motion of a two-level atom in a quantized light field

In the previous section, we discussed the quantum-mechanical motion of an atom or ion in a confining classical field. Now the quantum mechanics of the center-of-mass motion of an atom in the quantized field of an electromagnetic wave will be considered. The transition to a classical wave field will be straightforward. The electromagnetic field will interact with the atom and exert a force on it because the atom can be excited by the absorption of photons. We simplify our discussion by taking a one-mode electric wave field in resonance with the transition frequency of a two-level atom. The resulting model will be a slight generalization of the Jaynes–Cummings model. The structure of the Hamiltonian in the Heisenberg picture will be

$$H = \hbar\Omega a^\dagger a + \frac{1}{2}\hbar\omega\sigma_3 + \frac{\mathbf{P}^2}{2M} + \mathbf{D} \cdot \mathbf{E}(\mathbf{Q}). \quad (15.167)$$

The new term as compared to eqs. (15.3) and (15.7) is the kinetic energy term $\mathbf{P}^2/2M$ of the center-of-mass motion of the atom. Here $\mathbf{D}$ is the dipole operator of the atom, and, in contrast to eq. (15.8), the atom is not assumed to be in a fixed position. So, the electric field $\mathbf{E}$ has to be taken at the position $\mathbf{Q}$ of the center of mass of the atom.

Expanding the field $\mathbf{E}$ into creation and annihilation operators, and also applying the rotating phase approximation, we obtain

$$\begin{aligned} H &= \hbar\Omega a^\dagger a + \frac{1}{2}\hbar\omega\sigma_3 + \frac{\mathbf{P}^2}{2M} + \hbar g(\mathbf{Q})\sigma_+ a + \hbar g^*(\mathbf{Q})\sigma_- a^\dagger \\ &= \hbar\Omega a^\dagger a + \frac{1}{2}\hbar\omega\sigma_3 + H_{\text{int}}. \end{aligned} \quad (15.168)$$

This Hamiltonian differs from the Jaynes–Cummings Hamiltonian in eq. (15.59) in two respects: first, the kinetic energy of the atom contributes; and secondly, the coupling,

$$\hbar g(\mathbf{Q}) = i\sqrt{\frac{\hbar\Omega}{2\epsilon_0}} \mathbf{u}(\mathbf{Q}) \langle \varphi_1 | \mathbf{D} | \varphi_2 \rangle, \quad (15.169)$$

depends on the position of the atom. The position dependence is given by the shape of the mode function $\mathbf{u}(\mathbf{Q})$. For a standing wave, $\mathbf{u}(\mathbf{Q})$ will have a constant, position-independent phase, and $g(\mathbf{Q})$ can be made real by a simple rotation of a and $a^\dagger$. In the following, and for simplicity, we shall assume $g(\mathbf{Q})$ to be real. The total excitation degree $C = a^\dagger a + \frac{1}{2}\sigma_3$ is still conserved. The transition to the interaction picture is now readily performed. Assuming $\Delta = \omega - \Omega = 0$, we obtain

$$H_I = \frac{\mathbf{P}^2}{2M} + \hbar g(\mathbf{Q})(\sigma_+ a + \sigma_- a^\dagger). \quad (15.170)$$

Note that we have lumped the kinetic energy into the interaction part. The evolution equation,

$$i\hbar \frac{d}{dt} |\Psi(t)\rangle = H_I |\Psi(t)\rangle, \quad (15.171)$$

in the interaction picture will be written in the basis

$$\{|n\rangle \otimes |\varphi_i\rangle \otimes |\mathbf{x}\rangle\}, \quad n = 0, 1, 2, \dots, \quad i = 1, 2, \quad \mathbf{x} \in \mathbb{R}^3 \quad (15.172)$$

of the Hilbert space H of the system, where, as before, n and i describe the state of the wave field and the excitation state of the atom, respectively, and $|\mathbf{x}\rangle$ is an eigenstate of the position operator $\mathbf{Q}$. Writing $|n, i, \mathbf{x}\rangle$ for $|n\rangle \otimes |\varphi_i\rangle \otimes |\mathbf{x}\rangle$, the state vector $|\Psi(t)\rangle$ will be of the form

$$|\Psi(t)\rangle = \sum_{j=1}^2 \sum_{m=0}^{\infty} \int d^3x \Phi_{m,j}(\mathbf{x}', t) |m, j, \mathbf{x}'\rangle, \quad (15.173)$$

and eq. (15.171) will turn into

$$i\hbar \frac{d}{dt} \langle n, i, \mathbf{x} | \Psi(t) \rangle = \langle n, i, \mathbf{x} | H_I | \Psi(t) \rangle \quad (15.174)$$

or

$$\begin{aligned} i\hbar \frac{d}{dt} \Phi_{n,1}(\mathbf{x}, t) &= -\frac{\hbar^2}{2M} \Delta \Phi_{n,1} + \hbar \sqrt{n+1} g(\mathbf{x}) \Phi_{n+1,2}(\mathbf{x}, t), \\ i\hbar \frac{d}{dt} \Phi_{n,2}(\mathbf{x}, t) &= -\frac{\hbar^2}{2M} \Delta \Phi_{n,2} + \hbar \sqrt{n} g(\mathbf{x}) \Phi_{n-1,1}(\mathbf{x}, t). \end{aligned} \quad (15.175)$$

Here Δ is, of course, the Laplace operator. Exploiting the real-valuedness of $g(\mathbf{x})$, we can decouple the system of eq. (15.175) by introducing

$$\Phi_n^{\pm}(\mathbf{x}, t) = \Phi_{n-1,1}(\mathbf{x}, t) \pm \Phi_{n,2}(\mathbf{x}, t) \quad (15.176)$$

and obtain

$$i\hbar \frac{d}{dt} \Phi_n^{\pm}(\mathbf{x}, t) = -\frac{\hbar^2}{2M} \Delta \Phi_n^{\pm} \pm \hbar \sqrt{n} g(\mathbf{x}) \Phi_n^{\pm}(\mathbf{x}, t). \quad (15.177)$$

Equation (15.177) is a Schrödinger equation for the position wave functions $\Phi_n^{\pm}(\mathbf{x}, t)$ of the atom. The potential term $\hbar \sqrt{n} g(\mathbf{x})$ is determined by the shape of the mode function of the wave field. As is to be expected, the potential is proportional to $\sqrt{n}$ and, hence, to the expectation value of the amplitude of the electromagnetic field. The classical limit for the electromagnetic field is obtained by simply replacing n and $n \pm 1$ by a fixed large value N proportional to the intensity of the classical field. In the classical wave field, the atoms will experience a force proportional to $\nabla g(\mathbf{x})$.

Equation (15.177) is fundamental for any discussion of the interaction of matter with electromagnetic wave fields. For a treatment of laser cooling, the leakage of the electromagnetic field into different modes has to be represented by a loss term in the equation of motion for the density operator ρ .

Equation (15.177) is also the starting point for the rapidly developing field of *atom optics*. It is a wave equation for the Schrödinger amplitude of the atom similar to the wave equation for the electromagnetic field in a diffractive medium. The quantity $\sqrt{n} g(\mathbf{x})$ is in close analogy to the index of refraction.

By suitably shaping the wave field mode function $u(\mathbf{x})$, one can construct mirrors and beam splitters for atomic rays. With a mode function $u(\mathbf{x})$ periodic in the 3-direction, an atomic Fabry–Perot interferometer can be constructed. Atomic beam splitters can be used to

build atomic Mach–Zehnder interferometers and eight-port interferometers. Atom interferometers have the perspective to achieve even better resolution than optical interferometers. Note that, in atom optics, light and matter have exchanged their roles: in light optics, matter is used for shaping light wave fields, while in atom optics, matter wave fields are manipulated with light.

16 Quantum optics and fundamental quantum theory

In quantum optics, experiments with single photons and atoms have become feasible. This opens the way for direct tests of basic features of quantum theory, which until recently only enjoyed the status of “Gedanken” experiments. Atomic transitions can be monitored, the measurement process can be investigated, and entanglement correlations can be observed in detail. Quantum computers are one of many spectacular applications in sight. They represent the ultimate state of miniaturization, where the components of a computer are of atomic size. Novel quantum algorithms may speed up certain calculations, like factorization into prime numbers, dramatically beyond the limits of normal computation.

16.1 Quantum entanglement

Entangled quantum states give rise to particularly drastic genuinely quantum-theoretical phenomena. We have met entanglement already on several occasions, for instance in connection with the Jaynes–Cummings model. This section is devoted to a more systematic treatment of entanglement.

Let us assume that the Hilbert space $\mathcal{H}$ of a quantum system Σ is the tensor product of two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$:

$$\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2. \quad (16.1)$$

Such a tensor product decomposition will be natural if Σ is composed of two subsystems Σ_1 and Σ_2 with Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. A tensor product decomposition also arises if the system has independent commuting degrees of freedom, such as excitation and center-of-mass motion of an atom.

A state $|\Psi\rangle$ in $\mathcal{H}$ with

$$|\Psi\rangle = |\varphi\rangle \otimes |\psi\rangle \quad \text{and} \quad |\varphi\rangle \in \mathcal{H}_1, \quad |\psi\rangle \in \mathcal{H}_2 \quad (16.2)$$

is called a *product state* or a *decomposable state*. Generally, states in $\mathcal{H}$ are not product states but only linear combinations of product states, so-called *entangled states*. Every state in $\mathcal{H}$ can be represented as a superposition of product states. Let $\{|\varphi_i\rangle\}_{i \in I}$ and $\{|\psi_j\rangle\}_{j \in J}$ be orthonormal bases of $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Then $\{|\varphi_i\rangle \otimes |\psi_j\rangle\}_{i \in I, j \in J}$ is an orthonormal basis of $\mathcal{H}$, and every vector in $\mathcal{H}$ can be written as

$$|\Psi\rangle = \sum_{i,j} c_{ij} |\varphi_i\rangle \otimes |\psi_j\rangle. \quad (16.3)$$

Notice that a state can be called “entangled” or “decomposable” only with respect to a given tensor product decomposition (eq. (16.1)).

If A is an observable on $\mathcal{H}_1$ and B is an observable on $\mathcal{H}_2$, we can define a product observable $A \otimes B$ by

$$(A \otimes B)|\Psi\rangle = \sum_{i,j} c_{ij} A|\varphi_i\rangle \otimes B|\psi_j\rangle. \quad (16.4)$$

Again, we can distinguish between decomposable and non-decomposable, entangled observables with respect to the tensor product decomposition (eq. (16.1)).

A mixed state on $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ is given by a density operator ρ on $\mathcal{H}$. This ρ is called a product state if

$$\rho = \rho_1 \otimes \rho_2, \quad (16.5)$$

where ρ_1 and ρ_2 are density operators on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, and otherwise it is called an entangled state. The partial traces Tr_1 and Tr_2 with respect to the factors $\mathcal{H}_1$ and $\mathcal{H}_2$ are uniquely defined by the properties

$$\text{Tr}_1(A \otimes B) = (\text{Tr}_{\mathcal{H}_1} A)B, \quad \text{Tr}_2(A \otimes B) = A(\text{Tr}_{\mathcal{H}_2} B). \quad (16.6)$$

For

$$C = \sum_{i,j,k,r} c_{ikjr} |\varphi_i\rangle \otimes |\psi_k\rangle \langle \varphi_j| \otimes \langle \psi_r|, \quad (16.7)$$

we have

$$\text{Tr}_1 C = \sum_{i,k,r} c_{ikir} |\psi_k\rangle \langle \psi_r|, \quad \text{Tr}_2 C = \sum_{i,j,k} c_{ikjk} |\varphi_i\rangle \langle \varphi_j|. \quad (16.8)$$

Obviously, we have

$$\text{Tr}_1 \text{Tr}_2 C = \text{Tr}_2 \text{Tr}_1 C = \sum_{i,k} c_{ikik} = \text{Tr} C.$$

If ρ is a density operator on $\mathcal{H}$, then $\text{Tr}_2 \rho$ and $\text{Tr}_1 \rho$ are density operators on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. They are said to arise from ρ by *reduction* to $\mathcal{H}_1$ and $\mathcal{H}_2$.

Let us now consider in detail the simplest case of two-dimensional Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with bases $\{|\varphi_1\rangle, |\varphi_2\rangle\}$ and $\{|\psi_1\rangle, |\psi_2\rangle\}$. The space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ is four-dimensional and has an orthonormal base $\{|\varphi_i\rangle \otimes |\psi_j\rangle\}_{i,j=1,2}$. There exists another orthonormal basis of $\mathcal{H}$ consisting of the following entangled states, the so-called *Bell states*.

$$\begin{aligned} |\Phi_1\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\psi_2\rangle - |\varphi_2\rangle \otimes |\psi_1\rangle), \\ |\Phi_2\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\psi_2\rangle + |\varphi_2\rangle \otimes |\psi_1\rangle), \\ |\Phi_3\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\psi_1\rangle - |\varphi_2\rangle \otimes |\psi_2\rangle), \\ |\Phi_4\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\psi_1\rangle + |\varphi_2\rangle \otimes |\psi_2\rangle). \end{aligned} \quad (16.9)$$

Consider the observables

$$\begin{aligned} A_1 &= |\varphi_1\rangle\langle\varphi_1| - |\varphi_2\rangle\langle\varphi_2|, \\ A_2 &= |\psi_1\rangle\langle\psi_1| - |\psi_2\rangle\langle\psi_2|, \end{aligned} \quad (16.10)$$

on $\mathcal{H}_1$ and $\mathcal{H}_2$, which have $|\varphi_1\rangle$ and $|\psi_1\rangle$ as eigenstates with eigenvalue 1 and $|\varphi_2\rangle$ and $|\psi_2\rangle$ as eigenstates with eigenvalue -1 . The Bell states are not eigenstates of $A_1 \otimes \mathbf{1}$ or $\mathbf{1} \otimes A_2$; but if these observables are measured in a Bell state, there is a strict correlation of the outcomes of the measurements. If and only if a measurement of $A_1 \otimes \mathbf{1}$ gives the result ± 1 , then a measurement of $\mathbf{1} \otimes A_2$ will also give ± 1 for $|\Phi_2\rangle$ and $|\Phi_4\rangle$ with certainty, and ∓ 1 for $|\Phi_1\rangle$ and $|\Phi_3\rangle$.

