FACULTEIT INGENIEURSWETENSCHAPPEN

Fotonica

Photonics

Roel Baets

Günther Roelkens

Cursus voor de derde bachelor ingenieurswetenschappen: elektrotechniek

en voor de derde bachelor ingenieurswetenschappen: toegepaste natuurkunde

Course for the Erasmus Mundus Master of Science in Photonics

Academiejaar 2009-2010

Universiteit Gent

Vakgroep informatietechnologie

Faculteit Ingenieurswetenschappen

Academic year 2009-2010

Ghent University

Department of Information Technology

Faculty of Engineering

Preface

Photonics is a multidisciplinary discipline with strong roots in fundamental physics and with a rapidly increasing range of engineering applications in information technology, energy, lighting, manufacturing and materials processing, metrology and sensing, medicine and biotechnology etc.

The course Photonics (Fotonica) is set up as a basic course for the last year of a bachelor program (or as an introductory course at the master level). Its ambition is to introduce most of the basic concepts used in photonics as well as to teach some basic design approaches. Furthermore it will confront the student with a (limited) amount of factual knowledge about "real-life" photonic materials, components and systems. At the end of this course the student will have gained a broad introductory knowledge in photonics, in such a way that it will serve both those who will further specialize in photonics and those who will not.

The course is taught (in Dutch) as a compulsory course for the 3rd year Bachelor in Electrical Engineering as well as for the 3rd year Bachelor in Engineering Physics. In view of the different background of both groups there is unavoidably some overlap between this course and earlier courses taken by one of both groups. For this reason a limited number of chapters are taught separately to the two groups.

The course is also taught as preparatory access course (in English) for the 1st year Erasmus Mundus Master of Science in Photonics. Originally the course text was written in Dutch and it was used as such from the academic year 2003-2004 onwards. In view of the start of the international Erasmus Mundus program in 2006-2007 the text has been translated to English.

The writing of a course text of this volume is obviously an extensive task. We are indebted to a large group of co-workers in the Photonics Research Group for their help both for the contents and for the editing and layout. More in particular we thank Wim Bogaerts, Pieter Dumon, Hannes Lambrecht, Gino Priem, Olivier Rits, Joris Van Campenhout and Lieven Van Holme. We also thank Danae Delbeke, Pieter Dumon, Bjorn Maes and Karel Van Acoleyen for their contributions to the translation of the text. Finally we thank numerous students for feedback on the course and for reporting various errors and shortcomings.

We wish all students in this course an exciting ride into the world of photonics.

Gent, 24 September 2009

Günther Roelkens and Roel Baets

Contents

Part I

Introduction

Chapter 1

Introduction

Contents

1.1 Photonics - what's in a name?

The term *photonics* is relatively new and originated in the eighties of the 20th century. Its original use was in the field of information technology. The term could be seen as an analogy to the term electronics and corresponded to the application of optics and opto-electronics in electronic and telecommunications systems. However, in due course the term acquired a broader meaning, and now it refers to that field of science and technology where the fundamental properties of light and its interaction with matter are studied and applied. The term is thus broader than optics or opto-electronics.

Examining the dictionaries for the word photonics, we find amongst others:

- Merriam-Webster dictionary: ***photonics****: a branch of physics that deals with the properties and applications of photons especially as a medium for transmitting information*
- American Heritage Dictionary: ***photonics*** *: the study or application of electromagnetic energy whose basic unit is the photon, incorporating optics, laser technology, electrical engineering, materials science, and information storage and processing.*

And in The Photonics Dictionary (http://www.photonics.com/dictionary/), an encyclopedia of all the terms in this field, we obtain: ***photonics****: The technology of generating and harnessing light and other forms of radiant energy whose quantum unit is the photon. The science includes light emission, transmission, deflection, amplification and detection by optical components and instruments, lasers and*

other light sources, fiber optics, electro-optical instrumentation, related hardware and electronics, and sophisticated systems. The range of applications of photonics extends from energy generation to detection to communications and information processing.

This last definition is not limited to information technology anymore, but it also includes optical instrumentation, energy applications, etc. It is this broader meaning that we apply in this course.

1.2 Photonics - a historical outline

Light has always played a special role in the development of mankind, and there are valid reasons for this. The light of the sun is the most important energy source for the earth. This light is essential for most life forms and as it happens we are able to see a part of that light with our own eyes. Clearly this has always sparked the imagination. However, understanding the character of light and of all light-related phenomena, is a tremendous journey of ups and downs. It started in antiquity, accelerated in the 17th century and underwent a revolution in the 20th century with the discovery of the photon nature of light. The contribution of the 21th century remains yet unknown.

In the following we present a short outline of the evolution of photonics. The names of important discoverers and scientists are mentioned. However, many are left out in order to keep this overview brief.

1.2.1 Antiquity and the Middle Ages

During Greek antiquity there was ample philosophy about the nature of light. One knew that light propagated along a straight line, and about the phenomena of reflection and refraction. One had also toyed with curved pieces of glass, and realized this could ignite a fire. But that was about it. It was Euclid (300 BC) among others who put some things systematically on paper in his Optica. He thought light beams originated from the eye and "scanned" an object. Later Aristotle disputed this hypothesis. In the first century BC Hero proposed that light always followed the shortest path - indeed not far from the truth.

The Romans did not contribute much to optics and in the dark Middle Ages - a fitting wordplay - they got stuck completely. Around the 13th century thoughts about using lenses as glasses started to appear. There was also a first correct explanation for the occurrence of the rainbow. Furthermore it started to dawn that the speed of light had to be finite. One compared the propagation of light to the propagation of sound. Therefore one needed a medium, which they called aether. The Englishman Roger Bacon was pivotal in these developments. However, for a serious breakthrough we have to wait until...

1.2.2 The 17th century

In the beginning of the 17th century the construction of telescopes started. With one of these first telescopes Galileo Galilei discovered the four moons of Jupiter, among other things. On the theoretical field the understanding of ray optics - or geometrical optics - started to expand. Willebrord Snell (Snellius) uncovered the law of refraction, but died before making it public. However,

Descartes knew about the finding of Snellius, and published it using his own name. Today the French sometimes speak of "la loi de Descartes", but the rest of the world acknowledges Snellius. Pierre de Fermat developed his famous Principle of Least Time, which states that beams of light always follow "the shortest path" (in time). There has been a lot of debate about this principle, also in philosophy (how do beams *know* which is the shortest path?). With this expansion of ray optics the construction of optical instruments became more and more sophisticated. Antonie van Leeuwenhoek developed the first microscope.

In the same era the study of interference and diffraction commenced. Because of the peculiar colors of a thin film (such as an oil film on water) - often called Newton rings - one could not escape it any longer: light behaves like a propagating wave. And waves can exhibit constructive and destructive interference phenomena. Based on this wave principle Christian Huygens started to work on diffraction theory. He considered every point in the aether where light passes, as a point source itself, from which a spherical wave emanates. This is uncannily close to the truth. Moreover Huygens discovered that light has a polarization and he experimented with birefringent crystals. Around this time the finiteness of the speed of light was proven, via a study of the eclipses of the moons of Jupiter. Furthermore, Isaac Newton showed that white light could be split in its color components by a prism. However, Newton had a hard time accepting the wave character of light and he proposed that light consisted of particles which propagate linearly through the aether. Because of his authority his corpuscular theory put many scientists on the wrong track during decades. It took until the first half of the 19th century before the wave character of light would be generally accepted.

1.2.3 The 18th century

This century is often called the age of enlightenment, but that is about the main achievement for photonics.

1.2.4 The 19th century

In the beginning of the 19th century the situation suddenly accelerated. Diverse scientists pieced together the puzzle of interference and diffraction. Thomas Young, Auguste Jean Fresnel, Josef Fraunhofer, Karl Friedrich Gauss, Lord Rayleigh, George Airy: they all made their contribution, be it to the physics of the optical phenomena, or to the necessary mathematics. Around this period Fraunhofer discovered - actually rediscovered - the dark lines in the solar spectrum, and therefore one had to examine the interaction between light and matter. The world of spectroscopy was born.

In the meantime Fresnel stated his Fresnel laws, which for the first time provided a quantitative description of the strength of reflection and refraction at an interface between two media. Furthermore Johann Christian Doppler uncovered the Doppler effect by studying the spectrum of binary stars.

In 1850 J.L. Foucault (also known for the pendulum) devised a method to accurately measure the speed of light. He also discovered that light propagated slower in a transparent medium such as water or glass, than in air or vacuum (in contrast with sound waves). It was the end for the corpuscular theory, even though pastor Sir David Brewster vehemently defended the theory to the end. Luckily he is better known for the discovery of the Brewster effect. Independent of these

"opticians" - photonicists? - Michael Faraday, James Clerk Maxwell and others experimented with electricity and magnetism. The Maxwell equations saw the light of day. It was a big surprise that these electromagnetic waves traveled with the speed of... light, and James Maxwell rapidly concluded that light waves had to be electromagnetic waves. Heinrich Hertz, just after discovering the photo-electric effect (the effect where electrons escape from a material upon illumination with short wavelength light), would experimentally show the electromagnetic character of light in 1888. These concepts were so revolutionary that during a long period people where divided into "believers" and "non-believers" concerning the electromagnetic nature of light.

Meanwhile Lord Rayleigh (John Strutt was his real name) developed a theory describing the scattering of light from small particles. Finally it became clear why the sky is blue. The occurrence of total internal reflection was discovered by John Tyndall. In 1879 Thomas Alvin Edison constructed the first usable electric lamp, the incandescent lamp. Following the telescope and the microscope the next practical application of photonics, lighting, was kickstarted. On the other hand interferometry slowly evolved from a curious phenomenon into a useful technique. Armand-Hippolyte-Louis Fizeau, Albert Michelson, L. Mach and L. Zehnder, Charles Fabry and Alfred Perot : they all developed different types of interferometers carrying their name. Up until today these devices are part of the standard toolbox for a specialist in photonics.

At the end of the 19th century several experiments were conducted that paved the way for modern quantum mechanics. Josef Stefan and Wilhelm Wien studied blackbody radiation, Johann Jakob Balmer examined the hydrogen spectrum, Pieter Zeeman uncovered the broadening of spectral lines in a magnetic field. Step by step it became clear that classical mechanics was unable to explain all the phenomena. The 20th century dawned.

Joseph Plateau

Ghent, and more specifically Ghent University, has its famous optical scientist. Joseph Plateau(1801-1883) studied physiological optics and with his *phenakistiscoop*, mostly cited as the direct precursor to the movie, founded the basis for the movie industry. In 1835, when he becomes a professor at Ghent University, he already discovered the phenakistiscoop. This device is based on the slowness of sight, because of which a quick succession of pictures merges into a moving image. The phenakistiscoop consists of a support, mounted with a round disc with slightly differing drawings, separated by small slits. If one turns the disc in front of a mirror and looks at the passing images through the slits, motion appears. In a tragical twist of fate, analogous to Beethoven becoming deaf and never hearing the ninth symphony, Plateau became blind and conducted research as a blind person for forty years, partly in optics.

1.2.5 Twentieth century

The first half of the 20th century stands for the development of quantum physics. In 1900 Max Karl Planck could explain the blackbody radiation spectrum by postulating that the energy of an oscillator consists of a number of discrete quanta, with energy proportional to the oscillation frequency. Planck's constant h, the proportionality factor, was discovered. Using this Albert Einstein elucidated the photoelectric effect in 1905, by proposing that light itself consisted of quanta with energy $h\nu$. In a certain way, Newton's corpuscular theory was back, but without the misleading aether concept. The photon was born.

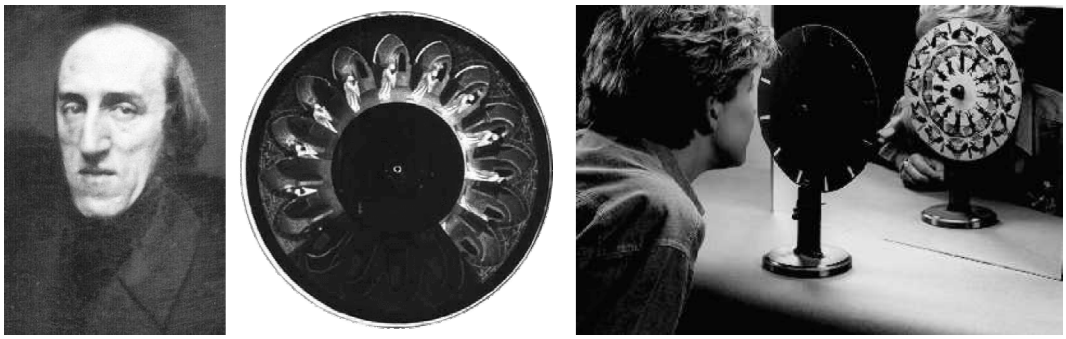

Figure 1.1: Joseph Plateau and his phenakistiscoop. Source: Museum for the History of Science - Ghent University (http://allserv.ugent.be/ ivhaeghe/mhsgent/)

A few years later Niels Bohr applied these quantum principles to explain the line spectra of gases. After this a delegation of physicists built and expanded quantum mechanics: Heisenberg, Born, de Broglie, Schrödinger, Dirac... From then on we had to accept that light has both a wave and a particle character. One phenomenon is better explained by one characteristic, another experiment by the other property. To this day the union of both pictures is not always harmonious, and some still struggle with fundamental problems in reconciling both.

In 1916 Albert Einstein had another remarkable proposal: apart from absorption and spontaneous emission of photons there needed to be another third interaction, stimulated emission. This process amplifies light and forms the basis of one of the greatest inventions of the 20th century: the laser. The optical laser was described theoretically in 1958 by Nobel Prize laureates Arthur Shawlow and Charles Townes, extending upon the concept of the microwave laser or maser. After this the race was on towards an experimental demonstration. In 1960 Theodore Maiman constructed the first working laser, which was a ruby laser. Two years later the semiconductor version was independently developed by three American groups. The importance of the laser can not possibly be overestimated. During the last decades of the 20th century, mainly because of the laser, photonics has developed into a discipline with enormous impact on fundamental physics and on diverse application areas.

We mention here several other fundamental discoveries made in the 20th century: discharge lamps, Raman and Brillouin scattering, acousto-optical interaction, holography, nonlinear optics, solitons, surface plasmons, liquid crystals, photonic crystals, quantum wells and quantum dots, etc.

Writing about recent history is difficult, because the topics are so diverse and because one cannot distinguish the most important subjects yet. Instead, in the next section we succinctly describe some application areas.

1.2.6 21st century

It is even more difficult to predict the future. What will the 21st century bring for photonics? From scientific research one can make some careful predictions for the next 10 to 20 years.

In analogy with micro-electronics, one can expect a major industrial advance for micro-photonics in the next 5 to 10 years. Micro-photonics means that photonic functions are integrated in mi-

crosystems, consisting of chips (processing optical signals, possibly combined with electrical signals), micro-optical elements, micromechanical parts etc.

In research there is a lot of activity on nanophotonics, where light interacts with nanoscale material structures, opening a new world of material properties and applications. Although there are still a lot of theoretical, conceptual and technological hurdles, we can expect concrete applications in the next 10 to 20 years.

From nonlinear optics one expects a huge impact, as it allows to introduce (digital) data-processing concepts into photonics.

The world of light sources will continue its rapid evolution of the last decades: more power, better efficiency, shorter pulses, different wavelengths, cleaner spectrum, higher modulation bandwidths etc.

Organic materials, including all kinds of liquid crystals, biomolecules and polymers, will play an important role in photonics.

Quantum optics has a potentially great impact in quantum communications and eventually even in quantum computing. But this may take 20 years or more.

1.3 Photonics - applications

Around the middle of the 20th century one had a good understanding of all kinds of optical phenomena, but the applications were lacking. There were diverse optical visualisation instruments, such as the telescope and the microscope, a variety of spectroscopic tools, and lighting sources, with the lightbulb and the gas discharge lamp. The layman only knew this last device.

Only during the second half of the 20th century, and more specifically during the last quarter, one started developing applications at a rapid pace. We give a short overview and distinguish five large application areas (with inevitable overlap): the energy sector, the medical area, measurement and sensor applications, visualisation, and information technology.

1.3.1 Energy applications

Electromagnetic radiation - and light in particular - transports energy from location A to location B without the need for material contact. This energy can be used in various ways: after absorption it is converted to heat, or the photons interact, if they have sufficient energy, with materials to start a chemical reaction or an electrical current. An important example of the last conversion is the solar cell. Today photovoltaic energy is mostly relevant for energy production in remote areas (on earth or in space), but it is probably only a matter of time before there are more widespread applications. Today's commercial cells are made of semiconductor, but in the future we can expect the use of plastics.

Lighting is a second important energy application. In the last decade the area of lamps has evolved from a classical market into a rapidly evolving high-tech marketplace. Efficiency, lifetime and compactness are the driving forces behind this innovation. Apart from the classical lightbulbs,

discharge and fluorescence lamps we more and more see the appearance of LEDs - Light Emitting Diodes - in various lighting and signalisation devices.

The advent of the laser has had a huge impact on energy applications. Because of the coherence of laser light it is possible to focus the energy of a beam onto a spot with a diameter on the order of the wavelength of the light. A laser with 1 KWatt power, focussed onto an area of 1 square micrometer delivers a power density of 100 GWatt per square centimeter. Thus the applications are spectacular. With a high power laser one is able to cut, weld or harden diverse materials - even thick steel plates. Three-dimensional modeling is also possible in a myriad of ways. With stereolithography one can prototype layer upon layer of a three-dimensional object. With laser ablation material of a surface can be vaporised in a very precise way, leading to formation of a shape. Since a couple of years one can directly "sculpt" a three-dimensional form in a transparent volume. This process requires lasers with pulse lengths on the order of 10 to 100 femtoseconds, and it employs nonlinear optical phenomena.

The whole world of micro-electronics exists today because of optical lithography. The patterns of almost all electronic ICs are realised by optically imaging a mask on a semiconductor wafer. The narrowest linewidth is on the order of the wavelength used. To obtain even finer lines - nowadays about 100 nm - one uses light sources with wavelengths deeper and deeper in the ultraviolet. Previously one used lamps, now lasers are more and more common.

1.3.2 Medical applications

The laser has profoundly changed many therapies in medicine, most of all because laser therapy is often less invasive than other techniques. The penetration depth of the light and its effect on tissue is strongly dependent on the light wavelength. The pulse length is another important parameter. The biggest impact of the laser is probably in the area of ophthalmology. With lasers one can repair damage to the retina or decrease eye pressure, which can injure the optic nerve. Of course one can correct near- or farsightedness by changing the curvature of the cornea with laser ablation. Lasers are also often used in general surgery, e.g. to evaporate tissue, to make non-bleeding cuts or to clot blood. In dermatology lasers are employed to treat diverse skin conditions, for medical as well as cosmetic reasons. There is also the photodynamic therapy to treat some types of cancers. For this therapy a photosensitive material is injected in the body, which preferentially locates itself in the cancer tumors. Upon illumination of the tumor with a laser the cancer cells will die more rapidly than healthy cells.

Diagnostics in medicine is also performed with light. A popular and recent form of monitoring is the non-invasive oxygen saturation meter. This device consists of a probe with a light emitting diode (LED) and a photodetector. It can be attached to a fingertip or an earlobe, so that light is radiated through tissue. The light emits in two different wavelengths, one in the visible and one in the infrared region. This light is absorbed with different rates by the hemoglobin. From this one can deduce the oxygen saturation level in the blood. Another frequently used, though "unpopular", diagnostic is endoscopy. Here one uses a flexible tube filled with optical fibres to look inside of the body, e.g. the stomach. The fiber bundle transports an optical image to a camera outside of the body. In recent years tomographic imaging of tissues by use of light is gaining importance. Despite the strong scattering of light during propagation through the body one can realise imaging. To this end one employs near-infrared wavelengths as tissue, even the skull, is

sufficiently transparent in this spectral region. In this way it is e.g. possible to map the oxygen concentration in the brain.

1.3.3 Measurement and sensor applications

Light is incredibly versatile for measuring many properties of materials and systems. First of all there is the broad field of spectroscopy - the measurement of light spectra. The light absorption or emission of materials (after some excitation) often shows a characteristic spectrum with very sharp peaks. This is employed both in fundamental and applied research on a large scale. As a consequence there is a multitude of spectroscopic instruments.

Besides this, one uses light to measure all kinds of physical quantities, such as temperature, distance, displacement, elastic strain, gas concentration, material composition etc. Most geometrical measurements (distance, displacement, deformation) are simply based on a change of the optical path length. If one works interferometrically it is possible to measure distances with an accuracy that is a very small fraction of the wavelength, e.g. a few nanometer. Ultra-precise translation stages are often equipped with a laser-interferometer. Other physical quantities such as temperature are often measured because of their influence on the refractive index, giving rise again to a change in optical path length.

Glass fibre sensors constitute a special class of optical sensors. Here the sensing is built in a certain way into the fibre, and the light is transported by this same fibre to the sensor. This method is used to monitor the safety of large constructions. The bridge over the Gentse Ringvaart near Flanders Expo e.g. is equipped with fibre sensors embedded into the concrete.

Another important class of sensors are the biomolecular detectors. They sense different kinds of biomolecules such as antibodies, proteins or DNA. Many techniques here work optically. One way is to selectively attach a fluorescent molecule to the biomolecule, so that emission shows the presence of this biomolecule. Another way is to attach the biomolecule to another one, and to detect the change of refractive index caused by this attachment. These techniques are routinely used today in labs, but they are often expensive and bulky to use in a normal medicine practice. Microphotonics could change this in the future by building “labs-on-a-chip”.

1.3.4 Visualisation

Because we have eyes and are used to see three-dimensional objects projected onto two dimensions (and to interpret them three-dimensionally with our brain), visualisation systems are a very important application of photonics.

The purely optical systems for direct observation with the eye were historically the first systems, and they remain important today. The field of microscopes, telescopes and projectors encompasses many variants. The art is to build a system that transports as much light as possible from object to image plane, combining a resolution as high as possible with an imaging area as large as possible. This combination of demands leads to very complex lens systems. For classical systems the resolution has always been limited by the diffraction limit. For a microscope this means that one is unable to resolve details smaller than the wavelength. In recent years, however, one has built systems that break this barrier, the Scanning Near-field Optical Microscopes (SNOM).

With the emergence of photography it became possible to capture images on film via photochemical methods. Later the vacuum systems such as vidicon or cathode ray tube appeared. Here photons are converted into electrons in vacuum (photoelectric effect) and subsequently into electric current. Inversely, current is converted into electrons in vacuum and then into photons via phosphorescence. By scanning images sequentially in these systems, a two-dimensional picture (or a sequence of pictures) is converted into a time signal, or the other way around.

Gradually these high-voltage vacuum systems are replaced by electronic low-voltage systems. Modern cameras work with Silicon chips (CCD-chips), that port the image into a sequential electronic signal. In recent years CMOS-chips with built-in detectors are used more and more. For displays the cathode ray tubes are replaced by flat panel displays, mainly based on liquid crystal technology (LCD). For big screen projection LCD-technology is also used, unless the picture has to be very bright (e.g. in daylight), in which case one employs huge LED-matrix panels.

In most applications the visible spectrum is used. However, there are special cameras that capture light in other wavelength regions, such as the infrared radiation. In this way one can detect the thermal radiation of an object and design night vision systems.

The field of graphics uses many optical techniques to convert electronic information onto film or paper. In laser- and LED-printers a photosensitive drum is illuminated line after line, attracting toner, and transferring it to paper. For professional printing techniques such as offset printing the printing plate is fabricated with laser illumination.

1.3.5 Information technology

Of all photonics' applications optical communication probably has the deepest impact on society. The internet works because of the optical fibres that transport the data streams between continents, countries, regions and cities. The story of optical communication is a combination of the semiconductor laser and the glass fibre itself. This fibre is able to transport incredibly high datarates over very long distances. Nowadays one laser typically sends 10 Gigabit/s in a fibre, and this will evolve to 40 Gigabit/s. If one combines the light of various lasers, with different wavelengths, into a fibre, one can reach datarates of over 1 Terabit/s. In the fibre itself light can propagate about twenty kilometers before being attenuated by a factor of 2. After 100 km this attenuation is about a factor of 30, and this is the typical length of a telecom-link. To proceed over longer distances, it is possible to use fibre based optical amplifiers. In these devices light is amplified directly by stimulated emission.

In recent years optical communication is increasingly employed for shorter links. As datarates become higher, or the number of connections within a volume becomes larger, the distance above which fibre is more interesting than electrical copper wire becomes smaller. More and more local networks are designed optically, especially for connections above about 100 meters. Fiber-to-the-home will undoubtedly arrive in the future, although introduction is hampered by large scale infrastructure investments. For very broadband connections between electronic hubs one often uses parallel optical links. Slowly work has been done to change the wiring on the level of a printed board into an optical connection. There is also research to implement the highest level of wiring within an integrated circuit by means of dense optical waveguides.

A second important application of photonics for information technology is optical data storage. With the technology of CD and DVD one can store Gigabytes of data on a cheap and portable

plastic disc. There are three forms: read only, write once and rewritable. Roughly speaking the capacity is limited by the diameter of the used laser spot, which is on the order of the light wavelength. Thus it is logical that the generations of optical data storage move to shorter wavelengths: from infrared (CD), to red (DVD) to blue (e.g. BlueRay).

1.4 Photonics - education

A few decades ago, the range of light applications was fairly limited. Therefore photonics never constituted the core of an engineer's training. Instead it was a side aspect of educations in electronics, physics and partly material science. However, because of the wide spectrum of photonic applications, as we described previously, there are more and more course programs with photonics as its main discipline, and in which electronics, physics and material science move to the periphery. This mostly happens on the master level, although examples of bachelors in optics or photonics also exist. These courses have a multidisciplinary approach. Besides optics one needs knowledge and techniques of material science, technology, mathematics and numerics, measurement and systems, etc.

1.5 Photonics - this course

The goal of this course is to provide insight into the basic principles and concepts of photonics. Moreover, it gives information about the significant materials, components and systems. The target audience consists of two groups: those who will not specialize in photonics and those who will. This course keeps both groups in mind.

There are many ways to arrange an introductory photonics text. For this course we chose to roughly follow the historical development. In this way the various models and techniques appear in increasing order of complexity: ray optics, scalar wave theory, electromagnetism and quantum electrodynamics (or quantum optics). It is important to stress that all models remain relevant and are used nowadays to design photonic systems. Indeed, it is a rule that one should not use a more complex model than necessary for the problem at hand. The same rule is applied to the mathematical descriptions in this course. They are as involved as necessary in order to understand the basic principles or to make simple designs. Figure 1.2 schematically depicts the four basic models in optics. It is clear that the simpler (and older) models are considered an approximation that is usable for a certain subclass of problems.

This course builds upon other courses of the bachelor electrotechnical engineering and applied physics, such as Physics, Electrical networks, Electromagnetism, Quantum mechanics and Semiconductor physics. Because students followed different curricula there will be a redundancy for some subjects, especially for quantum mechanics.

Any scientific discipline, including photonics, has its major reference works. Here we shortly describe a selection of important books. The next paragraph provides a complete reference. This course was inspired by these works.

Fundamentals of Photonics by **Saleh and Teich** [ST91] is closest to this course, both regarding scope and depth of description. It is recommended for anyone who needs a basic book about

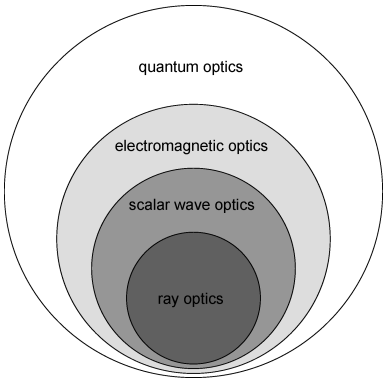

Figure 1.2: The four basic models of optics

photonics. This work provides a fairly complete overview and the level is perfect for a third year bachelor. About two thirds of its topics are presented in the course.

Principles of Optics by **Born and Wolf** [BW99] is a classic in this area. The first edition appeared in 1959 and there is already a 7th edition. This work describes classical optics (ray optics, scalar wave and electromagnetic optics) in a very thorough way (much more complete than in this course). Many topics however are not mentioned (e.g. lasers or semiconductor opto-electronic components).

The books **Optics** by **Möller** [Möl88] and **Modern Optics** by **Guenther** [Gue90] are comparable in detail to Saleh and Teich. However they are roughly restricted to the same topics as Born and Wolf. For some topics they provide an interesting alternative approach to Saleh and Teich.

The booklet **An introduction to theory and applications of quantum mechanics** by **Amnon Yariv** [Yar82] explains the difficult world of quantum mechanics with a modest amount of mathematics, without simplifying too much however.

The book **Principles of Lasers** by **Orazio Svelto** [Sve98] restricts itself to lasers and is one of the most important basic works in the field. The first edition dates from 1976 and there is already a fourth edition. The book assumes no previous knowledge about lasers but goes into much more detail than possible in this course.

The Photonics Directory [PDi03] is a book in four parts with a new edition every year. It offers a window to the "real" world of products, technology and photonics-related companies. Part 3 contains about 200 short tutorials on various concrete products. In this way one obtains a picture of the state of the art quickly and efficiently. Part 4 is a dictionary of terms and acronyms in photonics. The Photonics Directory is freely available on the web, except part 3.

Bibliography

[BW99] M. Born and E. Wolf. *Principles of Optics*. Cambridge University Press, ISBN 0-521-642221, 7th (expanded) edition, 1999.

[Gue90] R. Guenther. *Modern Optics*. John Wiley and Sons, ISBN 0-471-60538-7, 1990.

[Möl88] K.D. Möller. *Optics*. University Science Books, ISBN 0-935702-145-8, 1988.

[PDi03] *The Photonics Directory, Book 1-4, www.photonicsdirectory.com*. Laurin Publishing Company, Book 1: The Photonics Corporate Guide, Book 2: The Photonics Buyers' guide, Book 3: The Photonics Handbook, Book 4: The Photonics Dictionary, 2003.

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

[Sve98] O. Svelto. *Principles of Lasers*. Plenum Press, ISBN 0-306-45748-2, 4th edition, 1998.

[Yar82] A. Yariv. *An Introduction to Theory and Applications of Quantum Mechanics*. John Wiley and Sons, 1982.

Part II

Propagation of Light

Chapter 2

Quantities and Units of Light

Contents

In this chapter we present a few basic concepts about light units, illumination and color. In the first part we discuss the various quantities that are used to characterize light. This is a fairly complicated system, expressed in lesser known units. Moreover, for the *photometric* quantities one implicitly takes the properties of the eye into account.

2.1 The Electromagnetic Spectrum

Light is mostly defined as electromagnetic radiation with a frequency close to the part that is visible to the human eye. Therefore, one separates apart from visible light, infrared light (lower frequency) and ultraviolet light (higher frequency). Depending on the discipline the frequency is described by a number of different units. The frequency itself, noted as f or ν (the latter in a physics context), is evidently the clearest. However, it is used less commonly, probably because the numbers are rather big: typically $10^{14}Hz$ or $100THz$ (TeraHertz). The most used quantity is wavelength λ, defined as the distance the (monochromatic) light traverses during one period of the sinusoidal time signal:

$$\lambda = \frac{c}{nf}, \tag{2.1}$$

with c the speed of light ($c = 299792458m/s$) in vacuum and n the refractive index of the material (n is the square root of the relative permittivity ϵ_r). One notices that the wavelength depends on the refractive index, because the speed of light in the material depends on it. Thus, the wavelength

quantity	λ	f of ν	E	σ
unit	μm	THz	eV	$1/cm$
value	1	300	1.24	10000
trend w.r.t. λ	$\propto \lambda$	$\propto 1/\lambda$	$\propto 1/\lambda$	$\propto 1/\lambda$

Table 2.1: Electromagnetic quantities at a wavelength of $1\mu m$.

changes during propagation from one material to the next (while the frequency remains constant). Therefore, one often uses the vacuum wavelength, which is the wavelength of the light should it propagate through vacuum (index $n = 1$). Most of the time the vacuum wavelength is simply called wavelength. It is commonly expressed in μm or nm. The unit $\mathring{A}$ (Angström) is frequently used, but is to be avoided ($1\mathring{A} = 0.1nm$). Another measure of frequency is the photon energy. Light has both a wave and a particle character. It consists of elementary quanta, called photons. They have an energy E proportional to the frequency:

$$E = hf = h\nu \tag{2.2}$$

with h Planck's constant ($h = 6.626 \times 10^{-34} Js$). This can be expressed in Joule, but electron-volt (eV) is used more frequently ($1eV = 1.602 \times 10^{-19} J$). Often one needs to switch between (vacuum) wavelength (in μm) and photon energy (in eV). From the above equations we obtain the following relation:

$$E\,[eV] = \frac{1.24}{\lambda\,[\mu m]}. \tag{2.3}$$

Finally, it is common in spectroscopy to use the reciprocal quantity of wavelength, called wavenumber and noted as σ:

$$\sigma = \frac{1}{\lambda}. \tag{2.4}$$

This quantity is often expressed in $1/cm$ and shows how many wavelengths fit in 1 cm. The term wavenumber is somewhat confusing because in electromagnetism the wavenumber (k) is defined by:

$$k = \frac{2\pi}{\lambda}. \tag{2.5}$$

It is useful to memorize the values of the other quantities corresponding to a wavelength of $1\mu m$. In this way one can quickly find the magnitudes for other wavelengths, without knowing the fundamental constants. They are presented in table 2.1.

Figure 2.1 depicts the various frequency bands. The visible part is also shown in more detail. Notice that the human eye is sensitive to a limited part of this spectrum, more specifically from $380nm$ to $760nm$. These are mean values dependent on the observer and the light intensity. In optics one is mostly interested in the ultraviolet, visible and infrared parts, so spanning from $10nm$ to $100\mu m$.

Purely sinusoidal radiation does not exist in reality. So every radiation has a certain bandwidth and one refers to its spectral distribution. Figure 2.2 shows the spectral distribution of different sources with both continuous and line spectra.

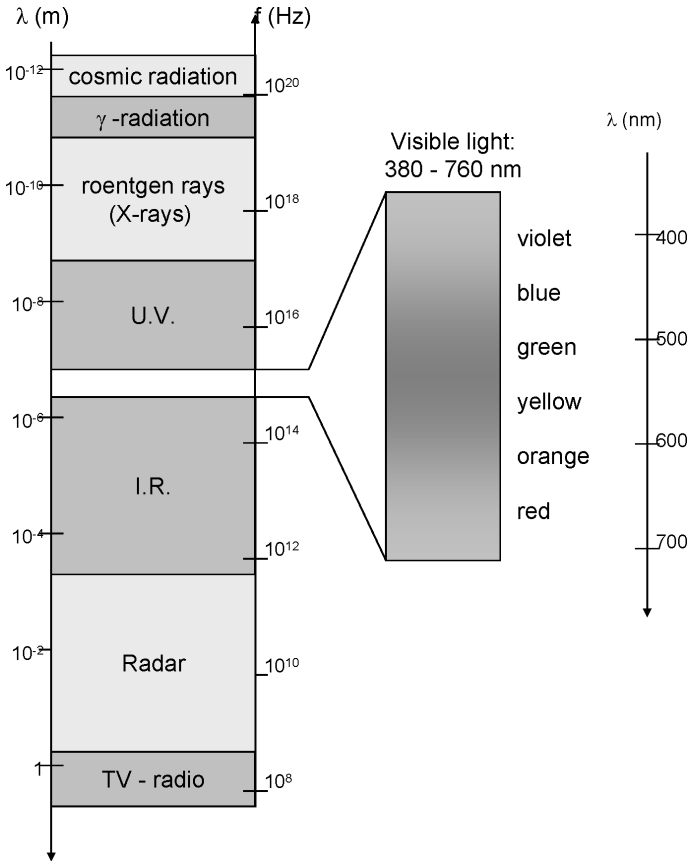

Figure 2.1: The electromagnetic spectrum.

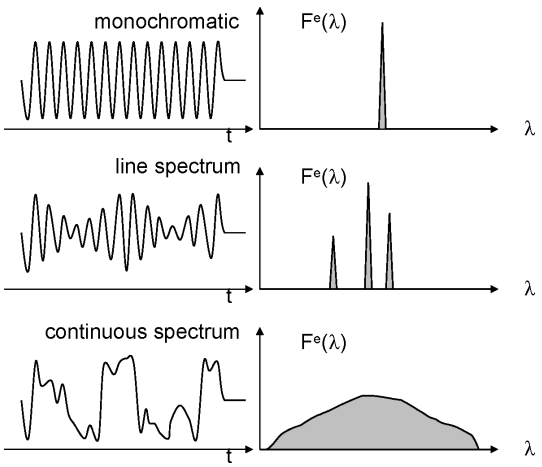

Figure 2.2: Spectra of different sources. From top to bottom: A monochromatic source (a single spectral line), a source consisting of various spectral lines, and a source with a continuous spectrum.

2.2 Units for Optical Radiation

Optical radiation is characterized with two different kinds of units. The energetic units are equivalent to the units used in physics to measure electromagnetic radiation. Examples are Watt (W), Joule (J) and the like. If one wants to describe light properties that are dependent on characteristics of the human eye, one uses photometric units. Thus, these units are a measure for the visual impression we get from the electromagnetic radiation.

2.3 Energetic Quantities

2.3.1 Radiant energy

- Symbol: Q^e
- Unit: $Joule$

This is the amount of energy transferred by electromagnetic waves (propagated during a given time through a given surface, or inside a given volume at a given instant).

2.3.2 Radiant flux

- Symbol: F^e
- Unit : $Watt$

This is the amount of radiation per time unit (the infinitesimal amount of radiation dQ^e that propagates through a given surface in an infinitesimal time dt, divided by this time dt):

$$F^e = \frac{dQ^e}{dt} \tag{2.6}$$

2.3.3 Radiant intensity

- Symbol: I^e
- Unit: $Watt/str$

For a point source (figure 2.3a) it is the radiant flux in a given direction per unit solid angle (it is the radiant flux dF^e in an infinitesimal solid angle $d\Omega$ around a given direction, divided by this solid angle).

$$I^e = \frac{dF^e}{d\Omega}. \tag{2.7}$$

The radiant intensity depends on direction.

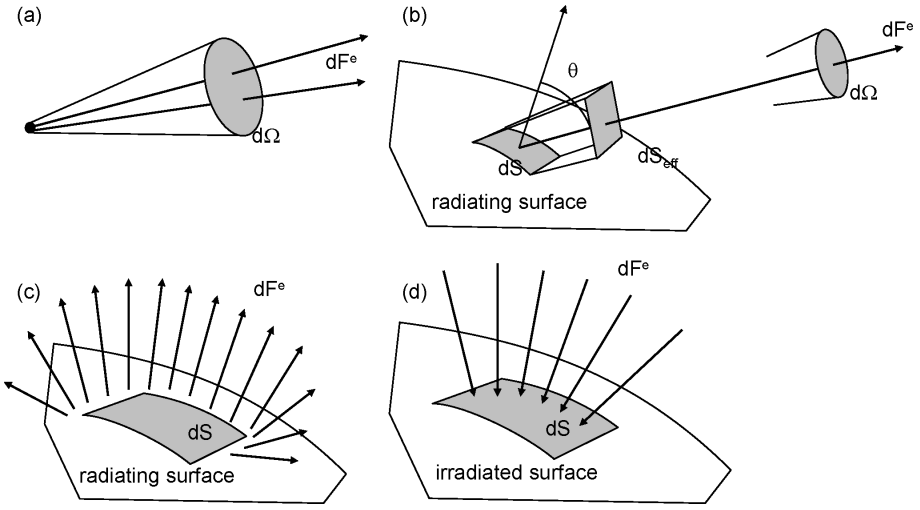

Figure 2.3: Illustration of the energetic quantities of radiation. (a) Radiant intensity I^e of a point source, (b) radiance L^e of a radiating surface, (c) radiant exitance M^e of a radiating surface and (d) irradiance E^e of a radiating surface.

Note: the unit of solid angle is *steradian* (str). A solid angle is $1str$ if for a sphere of radius 1 the part on the surface inside the solid angle has a surface area 1. Thus, the whole space around a point has a solid angle of $4\pi str$.

2.3.4 Radiance

- Symbol: L^e
- Unit : $Watt/str/m^2$

Radiance is the radiant intensity of a surface around a given point in a given direction, per unit of effective area of the surface in that direction (figure 2.3b). So, it is the radiant flux of a given infinitesimal surface dS in an infinitesimal solid angle $d\Omega$ around a given direction, divided by the effective area of dS and divided by the solid angle $d\Omega$:

$$L^e = \frac{dI^e}{dS_{eff}} \tag{2.8}$$

Here the effective area is given by $dS_{eff} = dScos\theta$ with θ the angle between the normal of the surface and the chosen direction. Thus, radiance depends on position and direction.

2.3.5 Radiant exitance

- Symbol: M^e
- Unit: $Watt/m^2$

Radiant exitance is the radiant flux per unit area, radiated by a surface (figure 2.3c) :

$$M^e = \frac{dF^e}{dS} \tag{2.9}$$

Thus it is dependent on position.

2.3.6 Irradiance

- Symbol E^e
- Unit : $Watt/m^2$

Irradiance is the opposite of radiant exitance. It is the radiant flux per unit area received by a surface (figure 2.3d):

$$E^e = \frac{dF^e}{dS} \tag{2.10}$$

2.3.7 Spectral density

It is possible to define a spectral density for all these quantities.

Example : spectral density of the radiant flux F^e:

$$F_s^e(\lambda) = \frac{dF^e}{d\lambda} \tag{2.11}$$

It is the radiant flux in an interval $d\lambda$ around a wavelength λ, divided by this wavelength interval (in $Watt/nm$). In the following we drop the subscript s of spectral density for notational simplicity.

2.4 The human eye

Figure 2.4 depicts a cross section of the human eye. The retina is the photosensitive part of the eye and consists of two kinds of light sensitive cells : rods and cones. The impression of light is a chemical reaction in these nerve cells. The cones provide sight at normal illumination (*photopic* sight) and are sensitive to color.

There are three kinds of cones:

- sensitive to red : maximum sensitivity at $580nm$
- sensitive to green : maximum sensitivity at $540nm$
- sensitive to blue : maximum sensitivity at $440nm$

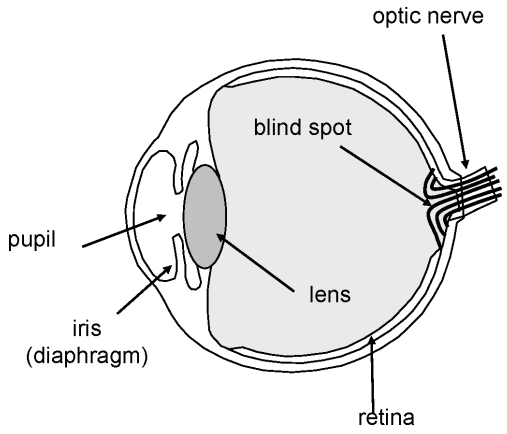

Figure 2.4: Schematic depiction of the human eye.

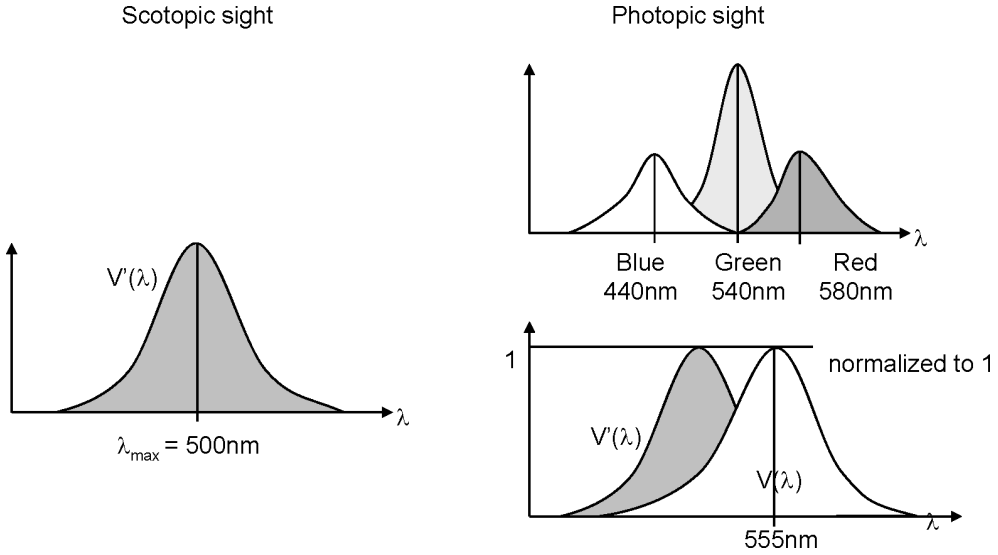

Figure 2.5: Sensitivity of rods and cones in the human eye.

At low illumination the cones become insensitive and the rods take over: this is night vision or *scotopic* sight. Rods are not sensitive to color, however they are much more sensible to light than cones. The transition area between photopic and scotopic sight is called *mesopic* sight. The light sensitivity of the human eye depends strongly on the wavelength. After extensive testing on many persons an internationally accepted spectral eye sensitivity curve was established in 1933, see figure 2.5. For photopic sight the curve $V(\lambda)$ has a maximum at $555nm$ (one obtains this curve by taking a weighted average of the spectral sensitivity curve for the three kinds of cones). Thus, as an example $2Watt$ of light with a wavelength of $610nm$ will appear as bright as $1Watt$ at $550nm$. So, the response of the eye to a radiant flux with spectral density $F^e(\lambda)$ will be proportional to:

$$\int F_s^e(\lambda)\, V(\lambda)\, d\lambda \tag{2.12}$$

For scotopic sight the eye sensitivity curve shifts to the blue.

2.5 Photometric quantities

With the spectral eye sensitivity curve $V(\lambda)$ it is possible to convert the (objective) energetic quantities into photometric quantities. Thus the latter take the response of the eye into account. The notation of photometric quantities is analogous to the energetic ones, although without the superscript ^{e}. The conversion from the radiant flux F_s^e into the luminous flux F is done with the equation:

$$F = K \int F_s^e(\lambda)\, V(\lambda)\, d\lambda \tag{2.13}$$

with $K = 680 lumen/Watt$

2.5.1 Luminous flux

- Symbol: F
- Unit : $lumen$

Using $F_s^e(\lambda)$ and $V(\lambda)$ one obtains that a radiant flux of $1Watt$ at $550nm$ corresponds to a luminous flux of $680lumen$.

2.5.2 Luminous intensity

- Symbol: I
- Unit: $candela = lumen/str$

For a point source the luminous intensity is defined as the luminous flux in a given direction per unit solid angle (see figure 2.3a):

$$I = \frac{dF}{d\Omega} \tag{2.14}$$

Historically the term candela (and the magnitude of 1 candela) originates from the luminous intensity of a candle. Today the candela is considered as one of the seven basic units of the international system of units (SI). Other photometric units are deduced from the candela. The international definition of the candela has been often changed in the 20th century. Nowadays the definition is: "The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency $540x10^{12}$ hertz and that has a radiant intensity in that direction of 1/683 watt per steradian."

2.5.3 Luminance

- Symbol: L
- Unit: $candela/m^2 = nit$

The luminance or brightness is the luminous intensity radiated by a surface around a given point in a given direction, per unit effective area of the surface in that direction. (figure 2.3b):

$$L = \frac{dI}{dS_{eff}} \tag{2.15}$$

2.5.4 Luminous exitance

- Symbol: M
- Unit: $lumen/m^2$

The luminous exitance is the luminous flux radiated by the surface per unit area (figure 2.3c):

$$M = \frac{dF}{dS} \tag{2.16}$$

2.5.5 Illuminance

- Symbol: E
- Unit[1] : $lux = lumen/m^2$

[1] In anglo-saxon literature one often finds the unit *footcandela* (fc) instead of lux. This is a unit of illuminance that corresponds to a uniform illumination of $1lumen$ on a surface of 1 square foot.

$$1fc = 1\frac{lumen}{ft^2} = 1\frac{lumen}{(0.304m)^2} = 10.76lux$$

The unit *foot-Lambert* is also employed. It is used as a measure for the luminance of lambertian emitters (see below). 1 foot-Lambert means that the surface has a luminous exitance of $1lumen$ per square foot. Furthermore, 1 foot-Lambert corresponds with a luminance of 3.426 Nit. 1 Lambert expresses that the surface has a luminous exitance of 10.000 lumen per square meter, which corresponds to 3183.1 Nit.

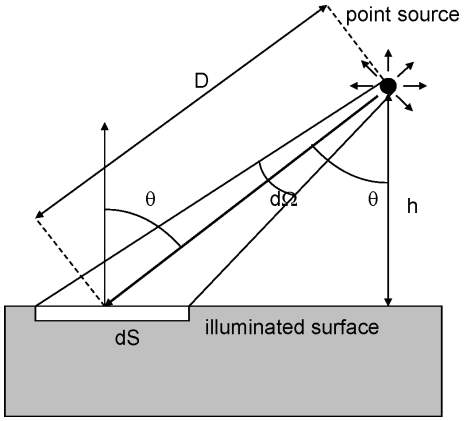

Figure 2.6: Calculation of the illuminance on a surface dS by a point source.

The illuminance is the opposite of luminous exitance: it is the luminous flux per unit area, received by the surface (figure 2.3d):

$$E = \frac{dF}{dS} \tag{2.17}$$

2.5.6 Spectral density

For all these quantities it is again possible to define a spectral density. Example: the spectral density of the luminous flux:

$$F_s(\lambda) = \frac{dF}{d\lambda} \tag{2.18}$$

or

$$F_s(\lambda) = KV(\lambda) F_s^e(\lambda). \tag{2.19}$$

2.6 Relations between different quantities

We only consider incoherent sources in this section. This means that light from different sources, or light that reaches the same point via different paths, can simply be added. The total luminous flux is then equal to the sum of all separate flux contributions. In later chapters it will become clear that this is not the case for coherent sources, such as lasers.

2.6.1 Calculating the illuminance

Point source

In practice a source is considered pointlike if the distance to the illuminated surface is at least six times as large as the largest size of the source.

First we determine the illuminance E at P on a surface at a distance D from the source, given the luminous intensity $I(\theta, \phi)$ of that source (figure 2.6).

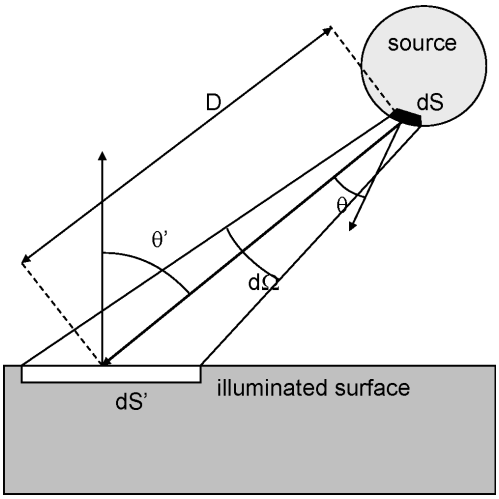

Figure 2.7: Calculation of the illuminance on the surface dS' from an extended source.

Solid angle:

$$d\Omega = \frac{dS \cos\theta}{D^2} \tag{2.20}$$

Luminous flux dF on dS :

$$dF = I\,(\theta, \phi)\,\frac{dS \cos\theta}{D^2} \tag{2.21}$$

Illuminance E on the surface :

$$E = \frac{dF}{dS} = \frac{I \cos\theta}{D^2} = \frac{I \cos^3\theta}{h^2} \tag{2.22}$$

For perpendicular incidence ($\theta = 0°$) one obtains

$$E = \frac{I}{h^2} \text{(square law)} \tag{2.23}$$

Thus, the illuminance on a surface decreases with the square of the distance D to the source (the latter being sufficiently small).

Extended source

Given the luminance $L(\theta, \phi)$ of a source, we obtain the illuminance E on a surface dS' (figure 2.7).

Luminous flux dF in a solid angle $d\Omega$:

$$dF = L\,(\theta, \phi)\ \ (dS\ \cos\theta)\ \ d\Omega \tag{2.24}$$

with

$$d\Omega = \frac{dS' \cos\theta'}{D^2} \tag{2.25}$$

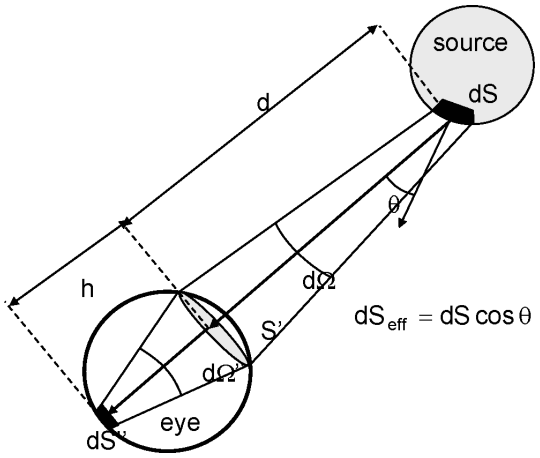

Figure 2.8: Illumination of the retina by an extended source.

Illuminance dE on dS' due to the surface dS on the source:

$$dE = L \frac{\cos\theta \ \cos\theta'}{D^2} dS \tag{2.26}$$

Total illuminance E by the complete source:

$$E = \int\int_{source} L \frac{\cos\theta \ \cos\theta'}{D^2} \ dS \tag{2.27}$$

2.6.2 Relation between the illuminance on the retina and the luminance of a light source

It is easy to show that the eye is sensitive to the luminance L of a source (figure 2.8).

The effective area of the surface dS on the source is:

$$dS_{eff} = dS \cos\theta \tag{2.28}$$

The luminous flux incident on the eye lens with surface S' from dS is :

$$\begin{aligned} dF &= LdS\cos\theta d\Omega \text{ with } d\Omega = \frac{S'}{d^2} \\ &= \frac{LdS \ \cos\theta S'}{d^2}, \end{aligned} \tag{2.29}$$

with d the distance from source to lens. The image magnification of the eye is given by h/d (with h the distance lens-retina). The surface magnification is thus :

$$\frac{dS''}{dS_{eff}} = \frac{h^2}{d^2}, \tag{2.30}$$

with dS'' the image of dS on the retina.

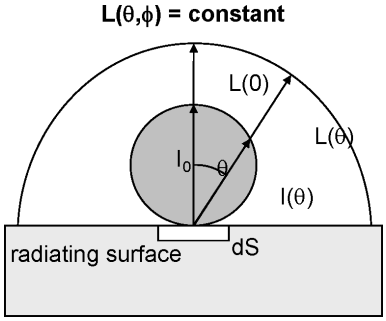

Figure 2.9: Lambert's law.

The luminous flux dF incident on the lens from dS is also the flux incident on dS''. We obtain:

$$dF = L\frac{dS''S'}{h^2} \tag{2.31}$$

The illuminance on dS'' becomes:

$$E = \frac{dF}{dS''} = L\frac{S'}{h^2} = Ld\Omega'', \tag{2.32}$$

with $d\Omega''$ a constant for the eye. This last expression implies that the illuminance on the retina is independent of the distance from the source to the eye. At first glance this seems surprising, as a source appears less bright as the distance increases. However, with increasing distance the image of the source on the retina becomes smaller. The amount of light incident on the eye decreases too, with the same rate. Therefore the illuminance (luminous flux on the retina divided by unit area!) remains constant.

Because the eye is a "measurement device" for luminance, this explains why "brightness" is a synonym for this quantity.

2.6.3 Lambert's law

A radiating surface is a lambertian emitter if the luminance of a point on the surface is independent of the direction. Consider dS on this surface (figure 2.9) then $L(\theta, \phi) = cst$.

The luminous intensity dI emitted by the surface dS is:

$$\begin{aligned} dI &= LdS_{eff} \\ &= LdS\cos\theta \end{aligned} \tag{2.33}$$

or

$$dI(\theta) = dI_0\cos\theta, \tag{2.34}$$

with dI_0 the luminous intensity for the normal.

	Recommended illumination (lux)
offices	$500 - 1000$
very precise work	$1000 - 5000$
living space - local	$500 - 1000$
living space - ambient	$50 - 100$

Table 2.2: Recommended illumination for artificial light.

For a source that complies with Lambert's law one can derive a relation between the luminance and the luminous exitance by calculating the total luminous flux dF radiated by dS in a half space:

$$\begin{aligned} dF &= \int dI d\Omega = dI_0 \int_0^{\pi/2} \cos\theta \sin\theta d\theta \int_0^{2\pi} d\varphi \\ &= \pi dI_0 \\ &= \pi L dS \end{aligned} \tag{2.35}$$

Thus, the luminous exitance is:

$$M = \frac{dF}{dS} = \pi L \tag{2.36}$$

Many incoherent sources have a surface that is described well with Lambert's law. Examples: the sun, the filament in a tungsten lamp, a light emitting diode. Many diffusely reflecting surfaces reflect according to Lambert's law, so quasi independent of the direction of incidence of light on the reflector. This is why a piece of paper appears as bright from every viewing angle. It also explains why the sun appears as a uniformly lit disk (the edges remain bright). However, the (full) moon may seem like a uniform disk but it is not: the illuminance decreases with the cosine of the incidence angle of the sunlight (with the normal). Thus, the luminance decreases according to this rate.

2.6.4 Typical values for luminance and illuminance

Illumination is by far the easiest and most used parameter for good lighting. One expresses the luminous flux ($lumen$) that reaches the surfaces surrounding the person, per unit area ($lumen/m^2$ of lux). In fact, this is not such a good measure, because the eye is sensitive to the luminance ($candela/m^2$). However, it is much easier to measure illumination (lux), and therefore the latter is mostly used.

In daylight the illumination is between 1000 and 100000 lux. From about 10000 lux the eye functions optimally. This means that a minimal effort gives a maximal performance. For technical and economical reasons the illumination of artificial light is smaller. A few typical values are given in table 2.2.

Some approximate illumination values are given in table 2.3. As mentioned earlier the luminance is the most important parameter for the eye. In practice it is important to avoid large luminance differences in the visual field. Depending upon the nature of the work the contrasts should be between 1:3 and 1:40. A number of typical luminance values are shown in table 2.4.

	Illumination (lux)
summer sun	100000
winter sun	10000
sunrise	500
full moon	0.25
retina sensitivity	10^{-9}
400 ISO film sensitivity (1 second)	10^{-2}

Table 2.3: Typical illumination values.

	Luminance ($candela/m^2$)
sun	$1.65\ 10^9$
moon	$2.5\ 10^3$
filament in an incandescent lamp	$7\ 10^6$
fluorescent lamp	$8\ 10^3$
LED	$10^4 - 10^6$
laser (1 Watt, green)	10^{15}
white paper (80% reflection, 400 lux)	10^2
grey paper (40% reflection, 400 lux)	50
black paper (4% reflection, 400 lux)	5
luminance needed for photopic sight	$1 - 10$ (visual field average)
luminance needed for scotopic sight	$0.01 - 0.1$ (visual field average)

Table 2.4: Typical luminance values.

The following methods can be used to limit the luminance ratios:

- use of large light sources with low luminance,
- partial screening of sources by armature or architectural features,
- avoidance of mirroring surfaces that bring sources into the visual field.

Collimated light is much more problematic than diffuse light (from all directions). However, the former is often used, because of economical or aesthetic reasons (shadow contrast).

2.7 Summary

The most important classification of electromagnetic radiation happens by means of the wavelength, or related to that, frequency or photon energy. The electromagnetic spectrum changes from radiowaves with low energy, over visible light to gamma radiation with very high photon energy.

The amount of light or radiation is expressed with energetic or photometric quantities. An overview of these quantities and their corresponding units is shown in figure 2.10 on page 2–16. Every energetic quantity has an equivalent photometric quantity. For conversion one uses the spectral density and the eye sensitivity curve $V(\lambda)$.

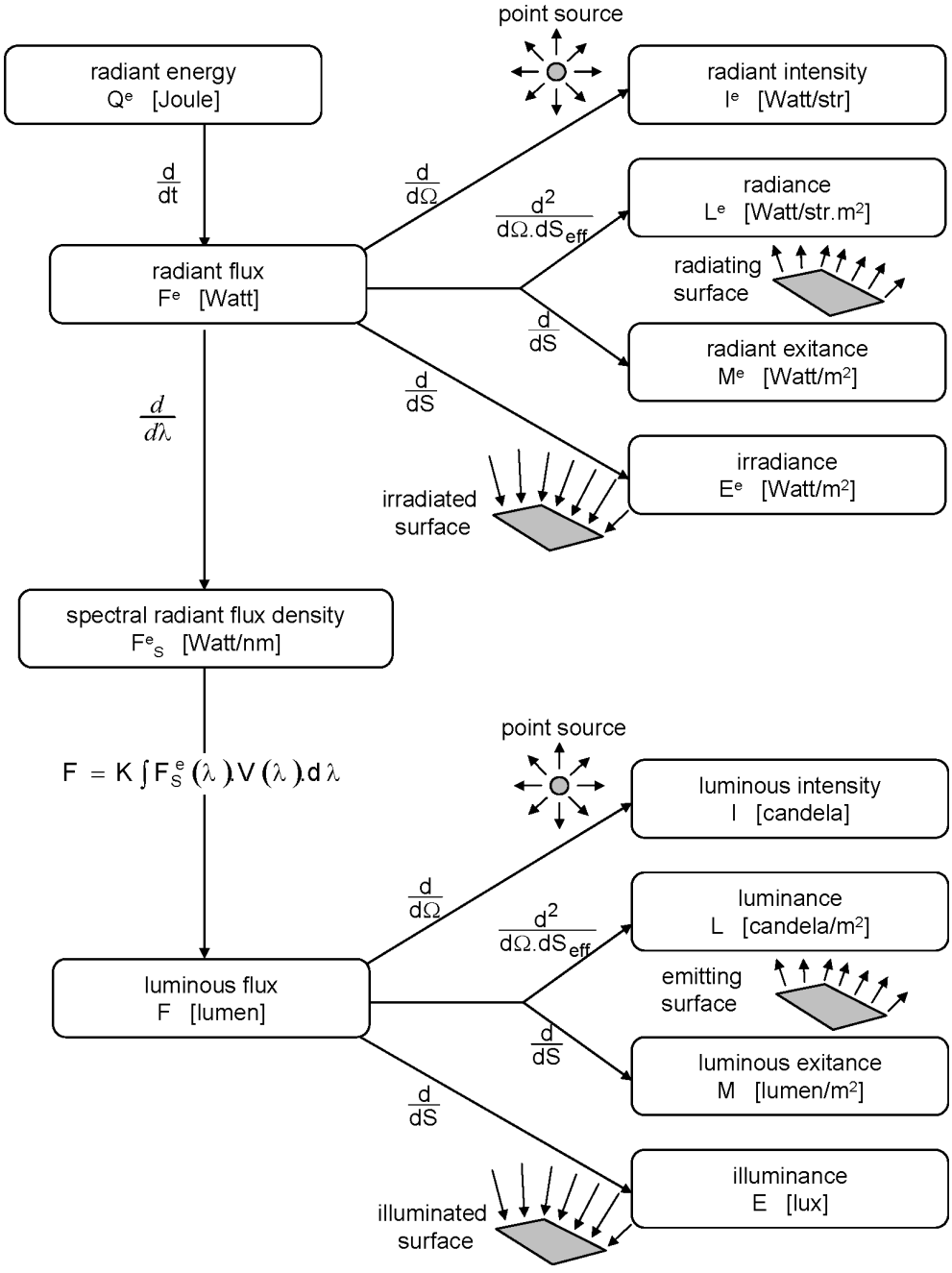

Figure 2.10: Energetic and photometric quantities.

We have seen how we can calculate the irradiance and the illuminance on a given surface, for a point source and an extended source. We have also considered a special kind of radiating surface, the lambertian emitter, with a constant luminance (or radiance) in all directions.

Bibliography

[BW99] M. Born and E. Wolf. *Principles of Optics*. Cambridge University Press, ISBN 0-521-642221, 7th (expanded) edition, 1999.

Chapter 3

Geometric Optics

Contents

In this chapter we describe light in a macroscopic environment, on a scale much larger than the wavelength. This enables us to make a number of assumptions, so that we can treat light as rays. With this method we can describe the refraction of light at an interface between two materials, and understand optical systems made of lenses and mirrors. At the end of the chapter we discuss a number of applications.

3.1 Introduction

Any form of electromagnetic energy, including light, can be viewed as beams of energy, or rays, that are emitted from an energy source. This view is different from the wave or particle character. In free space the rays have straight paths, whereas they can be reflected and/or bent (refracted) at a change in the medium. In fact, this corresponds to a simplified depiction of wave theory (a ray is a kind of local plane wave), and various optical phenomena can be easily explained using this theory. However, other phenomena (such as diffraction and interference) cannot be described by ray optics.

Roughly speaking, the ray model is accurate if the dimensions of the structural variations are much larger than the wavelength, and one is not concerned about the intensity distribution at the convergence point of different rays. Furthermore, it has no sense to try to determine the diameter of a beam or ray. Physically, this diameter cannot be infinitely small. Indeed, a ray cannot pass through a small aperture, as diffraction would occur. Note that the wavelength has no importance in ray optics (except for the wavelength dependance or dispersion of the refractive index), because

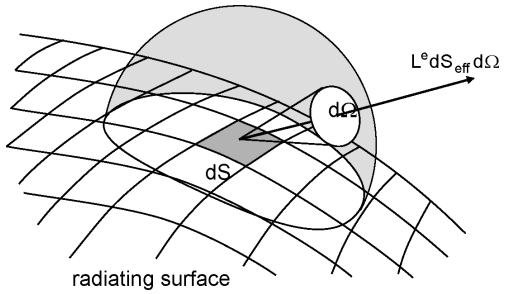

Figure 3.1: Ray presentation of a radiating surface.

reflection and refraction are not influenced by the wavelength. Another way to view ray optics is that it is a high frequency approximation. Thus the wavelength is considered infinitely small, which changes nothing to the laws of reflection or refraction, as previously mentioned. With an infinitesimal wavelength one can then assume that the rays are thin. Graphically they are depicted as lines, that show the traversed path through the system (*ray tracing*). Hence the name geometric or ray optics. Despite the limitations and approximations, geometric optics remains a very useful theory for the analysis of many optical systems, especially lens and mirror systems. Indeed, the design of complex lens systems is conducted by means of ray tracing. Here the objective is to obtain perfect imaging, which means that all rays starting from a point of the object cross each other at the same point in the image plane, and this for all points on the object (this is called a stigmatic image). Shortcomings on this are called aberrations. Furthermore, one desires that the image is an undeformed copy of the object (with optional scaling), and that a planar object is imaged into a planar image. If an imaging system is perfect (no aberrations, no deformation), this does not mean that the resolution is also perfect (so that a point in the object plane is imaged onto an infinitely small point in the image plane). Ray optics alone is not sufficient to determine the resolution, and one also needs diffraction theory. In reality the resolution can be limited both by aberrations and diffraction. The factor that is most limiting depends on the situation. A system with few or no aberrations will mainly be limited by diffraction, and this is called a diffraction limited system. Diffraction theory is not treated in this chapter. The simplest way to study diffraction is by means of Gaussian beams. This is the subject of chapter 5.

3.2 General concepts of ray theory

3.2.1 Ray representations of radiating objects

Rays originate from radiating objects. A radiating object can be a point source (that produces light energy itself), or simply an object that reflects or transmits incident light. The question is how to deduce a good ray representation for a given radiating object. Clearly, it is easy if the radiance (or luminance) of each point on the source is known for all directions. Then, one has to discretize the surface of the source into a finite number of points, and for each point one has to discretize the solid angle into a finite number of angles (figure 3.1).

In this way one assigns to every point-angle combination a ray with radiant (or luminous) flux equal to the corresponding radiance (or luminance) multiplied by the discretized surface (dS_{eff}) and solid angle unit ($d\Omega$). As the discretization becomes finer, the ray representation improves.

However, using diffraction theory one can prove that it is useless to choose $dSd\Omega$ smaller than λ^2. This would not enhance the results. Usually $dSd\Omega$ is much larger than λ^2 for ray calculations.

Conversely, we have to know how to calculate the irradiance (or illuminance) on an object, based on a discrete set of incident rays. In practice, one discretizes the surface of the irradiated object and for every small section the total incident flux is the sum of the fluxes of all incident rays. For a fine discretization of the irradiated object, the number of rays per surface section will be small, which leads to large discretization errors. Furthermore, we assume that the total power of a set of rays is equal to the sum of the individual ray powers. This assumption is correct if the electromagnetic fields associated with the rays have no phase relation with each other (are *incoherent*).

3.2.2 Postulates of ray optics

- Light propagates as rays. The rays are emitted by a source and can be perceived if they reach a detector (e.g. the eye).
- An optical medium is characterized by its *refractive index* $n \geq 1$. The refractive index is the ratio between the propagation speed of light v in the medium and the propagation speed in vacuum c. The time needed for light to traverse a certain distance d is equal to d/v or nd/c. The product nd is called the *optical path length*.
- In an inhomogeneous medium the refractive index $n(\mathbf{r})$ is a function of the position $\mathbf{r} = (x, y, z)$. Then the optical path length between two points P and P' along a certain path becomes

$$\text{Optical path length} = \int_P^{P'} n\left(\mathbf{r}\right) ds, \tag{3.1}$$

  with ds the differential length along the path. The time necessary to traverse the path is proportional to the optical path length.
- **Fermat's principle.** To propagate from point P to P' rays will follow a path so that the optical path length is an extremum over neighboring paths. This extremum can be a maximum, a minimum or an inflection point. In practice, we mostly encounter minima, so that *light follows the trajectory with least optical path length*. In another way:

$$\delta \int_P^{P'} n\left(\mathbf{r}\right) ds = 0. \tag{3.2}$$

  Sometimes the previous is true for different paths, and light propagates simultaneously along these trajectories.

Fermat's principle contains information about the path of a ray from P to P'. However, no fundamental law should be inferred from this, as it is explained perfectly by the wave character of light (thus from Maxwell's equations). Wave theory shows that the trajectory of least optical path length corresponds with the path along which the waves interfere constructively.

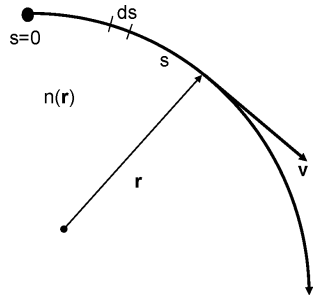

Figure 3.2: Propagation of a ray of light.

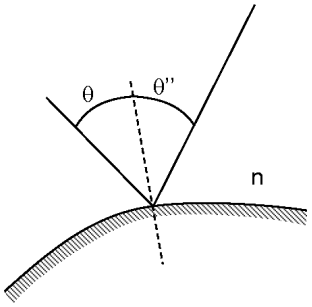

Figure 3.3: Reflection of light on a mirror

3.2.3 Propagation in a homogeneous medium

In a homogeneous medium the refractive index, and thus the speed of light v, is the same everywhere. Therefore, the *shortest optical path length* corresponds to the *shortest distance*. This is known as **Hero's principle**. In a homogeneous medium light propagates along a straight line.

3.2.4 Mirror reflection

Consider a homogeneous medium with a perfectly reflecting surface. This can be made of polished metal, or dielectric films deposited on a substrate. The mirror surface will reflect light according to the law of reflection:

- The reflected ray lies in the same plane as the incident ray and the normal on the mirror surface.
- The angle θ'' of the reflected ray with the normal is the same as the angle θ of the incident ray (figure 3.3).

A few specific cases of mirrors are depicted in figure 3.4. A plane mirror reflects light coming from P so that the reflected rays converge at point P', on the other side of the mirror. P' is called the image of P. As discussed later on, this is a virtual image: the reflected rays never really cross P'.

In a parabolic mirror all the rays that are parallel to the axis are *focused* on one point of the axis, the *focus*. These mirrors are used in telescopes, or to generate parallel beams.

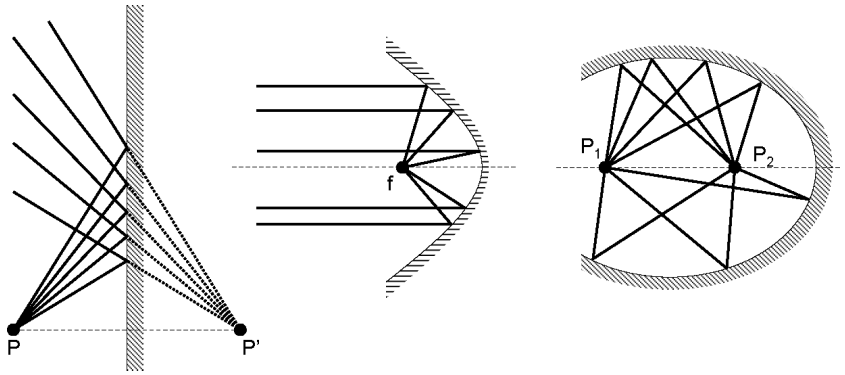

Figure 3.4: Examples of reflection. From left to right: Plane mirror, parabolic mirror, elliptical mirror.

An elliptical mirror has two foci P_1 and P_2. All the light from one point is focused on the other, and vice versa. The optical path length between P_1 and P_2 is the same for all trajectories.

3.2.5 Interface between homogeneous media

In principle the ray trajectory through a system with piecewise constant media is simple. Inside the media the rays follow a straight line. At an interface between media with indices n and n' the incident ray is split into a reflected ray and a refracted ray that propagates on the other side (figure 3.5).

Snell's law

At an interface the angle of the incident ray and the angle of the refracted ray are different. The ray is *refracted* according to the refraction law (figure 3.5):

- The refracted ray lies in the same plane as the incident ray and the normal on the interface.
- The angle θ' of the refracted ray with the normal relates to the angle θ of the incident ray according to *Snell's law*:

$$n \sin\theta = n' \sin\theta' \tag{3.3}$$

The curvature of the surface at the point of incidence has no influence on this law.

In a prism, see figure 3.6, light is refracted twice by a flat interface. The angle θ_d of the output ray relative to the input ray is calculated by applying Snell's law two times:

$$\theta_d = \theta - \alpha + \arcsin\left(\sin\alpha\sqrt{\frac{n'^2}{n^2} - \sin^2\theta} - \cos\alpha\sin\theta\right). \tag{3.4}$$

For a thin prism (α small) and paraxial incidence (θ small) the expression simplifies to:

$$\theta_d \approx \left(\frac{n'}{n} - 1\right)\alpha \tag{3.5}$$

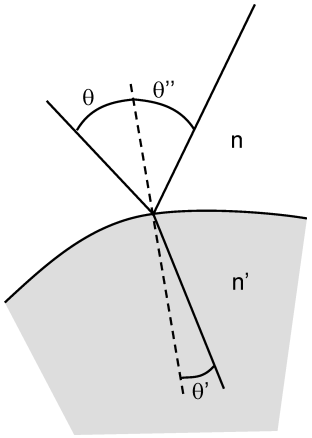

Figure 3.5: Refraction of light at an interface: Snell's law.

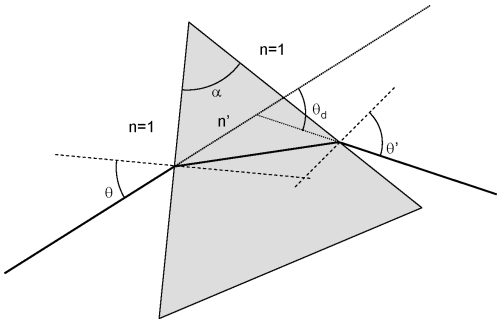

Figure 3.6: Refraction of light in a prism.

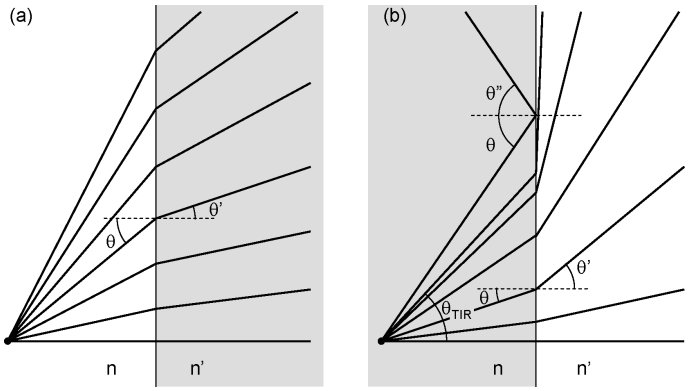

Figure 3.7: (a) External and (b) internal refraction and total internal reflection.

Reflection and transmission

Upon reaching an interface, part of the light power is refracted, while the rest is reflected. The reflection and transmission coefficients are given by the Fresnel laws for plane waves. However to derive these, we need a rigorous electromagnetical approach (see chapter 6). For example for perpendicular incidence one obtains for the power reflection and transmission:

$$R = \left(\frac{n-n'}{n+n'}\right)^2 \tag{3.6}$$

$$T = \frac{4nn'}{(n+n')^2} \tag{3.7}$$

For an air-glass (or glass-air) interface and perpendicular rays there is a power transmission loss of about 4% (most glasses have a refractive index n of about 1.5). The loss does not influence the trajectory of a ray, but of course it can lead to a drastic power reduction.

The reflections itself can cause problems. Therefore one often uses anti-reflection layers. Unfortunately these only work well over a limited wavelength range.

3.2.6 Total Internal Reflection

When light reaches an interface, it is refracted according to Snell's law. If the rays propagate from a low index material n into a higher index material n', one can find a refraction angle θ' for every incidence angle θ, see figure 3.7a. This is called *external refraction*, because the interface refracts 'from the outside to the inside'.

In the opposite case, see figure 3.7b, it is sometimes impossible to find an exit angle θ' corresponding to an incidence angle θ, according to Snell's law. Because we go 'from the inside to the outside' of the material (*internal refraction*), the exit angle θ' will always be larger than the incoming angle. For incidence angle $\theta = \theta_{TIR}$ the exiting ray will propagate at an angle $\theta' = 90^\circ$ with the normal. θ_{TIR} is called the *critical angle* and obeys:

$$\theta_{TIR} = \arcsin\frac{n'}{n}. \tag{3.8}$$

If $\theta > \theta_{TIR}$ Snell's law no longer applies. Then the interface behaves as a perfect mirror, and the incoming ray is reflected with $\theta'' = \theta$. This phenomenon is called *total internal reflection* (TIR). It is often used to replace metallic mirrors, as in a reflection prism (see section 3.6.4). Various waveguides (see chapter 7) are based on this principle.

3.2.7 Curved surfaces

At a flat surface diverging rays continue to diverge. Thus, in this way rays that originate from a point cannot be focused on another point. To change the converging or diverging character of a bundle of rays one has to use curved surfaces. This is employed in lenses. Usually one employs spherical surfaces, for technological reasons. Materials are easily polished into the spherical shape. In special situations, especially when strong refraction is desired, one uses *aspherical* surfaces.

Despite the simplicity of Snell's law, it is clear that it is not straightforward to obtain analytical expressions for ray trajectories with aspherical interfaces. For spherical surfaces the situation is manageable, except when there are multiple interfaces. Then the expressions quickly become cumbersome, because of successive sines and arcsines. For evaluation of these equations one needs a computer. Therefore, it is more useful to employ software that directly calculates the ray paths through an arbitrary lens system. However, such an approach does not deliver general and simple (albeit approximate) rules, that allow to get intuitive insight into the behavior of a system. In section 3.3 we describe such an approximate theory: *paraxial optics*.

3.2.8 Rays in inhomogeneous media - the ray equation

In a medium where the refractive index $n(\mathbf{r})$ depends on the position $\mathbf{r} = (x, y, z)$ light does not necessarily propagate along a straight line. If $n(\mathbf{r})$ is continuous the material is called *graded index* (GRIN). Often these materials are manufactured by gradually doping an optical material (e.g. glass). By carefully choosing the index profile of the GRIN-material it is possible to reach the same effect as with a piecewise constant component, such as a lens or a prism (see section 3.6.7).

To determine the ray trajectory in such a medium, we start from Fermat's principle, which states that light follows a path with minimal optical path length, in relation to neighboring paths:

$$\delta \int_{P}^{P'} n\left(\mathbf{r}\right) ds = 0,$$

with ds the differential along a path between P and P' (see figure 3.2 on page 3–4). Describing this path with the vector $\mathbf{r}(s)$, variational calculus shows us that the components $x(s)$, $y(s)$ and $z(s)$ have to obey the following differential equations [Wei74]:

$$\frac{d}{ds}\left(n\frac{dx}{ds}\right) = \frac{\partial n}{\partial x}, \quad \frac{d}{ds}\left(n\frac{dy}{ds}\right) = \frac{\partial n}{\partial y}, \quad \frac{d}{ds}\left(n\frac{dz}{ds}\right) = \frac{\partial n}{\partial z}, \tag{3.9}$$

or

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n. \tag{3.10}$$

This is the *ray equation*.

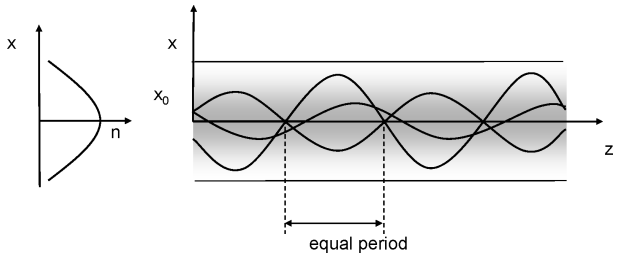

Figure 3.8: Propagation of light in a medium with parabolic index profile.

In the paraxial approximation (if all rays have a small angle with the optical z-axis) we obtain, as z is close to s:

$$\frac{d}{dz}\left(n\frac{d\mathbf{r}}{dz}\right) = \nabla n \tag{3.11}$$

As an example we calculate the ray trajectories in a system with parabolic index profile, shown in figure 3.8:

$$n = n_0 - \frac{1}{2}n_1 x^2 \tag{3.12}$$

Here n is independent of z so:

$$\frac{d^2x}{dz^2} = \frac{1}{n}\frac{dn}{dx} \tag{3.13}$$

$$= \frac{-n_1 x}{n} \tag{3.14}$$

$$\approx \frac{-n_1 x}{n_0}, \tag{3.15}$$

provided $|n - n_0|$ is small. The solution for x gives:

$$x = x_0 \cos\sqrt{\frac{n_1}{n_0}}z + x_0'\sqrt{\frac{n_0}{n_1}}\sin\sqrt{\frac{n_1}{n_0}}z \tag{3.16}$$

with x_0 and x_0' resp. the location and the slope of the incident ray at $z = 0$. Thus, the path of the ray is a sine, with a period determined exclusively by the index profile and not by the position or slope of incidence. The presentation here is two-dimensional, as if the structure were y-independent. However, in a circularly symmetric structure the previous applies in the case of meridional rays. These are rays that cross the optical (symmetry) axis. The analysis is somewhat more complex for other rays. Some rays will have a helical (spiral) trajectory around the axis, with a constant distance to the axis.

In practice the profile is only parabolical close to the axis, and constant at larger distances. This implies that only the rays that are incident on the graded part with small enough angle w.r.t. the optical axis are trapped inside the structure. The other rays escape.

The previous has major relevance to two practical situations: to optical fibers with parabolic index profile (*graded index fibres*) and to GRIN-lenses (see section 3.6.7). Also, some types of semiconductor lasers use waveguides with a parabolic index profile (see chapter 7).

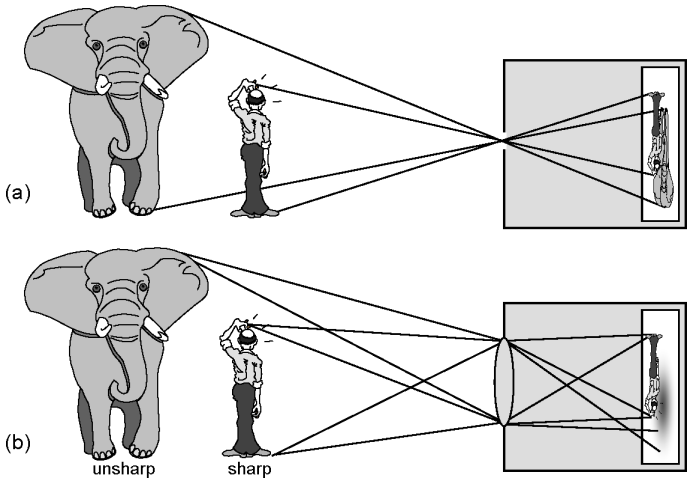

Figure 3.9: Camera Obscura (a) with *pin-hole*, (b) with lens.

3.2.9 Imaging systems

The purpose of an imaging system is to give a presentation as faithful as possible of a three-dimensional object. Ideally, the image should contain three-dimensional information about the object, so that all sides can be seen, as one can with the object itself. This is extremely difficult with purely optical techniques. Holography is one of the few techniques that allow this, however it has a lot of limitations.

Most imaging systems are *projecting* systems. So the 3-dimensional object is projected onto a 2-dimensional surface, with loss of information about depth in the direction of projection. This is not a large problem, because the eye itself is a projecting image system, and the brain is especially trained to reconstruct an imaginary 3-dimensional scene from a 2-dimensional projection. This is further aided by using two eyes (and thus two slightly different projections), and by interpreting parallax-changes during movement into information about depth.

A very simple - and in a sense perfect - projecting system is the (original) *camera obscura* (figure 3.9a). A box with a small aperture in the front and a photographic film in the back. Only "*one*" ray from every point in the object space can enter the box.

One obtains a sharp image for every object, independent of the position of the image surface. The drawback of this technique is that only a small fraction of the rays contribute to the image, so the film has to be very sensitive. Therefore the small aperture is replaced by a large opening with a lens (figure 3.9b).

The purpose of the lens is to make sure that all rays from an object point are focussed onto one point of the film surface. Unfortunately this is not possible for all points in the object space, but only for points at a certain distance from the camera. For other distances the image is not sharp. Better light efficiency is thus traded for depth of focus.

To conclude this section we introduce the concepts of real and virtual images (figure 3.10). In a real imaging system the rays that diverge from a point on the object are bent by a lens into rays that converge onto a real image. Therefore, real light is present at the location of the image, and it is possible e.g. to use a photographic film to capture the image. In a virtual system the rays from a

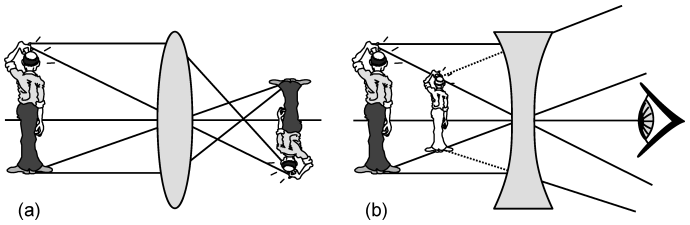

Figure 3.10: (a) Real image. (b) Virtual image.

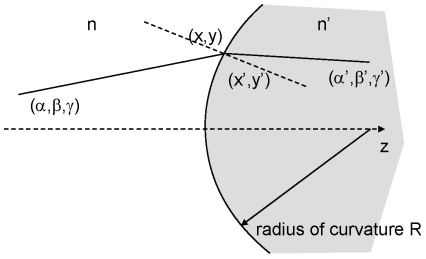

Figure 3.11: An interface between two homogeneous media in the paraxial approximation.

point on the object remain divergent after passing through the lens. One can imagine elongating the rays into the object space until they cross. However, no light is focussed onto that point. If one observes from behind the lens, the object point seems to be at the location of the virtual image point.

3.3 Paraxial theory of imaging systems

3.3.1 Introduction

The description of the ray paths is enormously simplified if we only consider rays with a small angle to the optical axis. Furthermore, we also assume that the angle between the rays and the normal to the surfaces, which the rays cross, is small. These rays are called paraxial rays. We will show for these rays that a perfect stigmatic image is formed in a system with spherical surfaces. This imaging is considered the nominal imaging of the lens system. If other rays lead to another image, then this is a shift from the nominal situation.

For paraxial rays we can approximate $\sin\theta$ by θ. For Snell's law we obtain:

$$n\theta = n'\theta'. \tag{3.17}$$

Thus, we make use of the first term in the sine series expansion. Therefore the paraxial theory is called a first order theory.

Consider the refraction of a paraxial ray on a single interface with radius R, between a medium with index n and another medium with index n' (figure 3.11). A ray with direction cosines (a, b, c) is incident on the surface at coordinates (x, y) (direction cosines are the cosines of the angles between a direction and the three coordinate axes, therefore $a = cos(\alpha), b = cos(\beta)$ and $c = cos(\gamma)$). After refraction the ray starts from (x', y') with direction cosines (a', b', c') (with corresponding angles $(\alpha', \beta', \gamma')$).

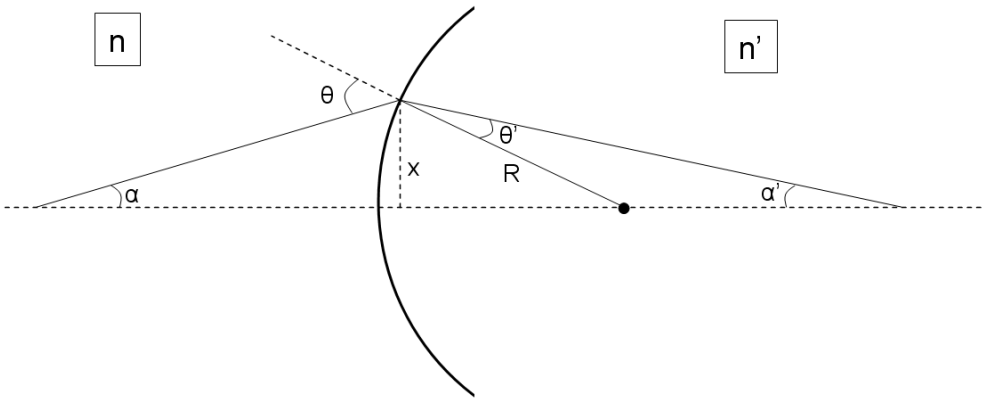

Figure 3.12: Calculation of the direction cosine α'.

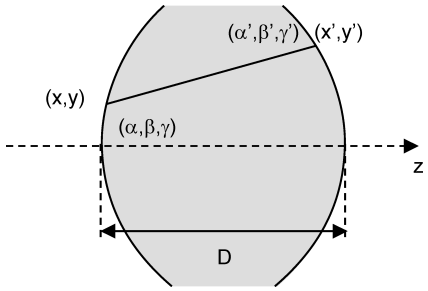

Figure 3.13: Propagation in a homogeneous medium in the paraxial approximation.

Starting from Snell's law in paraxial approximation we find (see figure 3.12) that

$$n'(arcsin\frac{x}{R} - (-\alpha')) = n(\alpha + arcsin\frac{x}{R}). \tag{3.18}$$

With the paraxial approximation, this leads to

$$\alpha' n' = \alpha n + (n - n')\frac{x}{R} \tag{3.19}$$

$$\alpha' = \frac{n}{n'}\alpha + \frac{n - n'}{n'R}x. \tag{3.20}$$

Analogously we find that

$$\beta' = \frac{n}{n'}\beta + \frac{n - n'}{n'R}y. \tag{3.21}$$

Furthermore we see that

$$\begin{aligned} x' &= x \\ y' &= y. \end{aligned} \tag{3.22}$$

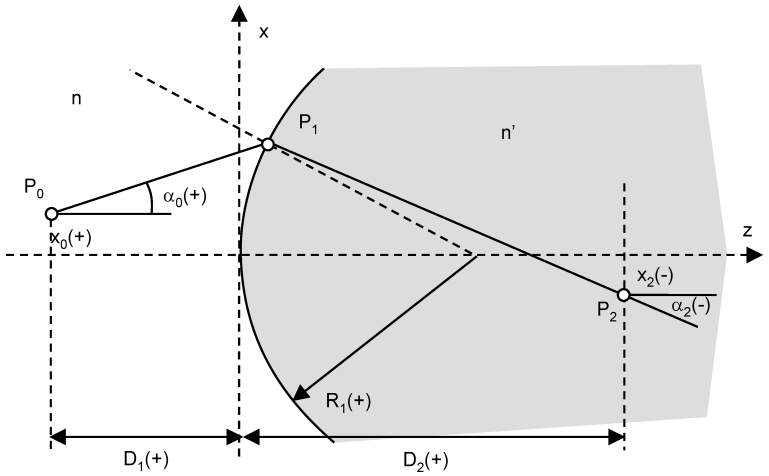

Figure 3.14: Propagation of light in the paraxial approximation between two points on both sides of an interface.

To calculate the trajectories through a lens system, we also need equations for the propagation within a medium with constant refractive index (e.g. between two interfaces). These are the translation equations (figure 3.13). Within the paraxial approximation we easily obtain:

$$\begin{aligned} \alpha' &= \alpha \\ \beta' &= \beta \\ x' &= x + D\alpha \\ y' &= y + D\beta \end{aligned} \tag{3.23}$$

with D the distance between the interfaces (measured on the z-axis). These equations are also linear and separated with respect to the (x, z) and (y, z) planes. They can be considered dual to the refractive equations. The latter contain an angle transformation, while the translation equations perform a location transformation.

Consider now the imaging of a point P_0 via one spherical interface to a point P_2 (figure 3.14). We follow a ray leaving P_0 with angle α_0 and going through P_2. This ray follows a sequence of translation, refraction and another translation.

With simple algebra we obtain the complete transformation:

$$x_2 = \left(\frac{(n-n')\,D_2}{n'R_1} + 1\right) x_0 + \left(D_1 + \frac{nD_2}{n'} + \frac{(n-n')\,D_1D_2}{n'R_1}\right) \alpha_0 \tag{3.24}$$

$$\alpha_2 = \left(\frac{n-n'}{n'R_1}\right) x_0 + \left(\frac{n}{n'} + \frac{(n-n')\,D_1}{n'R_1}\right) \alpha_0 \tag{3.25}$$

For the previous equations we did not use the fact that P_2 is the image of P_0. If this is the case, then x_2 has to be independent of α_0, all rays from P_0 have to arrive in P_2. Therefore:

$$D_1 + \frac{nD_2}{n'} + \frac{(n-n')\,D_1D_2}{n'R_1} = 0 \tag{3.26}$$

which can be written as:

$$\frac{n'}{D_2} + \frac{n}{D_1} = \frac{n' - n}{R_1}. \tag{3.27}$$

For all this we adopt a sign convention as shown in the figure. The radius of curvature of the refracting interface is positive if the center lies to the right of the interface (for light coming from the left) Therefore a positive radius means that light is incident on a convex surface. Object (resp. image) distance is positive if the object (resp. image) lies in the object (resp. image) space. The lateral distances to a point are positive if the point is above the axis, and the angles are positive if the angle of the ray to the right points upwards with respect to the optical axis. Notice that an image located in image space is called a real image, while an image in object space is a virtual image. As already mentioned, the term *virtual* stems from the fact that the rays do not converge to this image, but for an observer in image space they seem to originate from this image. Furthermore, we deduce that the lateral image magnification m_x and angular magnification m_α are given by:

$$m_x \stackrel{\Delta}{=} \frac{x_2}{x_0} = -\frac{n}{n'}\frac{D_2}{D_1} \tag{3.28}$$

$$m_\alpha \stackrel{\Delta}{=} \frac{\Delta\alpha_2}{\Delta\alpha_0} = -\frac{D_1}{D_2} \tag{3.29}$$

From the product of these expressions we obtain the important relation:

$$m_x.m_\alpha = \frac{n}{n'} \tag{3.30}$$

or

$$n'x_2\Delta\alpha_2 = nx_0\Delta\alpha_0. \tag{3.31}$$

This is the *Lagrange* or *Smith-Helmholtz* equation. It applies not only to a single interface, but also to a sequence of interfaces, and thus to a lens system. We conclude that a larger lateral magnification is obtained by reducing the angular magnification, and vice versa. For example, to image a light source on a point as small as possible, one will need a strong angular magnification. This also means that rays that depart with a large angle from the source are irretrievably lost. If object and image are both in air, it is thus impossible to image a source that radiates in all directions without power loss into an image which is smaller than the source itself! Consider now two special rays leaving the object point P_0, namely the chief ray and the marginal ray (figure 3.15). The *chief ray* is the ray that goes through the center of the optical system (for now we do not explain how this center is defined). The *marginal ray* is a ray through the outer edge of the optical system (for example the edge of a lens or a diaphragm). If θ_0 is the angle between these rays, the Lagrangian invariant is written as:

$$n'x_2\theta_2 = nx_0\theta_0. \tag{3.32}$$

For large (non-paraxial) angles one can prove that the Lagrangian invariant becomes more general:

$$n'x_2\sin\theta_2 = nx_0\sin\theta_0 \tag{3.33}$$

This is also called the *Abbe sine-relation*. This does not apply a priori to a general imaging system. But if it holds for all rays (thus not only for the marginal rays) this implies that the image is stigmatic. For the invariant quantity $(nx_{max}\sin\theta)$, with θ the angle between the marginal ray and

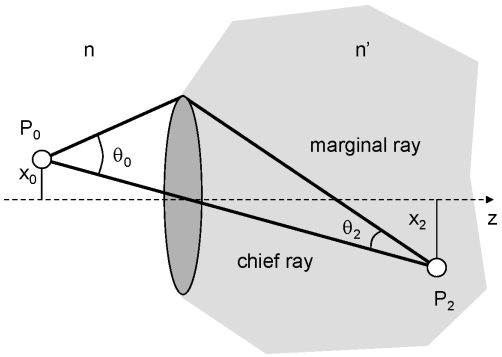

Figure 3.15: The chief ray and the marginal ray for imaging in the paraxial approximation.

the chief ray, and x_{max} the extreme lateral position of the object, there exist a host of names in the literature. Amongst others one uses *Throughput*, *Luminosity*, *Acceptance* and *étendue*. Indeed, these terms indicate that the quantity is a measure for the capacity of an optical system to image without loss of light.

3.3.2 Matrix formalism

The previously deduced ray equations are linear and contain two variables. Therefore they are easily put into a matrix form. A matrix performs a transformation (translation or refraction) from one plane to the other. The technique is elegant because multiple operations are simply presented by matrix multiplication.

We define the column matrices[1]:

$$\mathbf{r} = \begin{bmatrix} x \\ \alpha \end{bmatrix} \text{ and } \mathbf{r}' = \begin{bmatrix} x' \\ \alpha' \end{bmatrix} \tag{3.34}$$

A spherical interface

The refraction transformation at a spherical interface with radius R and between mediums n and n' is written as:

$$\mathbf{r}' = \mathbf{R}\mathbf{r}, \tag{3.35}$$

with

$$\mathbf{R} = \begin{bmatrix} 1 & 0 \\ -P & n/n' \end{bmatrix} \text{ with } P = \frac{n'-n}{n'R}. \tag{3.36}$$

P is called the refractive *power* of the interface. This power is expressed in diopters (1diopter $=$ $1\ m^{-1}$). The determinant of the matrix $\mathbf{R}$ is the ratio between the index of the start medium and

[1]There is an alternative convention for the matrix formalism of ray optics, with the column matrix $\mathbf{r}$ defined by

$$\mathbf{r} = \begin{bmatrix} x \\ n\alpha \end{bmatrix}.$$

$n\alpha$ is called the *optical direction cosine*. Both conventions have advantages and disadvantages. For this course we use the most accepted version.

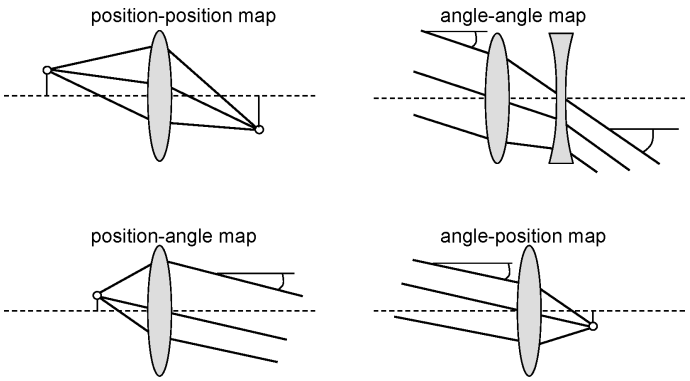

Figure 3.16: Different kinds of imaging.

the index of the end medium n/n'. The radius of curvature of a plane interface perpendicular to the optical axis is infinite, thus the matrix of the system becomes:

$$\mathbf{R}_{plane} = \begin{bmatrix} 1 & 0 \\ 0 & n/n' \end{bmatrix} \tag{3.37}$$

A translation

Analogously, in the paraxial approximation, a translation over a distance D_{12} in medium n is written as

$$\mathbf{r}' = \mathbf{T}\mathbf{r}, \tag{3.38}$$

with

$$\mathbf{T} = \begin{bmatrix} 1 & D_{12} \\ 0 & 1 \end{bmatrix} \tag{3.39}$$

The determinant of this matrix is 1, as the start and end index are the same.

Imaging

For a complete lens system one can define a system matrix $\mathbf{M}$ that describes the relation between rays departing from a certain plane and rays arriving at another plane. Thus, this matrix is the product of a number of $\mathbf{R}$ and $\mathbf{T}$ matrices. We note that the determinant of all system matrices is equal to the ratio between start and end index. If the start and end plane coincide with the object and image plane, respectively (these are called conjugate planes), then the system matrix has the following form by definition:

$$\mathbf{M} = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix} \tag{3.40}$$

Indeed, all rays from x have to arrive at x' independent of the angle α (figure 3.16a).

Other matrices with a zero element have an interesting function:

- $M_{22} = 0$: "imaging" from position to angle,

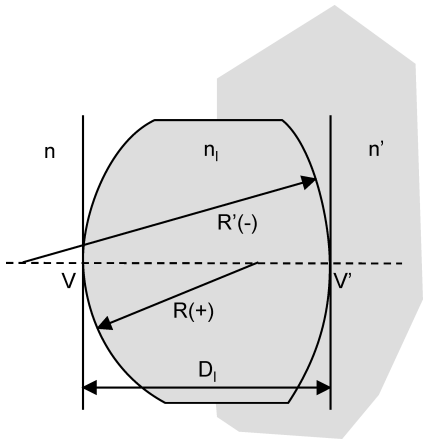

Figure 3.17: A single lens.

- $M_{21} = 0$: angle "imaging",
- $M_{11} = 0$: "imaging" from angle to position.

A single lens

Consider a single lens, as depicted in figure 3.17. The points V and V' are called the *vertices* of the lens. The two interfaces have a power P and P', respectively, given by:

$$P = \frac{n_l - n}{n_l R} \text{ and } P' = \frac{n' - n_l}{n' R'} \tag{3.41}$$

Thus, the system matrix $\mathbf{M}$, from input to output of the lens, becomes:

$$\begin{aligned} \mathbf{M} &= \mathbf{R'TR} \\ &= \begin{bmatrix} 1 & 0 \\ -P' & n_l/n' \end{bmatrix} \begin{bmatrix} 1 & D_l \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -P & n/n_l \end{bmatrix} \\ &= \begin{bmatrix} 1 - PD_l & D_l n/n_l \\ P'PD_l - Pn_l/n' - P' & n/n' - P'D_l n/n_l \end{bmatrix} \end{aligned} \tag{3.42}$$

A thin lens

In first order approximation we have $D_l = 0$ for a thin lens (figure 3.18). Thus, all refraction seems to take place in one plane. The system matrix becomes:

$$\mathbf{M_{thin}} = \begin{bmatrix} 1 & 0 \\ -P_{thin} & n/n' \end{bmatrix} \text{ with } P_{thin} = P' + Pn_l/n' \tag{3.43}$$

It has the same form as the matrix of a single interface. If we use the expressions for P' and P, we obtain the refractive power of a thin lens:

$$P_{thin} = \frac{n_l - n}{n' R} - \frac{n_l - n'}{n' R'} \tag{3.44}$$

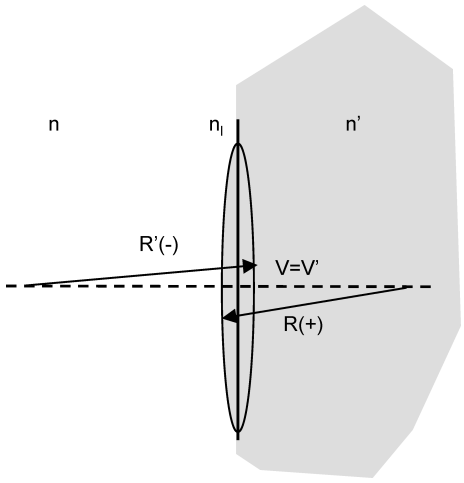

Figure 3.18: A thin lens.

By traversing the lens in the opposite direction, from medium n' to medium n, we get power P'_{thin} of the lens:

$$P'_{thin} = \frac{n_l - n'}{-nR'} - \frac{n_l - n}{-nR} \tag{3.45}$$

$$= \frac{n'}{n} P_{thin} \tag{3.46}$$

so that:

$$\frac{P_{thin}}{n} = \frac{P'_{thin}}{n'} \tag{3.47}$$

Note the minus sign in front of the curvatures of the interfaces. Because we move in the opposite direction a positive radius becomes negative, and vice versa. Therefore, the refraction in one direction does have the same sign as the refraction in the other direction.

A thin lens in air ($n' = n = 1$) has power:

$$P_{thin} = P'_{thin} = (n_l - 1)\left(\frac{1}{R} - \frac{1}{R'}\right). \tag{3.48}$$

This is the only quantity characterizing the thin lens (besides the diameter). If P_{thin} is positive, one calls it a positive lens. In the other case, it is a negative lens. Note also that if $n' = n$, nothing changes to the properties of the lens upon reversal. And this holds even if the lens has an asymmetrical form.

The *focal length* is determined by imposing that all rays with incidence angle $\alpha = 0$ converge to a point F' a length f' behind the lens:

$$\left.\begin{array}{l} \alpha' = \alpha - P_{thin} x = -P_{thin} x \\ \alpha' = -x'/f' \end{array}\right\} \Rightarrow f' = \frac{1}{P_{thin}} \tag{3.49}$$

An analogous result is obtained if we assume that all rays with $\alpha' = 0$ originate from the same point F a length f before the lens.

$$f = \frac{1}{P'_{thin}} \tag{3.50}$$

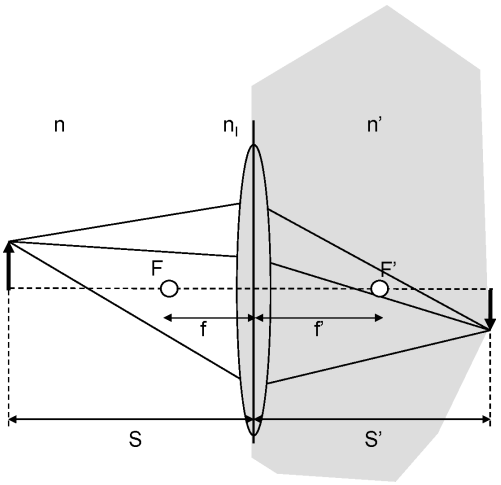

Figure 3.19: A position to position imaging with a thin lens.

so that

$$\frac{f}{n} = \frac{f'}{n'}. \tag{3.51}$$

Consider now in general the relationship between object and image distance (figure 3.19). We use a translation to the left and the right of the thin lens:

$$\mathbf{M}' = \mathbf{T}'\mathbf{M_{thin}}\mathbf{T}, \tag{3.52}$$

with

$$\mathbf{T} = \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{T}' = \begin{bmatrix} 1 & S' \\ 0 & 1 \end{bmatrix}. \tag{3.53}$$

The new M'_{12} element has to be zero, as we study imaging. Thus:

$$S + S'\frac{n}{n'} - P_{thin}SS' = 0, \tag{3.54}$$

so that $S = S' = 0$ or

$$\begin{aligned} \frac{n}{S} + \frac{n'}{S'} &= n'P_{thin} = \frac{n'}{f'} \\ &= nP'_{thin} = \frac{n}{f} \end{aligned} \tag{3.55}$$

This last expression is the well-known formula for a thin lens. Notice that for a thin lens for every location of the object plane it is possible to find a conjugate image plane. A special case occurs if the object plane coincides with the incidence plane of the lens ($S = 0$). The image plane is then the exit plane of the lens (that coincides with the incidence plane) and the magnification is 1.

A complex lens system

Consider again a more complex system as shown in figure 3.20, a thick lens or a lens system. Now the system matrix from plane V to plane V' (e.g. the vertices of the front and back lens resp.) has

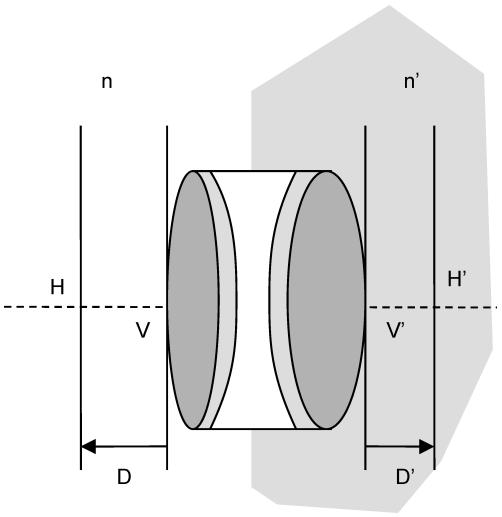

Figure 3.20: A complex system that can be treated as a thin lens with principal planes H and H'.

the following general form:

$$\mathbf{M} = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \text{ with } \det \mathbf{M} = n/n'. \tag{3.56}$$

Now we try to determine if it is possible to transform this matrix into the form for a thin lens, using only translations in front of and behind the system. Thus, we are looking for new reference planes (at the points H and H') with this property. These planes are called the *principal planes* of the system.

The new system matrix becomes:

$$\mathbf{M}' = \mathbf{T}'\mathbf{M}\mathbf{T} \tag{3.57}$$

with

$$\mathbf{T} = \begin{bmatrix} 1 & D \\ 0 & 1 \end{bmatrix} \text{ and } \mathbf{T}' = \begin{bmatrix} 1 & D' \\ 0 & 1 \end{bmatrix}, \tag{3.58}$$

or:

$$\mathbf{M}' = \begin{bmatrix} M_{11} + M_{21}D' & M_{22}D' + M_{21}DD' + M_{12} + M_{11}D \\ M_{21} & M_{22} + M_{21}D \end{bmatrix} \tag{3.59}$$

Because this matrix needs to have the form

$$\mathbf{M}' = \begin{bmatrix} 1 & 0 \\ M_{21} & n/n' \end{bmatrix} \tag{3.60}$$

we have three equations with only two unknowns D and D'. From M'_{11} and M'_{22} we obtain immediately:

$$\begin{aligned} D &= \left(n/n' - M_{22}\right)/M_{21} \\ D' &= \left(1 - M_{11}\right)/M_{21}. \end{aligned} \tag{3.61}$$

It is easy to prove that M'_{12} is 0, using $\det(\mathbf{M}) = n/n'$. Sometimes one obtains D and D' values so that H and H' lie on the inside of the lens. Moreover, it is possible that the incidence principal plane lies to the right of the exit principal plane.

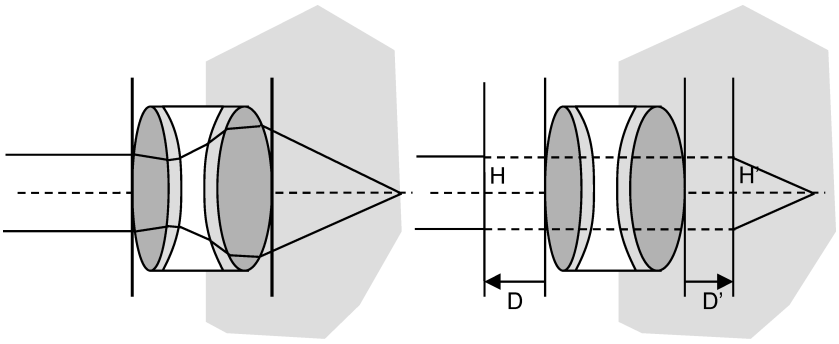

Figure 3.21: The imaging by a complex lens system is equivalent to a thin lens with two principal planes.

The M_{21} element remains invariant under the double translation and is the power of the entire system: $M_{21} = M'_{21} = -P_{syst}$. Finally it is important to remark that the lateral magnification for imaging from front to back principal plane is one, as $M'_{11} = 1$. Within the paraxial approximation a general lens system is characterized by the power and the location of the principal planes. The description based on principal planes is very elegant, as we can apply the simple equations for thin lenses to complex optical systems (in particular the expressions for f and f', and for the distance between object and image plane). We only need to realize that all lengths in the object and image space are referenced to the H and H' principal plane respectively, while for a real thin lens these planes coincide with each other and with the lens (figure 3.21). In practice, if one needs to choose or specify a lens, it is important to pay attention to the reference used for the lengths, especially for the focal length (relative to a principal plane or to a vertex). In some cases the distance between a vertex and a principal plane can be relatively large.

3.3.3 Spherical mirrors

A spherical mirror is an alternative to a lens (figure 3.22). For such a reflecting system we can again deduce a paraxial system matrix (where the previous sign convention has to be expanded for both propagation directions and for the radius of curvature of the mirror). For reflection at the mirror surface one obtains (within the paraxial approximation):

$$\begin{pmatrix} x' \\ \alpha' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -P & 1 \end{pmatrix} \begin{pmatrix} x \\ \alpha \end{pmatrix}, \tag{3.62}$$

with

$$P = \frac{2}{R} \tag{3.63}$$

By reintroducing the focal length f (as $P = 1/f$), we obtain:

$$f = \frac{R}{2}. \tag{3.64}$$

Thus, parallel rays are focussed halfway between the center of the sphere and the mirror. From the previous it (obviously) appears that the index n has no influence on the ray trajectory upon reflection on the spherical surface.

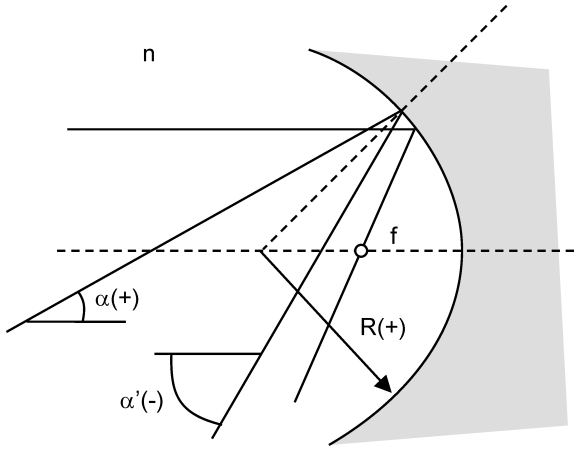

Figure 3.22: The spherical mirror

Notice that this behavior of the spherical mirror only holds within the paraxial ray optics approximation. In fact, the spherical mirror is a paraxial approximation of the parabolic mirror, discussed briefly in section 3.2.4.