This is a typical example of so-called *entanglement correlations* for measurements pertaining to the subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$, which occur if the measurements are applied to an entangled state. These entanglement correlations are particularly striking if subsystems described by $\mathcal{H}_1$ and $\mathcal{H}_2$ are spatially separated by a large distance. A measurement on one subsystem will immediately have a bearing on the result of a measurement on the other subsystem. It is highly implausible that this correlation is caused by an instantaneous interaction at a distance between the two subsystems. This would violate the laws of relativity physics, and we shall adopt for our following discussion the hypothesis of *locality* that such interactions are impossible. Then we are forced to attribute the correlations to the entangled state itself and admit interactionless correlations as a fundamental feature.

Before we proceed, we shall consider physical realizations of Bell states. First, let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the state spaces of two particles with spin $1/2$, where only the spin variables of the particles are considered. Taking the quantization axis in the 3-direction, $|\varphi_1\rangle$ and $|\varphi_2\rangle$ will be given by

$$\sigma_3|\varphi_{1,2}\rangle = \pm|\varphi_{1,2}\rangle, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and likewise for $|\psi_1\rangle$ and $|\psi_2\rangle$. The observable $\mathbf{n} \cdot \boldsymbol{\sigma} = n_1\sigma_1 + n_2\sigma_2 + n_3\sigma_3$, with $\mathbf{n}^2 = 1$, on $\mathcal{H}_1$ will have the value $+1$ if the spin is oriented in the $+\mathbf{n}$ direction and -1 if the spin points in the $-\mathbf{n}$ direction. With these assignments, the Bell state $|\Phi_1\rangle$ will be the singlet state and the other three Bell states are triplet states of the two-spin system.

The next example is fundamental for quantum optics. Let the system consist of two photons of given momenta and let $\mathcal{H}_1$ and $\mathcal{H}_2$ be the two-dimensional Hilbert spaces of their polarization degrees of freedom. For $\mathcal{H}_{1,2}$, let $\mathbf{e}_x$ and $\mathbf{e}_y$ be two vectors orthogonal to the momentum of the photon, and let $|x\rangle$, $|y\rangle$ be the planar polarization states along the line given by $\mathbf{e}_x$ and $\mathbf{e}_y$, respectively. We have $\langle x|y\rangle = 0$, and in the basis $\{|x\rangle, |y\rangle\}$ the observable with value $+1$ for $|x\rangle$ and -1 for $|y\rangle$ is given by

$$\sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

States with oblique planar polarization in the directions $(1/\sqrt{2})(\mathbf{e}_x \pm \mathbf{e}_y)$ are given by

$$\left| \frac{1}{\sqrt{2}}(x \pm y) \right\rangle = \frac{1}{\sqrt{2}}(|x\rangle \pm |y\rangle), \quad (16.11)$$

and they are eigenstates of

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

States of right and left circular polarization are given by

$$|R\rangle = \frac{1}{\sqrt{2}}(|x\rangle + i|y\rangle) \quad \text{and} \quad |L\rangle = \frac{1}{\sqrt{2}}(|x\rangle - i|y\rangle) \quad (16.12)$$

and are eigenstates of

$$\sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}.$$

A linear polarization in the line given by $\mathbf{n} = (\cos \theta, \sin \theta)$ is measured by the operator

$$\begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{pmatrix}. \quad (16.13)$$

The Bell states of eq. (16.9) assume the form

$$\begin{aligned} |\Phi_1\rangle &= \frac{1}{\sqrt{2}}(|R\rangle \otimes |L\rangle - |L\rangle \otimes |R\rangle) = \frac{1}{i\sqrt{2}}(|x\rangle \otimes |y\rangle - |y\rangle \otimes |x\rangle), \\ |\Phi_2\rangle &= \frac{1}{\sqrt{2}}(|R\rangle \otimes |L\rangle + |L\rangle \otimes |R\rangle) = \frac{1}{\sqrt{2}}(|x\rangle \otimes |x\rangle + |y\rangle \otimes |y\rangle), \\ |\Phi_3\rangle &= \frac{1}{\sqrt{2}}(|R\rangle \otimes |R\rangle - |L\rangle \otimes |L\rangle) = \frac{i}{\sqrt{2}}(|x\rangle \otimes |y\rangle + |y\rangle \otimes |x\rangle), \\ |\Phi_4\rangle &= \frac{1}{\sqrt{2}}(|R\rangle \otimes |R\rangle + |L\rangle \otimes |L\rangle) = \frac{1}{\sqrt{2}}(|x\rangle \otimes |x\rangle - |y\rangle \otimes |y\rangle). \end{aligned} \quad (16.14)$$

The directions x and y can be replaced by any two orthogonal directions x' and y' , and the form of the Bell states will remain unchanged. The various states of polarization of a photon can easily be transformed into each other with the devices described in chapter 12.

Bell states can nowadays be produced in a routine way. The first example were the photons arising from the decay of positronium into two photons. They have energies in the γ -radiation range of 0.5 MeV and are difficult to handle. Nowadays, entangled photon pairs are obtained from *downward conversion*. This is the inverse process of frequency doubling as described in chapter 6. A crystal with nonlinear optical properties converts photons of frequency ω into photon pairs of frequency $\omega/2$. Such photon pairs can be conveniently manipulated and used for entanglement experiments.

We conclude this section by showing that instantaneous entanglement correlations cannot be used to transmit signals with superluminal velocity. This, of course, strongly supports our locality assumption. Consider any density matrix ρ on $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ and let A and B be two

observables on $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, with spectral decomposition

$$\begin{aligned} A &= \sum_i a_i P_i, & B &= \sum_j b_j Q_j, \\ P_i P_j &= \delta_{ij} P_i, & Q_i Q_j &= \delta_{ij} Q_j, \\ \sum_i P_i &= \mathbf{1}_{\mathcal{H}_1}, & \sum_j Q_j &= \mathbf{1}_{\mathcal{H}_2}. \end{aligned} \quad (16.15)$$

The observables $A \otimes \mathbf{1}$ and $\mathbf{1} \otimes B$ are commensurable and can be measured simultaneously.

The probability of measuring the pair (a_i, b_j) of values of A and B is given by

$$w_{ij}^{(1,2)} = \text{Tr}(P_i \otimes Q_j) \rho = \text{Tr}_1(P_i) \text{Tr}_2(\mathbf{1} \otimes Q_j) \rho = \text{Tr}_2(Q_j) \text{Tr}_1(P_i \otimes \mathbf{1}) \rho. \quad (16.16)$$

If only $A \otimes \mathbf{1}$ or $\mathbf{1} \otimes B$ is measured, the probabilities for the result a_i or b_j are

$$\begin{aligned} w_i^{(1)} &= \text{Tr}(P_i \otimes \mathbf{1}) \rho = \text{Tr}_1(P_i) \text{Tr}_2(\rho), \\ w_j^{(2)} &= \text{Tr}(\mathbf{1} \otimes Q_j) \rho = \text{Tr}_2(Q_j) \text{Tr}_1(\rho). \end{aligned} \quad (16.17)$$

From eq. (16.15) we see that

$$w_i^{(1)} = \sum_j w_{ij}^{(1,2)}, \quad w_j^{(2)} = \sum_i w_{ij}^{(1,2)}. \quad (16.18)$$

Imagine, now, that a measurement of B has yielded the result b_j . The probability $w_{i|j}^{(1)}$ that a subsequent measurement of A will yield a_i is given by

$$w_{ij}^{(1,2)} = w_{i|j}^{(1)} w_j^{(2)}. \quad (16.19)$$

Evidently, we have

$$\sum_i w_{i|j}^{(1)} = 1. \quad (16.20)$$

The $w_{i|j}^{(1)}$ may depend strongly on j , which just describes the entanglement correlations. If, however, the outcome of the measurement of the observable B is unknown to an observer measuring A , the observer will see the distribution

$$\bar{w}_i^{(1)} = \sum_j w_{i|j}^{(1)} w_j^{(2)} = \sum_j w_{ij}^{(1,2)} = w_i^{(1)}. \quad (16.21)$$

This means that the observer measuring A cannot decide from the probability distribution obtained whether a measurement of B has been performed or not. So, no signal can be transferred by measuring B .

16.2 Bell's inequalities

In classical physics, precise knowledge of the state of a system Σ determines the result of every measurement performed on Σ . This fact is often formulated as the *principle of realism*:

The outcome of every measurement is predetermined by an objective feature of the state of the system.

In quantum theory, a quantum state in general only allows for statistical predictions of measurement results. Moreover, there are complementary observables, such as position and momentum or different components of angular momentum, which cannot simultaneously have sharp values in any quantum state. If the principle of realism is maintained also for quantum theory, indeterminacy and complementarity of quantum theory have to be interpreted in the following way:

1. The stochastic features of quantum theory are due to incomplete knowledge of the “true” state of a quantum system. In other words, a quantum state, in which the outcome of some measurement is always uncertain, only gives incomplete information about the “true” state of the system. In this sense, the quantum-mechanical description is incomplete, even if complete information about the “true” state may be inaccessible.
2. The measurement of an observable disturbs the “true” state in such a way that the result of a measurement of a complementary observable is uncertain.

In a famous article that appeared in 1935, Einstein, Podolsky, and Rosen sharpened the argument for the incompleteness of quantum theory by considering pairs of particles in an entangled state.

Imagine two particles that are widely separated. Assume the system to be in an entangled state such that the results of a measurement of an observable A for the two particles are strictly correlated. In other words, the value $a^{(2)}$ for A measured at particle 2 strictly determines the value $a^{(1)}$ for A measured at particle 1. If we adopt the locality principle of the previous section, a measurement of A at particle 2 cannot influence the state of the well-separated particle 1. Assume furthermore that the same strict correlation also holds for another observable B that does not commute with A . As an example we can take a Bell state (eq. (16.14)) and let A and B correspond to linear polarization measurements of two oblique directions x and x' .

Now we measure the observable A on particle 1 and B on particle 2. Under the assumption of locality, the A measurement does not disturb the B measurement, and the principle of realism allows us to interpret the result of the B measurement as information about the real state of particle 1. Repeating the A and B measurements many times, we can measure a joint distribution function of the values of the incommensurable observables A and B for particle 1. Such a quantity is not defined in ordinary quantum theory, which, assuming locality and realism, thus gives an incomplete description of the true state of particle 1. A complete theory would be a theory with local realistic hidden parameters describing true states such that a quantum state corresponds to a statistical ensemble of true states.

For a long time it was believed that such hidden parameter theories were empirically indistinguishable from ordinary quantum theory. Quite surprisingly, J. Bell in 1964 came out

with his inequalities on correlations or measurement results, allowing one to test the hypothesis of locality and realism experimentally. The experimental results obtained thereafter are all in accordance with quantum theory and strongly disfavor or exclude local realistic hidden parameter theories.

Consider again a well-separated two-particle system and three observables A , B , and C , each of which can be measured on both particles and can assume only two values, a_i , b_i , c_i ($i = 1, 2$). Let the system be in an entangled state with strict correlations between the measured values for the observables A , B , and C for the particles 1 and 2. For example, we can take a system of two spin-1/2 particles in a Bell state (eq. (16.14)) and take the observables $A = \mathbf{a} \cdot \boldsymbol{\sigma}$, $B = \mathbf{b} \cdot \boldsymbol{\sigma}$, $C = \mathbf{c} \cdot \boldsymbol{\sigma}$, which have eigenvalues ± 1 and measure the spin component in the direction of the unit vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$. Alternatively, we can consider a two-photon system in a Bell state and the linear polarization variables $\sigma(\alpha)$, $\sigma(\beta)$, and $\sigma(\gamma)$, with

$$\sigma(\theta) = \begin{pmatrix} \cos(2\theta) & \sin(2\theta) \\ \sin(2\theta) & -\cos(2\theta) \end{pmatrix}, \quad (16.22)$$

which have eigenvalue $+1$ if the linear polarization points in the θ direction and -1 if it is directed orthogonal to θ . Assuming local realism, the pair probabilities for particles 1 and 2, $w(a_i, b_i)$, are also probabilities for the joint distribution of values of A and B for particle 1. Now by elementary probability theory

$$w(a_i, b_j) = w(a_i, b_j, c_k) + w(a_i, b_j, c_{\bar{k}}) \leq w(a_i, c_k) + w(b_j, c_{\bar{k}}), \quad (16.23)$$

where $c_{\bar{1}} = c_2$ and $c_{\bar{2}} = c_1$. Evaluating these probabilities for the observables $\sigma(\alpha)$, $\sigma(\beta)$, and $\sigma(\gamma)$ for the Bell state

$$|\Phi_2\rangle = \frac{1}{\sqrt{2}}(|x\rangle \otimes |x\rangle + |y\rangle \otimes |y\rangle), \quad (16.24)$$

we obtain

$$\begin{aligned} w(a_1, b_1) &= \langle \Phi_2 | \frac{1}{2} [1 + \sigma(\alpha)] \frac{1}{2} [1 + \sigma(\beta)] | \Phi_2 \rangle = \frac{1}{4} [1 + \cos(2(\alpha - \beta))] \\ &= \frac{1}{2} \cos^2(\alpha - \beta) = w(a_2, b_2), \\ w(a_1, b_2) &= w(a_2, b_1) = \frac{1}{4} [1 - \cos(2(\alpha - \beta))] = \frac{1}{2} \sin^2(\alpha - \beta). \end{aligned} \quad (16.25)$$

The inequality eq. (16.23) implies that

$$w(a_1, b_1) \leq w(a_1, c_1) + w(b_1, c_2), \quad (16.26)$$

but, choosing $\alpha = 0$, $\beta = \pi/6$, and $\gamma = \pi/3$, we obtain

$$w(a_1, b_1) = \frac{3}{8}, \quad w(a_1, c_1) = w(b_1, c_2) = \frac{1}{8}, \quad (16.27)$$

and the inequality eq. (16.26) is violated. This violation is confirmed by experiment.