3.3.4 The graphical formalism

With the definition of principal planes one does not have to depict a lens or lens system exactly with its refractive surfaces, but only with the principal planes. Everything happening between the planes is not shown, as if every refraction takes place on the positions of the principal planes. To construct the image of a point in the object plane we only have to obey the following rules:

- a ray parallel with the axis, incident on the first principal plane, leaves the second principal plane at the same height and in the direction of the focal point F'.
- a ray through the principal point H leaves the second principal plane from H' with an angle equal to the incidence angle (apart from a factor n/n'). This ray is a *chief ray*.
- a ray through the focal point F and incident on the first principal plane, leaves the second principal plane at the same height and parallel to the axis.

This is illustrated in figure 3.23.

To make these drawings, it is in principle necessary to know the location of the principal points. Of course, these can be calculated using the methods of the previous section. However, it is useful to know these locations approximately for a number of common lens types. Figure 3.24 shows some examples.

For symmetrical lenses (convex or concave) the principal points H and H' divide the distance between the vertices V and V' approximately in three equal parts. For plano-convex or plano-concave lenses one principal point is located on the curved vertex, whereas the other is at about one third of $|VV'|$ from the curved vertex. Finally, for meniscus (or convex-concave) lenses one principal point will always lie outside of the lens.

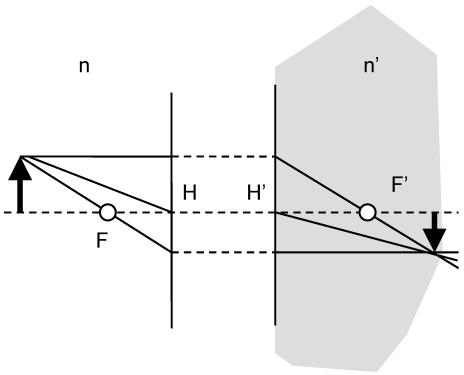

Figure 3.23: The graphical formalism. For some rays (the chief ray, rays parallel to the optical axis and rays through the focal point) the trajectories are easily drawn.

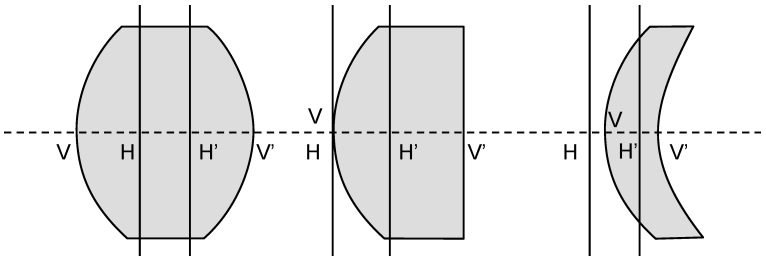

Figure 3.24: Location of principal planes for common lens types. From left to right: a double-concave lens, a plano-concave lens, a meniscus lens

To conclude this section we define some useful concepts. The f-number, or relative aperture, of the lens system is defined as:

$$f - \text{number} = \frac{f}{D} \tag{3.65}$$

with f the focal length and D the diameter of the lens (or the diaphragm in front of it: figure 3.25). An f-number of e.g. 4 is denoted as $f/4$. Common values in photography are 2, 2.8, 4, 5.6, 8, 11, 16 and 22. Large values indicate small diaphragms.

A quantity related to the f-number is the *numerical aperture* of the system. The numerical aperture (NA) is the sine of the angle between the marginal ray through the focal point and the optical axis. One obtains (for small angles):

$$NA = \frac{1}{2\,(f - \text{number})}. \tag{3.66}$$

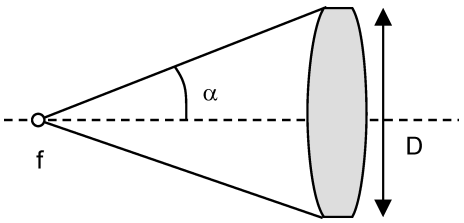

Figure 3.25: The lens parameters that determine the f number and the numerical aperture.

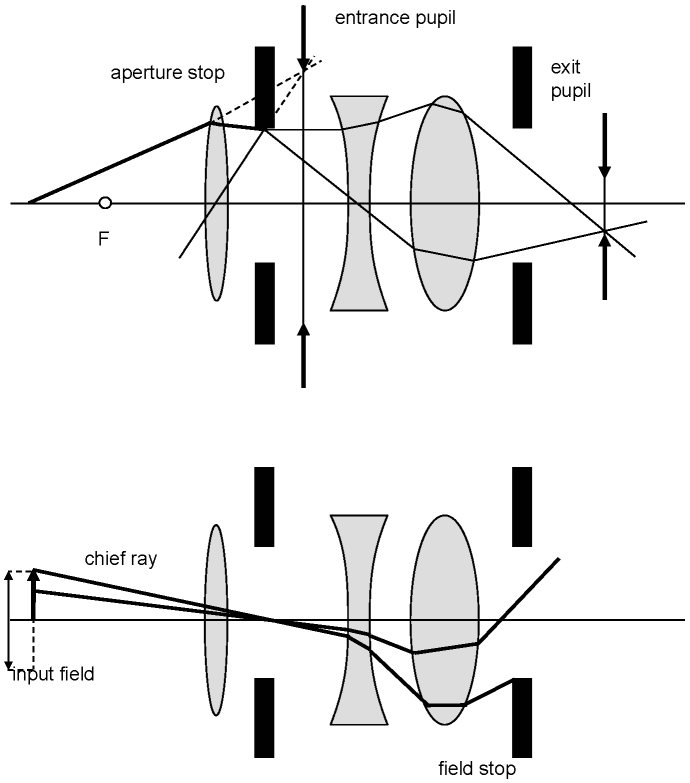

Figure 3.26: Illustration of (a) aperture stop and (b) field stop.

Thus, a large numerical aperture corresponds to a small f-number, and vice versa.

For complex lens systems D is not necessarily the diameter of the first lens or diaphragm. It is possible that the marginal ray through an object point on the axis is not determined by the first lens surface or diaphragm, but by a lens or diaphragm somewhere in the middle of the system. This limiting element is called the *aperture stop* (see figure 3.26). The image of this element by the part of the lens system to the left or the right of it is called the entrance or exit pupil, respectively (if the aperture stop is completely on the left or the right of the system, then it coincides with the entrance or exit pupil). The entrance pupil determines the cone of rays that leave the object point on the axis. Analogously, the exit pupil determines the cone of rays that arrive at the image point on the axis. Note that entrance and exit pupil may be real or virtual images of the aperture stop. In practice one can determine the aperture stop by imaging all elements of the system to the left. In this way one obtains a number of real or virtual images. The image that, seen from the object, forms the smallest cone corresponds to the aperture stop. In the same way one can find the aperture stop by imaging to the right. This has to lead to the same result. In a first approximation this calculation can be done paraxially.

For object points away from the axis, not all rays through the entrance pupil will reach their respective image point (figure 3.26b). The *number* of rays that reach the image point, decreases as the object point moves away from the axis. In this regard one defines a *field stop*. It is the lens or diaphragm of the system that first blocks chief rays from the object plane (the chief ray has a slightly different definition than in the previous paraxial approximation; it is the ray from an object point through the middle of the aperture stop). With this field stop corresponds a circular

area in the object plane (*field of view*) for which the chief rays just passes through the system. In the object plane one finds an accompanying circular area that obtains an image with reasonable intensity. The field stop does not necessarily coincide with the aperture stop. The image of the field stop in object and image space is called the entrance and exit window, respectively. Together, the aperture and field stop control the *étendue* of the optical system.

3.4 Aberrations in imaging systems

3.4.1 Introduction

When rays can no longer be considered paraxial, which is often the case for marginal rays, the imaging will differ from the paraxial imaging. This results in *aberrations*. Aberrations are deviations from perfect (stigmatic and distortion-free) imaging. It is easy to understand that spherical surfaces, either refracting or reflecting, will lead to aberrations. Consider for example the case of a curved reflector. To transform a beam of rays coming from the focal point into a parallel beam (thereby imaging the source at infinity) the reflector should have a parabolic shape. It is clear that a spherical mirror will not do this collimation in a perfect way and hence aberrations will arise. For paraxial rays only the central part of the mirror is used and hence there is little difference between a parabolic mirror and a spherical mirror with the same central radius of curvature.

Paraxial theory originates from a first order approximation of the sine function. Classically, the first study of aberration was thus performed by including a third order term in the sine series expansion. In this way, one analyzes third-order or *Seidel* aberrations. Seidel developed a formalism to describe the aberrations without explicitly calculating the ray trajectories through the system. He divided the aberrations in different categories. For monochromatic light there are aberrations that result in an image that is no longer stigmatic, such as *spherical aberrations*, *astigmatism* and *coma*. On the other hand, there are those that allow a stigmatic image, but still lead to deformation, such as *field curvature* and *distortion*. For polychromatic light there are also chromatic aberrations, created by *dispersion* of the lens material.

The next step is to include higher order terms in the series expansion, i.e. the fifth term, the seventh term etc. Although this can be relevant for systems with stringent demands, it is complicated, because it is no longer possible to subdivide the aberrations and to calculate them easily. In the following we largely refrain from the analytical calculations, and instead focus on the general characteristics of the different types of aberrations. It is significant to note that the importance of the various aberrations not only depends on the system itself, but also on the use of the system. This mainly depends on the *ratio of the image distance to the object distance* (or *conjugate ratio*), which is also the lateral magnification. A lens system that performs well (which means is free of aberrations) for a lateral magnification of 1, does not necessarily perform equally well for a very large (or very small) lateral magnification.

3.4.2 Spherical aberration

Spherical aberration relates to the imaging on the optical axis itself. Rays at a large angle with the axis will focus on a different location of the axis than paraxial rays (figure 3.27). The deviation is called the *longitudinal spherical aberration* (LSA), if measured along the axis, or *transversal spherical*

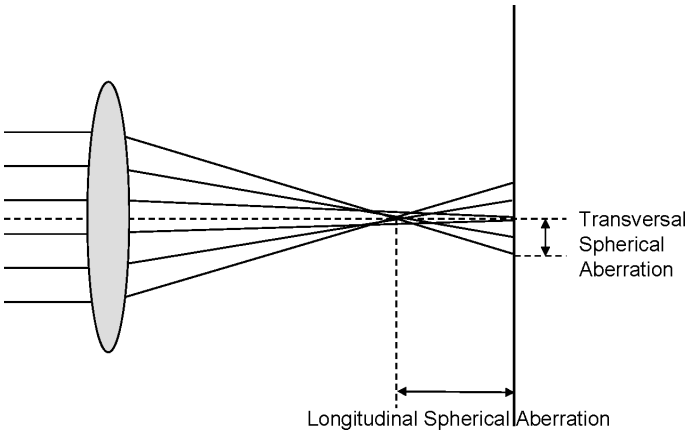

Figure 3.27: Spherical aberration.

aberration (TSA), if measured in the focal plane. They increase with the square and cube, respectively, of the lens aperture. Therefore, lenses with small f-numbers suffer most from spherical aberration.

There are three possible techniques to counter spherical aberrations, depending on the specifications and available resources. The first one is to use an ordinary spherical lens with a *best shape*. This means that one optimizes the two radii of curvature R_1 and R_2, at a given refractive power. In this context one defines a shape factor q:

$$q = \frac{R_2 + R_1}{R_2 - R_1} \tag{3.67}$$

With variation of q (at equal refractive power) one changes continuously from a symmetrical lens ($q = 0$), via a plano-convex lens ($q = \pm 1$), to a meniscus lens. For systems with a 1:1 magnification ($s = s' = 2f$) the optimal (but not perfect) shape is the symmetrical biconvex lens. In situations with infinite magnification (or reduction), as in the focusing of a parallel laser beam or collimation of light from a point source, the optimal q factor is in the neighborhood of ± 1. Here, the convex side of the plano-convex lens has to be on the side of the parallel beam. This is illustrated in figure 3.28.

The second technique involves using a combination of different lenses (figure 3.29). In this way one can get much better results than with a single lens (*singlet*). We discuss some common *doublets*. For applications with infinite magnification one often uses achromatic doublets. They consist of a positive lens glued to a negative meniscus having another refractive index. The spherical aberration of the negative lens counteracts the one from the positive lens, so that compensation takes place. For an achromatic doublet with positive refraction the index of the positive lens is smaller than the index of the negative lens. Again the parallel beam has to be incident on the most convex side of the doublet. If the materials are chosen correctly, one can obtain that the chromatic aberration is minimal, hence the name *achromat*. For 1:1 applications the symmetrical biconvex lens can be replaced by two identical plano-convex lenses, with their convex sides towards each other (figure 3.30). Of course, it is even better to use two achromatic doublets.

Finally, the third technique consists of using a single lens, but with an aspherical surface (figure 3.31). In principle spherical aberration can be perfectly eliminated in this way. Unfortunately it is technologically very difficult to produce an aspherical surface with good quality. Indeed,

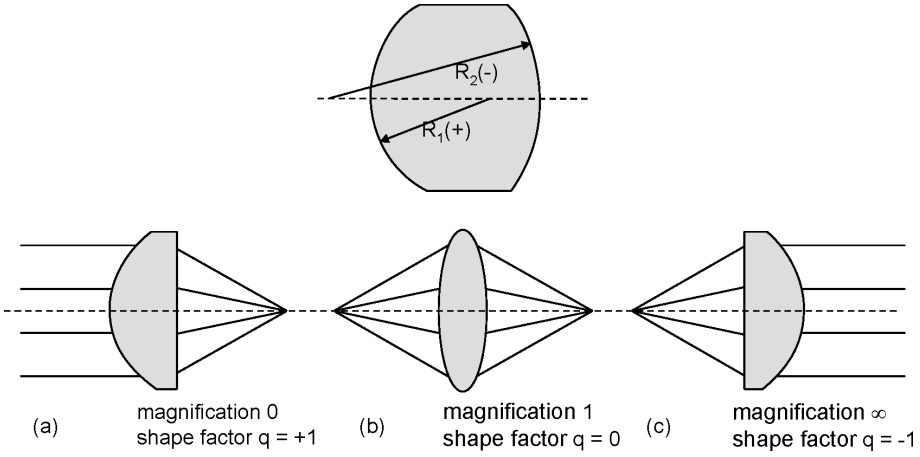

Figure 3.28: Shape factor q for various lenses.

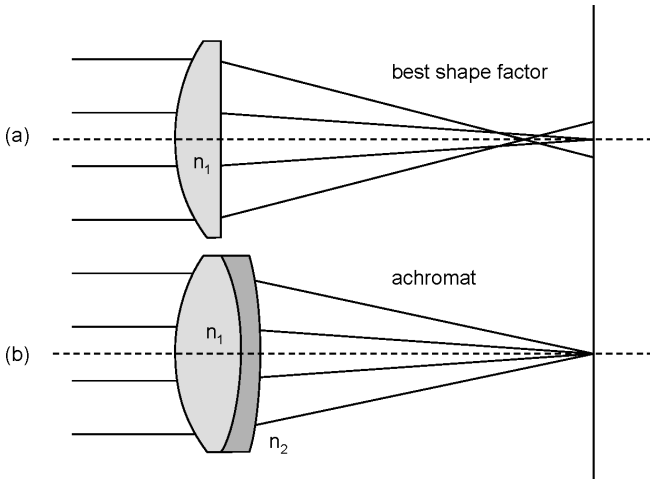

Figure 3.29: Correction of spherical aberration with a lens combination for inf:1 imaging. (a) Plano-convex singlet with aberration. (b) Achromatic doublet with much less aberration.

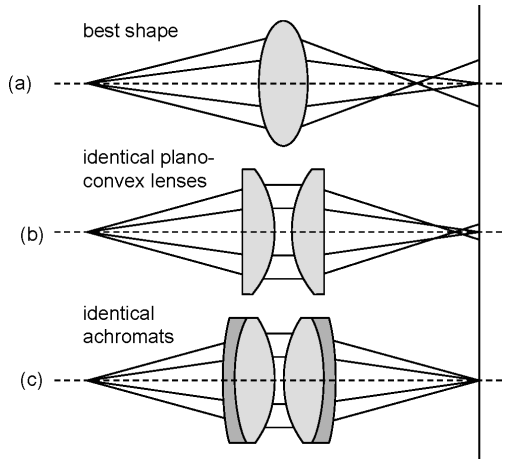

Figure 3.30: Optimization of spherical aberration for 1:1 imaging (a) with a single spherical lens with optimized shape, (b) with a pair of plano-convex lenses, (c) with a pair of identical achromats.

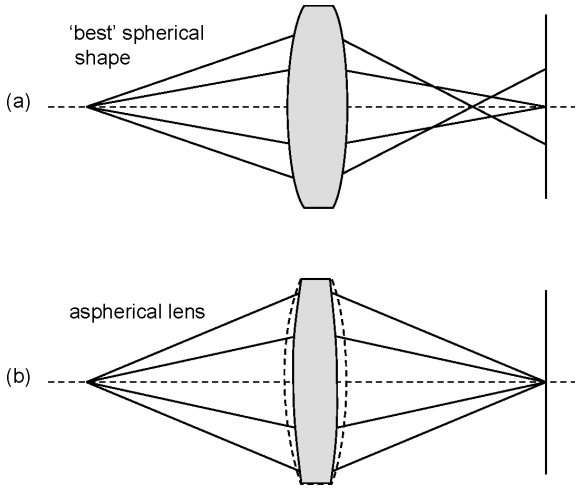

Figure 3.31: Correction of spherical aberration with an aspherical lens with optimized shape. (a) Optimal spherical lens, (b) optimal aspherical lens.

aspherical lenses are poured in a mold and not polished as with spherical lenses. However, for certain applications the aspherical lens still has the best price-quality value.

3.4.3 Astigmatism

Previously we mentioned that in non-paraxial circumstances the *non-meridional rays* (or *skew rays*) do not necessarily behave as meridional rays. Consider an object point not located on the axis. The plane through this point and the optical axis is the *meridional* plane (or *tangential* plane). The perpendicular plane that contains both the object point and the image point is the *sagittal* plane (or *radial* plane). Astigmatism means that rays in the sagittal plane focus closer or further than those in the meridional plane (figure 3.32). In that case one never achieves a sharp focus point. As one moves the image plane one obtains a horizontal focus line, followed by a fuzzy phase, and next a vertical focus line.

For lenses that are rotationally invariant due to symmetry, astigmatism only occurs for object points not located on the optical axis. However if the lens is not perfectly rotationally invariant, astigmatism will also occur for axial object points. Astigmatism is a common deviation of eye lenses and has to be corrected by glasses that are not rotationally invariant as well.

3.4.4 Coma

Even if the system is perfectly corrected for spherical aberration and astigmatism, it is still possible to have a blurred image. This can happen because of *coma*. It relates to object points that are distanced from the optical axis, like astigmatism. Rays through the edge of the optical system have a different lateral magnification than those close to the axis (figure 3.33). Furthermore, meridional rays obtain a different magnification than sagittal rays. It appears that every concentric ring of the system gives rise to a circle in the image plane. The center of this ring moves and the diameter increases as the concentric ring is magnified, leading to a comet-like image. Hence the name coma.

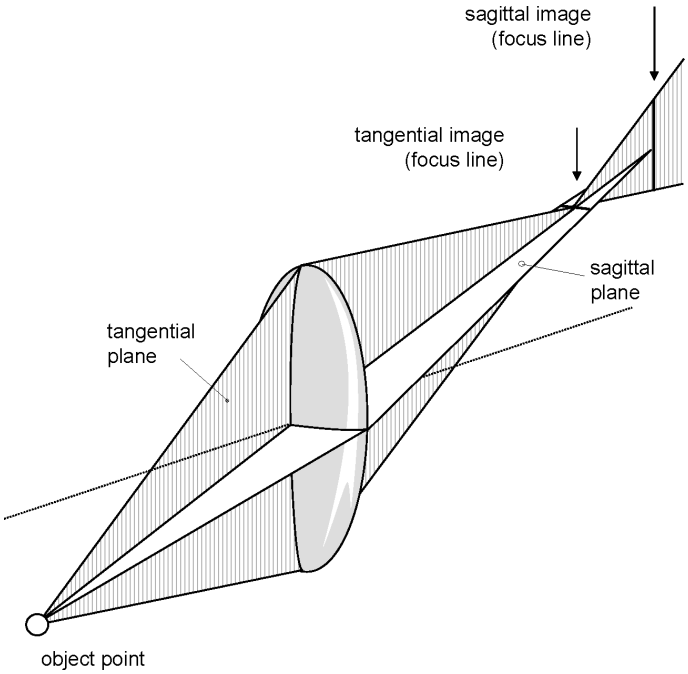

Figure 3.32: Astigmatism.

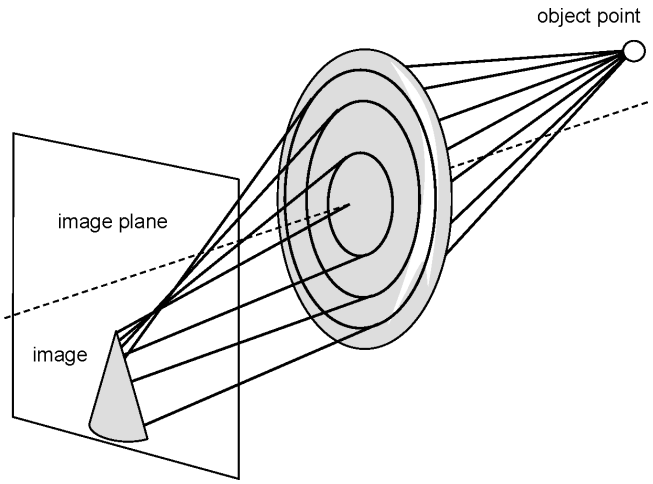

Figure 3.33: Coma.

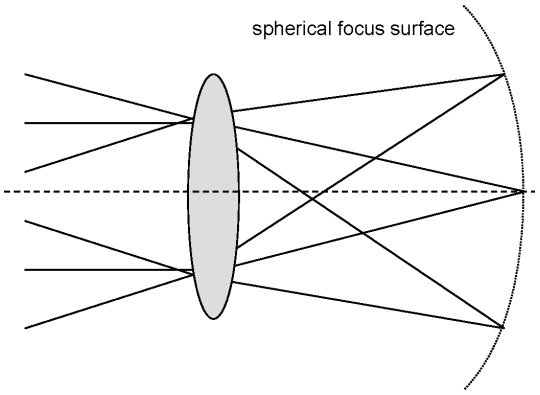

Figure 3.34: Field curvature.

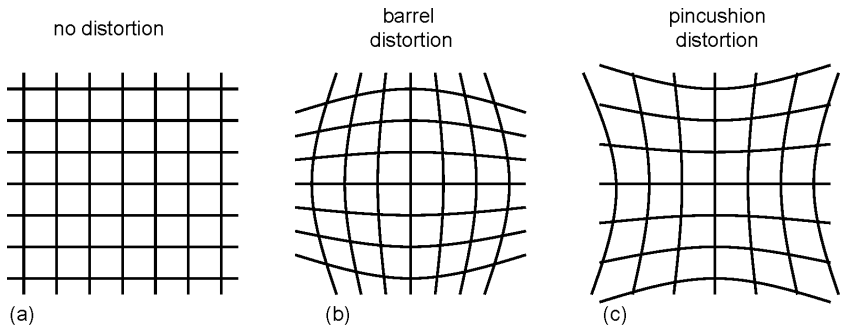

Figure 3.35: Distortion. (a) No distortion, (b) barrel distortion, (c) pincushion distortion.

3.4.5 Field curvature

A stigmatic system (corrected for spherical aberration, astigmatism and coma) will generally image in a different way than paraxial imaging. The image points are at a different location than predicted by the paraxial theory. The deviation in the longitudinal direction is called field curvature (figure 3.34). Indeed, one notices that most systems tend to image a plane object onto a curved surface, which is called the *Petzval* surface.

3.4.6 Distortion

In addition there is also a deviation in the lateral direction, which means a variation of the lateral magnification over the image. This leads to distortion in the image (figure 3.35). Most often one encounters pincushion or barrel distortion. A symmetric system with 1:1 magnification has no distortion. Furthermore, one can understand that a system with pincushion distortion will display barrel distortion upon reversal of the rays (and vice versa).

3.4.7 Chromatic aberration

Because the refractive index of materials depends on the wavelength (material dispersion), the refractive power will also depend on it (figure 3.36). For most materials (and in particular for

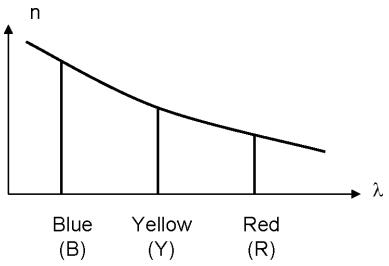

Figure 3.36: Wavelength dependence of the refractive index n.

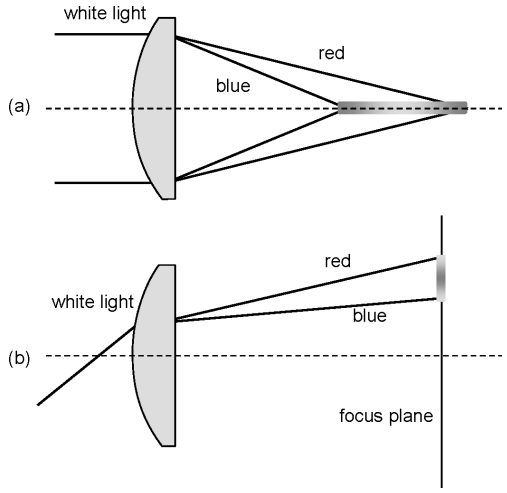

Figure 3.37: Chromatic aberration. (a) Dependance of the focus point; (b) dependance of the lateral magnification.

glass) the index decreases as wavelength increases. Thus, a lens system in air will show a stronger refraction at shorter wavelengths.

Chromatic aberration appears in two ways. For object and image points on the axis the focus point depends on the wavelength (figure 3.37a). Restricting ourselves to visible colors, blue light will focus closer to the lens than red light. On the other hand, the lateral magnification for points not on the axis differs for red and blue (figure 3.37b).

As positive and negative lenses have an opposite chromatic aberration, this allows to compensate for the effect. Indeed, this happens in achromatic doublets, as previously mentioned.

3.4.8 Aberrations in function of aperture and object size

It is clear that aberrations increase if the rays are less paraxial. This implies that they grow as the lens has a larger diameter D (so that it becomes brighter), and also when the object itself becomes larger. In this regard one defines the angle θ (*field angle*) with which the system sees the object. Table 3.1 indicates the power of D resp. θ of increase for the various aberrations. For example: lateral spherical aberration scales as D^3.

Aberration	Aperture D	Angle θ
Lateral spherical	3	0
Longitudinal spherical	2	0
Coma	2	1
Astigmatism	0	2
Field curvature	0	2
Distortion	0	3
Chromatic	0	0

Table 3.1: Power of aperture D and angle θ for various aberrations.

3.4.9 Vignetting

Often one will find *diaphragms* (or *stops*) on one or multiple locations in an optical system. They are very useful, on the one hand to stop scattered light, on the other hand to decrease the aberrations. In addition, every lens functions as a diaphragm because of its finite size.

However, diaphragms and lenses lead to the effect that some rays (especially from the outer object points) do not pass through the system. This decreases the light intensity of the corresponding image points. The phenomenon is called *vignetting*. Although not a real aberration, it corresponds to a deviation between object and image, with respect to intensity instead of sharpness. In practice one will often compromise between image sharpness and vignetting.

The example in figure 3.38 depicts a 1:1 symmetric lens system. It is clear that some rays do not reach the second lens surface. In this case it is easily remedied by putting an extra lens in the middle (*field lens*). This lens is located in an internal image plane, therefore it has no influence on the paraxial imaging, but drastically improves vignetting.

3.4.10 Depth of field

For a given image plane a system shows a sharp image for only one object plane. If the object is before or behind this object plane the image in the given image plane is unsharp. The depth of field determines the distance through which one may move the image plane to view a given object with acceptable sharpness. From figure 3.39 one notices that, for a given focal length, the depth of field is worse for a lens with larger aperture. Again there is a difficult compromise (!): a larger aperture leads to more light in the image, but to a smaller depth of field (and in general more aberrations). One obtains an infinite depth of field by employing a small hole in a screen: all objects are imaged sharply because one object point corresponds to only one ray through the system. Unfortunately the image will be dark.

3.5 Materials

An optical material is characterized first by its refractive index and absorption, both as a function of wavelength. In addition, a number of other attributes is important, such as hardness, uniformness, thermal expansion coefficient, chemical resistance etc. Glass is by far the most used lens material. The index of most common kinds of glass lies between 1.4 and 1.9. These indices are

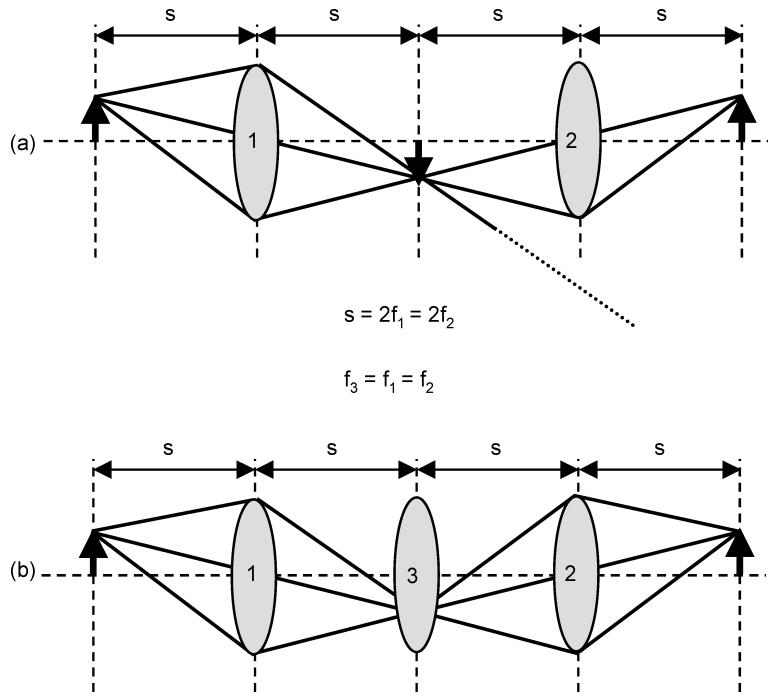

Figure 3.38: Vignetting illustrated by ray trajectories. (a) Some rays do not pass the system. (b) With an additional lens they do pass.

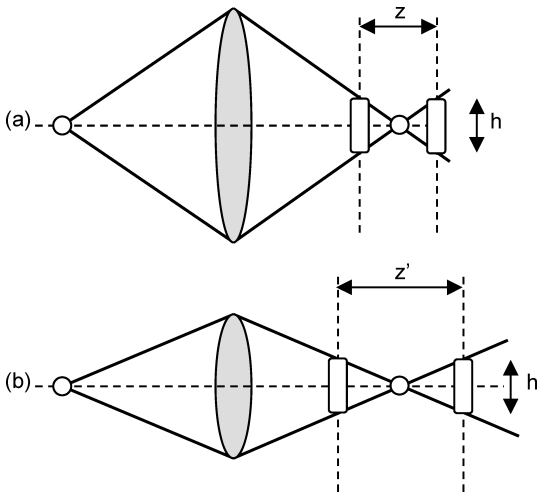

Figure 3.39: Depth of field. For a larger aperture (a) one obtains a smaller depth of field than for a smaller aperture (b).

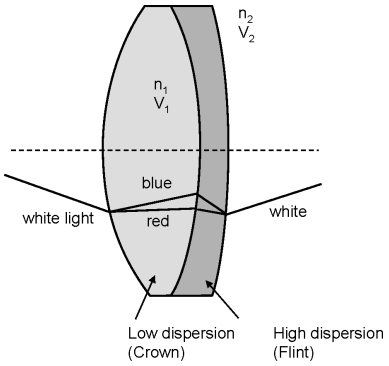

Figure 3.40: An achromat.

high enough to obtain a sufficient refractive power with respect to air, while they are low enough to control reflection losses, even without anti-reflection coating.

3.5.1 Dispersion

The wavelength dependence of the index (*dispersion*) is often described by various analytical formulas, for example:

$$n^2 = A_0 + A_1\lambda^2 + A_2\lambda^{-2} + A_3\lambda^{-4} + A_4\lambda^{-6} + A_5\lambda^8. \tag{3.68}$$

If the wavelength is not near an absorption band of the material, the index decreases monotonically with increasing wavelength. To simplify matters the dispersion is often described by one number, the Abbe constant or V-value, defined as:

$$V = \frac{n_Y - 1}{n_B - n_R} = \frac{P_Y}{P_B - P_R} \tag{3.69}$$

Here Y refers to *Yellow*, B to *Blue* and R to *Red*. In this respect the standard wavelengths are: $Y = 587.6nm$ (helium line), $B = 486.1nm$ and $R = 656.3nm$ (both hydrogen lines). A smaller V-value indicates a more dispersive material. Roughly speaking glass is divided into two categories with respect to dispersion. Low dispersion glass is called *crown* glass, whereas high dispersion glass is called *flint* glass. The division is made at a V-value of about 50. Often crown glass has a relatively low index ($n < 1.55$), while flint glass has a high index ($n > 1.6$). However, this is not a general rule.

One can easily prove that a combination of two thin lenses against each other can only be achromatic if the dispersion of the two kinds of glasses is different (figure 3.40). If one demands that the refractive powers for wavelengths R and B are the same, one obtains:

$$P_B = P_{B1} + P_{B2} = P_R = P_{R1} + P_{R2} \tag{3.70}$$

The indices 1 and 2 refer to (thin) lens 1 resp. 2. Furthermore:

$$(P_{B1} - P_{R1}) + (P_{B2} - P_{R2}) = 0. \tag{3.71}$$

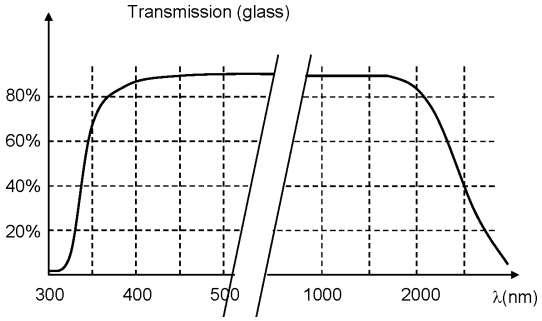

Figure 3.41: Transmission of glass as a function of wavelength.

This is equivalent to:

$$\frac{P_{Y1}}{V_1} + \frac{P_{Y2}}{V_2} = 0. \tag{3.72}$$

We also know that:

$$P_{Y1} + P_{Y2} = P_Y, \tag{3.73}$$

with P_Y the refractive power of the combination at wavelength Y. Both equations can be satisfied only if the V-values of both materials differ, and if the refractive powers have a different sign. Solving the system for P_{Y1} and P_{Y2} one gets:

$$\begin{aligned} P_{Y1} &= P_Y \frac{V_1}{V_1 - V_2} \\ P_{Y2} &= -P_Y \frac{V_2}{V_1 - V_2}. \end{aligned} \tag{3.74}$$

This means that a positive achromatic doublet has to consist of a positive lens with low dispersion (usually crown) and a negative lens with high dispersion (usually flint).

It is clear that the achromatic doublet is not yet completely free of chromatic aberration. As it is corrected only for two distant wavelengths (B and R). Sometimes one corrects for three wavelengths (B, Y and R). This is called an *apochromatic* system. One typically needs a triplet for this.

3.5.2 Absorption

Good quality glass has a low absorption in the entire visual range ($400 - 700nm$). In the UV-range the absorption quickly increases however. At $300nm$ absorption is often unacceptably strong. Also in the IR-range the absorption grows from about 2 to $3\mu m$. Figure 3.41 shows a typical transmission characteristic.

To work in the deep UV or IR *synthetic quartz* (*synthetic fused silica*) is often used. This is amorphous SiO_2. With this material one typically works until $200nm$ and $3.5\mu m$ respectively (although some absorption peaks show up in the IR). In addition, quartz has a lower expansion coefficient and it is thermally more stable and harder. The refractive index is about 1.46 (at Y) and the V-value is approximately 65. If quartz is too expensive for a certain application, but one works in thermally difficult circumstances, sometimes pyrex-glass is used. This also has a low thermal expansion

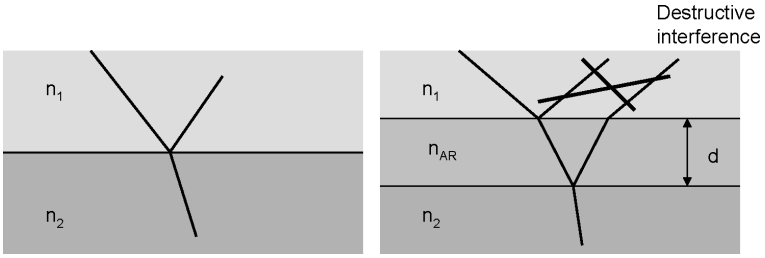

Figure 3.42: Reflection at an interface. (a) Without anti-reflection coating. (b) With anti-reflection coating.

coefficient. However, the optical quality (e.g. uniformness of the index) is less than for normal optical glass. The index typically measures 1.48.

In some cases one uses *sapphire*, which is crystalline Al_2O_3. The properties are comparable to those of quartz, but it is harder, stronger and especially chemically inert (very hard, small expansion). Moreover, transmission is very good from $200nm$ to $5\mu m$. The index is about 1.76. For special applications one will use mono- or polycrystalline semiconductors. Pure *silicon* e.g. has a good transmission from about $1\mu m$ until $7\mu m$. Germanium has a good transmission for even longer wavelengths and is used in optics for CO_2 high-power lasers, at a wavelength of $10.6\mu m$. Both silicon and germanium have a high refractive index ($n > 3$). Another semiconductor is *zinc selenide*, which is one of the few materials that has a good transmission for visible wavelengths (larger than $600nm$) and the far infrared, at the same time. This is very important for some applications. The refractive index of this material about 2.5.

3.5.3 Reflection at an interface

Although all high-index materials instigate reflection losses, the use of anti-reflection coatings can be very efficient (figure 3.42). The simplest AR coating between air and an element with index n consists of a single quarter wavelength layer with index equal to the square root of n. In practice the available materials are limited. For example, for glass with $n = 1.5$ a material with $n = 1.225$ would be needed. Often, the best choice for the coating is *magnesium fluoride* with an index of about 1.38. For materials with higher index it is easier to find the right coating material.

3.6 Applications

There are many different imaging systems, such as the eye and glasses, the magnifying glass and the microscope, binoculars and the telescope, the camera, copiers, optical scanners (read and write), projectors etc. From a paraxial imaging viewpoint these devices distinguish themselves only by the magnification and by the real or virtual character of the image. In practice there are many differentiating factors. Depending on the application one or more of the following specifics will play a role in the design:

- constant or variable magnification
- field of view

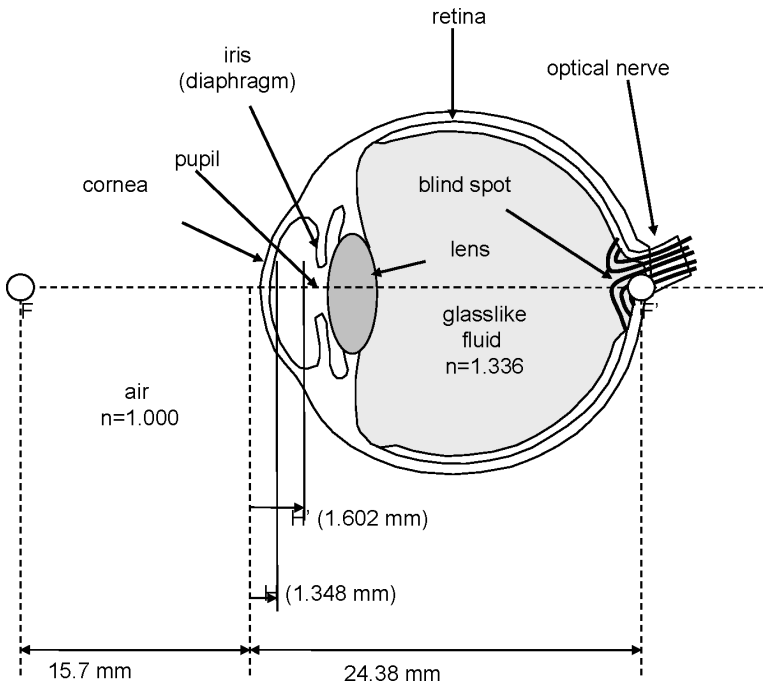

Figure 3.43: The eye.

- brightness
- monochromatic aberrations
- chromatic aberrations
- size and shape of the system
- geometric performance sensitivity (ease of alignment, thermal expansion...)

Here we succinctly describe the operation principles of some common imaging systems.

3.6.1 The eye

The refraction in the eye (see figure 3.43) is caused by the curved cornea interface (from $n = 1$ to $n = 1.34$) on the one hand, and the crystalline lens (from $n = 1.37$ to $n = 1.42$) on the other hand. The refractive power of the combination is about 58 diopters. For young people the adjustable character of the lens can increase the power with about 10 diopters. This adaptive power decreases with age. The *field of view* of the eye is very large, but because of the structure of the retina there is a high resolution only for a small area around the optical axis. The image on the retina is upside-down (the brain affects another reversal). The eye can accommodate an extraordinary range of intensity levels. This is possible partly because of the iris, but mainly because of the presence of two types of receptors on the retina.

For a nearsighted person the refractive power of the eye is too large. The eye cannot focus on distant objects. By employing glasses with negative power the global refractive power is decreased.

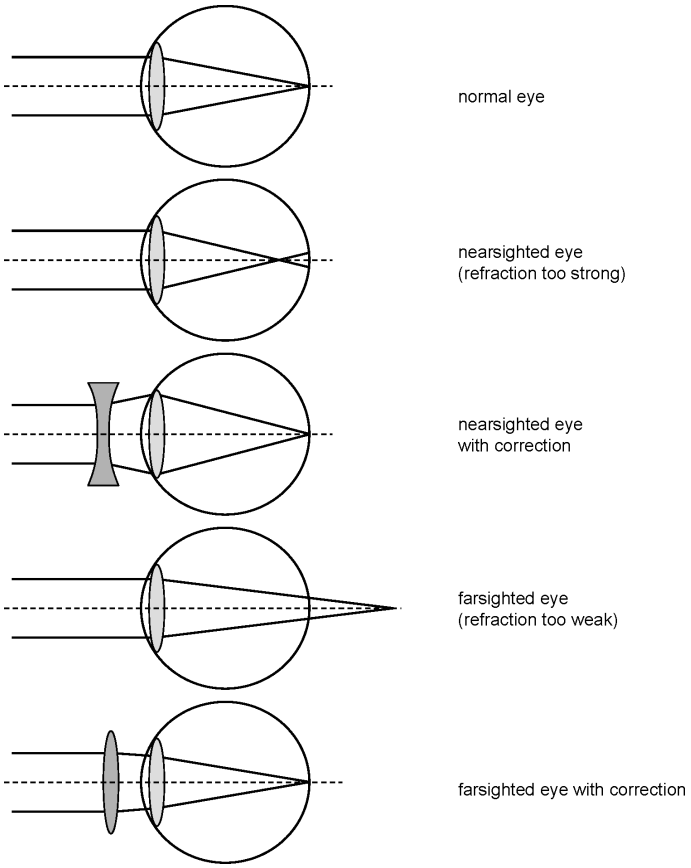

Figure 3.44: Nearsightedness, farsightedness and the necessary correcting lenses.

The glasses provide a virtual image that is closer to the eye than the object itself. For a farsighted person the opposite happens. Now glasses with a positive lens are used, which create a virtual image further away (figure 3.44).

The eye is relaxed the most if it looks at distant objects. Therefore, instruments for visual observation are designed so that a real or virtual image is created at an appreciable distance from the eye.

The resolution of our eyes is limited by three factors. First of all there are the aberrations of the optical system which limit the resolution. As will be discussed in the chapter on Gaussian beam optics, the diffraction limit (due to the finite size of the lens) imposes an additional restriction in resolution. A third factor influencing the resolution is the fact that the retina consists of discrete "pixels" (comparable to a digital camera). Away from the center of the retina, this imposes the largest resolution limit, since the density of the pixels is much lower there. In the center of the retina all three effects are of comparable magnitude.

3.6.2 Magnifying glass and eyepiece

The magnifying glass and the eyepiece (or ocular) are positive lenses or systems that are used when the object lies between the focal point in the object space and the system. This creates a

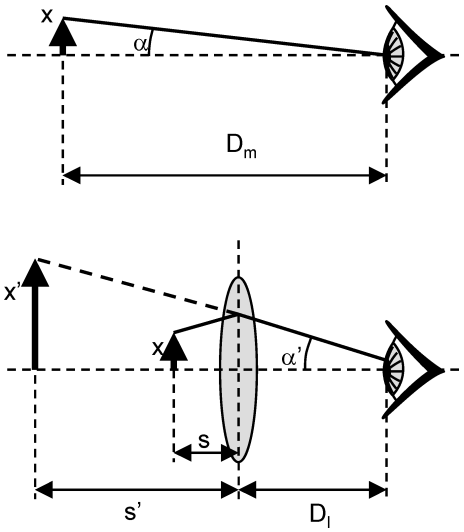

Figure 3.45: The eyepiece. (a) Imaging without eyepiece. (b) Imaging with eyepiece.

virtual image (generally at large distance before the system) without upside-down reversal. The term eyepiece is used for a magnifying glass held closely to the eye (with appropriate dimensions thereto). This is particularly the case for many optical instruments (microscope, telescope, binoculars, etc.), where the eyepiece serves to create a magnified virtual image of the real image obtained by the objective. In principle a magnifying glass or eyepiece can realize any magnification (defined as the ratio between image and object size) by correctly choosing the object distance. The magnification M is given by:

$$M = -\frac{s'}{s} = \frac{|s'|}{s} = |s'| \left(\frac{1}{f} + \frac{1}{|s'|} \right) = 1 + \frac{|s'|}{f} \tag{3.75}$$

This definition is often not very useful, as it indicates nothing about the visually perceived magnification by the eye. The following is a better definition: the magnification is the ratio between the size - as perceived by the eye - of the virtual image using the eyepiece and the object size without the eyepiece, taking the maximum size for both values. Figure 3.45 depicts both situations.

The size of an object perceived by the eye is determined by the angle α of the object with the axis, seen from the eye. Without lens this becomes:

$$\alpha = \frac{x}{D} \tag{3.76}$$

The closer the object is to the eye, the larger it seems. However, there is a minimum distance D_m, beyond which the image becomes unsharp:

$$\alpha_{\max} = \frac{x}{D_m} \tag{3.77}$$

Employing a lens, the angle for the virtual image becomes α' :

$$\alpha' = \frac{x'}{|s'| + D_l} = \frac{x \frac{|s'|}{s}}{|s'| + D_l} = \frac{x\,|s'|}{|s'| + D_l} \left(\frac{1}{f} + \frac{1}{|s'|} \right) \tag{3.78}$$

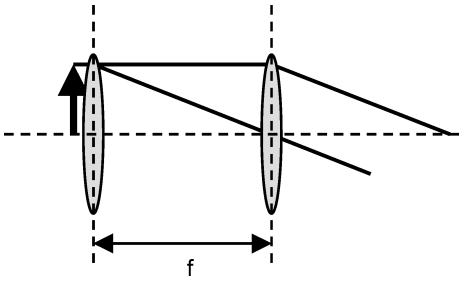

Figure 3.46: The *Ramsden* eyepiece.

The angle becomes larger as D_l decreases. Therefore we put D_l equal to 0. In practice we often look with the eye very close to the eyepiece, as in a microscope. For the magnification M we obtain:

$$M = D_m \left(\frac{1}{f} + \frac{1}{|s'|} \right) \tag{3.79}$$

We can still choose the image distance $|s'|$. Consider the two extreme situations. The largest distance is infinity, whereas the smallest is D_m. The magnifications for the two cases are:

$$M = \frac{D_m}{f} \text{ for } |s'| = \infty \tag{3.80}$$

$$M = \frac{D_m}{f} + 1 \text{ for } |s'| = D_m \tag{3.81}$$

$$\tag{3.82}$$

Thus, if the focal length f is small compared to the minimal distance D_m, the two expressions do not differ very much. The quantity $M = D_m/f$ is considered the nominal magnification of the eyepiece. Here D_m is standardized at $25cm$ (approximately the smallest distance still pleasant to the eye). Thus, an eyepiece with magnification $10\times$ has a focal length of $25mm$.

An eyepiece consisting of one lens will introduce an unacceptable amount of chromatic aberration in a microscope or telescope. Therefore one will often use two lenses. One possibility is to use an achromatic doublet, however this proves rather expensive. It is much simpler to use two identical lenses at focal length from each other (*Ramsden* eyepiece - figure 3.46). One can indeed prove that two lenses of the same glass behave achromatically if their distance is equal to half the sum of the respective focal lengths. In such a configuration the object plane is at the first lens (if we put the virtual image at infinity). A drawback is that dust on the first lens surface is imaged sharply. Therefore, we generally deviate slightly from the optimal achromatic design.

3.6.3 Objectives

An objective produces a real inverted image of the object. This image is created at a film plane or is viewed by an eyepiece (figure 3.47).

In a microscope the object is magnified by the objective. The magnification of the objective is given by:

$$M = -\frac{s'}{s} = -s' \left(\frac{1}{f} - \frac{1}{s'} \right) = 1 - \frac{s'}{f} \tag{3.83}$$

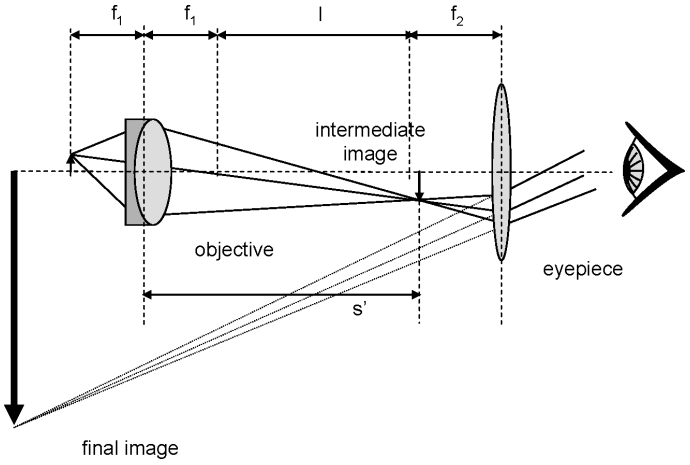

Figure 3.47: Objective + eyepiece.

Generally a large magnification is desired, which implies that $s' \gg f$. Thus, the object is approximately in the focal plane of the objective. The distance s' is standardized for microscopes at $16cm$. We get:

$$M \approx -\frac{s'}{f} = -\frac{16}{f\,[cm]}. \tag{3.84}$$

Thus, a microscope objective with a magnification $100\times$ has a focal length of $1.6mm$. The magnification and the numerical aperture are always indicated on the microscope objective. The global magnification of the microscope is the product of the objective and eyepiece magnifications, so

$$M_{tot} = -\frac{25}{f_{oc}\,[cm]}\frac{16}{f_{ob}\,[cm]}. \tag{3.85}$$

Thus, this is the size of the image seen by the eye in comparison to the size of the object itself, if it would be located at $25cm$ from the eye.

In a telescope the object (at very large distance) is shrunk, while the angles are enlarged. Now the image is approximately in the focal plane and the magnification of the objective is given by:

$$M = -\frac{s'}{s} = -\frac{1}{s}\left(\frac{1}{f} - \frac{1}{s}\right)^{-1} \approx -\frac{f}{s}. \tag{3.86}$$

The simplest type of telescope (figure 3.48) consists of watching this image with an eyepiece, so that the virtual image is again at a large distance.

If we assume that the virtual image is in the same plane as the object itself, the total angle magnification simply becomes:

$$M_{tot} = -\frac{f_{ob}}{f_{oc}} \tag{3.87}$$

This is also the angle magnification of the system. Such a telescope - called an astronomical telescope - has a global refractive power of zero: a ray arriving at the image has an angle that only depends on the starting angle. This type has an inverted image. To obtain a normal image a Galilean telescope should be used (figure 3.49), where a negative eyepiece converts the converging rays from the objective into a parallel beam before forming a real image.

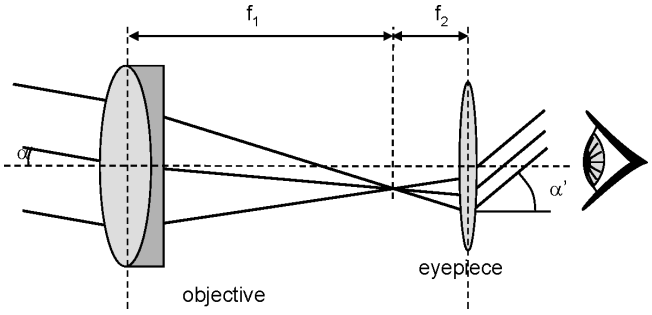

Figure 3.48: A simple telescope.

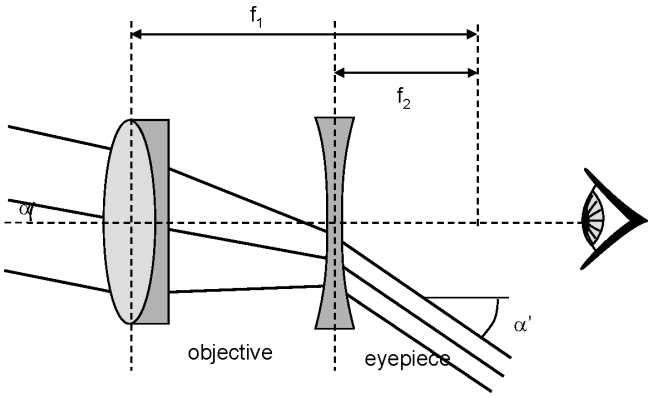

Figure 3.49: A Galilean telescope.

In a normal photographic camera the objective is used in approximately the same way as for the telescope: the object distance is large compared to the image distance. The film plane is thus equal to or slightly past the focal plane of the objective.

The focal length (in mm) and the f-number of a photographic objective are always indicated on the lens. A typical focal length is 50 mm. It determines the typical physical dimensions of a camera. If the image of an object has to be enlarged, there are two options: decrease the object distance or increase the focal length by employing another lens. For a strong tele-objective the length in case of a single lens would be impractically large. Therefore one uses a lens combination with a larger focal length, but which is relatively short because both principal planes are located on the object side of the lens, as illustrated in figure 3.50.

3.6.4 Camera

The most important part of the *camera* is the objective that creates a real inverted image on the film, as previously described (figure 3.51). Also important is the ability to visually observe the scene that is photographed. In the reflex camera this is done via a 45 degree mirror (which is removed at the moment the picture is taken) that reflects the image upwards. This creates a real image at a certain location, that one could see with an eyepiece. In practice one puts a diffuse or ground glass at the position of the real image, that scatters the incident rays. This image is then observed virtually, again by using an eyepiece. The use of the diffuser has a number of advantages. First

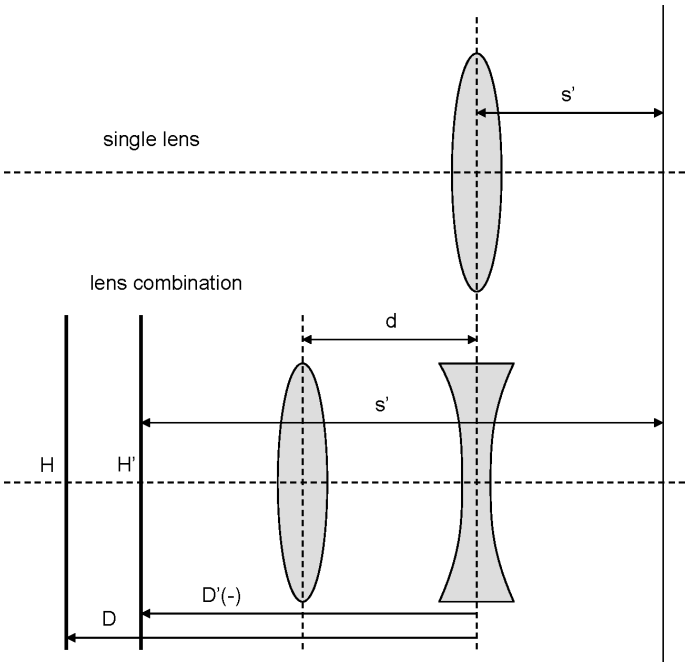

Figure 3.50: Increasing the object distance of an objective by using a lens combination. (a) Single lens with short image distance. (b) Combination with limited thickness but much larger image distance.

of all it allows for easy focusing. The location of the diffuser corresponds with the location of the film plane. Upon bad focusing there is a fuzzy image on the diffusing glass. Without diffuser the eye would be able to see the image sharply, because of its accommodation capacity.

Furthermore, a ground glass screen allows for easy incorporation of auxiliary focusing aids (e.g. microprisms). Finally, without diffuser one would obtain a very dark image in the corners of the screen, because these corner rays have a large angle with the optical axis and are not captured by the simple eyepiece. However, even with diffuser there is still a relevant dimming towards the corners. To decrease this one sometimes places a Fresnel lens in front of the diffusing glass, which makes that oblique rays travel parallel again with the optical axis.

In some older cameras one had to look vertically to the ground glass screen. The image was upright but left-right flipped. To look horizontally one would need another 45 degree mirror. However, this would make the image both left-right and upside-down inverted. The solution of these problems was brought with the *pentaprism*, in which every ray reflects on three faces of this multifacetted prism (figure 3.52). This creates a correct image.

3.6.5 Binoculars

Binoculars are based on the simple principle of the astronomical telescope. This means the image is inverted again (in both directions), which can be solved in several ways. One can insert an extra lens that creates reversal, however this lengthens the instrument and increases chances of aberrations. Another solution is to use two mirrors, that flip the image in two steps (left-right and upside-down). Unfortunately, the direction of observation would not coincide with the direction of the object, which is again unpractical.

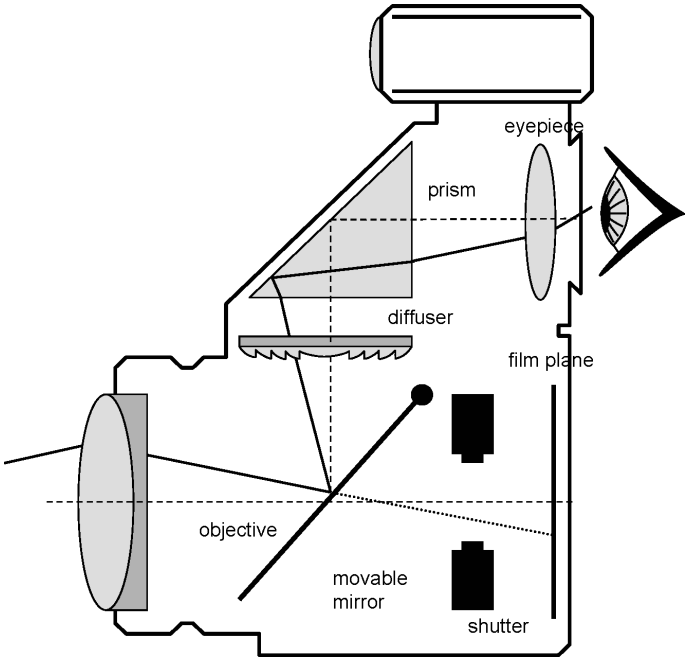

Figure 3.51: The reflex camera.

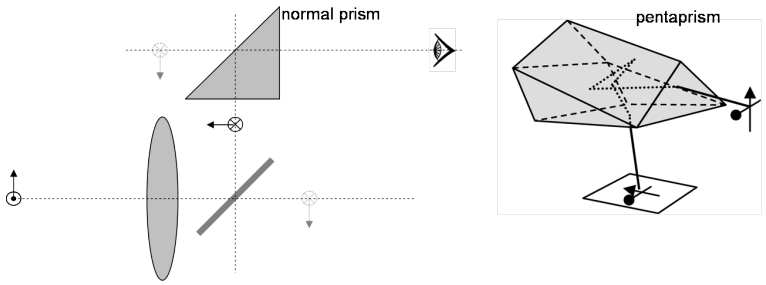

Figure 3.52: (a) A reflex camera with normal triangular prism (2D) and (b) a pentaprism.

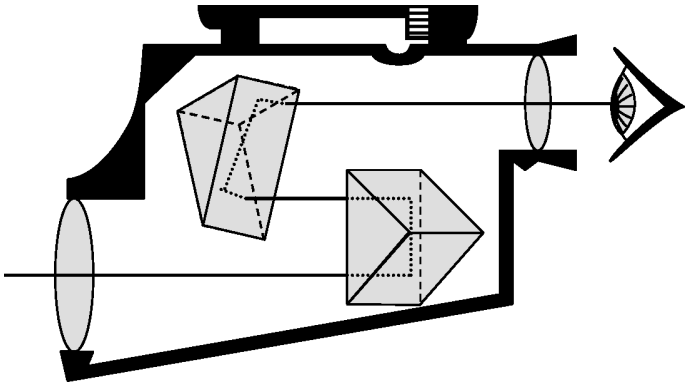

Figure 3.53: Binoculars.

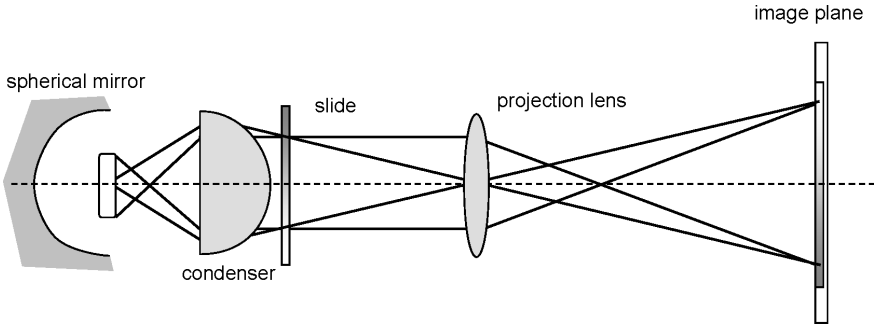

Figure 3.54: The slide projector.

The good solution is to use two prisms, where every ray reflects in each prism on two faces, see figure 3.53. In this way the image is reversed but the observation direction is the same as the object direction. Moreover, this approach folds the ray trajectories, so that the binoculars become more compact.

3.6.6 Projection systems

In *projectors* (slide projector, overhead projector) an image from the transparent object has to be created. Furthermore the light of the source has to go through the object so that the image is as strongly lit and uniform as possible. To achieve this in a slide projector a condenser lens is put in front of the slide (figure 3.54). This lens captures as many rays as possible from the source, and refracts them in the direction of the projecting lens. Actually, the source is imaged by the condenser into the plane of the projector lens. Thus, the latter has to have at least the size of this image. The condenser needs to be at least as large as the slide and evidently needs as large a numerical aperture as possible. In practice an aspherical lens is often used.

The overhead projector, depicted in figure 3.55, does the same in principle. However, because of the size of the transparent object the use of a condenser lens is quasi impossible. Instead a Fresnel lens (see below) is mostly used, which is not at all perfect with respect to imaging, but still deflects a large part of the power in the right direction.

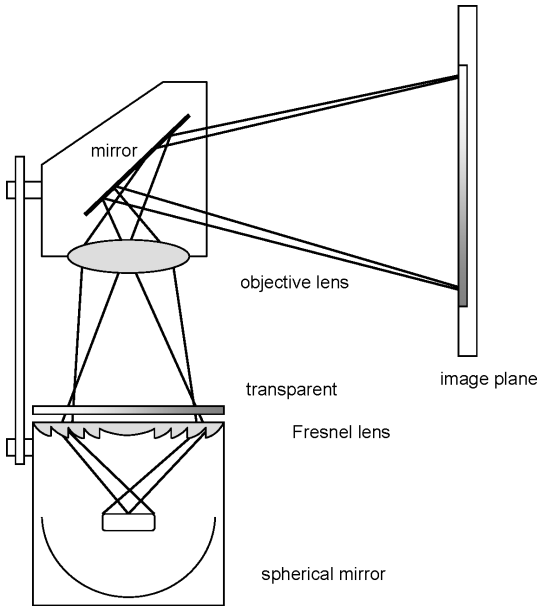

Figure 3.55: The overhead projector.

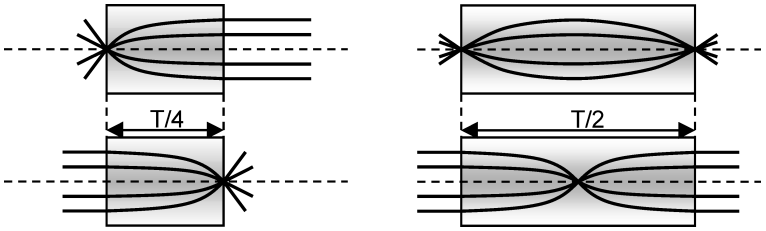

Figure 3.56: GRIN-lenses.

3.6.7 GRIN lenses

In fibers with parabolic index profile the ray trajectories are sinusoidal with period independent of the location and angle of incidence (see page 3–9 and chapter 7). This property is used for a special kind of lens: the *GRIN* (GRaded INdex) or SELFOC lens (figure 3.56). It consists of a thick graded index fiber with length equal to a fraction (e.g. $1/4$ or $1/2$) of the sine period. In this way the system creates a 1:1 image, or it transforms a point source into a parallel beam (or vice versa). A main advantage of the GRIN lens is the ease of component connection.

3.6.8 Fiber bundles

In geometrically challenging circumstances (flexible system, limited space) it can be useful to employ an ordered fiber bundle, where each object (and image) point corresponds to a distinct optical fiber. (More details about guiding in optical fibers can be found in chapter 7.) The number of *pixels* is thus limited by the number of fibers. Fiber bundles are often used as 1:1 imaging systems in medicine (e.g. endoscope). An alternative application is to transform a source with a certain shape into a source with another shape.

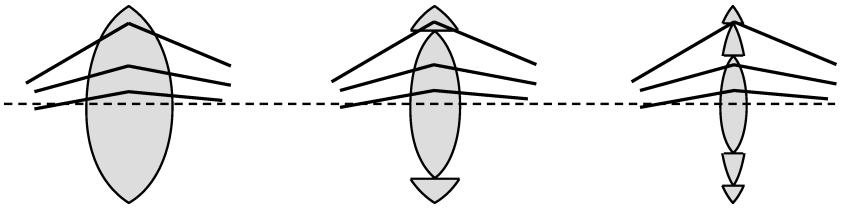

Figure 3.57: Transformation from a classical lens to a Fresnel lens.

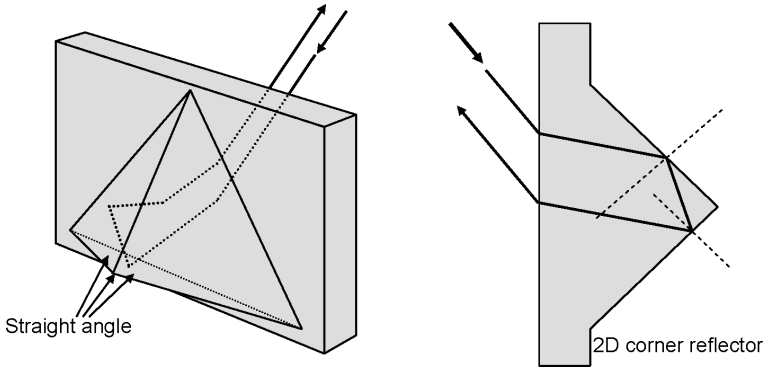

Figure 3.58: The corner reflector.

3.6.9 Fresnel lenses

Lens operation originates from the refraction of rays at surfaces. For this refraction only the angle between the ray and the surface is important. This means that the lenses in figure 3.57 have about the same functionality, provided we do not concern ourselves with rays incident on the discontinuous transitions.

Such a lens is called a *Fresnel lens* and often looks like a plane plate with a surface profile. It is used in cases where the equivalent normal lens would be too thick, often for lenses with large diameter. The Fresnel lens is commonly used to focus the light of a lamp in a certain direction (car lights, traffic lights, etc.). As previously mentioned, it is also employed in the camera and the overhead projector. These applications are not exigent with respect to aberrations (with respect to the function of the Fresnel lens), so that scattering at the transitions does not pose a problem.

3.6.10 Corner reflector

A corner reflector or corner-cube prism consists of three perpendicular mirrors. Incident light will be reflected in the same direction because of the three reflections. Reflectors in traffic (on bikes, road markings, etc.) contain a large number of corner reflectors next to each other. Instead of mirrors, the light is reflected here by total internal reflection (figure 3.58)). Corner reflectors lose their function partly for coherent light, as the phase relations change because of the different reflections. A plane wave is not reflected into a plane wave.

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

[Wei74] R. Weinstock. *Calculus of Variation*. Dover, New York, 1974.

Chapter 4

Scalar Wave Optics

Contents

In the second half of the 17th century the Dutch physicist Christiaan Huygens postulated that light was a wave phenomenon. Thus, light propagates as waves and each wave has an associated specific wavelength, just as a wave propagating along a rope.

In this chapter we describe the behavior of light by employing a scalar function, the **wave function**, which satisfies the wave equation. This wave theory encompasses the entire ray theory, and allows to present aspects of light unexplainable by the ray concept.

Theoretically speaking ray optics is the limit of wave theory in which the wavelength becomes infinitely small, or the frequency infinitely large. In practice, ray theory is rather accurate if the light propagates through objects that have much larger dimensions than the wavelength of the light.

As later described in chapter 6, light is actually an electromagnetic wave with a transverse vectorial character. However, the scalar description is mathematically much simpler than the electromagnetic theory. Nonetheless, the scalar approach allows us to present certain aspects of light in an easy way. Thus, here we neglect the vectorial nature of light, and we assume that the wave function represents any component of the electric or magnetic field. However, we use certain postulates to define physically observable quantities. In chapter 6 we will check the postulates by using the electromagnetic theory.

4.1 The postulates of wave optics

4.1.1 The wave equation

1. Light waves propagate in free space with the speed of light

$$c \approx 3.0 \times 10^8 m/s = 30 cm/ns \tag{4.1}$$

2. Homogeneous, isotropic, transparent media (such as glass) are characterized by a single constant, the refractive index $n(\geq 1)$. In a medium with refractive index n light propagates at a speed of $v = c/n$.

3. A light wave is described by a real scalar function $u(\mathbf{r}, t)$, the wave function, which satisfies the wave equation

$$\nabla^2 u - \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{4.2}$$

with $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$, the Laplacian operator.

4. Each function satisfying the wave equation describes a possible light wave.

5. The wave equation is linear, thus the superposition principle holds. If $u_1(\mathbf{r}, t)$ and $u_2(\mathbf{r}, t)$ are solutions, then $u_1(\mathbf{r}, t) + u_1(\mathbf{r}, t)$ is also a solution.

6. The wave function is continuous on the boundary between two different media. This assumption is actually the biggest approximation of scalar wave theory. The exact description of the behavior at an interface can only happen with inclusion of the vectorial nature of the light waves, and is described in chapter 6.

7. The scalar wave equation is approximately applicable for media with location dependent index (e.g. GRIN material), provided that the index variation is small over distances on the order of the wavelength of the light. In effect, these media are locally-homogeneous. Their index is described by a location dependent index $n(\mathbf{r})$, and thus with a location dependent speed of light $v(\mathbf{r})$.

4.1.2 Intensity and power

Light intensity ($Watt\ cm^{-2}$)

One can measure the intensity of light, in contrast to the wave function itself. The expression for intensity[1] connects the postulated wave function to a physically observable quantity

$$I(\mathbf{r}, t) = 2 < u^2(\mathbf{r}, t) > \tag{4.3}$$

The brackets $< . >$ stand for the calculation of the average over a time interval much larger than the period of an optical cycle. Because I is a physical quantity and u is not, the choice of the factor 2 is arbitrary. However, because of this choice equation (4.13) will look nice.

[1] Note that intensity has a different meaning here as compared to the luminous intensity defined in chapter 2

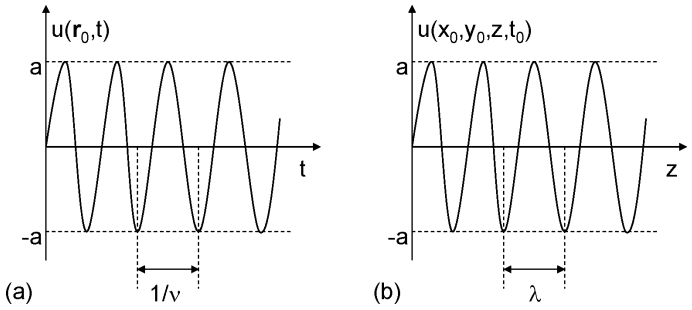

Figure 4.1: (a) monochromatic wave at a fixed location. (b) Monochromatic wave at a fixed time.

Optical power of a light beam propagating through an arbitrary surface A normal to the propagation direction (*Watt*)

$$P(t) = \int_A I(\mathbf{r}, t) dA \tag{4.4}$$

4.2 Monochromatic waves

A monochromatic wave can be described by a wave function with harmonic time dependence

$$u(\mathbf{r}, t) = a(\mathbf{r}) \cos\left[2\pi\nu t + \varphi(\mathbf{r})\right] \tag{4.5}$$

with

- $a(\mathbf{r})$ = amplitude
- $\phi(\mathbf{r})$ = phase
- ν = frequency (in Hz)
- $\omega = 2\pi\nu$ = angular frequency (in rad/s)

Although amplitude and phase can be location dependent, the wave function varies harmonically with the same frequency ν at all locations.

4.2.1 Complex representation and Helmholtz equation

As will become clear, it is usually easier to describe the wave function $u(\mathbf{r}, t)$ with a complex function, also called the analytic signal

$$U(\mathbf{r}, t) = a(\mathbf{r}) e^{j\varphi(\mathbf{r})} e^{j2\pi\nu t} \tag{4.6}$$

so that

$$\begin{aligned} u(\mathbf{r}, t) &= \text{Re}\left\{U(\mathbf{r}, t)\right\} \\ &= \frac{1}{2}\left[U(\mathbf{r}, t) + U^*(\mathbf{r}, t)\right] \end{aligned} \tag{4.7}$$

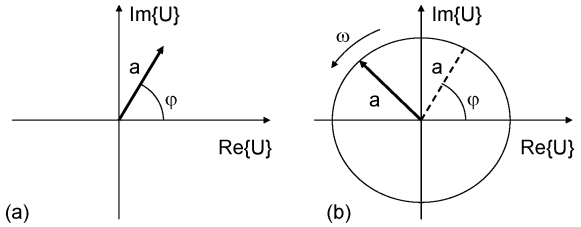

Figure 4.2: (a) Phasor diagram. (b) Rotating phasor.

This also means that the complex function $U(\mathbf{r}, t)$ obeys the wave equation (4.2) as well, thus

$$\nabla^2 U - \frac{1}{v^2}\frac{\partial^2 U}{\partial t^2} = 0 \tag{4.8}$$

The complex amplitude

Equation (4.6) can be written as

$$U(\mathbf{r}, t) = U(\mathbf{r})e^{+j2\pi\nu t} \tag{4.9}$$

The time independent factor $U(\mathbf{r}) = a(\mathbf{r})e^{j\phi(\mathbf{r})}$ is called the complex amplitude. $U(\mathbf{r})$ describes the time invariant envelope of the propagating wave, and this is a complex variable with

- $|U(\mathbf{r})|$ = the amplitude of the wave
- $arg\,[U(\mathbf{r})] = \phi(\mathbf{r})$ = the phase of the wave

Geometrically the complex amplitude can be represented in a phasor diagram, as shown in figure 4.2(a). The complex wave function is then depicted as a phasor turning around with circulation frequency ν (see figure 4.2(b)).

The Helmholtz equation

Because of the linear character of the wave equation we often eliminate the time factor $e^{+j2\pi\nu t}$ and thus the time dependance. If we substitute the function $U(\mathbf{r}, t)$ from equation (4.9) in the wave equation (4.2), we obtain the Helmholtz equation

$$(\nabla^2 + k^2)U(\mathbf{r}) = 0 \tag{4.10}$$

with

$$k = \frac{2\pi\nu}{v} = \frac{\omega}{v} \tag{4.11}$$

the propagation constant.

Wave front

A wave front is a surface of equal phase, so that $\phi(r) = cst$. The value of this constant is often taken as an integer times 2π, thus $\phi(\mathbf{r}) = 2\pi q$ with q integer. The normal to the wave front in the point $\mathbf{r}$ is parallel with the gradient $\nabla\phi(\mathbf{r})$ in that point. It is the direction in which the phase changes most rapidly.

Intensity

To calculate the intensity of a monochromatic wave, we substitute the function (4.5) in equation (4.3)

$$\begin{aligned} 2u^2(r,t) &= 2a^2(\mathbf{r})\cos^2\left(2\pi\nu t + \varphi(\mathbf{r})\right) \\ &= |U(\mathbf{r})|^2\left\{1+\cos\left(2\left[2\pi\nu t+\varphi(\mathbf{r})\right]\right)\right\} \end{aligned} \tag{4.12}$$

If we take the average over a time interval equal to an integer times the optical period, $1/\nu$, the cosine term vanishes.

$$I(\mathbf{r}) = |U(\mathbf{r})|^2 \tag{4.13}$$

The intensity of a monochromatic wave is equal to the modulus squared of its complex amplitude. Furthermore, the intensity of a monochromatic wave does not vary in time.

4.2.2 Elementary waves

The Helmholtz equation has a number of relatively simple solutions, which will be described here.

The plane wave

A plane wave has a complex amplitude

$$U(\mathbf{r}) = Ae^{-j\mathbf{k}\cdot\mathbf{r}} = Ae^{[-j(k_x x+k_y y+k_z z)]} \tag{4.14}$$

A is a complex constant, called the complex envelope, and $\mathbf{k} = (k_x, k_y, k_z)$ is the wave vector. In order for the plane wave to satisfy the wave equation (4.2), it is necessary that ${k_x}^2+{k_y}^2+{k_z}^2 = k^2$, so that the magnitude of the wave vector $\mathbf{k}$ is equal to the propagation constant k.

- The wave fronts are defined as surfaces of constant phase. From (4.14) we get $arg(U(\mathbf{r})) = arg(A) - \mathbf{k}.\mathbf{r}$ so that the wave fronts are determined by

  $$\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y + k_z z = 2\pi q + arg\{A\} \tag{4.15}$$

  with q an integer. Thus, the wave fronts are parallel planes perpendicular to the wave vector $\mathbf{k}$. The wavelength λ is the distance between two consecutive wave fronts ($q = 0, 1, 2, 3\ldots$), and is given by

  $$\lambda = \frac{v}{\nu} \tag{4.16}$$

- If we assume that $\mathbf{k}$ is in the positive z-direction, then $U(\mathbf{r}) = Ae^{-jkz}$ and the corresponding wave function is given by (4.7)

  $$\begin{aligned} u(r,t) &= |A|\cos\left[2\pi\nu t - kz + \arg\{A\}\right] \\ &= |A|\cos\left[2\pi\nu\left(t-\frac{z}{v}\right)+\arg\{A\}\right] \end{aligned} \tag{4.17}$$

  We observe that the wave function is periodic in time with period $\frac{1}{\nu}$ and periodic in space with period $\frac{2\pi}{k} = \lambda$. Note that Ae^{-jkz} (with positive k) represents a wave propagating in

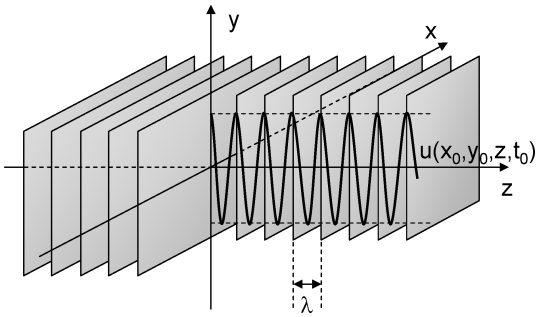

Figure 4.3: Wave fronts and amplitude of a plane wave.

the positive z-direction, as a consequence of the (arbitrary) choice of the plus-sign in the exponent of equation (4.9). If the other convention is used (as in some books), then Ae^{+jkz} is a forward wave.

- The phase of the complex wave function $\left[arg(U(\mathbf{r},t)) = 2\pi\nu\left(t - \frac{z}{v}\right) + arg(A)\right]$ varies in time and space as a function of $\left(t - \frac{z}{c}\right)$. Thus, v is called the phase velocity of the wave, because the wave fronts (surfaces of constant phase) propagate with speed v in the direction of the **k**-vector.

- If the wave propagates in a medium with refractive index n, the phase velocity becomes $v = \frac{c}{n}$, so that $\lambda = \frac{v}{\nu} = \frac{c}{n\nu} = \frac{\lambda_0}{n}$. **Thus, if a monochromatic wave propagates in a medium with index n the frequency remains the same, but the phase velocity, wavelength and propagation constant change according to**

$$\begin{aligned} v &= \frac{c}{n} \\ \lambda &= \frac{\lambda_0}{n} \\ k &= nk_0 \end{aligned} \tag{4.18}$$

- A plane wave implies a constant intensity $I(\mathbf{r}) = |A|^2$ everywhere in space. Therefore a plane wave is clearly non-physical, as it is everywhere and always present and thus carries an infinite amount of power. Nonetheless the concept of a plane wave is very useful, and it is often employed to describe light propagation in various structures.

The evanescent plane wave

Until now the wave vector **k** was considered real, but it can also be complex $\mathbf{k} = \mathbf{k_R} + j\mathbf{k_I}$. Still, $k_x^2 + k_y^2 + k_z^2 = |k|^2 = k^2 = nk_0^2$ has to remain valid, where n becomes complex. Applied to equation (4.14) one obtains

$$\begin{aligned} U(\mathbf{r}) &= Ae^{-j\mathbf{k}\cdot\mathbf{r}} \\ &= Ae^{\mathbf{k_I}\cdot\mathbf{r}}e^{-j\mathbf{k_R}\cdot\mathbf{r}} \end{aligned} \tag{4.19}$$

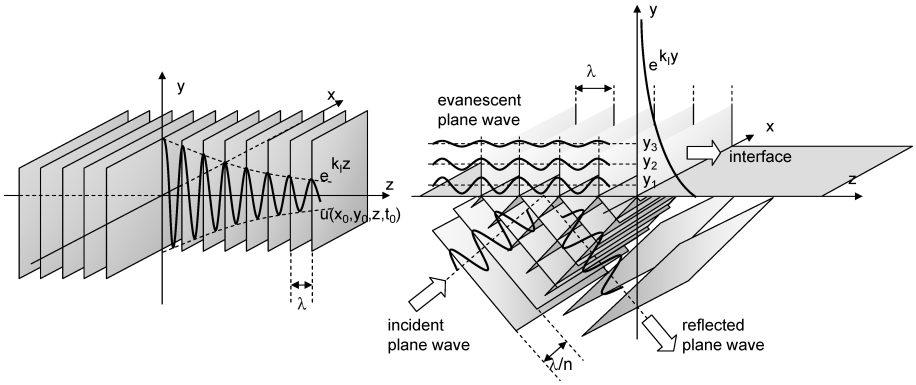

Figure 4.4: Wave fronts and amplitude of an evanescent plane wave for (a) $\mathbf{k_R} \| \mathbf{k_I}$ and (b) $\mathbf{k_R} \perp \mathbf{k_I}$.

This equation represents a plane wave that propagates in the direction $\mathbf{k_R}$, and exponentially increases or decreases in the direction $\mathbf{k_I}$. We discuss two extreme cases:

1. $\mathbf{k_R} \| \mathbf{k_I}$
   Assume both $\mathbf{k_R}$ and $\mathbf{k_I}$ parallel with the z-axis, then we get (4.14)

$$U(\mathbf{r}) = Ae^{k_I z}e^{-jk_R z} \tag{4.20}$$

   As the wave propagates deeper into the medium its amplitude decreases or increases exponentially. This corresponds to the propagation of a plane wave in an absorbing or amplifying medium.

2. $\mathbf{k_R} \perp \mathbf{k_I}$
   Assume $\mathbf{k_R}$ parallel to the z-axis and $\mathbf{k_I}$ parallel to the x-axis, thus perpendicular to $\mathbf{k_R}$, then (4.14) becomes

$$U(\mathbf{r}) = Ae^{k_I x}e^{-jk_R z} \tag{4.21}$$

   The wave propagates in the z-direction, while its amplitude decreases exponentially in the x-direction. This situation appears when total internal reflection occurs at an interface, and this will be extensively discussed in chapters 6 and 7. We mention here that if the angle of the wave vector of the incident light with the interface is smaller than a certain value, all energy will be reflected and no light is transmitted through the interface. This total internal reflection does generate a wave on the other side of the interface, that propagates parallel to the interface and decreases exponentially perpendicular to it.

The spherical wave

The complex amplitude of the spherical wave is

$$U(\mathbf{r}) = \frac{A}{r}e^{-jkr} \tag{4.22}$$

with r the distance to the origin and $k = \frac{2\pi\nu}{v} = \frac{\omega}{v}$ the wave number.

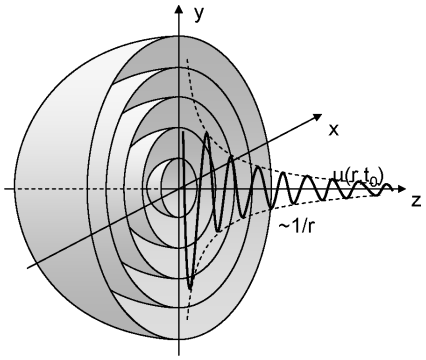

Figure 4.5: Wave fronts and amplitude of a spherical wave.

- If, for simplicity, we assume $arg(A) = 0$. We can now determine the wave fronts from equation (4.22): $kr = 2\pi q$ or $r = q\lambda$, with q an integer. Thus the wave fronts are concentric spheres separated by a radial distance $\lambda = 2\pi/k$, that propagate radially with phase velocity v.

- The $-$ sign in the exponential of equation (4.22) implies that the wave fronts start at the origin (a point source) and grow as they propagate away from the origin. Changing the sign into $+$ describes a spherical wave propagating towards the origin. (This applies with the convention $e^{+j2\pi\nu t}$. With the use of $e^{-j2\pi\nu t}$ one has to switch the signs in the previous explanation.)

- A spherical wave originating from the point $\mathbf{r_0}$ has the complex amplitude

$$U(\mathbf{r}) = \left(\frac{A}{|\mathbf{r} - \mathbf{r_0}|}\right) e^{(-jk|\mathbf{r}-\mathbf{r_0}|)} \tag{4.23}$$

  The wave fronts are concentric spheres centered on $\mathbf{r_0}$.

- The intensity of a spherical wave is inversely proportional with the square of the distance to the point source

$$I(\mathbf{r}) = \frac{|A|^2}{r^2} \tag{4.24}$$

The Fresnel approximation of the spherical wave: the parabolic wave

We consider again a spherical wave and assume an optical system where the z-axis is the main light propagation axis, so that we are interested in the behavior of the spherical wave along this axis. We examine the wave at a point $\mathbf{r} = (x, y, z)$ far from the source (large z), but close to the propagation axis (small x and y), so that $(x^2 + y^2)^{\frac{1}{2}} \ll z$. Thus we have $\theta^2 = \left(x^2 + y^2\right)/z^2 \ll 1$

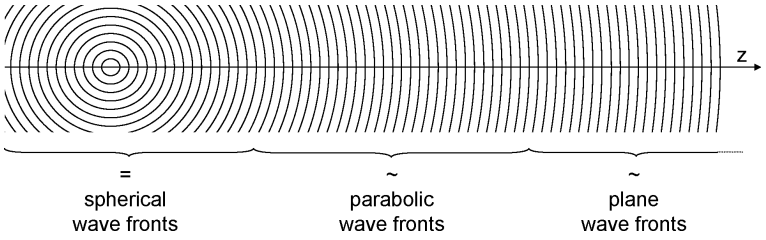

Figure 4.6: Evolution of a spherical wave near the propagation axis.

and we can use a Taylor expansion

$$\begin{aligned} r = \left(x^2+y^2+z^2\right)^{\frac{1}{2}} &= z\left(1+\theta^2\right)^{\frac{1}{2}} \\ &= z\left(1+\frac{\theta^2}{2}-\frac{\theta^4}{8}+...\right) \\ &\approx z\left(1+\frac{\theta^2}{2}\right) = z+\frac{x^2+y^2}{2z} \end{aligned} \tag{4.25}$$

After substituting $r = z + \left(x^2+y^2\right)/2z$ in the phase and $r = z$ in the amplitude of equation (4.22), we obtain

$$U(\mathbf{r}) \approx \frac{A}{z} e^{(-jkz)} e^{-jk\frac{x^2+y^2}{2z}} \tag{4.26}$$

We used a more precise approximation of r for the phase, as it is more sensitive to small perturbations. The previous expression is called the Fresnel approximation of the spherical wave.
This approximation consists of two parts. The first part describes a normal spherical wave propagating along z. The second part of expression (4.26) is a pure phase factor and determines the wave fronts as the spherical wave propagates along z. The phase factor induces that the wave fronts are bent into paraboloids, as one needs that $z + \left(x^2+y^2\right)/2z = constant$.
If z becomes very large we can assume $\left(x^2+y^2\right)/2z \approx 0$. Thus the curvature of the wave fronts disappears and we have plane waves. Therefore, the light of a star (a point source emitting spherical waves) has plane wave fronts.

4.2.3 Paraxial waves

Just as we have considered paraxial rays in ray optics, we can use paraxial waves in wave optics. Starting from a plane wave propagating in the z-direction as a carrier wave, we obtain a paraxial wave if we modulate the complex envelope A in such a way that it is a slowly varying function of location $A(\mathbf{r})$

$$U(\mathbf{r}) = A(\mathbf{r}) e^{-jkz} \tag{4.27}$$

The change of $A(\mathbf{r})$ with position has to be slow compared to the wavelength $\lambda = 2\pi/k$, for the wave to maintain a plane wave character.

Figure 4.7 illustrates the wave function

$$u(\mathbf{r},t) = |A(\mathbf{r})|\, cos\left[2\pi\nu t - kz + arg\left[A(\mathbf{r})\right]\right] \tag{4.28}$$

of a paraxial wave. Along the z-axis it is a sine function with amplitude $|A(0,0,z)|$ and phase $arg\left[A(0,0,z)\right]$ that vary slowly in function of z. As the phase $[arg\left[A(x,y,z)\right]]$ varies slowly with

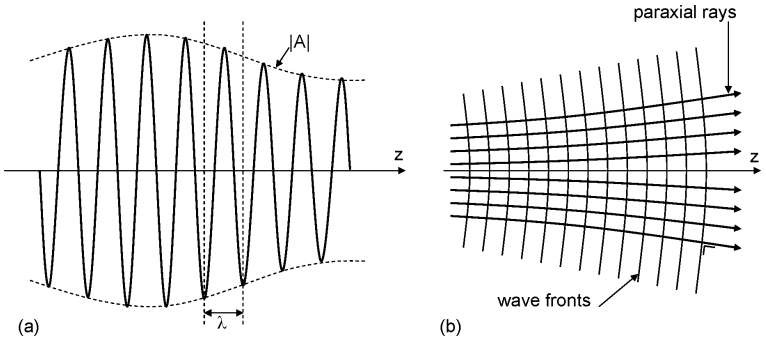

Figure 4.7: (a) The amplitude of a paraxial wave in function of the axial distance z, (b) wave fronts and wave front normals of a paraxial wave.

z over a distance λ, the plane wave fronts $kz = 2\pi q$ of the carrier are curved slightly, so that the normals are paraxial rays.

The paraxial Helmholtz equation

The assumption that $A(\mathbf{r})$ varies slowly with z implies that the change of A, ΔA, over a length $\Delta z = \lambda$ is much smaller than A itself, so $\Delta A = (\partial A/\partial z)\Delta z = (\partial A/\partial z)\lambda \ll A$. As $A/\lambda = kA/2\pi$ we deduce that

$$\frac{\partial A}{\partial z} \ll kA \tag{4.29}$$

Analogously, the derivative of $\partial A/\partial z$ varies slowly over a length λ, so that $\partial^2 A/\partial z^2 \ll k\partial A/\partial z$, and

$$\begin{aligned} \frac{\partial^2 A}{\partial z^2} &\ll k\partial A/\partial z \\ &\ll k^2 A \end{aligned} \tag{4.30}$$

To obtain the Helmholtz equation for paraxial waves we first substitute (4.27) into the Helmholtz equation (4.10), where we split the transversal and longitudinal components of the Laplacian

$$\nabla_T^2 U(\mathbf{r}) + \frac{\partial^2 U(\mathbf{r})}{\partial z^2} + k^2 U(\mathbf{r}) = 0 \tag{4.31}$$

with $\nabla_T^2 = (\partial^2/\partial x^2) + (\partial^2/\partial y^2)$. Working with the second term in the equation above we get

$$\frac{\partial^2 U(\mathbf{r})}{\partial z^2} = e^{-jkz}\left[-2jk\frac{\partial A(\mathbf{r})}{\partial z} + \frac{\partial^2 A(\mathbf{r})}{\partial z^2} - k^2 A(\mathbf{r})\right] \tag{4.32}$$

After substitution in equation (4.31) and use of (4.29) and (4.30) we obtain an equation for the slowly varying envelope of a paraxial wave, called the paraxial Helmholtz equation

$$\nabla_T{}^2 A(\mathbf{r}) - 2jk\frac{\partial A(\mathbf{r})}{\partial z} = 0 \tag{4.33}$$

This equation is much easier to solve (both analytically and numerically) compared with the Helmholtz equation. Starting with $A(x, y, 0)$ one can find $A(x, y, z)$ by integration of equation (4.33). In the next chapter we discuss some solutions, namely the Gaussian and the Hermite-Gaussian beams.

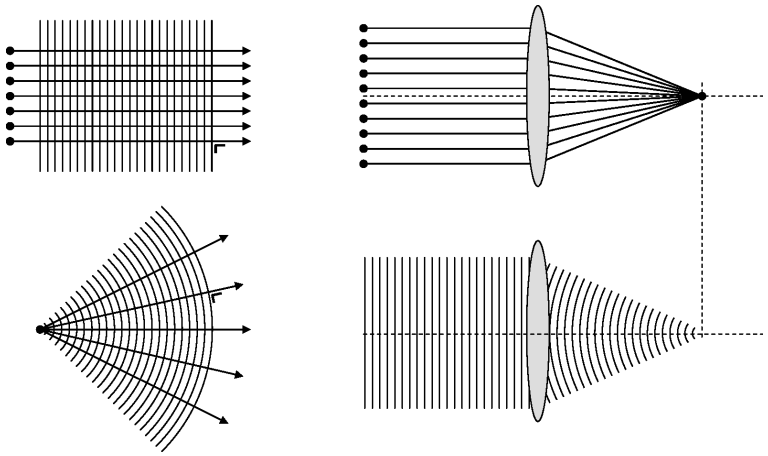

Figure 4.8: (a) The rays are perpendicular to the wave fronts. (b) The effect of a lens on the rays and wave fronts.

4.3 Deduction of ray theory from wave theory

The ray theory is the limit of wave theory as the wavelength $\lambda_0 \to 0$, as mentioned in the introduction to this chapter. To illustrate this, we consider a monochromatic wave with wavelength λ_0 in free space. The wave propagates in a medium with position dependent but slowly varying refractive index $n(\mathbf{r})$, so that the medium is considered locally homogeneous. The complex amplitude of the wave is given by

$$U(\mathbf{r}) = a(\mathbf{r})e^{[-jk_0S(\mathbf{r})]} \tag{4.34}$$

with $a(\mathbf{r})$ the amplitude, $-k_0S(\mathbf{r})$ the phase and $k_0 = 2\pi/\lambda_0$ the wave number. We assume that $a(\mathbf{r})$ varies slowly with $\mathbf{r}$, so we consider $a(\mathbf{r})$ constant over a length λ_0.

The wave fronts are surfaces determined by $S(\mathbf{r}) = cst$, and the normal to the wave fronts points in the direction of the gradient ∇S. In the vicinity of a point $\mathbf{r_0}$ we consider the wave as a plane wave with amplitude $a(\mathbf{r_0})$, wave vector $\mathbf{k}$ with size $k = n(\mathbf{r_0})k_0$ and direction parallel to the gradient vector ∇S in $\mathbf{r_0}$. Another neighborhood implies a different local plane wave with different amplitude and wave vector.

We associate the local wave vectors (normals to the wave fronts) in scalar wave optics with the rays in ray optics. This analogy shows that ray optics can be used to approximately determine the effect of optical components on the normals of the wave fronts, as illustrated in figure 4.8.

The eikonal equation

After substitution of (4.34) in the Helmholtz equation we obtain

$$k_0{}^2\Big[n^2 - |\nabla S|^2\Big]a + \nabla^2 a - jk_0\big[2\nabla S \cdot \nabla a + a\nabla^2 S\big] = 0 \tag{4.35}$$

with $a = a(\mathbf{r})$ and $S = S(\mathbf{r})$. Both the real and imaginary part have to be equal to zero. The real part leads to

$$|\nabla S|^2 = n^2 + \left(\frac{\lambda_0}{2\pi}\right)^2 \nabla^2 a/a \tag{4.36}$$

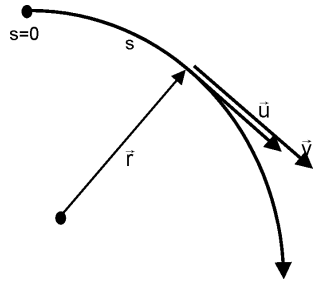

Figure 4.9: Propagation of a ray in space.

The assumption that a is slowly varying over a length λ_0 means that

$$\lambda_0{}^2\nabla^2 a/a \ll 1 \tag{4.37}$$

so the second term on the left side of equation (4.36) can be neglected in the limit $\lambda_0 \to 0$

$$|\nabla S|^2 \approx n^2 \tag{4.38}$$

This is the so-called eikonal equation and the scalar function S(r) is called the eikonal.

If we put the imaginary part of (4.35) equal to zero, we get a relation between a and S that allows us to determine the wave function.

The ray equation

The eikonal equation determines the surfaces with constant phase $S(x, y, z) = constant$. A ray of light can be considered as a local plane wave that propagates perpendicular to the surfaces of constant phase. Thus, the rays are orthogonal lines to the wave fronts $S(x, y, z) = constant$. If s is the length along the ray and $\mathbf{r}(s)$ is the vector function describing the propagation of the ray (see figure 4.9), then the vector $\mathbf{u}(s)$, defined as

$$\mathbf{u} = \frac{d\mathbf{r}}{ds}, \tag{4.39}$$

is the unit vector along the ray. The vector $\mathbf{v} = \nabla S$ is also perpendicular to the phase fronts, and thus it is parallel with $\mathbf{u}$. Because the size of $\mathbf{v}$ is given by n, the refractive index, one obtains $\mathbf{u} = \mathbf{v}/n$ or

$$\mathbf{v} = n\frac{d\mathbf{r}}{ds} \tag{4.40}$$

Taking the gradient of the eikonal equation, one gets

$$2\nabla S.\nabla\nabla S = 2n\nabla n \tag{4.41}$$

or after substitution of equation (4.39) and (4.40) into (4.41)

$$n\frac{d\mathbf{r}}{ds}.\nabla n\frac{d\mathbf{r}}{ds} = n\nabla n \tag{4.42}$$

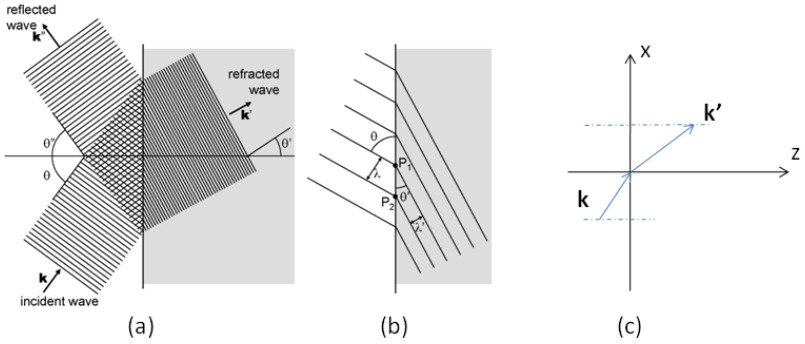

Figure 4.10: (a) refraction and reflection at an interface, (b) the agreement of phase fronts at the boundary implies $\lambda_1/sin\theta_1 = \lambda_2/sin\theta_2$ with $\lambda_1 = \lambda_0/n_1$ and $\lambda_1 = \lambda_0/n_1$, which leads to Snell's law, (c) continuity of the tangential component of the wavevector.

Because

$$\frac{d}{ds} = \frac{dx}{ds}\frac{\partial}{\partial x} + \frac{dy}{ds}\frac{\partial}{\partial y} + \frac{dz}{ds}\frac{\partial}{\partial z} = \frac{d\mathbf{r}}{ds}.\nabla \tag{4.43}$$

we immediately obtain

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n \tag{4.44}$$

This is the ray equation that we discussed in chapter 3. Thus, this shows that ray optics and Fermat's principle can be deduced from wave optics, and that all principles of ray optics are applicable to normals to the wave fronts of wave optics!

4.4 Reflection and refraction

4.4.1 Reflection and refraction at a planar dielectric boundary

We consider a plane wave with wave vector **k**, incident on a plane interface between two homogeneous media with indices n_1 and n_2, located in the plane $z = 0$. Refraction and reflection leads to waves with wave vectors $\mathbf{k}'$ and $\mathbf{k}''$ (figure 4.10). The combination of these three waves satisfies the Helmholtz equation so that $k = k'' = n_1 k_0$ and $k' = n_2 k_0$. Continuity of the wave function implies that the phase of the three waves at the boundary has to be equal, so

$$\begin{aligned} \mathbf{k}\cdot\mathbf{r} = \mathbf{k}'\cdot\mathbf{r} = \mathbf{k}''\cdot\mathbf{r}, \quad \mathbf{r} = (x, y, 0) \\ k_x x + k_y y = k'_x x + k'_y y = k''_x x + k''_y y \end{aligned} \tag{4.45}$$

This is true for all x and y values, so

$$\begin{aligned} k_x = k'_x = k''_x \\ k_y = k'_y = k''_y \end{aligned} \tag{4.46}$$

We can say that the tangential component of the wavevector is continuous at the interface. The vectors $\mathbf{k}$, $\mathbf{k}'$ and $\mathbf{k}''$ are given by

$$\begin{aligned} \mathbf{k} &= (n_1 k_0 sin(\theta), 0, n_1 k_0 cos(\theta)), \\ \mathbf{k}'' &= (n_1 k_0 sin(\theta''), 0, -n_1 k_0 cos(\theta'')), \\ \mathbf{k}' &= (n_2 k_0 sin(\theta'), 0, n_2 k_0 cos(\theta')), \end{aligned} \tag{4.47}$$

with θ, θ' and θ'' the angles of incident, refracted and reflected wave, respectively. This leads to $\theta = \theta''$ and $n_1 sin\theta = n_2 sin\theta'$. **Thus, the laws of reflection and refraction (Snell's laws) for ray optics are also applicable to wave vectors.** Note that it is impossible to calculate correctly the amplitudes of reflected and refracted waves with scalar wave theory. Therefore one needs to include the vectorial character of light waves, which is discussed in a later chapter.

4.4.2 Paraxial transmission through a thin plate and a thin lens

We consider a thin plate with variable thickness $d(x, y)$ and index $n(x, y)$. A plane wave is incident along the z-axis (see figure 4.11(a)). If we describe the plane wave with Ae^{-jkz}, then the transmitted wave just after the plate is well approximated by

$$A'' e^{-jkn(x,y)d(x,y)}, \tag{4.48}$$

where A'' is smaller than A, because of reflections (multiple reflections are neglected here). Thus, in a plane $z = d_0$ closely behind the plate the wave function is

$$A'' e^{[-jk(n(x,y)d(x,y)+(d_0-d(x,y)))]} = A'' e^{-jkd_0} e^{[-jk(n(x,y)-1)d(x,y)]} \tag{4.49}$$

This means that the deformation of the wave front because of the plate scales with the variation of optical path length of the plate relative to the path length in vacuum.

As an example we apply this to a thin plano-convex lens with spherical surface and radius of curvature R, as depicted in figure 4.11(b). The lens has a thickness

$$d(x, y) = d_0 - \left[R - \sqrt{R^2 - (x^2 + y^2)}\right]. \tag{4.50}$$

For small x and y (paraxial treatment!) this function is approximated by a paraboloid. After transmission of a plane wave through the lens the wave front is not exactly spherical, but it can also be approximated by a paraboloid. According to the Fresnel approximation this approximates a spherically converging wave. It is easy to prove that this wave converges to a point - the focal point - at the same location as the focal point predicted by geometric optics.

4.5 Interference

If two or more waves are present at the same place simultaneously, the superposition principle dictates that the total wave function is equal to the sum of the individual wave functions. When all the waves are monochromatic with the same frequency, we can eliminate the time factor and use the Helmholtz equation for the complex amplitude. Because of the linearity of this equation

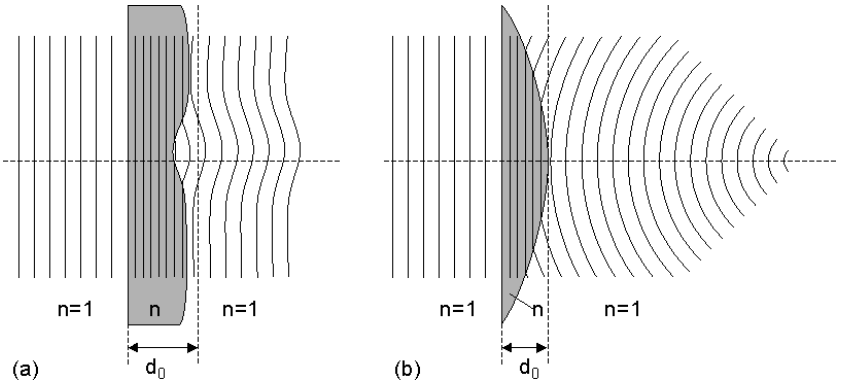

Figure 4.11: (a) Transparent plate with variable thickness. (b) Thin plano-convex lens.

the superposition principle is also applicable to the complex amplitude. From the relation between intensity and complex amplitude we deduce that the intensity of two or more waves is *not* necessarily equal to the sum of the individual intensities. The difference between both is ascribed to interference between the superposed waves. This interference is new and it cannot be explained with ray theory, because it is described by the phase relations of the contributing waves.

4.5.1 Interference between two waves

We consider the superposition of two monochromatic waves with the same frequency ν and complex amplitudes $U_1(\mathbf{r})$ and $U_2(\mathbf{r})$, respectively. This superposition results in a monochromatic wave with the same frequency, but with complex amplitude

$$U(\mathbf{r}) = U_1(\mathbf{r}) + U_2(\mathbf{r}). \tag{4.51}$$

By using equation (4.3) we get the intensity of the individual waves $I_1 = |U_1|^2$ and $I_2 = |U_2|^2$, but the intensity of the total wave is

$$I = |U|^2 = |U_1|^2 + |U_2|^2 + U_1^* U_2 + U_1 U_2^* \tag{4.52}$$

where we have dropped the $\mathbf{r}$-dependence for simplicity. We substitute

$$\begin{aligned} U_1 &= \sqrt{I_1} e^{j\phi_1} \\ U_2 &= \sqrt{I_2} e^{j\phi_2} \end{aligned} \tag{4.53}$$

in (4.52), where ϕ_1 and ϕ_2 are the phases of the two waves, and we obtain

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} cos(\phi) \tag{4.54}$$

with

$$\phi = \phi_2 - \phi_1. \tag{4.55}$$

Equation (4.54) is the interference equation. It is also easily deduced from a phasor diagram as in figure 4.12. This figure clearly shows that the size of the phasor U not only depends on the sizes

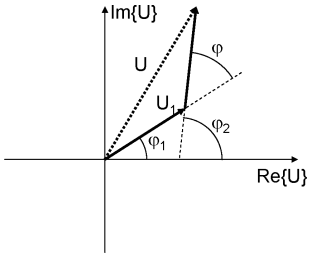

Figure 4.12: Phasor diagram for superposition of two waves with intensities I_1 and I_2, and phase $\phi = \phi_2 - \phi_1$

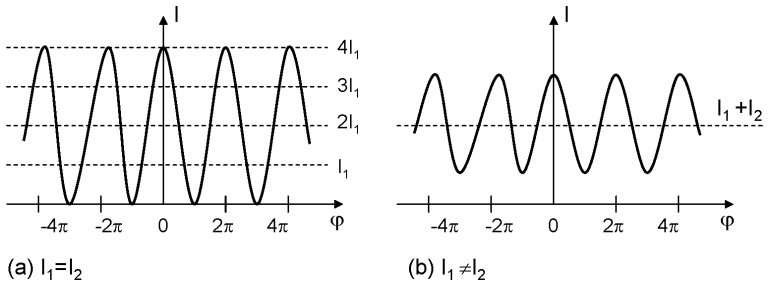

Figure 4.13: Interference between two waves with (a) $I_1 = I_2 = I_0$, (b) $I_1 \neq I_2$.

of the individual phasors, but also on the phase between these phasors. The interference term can be both positive and negative and this is called constructive and destructive interference, respectively.

Assume that $I_1 = I_2 = I_0$, then (4.54) becomes

$$I = 2I_0(1 + cos(\phi)) = 4I_0 cos^2\left(\frac{\phi}{2}\right) \tag{4.56}$$

so that $I = 4I_0$ if $\phi = 0$, and $I = 0$ if $\phi = \pi$. If $\phi = \pi/2$ or $\phi = 3\pi/2$ the interference term disappears, and the total intensity is the sum of the individual intensities $I = 2I_0$. This strong dependence on phase of the intensity allows one to determine phase differences by measuring the intensity.

Remark 1
Interference is caused by the simultaneous action of different waves. In no way does it mean that the waves interact and influence each other. The individual waves remain unchanged, but the total intensity is no longer simply the sum of individual intensities.

Remark 2
With interference the total power varies between 0 and 4 times the power of individual waves, dependent on the phase difference. It is important to realize that interference does not violate the law of conservation of power. It merely means a spatial redistribution of the optical power. Two waves can have the same phase in a certain position, but because of the position dependence of the phase, and thus of the phase difference ϕ, the total intensity at some place will be larger than $I_1 + I_2$ and at other places will be smaller than $I_1 + I_2$.

Remark 3
To observe interference one needs a fixed phase relation between the different waves. Normal

lamps emit light that is not monochromatic at all, but with chaotically varying phase. This leads to fluctuations in both ϕ_1 and ϕ_2, so that the difference ϕ varies quickly and randomly in time. By averaging (see equation (4.3)) the cosine term in equation (4.54) will disappear so that the interference term is absent. This light is called incoherent. Coherence of light is treated in chapter 13. In this chapter we limit ourselves to fully coherent light, and we assume the phases of individual waves to be constant at every position.

Interferometers

Assume two identical plane waves, each with intensity I_0, propagating in the z-direction. One wave is retarded over a distance d, respective to the other wave, so

$$\begin{aligned} U_1 &= \sqrt{I_0} e^{-jkz} \\ U_2 &= \sqrt{I_0} e^{[-jk(z-d)]} \end{aligned} \tag{4.57}$$

Then, the interference of the two waves is determined by substituting $I_1 = I_2 = I_0$ and $\phi = kd = 2\pi d/\lambda$ into the interference equation (4.54)

$$I = 2I_0 \left[1 + cos\left(2\pi \frac{d}{\lambda} \right) \right] \tag{4.58}$$

The dependence of the intensity I on the delay d is illustrated in figure 4.14. If the delay is an integer times the wavelength λ, we get constructive interference and the total intensity is $I = 4I_0$. On the other hand, if d is an odd integer times the half wavelength $\lambda/2$, then we get destructive interference and the total intensity is $I = 0$.

An interferometer uses the above principle. It is an optical instrument that splits a wave into two waves, delays them over an unequal distance, and combines them together to measure the intensity of their superposition. Because of the strong sensitivity of the intensity to the phase difference

$$\phi = 2\pi d/\lambda = 2\pi n d/\lambda_0 = 2\pi n \nu d/c \tag{4.59}$$

with d the difference in propagation distance between the two waves, one can use an interferometer to measure small variations of distance d, index n or wavelength λ_0 (or frequency ν). If $d/\lambda = 10^4$, then an index variation of $\delta n = 10^{-4}$ realizes a phase difference $\delta\phi = 2\pi$. Analogously, the phase changes over 2π, if d increases with a wavelength $\delta d = \lambda$. An increase of the frequency $\delta\nu = c/d$ has the same effect.

Three important examples of interferometers are the Mach-Zehnder interferometer, the Michelson interferometer and the Sagnac interferometer. They are shown in figure 4.14. In a Sagnac interferometer the optical path of the two waves is identical but opposite, so that a rotation of the interferometer results in a phase change proportional to the angular velocity of the rotation. This system is used as a gyroscope.

Interference of two oblique plane waves

Consider the interference between two plane waves with equal intensities, where both waves propagate at an angle θ with the z-axis, see figure 4.15. $U_1 = I_0^{1/2} e^{-j(kcos(\theta)z + ksin(\theta)x)}$ and $U_2 =$

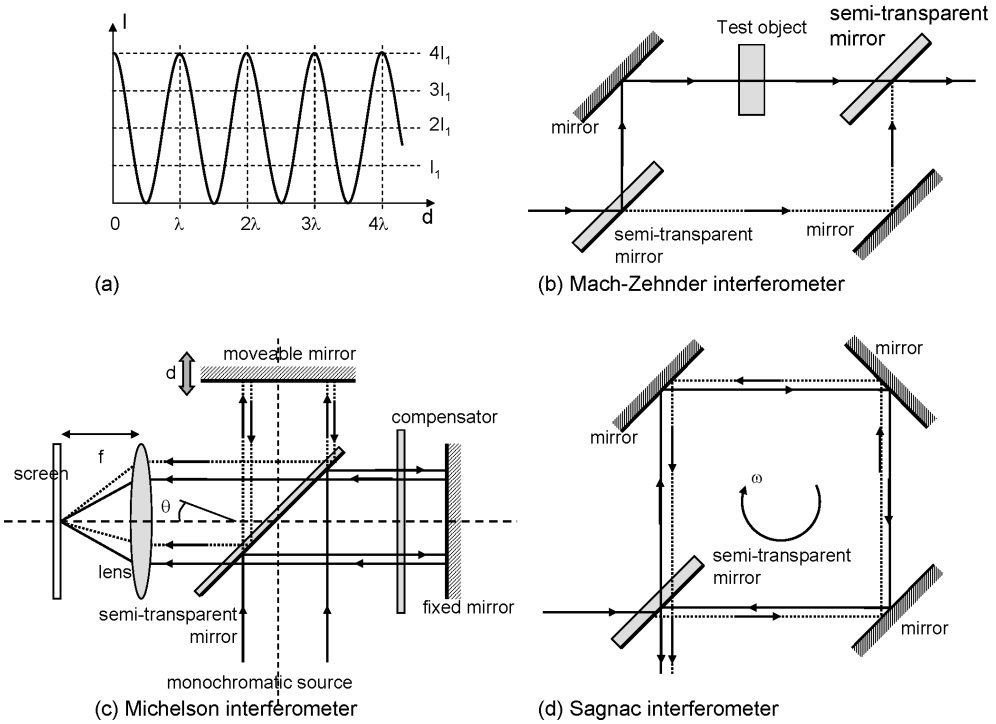

Figure 4.14: (a) Dependence of the intensity I on the delay d (b) Mach-Zehnder interferometer (c) Michelson interferometer (d) Sagnac interferometer.