This is a really spectacular result of “experimental philosophy”. The inequality (16.23) is derived using only the postulates of locality and realism. Its violation means that both of these intuitively natural properties cannot be true for quantum theory. If one insists on locality, then the realistic interpretation of measurement results on correlated pairs as results on one of its

members is not justified. It is not legitimate to assume that the result of a measurement on a quantum system is predetermined by (however unknown) features of the “true state” of the system. There is no way to consider the quantum state as an incomplete description of a true state. The measured value in a quantum system is in general not predetermined but created by the act of measurement.

Equation (16.23) is just one single example of a large family of Bell’s inequalities derivable under the hypothesis of local realism. One frequently used inequality employs correlations for four observables A, B, C , and D , which can have two values ± 1 . Under the assumption of local realism, the correlation $\langle ab \rangle$ should be given by

$$\langle ab \rangle = \int d\lambda \rho(\lambda) a(\lambda) b(\lambda), \quad (16.28)$$

where λ denotes the “real state” and $\rho(\lambda)$ is a probability distribution with $\rho(\lambda) \geq 0$ and $\int d\lambda \rho(\lambda) = 1$. The quantity

$$\begin{aligned} S &= \langle cb \rangle + \langle cd \rangle + \langle ab \rangle - \langle ad \rangle \\ &= \int d\lambda \rho(\lambda) \{c(\lambda)[b(\lambda) + d(\lambda)] + a(\lambda)[b(\lambda) - d(\lambda)]\} \end{aligned} \quad (16.29)$$

obeys the inequality

$$|S| \leq 2, \quad (16.30)$$

because $|a(\lambda)| = |b(\lambda)| = |c(\lambda)| = |d(\lambda)| = 1$ and, hence, either $b(\lambda) + d(\lambda) = 0$ or $b(\lambda) - d(\lambda) = 0$.

On the other hand, the quantum theory of the two-photon system described above gives

$$\langle ab \rangle = \langle \Phi_2 | \sigma(\alpha) \otimes \sigma(\beta) | \Phi_2 \rangle = \cos(2(\alpha - \beta)) \quad (16.31)$$

and, with $\alpha = 0$, $\beta = \pi/8$, $\gamma = \pi/4$, and $\delta = 3\pi/8$, we obtain

$$S = 2\sqrt{2} > 2. \quad (16.32)$$

Again, quantum theory predicts a violation of local realism in accordance with experimental findings.

The inequalities of Bell type imply correlations and probabilities of measurement results. More recently, Greenberger, Horn, and Zeilinger gave a first example of a system for which a single measurement can distinguish between quantum theory and local realistic theories.

Consider the three-fold tensor product of a two-dimensional Hilbert space $\mathcal{H}_2$ with basis $|\varphi_1\rangle, |\varphi_2\rangle$, and $\sigma_3|\varphi_{1,2}\rangle = \pm|\varphi_{1,2}\rangle$. On $\mathcal{H} = \mathcal{H}_2 \otimes \mathcal{H}_2 \otimes \mathcal{H}_2$ we define the observables

$$A_1 = \sigma_1 \otimes \sigma_2 \otimes \sigma_2, \quad A_2 = \sigma_2 \otimes \sigma_1 \otimes \sigma_2, \quad A_3 = \sigma_2 \otimes \sigma_2 \otimes \sigma_1. \quad (16.33)$$

These three observables mutually commute with each other, for instance

$$A_1 A_2 = A_2 A_1 = \sigma_3 \otimes \sigma_3 \otimes \mathbf{1}, \quad (16.34)$$

and can be simultaneously diagonalized. One readily sees that

$$|\Psi\rangle = \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\varphi_1\rangle \otimes |\varphi_1\rangle - |\varphi_2\rangle \otimes |\varphi_2\rangle \otimes |\varphi_2\rangle) \quad (16.35)$$

is the only state in $\mathcal{H}$ with

$$A_1|\Psi\rangle = A_2|\Psi\rangle = A_3|\Psi\rangle = |\Psi\rangle. \quad (16.36)$$

This implies that for the observable

$$B = \sigma_1 \otimes \sigma_1 \otimes \sigma_1 \quad (16.37)$$

we have

$$B|\Psi\rangle = -|\Psi\rangle, \quad (16.38)$$

because $B = -A_1A_2A_3$.

According to a local realistic reasoning, in a true state the values $s_1^{(i)}$ and $s_2^{(i)}$, $i = 1, 2, 3$, for the observables σ_1 and σ_2 must be predetermined for each of the three particles. Altogether there are eight combinations for the values of $s_1^{(i)}$ and $s_2^{(i)}$ for which the three variables

$$a_1 = s_1^{(1)} s_2^{(2)} s_2^{(3)}, \quad a_2 = s_2^{(1)} s_1^{(2)} s_2^{(3)}, \quad a_3 = s_2^{(1)} s_2^{(2)} s_1^{(3)} \quad (16.39)$$

all have the value unity. We list the allowed combinations in Table 16.1. In all of these combinations we have

$$b = s_1^{(1)} s_1^{(2)} s_1^{(3)} = +1, \quad (16.40)$$

just the opposite to the value obtained from quantum theory.

Table 16.1: Allowed combinations such that $a_i = 1$

Particle 1		Particle 2		Particle 3	
$s_1^{(1)}$	$s_2^{(1)}$	$s_1^{(2)}$	$s_2^{(2)}$	$s_1^{(3)}$	$s_2^{(3)}$
+1	+1	+1	+1	+1	+1
+1	+1	-1	-1	-1	-1
-1	-1	+1	+1	-1	-1
-1	-1	-1	-1	+1	+1
+1	-1	-1	+1	-1	+1
-1	+1	+1	-1	-1	+1
-1	+1	-1	+1	+1	-1
+1	-1	+1	-1	+1	-1

The GHZ (Greenberger–Horn–Zeilinger) state $|\Psi\rangle$ of eq. (16.35) can in principle be constructed both for systems of three photons and for systems of three spin-1/2 particles. Meanwhile, other GHZ states with four particles have also been proposed. The actual preparation of a GHZ state remains an experimental challenge.

16.3 Quantum erasers and measurement without interaction

In the well-known double-slit experiment, the interference pattern of particle detection probabilities on the screen is produced by the interference of two amplitudes $\psi_1(\mathbf{x})$ and $\psi_2(\mathbf{x})$ corresponding to the transition of the particles through slits 1 and 2:

$$w(\mathbf{x}) \sim \frac{1}{2} |\psi_1(\mathbf{x}) + \psi_2(\mathbf{x})|^2. \quad (16.41)$$

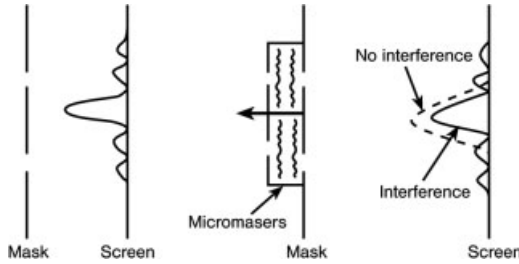


Figure 16.1: Double-slit experiment and quantum eraser.

The interference pattern disappears if “which path” information (which determines through which of the holes the particles actually passed) is available. A conceptually very clear although not yet realized way to achieve this is to perform the double-slit experiment with incoming excited atoms and to place micromasers in the vacuum state in front of each of the holes. The micromasers are tuned such that an atom passing them becomes de-excited with unit probability and is not deflected in any way by the micromaser. Then an atom will leave a photon and “which path” information in one of the two cavities, and the interference pattern on the screen will disappear. On the side of the screen, the total state of the atom and the two micromasers will be given by

$$|\Psi\rangle \sim \frac{1}{\sqrt{2}} |1, 0\rangle \psi_1(\mathbf{x}) + \frac{1}{\sqrt{2}} |0, 1\rangle \psi_2(\mathbf{x}), \quad (16.42)$$

and because of the orthogonality of the resonator states $|1, 0\rangle$ and $|0, 1\rangle$, we have indeed

$$\| |\Psi\rangle \|^2 \sim \frac{1}{2} [|\psi_1(\mathbf{x})|^2 + |\psi_2(\mathbf{x})|^2], \quad (16.43)$$

without interference. Note that it is not necessary to look up in which of the cavities the photon is really to be found. For the vanishing of the interference pattern, it is sufficient that the “which path” information is available in principle. By removing the separating wall between the two cavities, the “which path” information can be destroyed and the interference pattern reappears. This is the principle of the so-called *quantum eraser*.

After removing the separating wall the state is

$$|\Psi'\rangle \sim \frac{1}{\sqrt{2}} [|1\rangle \psi_1(\mathbf{x}) + |1\rangle \psi_2(\mathbf{x})], \quad (16.44)$$

and

$$\| |\Psi'\rangle \|^2 \sim \frac{1}{2} |\psi_1(\mathbf{x}) + \psi_2(\mathbf{x})|^2. \quad (16.45)$$

The “which path” information can be erased when the atoms have already passed the double-slit mask and are on their way to the screen.

One can even go one step further. Introducing the symmetrized and anti-symmetrized states

$$|\Phi_{\pm}\rangle = \frac{1}{\sqrt{2}}(|1, 0\rangle \pm |0, 1\rangle) \quad \text{and} \quad \psi_{\pm}(\mathbf{x}) = \frac{1}{\sqrt{2}}[\psi_1(\mathbf{x}) \pm \psi_2(\mathbf{x})], \quad (16.46)$$

we can write the state $|\Psi\rangle$ of eq. (16.42) in the form

$$|\Psi\rangle \sim \frac{1}{\sqrt{2}} [|\Phi_+\rangle \psi_+(\mathbf{x}) + |\Phi_-\rangle \psi_-(\mathbf{x})]. \quad (16.47)$$

Rather than determining in which cavity the photon is found, one can measure whether the two cavities are in the state $|\Phi_+\rangle$ or $|\Phi_-\rangle$. After such a measurement, the distribution on the screen will be

$$w_{\pm}(\mathbf{x}) \sim \frac{1}{2} |\psi_{\pm}|^2 = \frac{1}{2} |\psi_1(\mathbf{x}) \pm \psi_2(\mathbf{x})|^2. \quad (16.48)$$

A measurement of $|\Phi_+\rangle$ or $|\Phi_-\rangle$ erases the “which path” information and, correspondingly, $w_{\pm}(\mathbf{x})$ shows an interference pattern. The interference patterns $w_{\pm}(\mathbf{x})$ are shifted with respect to each other such that

$$\frac{1}{2} [w_+(\mathbf{x}) + w_-(\mathbf{x})] \sim \frac{1}{2} [|\psi_1(\mathbf{x})|^2 + |\psi_2(\mathbf{x})|^2]. \quad (16.49)$$

The sum shows no interference pattern. If an observer at the screen knows the result of the $|\Phi_{\pm}\rangle$ measurement, he will observe the distributions $w_{\pm}(\mathbf{x})$. If he does not know the result, he will observe the distribution of eq. (16.49) without interference. But even after measuring this distribution, the interference can be recovered. If the observer has carefully registered the order and position of the atoms impacting on the screen and if, in hindsight, he learns about the outcomes of the $|\Phi_{\pm}\rangle$ measurements, he can perform a classification of the atoms hitting the screen into a (+) class and a (−) class. Sampling the distributions of the classes separately, he will rediscover interference.

Quantum erasers have been realized in various forms. In 2002, S. P. Walborn, M. O. Terra Cunha, S. Pádua, and C. H. Monken (2002 *Phys. Rev. A* **65**, 033818) for the first time constructed a quantum eraser based on the double-slit experiment. It works with polarized photons.

In addition to a pair of orthogonal linear polarization states $|x\rangle$, $|y\rangle$, another oblique orthogonal pair

$$\begin{aligned} |x'\rangle &= \frac{1}{\sqrt{2}}(|x\rangle + |y\rangle), & |y'\rangle &= \frac{1}{\sqrt{2}}(|x\rangle - |y\rangle), \\ |x\rangle &= \frac{1}{\sqrt{2}}(|x'\rangle + |y'\rangle), & |y\rangle &= \frac{1}{\sqrt{2}}(|x'\rangle - |y'\rangle), \end{aligned} \quad (16.50)$$

of polarization states, as well as circular polarization states

$$|R\rangle = \frac{1}{\sqrt{2}}(|x'\rangle + i|y'\rangle), \quad |L\rangle = \frac{1}{\sqrt{2}}(|x'\rangle - i|y'\rangle) \quad (16.51)$$

are employed. The “which path” information is provided by two differently oriented $\lambda/4$ plates immediately behind the holes, which are described by unitary operators U_1 and U_2 with

$$\begin{aligned} U_1|x'\rangle &= |x'\rangle, & U_1|y'\rangle &= i|y'\rangle, \\ U_2|x'\rangle &= i|x'\rangle, & U_2|y'\rangle &= |y'\rangle. \end{aligned} \quad (16.52)$$

Hence,

$$\begin{aligned} U_1|x\rangle &= |R\rangle, & U_1|y\rangle &= |L\rangle, \\ U_2|x\rangle &= i|L\rangle, & U_2|y\rangle &= i|R\rangle. \end{aligned} \quad (16.53)$$

An incident photon in the polarization state $|x'\rangle$ or $|y'\rangle$ emerges from the double-slit apparatus in the state

$$|\Psi_{x'}\rangle = |x'\rangle(\psi_1 + i\psi_2) \quad \text{or} \quad |\Psi_{y'}\rangle = |y'\rangle i(\psi_1 - i\psi_2). \quad (16.54)$$

These states carry no “which path” information and show complementary interference patterns.

On the other hand, planar polarization states $|x\rangle$ or $|y\rangle$ emerge as

$$|\Psi_x\rangle = |R\rangle\psi_1 + i|L\rangle\psi_2 \quad \text{or} \quad |\Psi_y\rangle = |L\rangle\psi_1 + i|R\rangle\psi_2. \quad (16.55)$$

Now the polarization state of the photon carries the “which path” information and there is no interference pattern:

$$\| |\Psi_x\rangle \|^2 = \| |\Psi_y\rangle \|^2 \sim |\psi_1|^2 + |\psi_2|^2. \quad (16.56)$$

The “which path” information can be erased by placing polarization filters with equal orientation behind the $\lambda/4$ plates.