$I_0^{1/2} e^{-j(kcos(\theta)z - ksin(\theta)x)}$. In the plane $z = 0$ the waves have a phase difference $\phi = 2kxsin(\theta)$, so with equation (4.54)

$$I = 2I_0[1 + cos(2ksin(\theta)x)]. \tag{4.60}$$

Interference creates a pattern that varies sinusoidally with x with a period $2\pi/2ksin(\theta) = \lambda/2sin(\theta)$, see figure 4.15. This effect can be used to create a sine pattern with high resolution to fabricate a diffraction grating. Another application is to determine the angle of an incident wave by superposing it with a reference wave and measuring the interference distribution. This is the basic principle of holography. One should note that in the special case of $\theta = \pi/2$ we find the standing wave pattern caused by the interference of a forward and backward wave. The period of this standing wave pattern is $\lambda/2$ and this is the smallest period an interference pattern can have for a given wavelength.

4.5.2 Interference between multiple waves

When M monochromatic waves with complex amplitudes $U_1, U_2, \ldots, U_M$ and the same frequency are superposed, this results in a monochromatic wave with the same frequency and amplitude $U = U_1 + U_2 + \ldots + U_M$. The intensities of the individual waves $I_1, I_2, \ldots, I_M$ are insufficient to determine the total intensity $I = |U|^2$. The relative phases have a major impact on the total intensity, as the next examples show.

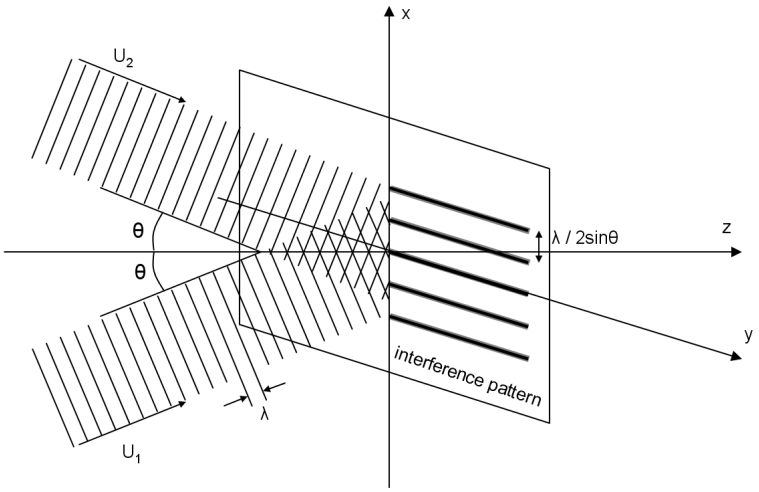

Figure 4.15: Interference between two inclined plane waves.

Interference of M waves with equal amplitudes and phase difference

We assume M waves with complex amplitudes

$$U_m = \sqrt{I_0} e^{[j(m-1)\phi]}, \quad m = 1, 2, \ldots, M. \tag{4.61}$$

The waves have equal intensity I_0 and constant phase difference ϕ between consecutive waves, as illustrated in figure 4.16(a). To derive an expression for the total intensity it is convenient to introduce $h = e^{j\phi}$, so $U = {I_0}^{1/2} h^{m-1}$. The complex amplitude of the total wave becomes

$$\begin{aligned} U &= \sqrt{I_0}\left(1 + h + h^2 + \ldots + h^{M-1}\right) \\ &= \sqrt{I_0}\frac{1 - h^M}{1 - h} \\ &= \sqrt{I_0}\frac{1 - e^{jM\phi}}{1 - e^{j\phi}} \end{aligned} \tag{4.62}$$

and the intensity is

$$I = |U|^2 = I_0 \left| \frac{e^{-jM\phi/2} - e^{jM\phi/2}}{e^{-j\phi/2} - e^{j\phi/2}} \right|^2 \tag{4.63}$$

so that

$$I = I_0 \frac{sin^2\left(M\phi/2\right)}{sin^2\left(\phi/2.\right)} \tag{4.64}$$

It is clear from figure 4.16(b) that the intensity I strongly depends on the phase difference ϕ.

- If $\phi = 2\pi q$, with q an integer, all phasors are aligned and the intensity reaches a peak $I = M^2 I_0$. The average intensity (averaged over a uniform ϕ distribution) is $\bar{I} = (1/2\pi)\int_0^{2\pi} I d\phi = M I_0$, which is the intensity without interference. Thus the peak intensity is M times larger than the average intensity, and the larger the number of waves M the more pronounced the effect is (compare figure 4.16(b) with 4.13).

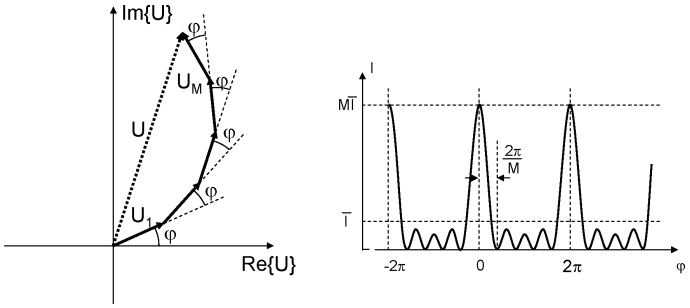

Figure 4.16: (a) the sum of M phasors with equal amplitude and equal phase difference (b) intensity I in function of phase difference ϕ.

- For a phase difference slightly off $2\pi q$, we get a steep decline in intensity I.
- If the phase difference is $2\pi/M$, the intensity becomes zero.

This example of interference between M waves is common in practice. Probably the most well-known case is the illumination of a screen through M slits by a plane wave. The diffracted field depicts the behavior described above, in function of the angle.

Interference of an infinite number of waves with progressively declining amplitude and equal phase difference

$$U_1 = \sqrt{I_0}, \quad U_2 = hU_1, \quad U_3 = hU_2 = h^2U_1, \quad \ldots \tag{4.65}$$

with $h = |h|\, e^{j\phi}$, $|h| < 1$ and I_0 the intensity of the initial wave. The phasor diagram is shown in figure 4.17. The superposition of all these waves has complex amplitude

$$\begin{aligned} U &= U_1 + U_2 + U_3 + \ldots \\ &= \sqrt{I_0}(1 + h + h^2 + \ldots) \\ &= \frac{\sqrt{I_0}}{1-h} \\ &= \frac{\sqrt{I_0}}{1 - |h|\, e^{j\phi}}. \end{aligned} \tag{4.66}$$

The intensity $I = |U|^2 = I_0/\left|1 - |h|\, e^{j\phi}\right|^2 = I_0/[(1 - |h|\, cos(\phi))^2 + |h|^2\, sin^2(\phi)]$ so

$$I = \frac{I_0}{(1-|h|)^2 + 4\,|h|\, sin^2(\phi/2)}. \tag{4.67}$$

The previous equation is often written as

$$I = \frac{I_{max}}{1 + \left(\frac{2\mathfrak{F}}{\pi}\right)^2 sin^2\left(\frac{\phi}{2}\right)} \tag{4.68}$$

with

$$I_{max} = \frac{I_0}{(1-|h|)^2} \tag{4.69}$$

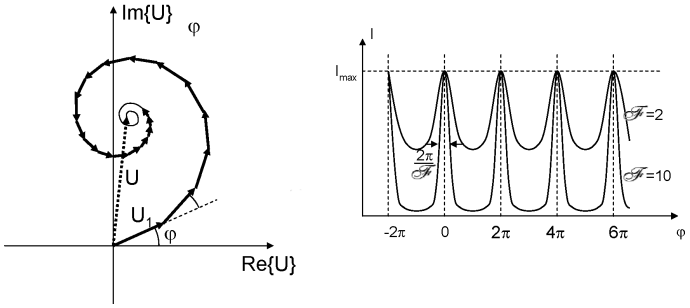

Figure 4.17: (a) the sum of M phasors with progressively declining amplitude and equal phase difference (b) intensity I in function of phase difference ϕ

and

$$\mathfrak{F} = \frac{\pi \left|h\right|^{\frac{1}{2}}}{1 - \left|h\right|} \tag{4.70}$$

a parameter called finesse.

As illustrated in figure 4.17 the intensity is a periodic function of ϕ with period 2π. It reaches the maximum I_{max} for $\phi = 2\pi q$, with q an integer. For this ϕ all phasors are aligned. When the finesse $\mathfrak{F}$ is large (so $|h|$ is close to one), the function I is sharply peaked. As the finesse $\mathfrak{F}$ decreases the peaks become less sharp and they disappear when $|h| = 0$.

As an example consider a value of ϕ close to the peak $\phi = 0$. For $|\phi| \ll 1$ one obtains $sin\phi/2 \approx \phi/2$ and equation (4.68) is approximated by

$$I \approx \frac{I_{max}}{1 + (\mathfrak{F}/\pi)^2 \phi^2}. \tag{4.71}$$

The intensity I decreases to half of its peak value when $\phi = \pi/\mathfrak{F}$, so the *Full Width at Half Maximum* (FWHM) of the peak is equal to

$$\delta\phi = \frac{2\pi}{\mathfrak{F}}. \tag{4.72}$$

If $\mathfrak{F} \gg 1$, then $\delta\phi \ll 2\phi$ and the assumption $\phi \ll 1$ is correct. Thus, the finesse $\mathfrak{F}$ is the ratio between the period 2π of the peaks and the FWHM of the interference pattern. So, $\mathfrak{F}$ is a measure for the sharpness of the interference function, or for the sensitivity of the intensity to phase deviations from the peak values $2\pi q$.

This example is especially relevant in practice. In particular for the Fabry-Perot interferometer, that consists of two parallel semi-transparent mirrors. The total transmission is realized by an infinite number of contributions from the multiple reflections between the mirrors.

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Chapter 5

Gaussian Beam Optics

Contents

In wave optics the wave functions in free space satisfy the Helmholtz equation $\nabla^2\phi + k^2\phi = 0$. The plane waves and the spherical waves are examples of solutions that 'oppose' each other. The plane wave has one direction but extends over the entire space. Whereas the spherical wave originates from one point but propagates in all directions. In this chapter we examine solutions that lie between these two extremes. They are finite both in space and direction.

The wave character of light prohibits that a beam with finite cross section propagates in free space without spreading. A perfectly collimated beam would have many practical applications. However, there are solutions of the Helmholtz equation that approximate this behavior. In this chapter we will study Gaussian beams. At the origin they exhibit the character of a plane wave, but at a distance they behave as a spherical wave. A laser beam is often a good approximation of a Gaussian beam.

5.1 Diffraction of a Gaussian light beam

Consider a monochromatic beam with a finite cross section propagating along the z-direction, as depicted in figure 5.1. Here the beam is represented as a number of arrows, of which the length indicates the local amplitude. Because of the wave character of light this beam will fan out. We are going to analyze this phenomenon in this section.

We assume the beam has a paraxial behavior around the z-axis, so we can employ the paraxial Helmholtz equation. The field $U(\mathbf{r})$ is written as

$$U(\mathbf{r}) = A(\mathbf{r})e^{-jkz} \tag{5.1}$$

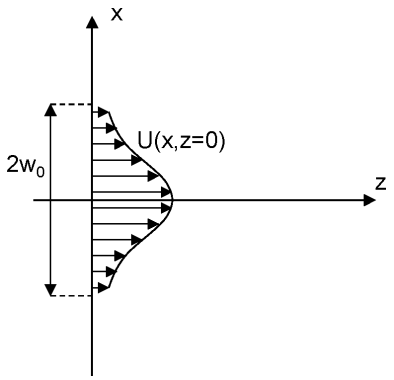

Figure 5.1: Gaussian beam profile.

with

$$\nabla_T^2 A(\mathbf{r}) - 2jk\frac{\partial A(\mathbf{r})}{\partial z} = 0. \tag{5.2}$$

We are looking for a solution of this equation so that the amplitude function U has a Gaussian amplitude profile and a plane phase front at $z = 0$:

$$U(x, y, 0) = A(x, y, 0) = e^{-\frac{x^2+y^2}{w_0^2}} = e^{-\frac{\rho^2}{w_0^2}},\ \rho^2 = x^2 + y^2 \tag{5.3}$$

w_0 is half of the $1/e$ width of the Gaussian profile (thus the $1/e^2$ width of the intensity). In a three-dimensional situation with a circular Gaussian beam 86% of the power propagates in a circle with radius w_0. For the solution at $z \neq 0$ we will notice that the function:

$$A(\mathbf{r}) = e^{-j\left[p(z)+k\frac{\rho^2}{2q(z)}\right]} \tag{5.4}$$

satisfies the paraxial Helmholtz equation. This function keeps a Gaussian amplitude profile during propagation. This is an important characteristic: a Gaussian beam is one of the few profiles that maintains its function profile during propagation — except for a widening. Here $p(z)$ can be considered as a complex phase shift along the z-axis, while $1/q(z)$ corresponds to the phase curvature (with respect to the real part of $1/q$) and the amplitude profile (with respect to the imaginary part) in the transversal plane. Substitution of (5.4) in (5.2) obtains:

$$2k\left(\frac{dp}{dz} + \frac{j}{q}\right) + \left(\frac{k\rho}{q}\right)^2\left(1 - \frac{dq}{dz}\right) = 0. \tag{5.5}$$

This has to hold for all x and z, so:

$$\frac{dq}{dz} = 1 \tag{5.6}$$

$$\frac{dp}{dz} = \frac{-j}{q} \tag{5.7}$$

At $z = 0$ the equations (5.4) and (5.3) have to be equal. This leads to the boundary conditions:

$$q(0) = j\frac{kw_0^2}{2} \tag{5.8}$$

$$p(0) = 0 \tag{5.9}$$

Integration of (5.6) gives:

$$q(z) = z + \frac{jkw_0^2}{2}. \tag{5.10}$$

We split $1/q(z)$ into its real and imaginary part:

$$\frac{1}{q(z)} = \frac{1}{R(z)} - \frac{2j}{kw(z)^2} \tag{5.11}$$

Within the paraxial approximation $R(z)$ represents the radius of curvature of the phase front (a quadratic front is approximated by a spherical front close to the z-axis), while $w(z)$ indicates the half width of the Gaussian beam at each location z. Equalizing the real and imaginary parts in the previous two expressions for $q(z)$ leads to:

$$R(z) = z\left(1 + \frac{b_0^2}{z^2}\right) \tag{5.12}$$

$$w(z) = w_0\sqrt{1 + \frac{z^2}{b_0^2}} \tag{5.13}$$

with

$$b_0 = \frac{kw_0^2}{2}. \tag{5.14}$$

After integration we get for $p(z)$:

$$jp(z) = -ln\left[\frac{w_0}{w(z)}\right] - j\ arctan\frac{z}{b_0} \tag{5.15}$$

so the entire function $U(\mathbf{r})$ finally becomes:

$$U(\mathbf{r}) = \frac{w_0}{w(z)} e^{-\frac{\rho^2}{w(z)^2}} e^{-j\frac{k\rho^2}{2R(z)}} e^{j\ arctan\frac{z}{b_0}} e^{-jkz} \tag{5.16}$$

Let us analyze the behavior of the radius of curvature $R(z)$ and the half width $w(z)$ in function of z. At $z = 0$ the radius of curvature is infinite (corresponding to our boundary condition), thus we have a kind of plane wave with finite width. For very large z we have $R = z$, so this approximates a spherically expanding wave from the origin. In between $R(z)$ reaches a minimum (see figure 5.2):

$$R(z)_{min} = 2b_0 \text{ for } z = b_0. \tag{5.17}$$

This means the center of the sphere is located at $z = -b_0$. Notice also that for all z the radius of curvature is larger than or equal to z. Thus, the center is always to the left of the origin at:

$$z_{centrum} = z - R(z) = -\frac{b_0^2}{z} \tag{5.18}$$

Concerning the half width $w(z)$ we note that it always increases from the minimum w_0 at the origin. This minimum is called the 'waist' of the Gaussian beam.

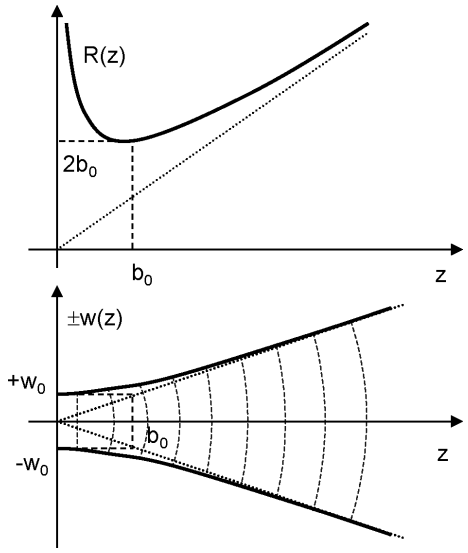

Figure 5.2: The Gaussian beam, radius of curvature and width.

Figure 5.2 shows the evolution of the radius of curvature and the width (one can consider this figure as being the contour line of the 2D amplitude profile). The hyperboles have an asymptote with an angle given by:

$$\theta = \pm arctan\frac{w_0}{b_0} \approx \pm\frac{w_0}{b_0} = \pm\frac{2}{kw_0} = \pm\frac{\lambda}{\pi w_0} \tag{5.19}$$

This is probably the most elementary result concerning the diffraction of waves in free space. The angle along which the wave spreads is inversely proportional to the width of this wave, scaled with the wavelength. In a first approximation, we now see that the Gaussian beam consists of two parts (figure 5.2). First, it will propagate over a distance b_0 with a quasi constant diameter (more precisely: the width increases with a factor $\sqrt{2}$). Subsequently, it will fan out spherically with angle θ. The distance b_0 is called the *Rayleigh range* (after Lord Raleigh), and the angle θ is the *beam divergence angle*.

There is an alternative — and extremely elegant — way to derive the Gaussian beam expressions. In chapter 4 we saw that the function

$$A(\mathbf{r}) = \frac{1}{z}e^{-jk\frac{\rho^2}{2z}} \tag{5.20}$$

represents the parabolic approximation of a spherical wave propagating from the origin. If we substitute z by $q(z) = z - z_1$,

$$A(\mathbf{r}) = \frac{1}{q(z)}e^{-jk\frac{\rho^2}{2q(z)}} \tag{5.21}$$

then this approximates a spherical wave departing from the point $z = z_1$. However, if we assume z_1 to be imaginary, $z_1 = -jb_0$, then the solution 5.21 is also a correct solution of the paraxial Helmholtz equation. This expression has a totally different character compared to the one with real z_1. Surprisingly, it corresponds to the Gaussian beam with b_0 the Rayleigh range and $\sqrt{\lambda b_0/\pi}$ the width at $z = 0$.

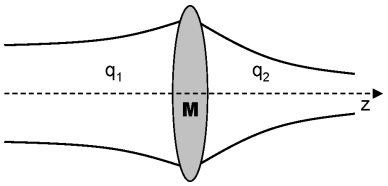

Figure 5.3: Gaussian beam incident on a lens.

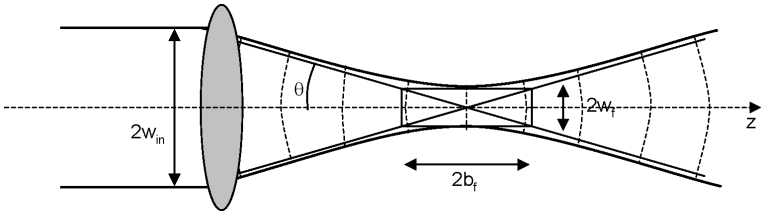

Figure 5.4: Waist of a focused laser beam.

5.2 Gaussian beams in lens systems

We can use the Gaussian beam theory to analyze the behavior of lens systems for coherent fields. It turns out we can use the paraxial matrix theory again (see chapter 3). If a Gaussian beam passes through a thin spherical lens, one expects that only the phase curvature changes slightly, while the beam remains Gaussian (figure 5.3). More generally one can prove that a Gaussian beam with q-value q_1 perpendicularly incident on a lens system with system matrix:

$$\mathbf{M} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{5.22}$$

obtains a q-value at the exit side of the lens given by:

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D}. \tag{5.23}$$

We can use the previous to determine the spot size if we focus a Gaussian beam with an (ideal, aberration free) lens (figure 5.4). w_{in} is the half width of the incident Gaussian beam. After propagating through the lens the beam converges with (half) angle:

$$\theta = \frac{w_{in}}{f}. \tag{5.24}$$

From (5.19) we know that this angle is correlated with the half width w_f in the focal plane according to:

$$\theta = \frac{\lambda}{\pi w_f}, \tag{5.25}$$

so that

$$w_f = \frac{\lambda f}{\pi w_{in}}. \tag{5.26}$$

Thus it is possible to obtain a small spot only if the focal distance is small (strong refraction). When the Gaussian beam has a width almost equal to the size of the lens, one can write alternatively:

$$\theta = NA = \frac{\lambda}{\pi w_f}, \tag{5.27}$$

$$2w_f = \frac{2}{\pi}\frac{\lambda}{NA}. \tag{5.28}$$

Here NA ($= sin\theta$) is approximated paraxially as θ. We notice that in the best case (large NA and ideal lens) a focused beam is never smaller than (approximately) the wavelength.

The previous leads to an expression for the depth of field (*Rayleigh range*) in the focal plane. It is given by:

$$2b_f = kw_f^2. \tag{5.29}$$

Thus, unfortunately a small spot correlates with a small depth of field. If a lens has aberrations, the focusing properties will be worse than described here. The behavior given by formula (5.28) – ideal incident beam and ideal lens – is called *diffraction limited*.

Data storage on a CD

An application of coherent laser light is the data-reading from a CD or CD-ROM (figure 5.5), where the bit density is close to the diffraction limit. The information is stored as a series of small pits about 150nm deep. Assuming one CD holds 750Megabytes, and has a surface area of $\pi(5.8^2 - 2.5^2) = 86$ cm^2 (diameter 11.6cm and an unused center hole of 5cm), we can easily calculate the bit density. Because of error correction the stored data is about 2 times the given data (17 physical bits to code 8 information bits). On a higher level there is another error correcting code with a factor 1.5 bit increase (36 bits for 24 data bits). Therefore a CD fits

$$750\text{MB} \times \frac{36}{24} \times \frac{17}{8} = 1.9\ 10^{10}\text{bits}. \tag{5.30}$$

This makes the physical bit density

$$\text{bit density} = \frac{1.9\ 10^{10} bits}{86\ \text{cm}^2} = 2.2\frac{bits}{\mu m^2}. \tag{5.31}$$

In practice a bit on a track has a length of 0.28μm, and the tracks are spaced by 1.6μm, which indeed means 2.2 bits per μm^2.

The theoretical maximal bit density is determined by the diffraction limit. The wavelength of the laser diode is about 780nm. According to diffraction theory we are able to store 1 bit only per λ^2 surface area. Thus, the theoretical bit density is $\frac{1}{\lambda^2}$ ($\approx 2.1\frac{bit}{\mu m^2}$). We notice that the bit density slightly exceeds the theoretical limit. This is possible because coding ensures that the minimal size of a pit on a CD is at least 3 bits.

Watching the stars

In this section we saw that a Gaussian beam can be focused to a spot with a waist width that is determined by the ratio of the wavelength λ of the beam and the numerical aperture of the lens system. Instead of focusing a Gaussian beam, one can also consider the case where a point source is imaged by a lens system. In this case, a spherical wave originates from the point source. Due to the finite transversal dimensions of the lens system only a fraction of the

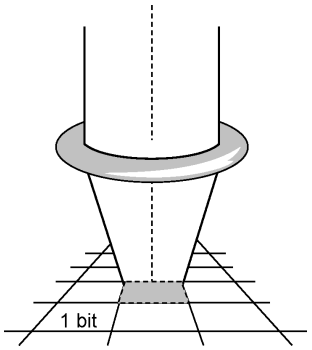

Figure 5.5: Focus spot of a CD-reader.

spherical wavefront is converted into a converging wavefront. Due to the finite dimensions of this converging wavefront, this can be in first order also be approximated by a gaussian beam, showing that a point source cannot be perfectly imaged onto a point, but instead onto a spot with lateral dimensions determined by the ratio of the wavelength and numerical aperture. Due to the fact that this spot has finite transverse dimensions, two point sources that are too close to each other in the object plane, will be indistinguishable in the image plane. Thereby the resolution of the imaging system is limited. In the case where the source has a broad spectrum, the longest wavelengths of the spectrum (and the numerical aperture of the lens system) will determine how well two point sources can be resolved.

A telescope (figure 5.6) is a fine example of gaussian beams in a lens system. In its simplest form, a telescope consists of two lenses: an objective and an eyepiece, with an intermediate real image between both (in the focal plane of both the objective and of the eyepiece).A star acts like a point source at infinity and will focus in that image plane with a spot size which can be approximated by the Gaussian formulas and hence is given by (5.28). Therefore two neighbouring stars can only be resolved if their image spots do not overlap too much. If we take as a criterion that the $1/e$ circles of the two spots should not overlap we find as a criterion for the minimum distance d_{min} between the spot centres:

$$d_{min} = 2w_f = 2\frac{\lambda}{\pi NA} = 0.64\frac{\lambda}{NA} \tag{5.32}$$

This minimum distance can be readily translated into a minimum angular separation of the two stars, leading to:

$$\Delta\alpha \approx 0.64\frac{\lambda/NA}{f_1} \approx 1.28\frac{\lambda}{D} \tag{5.33}$$

Hence we can see that the angular resolution of the telescope is determined by the diameter of the objective lens, and only by this diameter (if we assume that the lenses are free of aberrations). This explains why space telescopes are made as big as possible. A second reason for this is of course that a larger telescope objective collects a larger amount of light, and therefore one can see weaker stars.

5.3 Hermite-Gaussian beams

The Gaussian beam is not the only solution to equation (5.2) that keeps its form during propagation. It can be proven that there are higher order solutions with this property. They have the

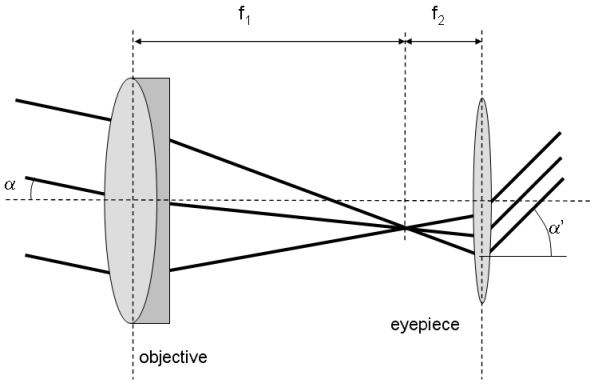

Figure 5.6: Principle of a simple telescope.

form:

$$\Psi_{lm}(x,y,z) = \frac{w_0}{w(z)} H_l\left(\frac{\sqrt{2}x}{w(z)}\right) H_m\left(\frac{\sqrt{2}y}{w(z)}\right) e^{-\frac{\rho^2}{w^2(z)}} e^{-jkz} e^{j(l+m+1)arctan\frac{z}{b_0}} e^{-j\frac{k\rho^2}{2R(z)}} \tag{5.34}$$

with $H_l(s)$ and $H_m(s)$ the Hermite polynomials:

$$\begin{aligned} H_0(s) &= 1 \\ H_1(s) &= 2s \\ H_2(s) &= 4s^2 - 2 \\ H_3(s) &= 8s^3 - 12s \\ &\cdots \\ H_{l+1}(s) &= 2sH_l(s) - 2lH_{l-1}(s) \end{aligned}$$

These solutions are called Hermite-Gaussian beams. The 0th order mode $(l = 0, m = 0)$ is the Gaussian beam. The optical intensity of the (l, m) Hermite-Gaussian beams is

$$I_{l,m}(x,y,z) = \left[\frac{w_0}{w(z)}\right]^2 H_l\left(\frac{\sqrt{2}x}{w(z)}\right)^2 H_m\left(\frac{\sqrt{2}y}{w(z)}\right)^2 e^{\frac{-2(x^2+y^2)}{w^2(z)}} \tag{5.35}$$

Figure 5.7 shows some of these modes. These show that the diameter of the beams increases for higher order solutions.

5.4 M^2 factor

Until now, we studied particular solutions of the paraxial Helmholtz equation, solutions that kept their form during propagation. There are of course an infinite number of beam solutions: one can use any function as a starting field.

For all these beams one will obtain that a beam with a finite size fans out because of diffraction. The Gaussian beam has the special property that for a given waist, it spreads out with the minimal

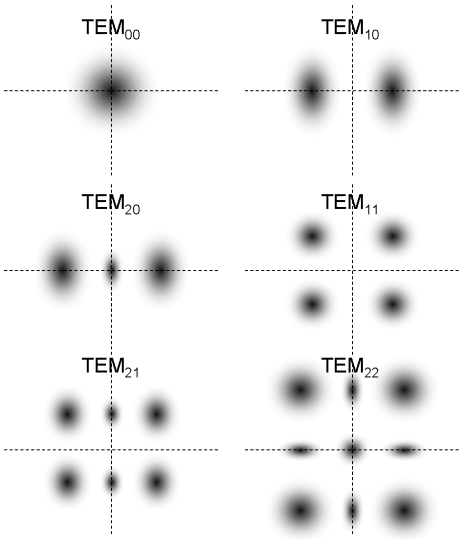

Figure 5.7: The intensity distribution in the transversal plane of some Hermite-Gaussian beams.

angle, given by equation 5.19. All other solutions (for the given waist) will diffract with a larger angle.

In this regard the M^2-factor is defined. This number expresses the speed of spreading of a certain beam, compared with a Gaussian beam with the same width. Thus the definition is:

$$M^2 = \pi\theta\frac{w_0}{\lambda}, \tag{5.36}$$

with θ the (half) divergence angle of the beam. For a Gaussian beam M^2 equals one, and that is the smallest possible value. M^2 is often used as a quality norm for laser beams.

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Chapter 6

Electromagnetic Optics

Contents

6.1 Introduction

Light is an electromagnetic wave phenomenon that is described by the same theoretical principles used for all electromagnetic radiation. Light or optical radiation (or optical frequencies) are all frequencies between infrared, visible and ultraviolet light, so all wavelengths (roughly) between $10nm$ and $1mm$. Propagation of electromagnetic radiation is expressed by two coupled symmetrical partial differential equations, coupling the electric field vector with the magnetic field vector. These equations were originally formulated by James Clark Maxwell in 1864. Maxwell's theory was not only a breakthrough in physics because it was the first example of unification (magnetism and electricity, at first sight separate phenomena, appeared to be fundamentally linked), but also because it led Einstein directly to his theory of relativity. From Maxwell's laws it follows that the speed of light is always 299792458 m/s. However, according to classical physics velocities can be added, so a light ray emitted by a fast object would have a speed larger than 299792458 m/s. This paradox made Einstein think, resulting in his famous theory of relativity.

The scalar wave optics theory discussed in chapter 4 is an approximation of Maxwell's equations, because light is described by one single scalar wave equation. This single scalar equation is sufficient for the paraxial approximation with certain conditions (explained later). By performing

another approximation, the short wavelength limit, we already arrived at geometrical optics, see chapter 3.

In this chapter we present a short overview of the important aspects of electromagnetic theory for optics. We start from Maxwell's equations and discuss some elementary waves. Then we describe properties of dielectric media. These two sections form the postulates of electromagnetic optics: a set of rules for the next sections. Furthermore we discuss polarization, absorption and dispersion, and the laws of reflection and refraction. To conclude a few layered structures are examined.

6.2 Maxwell's electromagnetic wave equations

The electric and magnetic field vectors $\mathbf{E}(\mathbf{r},\mathbf{t})$ (unit: V/m) and $\mathbf{H}(\mathbf{r},\mathbf{t})$ (unit: A/m) in a medium without free charges or currents, satisfy the following coupled partial differential equations which are function of space $\mathbf{r}$ and time t: Maxwell's equations.

$$\begin{aligned} \nabla \times \mathbf{H} &= \frac{\partial \mathbf{D}}{\partial t} \\ \nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \cdot \mathbf{D} &= 0 \\ \nabla \cdot \mathbf{B} &= 0 \end{aligned} \tag{6.1}$$

The vector fields $\mathbf{D}(\mathbf{r},t)$ (unit: C/m^2) and $\mathbf{B}(\mathbf{r},t)$ (unit: Wb/m^2) are the electric flux density (also called electric displacement vector or electrical induction) and the magnetic flux density (also called magnetic induction), respectively. The relation between $\mathbf{D}$ and $\mathbf{E}$ depends on the electrical properties of the medium. Analogously, the relation between $\mathbf{B}$ and $\mathbf{H}$ depends on the magnetic properties. They form the constitutive relations:

$$\begin{aligned} \mathbf{D} &= \epsilon_0 \mathbf{E} + \mathbf{P} \\ \mathbf{B} &= \mu_0 \mathbf{H} + \mu_0 \mathbf{M} \end{aligned} \tag{6.2}$$

The constants $\mu_o = 4\pi 10^{-7} H/m$ and $\epsilon_0 = \frac{1}{c^2 \mu_0} F/m$ are the permeability and the permittivity of the vacuum, respectively. $\mathbf{P}$ (unit: C/m^2) is the polarization density, and $\mathbf{M}$ (unit: A/m) is the magnetization density. In a dielectric medium the polarization density $\mathbf{P}$ is equal to the macroscopic sum of the electric dipole moments induced by the electric field. An analogous definition can be given for $\mathbf{M}$. Further on we will see that the fields $\mathbf{P}$ and $\mathbf{M}$ are related to $\mathbf{E}$ and $\mathbf{H}$, respectively, by relations dependent on the electrical and magnetic properties of the material.

In free space (= non-electrical and non-magnetic) we have: $\mathbf{P} = \mathbf{M} = 0$, so $\mathbf{D} = \epsilon_0 \mathbf{E}$ and $\mathbf{B} = \mu_0 \mathbf{H}$. Notice that in this case the Maxwell equations are reduced and decoupled to the scalar wave equation for all three vector components, because the permittivity or refractive index is constant.

6.2.1 Poynting vector and energy density

The current of electromagnetic energy (unit: W/m^2) is given by the vector:

$$\mathbb{P} = \mathbf{E} \times \mathbf{H} \tag{6.3}$$

known as the Poynting vector. The power follows the direction of this vector, that is perpendicular to both $\mathbf{E}$ and $\mathbf{H}$. The optical intensity[1] I, which is the power per surface area perpendicular to $\mathbb{P}$, is equal to the magnitude of the Poynting vector, averaged over a certain time, see section 4.1.2.

The energy density (unit: J/m^3) associated with an electromagnetic wave is given by

$$U = (\mathbf{E} \cdot \mathbf{D} + \mathbf{H} \cdot \mathbf{B})/2 \tag{6.4}$$

The first and second term represent the energy carried by the electric field and the magnetic field respectively.

6.3 Dielectric media

It is convenient to view the medium equation (eq. 6.2) between $\mathbf{E}$ and $\mathbf{P}$ as a system where the medium responds to an applied electric field $\mathbf{E}$ (input) and creates a polarization density $\mathbf{P}$ as output or response. We give a few definitions relating to dielectric media. A dielectric is:

- **Linear**: If $\mathbf{P}(\mathbf{r}, t)$ is linearly related to $\mathbf{E}(\mathbf{r}, t)$. Then the superposition principle applies.
- **Homogeneous**: If the relation between $\mathbf{P}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}, t)$ is independent of position $\mathbf{r}$.
- **Isotropic**: If the relation between $\mathbf{P}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}, t)$ is independent of the direction of $\mathbf{E}(\mathbf{r}, t)$, so the medium looks the same from all directions. Then, the vectors $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{P}(\mathbf{r}, t)$ have to be parallel.
- **Non-dispersive**: If the material response is instantaneous, so that $\mathbf{P}(\mathbf{r}, t)$ at a time t is determined by $\mathbf{E}(\mathbf{r}, t)$ at the same time t, and not by values of $\mathbf{E}(\mathbf{r}, t)$ at previous times. It is clear that this is an idealization, because an instantaneous response is physically impossible.
- **Spatially non-dispersive**: If the relation between $\mathbf{P}(\mathbf{r}, t)$ and $\mathbf{E}(\mathbf{r}, t)$ is local; if $\mathbf{P}(\mathbf{r}, t)$ at location $\mathbf{r}$ is only influenced by $\mathbf{E}(\mathbf{r}, t)$ at the same position $\mathbf{r}$. In this chapter we assume that all media are spatially non-dispersive.

6.3.1 Homogeneous, linear, non-dispersive and isotropic media

In this chapter we will use non-magnetic materials ($\mathbf{M} = 0$) without free electrical charges or currents. In addition, if the medium is linear, non-dispersive, homogeneous and isotropic we get:

$$\mathbf{P} = \epsilon_0 \chi \mathbf{E}. \tag{6.5}$$

Here the scalar constant χ is the electric susceptibility. It follows that $\mathbf{P}$ and $\mathbf{E}$ are parallel at each position and time, just like $\mathbf{D}$ and $\mathbf{E}$:

$$\mathbf{D} = \epsilon \mathbf{E} \tag{6.6}$$

[1] The use of the term 'Intensity' is a mess in optics. The term is used on the one hand for optical power density (W/m^2), but also for electric field energy density (J/m^3). To make matters worse the term is widely used in radiometry and photometry to denote radiant intensity (W/str) or luminous intensity ($Candela$). In all cases it has 'something' to do with power or energy.

with

$$\epsilon = \epsilon_0(1+\chi) \tag{6.7}$$

The scalar constant ϵ is the electrical permittivity of the medium. With the previous conditions the Maxwell equations reduce to:

$$\begin{aligned} \nabla \times \mathbf{H} &= \epsilon \frac{\partial \mathbf{E}}{\partial t} \\ \nabla \times \mathbf{E} &= -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \\ \nabla \cdot \mathbf{E} &= 0 \\ \nabla \cdot \mathbf{H} &= 0 \end{aligned} \tag{6.8}$$

Note that the equations are reduced and decoupled to the scalar wave equation for each of the three components of $\mathbf{E}$ and $\mathbf{H}$:

$$\nabla^2 u - \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} = 0 \quad \text{with} \quad v^2 = \frac{1}{\epsilon\mu_0} \tag{6.9}$$

The components of the electric and magnetic field propagate in the medium with velocity v, according to:

$$v = \frac{c}{n} \tag{6.10}$$

$$n = \sqrt{\frac{\epsilon}{\epsilon_0}} = \sqrt{(1+\chi)} \tag{6.11}$$

with c the speed of light in free space. The constant n is equal to the ratio of the speed of light in free space to the speed in the medium. It is called the refractive index of the material.

Boundary conditions at an interface The boundary conditions at an interface between two linear, isotropic, homogeneous and non-magnetic media with dielectric constants ϵ_1 and ϵ_2, are important. We get:

$$\mathbf{n} \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \tag{6.12}$$

$$\mathbf{n} \times (\mathbf{H}_1 - \mathbf{H}_2) = 0 \tag{6.13}$$

$$\mathbf{n} \cdot (\epsilon_1\mathbf{E}_1 - \epsilon_2\mathbf{E}_2) = 0 \tag{6.14}$$

$$\mathbf{n} \cdot (\mathbf{B}_1 - \mathbf{B}_2) = 0 \tag{6.15}$$

The tangential components of the electric and magnetic field, and the normal component of the magnetic field, are continuous. The normal component of the electric field makes a discontinuous jump.

6.3.2 Inhomogeneous, linear, non-dispersive and isotropic media

In a non-homogeneous medium the electrical susceptibility, the dielectric constant and thus refractive index are a function of the position $\mathbf{r}$. An example of a non-homogeneous medium is a

graded index medium. One can prove (by using $\nabla\times$ on the Maxwell equations) that the scalar wave equation of eq. (6.9) obtains an extra term:

$$\nabla^2\mathbf{E} - \frac{1}{c^2\left(\mathbf{r}\right)}\frac{\partial^2\mathbf{E}}{\partial t^2} + \nabla\left(\frac{1}{\epsilon\left(\mathbf{r}\right)}\nabla\epsilon\left(\mathbf{r}\right).\mathbf{E}\right) = 0 \tag{6.16}$$

Notice that the location dependent refractive index results in a location dependent speed of the wave in the medium.

For locally homogeneous media, so $\epsilon(\mathbf{r})$ varies slowly in space, the third term on the left side can be neglected.

6.3.3 Dispersive media

In dispersive media $\mathbf{E}$ will create $\mathbf{P}$ by inducing oscillations of bound electrons in atoms of the medium, so they can collectively and with a certain retardation build up a polarization density.

Because we assume a linear medium, an arbitrary electric field will induce a polarization density $P(t)$composed of the superposition of all $\mathbf{E}(t')$ with $t' < t$, or:

$$\mathbf{P}\left(t\right) = \epsilon_0 \int_{-\infty}^{+\infty} \chi\left(t - t'\right)\mathbf{E}\left(t'\right)dt' \tag{6.17}$$

which is a convolution integral, with $\epsilon_0\chi(t)$ the polarization density response to an impulse of electric field.

6.4 Elementary electromagnetic waves

6.4.1 Monochromatic electromagnetic waves

A monochromatic plane wave is a wave where all components of the electric and magnetic field are harmonic functions in time with the same frequency. To simplify notations these components are presented with their complex amplitudes, as in section 4.2

$$\mathbf{E}(\mathbf{r},t) = \mathrm{Re}\left\{\mathbf{E}(\mathbf{r})e^{j\omega t}\right\} \tag{6.18}$$

$$\mathbf{H}(\mathbf{r},t) = \mathrm{Re}\left\{\mathbf{H}(\mathbf{r})e^{j\omega t}\right\} \tag{6.19}$$

Here $\mathbf{E}(\mathbf{r})$ and $\mathbf{H}(\mathbf{r})$ are the complex amplitudes of the electric and magnetic field. In the same way the complex amplitudes of $\mathbf{P}(\mathbf{r},t)$, $\mathbf{D}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$ are denoted as: $\mathbf{P}(\mathbf{r})$, $\mathbf{D}(\mathbf{r})$ and $\mathbf{B}(\mathbf{r})$. Maxwell's equations (for linear, non-dispersive, homogeneous and isotropic media) for monochromatic waves are derived by substitution of the complex amplitudes in (6.8). If we also perform the

substitution: $\partial/\partial t = j\omega$ we obtain:

$$\begin{aligned}
\nabla \times \mathbf{H} &= j\omega\epsilon\mathbf{E} && (6.20)\\
\nabla \times \mathbf{E} &= -j\omega\mu_0\mathbf{H} && (6.21)\\
\nabla \cdot \mathbf{E} &= 0 && (6.22)\\
\nabla \cdot \mathbf{H} &= 0 && (6.23)\\
P(\mathbf{r}) &= \epsilon_0\chi(\mathbf{r},\omega)\mathbf{E}(\mathbf{r}) && (6.24)\\
\epsilon(\mathbf{r},\omega) &= \epsilon_0(1+\chi(\mathbf{r},\omega)) && (6.25)\\
n(\mathbf{r},\omega) &= \sqrt{(}\frac{\epsilon(\mathbf{r},\omega)}{\epsilon_0}) && (6.26)\\
& && (6.27)
\end{aligned}$$

Complex Poynting vector

We already know that the electromagnetic power flux is equal to the time averaged Poynting vector. With complex amplitudes we get:

$$\begin{aligned}
\mathbb{P} &= \text{Re}\left\{\mathbf{E}e^{j\omega t}\right\} \times \text{Re}\left\{\mathbf{H}e^{j\omega t}\right\} = \frac{1}{2}(\mathbf{E}e^{j\omega t} + \mathbf{E}^* e^{-j\omega t}) \times \frac{1}{2}(\mathbf{H}e^{j\omega t} + \mathbf{H}^* e^{-j\omega t})\\
&= \frac{1}{4}(\mathbf{E}\times\mathbf{H}^* + \mathbf{E}^*\times\mathbf{H} + \mathbf{E}\times\mathbf{H}e^{2j\omega t} + \mathbf{E}^*\times\mathbf{H}^* e^{-2j\omega t}) \qquad (6.28)
\end{aligned}$$

By averaging over time the exponential terms will disappear and we obtain:

$$\langle\mathbb{P}\rangle = \frac{1}{4}(\mathbf{E}\times\mathbf{H}^* + \mathbf{E}^*\times\mathbf{H}) = \frac{1}{2}(\mathbf{S}+\mathbf{S}^*) = \text{Re}\{\mathbf{S}\} \qquad (6.29)$$

with

$$\mathbf{S} = \frac{1}{2}(\mathbf{E}\times\mathbf{H}^*) \qquad (6.30)$$

The vector $\mathbf{S}$ is called the complex Poynting vector. The optical intensity is equal to the magnitude of the vector $\text{Re}\{\mathbf{S}\}$.

6.4.2 Transversal electromagnetic plane wave (TEM)

We consider a monochromatic plane wave in a medium (without sources) that is linear, non-dispersive, homogeneous and isotropic. For the electric and magnetic components with wave vector $\mathbf{k}$ we have the complex amplitudes:

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0\, e^{-j\mathbf{k}.\mathbf{r}} \qquad (6.31)$$

$$\mathbf{H}(\mathbf{r}) = \mathbf{H}_0\, e^{-j\mathbf{k}.\mathbf{r}} \qquad (6.32)$$

Here $\mathbf{E}_0$ and $\mathbf{H}_0$ are constant vectors. Each of these components satisfies the Helmholtz equation, where $\mathbf{k}$ is equal to $k = nk_0$, with n the refractive index of the medium. By substituting the

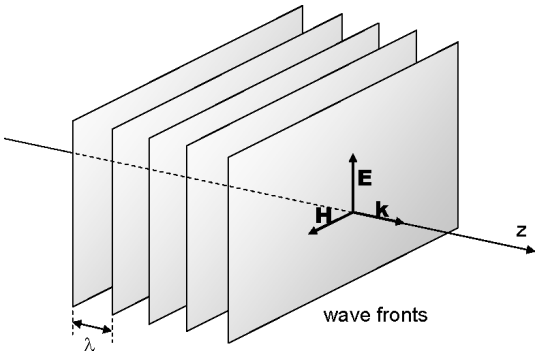

Figure 6.1: TEM plane wave. The vectors **E**, **H** and **k** are perpendicular. The wavefronts are normal to **k**.

previous amplitudes in the first two Maxwell equations (6.20) and (6.21) in the frequency domain, we get:

$$\mathbf{k} \times \mathbf{H}_0 = -\omega\epsilon\mathbf{E}_0 \tag{6.33}$$

$$\mathbf{k} \times \mathbf{E}_0 = \omega\mu_0\mathbf{H}_0 \tag{6.34}$$

This means **E** is perpendicular to both **k** and **H**. In addition **H** is perpendicular to **k** and **E**, see figure 6.1. Such a wave is called a transversal electromagnetic (TEM) wave. For the above equations to be consistent one needs:

$$\omega\epsilon/k = k/\omega\mu_0 \tag{6.35}$$

or

$$k = \omega\sqrt{\epsilon\mu_0} = \omega/v = n\omega/c = nk_0. \tag{6.36}$$

This is the condition for the wave to satisfy the Helmholtz equation. The ratio of the amplitudes gives:

$$\frac{E_0}{H_0} = Z = \frac{Z_0}{n} = \frac{\omega\mu_0}{k} \quad \text{with} \quad Z_0 = \sqrt{\frac{\mu_0}{\epsilon_0}} \approx 377\Omega, \tag{6.37}$$

with Z the impedance of the medium and Z_0 the impedance of free space.

6.4.3 Spherical wave

An example of an electromagnetic spherical wave is the field radiated by an electrical dipole. Such a spherical wave can be constructed by use of an auxiliary field **A**:

$$\mathbf{A}(\mathbf{r}) = A_0 U(\mathbf{r})\mathbf{e_x} \tag{6.38}$$

with $U(\mathbf{r})$ a scalar spherical wave with origin $r = 0$:

$$U(\mathbf{r}) = \frac{1}{r}e^{-jkr} \tag{6.39}$$

where $\mathbf{e_x}$ is the unit vector along the x-direction and also represents the direction of the dipole. We know that $U(\mathbf{r})$ satisfies the Helmholtz equation (see chapter 4), so $\mathbf{A}(\mathbf{r})$ is also a solution of the

Helmholtz equation, and it is called the electromagnetic vector potential. It can be proven that:

$$\mathbf{H} = \frac{1}{\mu_0}\nabla \times \mathbf{A} \tag{6.40}$$

$$\mathbf{E} = \frac{1}{j\omega\epsilon}\nabla \times \mathbf{H} \tag{6.41}$$

All these fields are proportional to $U(\mathbf{r})$.

6.5 Polarization of electromagnetic waves

The concept 'polarization' relates to the fact that the orientation of the electric field vector $\mathbf{E}(\mathbf{r}, t)$ of an electromagnetic wave changes in time if we look at the vector at a certain location in space. The state of polarization is completely known if we know how the orientation of the electric field vector changes in time.

The polarization of light has important consequences for the interaction of light with matter:

- The amount of reflected light at an interface depends on the polarization of the incident wave.
- The amount of absorption for some materials is polarization dependent.
- The refractive index of anisotropic materials depends on polarization. Waves with different polarization propagate with different speeds, thus they experience different phase changes so the polarization ellipse (see further) will transform.

Consider a monochromatic plane wave with frequency ν propagating in the z-direction with speed c. The electric field is in the xy-plane and is described in general by:

$$\mathbf{E}(z, t) = \text{Re}\left\{\mathbf{A}e^{j2\pi\nu(t-\frac{z}{c})}\right\} \tag{6.42}$$

with the complex vector

$$\mathbf{A} = A_x\mathbf{e_x} + A_y\mathbf{e_y} \tag{6.43}$$

with complex components A_x and A_y. To find the polarization of the wave we have to follow the end points of the vector $\mathbf{E}(z, t)$ at every position z and at every time t.

6.5.1 Elliptical polarization

Starting from the real representation of a monochromatic wave at a certain location in space we can see that the most general movement of the electric field vector in time is an ellipse. We call this an elliptical polarization state.

We write A_x and A_y with their magnitude and phase $A_x = a_x e^{j\phi_x}$, $A_y = a_y e^{j\phi_y}$. We substitute this into eq. 6.42 and obtain:

$$\mathbf{E}(z, t) = E_x\mathbf{e_x} + E_y\mathbf{e_y} \tag{6.44}$$

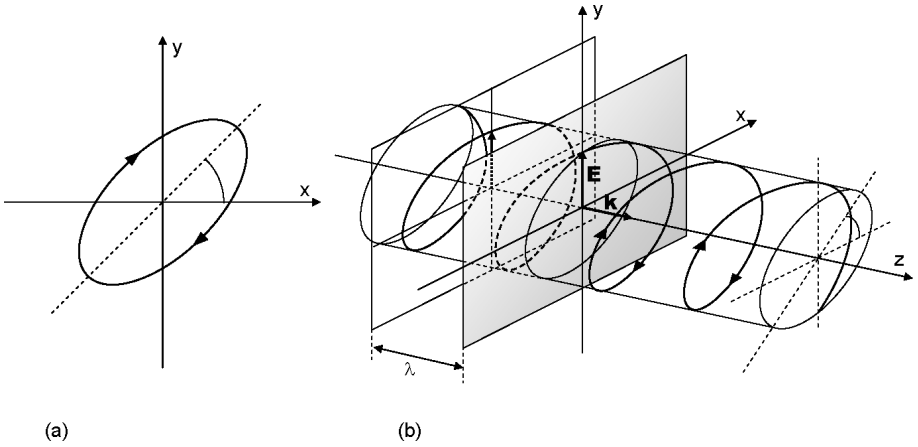

Figure 6.2: Elliptically polarized light. (a) Rotation of the end point of the electrical field vector in the xy-plane at a fixed location in space. (b) Trajectory in space at a fixed time t.

with

$$\begin{aligned} E_x &= a_x \cos\left[2\pi\nu\left(t - \frac{z}{c}\right) + \phi_x\right] \\ E_y &= a_y \cos\left[2\pi\nu\left(t - \frac{z}{c}\right) + \phi_y\right] \end{aligned} \tag{6.45}$$

The components E_x and E_y are periodic functions of $(t - z/c)$ and oscillate with frequency ν. These equations are the parameter equations of an ellipse. Indeed, by eliminating t we get:

$$\frac{E_x^2}{a_x^2} + \frac{E_y^2}{a_y^2} - 2\cos\phi\frac{E_x E_y}{a_x a_y} = \sin^2\phi \tag{6.46}$$

Here $\phi = \phi_y - \phi_x$ is the phase difference. At a fixed location z in space the end point of the electric field vector will rotate periodically in the xy-plane describing an elliptical trajectory, see figure 6.2a. At a fixed time t the location of the end point will follow a helical trajectory in space, see figure 6.2b. However when we travel along with the field at the speed of light ($t - \frac{z}{c}$ = constant), we will always see the same field orientation.

The complete state of polarization is known if we know the plane of the ellipse, the direction and magnitude of the main axes, the direction of revolution and the starting phase (thus the orientation of the electric field at time $t = 0$).

6.5.2 Linear polarization

If for elliptical polarization one of the components is dropped, e.g. $a_x = 0$, then the light is linearly polarized in the direction of the other component (e.g. y-direction). Light is also linearly polarized if the phase difference $\phi = 0$ or π, because then we obtain from eq. 6.46: $E_y = \pm(a_y/a_x)E_x$. This is the equation of a line with slope $\pm a_y/a_x$. These cases at a fixed position z and at a fixed time t are shown in figure 6.3.

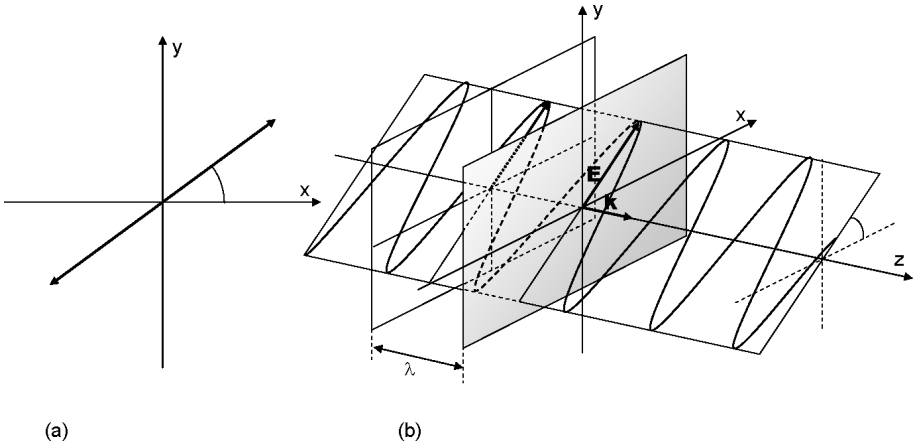

Figure 6.3: Linearly polarized light. (a) Time evolution at a fixed point in space. (b) Space evolution at a fixed time t.

6.5.3 Circular polarization

If $\phi = \pm\pi/2$ and $a_x = a_y = a_0$, then we obtain from eq. 6.46: $E_x^2 + E_y^2 = a_0^2$, which represents a circle. The elliptical cylinder of figure 6.2 will now be a circular cylinder and the wave is circularly polarized. If $\phi = +\pi/2$, the field at a fixed position z rotates clockwise, viewed from the direction the wave is propagating to. This is called right-hand circular polarization. The case $\phi = -\pi/2$ corresponds to a left-hand circularly polarized wave.

Unfortunately right- and left-handed polarization is not univocally defined in literature. In optics and physics the definition is used that right-handed corresponds to a clockwise movement when one looks into the bundle, while in the world of radiowaves and microwaves the reversed definition is used.

6.5.4 Superposition of polarizations

It is clear that the **E**-vector of an elliptical wave can be considered as the superposition of 2 linearly polarized waves. Because of linearity of Maxwell's equations, this means that the analysis of an optical system w.r.t. all possible polarizations can be limited to the behavior for 2 orthogonal linear polarizations.

6.5.5 Interference of electromagnetic waves

As was explained in the chapter on scalar waves, the superposition of two (or more) waves leads to interference effects in the intensity of those waves. At optical frequencies this means that an optical detector can 'see' fluctuations in the detected intensity: in certain locations in space there may be destructive interference and no signal is picked up by the detector, whereas in others there is constructive interference and therefore a strong signal is picked up. For most optical detectors (and also for the human eye) the relevant intensity is the energy density of the electric field, as given by $(\mathbf{E} \cdot \mathbf{E})/2$. If two fields $\mathbf{E}_1$ and $\mathbf{E}_2$ are present, the total energy density is given by

$$(\mathbf{E} \cdot \mathbf{E})/2 = (\mathbf{E}_1 + \mathbf{E}_2) \cdot (\mathbf{E}_1 + \mathbf{E}_2)/2 = |\mathbf{E}_1|^2/2 + |\mathbf{E}_2|^2/2 + \mathbf{E}_1 \cdot \mathbf{E}_2 \tag{6.47}$$

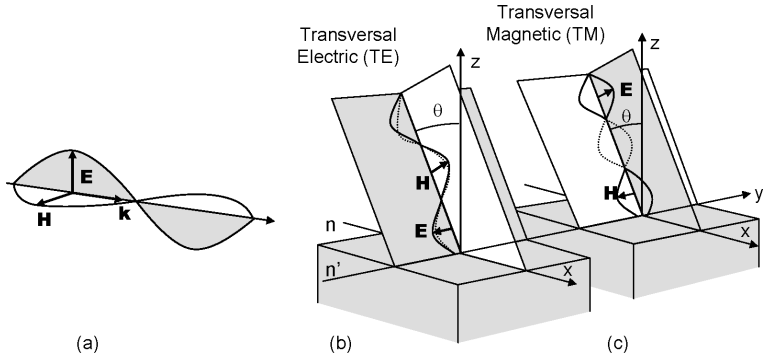

Figure 6.4: The problem of an electromagnetic wave (a) incident on an interface can be separated into a TE-problem (b) and a TM-problem (c). Both are decoupled.