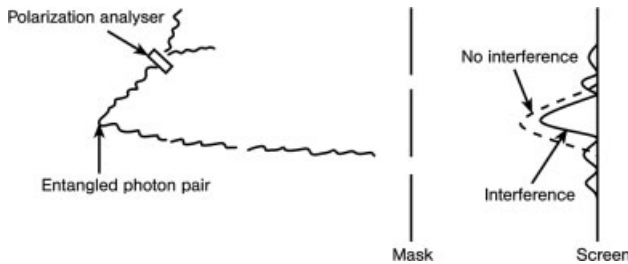


Figure 16.2: Double-slit experiment and quantum eraser with entangled pairs.

A more elegant realization of an eraser employed by the authors mentioned above uses an entangled photon pair in a Bell state (see fig. 16.2). One photon falls on the double-slit instrument, and the other one is free for further measurement and manipulation. The Bell state

$$|\Phi\rangle = \frac{1}{\sqrt{2}}(|x\rangle \otimes |x\rangle + |y\rangle \otimes |y\rangle) = \frac{1}{\sqrt{2}}(|x'\rangle \otimes |x'\rangle + |y'\rangle \otimes |y'\rangle) \quad (16.57)$$

is transformed into

$$\begin{aligned} |\Psi\rangle &= \frac{1}{\sqrt{2}}|x\rangle \otimes (|R\rangle\psi_1 + i|L\rangle\psi_2) + \frac{1}{\sqrt{2}}|y\rangle \otimes (|L\rangle\psi_1 + i|R\rangle\psi_2) \\ &= \frac{1}{\sqrt{2}}|x'\rangle \otimes (|x'\rangle\psi_1 + |x'\rangle i\psi_2) + \frac{1}{\sqrt{2}}|y'\rangle \otimes (|y'\rangle i\psi_1 + |y'\rangle\psi_2) \\ &= \frac{1}{\sqrt{2}}(|x'\rangle \otimes |x'\rangle + i|y'\rangle \otimes |y'\rangle)\psi_1 + \frac{i}{\sqrt{2}}(|x'\rangle \otimes |x'\rangle - i|y'\rangle \otimes |y'\rangle)\psi_2. \end{aligned} \quad (16.58)$$

The state $|\Psi\rangle$ gives no interference pattern on the screen because it carries “which path” information. Measuring the xy polarization of the additional photon yields the states $|\Psi_x\rangle$ and $|\Psi_y\rangle$ of eq. (16.55), which still carry “which path” information, whereas measuring the $x'y'$ polarization gives the states $|\Psi_{x'}\rangle$ and $|\Psi_{y'}\rangle$, for which the “which path” information is erased. As above, this measurement can be delayed, or its results can be used later to recover the interference pattern.

Measurement without interaction is another very surprising and interesting application of “which path” information. Consider a Mach–Zehnder interferometer with two symmetric beam splitters as depicted in fig. 16.3. Let the beam acquire a phase jump of $\pi/2$ on each reflection and no phase jump on transmission.

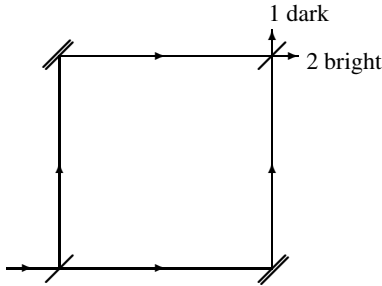


Figure 16.3: Mach–Zehnder interferometer.

Then, by destructive interference, an incoming photon will never emerge at the “dark” port 1 but always at the “bright” port 2. Now we block one of the paths by placing an absorber with 100% efficiency in it. (In the original proposal, the absorber is a bomb that explodes when it is hit by a photon!) The absorber now provides the “which path” information because the blocked path cannot be used. With the absorber, 50% of the photons will emerge at each of the ports 1 and 2.

Imagine now that an observer at the ports 1 and 2 does not know whether the absorber is in place. With 50% probability no photon will arrive at port 1 or 2. In this case (and also from the explosion!) the observer knows that the absorber was there and was hit by a photon. With

50% probability the observer will see a photon at either port 1 or 2 and with 25% probability each. If the photon arrives at the bright port 1, the observer cannot tell whether the absorber is in place. However, if the photon arrives at the dark port 1, the observer knows that the absorber was there, although he also knows that it was not hit by a photon. So, with 25% efficiency it is possible to secure the presence of the absorber without hitting it with a single photon.

The detection efficiency can even be pushed arbitrarily close to 100% by the following device (see fig. 16.4). Take a cavity with mirrors on both ends, which is divided into two halves by an asymmetric beam splitter with high reflectivity. A photon oscillating back and forth will meet the beam splitter at regular time intervals.

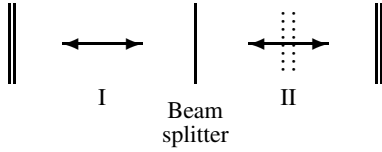


Figure 16.4: Detection without interaction.

Let

$$|\varphi_1\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad |\varphi_2\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

be the states representing a photon in, respectively, the left-hand part I and right-hand part II of the resonator. Let the photon be initially in part I. The beam splitter will be represented by a unitary matrix U of the form

$$U = \begin{pmatrix} \cos \varphi & i \sin \varphi \\ i \sin \varphi & \cos \varphi \end{pmatrix}. \quad (16.59)$$

Now, evidently, the N th power of U will satisfy

$$U^N \begin{pmatrix} 1 \\ 1 \end{pmatrix} = e^{iN\varphi} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad U^N \begin{pmatrix} 1 \\ -1 \end{pmatrix} = e^{-iN\varphi} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad (16.60)$$

and

$$U^N \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos(N\varphi) \\ i \sin(N\varphi) \end{pmatrix}. \quad (16.61)$$

This means that the probability of finding the photon in part I of the cavity after N interactions with the beam splitter is given by

$$w_I(N) = \cos^2(N\varphi). \quad (16.62)$$

Choosing $\varphi = \pi/2N$, we have $w_I(N) = 0$.

Now we place an absorber into part II, which absorbs a photon in part II with 100% efficiency and on absorption gives a clear signal (for example, an explosion). Then the lack of this signal corresponds to a measurement asserting that the photon is still in part I. After the first interaction, the probability for this is

$$\overline{w}_I(1) = \cos^2 \varphi = \cos^2 \left(\frac{\pi}{2N} \right), \quad (16.63)$$

and after N interactions with the beam splitter

$$\overline{w}_I(N) = \left[\cos\left(\frac{\pi}{2N}\right) \right]^{2N}. \quad (16.64)$$

Now

$$\lim_{N \rightarrow \infty} \left[\cos\left(\frac{\pi}{2N}\right) \right]^{2N} = 1.$$

Indeed,

$$\begin{aligned} \ln \left[\cos\left(\frac{\pi}{2N}\right) \right]^{2N} &= 2N \ln \left[\cos\left(\frac{\pi}{2N}\right) \right] \\ &= 2N \ln \left[1 - \frac{\pi^2}{8N^2} + \cdots \right] \\ &= \frac{\pi^2}{4N} + O\left(\frac{1}{N^2}\right). \end{aligned}$$

Hence, $\overline{w}_I(N)$ can be pushed arbitrarily close to 1 as opposed to $w_I(N) = 0$. This means that from $\overline{w}_I(N) \neq 0$ we can tell with certainty that the absorber is in place and that we can detect the absorber with arbitrarily high efficiency without hitting it with a single photon.

The mechanism to achieve $\lim_{N \rightarrow \infty} \overline{w}_I(N) = 1$ is known by the name of the *quantum Zeno effect*. It is possible to suppress a quantum-mechanical transition by repeatedly measuring whether it has occurred. The device just described can in principle be improved to photograph an object without a photon hitting it or to take an X-ray picture with zero radiation doses.

16.4 No cloning and quantum teleportation

In quantum theory, it is, at least in principle, possible to prepare any prescribed admissible quantum state of a system Σ . On the other hand, the identification of an unknown state $|\psi\rangle$ of Σ is subject to severe restrictions, because a measurement is always connected to a projection of $|\psi\rangle$ onto some subspace and, hence, generically with a change of $|\psi\rangle$ such that $|\psi\rangle$ cannot be completely reconstructed from the result of a measurement.

If the system is known to reside in one of the eigenstates of an observable

$$A = \sum_i |\varphi_i\rangle a_i \langle \varphi_i|, \quad a_i \neq a_j \text{ for } i \neq j, \quad (16.65)$$

then a measurement of A can identify this eigenstate. But in general, applying a measurement of A to an unknown state $|\psi\rangle$ will identify a component of $|\psi\rangle$ in exactly one of the directions $|\varphi_i\rangle$, whereas all other possible information about $|\psi\rangle$ is lost after the measurement. If and only if the state $|\psi\rangle$ can be produced repeatedly as many times as desirable, then, by an appropriate sequence of measurements, $|\psi\rangle$ can be determined to arbitrary precision. Related to this situation, quantum theory excludes the existence of a “state measurement apparatus” in the same sense that measuring devices exist for position or momentum.

To see this, let us briefly enter into the formal quantum-theoretical description of the measurement of an observable A like eq. (16.65). The Hilbert space of the system “measuring apparatus plus measured system Σ ” is a tensor product

$$\mathcal{H} = \mathcal{H}_{\text{AP}} \otimes \mathcal{H}_{\Sigma}. \quad (16.66)$$

Let $|\Phi\rangle \in \mathcal{H}_{\text{AP}}$ be the zero state of the measuring apparatus. For an ideal measurement of the observable A , the interaction with the measuring apparatus is described by a unitary operator

$$U : \mathcal{H} \rightarrow \mathcal{H}, \quad (16.67)$$

such that

$$U(|\Phi\rangle \otimes |\varphi_i\rangle) = |\Phi_{a_i}\rangle \otimes |\varphi_i\rangle \quad (16.68)$$

with

$$\langle \Phi_{a_i} | \Phi_{a_j} \rangle = 0 \quad \text{for } a_i \neq a_j. \quad (16.69)$$

An initial pure product state

$$|\Psi\rangle = |\Phi\rangle \otimes \sum_i c_i |\varphi_i\rangle \quad (16.70)$$

will go over into an entangled state

$$|\Psi'\rangle = U|\Psi\rangle = \sum_i c_i |\Phi_{a_i}\rangle \otimes |\varphi_i\rangle, \quad (16.71)$$

and after reduction to $\mathcal{H}_{\text{AP}}$, the density matrix will be

$$\rho = \text{Tr}_2 |\Psi'\rangle \langle \Psi'| = \sum_i |\Phi_{a_i}\rangle |c_i|^2 \langle \Phi_{a_i}|, \quad (16.72)$$

which just contains the standard quantum-theoretical probabilities $w_i = |c_i|^2$ to measure the value a_i .

A “state measurement apparatus” would have to correspond to a unitary operator U with

$$U|\Psi_\varphi\rangle = U(|\Phi\rangle \otimes |\varphi\rangle) = |\Phi_\varphi\rangle \otimes |\varphi\rangle \quad (16.73)$$

for all $|\varphi\rangle \in \mathcal{H}_{\Sigma}$ and

$$\langle \Phi_{\varphi'} | \Phi_\varphi \rangle = 0 \quad \text{for } |\varphi\rangle \neq |\varphi'\rangle. \quad (16.74)$$

Such a unitary operator cannot exist, because for $\langle \varphi | \varphi' \rangle \neq 0$ we would have

$$\langle \Psi_\varphi | \Psi_{\varphi'} \rangle = \langle \varphi | \varphi' \rangle \neq 0 \quad \text{but} \quad \langle U\Psi_\varphi | U\Psi_{\varphi'} \rangle = 0. \quad (16.75)$$

The so-called *no-cloning theorem* imposes another serious limitation on manipulating quantum states.

Theorem 16.1 (no-cloning theorem). *If a state $|\varphi\rangle \in \mathcal{H}_\Sigma$ can be prepared, it can also be prepared many times. It is, however, impossible in principle to build a universal “state doubling” or cloning machine that transforms every state $|\varphi\rangle \in \mathcal{H}_\Sigma$ into a doubled state $|\varphi\rangle \otimes |\varphi\rangle \in \mathcal{H}_\Sigma \otimes \mathcal{H}_\Sigma$.*

Proof: The proof of this important theorem is simple. A cloning machine would have to correspond to a linear operator V with

$$V(|\Phi\rangle \otimes |\varphi\rangle) = |\Phi'\rangle \otimes |\varphi\rangle \otimes |\varphi\rangle \quad \text{for all } |\varphi\rangle \in \mathcal{H}. \quad (16.76)$$

Such a linear operator is impossible because it would have to fulfill, on the one hand,

$$\begin{aligned} V(|\Phi\rangle \otimes (|\varphi_1\rangle + |\varphi_2\rangle)) &= V(|\Phi\rangle \otimes |\varphi_1\rangle) + V(|\Phi\rangle \otimes |\varphi_2\rangle) \\ &= |\Phi'\rangle \otimes |\varphi_1\rangle \otimes |\varphi_1\rangle + |\Phi''\rangle \otimes |\varphi_2\rangle \otimes |\varphi_2\rangle, \end{aligned} \quad (16.77)$$

and, on the other hand,

$$V(|\Phi\rangle \otimes (|\varphi_1\rangle + |\varphi_2\rangle)) = |\Phi'''\rangle \otimes (|\varphi_1\rangle + |\varphi_2\rangle) \otimes (|\varphi_1\rangle + |\varphi_2\rangle). \quad (16.78)$$

Obviously, eqs. (16.77) and (16.78) cannot be identical for arbitrary $|\varphi_1\rangle, |\varphi_2\rangle \in \mathcal{H}_\Sigma$. There are variants and generalizations of the no-cloning theorem to mixed states. $\square$

The impossibility of cloning is very deeply rooted in the foundations of quantum theory. In fact, if an unknown polarization state of a photon could be cloned, then the cloning could be repeated many times and the state could be converted into a classical light ray whose Stokes parameters, as described in chapter 12, could be measured, thus determining the unknown polarization state completely.

The no-cloning theorem only excludes a universal cloning machine cloning any unidentified state. States that can be identified by a single measurement can be cloned. If, for instance, the state is known to be a member of an orthonormal system, a measurement of the observable A in eq. (16.65) will identify it, and subsequently it can be prepared as many times as desired. Indeed, for an orthonormal system $\{|\varphi_i\rangle\}_{i \in I}$, there exist unitary operators

$$U : |\psi\rangle \otimes |\varphi_i\rangle \rightarrow |\varphi_i\rangle \otimes |\varphi_i\rangle. \quad (16.79)$$

In view of the limitations described above, it may be surprising but satisfying to hear that an unknown state $|\varphi\rangle \in \mathcal{H}_\Sigma$ can at least be transported from one place q_1 to another place q_2 without analyzing it (at the expense of annihilating it at q_1). The mechanism, which makes essential use of entanglement, goes by the name of *quantum teleportation*. Let us first describe it for a two-dimensional Hilbert space $\mathcal{H}_2$, where $|\varphi_1\rangle$ and $|\varphi_2\rangle$ are an orthonormal basis of $\mathcal{H}_2$.

First, one of the four Bell states in $\mathcal{H}_2 \otimes \mathcal{H}_2$,

$$\begin{aligned} |\Phi_1\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\varphi_2\rangle - |\varphi_2\rangle \otimes |\varphi_1\rangle), \\ |\Phi_2\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\varphi_2\rangle + |\varphi_2\rangle \otimes |\varphi_1\rangle), \\ |\Phi_3\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\varphi_1\rangle - |\varphi_2\rangle \otimes |\varphi_2\rangle), \\ |\Phi_4\rangle &= \frac{1}{\sqrt{2}}(|\varphi_1\rangle \otimes |\varphi_1\rangle + |\varphi_2\rangle \otimes |\varphi_2\rangle), \end{aligned} \quad (16.80)$$

is produced such that one member of the entangled pair remains at q_1 and the other one is sent to q_2 . For definiteness, we take the state $|\Phi_4\rangle$. Next we go over to the state

$$\begin{aligned} |\Psi\rangle &= |\varphi\rangle \otimes |\Phi_4\rangle = \frac{1}{\sqrt{2}}(|\varphi\rangle \otimes |\varphi_1\rangle \otimes |\varphi_1\rangle + |\varphi\rangle \otimes |\varphi_2\rangle \otimes |\varphi_2\rangle) \\ &\in \mathcal{H}_2 \otimes \mathcal{H}_2 \otimes \mathcal{H}_2, \end{aligned} \quad (16.81)$$

where

$$|\varphi\rangle = c_1|\varphi_1\rangle + c_2|\varphi_2\rangle \quad (16.82)$$

is the unknown state to be teleported and assumed to reside at the place q_1 . Expanding the state $|\Psi\rangle$ in a Bell state basis with respect to the first two factors of the tensor product, we see that $|\Psi\rangle$ is also given by

$$\begin{aligned} |\Psi\rangle &= \frac{1}{2}\{|\Phi_1\rangle \otimes (-c_2|\varphi_1\rangle + c_1|\varphi_2\rangle) + |\Phi_2\rangle \otimes (c_2|\varphi_1\rangle + c_1|\varphi_2\rangle) \\ &\quad + |\Phi_3\rangle \otimes (c_1|\varphi_1\rangle - c_2|\varphi_2\rangle) + |\Phi_4\rangle \otimes (c_1|\varphi_1\rangle + c_2|\varphi_2\rangle)\}. \end{aligned} \quad (16.83)$$

Now the observer at q_1 performs a measurement with respect to the Bell state basis in the first two factors. Depending on the outcome of this measurement, the state $|\psi_i\rangle$ at q_2 will be the following:

$$\begin{aligned} \text{for } |\Phi_1\rangle: \quad & |\psi_1\rangle = -c_2|\varphi_1\rangle + c_1|\varphi_2\rangle = U_1|\varphi\rangle, \\ \text{for } |\Phi_2\rangle: \quad & |\psi_2\rangle = c_2|\varphi_1\rangle + c_1|\varphi_2\rangle = U_2|\varphi\rangle, \\ \text{for } |\Phi_3\rangle: \quad & |\psi_3\rangle = c_1|\varphi_1\rangle - c_2|\varphi_2\rangle = U_3|\varphi\rangle, \\ \text{for } |\Phi_4\rangle: \quad & |\psi_4\rangle = c_1|\varphi_1\rangle + c_2|\varphi_2\rangle = U_4|\varphi\rangle = |\varphi\rangle. \end{aligned} \quad (16.84)$$

The states $|\psi_i\rangle$ ($i = 1, \dots, 4$) differ from the unknown state $|\varphi\rangle$ to be teleported by unitary transformations U_i ($i = 1, \dots, 4$), which are independent of $|\varphi\rangle$. Now the observer at q_1 informs the observer at q_2 about the outcome $i = 1, \dots, 4$ of his Bell basis measurement by a normal classical communication channel and asks him to apply a unitary transformation U_i^{-1} to the state $|\psi_i\rangle$ present at q_2 . As a result, the state $|\varphi\rangle$ will be present at q_2 , but no longer at q_1 .

Notice that the states $|\psi_i\rangle$ ($i = 1, \dots, 4$) are not orthogonal. If they were orthogonal, they could be distinguished by a single measurement, the classical information channel would be superfluous, and the apparatus could be used to transmit information faster than light.

For state spaces of arbitrary dimension N , this teleportation scheme can be generalized in the following way. Let $(|\varphi_n\rangle)_{n=1,\dots,N}$ be a basis of the Hilbert space $\mathcal{H}$. The orthonormal system

$$|\Phi_{m,n}\rangle = \frac{1}{\sqrt{N}} \sum_{j=1}^N \zeta^{jn} |\varphi_j\rangle \otimes |\varphi_{j+m}\rangle \quad (16.85)$$

with

$$\zeta = e^{2\pi i/N}$$

generalizes the Bell state basis to an orthonormal basis of $\mathcal{H}_N \otimes \mathcal{H}_N$. The sum $j+m$ is to be understood modulo N . The states $|\Phi_{mn}\rangle$ can be written in the form

$$|\Phi_{mn}\rangle = (\mathbf{1} \otimes U_{mn}) |\Phi_{00}\rangle, \quad (16.86)$$

where the unitary operator U_{mn} with

$$U_{mn} |\varphi_j\rangle = \zeta^{jn} |\varphi_{j+m}\rangle \quad (16.87)$$

is the product

$$U_{mn} = S_m P_n \quad (16.88)$$

of a phase operator

$$P_n |\varphi_j\rangle = \zeta^{jn} |\varphi_j\rangle \quad (16.89)$$

and a shift operator

$$S_m |\varphi_j\rangle = |\varphi_{j+m}\rangle. \quad (16.90)$$

Teleportation is now performed by projecting

$$|\Psi\rangle = |\varphi\rangle \otimes |\Phi_{0,0}\rangle = \sum_{k=1}^N c_k |\varphi_k\rangle \otimes |\Phi_{0,0}\rangle \quad (16.91)$$

$$\in \mathcal{H}_N \otimes \mathcal{H}_N \otimes \mathcal{H}_N$$

onto $|\Phi_{m,n}\rangle$ in the first two factors of the three-fold tensor product. Then the state left at q_2 will be

$$|\psi_{m,n}\rangle = \sum_k c_k \zeta^{-kn} |\varphi_{k+m}\rangle = U_{m,-n} |\varphi\rangle, \quad (16.92)$$

and $|\varphi\rangle$ can be recovered at q_2 by on demand undoing the unitary transformation $U_{m,-n}$. Again, the states $|\psi_{m,n}\rangle$ are not orthogonal.

16.5 Quantum cryptography

In quantum theory, a measurement of an observable will in general change the state of the system. More precisely, if the system is known to reside in an eigenstate of an observable A , the subsequent measurement of a complementary observable B will change the state. The fact that a measurement will in general leave a trace can be used to establish communication channels for which an eavesdropper can be recognized with arbitrary efficiency. Such channels are ideal for handing over information like the key for data encryption for future communication which is itself meaningless but must be transmitted such that eavesdropping can be excluded.

An emitter E measures at random one of several complementary observables $A, B, C, \dots$ of a quantum system Σ , thus preparing the system in an eigenstate of the observable just measured, and subsequently sends the system to a receiver R . Just for definiteness, we assume that the emitter sends photons to R and that the observables are planar polarization observables along or orthogonal to one of several mutually oblique directions.

The emitter does not tell which observable he measured each time and what the results of the measurements were. The receiver performs measurements with the photons arriving at him whereby he also chooses at random each time one member of the same set of complementary observables. After the transmission of a large number of photons, E and R exchange information about which observable they measured for each photon. This may be done by a public insecure channel. If they happened to measure the same observable, their results, which they still keep secret, are guaranteed to agree unless some eavesdropper or any other disturbance has intervened. To exclude such an unwelcome incident, they use part of the photons for which they measured the same observable as a control by comparing the results of their measurements, possibly even in public. If there are no discrepancies for a sufficiently large number of cases, they can be sure that no disturbance has occurred, and they can use the results from the other photons for which they measured the same observable to fix the key they wanted to exchange.

If there are discrepancies, they must suspect that an eavesdropper has performed measurements on the exchanged photons. Not knowing at the time of his measurement which observable the emitter had chosen, the eavesdropper will inevitably have taken the wrong observable many times and will have changed the photon states. In this case, E and R reject their data and try to communicate again.

The communication system can be improved by taking photon pairs in a Bell state rather than single photons. The emitter produces Bell states, for instance,

$$|\Phi_2\rangle = \frac{1}{\sqrt{2}}(|x\rangle|x\rangle + |y\rangle|y\rangle), \quad (16.93)$$

keeps one of the photons, and sends the other one to the receiver. Now, at least in principle, the photons can be stored and the measurements can be performed when they need a common key. This will reduce problems with the safe storage of a secret common key.

The transmission of information and the check for eavesdropping can proceed as described. If E and R measure the same observable, they are guaranteed to find the same result unless some disturbance or intervention has occurred.

The use of photon pairs in a Bell state opens up an additional possibility to secure oneself against eavesdropping. The complementary observables may be chosen such that a maximal

violation of a Bell-type inequality results. Now E and R may use the photon pairs for which they measured different observables to check whether the violation of the inequality is still maximal. The intervention of an eavesdropper will be detectable from deviations from maximal violation.

As an example, let us consider the Bell inequality

$$w(a_1, b_1) \leq w(a_1, c_1) + w(b_1, c_2), \quad (16.94)$$

derived and explained in section 16.2. For photon pairs in a Bell state $|\Phi_2\rangle$ we found

$$w(a_1, b_1) = \frac{1}{2} \cos^2(\alpha - \beta) \quad \text{and} \quad w(a_1, b_2) = \frac{1}{2} \sin^2(\alpha - \beta). \quad (16.95)$$

The inequality is maximally violated for $\alpha = 0$, $\beta = \pi/6$, and $\gamma = \pi/3$:

$$T = w(a_1, b_1) - w(a_1, c_1) - w(b_1, c_2) = \frac{1}{8}. \quad (16.96)$$

If an eavesdropper has performed a measurement of any polarization observable X at one of the photons in the Bell state, the state will subsequently be given by the density matrix:

$$\rho = \frac{1}{2}|x\rangle \otimes |x\rangle \langle x| + \frac{1}{2}|y\rangle \otimes |y\rangle \langle y| \otimes \langle y|. \quad (16.97)$$

From this we readily obtain the new pair probabilities

$$\begin{aligned} w_X(a_1, b_1) &= \frac{1}{2}(\cos^2 \alpha \cos^2 \beta + \sin^2 \alpha \sin^2 \beta), \\ w_X(a_1, b_2) &= \frac{1}{2}(\cos^2 \alpha \sin^2 \beta + \sin^2 \alpha \cos^2 \beta), \end{aligned} \quad (16.98)$$

with

$$w_X(a_1, b_1) + w_X(a_1, b_2) = \frac{1}{2}, \quad (16.99)$$

as necessary. With $\alpha = 0$, $\beta = \pi/6$, and $\gamma = \pi/3$ we have

$$T_X = w_X(a_1, b_1) - w_X(a_1, c_1) - w_X(b_1, c_2) = \frac{3}{8} - \frac{1}{8} - \frac{5}{16} = -\frac{1}{16}. \quad (16.100)$$

Bell's inequality is not violated any more. So, even if the eavesdropper does not intervene each time, his activity will be noticed by a reduction of the value of T below $1/8$.

In fact, the interactionless detection mechanism described in the previous section can be interpreted such that the absorber in the right-hand part of the cavity is a measuring instrument detecting the presence of a photon there. Even an unsuccessful measurement leaves a trace betraying the existence of the measurement apparatus.

16.6 Quantum computation

From a bird's eye perspective, a classical computer is a large collection of N classical two-state systems, called *bits*. Denoting the state of each bit by 0 or $1 \in \mathbb{Z}_2$, the total state of a computer, which also includes program and data, is given by an N -tuple:

$$(x) = (x_1, \dots, x_n) \in \mathbb{Z}_2^N. \quad (16.101)$$

With the addition,

$$0 + 0 = 1 + 1 = 0, \quad 0 + 1 = 1 + 0 = 1, \quad (16.102)$$

and the multiplication,

$$0 \cdot 0 = 0 \cdot 1 = 1 \cdot 0 = 0, \quad 1 \cdot 1 = 1, \quad (16.103)$$

$\mathbb{Z}_2$ is a field and $\mathbb{Z}_2^N$ an N -dimensional vector space over $\mathbb{Z}_2$ with 2^N different vectors. Sometimes it is advantageous to identify (x) with the integer

$$n_x = \sum_{i=0}^{N-1} x_{N-1-i} 2^i, \quad 0 \leq n_x < 2^N, \quad (16.104)$$

and to label the state by the number $n_x : (x) = n_x$.