From this expression it is clear that interference fringes will only be detectable if the constituting fields are not orthogonal. More in particular orthogonal polarizations will never interfere!

6.6 Reflection and refraction

In this section we examine reflection and refraction of a monochromatic plane wave with arbitrary polarization, incident on a plane interface between two dielectrics. We assume that these media are linear, homogeneous, isotropic, non-dispersive and non-magnetic. Figure 6.4 and 6.5 present an overview of the problem: we have two media with indices n and n', an incident wave, a reflected wave and a refracted wave. Already in chapter 4 we showed that the wave fronts of the incident and the reflected wave agree at the interface only if $\theta = \theta''$. Snell's law was also obtained: $nsin\theta = n'sin\theta'$.

Now we want to get the reflection and transmission coefficient for the reflected and refracted wave. Therefore we demand that the fields satisfy the boundary conditions at the interface. Previously we observed that the tangential components of $\mathbf{E}$ and $\mathbf{H}$, and the normal components of $\mathbf{D}$ and $\mathbf{B}$, have to be continuous at the boundary. Furthermore, we noted that the ratio of the amplitude of the magnetic field to the perpendicular electric field is equal to $E/H = Z_0/n$, with Z_0 the free space impedance ($Z_0 = \sqrt{\mu_0/\epsilon_0}$), and with n the refractive index of the medium in which the wave is propagating.

When solving Maxwell's equations at the interface, the problem reduces to a two-dimensional one, because the fields at an interface are y-invariant, see figure 6.4. One can prove (substitution of two-dimensional fields in Maxwell's equations) that the general solution of the equations for two-dimensional phenomena are separated into two partial problems: we get two decoupled sets of differential equations. One gives the solution for the components: $E_y(x,z), H_x(x,z)$ and $H_z(x,z)$. These are called TE or transversal electric solutions (sometimes also called s-polarization), because the single component of the electric field is transversal (=perpendicular) to the plane of incidence (being the plane containing the direction of incidence, and which is normal to the interface). The other differential equation set determines: $H_y(x,z), E_x(x,z)$ and $E_z(x,z)$, which are analogously called TM or transversal magnetic solutions (or sometimes p-polarization).

From the previous data it is possible to calculate the reflection and transmission coefficients for both TE and TM polarizations (do this yourself!). The results are:

$$r_{TE} = \frac{E''_{TE}}{E_{TE}} = \frac{n\cos\theta - n'\cos\theta'}{n\cos\theta + n'\cos\theta'} \tag{6.48}$$

$$t_{TE} = \frac{E'_{TE}}{E_{TE}} = 1 + r_{TE} = \frac{2n\cos\theta}{n\cos\theta + n'\cos\theta'} \tag{6.49}$$

$$r_{TM} = \frac{E''_{TM}}{E_{TM}} = \frac{n'\cos\theta - n\cos\theta'}{n'\cos\theta + n\cos\theta'} \tag{6.50}$$

$$t_{TM} = \frac{E'_{TM}}{E_{TM}} = \frac{n}{n'}(1 + r_{TM}) = \frac{2n\cos\theta}{n'\cos\theta + n\cos\theta'} \tag{6.51}$$

These coefficients are known as the Fresnel coefficients for TE and TM polarization. Note that according to Snell's law:

$$\cos\theta' = \sqrt{1 - \sin^2\theta'} = \sqrt{1 - \left(\frac{n}{n'}\right)^2 \sin^2\theta} \tag{6.52}$$

Thus, it is possible that the reflection and transmission coefficient are complex, because the expression under the root in the previous equation can be negative. The magnitudes of $|r_{TE}|$ and $|r_{TM}|$, as well as the phase shifts $\phi_{TE} = arg(r_{TE})$ and $\phi_{TM} = arg(r_{TM})$ are shown in figure 6.5 in function of the incidence angle θ. For each polarization we distinguish external ($n' > n$) and internal ($n > n'$) reflections.

For perpendicular incidence, there is no difference between the TE and TM case. Equation (6.48) differs however from equation (6.50) in this case (sign difference). This is caused by the different definition of the direction of the unit vectors for the E-field, in the TE and TM case. Figure 6.5 depicts the definition of the unit vectors for the E and H-field for the incident, reflected and refracted wave.

It is interesting to note the connection between the reflection r and transmission t (from the medium with index n and angle θ), and the reflection r' and transmission t' upon incidence from the other side (in medium n' and with angle θ'). By inspecting the Fresnel coefficients one obtains for both TE and TM polarization:

$$r = -r', \tag{6.53}$$

$$tt' - rr' = tt' + r^2 = 1, \tag{6.54}$$

so

$$tt' = 1 - r^2 = 1 - r'^2. \tag{6.55}$$

6.6.1 TE polarization

External reflection ($n' > n$). The reflection coefficient r_{TE} is always real and negative, which corresponds to a phase shift $\phi_{TE} = \pi$. The magnitude $|r_{TE}|$ for perpendicular incidence ($\theta = 0$) is equal to $\frac{n'-n}{n+n'}$. For $\theta = 90°$, $|r_{TE}| = 1$.

Internal reflection ($n' < n$). For small θ the reflection coefficient r_{TE} is real and positive. The magnitude $|r_{TE}|$ for perpendicular incidence ($\theta = 0$) is $\frac{n-n'}{n+n'}$. At a certain angle θ we get that

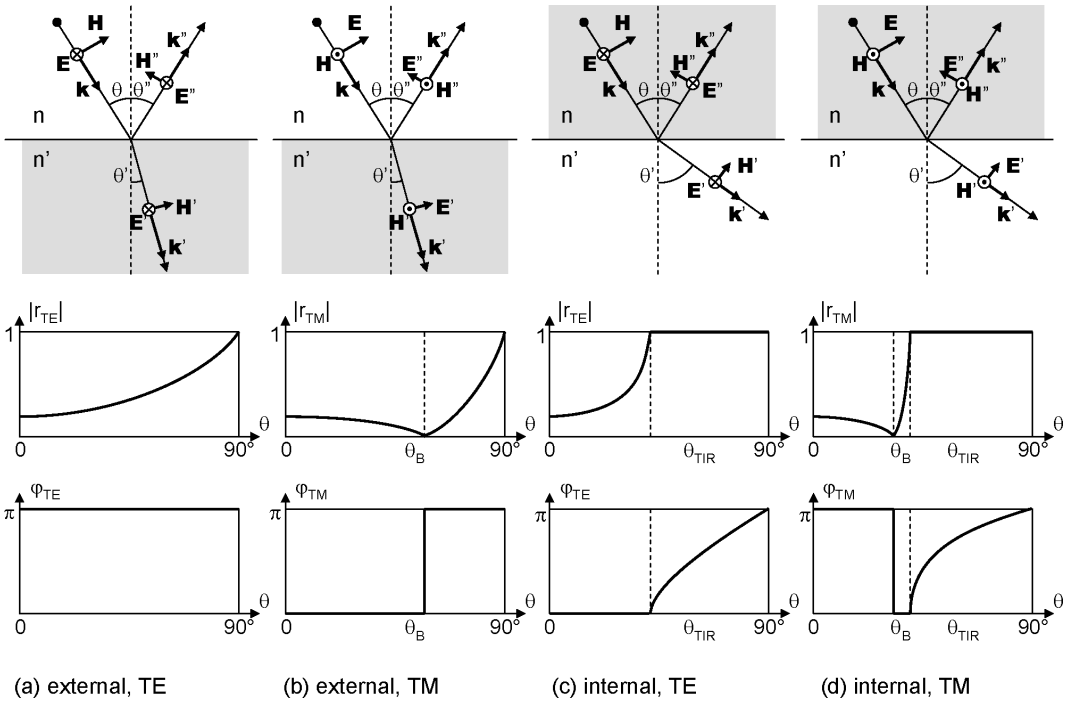

Figure 6.5: Magnitude and phase of the reflection coefficient in function of incidence angle for (a) external reflection ($n'/n = 1.5$) and TE polarization, (b) external reflection ($n'/n = 1.5$) and TM polarization, (c) internal reflection ($n/n' = 1.5$) and TE polarization and (d) internal reflection ($n/n' = 1.5$) and TM polarization.

$|r_{TE}| = 1$. This angle is called the critical angle:

$$\theta_{CRIT} = \sin^{-1}\left(\frac{n'}{n}\right). \tag{6.56}$$

For $\theta > \theta_{CRIT}$ one has $|r_{TE}| = 1$, which corresponds to total internal reflection (TIR) at the interface. Under conditions of TIR the electromagnetic field in the external medium is not zero but decays exponentially away from the interface. We call this decaying field tail an evanescent field. At the critical angle the tail extends infinitely into the external medium whereas at $\theta = 90°$ the tail becomes very short.

> Exercise: Derive an expression for the decay constant of the tail in the TIR regime as a function of angle of incidence.

6.6.2 TM polarization

External reflection ($n' > n$). The reflection coefficient r_{TM} is real. The magnitude $|r_{TM}|$ for perpendicular incidence ($\theta = 0$) is equal to $\frac{n'-n}{n+n'}$ and decreases for increasing θ, until $|r_{TM}| = 0$. This angle is called the Brewster angle, θ_B:

$$\theta_B = \tan^{-1}\left(\frac{n'}{n}\right). \tag{6.57}$$

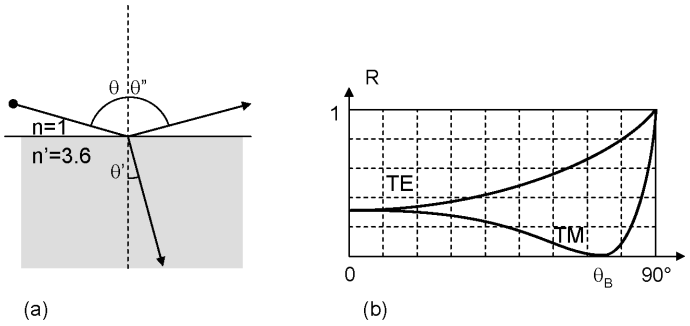

Figure 6.6: Reflectance for TE and TM polarization at an interface between air and GaAs ($n' = 3.6$).

For $\theta > \theta_B$ r_{TM} will change sign and its magnitude increases gradually until it reaches 1 at $\theta = 90°$. The fact that a TM-wave is not reflected at the Brewster angle is used for the fabrication of polarizers (devices that block a certain polarization and transmit another).

Internal reflection Analogous discussion.

6.6.3 Power reflection and transmission

The reflection and transmission coefficients r and t are ratios of complex field amplitudes. The power reflection (or reflectance) R and power transmission (or transmittance) T is defined as the ratio of optical flux densities (along a direction perpendicular to the surface) of reflected and transmitted wave, relative to the incident wave. Because the incident and reflected wave propagate in the same medium, and their angles with the interface are the same, we obtain:

$$R = |r|^2. \tag{6.58}$$

Power conservation dictates: $T = 1 - R$. Note that $T = \frac{n' \cos\theta'}{n \cos\theta}|t|^2$, which is *not* equal to $|t|^2$, as the power propagates along a different angle.

An important case is that of perpendicular incidence on an interface. The reflectance, resp. transmittance, is the same for TE and TM, both for internal and external reflection, and is equal to:

$$\begin{aligned} R &= \left(\frac{n - n'}{n + n'}\right)^2 \\ T &= \frac{4nn'}{(n + n')^2} \end{aligned}$$

Example: the reflectance and transmittance at the interface between glass ($n' = 1.5$) and air is 4% for perpendicular incidence. Figure 6.6 shows the reflectance for TE and TM between air and GaAs ($n' = 3.6$) in function of the incidence angle θ.

6.7 Absorption and dispersion

6.7.1 Absorption

Up until now we assumed that the dielectric media were completely transparent, there was no absorption of light by the material. For example, glass is very transparent in the visible part of the spectrum, but it strongly absorbs infrared and ultraviolet light. Dielectrics that absorb light are often described by a complex susceptibility:

$$\chi = \chi_R + j\chi_I \tag{6.59}$$

Correspondingly, there is a complex permittivity $\epsilon = \epsilon_0(1+\chi)$ and a complex wave number $k = k_0\sqrt{1+\chi}$.

Now assume a plane wave propagating in the z-direction in a certain medium, then its complex amplitude is equal to: Ae^{-jkz}. This is analogous to the description of the evanescent plane wave in section 4.2.2. Because k is complex both the phase and the amplitude of the wave will vary along z. We write k with its real and imaginary part:

$$k = k_0\sqrt{1+\chi_R+j\chi_I} = k_0(n_R + jn_I) = \beta - \frac{j}{2}\alpha \tag{6.60}$$

Thus $e^{-jkz} = e^{-\frac{1}{2}\alpha z}e^{-j\beta z}$, the intensity of the plane wave is attenuated (exponentially) by the coefficient α: attenuation coefficient, absorption coefficient or extinction coefficient. α is expressed in 1/m. We can say that the power of the light decreases exponentially with propagation distance: $P(z) = P_0e^{-\alpha z}$

Note about dB's

The ratio between optical power after a certain propagation (P_o) and initial optical power (P_i) is mostly expressed in dB:

$$10\ \log\frac{P_o}{P_i} \tag{6.61}$$

In a medium with absorption this results in

$$10\ \log\frac{P_o}{P_i} = 10\ \log e^{-\alpha z} = (10\ \log e)(-\alpha z). \tag{6.62}$$

The attenuation coefficient expressed in dB/m is

$$\alpha(dB/m) = (10\ \log e)\alpha(1/m) = 4.34\alpha(1/m) \tag{6.63}$$

The following table presents some important conversions between dB's and power ratios.

$0dB = 1$	
$+1dB \approx +25\%$	$-1dB \approx -20\%$
$+3dB \approx +100\%\ of\ 2\times$	$-3dB \approx -50\%\ or\ \div 2$
$+6dB \approx 4\times$	$-6dB \approx \div 4$
$+10dB \approx 10\times$	$-10dB \approx \div 10$
$+20dB \approx 100\times$	$-20dB \approx \div 100$

In the chapter about lasers we will see that α can be negative, which means that the medium amplifies the propagating light, instead of absorption!

The parameter β corresponds to the rate by which the phase changes with z and it is called the propagation constant. The plane wave propagates with phase velocity $v_p = c/n = \omega/(k_0 n)$.

6.7.2 Dispersion

Dispersive media are characterized by a frequency dependent (and wavelength dependent) susceptibility $\chi(\nu)$, refractive index $n(\nu)$ and speed of light $v(\nu) = c/n(\nu)$. Optical components such as prisms and lenses fabricated from dispersive media will refract waves of different wavelengths into different angles, which leads to chromatic abberation (see section 3.5.1).

Because the speed of light depends on the frequency in a dispersive medium, each frequency component constituting the wave will experience a different time retardation upon propagation through the dispersive material. Because of this a short pulse in time will spread out in time. This effect becomes important upon propagation through kilometers of optical fibers.

The quantity $\frac{dn}{d\lambda}$ is called the *material dispersion*. We noted previously that a monochromatic wave propagating with propagation constant β has a *phase velocity* equal to $v_p = \omega/\beta$. However, a perturbation of the wave, for example by amplitude modulation, travels with another velocity that is called the *group velocity*: $v_g = \frac{d\omega}{d\beta}$. Correspondingly one defines the group index as $N = c/v_g = n_{eff} - \lambda\frac{dn}{d\lambda}$. For most optical materials the refractive index decreases as the wavelength increases. Then the group index is larger than the effective index, so the group velocity will be smaller than the phase velocity. To better understand the concept of group velocity it is instructive to consider two optical signals with slightly different frequencies, and thus with slightly different phase velocities (because of material dispersion). The total field shows a beating pattern for the intensity. This pattern will propagate with a different speed than the two phase velocities.

6.8 Layered structures

6.8.1 Three-layer structure

If a wave is incident on a layered medium - a structure with a number of parallel layers and interfaces - there are interference effects that are a consequence of the multiple reflections in these structures. The global reflection and transmission of the structure is dependent on the incidence angle, the wavelength and the polarization of the incident wave.

The general case of a plane wave incident on a layered medium with N interfaces is treated elegantly by the transfer matrix method. However, this method is beyond the scope of this course. Here we discuss the simpler case of the three-layer structure (this means one layer in between two semi-infinite media), as depicted in figure 6.7(a). Such a structure with two parallel semi-transparent mirrors is a cavity where resonances can exist. It is called a Fabry-Perot etalon. The discussion is limited to lossless structures, thus with a real refractive index.

As in the case of a single interface, we consider one monochromatic plane wave incident on the layer structure from a given direction. We assume linear polarization - s-type or p-type - for

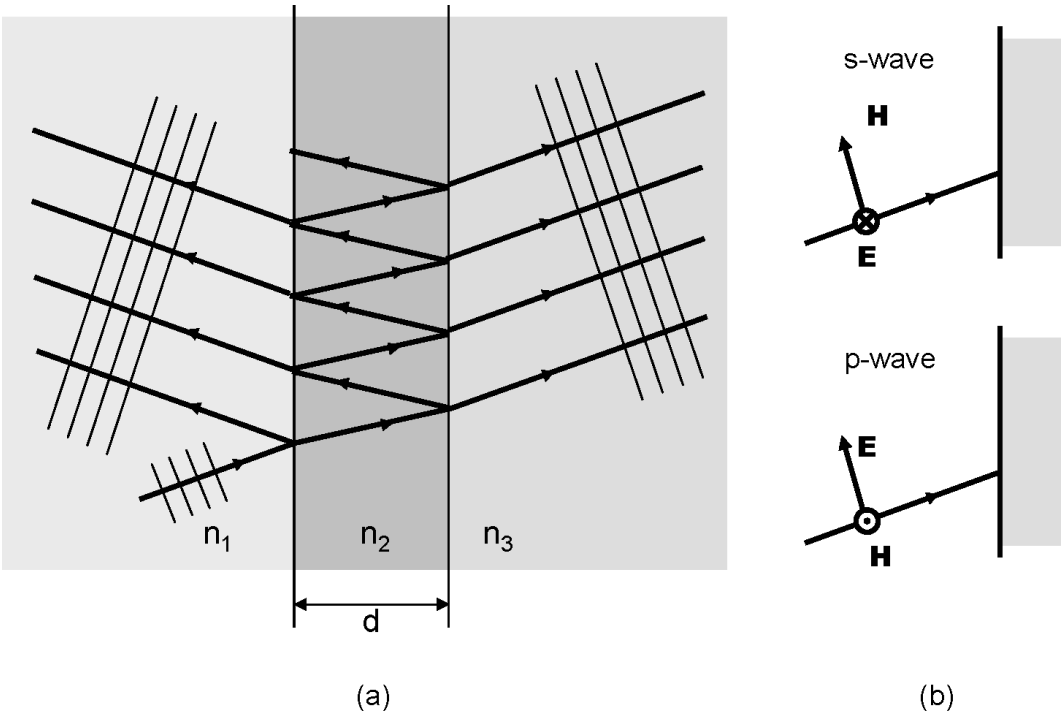

Figure 6.7: (a) Reflection and transmission at a plate. (b) The s-wave and the p-wave.

the E-field, as shown in figure 6.7(b). Any incident field may be considered as a superposition of such monochromatic linearly polarized plane waves. The following analysis is valid for both s- and p-polarization. The difference between both situations is contained in the reflection and transmission coefficients for the interfaces.

One can calculate the global reflection and transmission in two different ways. The first method closely resembles the physical process, whereas the second method is mathematically more elegant. In the first approach the 'consecutive' reflections at both interfaces are determined, and the global reflection and transmission are written as infinite sum series of these contributions. In the second approach one realizes that every layer contains one forward and one backward plane wave. By matching the boundary conditions at the interfaces one obtains a linear system that is easily solvable. Both methods are presented here, and they deliver the same result, of course. For layered media with more than three layers it is possible to work with both methods, in principle. However, the first 'sequential' method quickly becomes cumbersome, while the second method remains elegant. For this approach, the system to be solved scales linearly with the number of layers.

For the first method we consider figure 6.8(a): the plane wave impinges on the first interface, a part reflects and a part transfers to the second medium. The transmitted part hits the second interface, with again a partial reflection and transmission. The reflected part goes to the first interface, part of it goes to medium 1, the other part reflects back etc. etc. All contributions to this sequential story are indicated as arrows on the figure. However, it has to be clear that each arrow represents a plane wave that is present in the entire vertical layer. We write the linearly polarized E-field of the incident field as:

$$E_{F,1}(x,z) = Ae^{-j(k_{z,1}z+k_{x,1}x)}, \tag{6.64}$$

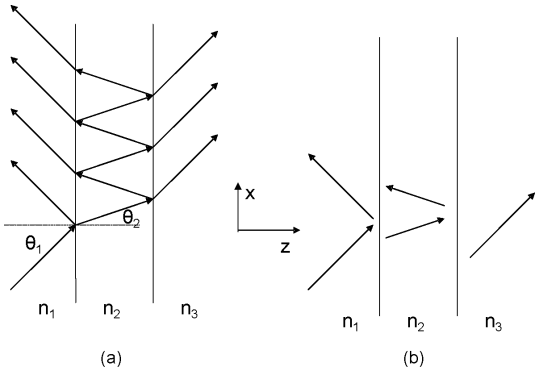

Figure 6.8: Two methods: (a) Sum series of contributions. (b) Global forward and backward plane waves.

with

$$k_{z,i} = k_0 n_i \cos\theta_i, \tag{6.65}$$

$$k_{x,i} = k_0 n_i \sin\theta_i. \tag{6.66}$$

The index F indicates 'forward', thus propagating in the positive z-direction. The total field in layer 2 is then easily written as the series:

$$\begin{aligned} E_{F,2}(x,z) &= A t_{12} e^{-j(k_{z,2}z + k_{x,2}x)} \left[1 + r_{23} r_{21} e^{-j2k_{z,2}d} + \left(r_{23} r_{21} e^{-j2k_{z,2}d} \right)^2 + \ldots \right] && (6.67) \\ &= \frac{A t_{12}}{1 - r_{23} r_{21} e^{-j2k_{z,2}d}} e^{-j(k_{z,2}z + k_{x,2}x)}. && (6.68) \end{aligned}$$

Here r_{ij} (t_{ij}) is the field reflection (transmission) coefficient for incidence from medium i on the interface with medium j. For the directions of the waves in the three layers one uses Snell's law:

$$k_{x,1} = k_{x,2} = k_{x,3}. \tag{6.69}$$

Analogously, one obtains for the backward field in layer 2:

$$\begin{aligned} E_{B,2}(x,z) &= A t_{12} r_{23} e^{-jk_{z,2}d} e^{-j(k_{z,2}(d-z) + k_{x,2}x)} \left[1 + r_{23} r_{21} e^{-j2k_{z,2}d} + \left(r_{23} r_{21} e^{-j2k_{z,2}d} \right)^2 + \ldots \right] \\ &= \frac{A t_{12} r_{23} e^{-j2k_{z,2}d}}{1 - r_{23} r_{21} e^{-j2k_{z,2}d}} e^{-j(-k_{z,2}z + k_{x,2}x)}. && (6.70) \end{aligned}$$

Based on the expressions for $E_{F,2}$ and $E_{B,2}$ it is easy to write the total reflected field $E_{B,1}$ and the total transmitted field $E_{F,3}$:

$$\begin{aligned} E_{F,3}(x,z) &= t_{23} E_{F,2}(x,d) e^{-jk_{z,3}(z-d)} && (6.71) \\ &= \frac{A t_{12} t_{23} e^{-j(k_{z,2} - k_{z,3})d}}{1 - r_{23} r_{21} e^{-j2k_{z,2}d}} e^{-j(k_{z,3}z + k_{x,3}x)}, && (6.72) \end{aligned}$$

and

$$\begin{aligned} E_{B,1}(x,z) &= r_{12} E_{F,1}(x,0) e^{+jk_{z,1}z} + t_{21} E_{B,2}(x,0) e^{+jk_{z,1}z} && (6.73) \\ &= A \left[r_{12} + \frac{t_{12} t_{21} r_{23} e^{-j2k_{z,2}d}}{1 - r_{23} r_{21} e^{-j2k_{z,2}d}} \right] e^{-j(-k_{z,1}z + k_{x,1}x)}. && (6.74) \end{aligned}$$

The second method starts from the insight that, upon incidence of a single plane wave, all contributions to the forward field in every layer have the same direction, and thus they form one plane wave. The same holds for the backward field in each layer. The situation is shown in figure 6.8(b). In each layer the total forward or backward field is represented by a single plane wave, that we can write as:

$$\begin{aligned} E_{F,1}(x,z) &= Ae^{-j(k_{z,1}z+k_{x,1}x)}, && (6.75)\\ E_{B,1}(x,z) &= A_{B,1}e^{-j(-k_{z,1}z+k_{x,1}x)}, && (6.76)\\ E_{F,2}(x,z) &= A_{F,2}e^{-j(k_{z,2}z+k_{x,2}x)}, && (6.77)\\ E_{B,2}(x,z) &= A_{B,2}e^{-j(-k_{z,2}z+k_{x,2}x)}, && (6.78)\\ E_{F,3}(x,z) &= A_{F,3}e^{-j(k_{z,3}z+k_{x,3}x)}. && (6.79) \end{aligned}$$

Determining the 4 complex coefficients $A_{B,1}$, $A_{F,2}$, $A_{B,2}$ and $A_{F,3}$ is possible in the following way. At each interface we write the fields that propagate away from it in function of the fields that propagate towards it. This amounts to applying the boundary conditions at the interface.

$$\begin{aligned} E_{F,2}(x,0) &= t_{12}E_{F,1}(x,0) + r_{21}E_{B,2}(x,0), && (6.80)\\ E_{B,1}(x,0) &= r_{12}E_{F,1}(x,0) + t_{21}E_{B,2}(x,0), && (6.81)\\ E_{B,2}(x,d) &= r_{23}E_{F,2}(x,d), && (6.82)\\ E_{F,3}(x,d) &= t_{23}E_{F,2}(x,d). && (6.83) \end{aligned}$$

Solving this system of 4 complex equations with 4 complex unknowns leads to the same result as with the first method.

For the power reflectance and transmittance of the Fabry-Perot etalon one obtains finally:

$$R = \left|\frac{E_{B,1}(x,0)}{E_{F,1}(x,0)}\right|^2 = \left|r_{12} + \frac{t_{12}t_{21}r_{23}e^{-j2k_{z,2}d}}{1 - r_{23}r_{21}e^{-j2k_{z,2}d}}\right|^2, \quad (6.84)$$

$$T = \frac{n_3\cos\theta_3}{n_1\cos\theta_1}\left|\frac{t_{12}t_{23}}{1 - r_{23}r_{21}e^{-j2k_{z,2}d}}\right|^2. \quad (6.85)$$

The latter expression is only valid if $k_{z,2}$ is real. It is left to the reader to generalize this expression for the case where $k_{z,2}$ is complex. This happens either when the layer is absorptive or when the field is evanescent in layer 2 due to total internal reflection.

In the next section we examine a symmetrical structure ($n_1 = n_3$), like the case of a transparent plate in air. Then, the previous expressions simplify to:

$$\begin{aligned} R &= \left|\frac{r_{12} - r_{12}\left(r_{12}^2 + t_{12}t_{21}\right)e^{-j2k_{z,2}d}}{1 - r_{12}^2e^{-j2k_{z,2}d}}\right|^2 && (6.86)\\ &= \left|\frac{r_{12}\left(1 - e^{-j2k_{z,2}d}\right)}{1 - r_{12}^2e^{-j2k_{z,2}d}}\right|^2 && (6.87)\\ &= 4\left|\frac{r_{12}}{1 - r_{12}^2e^{-j2k_{z,2}d}}\right|^2 \sin^2\left(k_{z,2}d\right), && (6.88)\\ T &= \left|\frac{t_{12}t_{21}}{1 - r_{12}^2e^{-j2k_{z,2}d}}\right|^2. && (6.89) \end{aligned}$$

We can simplify this to:

$$R = 4\left|\frac{r_{12}}{1-r_{12}^2e^{-j2\phi}}\right|^2\sin^2\phi, \tag{6.90}$$

$$T = \left|\frac{t_{12}t_{21}}{1-r_{12}^2e^{-j2\phi}}\right|^2 \tag{6.91}$$

with

$$\phi = k_{z,2}d = \frac{2\pi}{\lambda_0}n_2 d\cos\theta_2. \tag{6.92}$$

For media with a real refractive index we can write the reflectance and transmittance for one transition, respectively:

$$R_1 = |r_{12}|^2 = r_{12}^2, \tag{6.93}$$

$$T_1 = |t_{12}t_{21}| = 1 - R_1. \tag{6.94}$$

Then, for the transmission of the plate we obtain:

$$\begin{aligned} T &= \frac{T_1^2}{1+R_1^2-2R_1\cos 2\phi} \\ &= \frac{(1-R_1)^2}{(1-R_1)^2+2R_1-2R_1\cos 2\phi} \\ &= \frac{1}{1+F\sin^2\phi} \quad \text{with} \quad F=\frac{4R_1}{(1-R_1)^2} \end{aligned} \tag{6.95}$$

This last equation is also called the *Airy* equation. Note that we have already calculated this transmission for interference between multiple waves, see section 4.5.2. The maximum transmission is 1, and then we find for perpendicular incidence ($\cos\theta_2 = 1$) that $\phi = m\pi$ or $d = m\frac{\lambda}{2n_2}$, so the thickness of the layer is an integer times the half wavelength in the material. The minimal transmission is given by:

$$T_{\min} = \frac{1}{1+F} \tag{6.96}$$

For sharp maxima we need to ensure that T_{min} is as small as possible. Therefore F has to be large, so R_1 needs to be close to 1. In practice, this is difficult because of the available materials.

Figure 6.9 shows the reflectance R for perpendicular incidence on a layer with thickness d and index n, placed in air. R is presented in function of the wavelength (normalized to nd) for 4 values of n: 1.5, 2, 4 and 8 (this last value is unrealistic for normal materials). One notices that the reflections drop to 0 if the thickness is an integer times the half wavelength. The reflection dips become sharper as n increases (and thus R_1 increases) and they obtain the character of a resonance. Such a structure consisting of two semi-transparent mirrors is called a Fabry-Perot resonator.

It is interesting to consider what happens to the reflection or transmission spectrum if the light incidence is no longer perpendicular but oblique. It is sufficient to realize that the maxima and minima occur for certain values of ϕ. If the angle θ increases, then $\cos\theta$ decreases and thus the wavelength has to decrease also to keep ϕ constant. *This means that a reflection or transmission spectrum always shifts to shorter wavelengths as the light incidence becomes more oblique.* This contradicts the intuition: for oblique angles the light has to travel a longer distance in the layer, and thus one

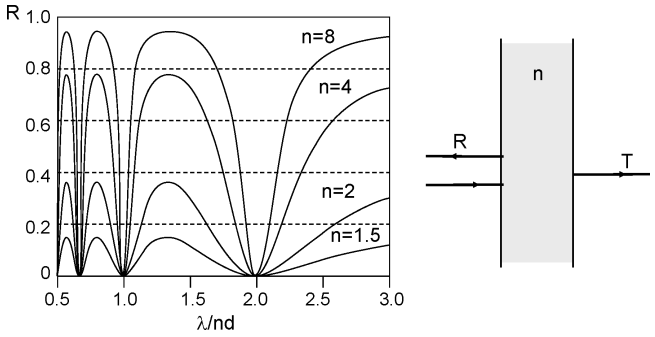

Figure 6.9: Reflection of a layer in air with indices $n = 1.5$ (lower curve), $n = 2$, $n = 4$, $n = 8$ (upper curve).

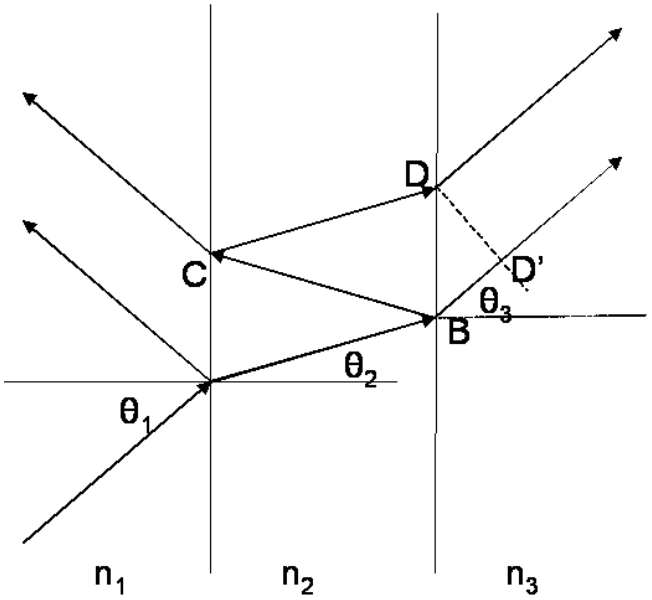

Figure 6.10: Fabry-Perot structure with oblique incidence.

could expect that the wavelength has to increase to remain in the same maximum or minimum. We show with figure 6.10 that this reasoning is incorrect: Consider the primary and secondary contribution to the total transmission. Both contributions are plane waves. To know the phase difference between them, we have to examine the phase at the same phase front, for example the plane DD'. Thus the phase difference is *not* determined by the path length $|BC|+|CD|$, but by the difference in path length between $|BC| + |CD|$ and $|BD'|$. This path length difference decreases as θ increases, while $|BC| + |CD|$ increases! It is an exercise for the reader to show how this path length difference translates into the phase 2ϕ.

6.8.2 Reciprocity

Upon careful inspection of equation 6.85 one can see that the power transmission T is invariant to an exchange of layer 1 by layer 3 and vice versa. More precisely: the power transmission is identical for transmission from left to right at an angle of incidence θ_1 and from right to left at an angle of incidence θ_3 (connected to θ_1 by Snell's law). This remarkable property is called

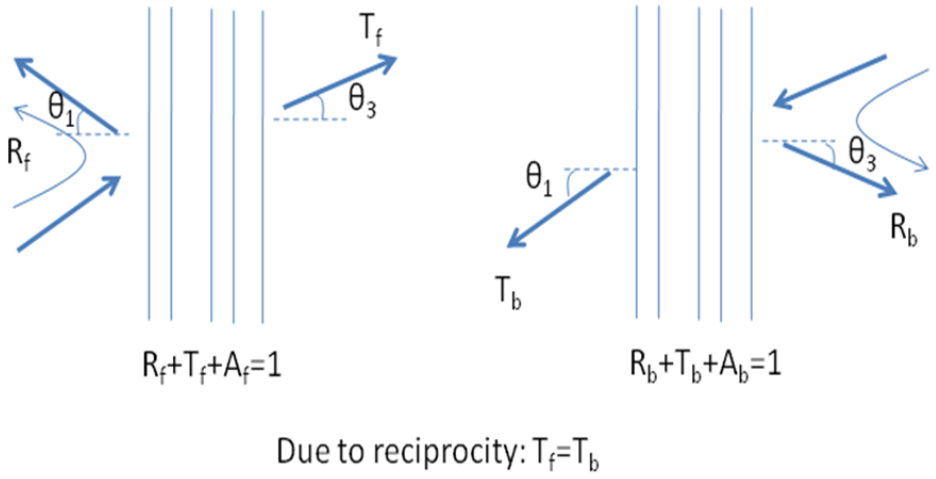

Figure 6.11: Reciprocity in an N-layer slab

reciprocity and is the consequence of a very general reciprocity theorem in electromagnetics. It is not only valid for a 3-layer structure with real refractive indices but also for an N-layer structure with complex refractive indices. Only in very special materials (non-reciprocal materials) - which cannot be described by a simple refractive index - this property is broken. In lossless structures the equality of $T_{forward}$ and $T_{backward}$ implies that the reflection R is also equal for incidence from the left (θ_1) and from the right (θ_3) respectively. However, in lossy structures this is not the case. The forward and backward transmission are equal, but the reflection can be different, in which case the absorption is also different. The situation is depicted graphically in figure 6.11, where A denotes the fraction of power which is absorbed.

6.8.3 Coatings

Layer structures can be employed to increase or decrease the reflection of a surface. This is useful for the design of anti-reflection coatings (AR-coatings) for lenses, and for the design of dielectric mirrors. Most of the time one uses perpendicular incidence.

AR-coatings: quarter-wave layer

In designing an AR-coating we ensure that the reflection at the front of the film interferes destructively with the reflection at the back of the film. If $n_1 < n_2 < n_3$ then one needs: (note: extra phase shift π for reflection at interface 1-2 and interface 2-3!)

$$d = \frac{1}{4}\frac{\lambda_0}{n_2} = \frac{\lambda_2}{4}, \tag{6.97}$$

hence the name quarter-wave layer. This is illustrated in figure 6.12 for the first two contributions of the reflected field. In practice these are the most important contributions (note that our analysis does take all reflections into account).

Combining the previous equation with the Fresnel coefficients for perpendicular incidence:

$$r_{ij} = \frac{n_i - n_j}{n_i + n_j} \tag{6.98}$$

$$t_{ij} = \frac{2n_i}{n_i + n_j} \tag{6.99}$$

and setting e.g. $T = 1$ in equation 6.85, one obtains:

$$n_2 = \sqrt{n_1 n_3}. \tag{6.100}$$

Example: AR-coating for GaAs-structures

An AR-coating to minimize the reflection between air and a medium with index $n = 3.2$ at $\lambda = 1550nm$ (e.g. an optical amplifier in GalliumArsenide). We get $n_2 = \sqrt{n_1 n_3} = 1.79$ and $d = 217nm$. In figure 6.12 we see indeed that for an AR-coating with these values no reflections occur at $\lambda = 1550nm$, and that reflection remains smaller than 0.5% in a wide interval (1450nm tot 1650nm). In practice it is not easy to fabricate a coating with the exact index and thickness (e.g. because only a limited number of materials are available). Moreover, a small error for d and n immediately leads to higher values for the reflection coefficient.

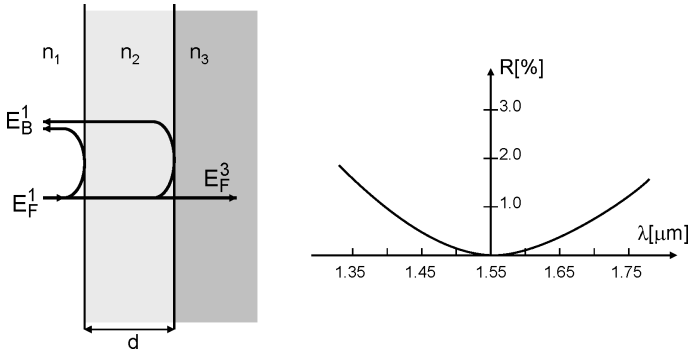

Figure 6.12: AR-coating consisting of one layer. (a) Principle, (b) reflection spectrum for an AR-coating designed for the telecom wavelength of 1550nm. $n_1 = 1$, $n_3 = 3.2$, $n_2 = \sqrt{3.2} = 1.79$, $d = 217$nm.

Highly reflective coatings

A HR-coating could consist of a quarter-wave layer made from a higher index than both considered media. In practice this is often not realizable, therefore one employs a periodic structure of quarter-wave layers alternating between high and low index, see figure 6.13. Together they behave as a Bragg reflector.

If the thickness of the consecutive layer can be controlled so that:

$$n_H d_H = n_L d_L = \frac{\lambda_0}{4} \tag{6.101}$$

then the reflected beams from the different interfaces will all interfere constructively, leading to a large reflection coefficient. Using the matrix method one can obtain for R:

$$R = \left(\frac{1 - \left(\frac{n_H}{n_L}\right)^{2N}}{1 + \left(\frac{n_H}{n_L}\right)^{2N}} \right)^2 . \tag{6.102}$$

R converges to 1 as N increases. The convergence improves as the ratio $\frac{n_H}{n_L}$ becomes larger.

Example: HR coating for a He-Ne laser

A HR-coating consisting of silver sulfide ($n_H = 2.32$) and magnesium fluoride ($n_L = 1.38$) has a reflection of 98.9% already after 13 layers, at $\lambda = 633nm$. Such a highly-reflective mirror is used for fabrication of a helium-neon laser cavity.

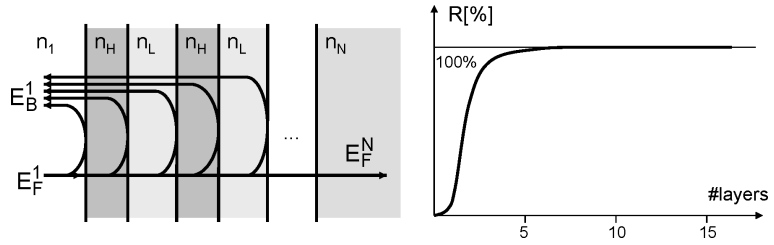

Figure 6.13: HR coating. (a) Principle. (b) Reflection of a coating for a He-Ne laser at wavelength λ =633nm. n_H =2.32 (ZnS), n_L=1.38 (MgF_2)

Exercise: Explain why in the AR-coating a quarter wavelength thick layer leads to destructive interference for the reflected light, whereas in the HR-coating quarter wavelength thick layers lead to constructive interference for the reflected light.

Design of complicated coatings

For more complicated applications (broadband and narrowband filters, power and polarization splitters...) one uses specialized CAD-software.

Example: coating for sunglasses

Figure 6.14 shows an example of a design for sunglasses. The demands were the following:

- Transmission $< 1\%$ for wavelengths between 400nm and 500nm.
- Transmission between 15% and 25% between 510nm and 790nm.
- Transmission $< 1\%$ between 800nm and 900nm.

The designed coating has 29 layers of SiO_2 and TiO_2 with thicknesses between 20nm and 200nm on a glass substrate.

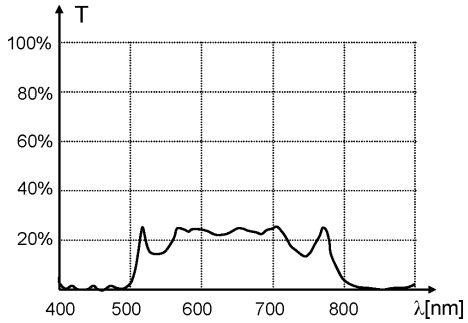

Figure 6.14: Transmission of sunglasses (© 1995-98 Software Spectra, Inc., http://www.sspectra.com/).

6.9 Scattering

The scattering of light can be seen as the deviation from a straight trajectory when the electromagnetic (EM) wave (light) encounters obstacles or non-uniformities in the medium in which it travels. The scattering mechanisms that we will discuss here involve scattering particles which can be assumed spherical. When the EM wave encounters a particle it will cause a periodic perturbation in the electron orbits within the molecules of the particle. This perturbation has the same frequency as the incoming EM-wave. The separation of the charges in the molecule due to the perturbation is called the induced dipole moment. This oscillating dipole moment is now a new EM source, resulting in scattered light.

When the wavelength of the scattered light is the same as the wavelength of the incoming wave, we say that the scattering is elastic. This means that no energy is lost in the scattering process. When energy is partly converted (e.g. to heat or vibrational energy) and the resulting wavelength is larger than the original wavelength, the scattering process is said to be inelastic. Examples of such inelastic scattering are Brillouin and Raman scattering. We will now discuss two elastic light scattering mechanisms: Rayleigh scattering and Mie scattering.

Rayleigh scattering (named after Lord Rayleigh) is caused by particles smaller than the wavelength of the incident light. It can occur in solids or liquids but it is mostly seen in gasses. The criterion for Rayleigh scattering is: $\alpha << 1$, with

$$\alpha = \frac{2\pi r}{\lambda}. \tag{6.103}$$

r is the radius of the particle and λ is the wavelength of the incident light. It can be shown that in the Rayleigh regime, shorter wavelengths are scattered more efficiently $(scaling 1/\lambda^4)$. This explaines why the daytime sky looks blue. The (shorter) blue wavelengths are redirected more efficiently towards earth than the (longer) red ones.

Mie scattering (named after Gustav Mie) is the general scattering theory without limitations on the particle size. For large particles, this theory converges to geometric optics. It can also be used for very small particles but in that case the Rayleigh theory is preferred due to the simplicity compared to the Mie theory. E.g. Mie scattering explains why clouds are white as it involves scattering of sunlight from particles (in this case water droplets) which are small but larger than the wavelength of the light. Other examples are scattering from dust, smoke, pollen,...

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Chapter 7

Waveguide optics

Contents

7.1 Introduction

Electromagnetic waves can transport energy, and thus also information, over very large distances. This has led to an explosive development of modern communications techniques. Transport through free space is inefficient however, because diffraction defocuses the energy. Therefore one looks for structures that guide the electromagnetic energy more efficiently. For light, one mainly employs dielectric waveguides. We will show in this chapter that such a waveguide gives rise to field distributions that propagate without change at their own speed. These field distributions are called the *eigenmodes* of the waveguide. The wave number corresponding to each mode is the propagation constant of the mode.

Optical fibers are a very important type of waveguide. They replace the electrical cables in modern communication networks (e.g. phone and internet). This is mainly the consequence of a much larger bandwidth and much smaller losses, compared to electrical connections.

Just like in electronics where one has moved from components on PCBs to monolithic ICs, the same move is happening in optics towards miniaturization and integration. Classical optical systems consist(ed) of a collection of components (lenses, mirrors, diffractive elements, sources) that had to be aligned carefully. Therefore they are often expensive and large. The idea of integrated optics originated in the sixties, and means that different optical functions (lasers, detectors, filters, couplers ...) are integrated on a single substrate. Waveguides, instead of free space, are used to guide light from one component to the next or within components. A major advantage of integra-

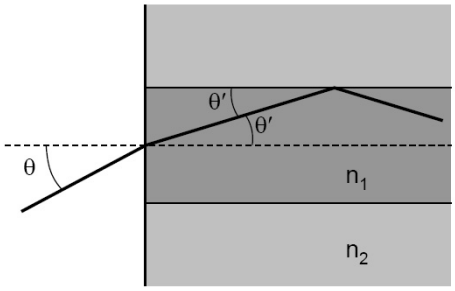

Figure 7.1: Step-index waveguide.

tion is that all components are collectively and precisely aligned by the lithographic processes at the moment of fabrication.

7.2 Waveguides with the ray approximation

Thus waveguides are optical systems that aim to confine the light. A simple type of waveguide is the step-index waveguide, see figure 7.1. In this waveguide the rays follow a zigzag route because of total internal reflection at the core-cladding interface (hence the term waveguide). To this end the angle θ' has to be sufficiently small:

$$\theta'_{\max} = \arccos\left(\frac{n_2}{n_1}\right) \tag{7.1}$$

This immediately means that the core has to have a higher index than the cladding. We calculate the maximum angle θ of the incident rays that undergo total internal reflection:

$$n_0 \sin\theta_{\max} = n_1 \sin\theta'_{\max} = n_1\sqrt{1-\left(\frac{n_2}{n_1}\right)^2} = \sqrt{n_1^2 - n_2^2} \tag{7.2}$$

In analogy with lenses we define a numerical aperture, NA. For $n_0 = 1$ one obtains:

$$NA = \sin\theta_{\max} = \sqrt{n_1^2 - n_2^2} \tag{7.3}$$

In the chapter about geometric optics (see section 3.2.8) we discussed another type of waveguide, the parabolic index or graded index waveguide (see figure 7.2). In these waveguides the trajectories are not zigzag, but sinusoidal, with every ray having the same period. Therefore, a special property of this type of guide is that all rays propagate with the same axial speed, regardless of their incidence angle. Again we can define a NA, but now it is dependent on the incidence position relative to the axis. The NA is largest at the axis.

The previous discussed two-dimensional waveguides. However, it is also possible to guide in three dimensions. For this purpose one needs a core surrounded on all sides by a lower index cladding. The most common types are the rectangular and the cylindrical (fiber) waveguide, see figure 7.3. The latter type can have a constant or parabolic core index profile.

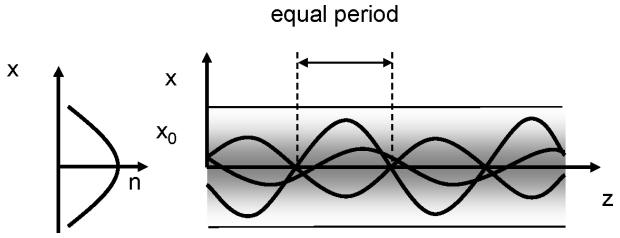

Figure 7.2: Graded index waveguide.

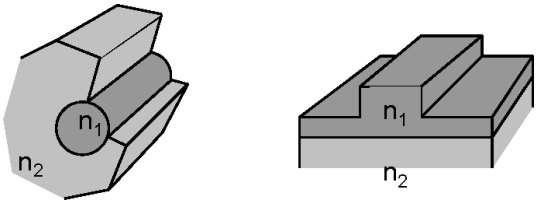

Figure 7.3: Three-dimensional waveguides.

One of the most important waveguide properties is that they can guide light around a bend (figure 7.4). If the radius of curvature R is not too small, then the majority of guided rays will be led through the bend. However, a small loss is unavoidable. If the index contrast $n_1 - n_2$ increases, the radius R can be smaller.

The most well known waveguide is the optical fiber. It usually consists of glass, but sometimes polymer is used (POF: polymer optical fiber). Typically the fiber has a core with diameter $50\mu m$, surrounded by a cladding with outer diameter $125\mu m$. The index difference between core and cladding is typically between 0.001 and 0.01. Because of the small diameter the fiber is very flexible, and it is used in many applications: optical communications, sensors, medical applications...More and more in communications one uses a fiber with a very small core ($< 10\mu m$). For this fiber geometric optics is not applicable anymore, and one has to use a more rigorous wave approach.

7.3 Modes in longitudinally invariant waveguide structures

In this section we consider structures that are invariant along the propagation direction z of the optical power (longitudinal direction), as shown in figure 7.5. Then the refractive index profile is written as: $n(\mathbf{r}) = n(\mathbf{r_t}) = n(x, y)$. An eigenmode of the waveguide structure is defined as a propagating or evanescent wave that keeps its transversal shape (thus the shape in the (x, y)-plane).

Therefore, a forward propagating eigenmode can be written as

$$\mathbf{E}(x, y, z) = \mathbf{E}(x, y)e^{-j\beta z} \tag{7.4}$$

$$\mathbf{H}(x, y, z) = \mathbf{H}(x, y)e^{-j\beta z} \tag{7.5}$$

Three parameters, all interdependent, are used to describe the propagation characteristics of the eigenmode. The first is the propagation constant β, the second is the effective refractive index

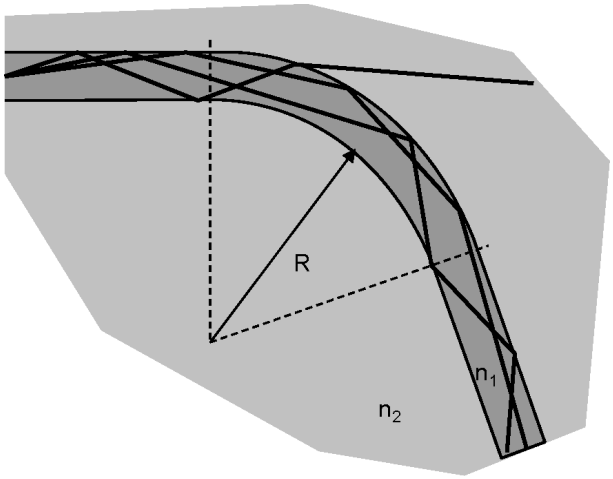

Figure 7.4: Bend in a waveguide.

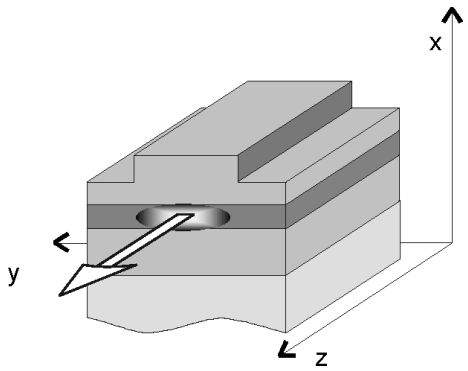

Figure 7.5: Example of a longitudinally invariant waveguide structure.

$n_{eff} = \beta/k_0$, and the third is the effective dielectric constant $\epsilon_{eff} = n_{eff}^2$. In the next section about slab waveguides it is shown that these are the eigenvalues of the eigenvalue equation derived from Maxwell's equations, with the eigenmodes as solutions.

Before embarking on a detailed study of the eigenvalue problem, we present a short overview of phenomena that are typical for lossless optical waveguides, so waveguides where:

$$Im(\epsilon(\mathbf{r}_t)) = 0 \tag{7.6}$$

We limit ourselves to the slab waveguide structure. This is a y-independent geometry, but the following properties are generally valid. Figure 7.6 shows the dielectric profile of a step index slab waveguide, with different types of eigenmodes. From Maxwell's equations one can derive the following characteristics:

- There are no eigenmodes with eigenvalue larger than the maximum of the dielectric function. So:
$$n_{eff} < \max(n(\mathbf{r_t})) \tag{7.7}$$
- The guided modes form a set of discrete eigenvalues, that are limited to the region:
$$n_{max} > n_{eff} > \max(n_{cladding}) \tag{7.8}$$
For these modes the radiation condition holds: $\lim_{\mathbf{r_t}\to\infty} \Psi(\mathbf{r_t}) = 0$, so the field profile is limited to the 'core' of the waveguide. We will see that there are waveguide structures without any guided mode.
- The radiation modes form a continuous set of eigenvalues, with:
$$n_{eff} < \max(n_{cladding}) \tag{7.9}$$
Radiation modes have an oscillating behavior along at least one side of the structure. One distinguishes between propagating (n_{eff} is real) and evanescent (n_{eff} is purely imaginary) radiation modes. In the latter case the field profile decays exponentially in the positive z-direction (hence there is no power transport by evanescent radiation modes in the z-direction).

An important property of eigenmodes is that they form a complete set. This means that an arbitrary field distribution can be described by a superposition of modes.

Note that there is a connection between the kinds of modes of a waveguide calculated with Maxwell's equations and with ray theory. Rays propagating in the core (=highest index) that are incident on the cladding (=lower index) will or will not be totally reflected. If the angle of the ray is larger than the critical angle for total internal reflection (TIR), then there will only be reflection (no transmission!) at the core-cladding interface. These rays are associated with guided modes, although the relation is not straightforward. If the incidence angle is smaller than the TIR critical angle there will be both reflection and transmission. This corresponds to radiation modes.

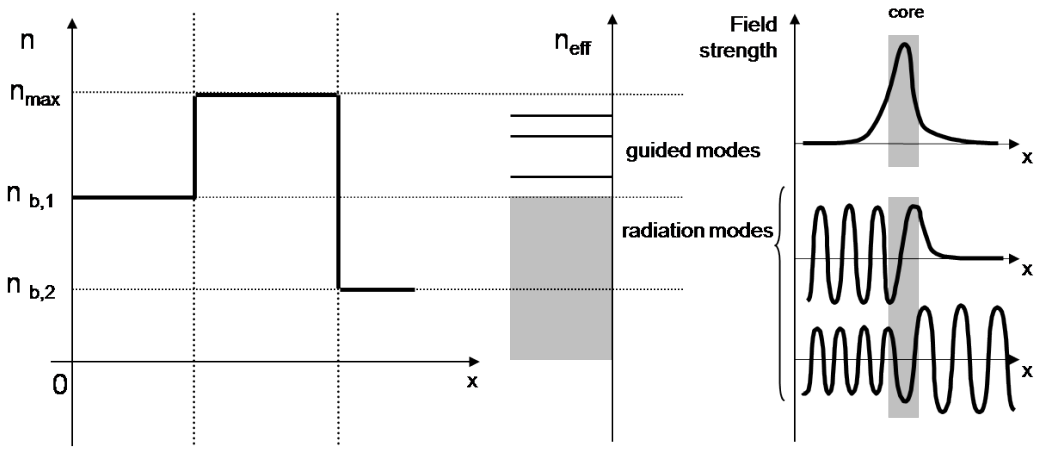

Figure 7.6: Propagating and radiation modes in a longitudinally invariant waveguide structure.

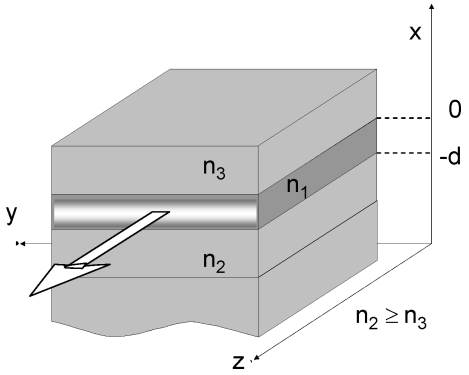

Figure 7.7: The three-layer slab waveguide.

7.4 Slab waveguide

7.4.1 Three-layer slab waveguide

To determine the eigenmodes of a longitudinally invariant waveguide we use an approximation. We consider a structure that is invariant both in the z and y-direction, see figure 7.7.