Computation consists in repeated application of Boolean functions,

$$f : \mathbb{Z}_2^N \rightarrow \mathbb{Z}_2^N, \quad (x) \rightarrow (f(x)) = (f_1(x), \dots, f_N(x)), \quad (16.105)$$

to the state (x) . The component functions f_i are Boolean functions,

$$f_i : \mathbb{Z}_2^N \rightarrow \mathbb{Z}_2, \quad (16.106)$$

with values in $\mathbb{Z}_2$, to which we can restrict ourselves for the following discussion.

There are exactly four $\mathbb{Z}_2$ -valued Boolean functions of one argument, namely the constant functions with values 0 and 1, the identity function and the negation function:

$$\text{NOT}(x) = \neg x = x + 1. \quad (16.107)$$

There are $2^4 = 16$ Boolean functions of two arguments, for instance,

$$\begin{aligned} \text{AND}(x, y) &= x \wedge y = xy, \\ \text{OR}(x, y) &= x \vee y = x + y + xy, \\ \text{NAND}(x, y) &= \neg(x \wedge y) = xy + 1, \\ \text{NOR}(x, y) &= \neg(x \vee y) = (x + 1)(y + 1), \\ \text{XOR}(x, y) &= x \vee y \wedge \neg(x \wedge y) = x + y. \end{aligned} \quad (16.108)$$

There are 2^n $\mathbb{Z}_2$ -valued Boolean functions of n arguments. Every $\mathbb{Z}_2$ -valued Boolean function can be constructed from the basic functions $\wedge$, $\vee$, and $\neg$, for example:

$$f(x, y) = f_{00} \wedge \neg x \wedge \neg y \vee f_{01} \wedge \neg x \wedge y \vee f_{10} \wedge x \wedge \neg y \vee f_{11} \wedge x \wedge y, \quad (16.109)$$

from which the generalization to an arbitrary number of arguments is immediate.

Because of the identity

$$x \vee y = \neg(\neg x \wedge \neg y), \quad (16.110)$$

the set of basic functions can be reduced to two functions $\wedge$ and $\neg$. It can even be reduced to the single function NAND, because

$$\begin{aligned}\neg x &= \text{NAND}(x, x), \\ x \wedge y &= \neg \text{NAND}(x, y) = \text{NAND}(\text{NAND}(x, y), \text{NAND}(x, y)).\end{aligned}\quad (16.111)$$

We have mentioned these simple facts of Boolean algebra just in order to contrast them to the more complicated quantum computation. Quantum computers will eventually arise if the miniaturization of the computer elements is pushed to its extreme. The fundamental unit of a quantum computer is a *quantum bit* or *qubit*, a quantum system with a two-dimensional Hilbert space $\mathcal{H}_2$. Let $|0\rangle$ and $|1\rangle$ be an orthonormal basis of $\mathcal{H}_2$. The quantum state space of a qubit consists of a continuous infinity of states, which are given by Hilbert space vectors

$$|\varphi\rangle = c_0|0\rangle + c_1|1\rangle \neq 0, \quad c_1, c_2 \in \mathbb{C}, \quad (16.112)$$

such that for $\alpha \neq 0$, $|\varphi\rangle$ and $\alpha|\varphi\rangle$ label the same quantum state. In this way, the set of quantum states of a qubit turns out to be isomorphic to the two-sphere S^2 . A quantum computer is a large arrangement of N quantum bits and the total set of states of the quantum computer is described by the 2^N -dimensional Hilbert space:

$$\mathcal{H}_2^{\otimes N} = \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_2 \quad (N \text{ factors}). \quad (16.113)$$

An orthonormal basis of $\mathcal{H}_2^{\otimes N}$ is given by the vectors

$$|x\rangle = |x_1, \dots, x_N\rangle = |x_1\rangle \otimes \cdots \otimes |x_N\rangle, \quad x_i \in \mathbb{Z}_2. \quad (16.114)$$

The states associated to the basis vectors are called *Boolean quantum states*.

Computation is a quantum-dynamical process that consists of repeated application of unitary transformations:

$$U : \mathcal{H}_2^{\otimes N} \rightarrow \mathcal{H}_2^{\otimes N}. \quad (16.115)$$

A unitary transformation is called “Boolean” if it transforms Boolean quantum states into Boolean quantum states.

Just as in the classical case, it is desirable to identify a set of generating transformations from which all unitary transformations can be obtained. Thus U is called a 1-bit transformation if it acts on only one factor in the tensor product $\mathcal{H}_2^{\otimes N}$, and an n -bit transformation if it acts on (at most) n factor bits. Using elementary matrix theory, one can show that every unitary transformation of $\mathcal{H}_2^{\otimes N}$ can be generated by 1- and 2-bit transformations.

The 1-bit transformations are of the simple form

$$U = \mathbf{1} \otimes \cdots \otimes \mathbf{1} \otimes V \otimes \mathbf{1} \otimes \cdots \otimes \mathbf{1}, \quad (16.116)$$

where V is a unitary 2×2 matrix. Boolean 1-bit transformations are given by the identity $V = \mathbf{1}$ and by

$$V = X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \quad (16.117)$$

Here, X describes quantum negation. In general, 1-bit transformations are not Boolean. Particularly important is the *Hadamard transformation*:

$$\begin{aligned} H|0\rangle &= \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle), & H|1\rangle &= \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle), \\ H &= \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}. \end{aligned} \quad (16.118)$$

Another important 1-bit transformation is the phase transformation:

$$Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \quad (16.119)$$

The most important 2-bit transformation is the *checked not* transformation U_{CN} , defined by

$$U_{\text{CN}}|y, x\rangle = |y, x+y\rangle \quad (16.120)$$

or

$$\begin{aligned} U_{\text{CN}}|0, 0\rangle &= |0, 0\rangle, & U_{\text{CN}}|0, 1\rangle &= |0, 1\rangle, \\ U_{\text{CN}}|1, 0\rangle &= |1, 1\rangle, & U_{\text{CN}}|1, 1\rangle &= |1, 0\rangle. \end{aligned}$$

Depending on the value of the first “control” bit, the second bit is left intact or negated. U_{CN} is a Boolean transformation. It can be shown that the special 2-bit transformations U_{CN} together with the total set of 1-bit transformations already generate all unitary transformations of $\mathcal{H}_2^{\otimes N}$. So, for the realization of quantum circuits, one only has to worry about 1-bit transformations and U_{CN} .

For example, the Bell states are obtained by applying $C = U_{\text{CN}}(H \otimes \mathbf{1})$ to the basis states $|x, y\rangle$:

$$\begin{aligned} C|0, 0\rangle &= U_{\text{CN}} \frac{1}{\sqrt{2}}(|0, 0\rangle + |1, 0\rangle) = \frac{1}{\sqrt{2}}(|0, 0\rangle + |1, 1\rangle), \\ C|0, 1\rangle &= U_{\text{CN}} \frac{1}{\sqrt{2}}(|0, 1\rangle + |1, 1\rangle) = \frac{1}{\sqrt{2}}(|0, 1\rangle + |1, 0\rangle), \\ C|1, 0\rangle &= U_{\text{CN}} \frac{1}{\sqrt{2}}(|0, 0\rangle - |1, 0\rangle) = \frac{1}{\sqrt{2}}(|0, 0\rangle - |1, 1\rangle), \\ C|1, 1\rangle &= U_{\text{CN}} \frac{1}{\sqrt{2}}(|0, 1\rangle - |1, 1\rangle) = \frac{1}{\sqrt{2}}(|0, 1\rangle - |1, 0\rangle). \end{aligned} \quad (16.121)$$

The four Bell states can be transformed into each other by applying to this first factor a negation and/or phase operator only to their first factor. This is the essence of *dense quantum coding*. Two bits of classical information can be transmitted by manipulating only one quantum bit.

For every n -bit transformation

$$V : \mathcal{H}_2^{\otimes n} \rightarrow \mathcal{H}_2^{\otimes n}, \quad (16.122)$$

one can define a checked version $V_C : \mathcal{H}_2^{\otimes(n+1)} \rightarrow \mathcal{H}_2^{\otimes(n+1)}$ by

$$V_C |y; x_1, \dots, x_n\rangle = \begin{cases} |y; x_1, \dots, x_n\rangle, & \text{for } y = 0, \\ |y\rangle \otimes V |x_1, \dots, x_n\rangle, & \text{for } y = 1. \end{cases} \quad (16.123)$$

In this sense, U_{CN} is just the checked version of the negation transformation X .

Classical Boolean functions $f : \mathbb{Z}_2^N \rightarrow \mathbb{Z}_2$ are never invertible for $N > 1$. Hence, it is not possible to associate to them unitary transformations $V_f : \mathcal{H}_2^{\otimes n} \rightarrow \mathcal{H}_2^{\otimes n}$ in an immediate way. This can, however, be achieved by means of the *Toffoli construction*. Let $f : \mathbb{Z}_2^n \rightarrow \mathbb{Z}_2^m$ be a classical Boolean function. Then

$$U_f : \mathcal{H}_2^{\otimes(n+m)} \rightarrow \mathcal{H}_2^{\otimes(n+m)}$$

is defined by

$$U_f |x\rangle \otimes |y\rangle = |x\rangle \otimes |y + f(x)\rangle, \quad (16.124)$$

for

$$|x\rangle \in \mathcal{H}_2^{\otimes n}, \quad |y\rangle, |f(x)\rangle \in \mathcal{H}_2^{\otimes m}.$$

Here U_f is unitary and Boolean. This construction guarantees that a quantum computer, restricted to Boolean states, can simulate every classical computer. Being members of an orthonormal set, Boolean states can be measured and cloned just like classical states.

The real power of quantum computing reveals itself only if non-Boolean states and operators are employed. To see this, let us start with the state

$$|\psi\rangle = \frac{1}{2^{N/2}} \sum_{x \in \mathbb{Z}_2^N} |x\rangle \in \mathcal{H}_2^{\otimes N}, \quad (16.125)$$

which is the starting point for many quantum computations. The state $|\psi\rangle$ is easily prepared by applying the Hadamard operator to each factor of $|\mathbf{0}\rangle = |0\rangle \otimes \dots \otimes |0\rangle$ (n factors):

$$|\psi\rangle = \frac{1}{2^{N/2}} \sum_x |x\rangle = H^{\otimes N} |\mathbf{0}\rangle = H|0\rangle \otimes \dots \otimes H|0\rangle. \quad (16.126)$$

The state $|\psi\rangle$ is not Boolean but separable. Applying U_f of eq. (16.124) for $m = 1$ to $|\psi\rangle \otimes |0\rangle \in \mathcal{H}_2^{\otimes(N+1)}$ gives

$$U_f |\psi\rangle \otimes |0\rangle = \frac{1}{2^{N/2}} \sum_x |x\rangle \otimes |f(x)\rangle. \quad (16.127)$$

This result is remarkable, because it shows that by one application of U_f the function f has been simultaneously evaluated for all values of x . This feature of *quantum parallelism* is the essence of the power of quantum computation.

Quantum parallelism, of course, has its price. The state $U_f |\psi\rangle \otimes |0\rangle$, the result of the computation, is, in general, not Boolean and can be neither copied nor completely read off

by a single measurement. One would think of repeating the calculation and the measurement of the result many times, but this would amount to giving away the advantage of quantum parallelism.

The art of quantum computation consists in steering the computation and in performing the final measurement such that the result gives as much useful information as possible. A single measurement can identify a component of the measured state in the direction of a given subspace. This may be sufficient to answer questions about several interesting global features of the result. Also, by choosing the measured observable appropriately, the largest component is found with the highest probability.

Using quantum parallelism, Fourier transformation can be performed faster than with any known classical algorithm. Let $f(n) = f_n$ be a complex-valued function defined for integer $0 \leq n < 2^N$. The Fourier transform $\tilde{f}_n$ is defined by

$$\tilde{f}_n = \frac{1}{2^{N/2}} \sum_{m=0}^{2^N-1} \zeta^{mn} f_m \quad \text{with} \quad \zeta = e^{2\pi i/2^N}. \quad (16.128)$$

The inverse transformation is given by

$$f_n = \frac{1}{2^{N/2}} \sum_m \zeta^{-mn} \tilde{f}_m. \quad (16.129)$$

Even for the so-called fast Fourier transform algorithm, the computation time increases exponentially with N . To describe the quantum algorithm, we represent the basis states $|x\rangle$ for $x \in \mathbb{Z}_2^N$ by non-negative integer numbers as described above:

$$|x\rangle = |n_x\rangle = |n\rangle. \quad (16.130)$$

Fourier transformation is performed by the unitary operator

$$F : \mathcal{H}_2^{\otimes N} \rightarrow \mathcal{H}_2^{\otimes N} \quad \text{with} \quad F|n\rangle = \frac{1}{2^{N/2}} \sum_{m=0}^{2^N-1} \zeta^{mn} |m\rangle. \quad (16.131)$$

Indeed,

$$F\left(\sum_n f_n |n\rangle\right) = \frac{1}{2^{N/2}} \sum_{m,n} \zeta^{mn} f_n |m\rangle = \sum_n \tilde{f}_n |n\rangle. \quad (16.132)$$

The calculation time grows polynomially with N , in fact only a bit faster than N .

If $f_n = f_{n+k}$, this periodicity reflects itself in a concentration of $\tilde{f}_n$ at positions $n = kr$, which can be detected with high efficiency by a suitable quantum measurement.

Fast period determination by quantum computation is the core of *Shor's quantum algorithms* for factorizing large numbers M into prime factors or for finding the discrete logarithm x such that $c^x = d \bmod p$. For all known classical algorithms for these problems, the computation time grows exponentially with the size (i.e. the number of digits) of the numbers, but for Shor's quantum algorithms the growth is only polynomial.

This is also of extreme practical importance, since the security of so-called *public key encryption schemes* relies on the intractability of such problems. Let us briefly describe the essential idea of Shor's factorization algorithm for large M . Take some number a and determine $n_0 \neq 0$ such that

$$a^{n_0} - 1 \equiv 0 \pmod{M}. \quad (16.133)$$

This is quickly done with a quantum computation of the period of the function

$$f(n) = f_n = a^n. \quad (16.134)$$

If n_0 turns out to be even, $n_0 = 2m_0$, then

$$(a^{2m_0} - 1) = (a^{m_0} - 1)(a^{m_0} + 1) \equiv 0 \pmod{M}. \quad (16.135)$$

Hence, at least one of the factors $(a^{m_0} - 1)(a^{m_0} + 1)$ will have a common divisor d with M , which can be quickly determined with the classical Euclidean algorithm. If $d < M$, a factor of M has been found. If $d = M$ or n_0 is odd, one tries a different number a . The search for a suitable a can be shown to lead to a positive result after a few trials with high probability, unless M is the power of a prime number: $M = p^r$. This case can be separately excluded with a fast classical algorithm.

Another important quantum algorithm of very promising prospects is a quantum search algorithm faster than any known classical search algorithm. The searching time for one item among n items is only $O(\sqrt{n})$, whereas it is $O(n)$ for classical search algorithms.

Quantum informatics and quantum computing are new fields whose quick development promises many surprises. On the theoretical side, all of informatics has to be reconsidered under the aspect of quantum theory. Much work has gone into the theoretical analysis of quantum channels for information transfer. A key feature is the investigation of the role of entanglement as a resource of quantum information theory. Even a fully satisfactory measure for entanglement in mixed states is still lacking. The comprehensive monograph of Nielsen and Chuang gives a very good overview of the field of quantum informatics.

A lot of progress has been achieved recently in devising quantum algorithms for error correcting and repair. Self-correction of errors is indispensable already in classical computing. Quantum states are even more vulnerable than classical two-state systems. In particular, entanglement, which is so vital for quantum computation, has a tendency quickly to get lost by "decoherence" as the effect of a small interaction with the environment of a system. For some time, it was widely believed that the problems related to decoherence were insurmountable. The advent of quantum correction algorithms was very encouraging in this respect.

On the experimental side, much work is in progress on the implementation of various components of future quantum computers. Here we only list some of the strategies being followed:

- Perhaps most advanced are linear ion traps in which several two-state ions are confined and are manipulated by letting them interact with laser light. For the realization of a two qubit gate like the controlled NOT gate, two ions have to be brought into contact such that their interaction depends on their states. Arrangements with a small number of qubits have been realized in this way.

- Atoms can also be confined in a regular standing laser light field, which, as described in section 15.8, exerts a force on them. Such a standing wave field may act as a regular lattice with atoms sitting at various lattice points. Computing may consist in changing the excitation states and/or the positions of atoms on this lattice.
- Nanostructures on the surface of a crystal on which individual atoms can be moved and manipulated provide another promising possibility.
- Also electromagnetic fields in cavities may serve as arrangements of qubits.
- Nuclear magnetic resonance can address spin variables in a solid or liquid material by microwaves. In order to obtain a signal of sufficient size, a qubit has to be represented by an assembly of several spins. In this sense, one might dispute whether nuclear magnetic resonance really provides an implementation of quantum computing in the strict sense.
- Quantum dots are point-like defects in semiconductors, which can be individually excited and de-excited. There are many promising techniques to manufacture architectures of quantum dots. The main problem will be to control their interactions with each other and with the environment.

So far, only small numbers of quantum bits can be controlled. A quantum computer should contain at least about 100,000 of them. This is still a challenge for the more remote future. The more modest aim of building a quantum simulator for several quantum systems of interest containing a few dozen qubits is still demanding but may be reached sooner.

Selected references

General optics

- L. Bergmann, C. Schäfer** *Lehrbuch der Experimentalphysik*, vol. III, *Optik*, Ed. H. Gobrecht. de Gruyter, Berlin, New York, 1987.
- M. Born, E. Wolf** *Principles of Optics*, 7th ed. 1999; Cambridge University Press.
- R. Guenther** *Modern Optics*. Wiley, New York, 1990.
- O. S. Heavens, R. W. Ditchburn** *Insight into Optics*. Wiley, Chichester, 1991.
- E. Hecht** *Optics*, 2nd edn. Addison Wesley, Reading, MA, 1987.
- S. G. Lipson, H. Lipson** *Optical Physics*. Cambridge University Press, London, 1969.
- M. Young** *Optics and Lasers*, 4th edn. Springer, Heidelberg, 1992.

History of optics (chapter 1)

- N. R. Hanson** Waves, Particles and Newton's "Fits", *Journal of the History of Ideas* **21** (1960) 370.
- I. Newton** New Theory About Light and Colours, *Philosophical Transactions* **80** (1672) 3075–3087.
- I. Newton** *Opticks*; Dover Publications, 1952.
- I. Newton** *The Principia – Mathematical Principles of Natural Philosophy*. A new translation by I. Bernard Cohen and Anne Whitman; University of California Press 1999.
- A. E. Shapiro** *The Optical Papers of Isaac Newton: The Optical Lectures, 1670–1672: 001*; Cambridge University Press, 1984.

Classical electrodynamics, electrodynamics of media (chapters 2–5)

- P. Gay** *Crystal Optics*. Longmans, Green and Co., London, 1967.
- J. D. Jackson** *Classical Electrodynamics*, John Wiley & Sons, 1998.

L. D. Landau, I. M. Lifschitz *Electrodynamics of Continuous Media*, Butterworth–Heinemann, 1995.

L. D. Landau, I. M. Lifschitz *Mechanics and Electrodynamics (Shorter Course of Theoretical Physics, Vol 1)*, Elsevier Science Ltd, 1972.

H. Römer, M. Forger *Elementare Feldtheorie*. VCH, Weinheim, 1993.

Nonlinear optics, solitons, nonlinear dynamical systems (chapter 6)

F. Abdullaev, S. Darmanyan, P. Khabibullaev *Optical Solitons*. Springer, Heidelberg, 1993.

D. K. Arrowsmith, C. M. Place *An Introduction to Dynamical Systems*. Cambridge University Press, Cambridge, 1990.

W. Demtröder *Laser Spectroscopy*, 3rd edn. Springer, Berlin, 2002.

P. G. Drazin *Nonlinear Systems*. Cambridge University Press, Cambridge, 1992.

G. Eilenberger *Solitons*. Springer, Heidelberg, 1981.

L. D. Faddeev, L. A. Takhtajan *Hamiltonian Methods in the Theory of Solitons*. Springer, Heidelberg, 1986.

M. L. Gower, D. Proch *Optical Phase Conjugation*. Springer, Heidelberg, 1991.

M. C. Gutzwiller *Chaos in Classical and Quantum Dynamics*. Interdisciplinary Applied Mathematics, vol. I, Springer, Heidelberg, 1990.

A. J. Lichtenberg, M. A. Lieberman *Regular and Stochastic Motion*. Springer, Heidelberg, 1983.

R. Meinel, G. Neugebauer, H. Steudel *Solitonen*. Akademie-Verlag, Berlin, 1991.

D. L. Mills *Nonlinear Optics*. Springer, Heidelberg, 1991.

S. N. Rasband *Chaotic Dynamics of Nonlinear Systems*. Wiley, New York, 1990.

D. Ruelle *Chaotic Evolution and Strange Attractors*. Cambridge University Press, Cambridge, 1987.

H. G. Schuster *Deterministic Chaos*, 2nd edn. VCH, Weinheim 1989.

Symplectic geometry, symplectic mechanics, geometry of short-wave asymptotics (chapters 7–10)

- R. Abraham, J. E. Marsden** *Foundations of Mechanics*, 2nd edn. Benjamin/Cummings, Reading, MA, 1978.
- V. I. Arnol'd** *Mathematical Methods of Classical Mechanics*. Graduate Texts in Mathematics, Springer, Heidelberg, 1978.
- V. I. Arnol'd** *The Theory of Singularities and its Applications*. Academia dei Lincei, Scuola Normale Superiore, Pisa, 1991.
- V. Guillemin, S. Sternberg** *Geometric Asymptotics*. Mathematical Surveys, vol. 14. American Mathematical Society, Providence, RI, 1977.
- V. Guillemin, S. Sternberg** *Symplectic Techniques in Physics*. Cambridge University Press, Cambridge, 1985.
- Yu. A. Kravtsov** *Caustics, Catastrophes and Wave Fields*. Series on Wave Phenomena, Springer, Heidelberg, 1993.
- E. Meinrenken** *Differentialgeometrische Methoden in der Kurzwellenasymptotik*. Diplomarbeit, Freiburg, 1990.
- W. Thirring** *Classical Mathematical Physics*, Springer, 2003.

Stochastic optics, coherence, Fourier optics, stochastic processes (chapters 10–12)

- B. Berne, R. Pecora** *Dynamic Light Scattering*. Wiley, New York, 1976.
- H. P. Breuer, F. Petruccione** *The Theory of Open Quantum Systems*; Oxford University Press, 2002.
- I. I. Gihman, A. V. Skorohod** *The Theory of Stochastic Processes*, vols. I–III. Grundlagen der Mathematischen Wissenschaften, vols. 210, 218, 323. Springer, Heidelberg, 1971, 1973, 1979.
- J. Honerkamp** *Statistical Physics*, 2nd ed., Springer Berlin, 2002.
- J. Peřina** *Coherence of Light*. Reidel, Dordrecht, 1985.

Quantum optics and quantum information (chapters 13–16)

- H. A. Bachor** *A Guide to Experiments in Quantum Optics*. Wiley-VCH, 1998.
- H.-D. Ebbinghaus, J. Flum** *Finite Model Theory. Perspectives in Mathematical Logic*; 2nd ed., Springer, Berlin, 1999.
- B. Fedosov** *Deformation Quantization and Index Theory*. Akademie Verlag, Berlin, 1996.

- L. Mandel, E. Wolf** *Optical Coherence and Quantum Optics*. Cambridge University Press, 1995.
- W. P. Schleich** *Quantum Optics in Phase Space*. Wiley-VCH, 2001.
- M. O. Scully, M. S. Zubairi** *Quantum Optics*. Cambridge University Press, 1997.
- A. E. Siegmann** *Lasers*. University Science Books, 1986.
- M. A. Nielsen, I. M. Chuang** *Quantum Computation and Quantum Information*. Cambridge University Press, 2002.
- Ch. H. Papadimitriou** *Computational Complexity*. Addison Wesley Longman, 1994.

Index

- Abbe's sine condition 131
- Abbe, Ernst 2, 131
- Aberrations 140
- Absorption, coefficient of 34
- Acoustic holography 206
- Action functional 118
- Adiabatic population transfer 299
- Airy functions 189
- Airy, G. B. 188
- Alhazen 2
- Amplitude holograms 205
- Amplitude objects 182
- Analytical signal 214
- Angle characteristic 127
- Angle eikonal 127
- Annihilation operator 246
- Anti Jaynes–Cummings model 316
- Anti-bunching condition 286
- Anti-normal ordering 260
- Anti-standard ordering 260
- Anti-Stokes line 90
- Aplanatic points 131
- Arago, Dominique Jean François 9
- Aristotle 1
- Armati, Salvino degli 2
- Arnol'd, V. I. 160
- Astigmatism 141
- Atom optics 321
- Atomists 1

- Babinet's principle 180
- Bacon, Roger 2
- Barrel distortion 142
- Bartholinus, Erasmus 2
- Basow, N.G. 13
- Beam splitters 273
- Bell states 324

- Bell, J. 328
- Biot, Jean Baptiste 9
- Bits 343
- Blaschke manifold 125
- Bogolyubov transformations 247
- Boolean functions 344
- Boolean quantum states 345
- Boolean transformations 345
- Boyle, Robert 3
- Bracciolini, Poggio 2
- Bradley, James 9
- Bragg's condition 191
- Breather solutions 89
- Brewster angle 72
- Brewster, D. 12
- Brillouin peaks 235
- Brillouin scattering 90
 - stimulated 90
- Bulk modulus 49
- Bunching condition 285

- Camera Obscura 2
- Canonical transformations 125, 161
- Carathéodory, C. 123
- Cauchy principal value 26
- Cauchy's integral formula 24
- Causality condition 24
- Caustics 109, 154
- Characteristic equation 144
- Characteristics 145
- Checked not transformation 346
- Chromatic aberration 142
- Chromatic dispersion 103
- Cittert–Zernicke theorem 229
- Classical limit 268
- Clausius, R. 131
- Co-isotropic subspaces 132
- Coherence 211

- Coherence function of order n 279
- Coherence length 211
- Coherence radius 212
- Coherence time 211
- Coherence volume 212
- Coherent states 252
- Collapse time 306
- Coma 141
- Combination frequencies 84
- Compensator 240
- Complete integrability 88
- Conductivity, dynamical 21
- Conductivity, static 21
- Conformally flat metric 119
- Conical refraction, inner and outer 66
- Conjugated planes 135
- Conjugated points 121
- Continuity conditions 32
- Corpuscular theory 4, 107
- Correlated emission laser 301
- Correlation functions 217
 - generalized 217
- Cotangent bundle 161
- Cotton-Mouton effect 82
- Coulomb gauge 248
- Covariance functions 217
- Creation operator 246
- Critical points 148
- Crystal classes 76
- Crystal optics 35, 55
- Crystal systems 57, 76
- Crystal, positive and negative 62
- Crystals
 - optically biaxial 65
 - optically uniaxial 62
- Darboux coordinates 162
- Darboux's theorem 161
- Darboux, G. 125
- Dark states 297
- Dark-ground method 194
- De rerum natura* (Lucretius) 1
- Decoherence 349
- Decomposable states 323
- Deformations, elastic 45
- Degenerate four-wave interaction 100
- Degree of coherence 225
- Democritus 1
- Dense quantum coding 346
- Density operator 266
- Descartes, René 2
- Detunings 291
- Dielectric tensor 32
- Diffraction 4
- Diffraction theory 167
- Dioptric power 136
- Dipole interaction 290
- Dispersion of the axes 56
- Dispersion relations 26, 36
- Distortion 142
- Dolland, John 9
- Double refraction 65
- Double-slit experiment 332
- Downward conversion 326
- Drift velocity 20
- Drude's theory of conductivity 20
- Drude, Paul 20
- Dynamical light scattering 233
- Eight-port interferometer 277
- Eikonal equation 110
- Eikonal function 108
- Einstein, Albert 10, 328
- Elastooptical constants 80
- Electric displacement density 16, 31
- Electric displacement energy 75
- Electric field vector 15
- Enantiomorphic form 79
- Energy density 47
- Energy density of the electromagnetic field 17
- Energy flux density 17
- Enlargement 135
- Entangled states 323
- Entanglement 323
- Entanglement correlations 325
- Envelopes 168
- EPR argument 328
- Ergodic process 219
- Error correction 349
- Error repair 349
- Ether 5
- Euclid 10
- Exact symplectic manifold 162
- Experimental philosophy 329
- Extraordinary waves 63
- Fabry-Perot étalons 73, 97
- Fabry-Perot interferometer 72, 97, 321