This waveguide consists of a plane dielectric layer (core), with thickness d, confined between two semi-infinite dielectric layers (cladding layers). The core material has index n_1, the lower dielectric or substrate has index n_2, and the upper dielectric or superstrate has index n_3, so that: $n_1 > n_2 > n_3$. We assume wave propagation in the z-direction. Because the structure is infinite in the y-direction all fields are independent of y. Thus, the complex amplitudes of the electric and magnetic field become:

$$\mathbf{E}(x,z) = \mathbf{E}(x)e^{-j\beta z} \tag{7.10}$$

$$\mathbf{H}(x,z) = \mathbf{H}(x)e^{-j\beta z} \tag{7.11}$$

Substitution of these two expressions in Maxwell's curl laws gives:

$$\begin{aligned} j\beta E_y &= -j\omega\mu_0 H_x \\ \frac{dE_y}{dx} &= -j\omega\mu_0 H_z \qquad \textbf{(TE)} \\ -j\beta H_x - \frac{dH_z}{dx} &= j\omega\epsilon_0 n^2 E_y \end{aligned} \tag{7.12}$$

and

$$\begin{aligned} -j\beta E_x - \frac{dE_z}{dx} &= -j\omega\mu_0 H_y \\ j\beta H_y &= j\omega\epsilon_0 n^2 E_x \qquad \textbf{(TM)} \\ \frac{dH_y}{dx} &= j\omega\epsilon_0 n^2 E_z \end{aligned} \tag{7.13}$$

The index n takes the values n_1, n_2 and n_3 in the corresponding media. The two-dimensional nature of the problem leads to two decoupled sets of equations: TE- and TM-polarized modes. In the TE case the H_x- and H_z-components are derived from the E_y-component. For TM E_x and E_z follow from H_y:

$$\begin{aligned} H_x &= -\frac{\beta}{\omega\mu_0} E_y \qquad \textbf{(TE)} \\ H_z &= \frac{j}{\omega\mu_0}\frac{dE_y}{dx} \end{aligned} \tag{7.14}$$

and

$$\begin{aligned} E_x &= \frac{\beta}{\omega\epsilon_0 n^2} H_y \qquad \textbf{(TM)} \\ E_z &= -\frac{j}{\omega\epsilon_0 n^2}\frac{dH_y}{dx} \end{aligned} \tag{7.15}$$

The y-components satisfy the following wave equations, obtained from (7.12) and (7.13) after eliminating the x- and z-components:

$$\frac{d^2E_y}{dx^2} + (n^2k_0^2 - \beta^2)E_y = 0 \qquad \textbf{(TE)} \tag{7.16}$$

$$\frac{d^2H_y}{dx^2} + (n^2k_0^2 - \beta^2)H_y = 0 \qquad \textbf{(TM)} \tag{7.17}$$

With k_0 the wave number in vacuum. Here $\mathbf{k}_0 n = \mathbf{k} = k_x\mathbf{e_x} + k_y\mathbf{e_y} + k_z\mathbf{e_z}$, with $k_z = \beta$.

Guided TE-modes of the three-layer slab waveguide

Calculating the TE-modes now amounts to solving the wave equation (7.16) for E_y, from which the other components follow with (7.14). For the eigenmodes we solve the transversal boundary condition problem. As a boundary condition we demand continuity of the tangential components of $\mathbf{E}$ and $\mathbf{H}$ at both interfaces $x = 0$ and $x = -d$. In addition we have the radiation condition for

$|x \to \infty|$: a guided mode has a field profile that exponentially decays towards infinity. All other field profiles are regarded as radiation modes.

From (7.14) we see that continuity of H_z implies that dE_y/dx is also continuous, in addition to E_y. Because we are looking for guided modes we demand that $E_y \to 0$ as $x \to \pm\infty$. Then the solution has the form:

$$\begin{aligned} E_y &= Ae^{-\delta x} && x \geq 0 \\ &= C\cos(\kappa x) + B\sin(\kappa x) && 0 \geq x \geq -d \\ &= De^{\gamma(x+d)} && -d \geq x \end{aligned} \tag{7.18}$$

with

$$\begin{aligned} \kappa &= \sqrt{n_1^2 k_0^2 - \beta^2} \\ \gamma &= \sqrt{\beta^2 - n_2^2 k_0^2} = \sqrt{(n_1^2 - n_2^2)k_0^2 - \kappa^2} \\ \delta &= \sqrt{\beta^2 - n_3^2 k_0^2} = \sqrt{(n_1^2 - n_3^2)k_0^2 - \kappa^2} \end{aligned} \tag{7.19}$$

Where we assume that the expression under the square root is positive, as we are now looking for guided modes, so $k_0 n_1 > \beta > k_0 n_2$. Continuity of E_y in $x = 0$ and $x = -d$ means:

$$A = C$$

$$C\cos(\kappa d) - B\sin(\kappa d) = D \tag{7.20}$$

so

$$\begin{aligned} E_y &= Ae^{-\delta x} && x \geq 0 \\ &= A\cos(\kappa x) + B\sin(\kappa x) && 0 \geq x \geq -d \\ &= (A\cos(\kappa d) - B\sin(\kappa d))\, e^{\gamma(x+d)} && -d \geq x \end{aligned} \tag{7.21}$$

Continuity of H_z or dE_y/dx implies:

$$\delta A + \kappa B = 0$$

$$(\kappa sin(\kappa d) - \gamma\cos(\kappa d))A + (\kappa\cos(\kappa d) + \gamma\sin(\kappa d))B = 0 \tag{7.22}$$

This homogeneous system has solutions that are not equal to zero if the determinant vanishes, thus:

$$(\kappa\cos(\kappa d) + \gamma\sin(\kappa d))\delta - (\kappa sin(\kappa d) - \gamma\cos(\kappa d))\kappa = 0 \tag{7.23}$$

We can write this eigenvalue equation in the following form:

$$\begin{aligned} \tan(\kappa d) &= F(\kappa d) \\ F(\kappa d) &= \frac{\kappa(\gamma + \delta)}{\kappa^2 - \gamma\delta} \end{aligned} \tag{7.24}$$

This transcendental expression is depicted in figure 7.8. Every intersection of the $\tan(\kappa d)$-curve with the $F(\kappa d)$-curve gives an eigenvalue or discrete mode β for the slab waveguide. The guided modes, and β, are a function of 5 parameters: λ_0, n_1, n_2, n_3 and d. We call $n_{eff} = \beta/k_0$ the effective index of a mode; this is the index that a mode 'feels'. It represents a kind of average

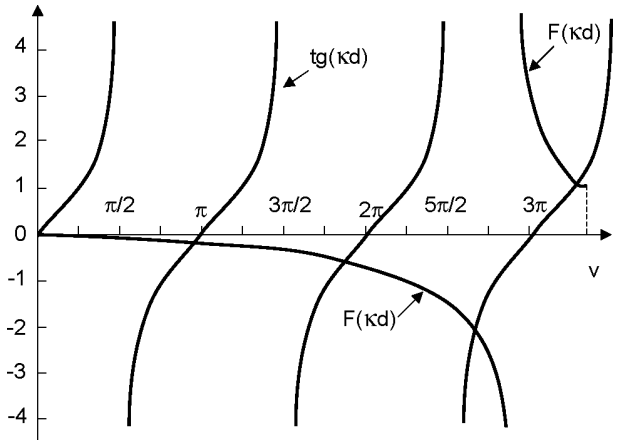

Figure 7.8: Solutions of the eigenvalue equation for a three-layer slab waveguide.

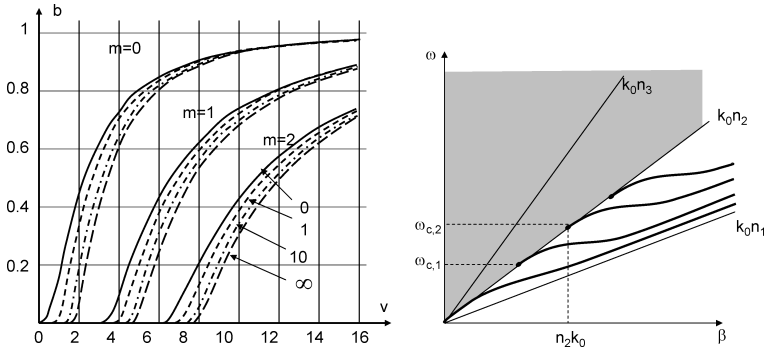

Figure 7.9: Schematic depiction of the dispersion relation for various TE-modes $m = 0, 1, 2, ...$ and asymmetry parameters $a = 0, 1, 10, \infty$.

refractive index, that corresponds to weighing the index at each location with the local strength of the TE-field.

One often uses the $\omega - \beta$-diagram, the graphical representation of the dispersion relation for the different modes, see figure 7.9b. We see on the figure that there are only discrete modes for $\beta > k_0 n_2$. The lowest order mode is called the fundamental mode or the 0^{th}-order mode. Each discrete mode also has a cutoff frequency (see below), for mode 2 it is $\omega_{c,2}$ (then β is equal to $k_0 n_2$). If the frequency increases then β increases too, and for large frequencies β approaches $n_1 k_0$. In that case the fields in the sub- and superstrate are strongly damped, and the field only 'sees' the core medium.

To obtain a graph as general as possible one employs normalized quantities: the normalized frequency V, the relative effective index b and the asymmetry parameter a:

$$V = k_0 d \sqrt{n_1^2 - n_2^2} \tag{7.25}$$

$$b = \frac{n_{eff}^2 - n_2^2}{n_1^2 - n_2^2} \tag{7.26}$$

$$a^{TE} = \frac{n_2^2 - n_3^2}{n_1^2 - n_2^2} \tag{7.27}$$

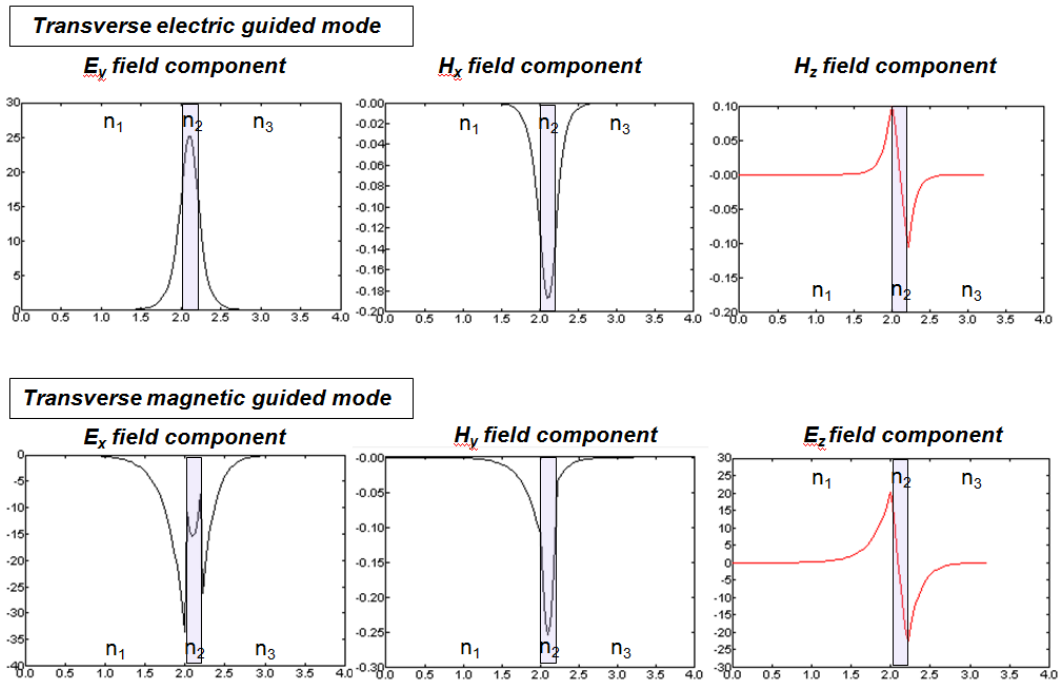

Figure 7.10: Field components of the guided TE and Tm mode in an asymmetric three-layer slab waveguide

This gives the dispersion curves in figure 7.9a, for the three lowest order TE-modes for different values of the asymmetry parameter a. Notice that for symmetrical waveguides $a = 0$ there is always at least one guided mode, and the next mode starts at $V = \pi$. For strongly asymmetrical waveguides there is no guided mode for $V < \pi/2$ and the second mode starts at $V = 3\pi/2$.

Guided TE and TM mode profiles in a single mode three-layer slab waveguide As an example, the guided transverse electric and transverse magnetic mode profiles in a single mode three-layer slab waveguide are plotted in figure 7.10. The slab waveguide structure considered is an asymmetric structure with $n_1 = 1.5$,$n2 = 3.5$ and $n_3 = 1$. Notice the discontinuity in the E_x field component of the transverse magnetic guide mode.

Cutoff frequency For certain frequencies if κd increases, or β decreases, γ becomes complex. If γ is complex, then $F(\kappa d)$ is complex too and the $F(\kappa d)$-curve stops, so there are no more intersections with $\tan(\kappa d)$. The value of κd where this happens is given by $\gamma = 0$ or

$$\beta = n_2 k_0 \tag{7.28}$$

The cutoff happens first with γ because we assumed that: $n_1 > n_2 > n_3$. Expression (7.24) becomes: $\tan(\kappa d) = \frac{\delta}{\kappa}$ (with $\beta = n_2 k_0$). Working this out we obtain:

$$k_{0,c,m} d \sqrt{n_1^2 - n_2^2} = \arctan\left(\sqrt{\frac{n_2^2 - n_3^2}{n_1^2 - n_2^2}}\right) + m\pi \tag{7.29}$$

With $\omega_{c,m} = k_{0,c,m} v$ as cutoff frequency of the m^{th} TE-mode. If ω decreases and approaches the cutoff frequency ω_c of a mode, then the longitudinal component β of the wave vector in the slab

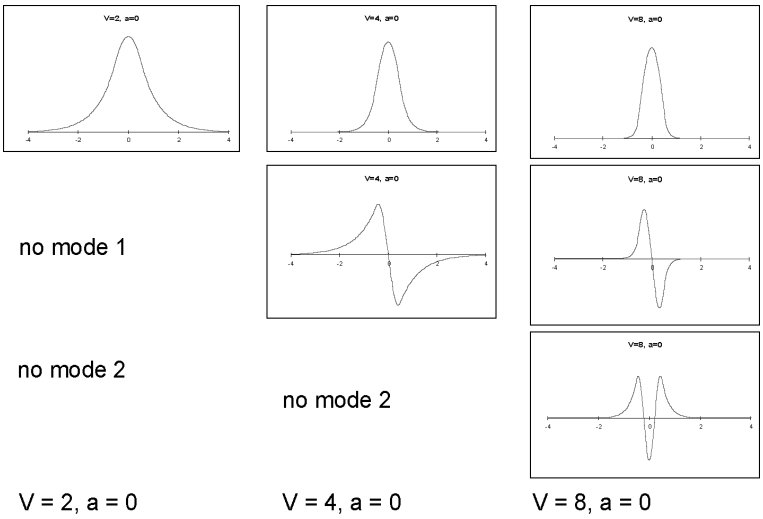

Figure 7.11: Field profiles of the lowest order TE-modes.

waveguide will decrease and become equal to $\beta = n_2 k_0$ at cutoff. If the frequency decreases further and becomes smaller than the cutoff frequency, the guided mode changes into a radiation mode.

In figure 7.11 we show the E_y field distributions of the lowest order TE-modes for a symmetrical slab waveguide for $V = 2,\ 4,\ 8$. As V increases the fundamental mode is more concentrated in the core. The field profiles for higher order modes decay less rapidly in the cladding than for lower order modes, at the same value of V.

Up until now we discussed the TE-problem. The solution of the TM-problem leads to analogous results: The field profiles of H_y for TM-modes are similar, but not identical to those of E_y. The same holds for the propagation constants of the i^{th} TM-mode versus the i^{th} TE-mode. This means in fact that a 'single mode' waveguide has two guided modes: The fundamental TE- and the fundamental TM-mode.

Radiation modes in a three-layer slab waveguide

In the previous section we derived a finite number of guided modes. These modes do not constitute a complete set as they are unable to represent radiation outside of the core. We also noticed that there are no more guided modes if $\beta \leq n_2 k_0$. Still we obtain standing wave patterns, as γ becomes complex in expression (7.18) (proof: see below). This pattern is a superposition of radiation incident from $x = -\infty$ and radiation going to $x = -\infty$. If $n_3 k_0 < \beta < n_2 k_0$ then there is still total internal reflection at the core-superstrate interface, and the field decays exponentially in the superstrate. However, on the side of the substrate there is a standing wave pattern. If $\beta < n_3 k_0$ then there is radiation from and to infinity at both sides of the core. See figure 7.6 for an overview.

As an example we calculate the TE radiation modes in the case $n_3 k_0 < \beta < n_2 k_0$. We obtain the solution for E_y, based on (7.21) and with a complex γ:

$$\begin{aligned} E_y &= Ae^{-\delta x} && x \geq 0 \\ &= A\cos(\kappa x) + B\sin(\kappa x) && 0 \geq x \geq -d \\ &= (A\cos(\kappa d) - B\sin(\kappa d))\cos(\rho(x+d)) + C\sin(\rho(x+d)) && -d \geq x \end{aligned} \tag{7.30}$$

With $\rho = j\gamma$ being:

$$\rho = \sqrt{n_2^2 k_0^2 - \beta_2} \tag{7.31}$$

Here we used $D\sin(\rho(x+d))+C\cos(\rho(x+d))$ instead of $D\,e^{\gamma(x+d)}$ in equation (7.18). The continuity conditions for H_z or dE_y/dx in $x = 0$ and $x = -d$ result in the following:

$$\delta A + \kappa B = 0 \tag{7.32}$$

$$\kappa\sin(\kappa d)A + \kappa\cos(\kappa d)B - \rho C = 0 \tag{7.33}$$

For the guided modes the continuity relations are a homogeneous system, and setting the determinant to zero one obtains an eigenvalue equation. Here however we have two equations for three unknowns, so we choose one and the other two are determined by solving the resulting inhomogeneous system. Thus, we have no eigenvalue equation and the values for β cover a continuum in the area $\beta < k_0 n_2$, see figures 7.6 and 7.9. The radiation modes do not satisfy the radiation condition and they have infinite energy, so these modes are not physical, but they are mathematical solutions of the eigenmode equation. We need these extra nonphysical solutions to obtain a complete set of eigenmodes. With this complete set we can describe every possible physical field propagating in the z-direction. This description consists of a discrete sum of guided modes, plus an integral over the radiation modes. The integration works because the contributions of waves from infinity cancel each other, so the total propagating energy is finite and equal to the energy of the field.

7.5 Optical fiber waveguides

7.5.1 Introduction

The optical fiber is the best medium for transporting a large number of signal with high bandwidth over large distances. In this section we discuss some main properties of fibers, where we focus on monomode and multimode glass fiber (or silica fiber: SiO_2), but we also mention other types such as polymer fibers (POF: polymer optical fiber). We describe the propagation of light in these fibers, using both the ray and the mode concept. In this way we introduce the phenomenon of dispersion. This concept incorporates all effects caused by the property that the propagation speed in a waveguide is not uniform but has a certain spread, which puts limits on the information capacity. However, if one uses the right type of fiber and source the capacity is virtually unlimited. Furthermore we describe the attenuation properties. Modern fibers exhibit a spectacularly low attenuation. A decay with a factor 2 happens only after a few tens of kilometers, if one operates at the correct wavelength.

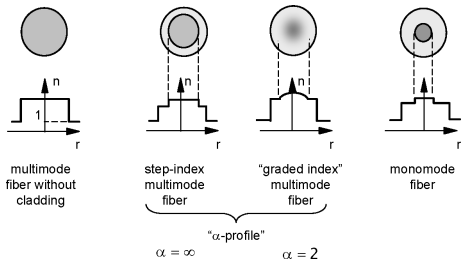

Figure 7.12: Overview of a number of different types of fiber.

7.5.2 Types of fibers

All optical fibers possess a cylindrical geometry with a core having a larger index than the surrounding medium. In step index multimode fibers the core has a diameter of $50\mu m$ to $300\mu m$, and it has a cladding material with lower index, see figure 7.12. The graded index (GRIN) multimode fiber however has a radially varying index profile for the core, that is approximately parabolic and the core diameter is on the same order as the core in the step index fiber. The index profile of both types can be described by the same mathematical expression:

$$n(r) = n_0(1 - 2\Delta(r/a)^\alpha)^{1/2} \tag{7.34}$$

With $\alpha = 2$ for graded index fiber and $\alpha = \infty$ for step index fiber.

The single-mode or mono-mode fiber has a much smaller core diameter: $5\mu m$ to $10\mu m$, where both step index and graded index (GRIN) are used as index profile. In most cases the index difference between core and cladding is on the order of 0.001 to 0.01. The fibers for long-distance communications are standardized: the core diameter for multimode fibers is $50\mu m$ to $62.5\mu m$, the one for single-mode fibers is $\pm 9\mu m$. The outer diameter is always $125\mu m$.

The most important fiber is fabricated of amorphous silicon, that has a refractive index of about 1.5. To achieve the index difference between core and cladding one adds impurities to the material during production (B_2O_3 and F are used for the cladding, P_2O_5 and GeO_2 for the core). The most important manufacturing method is the *preform*-method. Here one starts with a thick ($2cm$ diameter and $50cm$ length) cylindrical rod that is stretched under high temperature to a fiber that is 100 times thinner and 10000 times longer. This process needs extreme precision.

For communication over short distances ($< 100m$) one increasingly uses polymer (POF) fibers. They are cheaper and it is much easier to obtain a high quality fiber ending. Generally the core diameter of POFs is almost equal to the cladding diameter (so a very thin cladding), and therefore the fiber is always multimode. The cladding diameter varies from $125\mu m$ to a few mm. Usually the index difference between core and cladding is higher than for glass fibers, and the propagation losses are larger.

7.5.3 Optical fibers: ray model description

When a short light pulse is sent into an ideal optical fiber, there would be no losses nor deformation, as shown in the upper part of figure 7.13. Unfortunately, in reality each fiber shows attenuation and dispersion. Because of attenuation, the pulse will have to be amplified after a distance.

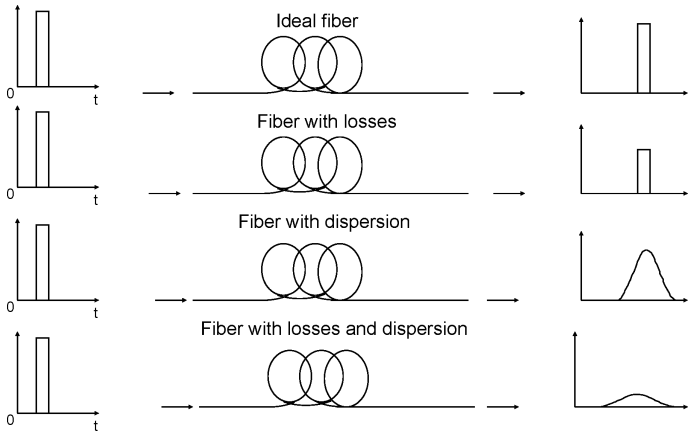

Figure 7.13: Propagation of a short light pulse in an optical fiber.

Otherwise losses would make detection impossible. Dispersion limits the data that can be sent through the fiber: the speed of light in an optical fiber varies a little bit, so that a short light pulse will spread out in time (this effect becomes worse with increasing fiber-length). Therefore, light pulses cannot be too close to each other in order not to overlap (which means loss of information).

Propagation of light in fibers

Propagation of light in fibers can be described by different theories. The most important theories are the ray theory and the electromagnetic theory. The ray theory is the high frequency limit of the electromagnetic theory and is valid if the optical variations are large in comparison with the wavelength. In the case of optical fibers, this means that the ray theory can be used for the description of multimode fibers, but fails for the description of monomode fibers (i.e. the ray theory in its simplest form).

Let us look again at figure 7.1 for the ray description of the multimode step-index fiber. We have seen that the numerical aperture of the fiber is defined as:

$$NA = \sin\theta_{\max} = \sqrt{n_1^2 - n_2^2} \approx \sqrt{2n\Delta n} \tag{7.35}$$

The numerical aperture of the fiber determines whether or not the incident rays at the air-(fiber)core interface, will be conducted via TIR at the core-cladding interface. Incident rays at the air-core interface with an angle smaller than the critical angle $\theta_{\max}$ will be conducted.

Example

For a typical optical fiber with $\Delta n = 0.01$ and $n \approx 1.5$, is $\theta_{\max} = 10^o$ and $NA = 0.17$.
For a typical POF with $\Delta n = 0.08$ and $n \approx 1.5$, is $\theta_{\max} = 30^o$ and $NA = 0.5$.

Multi-path dispersion in a step-index fiber

The simple ray model for a step-index multimode fiber is able to explain multi-path dispersion. This type of dispersion occurs because the different propagating rays through a fiber with length

L have a different propagation time. An axial ray will propagate with the highest speed, $v = c/n_1$, whereas a ray at the critical TIR angle will propagate slowest, with $v = c\cos\theta'_{\max}/n_1 = c\, n_2/n_1^2$. The time difference between both is given by:

$$\Delta T = \frac{L}{c}\frac{n_1}{n_2}\Delta n \approx \frac{L}{c}\Delta n. \tag{7.36}$$

Multi-path dispersion is defined as:

$$\frac{\Delta T}{L} = \frac{\Delta n}{c} \quad [ns/km]. \tag{7.37}$$

This time dispersion is proportional to the length L of the fiber and is expressed in ns/km. For a (typical) $\Delta n = 0.01$, the multi-path dispersion is equal to $34ns/km$. The maximum bit rate, B, is limited by this dispersion because it widens the pulses. The bandwidth, Δf, necessary for a given bit rate depends on the used coding technique, but it is at least equal to half the bit rate. As a rough rule of thumb one can assume that:

$$B_{\max} \approx 2\Delta f \approx \frac{1}{\Delta T}. \tag{7.38}$$

A quantity that is often employed in optical fiber communications is the bandwidth-length product, expressed in $MHz.km$:

$$\Delta f.L = \frac{c}{2\Delta n} \quad [MHz.km]. \tag{7.39}$$

For $\Delta n = 0.01$ we get a bandwidth-length product of $15MHz.km$. This means that $1km$ of fiber is limited to $30Mb/s$, while a length of $10km$ only reaches $3Mb/s$. It is clear that one has to keep the refractive index difference Δn small.

Multi-path dispersion in graded index (GRIN) fiber

We saw previously (section 3.2.8) that the rays in a graded index medium with a parabolic profile follow a sine trajectory, instead of a zigzag. Furthermore we noted that sine period is independent of the incidence angle! Thus in a GRIN medium the guiding does not happen because of TIR, but because of a gradual deflection. The maximum incidence angle $\theta_{\max}$ depends on the incidence position, and it is maximal in the middle of the fiber core, and zero at the core-cladding interface.

Two different rays from a point A to a point B inside the core propagate a different length at different speeds. However, Fermat's principle, $T = \int_A^B \frac{n}{c} ds$, teaches us that the elapsed time is the same for all possible neighboring rays. This means that all these rays have the same longitudinal velocity, thus *there is no multi-path dispersion!* This property is unique to a parabolic index profile.

7.5.4 Optical fibers: electromagnetic description

Guided modes for step-index fiber

Just like the description of the slab waveguide we can calculate the guided modes of the optical fiber via the rigorous Maxwell equations. We do not consider the mathematical details, but present an overview.

Consider a straight step-index fiber, where the z-direction corresponds to the propagation direction. We look for solutions to Maxwell's equations in the form of modes. This means we look for transversal field distributions that do not change upon propagation (in the z-direction). The only evolution is a periodic phase change, with a characteristic propagation constant β. We search for solutions in the form (complex amplitude representation):

$$\mathbf{E}(r, \phi, z) = \mathbf{E}(r, \phi)e^{-j\beta z}, \tag{7.40}$$

$$\mathbf{H}(r, \phi, z) = \mathbf{H}(r, \phi)e^{-j\beta z}. \tag{7.41}$$

Because of the fiber symmetry we work with cylindrical coordinates. The modes are rigorous solutions to Maxwell's equations, so they do not couple or exchange energy. Mathematically this means the modes are orthogonal. Analogous as for the slab waveguide, β is related to the propagation speed. Different modes have different β's, so dispersion will occur. This phenomenon is physically equivalent to multi-path dispersion, as discussed in the previous section with the ray model.

Although the cylindrical step-index fiber seems like a simple structure, the calculation of the modes is not easy. We give the most important results. Qualitatively the modes are equivalent to the solutions for the slab waveguide, mutatis mutandis. We obtain a discrete set of guided modes with different polarization states and field profiles. There are TE- and TM-modes (with $E_z = 0$ and $H_z = 0$ respectively), but also complex hybrid solutions that are called HE- and HM-modes ($E_z \neq 0$ and $H_z \neq 0$). Because of the two-dimensional character of the fiber, the modes are characterized by two numbers. One gets that the HE_{11}-mode is the lowest order mode. It has a profile with maximum at the core axis and consists of two variants with orthogonal polarization that have the same β, thus they are degenerate (the origin of this is the rotational symmetry). Every linear combination of these two degenerate modes is also a mode. The V-number is also important here:

$$V = k_0 R\sqrt{n_1^2 - n_2^2} \approx k_0 R\sqrt{2n\Delta n}. \tag{7.42}$$

in which n is the average of the core and cladding refractive index and Δn is the refractive index contrast between core and cladding. Figure 7.14 shows the V-number in function of the effective refractive index. We see indeed that HE_{11} is the lowest order mode, and that the fiber is monomode if $V < 2.405$. Thus, for a typical fiber with $\Delta n = 0.0025$ and $\lambda_0 = 1.5\mu m$ the fiber diameter has to be smaller than $13\mu m$ to be monomode. Remember that the standard single-mode fiber has a diameter of $9\mu m$. This fiber is no longer monomode if the wavelength is smaller than $1\mu m$, which is called the cutoff wavelength.

Dispersion

In the context of optical fibers, the term dispersion is used for all effects that cause the light to propagate not with one single speed, but with a variety of speeds (in general with very small speed differences!). All forms of dispersion create a widening in time of an optical pulse.

A monochromatic plane wave propagating in a uniform medium has a well determined velocity. If we couple this monochromatic wave into a multimode fiber, a number of different modes will be excited. Even in spite of the fact that only one wavelength was excited, multiple modes will propagate with slightly different speeds. This is multimode dispersion.

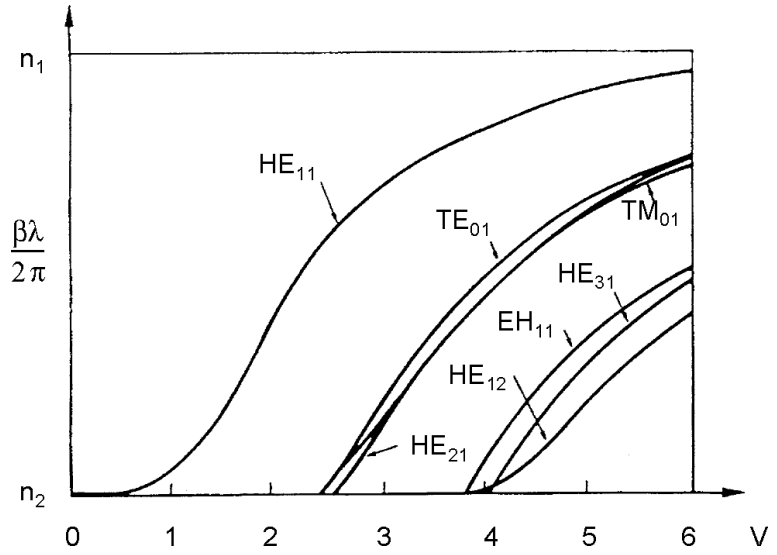

Figure 7.14: Effective index as a function of V-number for an optical fiber waveguide.

If the source is not perfectly monochromatic (which is generally the case), each mode will be excited at various frequencies. This leads to waveguide and material dispersion for each mode.

In chapter 6 we introduced the concepts effective refractive index (n_{eff}), phase velocity (v_p), group velocity (v_g) and group index (N):

$$v_p = \left(\frac{\beta}{\omega}\right)^{-1} \qquad \Rightarrow \qquad n_{eff} = \frac{\beta}{k} = \frac{c}{v_p} \tag{7.43}$$

$$v_g = \left(\frac{d\beta}{d\omega}\right)^{-1} \qquad \Rightarrow \qquad N = \frac{c}{v_g} = c\frac{d\beta}{d\omega} = n - \lambda_0 \frac{dn}{d\lambda_0} \tag{7.44}$$

We know that the information in a light wave propagates with the group velocity, so the propagation time equals:

$$t = \frac{L}{v_g} = L\frac{d\beta}{d\omega} = \frac{L}{c}N = \frac{L}{c}\left(n - \lambda_0 \frac{dn}{d\lambda_0}\right). \tag{7.45}$$

Material dispersion

The group velocity depends on the wavelength. Thus, if a source has a certain spectral width $\Delta\lambda_0$, the wavelength components will propagate with different speeds, and the pulse widens. We calculate the time difference because of $\Delta\lambda_0$ and $n(\lambda_0)$.

$$\Delta t = t_{\max} - t_{\min} = \frac{dt}{d\lambda_0}\Delta\lambda_0 = \frac{L}{c}\frac{dN}{d\lambda_0}\Delta\lambda_0 = -\frac{L}{c}\lambda_0 \frac{d^2n}{d\lambda_0^2}\Delta\lambda_0 \tag{7.46}$$

We call:

$$\frac{|\Delta t|}{L\Delta\lambda_0} = \frac{\lambda_0}{c}\frac{d^2n}{d\lambda_0^2} \qquad [\frac{ps}{km.nm}] \tag{7.47}$$

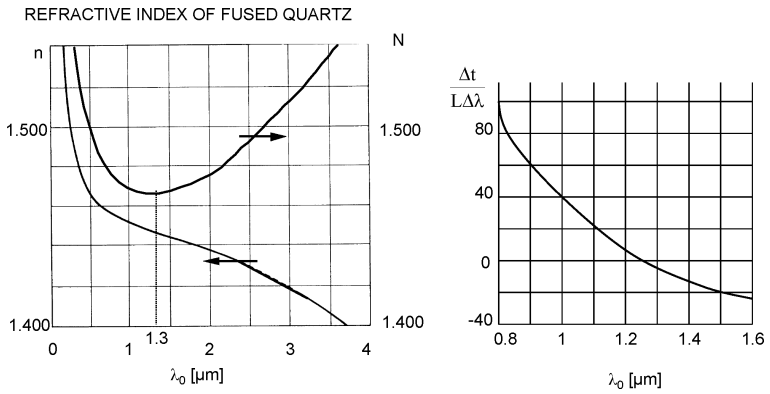

Figure 7.15: Material dispersion.

the material dispersion coefficient. Sometimes the dimensionless material dispersion coefficient is used: $Y_m = -\lambda_0^2 d^2 n / d\lambda_0^2$. Note that for a small L the time difference because of the spectral width is negligible. But, because a fiber is several km long, material dispersion is significant!

Figure 7.15a shows the refractive index and group index of quartz glass in function of the wavelength. Note that the group index reaches a minimum at $1.3\mu m$. This means that a pulse travels fastest through quartz glass at that wavelength.

From figure 7.15b the differential deceleration per unit of length and spectral width is equal to $0ps/(km.nm)$ at $\lambda_0 = 1.3\mu m$. At $\lambda_0 = 1.55\mu m$ the material dispersion coefficient is about $20ps/(km.nm)$. This means that a pulse from a laser with spectral width $1nm$ will widen about $20ps$ per traversed km.

Waveguide dispersion

Because an optical fiber is not a uniform medium, and a propagating light beam is not a plane wave but a set of guided modes, dispersion of a non-monochromatic wave is more complex than in the previous section. We have to take waveguide or intramodal dispersion of the propagation constants of all guided modes into account. Even if the refractive index is not wavelength dependent, each mode will disperse because the propagation constant is not perfectly linear with frequency (as in a uniform medium). The propagation constant is a kind of weighted average over core and cladding material, and this factor is frequency dependent.

Figure 7.16 shows the material dispersion (dimensionless material dispersion coefficient) Y_m and waveguide dispersion Y_w in function of wavelength for different fiber diameters. We note that for the most used fiber of $9\mu m$ the material dispersion dominates and it reaches a minimum at about $1.3\mu m$.

The table below gives an overview of values for the different dispersion effects (material, waveguide and multi-path dispersion) in a fiber for different sources (LED: Light Emitting Diode, LD: laser diode, SLD: single longitudinal mode laser diode) at three important wavelengths. The spectral width of the sources is approximately: LED: $\Delta\lambda_0/\lambda_0 \approx 0.04$, LD: $\Delta\lambda_0/\lambda_0 \approx 0.004$, SLD: $\Delta\lambda_0/\lambda_0 \approx 0.0004$.

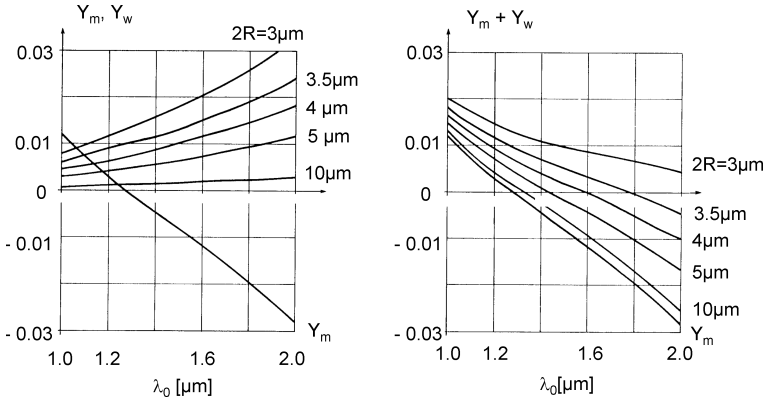

Figure 7.16: Material dispersion Y_m, and wavelength dispersion Y_w, in function of wavelength for different fiber diameters.

	$\Delta t/L \quad [ns/km]$	**fiber type** step-index	graded index	single mode
multimode dispersion		15	0.5-0.05	0
material + waveguide dispersion	LED @0.9μm	2	2	2
	LD @0.9μm	0.2	0.2	0.2
	LED @1.3μm	0.1	0.1	0.1
	LD @1.3μm	0.01-0.001	0.01-0.001	0.01-0.001
	LED @1.55μm	1	1	1
	LD @1.55μm	0.1	0.1	0.1
	SLD @1.55μm	0.01	0.01	0.01

7.5.5 Attenuation in optical fibers

In this section we succinctly describe propagation losses in waveguides. There are different loss factors: interaction of light and matter leads to absorption, imperfect guiding causes scattering and radiation. If the origin of the loss is spread evenly over the guide, the guided optical power decreases exponentially with propagation distance: $P(z) = P_0 e^{-\alpha z}$, with α the attenuation coefficient.

Absorption losses

Absorption of light during propagating through a material is caused by the interaction of photons with energy levels of the material. A dielectric such as SiO_2 has a number of important absorption peaks. The UV region has a strong absorption peak because of transitions between electron levels. The IR region has a peak from transitions associated with molecule vibrations (SiO bindings). Although these features lie outside of the optical window, their 'tails' cause important absorption in the optical domain. In figure 7.17 we see a region from about $1\mu m$ to $1.5\mu m$ between the tails with a low attenuation. However, on top of these UV and IR tails there are also narrow peaks originating from material impurities. There are small absorption peaks at $2.73\mu m$, $1.39\mu m$ (second harmonic) and $0.93\mu m$ (third harmonic: see figure 7.17) from OH bindings in the material.

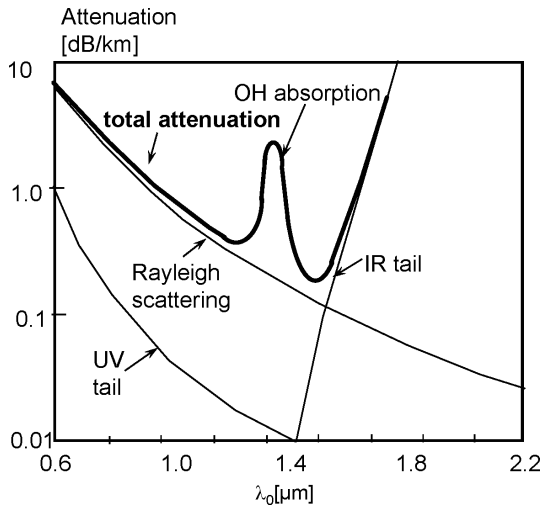

Figure 7.17: Attenuation in optical fiber by absorption and scattering.

Scattering losses

Scattering is caused by spatial variations of the refractive index (volume or Rayleigh scattering: $\alpha \sim 1/\lambda_0^4$, see figure 7.17) or by roughness of the boundary interfaces of the waveguide (surface scattering). These interfaces are the etched surfaces that determine the waveguide, or the interfaces between two layers grown on each other. In practice the surface scattering is the most important. Based on some simplifying assumptions one finds an approximate equation for the boundary surface scattering loss:

$$\alpha = \alpha_{scat} \frac{(\Delta n)^2 E_s^2}{P} \tag{7.48}$$

Here Δn is the index contrast, P is the optical power and E_s is the field strength at the boundary (larger for higher order modes!). The constant α_{scat} is determined empirically and depends on the fabrication process.

One finds that the total absorption minimum of a glass fiber is $0.15dB/km$ at $1.55\mu m$. There is another important local minimum of $0.4dB/km$ at $1.3\mu m$. Optical communications uses the following wavelengths almost exclusively: $1.55\mu m$ for the most demanding long distance applications, $1.3\mu m$ for less demanding medium-range systems and $0.85\mu m$ for short connections ($< 100m$).

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Chapter 8

Photon Optics

Contents

Classical electromagnetism succeeded in the explanation of a lot of optical phenomena, but failed in the description of other experiences. This became clear in the beginning of the twentieth century. It led to the development of a quantum-electromagnetic theory, often called quantum-electrodynamics (QED). In the optical world it is also called quantum optics or photon optics.

8.1 The photon

Light consists of particles, called photons. A photon has zero rest mass and carries electromagnetic energy. It also has a momentum and an intrinsic angular momentum (spin) that can be associated with the polarization of the light. The photon travels at the speed of light in vacuum and at a slower speed in a material. Photons also have a wavelike character that allows us to explain interference and diffraction. The fact that the blackbody radiation spectrum could not be explained with classical electromagnetism led to the concept of the photon. Max Planck solved the problem by postulating that the electromagnetic energy, radiated from a resonator, is quantized.

8.1.1 Photon energy

Photon optics states that the total energy in an electromagnetic mode is quantized in discrete energy levels separated by a finite interval. We then say that the mode contains a discrete number of photons. If the mode has a frequency ν, the energy difference between successive energy levels — thus the energy of the photon — is given by:

$$E = h\nu = \hbar\omega \tag{8.1}$$

with h Planck's constant ($h = 6.626\ 10^{-34} Js$) and $\hbar = h/2\pi$. The concept of a mode is not so trivial here. In a closed cavity with finite dimensions there are a number of electromagnetic modes satisfying the boundary conditions, each of them containing a discrete number of photons (at a given

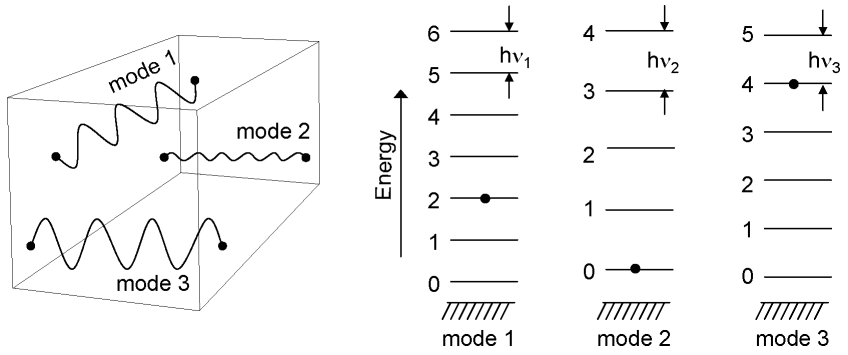

Figure 8.1: Electromagnetic modes.

time). Each of these modes has a different frequency, a different field distribution and a different polarization. This is illustrated in figure 8.1. In a waveguide there are — at a certain frequency — also a finite number of propagating modes and the power flux of each of these contains a discrete number of photons in a finite time interval. A Gaussian beam is a mode of the free space, and again we can state that a discrete number of photons will pass through a plane perpendicular to the direction of propagation in a finite time interval. We could say that the energy in an electromagnetic mode is given by the number of photons multiplied with the photon energy. This is however *not* correct. When a mode contains n photons, the energy E_n equals to:

$$E_n = (n + \frac{1}{2})h\nu, \quad n = 0, 1, 2 \ldots \tag{8.2}$$

When the mode does not contain any photons, there is still an energy $E_0 = \frac{1}{2}h\nu$ in this mode. This energy is called the zero-point energy and plays an important role in spontaneous emission in atoms. Because the energy of the photon is proportional to the frequency, it is logical that the particle nature of electromagnetic radiation becomes more important for increasing frequencies. In microwaves, the particle nature is seldom relevant, while X-rays and gamma–rays nearly always act as particles. Light is situated between these two extremes. Therefore, the wavelike character is apparent on some occasions, and the particle nature on others.

8.1.2 Photon position

A photon has both a spatially distributed and a localized character. The first is the consequence of the wavelike character, while the second is caused by its particle nature. When photons are converted into electric energy with a detector, we perceive the particle nature. No matter how small the detector is, it will either detect a photon in its vicinity or it will not detect it, even if the photon is carried by a long stretched electromagnetic mode. The probability that a detector with surface area dA, placed perpendicular to the light bundle, is going to detect a photon is proportional to the intensity (the Poynting vector) of the optical mode at that location. This means that if a photon is incident on a semi-transparent mirror that reflects 50% and transmits 50%, the photon has a 50% chance of being reflected and a 50% chance of passing through.

8.1.3 Photon momentum

The photon momentum is in a trivial way related to the wave vector concept: $\mathbf{p} = \hbar\mathbf{k}$. The photon propagates in the direction of the wave vector and the magnitude of its momentum is: $p = \hbar k = h/\lambda = E/c$. Energy and momentum are thus proportional to each other. If the photon is carried by a plane wave, the k-vector is uniquely defined and thus also the momentum. However, if the photon is carried by a more complex electromagnetic mode, its momentum becomes a statistical quantity that has a certain value with a certain probability. When photons interact with materials, there is always the conservation of energy and momentum. This means that if a photon is incident on a material and absorbed by this material, not only the energy of the photon is transferred to the material, but the material undergoes a force due to the momentum of the photon, and thus accelerates. This is called the radiation pressure exerted by the photons.

> Compare the forces on 2 plane plates, perpendicularly illuminated with photons: a black plate absorbing the incident photons perfectly and a perfectly mirroring plate reflecting the photons.

8.1.4 Photon polarization

Each elliptical polarization can be seen as the superposition of two linear polarizations or as the superposition of a right-handed and a left-handed circular polarization. With the latter we can associate the concept spin. We say that the spin of photons is quantized to two discrete values:

$$S = \pm\hbar \tag{8.3}$$

For a non-circularly polarized wave we can say that there is a certain probability the photons have the one spin, and for the rest the other spin.

8.1.5 Photon interference

In an interference situation, the wave-particle duality is completely apparent. When for example a plane wave is incident on a plate with two slits, an interaction pattern will arise behind the plate (see figure 8.2). Even if the plane wave only contains one single photon, a small detector will detect the photon with a probability proportional to the intensity distribution of the interference pattern. We can however determine that if we place the detectors at the slits, each photon is detected only by one detector. In other words, we can determine experimentally that the photon 'passes through both slits'. But also, a detector placed near one of the slits either detects the photon or does not detect it.

8.1.6 Photon time

If a monochromatic wave carries photons, the energy of these photons is known exactly. However, the wave is then infinitely long in time, and the time needed for detection of the photon is completely undetermined. If a light bundle has a finite duration, this automatically means that the light is not monochromatic, and the energy of the photons in that bundle can not be known

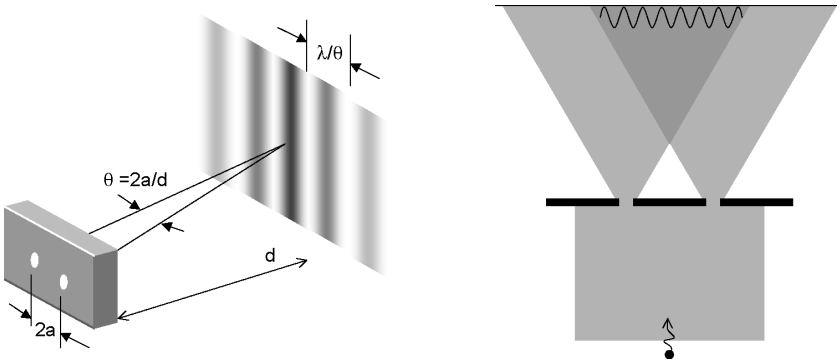

Figure 8.2: Young's two-slit experiment with one single photon.

exactly. The duration and spectral width are inversely proportional to each other and we can write:

$$\sigma_\omega \sigma_t \geqslant 1/2 \tag{8.4}$$

This relationship is rewritten for photons as:

$$\sigma_E \sigma_t \geqslant \frac{\hbar}{2} \tag{8.5}$$

It is called the time-energy uncertainty.

8.2 Photon streams

In the previous section we studied properties of a single photon. Now we treat photon streams.

8.2.1 Mean photon flux

The concepts optical power density, optical power and optical energy can be converted into a quantum-quantity by dividing by the photon energy. The optical power density (unit: W/m^2) is then converted into a mean photon flux density (unit: $fotonen/(s.m^2)$). Optical power (unit: W) is converted into a mean photon flux (unit: $fotonen/s$). Optical energy (unit: J) is converted into a number of photons.

Moonlight for example corresponds to a mean photon flux density of $10^8 fotonen/(s.cm^2)$. Thus, if the light of the moon is incident on a small aperture of $1\mu m^2$, one photon per second will pass through this aperture. A simple mnemonic for the conversion of optical power into photon flux is the following: for light with a wavelength of $0.2\mu m$, a power of $1nW$ corresponds to (on average) one photon per ns. For a wavelength of $1\mu m$, $1nW$ contains 5 photons per ns.

8.2.2 Photon flux statistics

The mean photon flux is proportional to the optical intensity, but the exact time on which photons arrive on the detector is random normally. If the intensity is high, the average 'arrival frequency' of

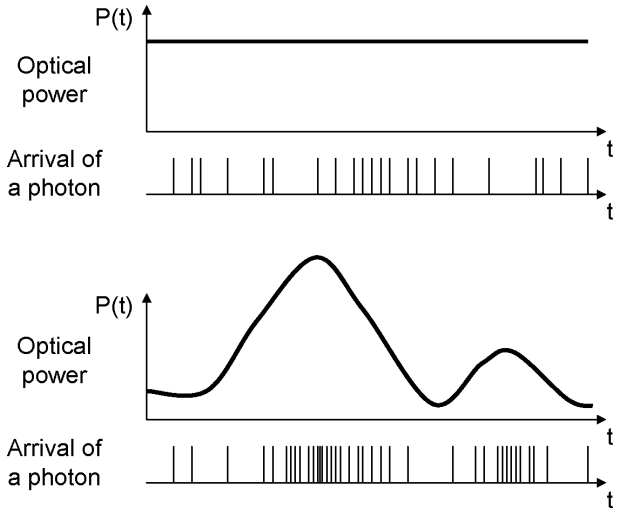

Figure 8.3: Optical power.

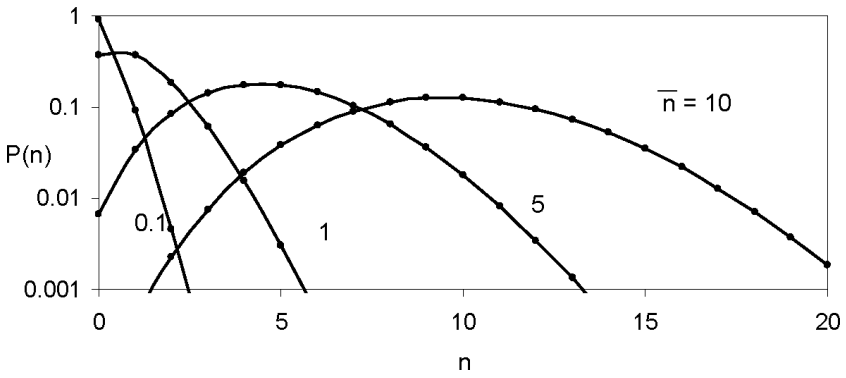

Figure 8.4: The Poisson distribution.

photons is high, while at low intensities only now and then a photon will arrive. This is illustrated in figure 8.3. The exact statistical distribution of the photon flux depends on the nature of the light and we have to make a difference between coherent light, in which the optical power is constant, and thermal light, in which the optical power fluctuates.

Coherent light

For coherent light (e.g. monochromatic light of an ideal laser), the light power is constant but the arrival of photons is caused by uncorrelated events and thus completely random. Under those circumstances the probability $p(n)$ that in a given time interval with duration T, n photons will arrive is given by a Poisson distribution:

$$p\left(n\right) = \frac{\overline{n}^n e^{-\overline{n}}}{n!}, n = 0, 1, 2 \ldots \tag{8.6}$$

This distribution is depicted in figure 8.4 for different values of the average number of photons arriving in the time interval T (this average is proportional to the optical power). The most im-

portant characteristics of a statistical distribution are the average and the variance, defined as:

$$\overline{n} = \sum_{n=0}^{\infty} np\,(n) \tag{8.7}$$

$$\sigma_n^2 = \sum_{n=0}^{\infty} (n - \overline{n})^2 p\,(n) \tag{8.8}$$

The standard deviation (square root of the variance) is a measure for the width of the distribution. For a Poisson distribution we easily find that the variance is equal to the average:

$$\sigma_n^2 = \overline{n} \tag{8.9}$$

This means that when the average number of photons increases (because of increasing power or increasing time interval), the standard deviation will also increase but not as fast as the average itself, and thus, there will be less 'noise' on the photon flux. This plays an important role in communication systems. If the light is only partially coherent, the light intensity will not be constant and there will be an extra fluctuation in the signal. The variance of the number of photons (in a certain time duration) will also be larger than predicted by the Poisson situation. Actually, a photon flux with a Poisson distribution represents a quantized particle stream with the smallest possible variance. This type of noise is also referred to as shot noise.

Thermal light

Thermal light is the other extreme of distributions that photon streams may show. A thermal radiator arises when an object at a temperature T emits photons in a situation of thermal equilibrium. Consider an optical cavity at equilibrium with walls at a temperature T. According to the laws of statistical mechanics, the probability distribution for the electromagnetic energy in one of the modes of the cavity will be a Boltzmann-distribution (figure 8.5).

$$P\,(E_n) \propto e^{-\frac{E_n}{k_b T}} \tag{8.10}$$

with k_B Boltzmann's constant ($k_B = 1.38\ 10^{-23} J/K$). As the energy in an electromagnetic mode is given by $E_n = (n + \frac{1}{2})h\nu$, the probability distribution for the number of photons in a mode is given by

$$\begin{aligned} P\,(E_n) &\propto e^{-\frac{nh\nu}{k_b T}} \\ &= \left(e^{-\frac{h\nu}{k_b T}} \right)^n , n = 0, 1, 2 \ldots \end{aligned}$$

As

$$\sum_{n=0}^{\infty} p(n) = 1 \tag{8.11}$$

we can obtain for the distribution $p(n)$:

$$p(n) = \frac{1}{\overline{n} + 1} \left(\frac{\overline{n}}{\overline{n} + 1} \right)^n \tag{8.12}$$

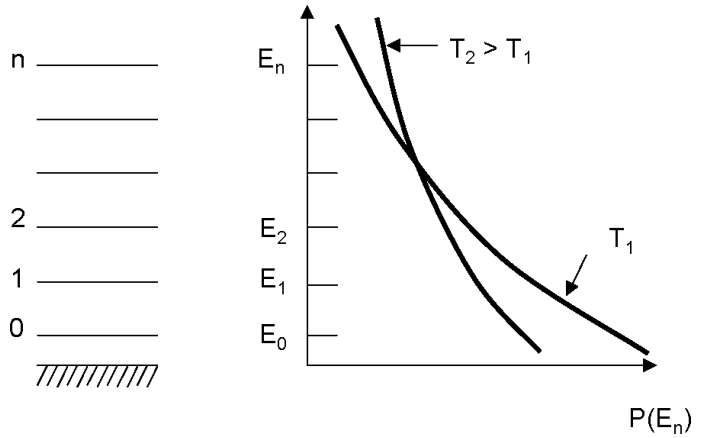

Figure 8.5: The Boltzmann probability distribution $P(E_n)$.