- Fano factor 287
- Faraday effect 82
- Faraday, Michael 9
- Fermat's principle 115, 116
- Fermat, Pierre de 2
- Fiber of a Maslov bundle 197
- Fiber optics 101
- Field operators 250
- First fundamental form 156
- First-order coherent states 280
- Fits 7
- Fizeau, H. 12
- Floquet exponents 317
- Floquet solutions 317
- Flux velocity 48
- Focal planes 136
- Focal points 109, 121, 154
- Focal points of lenses 155
- Foucault, Jean Bernard Léon 9
- Four-photon phenomena 89
- Fourier integral operator 200
- Fourier transform holograms 206
- Fourier transforms 192
- Fraunhofer diffraction 183
- Fraunhofer, Joseph 9
- Frequency doubling 89, 92
- Frequency multiplication 84
- Frequency set 199
- Fresnel diffraction 183
- Fresnel zones 169
- Fresnel's ellipsoid 55
- Fresnel's equations 71
- Fresnel's integral 186
- Fresnel, Augustin Jean 9
- Fundamental forms 156

- Gabor, D. 13, 205
- Galilean transformations 251
- Galilei, Galileo 2
- Gauge degrees of freedom 18
- Gauss, Carl Friedrich 2
- Gaussian distribution 220
- Gaussian optics 134
- Gaussian system of units 16
- Generalized correlation functions 217
- Generating function 126
- GEO 600 284
- Geodesics 116
- Geometrical optics 2, 108

- GHZ experiment 330
- Glass fibers 101
- Gouy, L. G. 13, 157
- Gradient profile fibers 101
- Graph of a function 126
- Gravitational wave detectors 283
- Gravitational waves 283
- Green's function 153
 - retarded 18
- Grimaldi, Francesco Maria 2
- Group velocities 48, 51
- Group velocity dispersion 103
- Guiding waves 6
- Gyration tensor 77

- $\hbar$ dependence 262, 268
- Hadamard transformation 346
- Haitham, Ibn al 2
- Half-densities 163
- Hall, Chester Moor 9
- Hamilton's equations of motion 116, 146
- Hamilton, William Rowan 2
- Hamilton–Jacobi equation 117
- Hamilton–Jacobi theory 116
- Hanbury Brown and Twiss interferometer 281
- Hanbury Brown, R. 231
- Hanle effect 297
- Harmonic oscillator, classical and quantum 245
- Heisenberg picture 249
- Helmholtz equation 112, 145
- Helmholtz, H. v. 12, 131
- Hermiticity of a star product 266
- Hero of Alexandria 1
- Herschel's condition 130
- Herschel, F. W. 11
- Hertz, Heinrich Rudolf 9
- Heterodyne detection 276
- Hexagonal 57
- Hologram 204
- Holography 101, 203
- Holohedral 76
- Homodyne detection 276
- Hooke's law 46, 49
- Hooke, Robert 4
- Huygens' principle 168
- Huygens, Christiaan 2
- Huygens–Fresnel principle 169, 171

- Idler wave 89
- Image 121
- Inclination factor 170
- Index ellipsoid 55
- Index of refraction 34, 51
- Intensity correlation spectroscopy 231
- Intensity functions 217
- Interference 226
- Interferometer
 - eight-port 277
 - Fabry–Perot 72, 321
 - Hanbury Brown and Twiss 281
 - Mach–Zehnder 335
 - Michelson 230, 284
 - Sagnac 285
- Inversion revival 307
- Isotropic submanifold 149
- Isotropic subspaces 132
- Janssen, Zacharias 2
- Jaynes–Cummings Hamiltonian 292
- Jaynes–Cummings model 302
- Joule heat 26
- Katoptrik* (Euclid) 10
- Kepler, Johannes 2
- Kerr effect 81
- Kink solution 88
- Kirchhoff's diffraction theory 179
- Kirchhoff's integral 175, 179
- Kirchhoff, Gustav Robert 9
- Kramers–Kronig relations 26
- Lagrange's equation of motion 116
- Lagrangian density 45
- Lagrangian submanifold 149
- Lagrangian subspaces 132
- Lamb–Dicke approximation 315
- Lamb–Dicke parameter 315
- Lamé coefficients 49
- Larmor frequency 82
- Laser 294
- Laser cooling 314
- Lasing without inversion 301
- Lens formula 137
- Leucippus 1
- Light metric 154
- Light quantum 249
- Light rays 108
- Light, visible 19
- Light-length element 115
- LIGO 284
- Lindblad equations 294
- Liouville equation
 - classical 271
 - quantum 270
- Lippershey, Hans 2
- Locality 325
- Longitudinal magnification 138
- Lorentz force 15
- Lorentz gauge condition 18
- Lorentz, Hendrik Antoon 10
- Lorentz-Lorenz formula 40
- Lucretius 1
- Mach–Zehnder interferometer 335
- Macroscopic fields 19
- Magnetic field strength 31
- Magnetic field vector 16
- Magnetic induction 15
- Magnetic permeability 16
- Maiman, T.H. 13
- Malus, Etienne Louis 9
- Maslov bundle 164, 197
- Maslov index 159, 165, 197
- Material quantities 35
- Mathieu's equation 317
- Maximally isotropic subspaces 132
- Maxwell's effect 80
- Maxwell's equations 15
 - covariant form 41
 - for media 32
 - homogeneous 16
 - inhomogeneous 16
- Maxwell's fish-eye 124
- Maxwell, James Clerk 9, 124
- Metal optics 34
- Michelson interferometer 230, 284
- Michelson stellar interferometer 281
- Michelson, Albert Abraham 10
- Micrographia* (Hooke) 4
- Microlasers 308
- Micromasers 308
- Microscope 2, 189
- Mixed characteristic 127
- MKS-system 16
- MKSA-system 16
- Mode dispersion 101

- Molecular fields 18
- Molecular polarizability 39
- Momentum of the electromagnetic field 17
- Monoclinic 57
- Morley, Edward Williams 10
- Morse family 196
- Morse index 154, 157
- Morse theory 155

- Natural optical activity 75, 79
- Newton's rings 5
- Newton, Isaac 4
- No-cloning theorem 338
- Nonlinear optics 83
- Nonlinear Schrödinger equation 88, 104
- Normal ellipsoid 55
- Normal ordering 260

- Occupation inversion 306
- Occupation number inversion 296
- Ohm's law 21
- Ohm's law, generalized 20
- Onsager's symmetry relations 27
- Onsager–Casimir relations 38
- Optic axes 57, 129
- Optical bistability 98
- Optical data storage 209
- Optical length 135
- Optical pumping 294, 299
- Optical rectification 84
- Optical solitons 101
- Optical transmission function 182
- Optically biaxial 58
- Optically uniaxial 58
- Opticks* (Newton) 6
- Order of a focal point 154
- Ordering procedures 259
- Ordinary waves 63
- Orthorhombic 57
- Oscillation intensities 39
- Oscillatory distribution 199
- Oscillatory function 198

- Pair distribution function 234
- Parallelepiped 57
- Parametric amplification 89, 91
- Parity transformation 31
- Passivity condition 26
- Pattern recognition 207

- Paul trap 313
- Penetration depth 23
- Perfect optical instrument 122
- Permeabilities 32
- Phase conjugated wave 98
- Phase conjugating mirror 99
- Phase conjugation 99
- Phase contrast method 195
- Phase holograms 205
- Phase matching 93
- Phase mismatch 87, 92
- Phase modulation 182
- Phase objects 182
- Phase self-modulation 96
- Phase space 125
- Phase velocity 50
- Phasonium 299
- Phonons 84
- Photon count correlation function 285
- Photonics 90
- Photons 84
- Pick functions 37
- Piezoelectricity 75
- Pin-cushion distortion 142
- Principia* (Newton) 6
- Planck, Max 10
- Plane of incidence 70
- Plasma frequency 21
- Plenists 3
- Pockels cell 81
- Pockels effect 80
- Point characteristic 126
- Point eikonal 126
- Poissonian light 283
- Poisson bracket 143, 265
- Poisson manifolds 265
- Polarization 9, 19
 - electric 30
 - magnetic 30
 - transverse 70
- Polarization degree 242
- Polarization filter 241
- Polarization states
 - circular 326
 - planar 325
- Polarization density 31
- Poynting vector 32
- Principal curvatures 156
- Principal maxima 191

- Principal planes 138
- Principal section 63
- Principal symbol 200
- Principal values 55
- Principle of Maupertuis 119
- Principle of realism 328
- Prochorow, A.M. 13
- Product states 323
- Pseudo-differential operators 143
- Ptolemaeus 2
- Public key encryption schemes 349
- Pumping wave 89
- Pupil 183
- Pupil function 182
- Pure coherent states 266
- Pure number states 267
- Pyroelectricity 75

- Quadratures 246
- Quantization 259
- Quantum bit 345
- Quantum cryptography 342
- Quantum eraser 332
- Quantum parallelism 347
- Quantum state engineering 310
- Quantum teleportation 339
- Quantum Zeno effect 337
- Quasi-monochromatic approximation 224
- Qubit 345

- Rabi frequencies, complex 292
- Radiation condition 176
- Radon transform 277
- Raman scattering 90
 - stimulated 90
- Random function 217
- Random variables, Gaussian 220
- Ray coordinates 131
- Ray surface 51
- Ray vector 51
- Rayleigh peak 235
- Reality condition 24
- Realization of a stochastic process 217
- Reconstruction wave 204
- Reduction 324
- Reference wave 203
- Reflection 66
 - total 68
- Reflection coefficient 71
- Reflection law 70, 115
- Refraction 66
- Refraction law 2, 115
- Refractive index 34, 51
- Refractive index enhancement 301
- Resistance, static 23
- Resonance conditions 291
- Revival 307
- Ritter, J.W. 12
- Rømer, Olaus 3
- Röntgen, W.C. 13
- Rotating phase approximation 87, 291
- Rydberg states 308

- s*-phonon Jaynes–Cummings model 316
- Sagnac interferometer 285
- Scalar potential 17
- Scalar wave equation 111
- Schlieren method 195
- Schrödinger picture 249
- Schrödinger cat states 313
- Second fundamental form 156
- Secondary maxima 191
- Seidel's theory of aberrations 140
- Self-focussing 96
- Shear contraction 50
- Shear modulus 49
- Shor's quantum algorithms 348
- Short-wave approximation 108
- SI-system 16
- Side maxima 191
- σ matrices 239
- Signal wave 89
- Signal, analytical 214
- Sine–Gordon equation 88
- Single-mode fibers 103
- Skin effect 23
- Snell's law of refraction 70, 114
- Snell, Willebrord 2
- Solitary waves 88
- Sommerfeld, Arnold 13
- Speckle effect 236
- Speckle pattern, objective and subjective 236
- Spherical aberration 141
- Spontaneous emission 305
- Squeezed states 255
- Squeezing operator 252
- Standard ordering 260
- Star product 261

- Stark effect, dynamical 303
- State measurement apparatus, impossibility of 337
- Stationary stochastic process 218
- Step profile fibers 101
- Stochastic process 217
- Stokes line 90
- Stokes parameters 240
- Stress tensor, elastic 49
- Stress tensor, Maxwell's 32
- Strong oscillator limit 276
- Structure function 234
- Sub-Poissonian light 287
- Super-Poissonian light 287
- Susceptibilities 32
 - electric 35
 - magnetic 35
- Symbol 143, 164, 200, 264
- Symbol calculus 202
- Symplectic basis 133
- Symplectic isomorphisms 132
- Symplectic manifolds 125, 161
- Symplectic subspaces 132
- Symplectic transformations 132, 161
- Symplectic vector spaces 132

- Telegraph equation 22
- Telescope 2, 189
- Telescopic mapping 135
- Tetragonal 57
- Thermal light 283
- Thermal states 267
- Thomas–Kuhn sum rule 39
- Three-photon phenomena 89
- Three-wave interaction 87, 89
- Time development operator 290
- Toeplitz, A. 195
- Toffoli construction 347
- Tomographic reconstruction 277
- Townes, C.H. 13
- Traité de la Lumière* (Huygens) 2
- Transmission 182
- Transmission coefficient 71
- Transmissivity 73
- Transparency, electromagnetically induced 299
- Transport equation 112, 144

- Transverse magnification 139
- Trapped states 297
- Triclinic 57
- Trigonal 57
- “True” state 328
- Twiss, R. Q. 231
- Two-photon absorption 90

- Uncertainty relation 223, 257
- Unitarity of a beam splitter 274
- Unitary transformations
 - Bogolyubov transformations 252
 - rotations 251
 - translations 251

- Vacuum permittivity 16
- Vector potential 17
 - quantum 250
- Verdet's constant 82
- VIRGO 284
- Visibility 227
- Volume holograms 207

- Wave equation 16
- Wave operator 18
- Wave surface 51
- Wave theory 6, 107
- Wave–particle duality 107
- Wavelets 167
- Weyl ordering 261
- “Which path” information 297, 332
- Whispering gallery modes 310
- White noise 222
- Wiedersehen manifold 125
- Wiener filter 237
- Wiener–Khinchin theorem 219
- Wigner function 263
- WKB ansatz 144, 271
- Wollaston, W.H. 12

- X-ray spectroscopy 235

- Young's interference experiment 225, 281
- Young's modulus 50
- Young, Thomas 9
- Yule–Walker equation 239

- Zernicke, F. 195