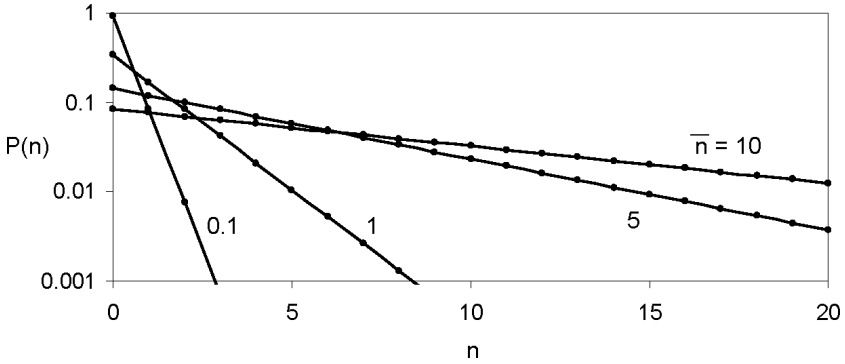

Figure 8.6: The Bose–Einstein distribution.

with

$$\overline{n} = \frac{1}{e^{(\frac{h\nu}{k_b T})} - 1} \tag{8.13}$$

This is a geometric distribution, in quantum-optical context also called the Bose-Einstein distribution. This distribution is shown in figure 8.6 and we immediately see that the variance of this distribution is a lot larger than the variance of the Poisson distribution. The variance is indeed given by:

$$\sigma_n^2 = \overline{n} + \overline{n}^2 \tag{8.14}$$

In practice, this distribution can be measured (approximately) if we filter the light of an incandescent lamp so that only a small spectral band is transmitted, and furthermore, if we only consider one mode of the free space (one plane wave with one direction). The light then fluctuates strongly and is not suitable as a communications carrier.

Partitioning of photon bundles

When a photon bundle is incident on a semi-transparent mirror with reflection R and transmission $T = 1 - R$, each photon will have a probability R of being reflected and a probability T of being

transmitted. We could think that this gives rise to a larger fluctuation of the resulting photon bundles compared to the incident bundle. If there is no correlation between the 'choice' the successive photons make at the mirror, we can however prove that Poisson distributed light remains Poisson distributed and thermal light remains thermal light, of course each time with a new *lower* mean photon flux.

Bibliography

Part III

Light-Material Interaction

Chapter 9

Material Properties

Contents

Properties of materials such as refraction and absorption were introduced in chapter 6. These properties were described by means of the quantity $\mathbf{P}(t)$ - the polarization. In most cases, $\mathbf{P}(t)$ is approximately proportional to the electric field $\mathbf{E}(t)$. In this chapter we will give a more detailed description of the polarization concept and deduce some simple classic models that describe polarization in dielectric structures and metals.

9.1 General definition of polarization

In chapter 6 polarization was defined in the frequency domain as:

$$\mathbf{P}(\omega) = \epsilon_0 \boldsymbol{\chi} \cdot \mathbf{E}(\omega) \tag{9.1}$$

For a general definition however, we will start in the time domain. As polarization is approximately proportional to the electric field, $\mathbf{P}(t)$ can be developed in a series in function of $\mathbf{E}(t)$:

$$\mathbf{P}(t) = \mathbf{P}^{(0)}(t) + \mathbf{P}^{(1)}(t) + \mathbf{P}^{(2)}(t) + \mathbf{P}^{(3)}(t) + ... \tag{9.2}$$

in which the number between brackets denotes the power of proportionality. In other words $\mathbf{P}^{(1)}(t)$ is the first-order polarization (proportional with $\mathbf{E}(t)$), $\mathbf{P}^{(2)}(t)$ the second-order polarization (proportional to the square of the electric field), etc. $\mathbf{P}^{(0)}(t)$ is the statical polarization, which is independent of the electric field. Statical polarization occurs for example in some crystals. Typically, the polarization is approximately proportional to the electric field. Higher order terms ($\mathbf{P}^{(2)}(t)$,$\mathbf{P}^{(3)}(t)$, etc.) are studied in the field of nonlinear optics. These terms are responsible for nonlinear optical effects such as frequency doubling (e.g. used to convert infrared light into the visible), intensity dependent propagation of light, etc. These effects will not be considered in this chapter and we will assume a linear relation between the electric field and the polarization.

9.1.1 Time invariance and causality

Since we are assuming a linear system, we can write the polarization $\mathbf{P}(t, \mathbf{r})$ as

$$\mathbf{P}(t, \mathbf{r}) = \epsilon_0 \int_{-\infty}^{\infty} dt_1 \mathbf{T}(t, \mathbf{r}, t_1) \cdot \mathbf{E}(t_1, \mathbf{r}) \tag{9.3}$$

in which $\mathbf{T}(t, \mathbf{r}, t_1)$ is a 3 by 3 matrix. It describes the response of the polarization when the material is excited with a dirac-impulse in the electric field at a time t_1. Furthermore, the response of a material system that does not undergo any changes, is time invariant, which means that when there is a time translation of the excitation, the dynamic response of the system shifts along with this translation. In other words, $\mathbf{T}(t, \mathbf{r}, t_1)$ depends only on the time differences $\tau = t - t_1$. This is expressed explicitly as,

$$\mathbf{T}(t, \mathbf{r}, t_1) \equiv \mathbf{R}(\mathbf{r}, \tau) \tag{9.4}$$

so that the polarization can be written as

$$\mathbf{P}(t, \mathbf{r}) \quad = \quad \epsilon_0 \int_{-\infty}^{\infty} d\tau \mathbf{R}(\mathbf{r}, \tau) \cdot \mathbf{E}(t - \tau, \mathbf{r}) \tag{9.5}$$

The quantity $\mathbf{R}(\mathbf{r}, \tau)$ is called the polarization response function of the first order. In addition, $\mathbf{R}(\mathbf{r}, \tau)$ is zero when these time differences become negative. Otherwise, if $\tau < 0$ and $\mathbf{R}(\mathbf{r}, \tau) \neq 0$, a field value at a time later on, would have an influence on the value of the polarization at time t. This is of course impossible, as polarization can only be influenced by field values at previous moments and not by future moments. This is called the principle of causality.

9.1.2 Polarization in the frequency domain - linear materials

The above equations in the time domain can easily be transformed into the frequency domain by applying

$$\mathbf{P}(t) \quad = \quad \int_{-\infty}^{\infty} d\omega \mathbf{P}(\omega) \exp(j\omega t) \tag{9.6}$$

$$\mathbf{P}(\omega) \quad = \quad \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \mathbf{P}(t) \exp(-j\omega t) \tag{9.7}$$

Using these transformation formulas it is straightforward to deduce (using the convolution theorem) that the electrical susceptibility $\boldsymbol{\chi}$ and the polarization response function are related as

$$\boldsymbol{\chi}(\omega) = \int_{-\infty}^{\infty} d\tau \mathbf{R}(\tau) \exp(-j\omega\tau) \tag{9.8}$$

As we have seen in chapter 6, $\boldsymbol{\chi}$ describes the refractive index as well as absorption and both effects are dispersive. In the most general case, $\boldsymbol{\chi}$ can only be represented as a symmetrical matrix. In this case the material is anisotropic. This means that the medium has certain preferential directions, which is the case with e.g. crystalline materials. However, in many cases this matrix can be reduced to one single number. So the first-order polarization can be written as

$$\mathbf{P}(\omega) = \epsilon_0 \chi(\omega) \mathbf{E}(\omega) \tag{9.9}$$

This is e.g. the case with amorphous materials. In such structures, the orientation of the different micro-components is random, so that there is no macroscopically preferential direction. Note that this is exactly the same χ as in relationship (6.5) in chapter 6.

9.1.3 Kramers-Kronig relations

As a result of causality, the real and imaginary parts of $\chi(\omega) = \chi_R(\omega) + j\chi_I(\omega)$ are not independent of each other. This means that a dispersive material (thus with a frequency-dependant $\chi_R(\omega)$) will also show absorption (described by $\chi_I(\omega)$) and vice versa. The relations between $\chi_R(\omega)$ and $\chi_I(\omega)$ are called the Kramers-Kronig relations and are given by

$$\chi_R(\omega) = \frac{2}{\pi} P \int_0^\infty \frac{\omega' \chi_I(\omega')}{\omega'^2 - \omega^2} d\omega' \tag{9.10}$$

$$\chi_I(\omega) = \frac{2}{\pi} P \int_0^\infty \frac{\omega \chi_R(\omega')}{\omega^2 - \omega'^2} d\omega' \tag{9.11}$$

with P the principal value of the integral.

With these relationships the real or imaginary part of $\chi(\omega)$ can be deduced if one part is known over the entire frequency range.

9.2 Models for linear, isotropic, dispersive materials

In this section we will formulate by means of a few simple models, relations between macroscopic quantities (namely the refractive index n_R and the extinction coefficient n_I, bundled in the complex susceptibility) and microscopic parameters, which describe the material. We will distinguish dielectric materials and metals.

9.2.1 Damped-oscillator model for dielectric structures

In general a material can be described as a collection of damped oscillators, which interact with the incident light and in that way give cause to refraction and absorption. In materials with no free charge carriers - named dielectrics - the microscopic response of the incident light is in its simplest form an oscillation of bounded charged particles (ions, electrons, ...) under the influence of the electromagnetic waves.

We assume that the material exists of N identical one-dimensional oscillators per unit of volume with mass m, charge e and dampening coefficient γ. A displacement of the particles from their

equilibrium state $u(t)$ will cause a number of forces that try to restore the equilibrium. As the bond between the particles is represented by a damped spring, we have on the one hand Hooke's law given by $F = -k_H u(t) = -m\omega_0^2 u(t)$, where ω_0 represents the resonance frequency associated with the spring constant k_H, and on the other hand a dampening force $-m\gamma\frac{du}{dt}$. Therefore Newton's law for the damped oscillation of these particle becomes

$$m\frac{d^2u}{dt^2}(t) = -m\gamma\frac{du}{dt}(t) - m\omega_0^2 u(t) \tag{9.12}$$

The interaction between these oscillators and the incident light can be described by adding another power term to this damped harmonic oscillator equation. This power term oscillates with the frequency of the incident light and is proportional to the charge of the oscillator as well as the magnitude of the electric field $E(t) = \text{Re}\{E\exp(j\omega t)\}$ of the incident light, thus

$$m\frac{d^2u}{dt^2}(t) = -m\gamma\frac{du}{dt}(t) - m\omega_0^2 u(t) + \text{Re}\{eE\exp(j\omega t)\} \tag{9.13}$$

By going to the frequency domain with $u(t) = \text{Re}\{u\exp(j\omega t)\}$ we get

$$(-m\omega^2 + mj\gamma\omega + m\omega_0^2)u = eE \tag{9.14}$$

or

$$u = \frac{eE}{m(\omega_0^2 - \omega^2 + j\gamma\omega)} \tag{9.15}$$

We conclude that the displacement u depends on the material parameters as well as on the incident light. The total effect of the oscillation of N identical oscillators is a polarization of the material, which is given by the number of electric dipoles per unit of volume. The induced dipole moment of one single oscillator is equal to eu. The polarization $P(t) = \text{Re}\{P\exp(j\omega t)\}$ is then

$$P = Neu = \frac{Ne^2E}{m(\omega_0^2 - \omega^2 + j\gamma\omega)} \tag{9.16}$$

Furthermore because $P = \epsilon_0\chi E$, the (linear) susceptibility becomes:

$$\chi = \frac{Ne^2}{m\epsilon_0}\frac{1}{(\omega_0^2 - \omega^2 + j\gamma\omega)} \tag{9.17}$$

The first factor on the right side of this relationship is usually indicated as the square of a frequency, the so-called plasma frequency ω_p, namely

$$\omega_p = \sqrt{\frac{Ne^2}{m\epsilon_0}} \tag{9.18}$$

Relationship (9.17) represents a resonance.

As seen in chapter 6, the relationship between the refractive index n_R, the extinction coefficient n_I and the linear susceptibility χ is given by,

$$(n_R + jn_I)^2 = 1 + \chi \tag{9.19}$$

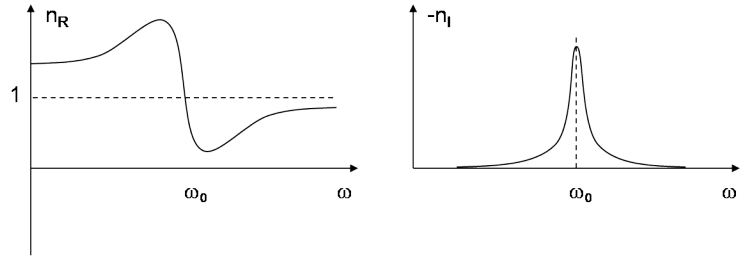

Figure 9.1: Example of the refractive index and extinction coefficient for a resonant dielectric.

or

$$n_R^2 - n_I^2 = 1 + \frac{\omega_p^2(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} \tag{9.20}$$

$$-2n_R n_I = \frac{\gamma\omega\omega_p^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} \tag{9.21}$$

It is obvious that the microscopic dampening coefficient results in a macroscopic extinction coefficient.

Exercise

Materials with a low density have a refractive index close to 1. Furthermore, their extinction coefficient is small usually. Approximate the equations (9.20) *and* (9.21) *in this situation and find a closed relationship for n_R as well as n_I. Check if the limit of n_R for $\omega = 0$ and $\omega = \infty$ is approximately 1 and try to explain this.*

A typical sketch of $n_R(\omega)$ and $n_I(\omega)$ is given in figure 9.1.

Above we implicitly assumed that the so-called local field, which is felt by the microscopical particles is equal to the electric field of the incident light. This applies only if the density of the oscillators in the material is small, such as in a gas. For dense materials we have to take the field caused by neighbouring oscillators into account, which causes the so-called Lorentz contribution.

A typical dielectric material has multiple resonances that correspond with different lattice and electron vibrations. The total susceptibility is then equal to the sum of the contributions of the different resonances. An example is given in figure 9.2. The interaction between the incident light and the ions, respectively electrons, can be seen.

Exercise

Can you explain the order of the interactions? In other words, why do the ions have a lower interaction frequency than the electrons?

The extinction coefficient only has contributions near the different resonance frequencies, while the refractive index is different from zero for all frequencies. In the limit for $\omega \to \infty$ one obtains $n = 1$, no particle is able to move along with the frequency of the incident light. In other words,

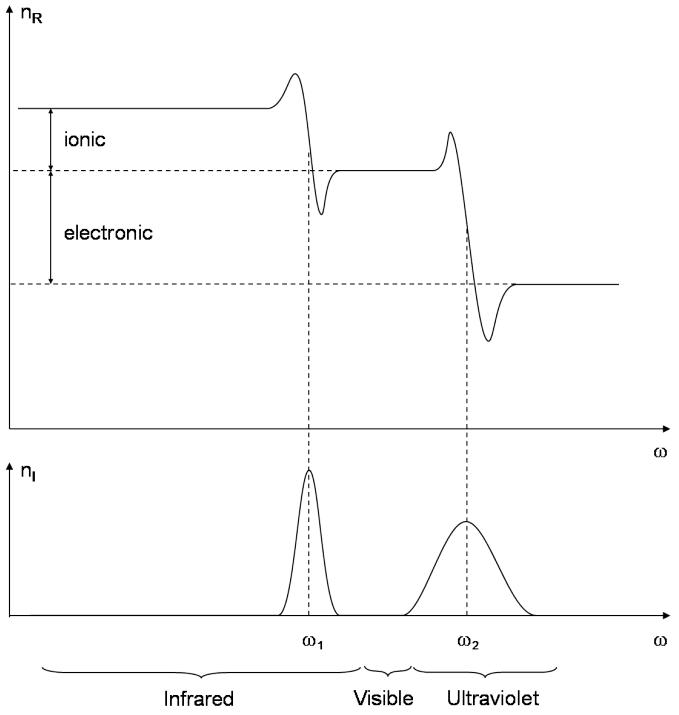

Figure 9.2: Example of the refractive index and extinction coefficient for a dielectric.

the light sees no particles. The successive resonances increase the refractive index for decreasing frequencies. Between the different resonances, the refractive index is approximately constant. We can also see that dispersion plays a role especially near the resonances, both for the refractive index and the extinction coefficient.

In reality, the two spectra show not only resonances, but also relaxations. Then there is no sudden augmentation of the refractive index, but a gradual transition between the two levels.

9.2.2 Drude-model for metals

Metals differ substantially from dielectric materials because they contain electrons that are not bound to the ion-cores. The incident light now interacts with these particles and moves them. But, contrary to the damped resonator model, they are (almost) not drawn back to their original position. These electrons are almost free particles and the incident light will now generate microscopic currents instead of oscillations.

The movement of the charges in a metal can be described by means of an equation that only incorporates the dampening force on these particles. We do not have to take the restoring Hooke force from the damped oscillator model into account, as the particles are not bound now. Thus, this gives

$$m\frac{d^2u}{dt^2}(t) = -m\gamma\frac{du}{dt}(t) \tag{9.22}$$

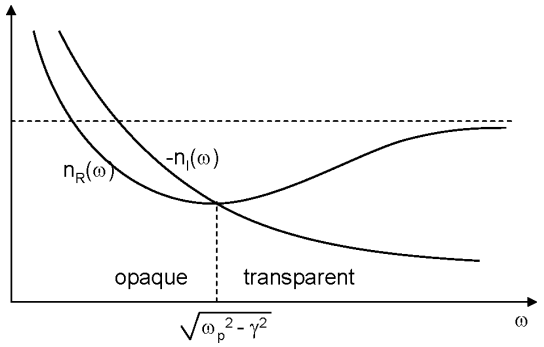

Figure 9.3: Example of the refractive index and extinction coefficient for a metal.

On the right side we now add the driving force from the incident light as a result of the interactions with the free electrons. Thus, we get

$$m\frac{d^2u}{dt^2}(t) = -m\gamma\frac{du}{dt}(t) + \text{Re}\left\{eE\exp(j\omega t)\right\} \tag{9.23}$$

Again we change to the frequency domain, with $u(t) = \text{Re}\left\{u\exp(j\omega t)\right\}$. We get

$$u = \frac{eE}{m(-\omega^2 + j\gamma\omega)} \tag{9.24}$$

The induced polarization P associated with the interaction between the incident light and the free electrons in metals becomes

$$P = \frac{Ne^2E}{m(-\omega^2 + j\gamma\omega)} \tag{9.25}$$

and the (linear) susceptibility is

$$\chi = \frac{Ne^2}{m\epsilon_0}\frac{1}{(-\omega^2 + j\gamma\omega)} = \frac{\omega_p^2}{(-\omega^2 + j\gamma\omega)} \tag{9.26}$$

Using the relationship between the refractive index n_R, the extinction coefficient n_I and the linear susceptibility, we get

$$n_R^2 - n_I^2 = 1 - \frac{\omega_p^2}{\omega^2 + \gamma^2} \tag{9.27}$$

$$-2n_Rn_I = \frac{\gamma}{\omega}\frac{\omega_p^2}{\omega^2 + \gamma^2} \tag{9.28}$$

A sketch of $n_R(\omega)$ and $n_I(\omega)$ is shown in figure 9.3. The realistic case of the metal Au (gold) is presented in figure 9.4[1].

These equations differ strongly compared to the case of dielectric structures, because the lack of a resonance term. The singularity that we obtained when $\gamma = 0$ has now vanished. On the other hand, the limit $\omega \to 0$ is now singular. Physically this means that metals are opaque for low

[1] reminder: $1\mu m = 1.24eV$

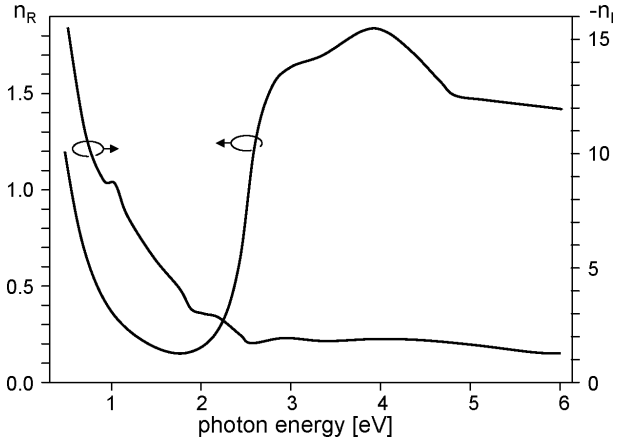

Figure 9.4: Example of the refractive index and extinction coefficient of gold.

frequencies. When we take the limit $\omega \to \infty$ we get $n_R = 1$ and $n_I = 0$ again. In other words, metals are also transparent at high frequencies, like e.g. X-rays.

We will now try to deduce a relationship for the penetration depth of low-frequency electromagnetic waves in metals. For very low frequencies we can approximate equation (9.26) as

$$\chi \approx \frac{\omega_p^2}{j\gamma\omega} \tag{9.29}$$

In addition, since $\sqrt{\frac{1}{j}} = \frac{1-j}{\sqrt{2}}$, we get

$$n_R \approx -n_I \approx \frac{\omega_p}{\sqrt{2\gamma\omega}} \tag{9.30}$$

The penetration depth is now defined as the distance by which the intensity of the incident light drops to a $1/e$ fraction of its original value. Assume that the propagation occurs along the z-axis, and that the metal extends along the positive z-axis, then the intensity is given by

$$I = I_0 \exp(-\alpha z) \tag{9.31}$$

with I_0 the intensity at $z = 0$. The absorption coefficient α is related to n_I as follows

$$\alpha = -2\frac{\omega}{c} n_I \tag{9.32}$$

In this way we get the penetration depth l

$$l = \frac{1}{\alpha} = -\frac{c}{2\omega n_I} = \frac{c}{\omega_p}\sqrt{\frac{\gamma}{2\omega}} \tag{9.33}$$

For a good conductor like copper, the penetration depth is

$$l_{Cu} = \sqrt{\frac{1}{48\pi\omega}} m = \frac{0.081}{\sqrt{\omega}} m \tag{9.34}$$

Exercise

Calculate the penetration depth of copper for an impinging wavelength of 1.55μm. What is the needed thickness of a copper layer in order to let 99% of the incident light through at the same wavelength?

Bibliography

[Möl88] K.D. Möller. *Optics*. University Science Books, ISBN 0-935702-145-8, 1988.

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Chapter 10

Photons and Atoms

Contents

10.1 Atoms and molecules

Matter consists of atoms. These atoms can be rather isolated from each other, like in a thin gas, or they can interact with each other and form molecules or crystal structures in the liquid or solid phase. The movement and mutual interaction of all particles is determined by the laws of quantum mechanics. The behavior $\Psi(\mathbf{r}, t)$ of one single particle with mass m in a potential energy $V(\mathbf{r}, t)$ is determined by the time-dependent Schrödinger equation.

$$\frac{-\hbar^2}{2m}\nabla^2\Psi(\mathbf{r}, t) + V(\mathbf{r}, t)\Psi(\mathbf{r}, t) = i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r}, t) \tag{10.1}$$

A system consisting of multiple particles satisfies a more extensive equation (different Ψ's). In addition, the potential energy contains all sorts of terms allowing interactions with other particles and exterior fields. The probability of finding the particle at a position $\mathbf{r}$ (volume cell $d\mathbf{r}$) during the interval $[t, t + dt]$ is

$$dP(\mathbf{r}, t) = |\Psi(\mathbf{r}, t)|^2 \, d\mathbf{r}dt \tag{10.2}$$

To determine the allowed energy states for a particle (assuming that the Hamiltonian and thus V(**r**) is independent of time), we can use separation of variables on equation (10.1) and we get

$$\Psi(\mathbf{r}, t) = \Psi(\mathbf{r}) \exp\left(i\frac{E}{\hbar}t\right) \tag{10.3}$$

whereby $\Psi(\mathbf{r})$ satisfies the time-independent Schrödinger equation

$$\frac{-\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{10.4}$$

These energy levels can be either discrete or continuous. For a system with multiple particles, a similar equation applies. For such systems, the energy levels can even form bands of discrete, but very closely spaced values, like e.g. in semiconductors. The interaction with an external field, like an incident light beam, can cause the system to transfer to another energy level by absorbing photons from this beam.

10.1.1 Energy levels

Isolated atoms

The energy levels of an atom with Z electrons can be determined approximately by solving the time-independent Schrödinger equation, which describes the movement of Z particles in the field caused by the nucleus (typically a Coulomb potential) as well as the Coulomb interaction between the electrons themselves. The simplest problem is that of the isolated hydrogen atom. After solving the Schrödinger equation, we ultimately get for the discrete energy levels

$$E_q = -\frac{m_r e^4}{2\hbar^2 q^2} \qquad q = 1, 2, 3, ... \tag{10.5}$$

with e the charge of the electron and m_r the reduced mass of the system, defined as

$$m_r = \frac{mM}{M+m} \approx m \tag{10.6}$$

with m the electron mass and M the mass of the hydrogen nucleus.

Molecular systems

The energy levels of systems with multiple atoms, like molecules, are a lot more complex. On the one hand they are the result of the valence electrons which can move freely in the field of the atomic nuclei and the other (bound) electrons. These electrons cause the bond between the different atoms. On the other hand the nuclei can (together with their strongly bound electrons) move w.r.t. each other, which causes rotational and vibrational energy levels. We will explain this in more detail below.

- The electronic states in molecules are a lot more difficult to determine than these in the case of isolated atoms. As mentioned before, they are the consequence of the movement of quasi-free electrons (valence electrons) in the field caused by the different atomic nuclei and they are the result of the interaction between the different valence levels of the valence electrons in the original atoms. The levels are discrete, as was the case for atoms. The energy difference between successive energy states is, like in isolated atoms, typically 1 to $10eV$.

- In additon, molecules can also vibrate because the mutual distance between the atomic nuclei can vary dynamically. This causes a splitting up of each electronic state in different vibrational levels. A diatomic molecule, such as CO, can for example be modeled as a system consisting of two masses that are connected with each other by a spring. The whole forms a harmonic oscillator with potential energy $V(x) = \frac{1}{2}kx^2$ with x the coordinate along the connecting axis. As seen in chapter 8, the energy levels of a harmonic oscillator are given by

$$E_q = \left(q + \frac{1}{2}\right)\hbar\omega \qquad q = 0, 1, 2, ... \tag{10.7}$$

  where $\omega = \frac{k}{m_r}$. Typical values of $\hbar\omega$ are $0.05 - 0.5eV$. This corresponds with energy levels in the infrared. More complicated molecules can display different kinds of vibration, according to the atoms that are moving. Each type of vibration is represented by its own quantum number q.

- Finally, each vibrational energy level demonstrates different rotational levels. These correspond to rotational movements of the molecule around different axes. For a diatomic molecule, only rotation around the gravitational point (which is located on the connecting axis between the two atoms) can occur. The energy levels are given by

$$E_q = q\,(q+1)\,\frac{\hbar^2}{2\Im} \qquad q = 0, 1, 2, ... \tag{10.8}$$

  with $\Im$ the moment of inertia. The differences between different rotational energy levels are situated between 0.001 and $0.01eV$. These values are in the far-infrared.

The transitions between all these different energy levels are submitted to a number of rules, the so-called selection rules, which ensure that not all transitions are allowed.

Solid-state systems

In solids there are typically a bunch of atoms and molecules located very close to each other. Like in molecules, the energy states can be determined by on the one hand considering the movement of the electrons and on the other hand by taking the possibility of vibrational and rotational states into account. We will only explain the role of the valence electrons a bit more thoroughly. In contrast to molecules, this causes a quasi-continuous spectrum that consists of very closely spaced energy levels as if they form bands. These bands are separated from each other by forbidden zones and they fundamentally determine the properties of the solid.

In a system consisting of N closely spaced atoms[1], such an energy band consists of N different energy levels. In the three-dimensional case, it is possible that these bands partially overlap. Because of the Pauli-principle, each energy level can contain 2 electrons, namely one spin-up and one spin-down. Thus, each energy band can contain $2N$ electrons. All this gives rise to three possible situations:

[1] In the case of N atoms located far away from each other, we have in fact one single level that is N-times degenerated. In other words, by bringing the atoms near to each other, an interaction arises that cancels out the overlap of the energy levels, but on the contrary forms a band of N levels.

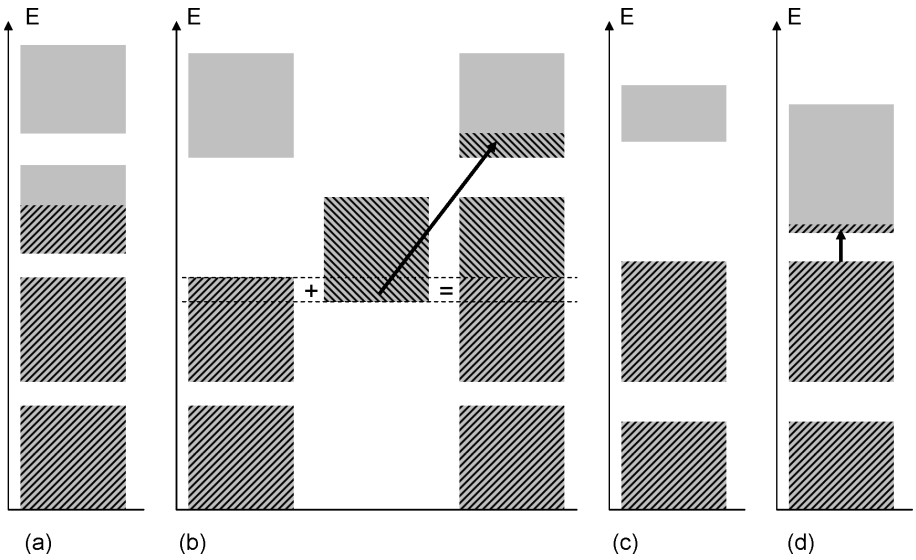

Figure 10.1: Energy levels of solid-state systems.

- Assume each atom has an odd ($2k + 1$ with k natural) number of electrons, so that the entire system consists of $(2k + 1)N$ electrons, then we have, besides completely filled (with $2N$ electrons) and completely empty bands, also a half filled energy band (with N electrons) (figure 10.1(a)). This band is called the valence band. The previous is also possible with an even number of electrons when the last filled band partially overlaps with the following band (figure 10.1(b)). Because of the many empty states, electrons can easily be excited under influence of e.g. an external electric field. Concretely, this means that they can move easily through the solid. In other words, these are metals.
- When each atom has an even number of valence electrons and no band overlap occurs, then we only have fully filled bands, that are separated by a forbidden zone from the empty bands (figure 10.1(c)). It takes a lot of energy to excite such electrons. These materials are called isolators.
- It is however also possible that this forbidden zone is not very large (figure 10.1(d)). So that these electrons can leap over this zone by thermal excitation, and end up in the first non-filled band, the so-called conduction band. These materials have some resistance, but it is not insurmountable. These are the so-called semiconductors.

10.1.2 Occupation of energy levels in thermal equilibrium

Each atom or molecule in a whole of atoms and molecules continually undergoes transitions between the different energy levels because of thermal excitation and relaxation; for kinetic energy (caused by temperature) is continually being exchanged when the different particles collide with each other. These random transitions are described with statistical physics and result in a number of thermal distributions.

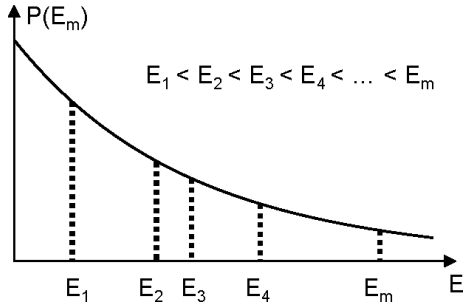

Figure 10.2: The Boltzmann distribution.

Boltzmann distribution

Consider a collection of identical atoms or molecules in a medium such as a dilute gas. Each atom is then located in one of the allowed energy states $E_0, E_1, E_2, ...$ If the system is in a state of thermal equilibrium[2] at a temperature T, then the probability that an arbitrary atom is in an energy state E_m, is given by the Boltzmann distribution

$$P(E_m) = A \exp\left(-\frac{E_m}{k_B T}\right) \tag{10.9}$$

with A chosen so that $\sum_m P(E_m) = 1$ and $k_B = 1.38 \times 10^{-23} JK^{-1}$, Boltzmann's constant. This is an exponentially decreasing function of the energy (see figure 10.2).

For a large number of atoms N, the number of atoms N_m in the energy state E_m is thus equal to

$$N_m = AN \exp\left(-\frac{E_m}{k_B T}\right) \tag{10.10}$$

and the proportion between the number of atoms in state E_i and the number in state E_j is thus

$$\frac{N_i}{N_j} = \exp\left(-\frac{E_i - E_j}{k_B T}\right) \tag{10.11}$$

The Boltzmann distribution clearly depends on the temperature. At $T = 0K$ all the atoms are in the ground state (logical). With rising temperature, the number of atoms occupying a higher energy state increases. At equilibrium, the occupation of a higher energy level is, averagely speaking, always lower than the occupation of a lower energy level. Thus, if $E_i < E_j$ then $N_i > N_j$. This is no longer necessarily true if the whole of the atoms is no longer in equilibrium. The situation where $E_i < E_j$ and $N_i < N_j$ is called population inversion and lies at the base of the operation of a laser. This will be explained in chapter 13.

Until now we assumed that an atom only has one state with energy E_m. However, this is not always the case: degenerate states are possible[3]. In general, we get

$$\frac{N_i}{N_j} = \frac{g_i}{g_j} \exp\left(-\frac{E_i - E_j}{k_B T}\right) \tag{10.12}$$

where g_m represents the number of states with energy E_m.

[2] E.g. by bringing the atoms in contact with a large reservoir at temperature T.
[3] Recall e.g. the different spin states.

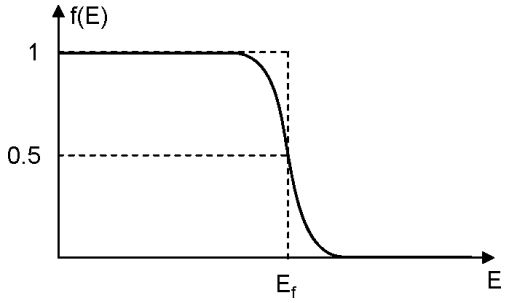

Figure 10.3: The Fermi-Dirac distribution

Fermi-Dirac distribution

Electrons in semiconductors satisfy another occupation distribution. As the atoms in such a situation are closely spaced to each other, the material has to be treated as one single system[4]. This means that each possible state is either occupied or unoccupied, whereas in a system of N isolated particles all particles can occupy the *same* state (e.g. at $T = 0$)[5]. The probability that a state with energy E is occupied, is then given by

$$f(E) = \frac{1}{\exp\left(\frac{E-E_f}{k_B T}\right) + 1} \tag{10.13}$$

This is called the Fermi-Dirac distribution, with E_f the Fermi-energy. We get that $f(E) = \frac{1}{2}$ if $E = E_f$. The Fermi-Dirac distribution is depicted in figure 10.3.

For $E \gg E_f$, we obtain

$$f(E) \propto \exp\left(-\frac{E - E_f}{k_B T}\right) \tag{10.14}$$

and thus we again get the Boltzmann distribution.

10.2 Interactions between photons and atoms

As mentioned before, an atom can be excited by absorption of a photon, and inversely it can be relaxed by emission of a photon. Now we go into a bit more detail.

Consider the energy levels E_1 and E_2 of an atom (with $E_2 > E_1$) in a cavity with volume V. We are especially interested in photons with an energy $h\nu_0 = E_2 - E_1$, as this corresponds with the energy difference between the two atomic levels. Such photon-atom interactions can formally be studied with quantum electrodynamics. Here, we only mention the results. The interactions between atoms and photons are separated into three types, namely spontaneous emission, stimulated emission and (stimulated) absorption.

[4]Recall the formation of energy bands by bringing the atoms closer to each other.

[5]The number of *different* occupied energy states occupied in this system is then also a lot larger than in the case of a system with isolated atoms, where often the *same* energy state is occupied

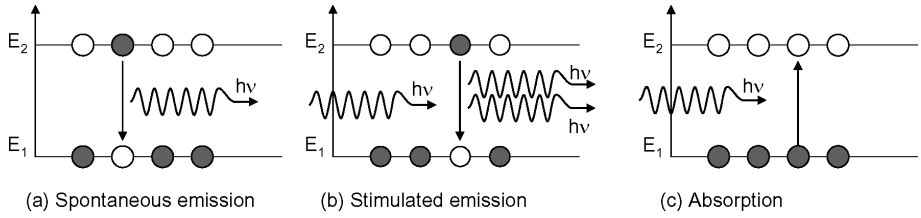

Figure 10.4: Spontaneous emission.

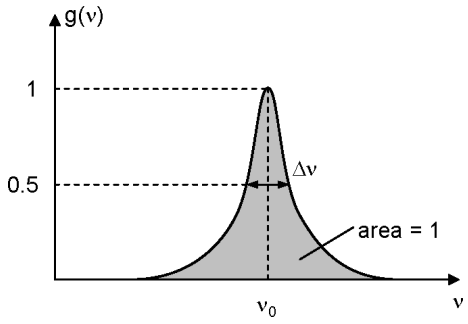

Figure 10.5: Example of a lineshape function.

10.2.1 Spontaneous emission

If an atom is initially in an energy state E_2, it can spontaneously make a transition to a lower energy state E_1 by emission of a photon in a radiation mode with a *specific* energy $h\nu \approx E_2 - E_1$ (figure 10.4(a)). The process in which this happens is called spontaneous emission, as the transition occurs independent of the number of photons with this energy that are already present in the cavity.

In a cavity with volume V, the probability density (per second) p_{sp} for spontaneous emission to occur, depends on the frequency ν:

$$p_{sp} = \frac{c}{V}\sigma(\nu) \tag{10.15}$$

with $\sigma(\nu)$ a function centered around the atomic resonance frequency $\nu_0 = \frac{E_2 - E_1}{h}$. This function is called the transition cross section and is expressed in m^2. This function can be determined by using the time-dependent Schrödinger equation. In practice, the characterization is usually done experimentally. The normalized version of this function is also called the lineshape function $g(\nu)$

$$g(\nu) = \frac{\sigma(\nu)}{\int \sigma(\nu) d\nu} \tag{10.16}$$

A typical example of a lineshape function is given in figure 10.5. The width of this function is called the linewidth $\Delta\nu$, defined as the full width of $g(\nu)$ at half its maximum (FWHM).

The term 'probability density' means that the probability of spontaneous emission between the times t and $t + dt$ is equal to $p_{sp}dt$. Thus, having N_2 atoms in the energy state E_2, the number of atoms that spontaneously emits a photon during the time interval dt becomes:

$$dN_2 = -p_{sp} N_2 dt \tag{10.17}$$

so that the population N_2 evolves as

$$N_2(t) = N_2(0)\exp(-p_{sp}t) \tag{10.18}$$

Until now, we have studied only spontaneous emission of photons into a *specific* cavity mode with frequency ν. The density of these modes (per unit of volume and frequency) in a three-dimensional cavity is given by $M(\nu) = \frac{8\pi\nu^2}{c^3}$. An atom can however emit a photon in *any* radiation mode of frequency $\nu \approx \frac{E_2-E_1}{h}$. We get the total spontaneous emission probability density P_{sp} by integrating over all frequencies, namely,

$$\begin{aligned} P_{sp} &= \int p_{sp}(\nu) V M(\nu) d\nu & (10.19)\\ &= c\int \sigma(\nu) M(\nu) d\nu & (10.20)\\ &\approx cM(\nu_0)\int \sigma(\nu) d\nu & (10.21)\\ &= \frac{8\pi}{\lambda_0^2}\int \sigma(\nu) d\nu & (10.22) \end{aligned}$$

This relationship is independent of V. The fact that $\sigma(\nu)$ is typically varying faster than $M(\nu)$ has been taken into account. The spontaneous lifetime τ_{sp} is defined as

$$\tau_{sp} = \frac{1}{P_{sp}} \tag{10.23}$$

$A = P_{sp} = \frac{1}{\tau_{sp}}$ is also called the A coefficient of Einstein. He deduced the expression for A by analyzing the photon-atom interactions in thermal equilibrium.

10.2.2 Stimulated emission

If an atom is initially in an energy state E_2 and the radiation mode with frequency $\nu \approx \frac{E_2-E_1}{h}$ contains a photon, then the atom can also make a transition to a lower energy state E_1 stimulated by this mode by emitting a photon that also belongs to this mode (figure 10.4(b)). This process is called stimulated emission. The newly emitted photon is in every aspect the same as the already existing photon of that mode. This lies at the base of laser operation.

The probability density p_{st} of this process in a cavity with volume V is, in the presence of one photon in the mode, the same as in the case of spontaneous emission, namely

$$p_{st} = \frac{c}{V}\sigma(\nu) \tag{10.24}$$

If the mode contains n photons, the total probability density becomes

$$P_{st} = n\frac{c}{V}\sigma(\nu) \tag{10.25}$$

The total emission probability of a photon in a cavity mode with frequency ν is

$$p_{sp} + P_{st} = (n+1)\frac{c}{V}\sigma(\nu) \tag{10.26}$$

In quantum electrodynamics, spontaneous emission is seen as the process stimulated by the zero-point energy of a mode (analogous to the zero-point energy of the harmonic oscillator.

Now we consider a cavity with a broadband spectral energy density (energy per unit of volume and frequency) given by $\rho(\nu)$. The number of photons with frequency between ν and $\nu + d\nu$ in the cavity is then $\frac{\rho(\nu)}{h\nu} V d\nu$, so that the total stimulated emission probability density P_{st} becomes

$$\begin{aligned} P_{st} &= \int p_{st}(\nu) \frac{\rho(\nu)}{h\nu} V d\nu && (10.27) \\ &= c \int \frac{\rho(\nu)}{h\nu} \sigma(\nu) d\nu && (10.28) \\ &\approx \frac{\rho(\nu_0)\lambda_0}{h} \int \sigma(\nu) d\nu && (10.29) \\ &= \frac{\lambda_0^3}{8\pi h \tau_{sp}} \rho(\nu_0) && (10.30) \end{aligned}$$

Here we have again taken into account that $\sigma(\nu)$ is considered much more narrow than $\rho(\nu)$. If we now define the average number of photons per mode as

$$\bar{n} = \frac{\lambda_0^3}{8\pi h} \rho(\nu_0) \tag{10.31}$$

we get

$$P_{st} = \frac{\bar{n}}{\tau_{sp}} = \bar{n} P_{sp} \tag{10.32}$$

The quantity $\frac{\lambda_0^3}{8\pi h \tau_{sp}}$ is also called Einstein's B coefficient. As mentioned before, Einstein used a different approach to deduce this.

10.2.3 Absorption

If an atom is initially in an energy state E_1 and a radiation mode with frequency $\nu \approx \frac{E_2 - E_1}{h}$ contains a photon, then the atom can make a transition to a higher energy state E_2 by absorbing this photon (figure 10.4(c)). Thus, absorption is a process stimulated by the presence of a photon with an appropriate frequency.

The probability density p_{ab} for absorption of a photon from a given mode with frequency ν in a cavity with volume V, is the same as the one for spontaneous and stimulated emission, namely

$$p_{ab} = \frac{c}{V} \sigma(\nu) \tag{10.33}$$

Now if there are n photons in this mode, then the total absorption probability density P_{ab} is equal to

$$P_{ab} = n \frac{c}{V} \sigma(\nu) \tag{10.34}$$

as only one atom can be absorbed and the events are mutually exclusive. Note $P_{ab} = P_{st}$.

Analogously to stimulated emission, we can prove that in the presence of a broadband spectral energy density $\rho(\nu)$, the total absorption density P_{ab} is also given by

$$P_{ab} = \frac{\bar{n}}{\tau_{sp}} = \bar{n} P_{sp} \tag{10.35}$$

so that again $P_{ab} = P_{st}$.

10.3 Thermal light

Light emitted by atoms, molecules and solids under the condition of thermal equilibrium and *in the absence of other energy sources*, is called thermal light. We will study the properties of thermal light based on the interactions between photons and atoms.

10.3.1 Thermal equilibrium between atoms and photons

Consider a cavity of unit volume with the walls consisting of a large number of atoms that have two different energy levels E_1 and E_2 (with again $E_2 > E_1$). Denote the number of atoms per unit volume that are at the time t in state 1 by $N_1(t)$ and in state 2 by $N_2(t)$. Spontaneous emission will cause electromagnetic radiation in the cavity, assuming that the population of the second energy level is initially not equal to zero. At its turn, the radiation causes stimulated emission as well as absorption. These three processes result in thermal equilibrium between on the one hand the atoms and on the other hand the radiation of photons. We assume that *each* radiation mode with frequency lying in the linewidth of $g(\nu)$ is occupied by an average number of photons $\bar{n}$. This means that

$$P_{st} = P_{ab} = \frac{\bar{n}}{\tau_{sp}} \tag{10.36}$$

Let us consider spontaneous emission. Analogous to section 10.2.1, the number of atoms spontaneously emitting a photon during the time interval dt is equal to

$$dN_2 = -\frac{N_2}{\tau_{sp}} dt \tag{10.37}$$

so that the population N_2 evolves as an exponentially decreasing function

$$N_2(t) = N_2(0) \exp(-\frac{t}{\tau_{sp}}) \tag{10.38}$$

However, spontaneous emission is not the only interaction that occurs. In the presence of radiation stimulated emission and absorption will happen, which influences the occupations N_1 and N_2. Let us first consider absorption. At a time t N_1 atoms per unit volume are able to absorb a photon. During the time interval dt this will cause a rise of the number of atoms at the energy level E_2 with $dN_2(t)$:

$$dN_2 = N_1 P_{ab} dt = \frac{N_1 \bar{n}}{\tau_{sp}} dt \tag{10.39}$$

Analogously, stimulated emission causes a decrease of the number of atoms in state 2, given by

$$dN_2 = -N_2 P_{st} dt = -\frac{N_2 \bar{n}}{\tau_{sp}} dt \tag{10.40}$$

All these processes (spontaneous emission, stimulated emission and absorption) together give rise to the equation for the rate of change of the population density $N_2(t)$ of the energy level E_2

$$\frac{dN_2}{dt} = \frac{N_1 \bar{n}}{\tau_{sp}} - \frac{N_2(\bar{n}+1)}{\tau_{sp}} \tag{10.41}$$

This relationship does not take the interaction of atoms transferring from/to other energy levels than E_1 and E_2 into account, so the following also applies

$$\frac{dN_1}{dt} = -\frac{dN_2}{dt} \tag{10.42}$$

Neither does this relationship take non-radiative processes and external excitations into account. The solution at equilibrium $\frac{dN_2}{dt} = 0$ gives

$$\frac{N_2}{N_1} = \frac{\bar{n}}{\bar{n}+1} \tag{10.43}$$

which clearly proves that $N_2 < N_1$ as expected. Furthermore, when the atoms are at thermal equilibrium, the following applies according to Boltzmann (assuming no degenerate states)

$$\frac{N_2}{N_1} = \exp\left(-\frac{E_2 - E_1}{k_B T}\right) = \exp\left(-\frac{h\nu}{k_B T}\right) \tag{10.44}$$

so that the average number of photons in the mode with frequency ν is

$$\bar{n} = \frac{1}{\exp\left(\frac{h\nu}{k_B T}\right) - 1} \tag{10.45}$$

The previous derivation applies for a system with two energy levels. The validity of formula (10.45) goes a lot further. Consider a cavity occupied with atoms having a continuum of energy levels. Again these will interact with a radiation field through spontaneous emission, stimulated emission and absorption so that finally thermal equilibrium arises. The average number of photons with a frequency ν will be given by formula (10.45).

Remark: This is the average of the Bose-Einstein probability distribution[6].

10.3.2 Blackbody radiation spectrum

In addition, relationship (10.45) tells us that the average energy of a radiation mode with frequency ν at thermal equilibrium equals to

$$\bar{E} = \bar{n} h\nu = \frac{h\nu}{\exp\left(\frac{h\nu}{k_B T}\right) - 1} \tag{10.46}$$

[6]This is the equivalent of the Fermi-Dirac distribution (that is only valid for fermions) for bosons. Bosons are particles like photons, for which the anti-particle is equal to the particle itself. Fermions have different particles and anti-particles

Figure 10.6: The blackbody radiation spectrum.

If $h\nu << k_BT$ this becomes, with $\exp\left(\frac{h\nu}{k_BT}\right) \approx 1 + \frac{h\nu}{k_BT}$,

$$\bar{E} \approx k_BT \tag{10.47}$$

This is nothing else than the classic value of the average energy of a radiation mode.

If we multiply the expression of $\bar{E}$ with the mode density (per unit of volume and frequency) of a three-dimensional cavity $M(\nu) = \frac{8\pi\nu^2}{c^3}$, we get the spectral energy density (energy per unit of volume and frequency), namely

$$\rho(\nu) = \frac{8\pi h\nu^3}{c^3} \frac{1}{\exp\left(\frac{h\nu}{k_BT}\right) - 1} \tag{10.48}$$

This relationship is called the blackbody radiation spectrum and is depicted in figure 10.6. This is the same expression proposed by Planck in order to solve the problem of the ultraviolet catastrophe. Classically one obtains

$$\bar{E} \approx \frac{8\pi\nu^2}{c^3} k_BT \tag{10.49}$$

which is indeed nothing else than the relationship of Rayleigh-Jeans.

10.4 Luminescent light

An *external energy source*[7] brought into contact with an atomic or molecular system may cause transitions to higher energy levels. As a consequence, during the decay of these high energy levels to lower energy levels, the system can emit optical radiation. This non-thermal radiation is called luminescent radiation and the process is called luminescence. Luminescent radiators are usually classified according to the source of excitation energy:

- Cathodoluminescence is caused by accelerated electrons that collide with the atomic system such as in e.g. a cathode ray tube in which the electrons transfer their energy to the phosphor atoms. The term betaluminescence is used when electrons are generated by nuclear β-decay instead of an electron gun.
- Photoluminescence is caused by energetic optical photons, for example the radiation caused by some crystals after illumination with ultraviolet light. The term radioluminescence is applied when the energy source is an X- or γ-radiator, or other ionizing radiation. Photoluminescence is discussed in more detail below.

[7] In contrast with the situation of thermal light.

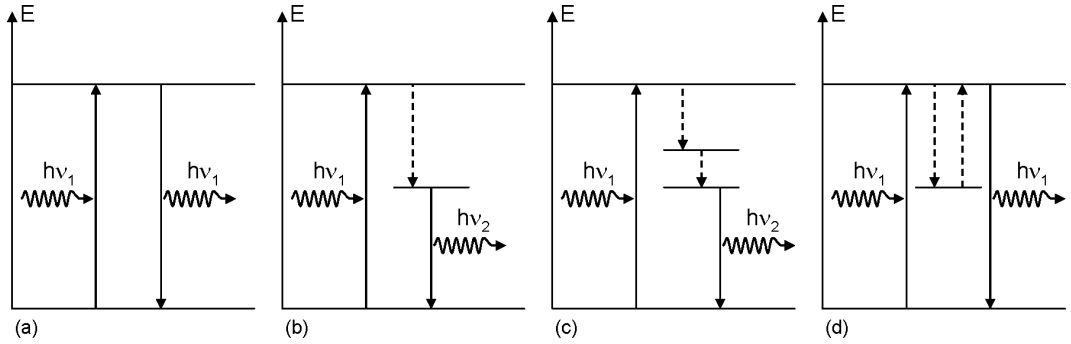

Figure 10.7: Examples of photoluminescent processes.

- Chemiluminescence provides energy by means of chemical reactions. An example is the radiation of phosphor when it oxidizes in air. Bioluminescence - light emitted by some living organisms such as fireflies - is another example.
- Electroluminescence is caused by the energy provided by establishing an electric field. An important example is injection electroluminescence. This occurs when an electric current is injected into a forward-biased semiconductor junction diode. When the injected electrons drop from the conduction band to the valence band, they emit photons. This is e.g. the case in LEDs.
- Sonoluminescence is caused by the energy acquired from a sound wave. Light emitted by water that is irradiated by a strong ultrasonic source is an example.

10.4.1 Photoluminescence

As mentioned above, photoluminescence occurs when an atomic system is excited to a higher energy level by absorption of photons, and then spontaneously decays to a lower energy level by emitting a photon. This emitted photon can not have a higher energy than the original exciting photon, unless multiple photons are together responsible for the excitation of an atom or molecule. A number of examples of luminescence are shown in figure 10.7. Intermediate nonradiative processes are also possible, indicated by the dashed line in figure 10.7. The electron can also temporarily end up in a quasi-stable state and then later decay with emission of a photon. This causes so-called delayed luminescence.

On the other hand, in photoluminescence we can distinguish radiative transitions allowed by the selection rules - this is called fluorescence - and radiative transitions forbidden by the selection rules - this is called phosphorescence. The lifetime of the electron after excitation is a lot smaller (order $0.1 - 10ns$) in the case of fluorescence compared to the lifetime in the case of phosphorescence (typically of the order $1ms - 10s$).

Photoluminescence occurs in many materials, including a few simple inorganic molecules, such as $N_2,\ CO_2,\ Hg$. It also happens in noble gases, inorganic crystals like diamond, zinc sulfide, ruby and different aromatic crystals. Even semiconductors can act as photoluminescent materials.

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

Part IV

Light as information carrier

Chapter 11

Analog and digital modulation of an optical carrier

Contents

11.1 Introduction

Optics is often used for communication purposes. In this application, an optical carrier (an electromagnetic wave at optical frequencies) is modulated by an information signal at the transmitter side, transported over a medium (e.g. free space or optical fiber) and demodulated at the receiver to retrieve the information encoded on the optical carrier. The high frequency of the optical carrier, typically around 200THz, together with the high bandwidth of the channel medium, allows transporting huge amounts of information in a given time frame, which is one of the major advantages of optical communication. In this chapter we will elaborate on how information can be transferred onto an optical carrier at the transmitter side, and how this information is retrieved at the receiver. Both the modulation of analog and digital data signals will be considered.

11.2 Analog versus digital modulation

Both analog and digital data signals can be imprinted on an optical carrier. The principal feature of a digital communication system is that during a finite interval of time, it sends a waveform from a finite set of possible waveforms. This is in contrast to an analog communication system, which sends a waveform from an infinite variety of waveform shapes with theoretically infinite resolution. While analog communication systems can in principle provide communication with infinite resolution, these communication systems are much more prone to distortion of the signal during the transport over the medium between transmitter and receiver than in digital communications. In a digital communication system, the objective at the receiver is not to reproduce a transmitted waveform with precision; it is, instead, to determine from a noise and other impairments perturbed signal which waveform from the finite set of waveforms has been sent by the transmitter. This makes digital modulation more robust than analog modulation. A drawback of digital modulation systems is that they often require a larger bandwidth to send the same information, which makes them less bandwidth efficient (although compression and coding can reduce the required bandwidth). In the following sections we will outline both the analog and digital modulation formats that are often used in practical communication systems.

11.3 Spectral content of a modulated optical carrier

An important property of a modulated optical signal is its frequency spectrum or spectral content. This spectrum can be obtained by calculating the Fourier transform $F(\omega)$ of the optical signal $f(t)$ as

$$F(\omega) = \int_{-\infty}^{+\infty} f(t) \exp(-j\omega t) dt \tag{11.1}$$

The optical signal $f(t)$ represents the time variation of an arbitrary component of the electromagnetic field.

For a real signal $f(t)$, one can easily find that $F(-\omega) = F^*(\omega)$.

For the monochromatic optical carrier $\psi(t)$, with

$$\psi(t) = A \sin(\omega_c t + \varphi_c) \tag{11.2}$$

the frequency spectrum is given by

$$F(\omega) = \frac{A}{2j} \left[e^{j\varphi_c} \partial(\omega - \omega_c) - e^{-j\varphi_c} \partial(\omega + \omega_c) \right] \tag{11.3}$$

with $\partial(\omega)$ the Dirac-function.

11.4 Analog modulation of an optical carrier

In this section we will describe various methods for analog modulation of an optical carrier. Modulation of the amplitude, the phase and frequency of the optical carrier signal will be discussed.

11.4.1 Amplitude modulation

Sinusoidal modulation of an optical carrier

A basic example of amplitude modulation of a monochromatic optical carrier is the sinusoidal modulation. In this case we can write the field as

$$\psi(t) = A(1 + m\sin(\omega_s t + \varphi_s))\sin(\omega_c t + \varphi_c) \tag{11.4}$$

in which ω_c is the radial frequency of the optical carrier, while ω_s is the radial frequency at which the amplitude of the optical carrier is modulated. m determines the depth of the amplitude modulation ($0 < m < 1$). The spectral content of this signal is characterized by its Fourier transform, which can be written in the case of sinusoidal amplitude modulation as

$$F(\omega) = \mathbf{F}[A(1 + m\sin(\omega_s t + \varphi_s))\sin(\omega_c t + \varphi_c)] \tag{11.5}$$

or

$$F(\omega) = \mathbf{F}[A\sin(\omega_c t + \varphi_c) + \frac{Am}{2}(\cos((\omega_c - \omega_s)t + (\varphi_c - \varphi_s)) - \cos((\omega_c + \omega_s)t + (\varphi_c + \varphi_s)))] \tag{11.6}$$

From this equation we can conclude that the sinusoidal amplitude modulated signal contains three frequency components (sometimes also referred to as tones), which results in three dirac-functions in the frequency spectrum (considering only the positive frequencies): ω_c, $\omega_c - \omega_s$ and $\omega_c + \omega_s$. Since the carrier frequency ω_c in practice is much larger than the modulation frequency ω_s, the three frequency components are closely spaced.

While the weight of the dirac function around the radial frequency ω_c is independent of the modulation depth, the weight of the dirac functions around $(\omega_c - \omega_s)$ and $(\omega_c + \omega_s)$ linearly scales with this modulation depth.

A sinusoidally modulated optical carrier and its Fourier transform is shown in figure 11.1(a).

Amplitude modulation of an optical carrier using a deterministic signal

A more general amplitude modulated signal can be written as

$$\psi(t) = f(t)\sin(\omega_c t + \varphi_c) \tag{11.7}$$

The spectral content of this signal can be written as

$$F(\omega) = \mathbf{F}[f(t)\sin(\omega_c t + \varphi_c)] = \mathbf{F}[f(t)] * \mathbf{F}[\sin(\omega_c t + \varphi_c)] \tag{11.8}$$

using the Fourier transform convolution property of a product of two time varying functions. Using equation 11.3, the spectral content can be written as

$$F(\omega) = \mathbf{F}[f(t)] * [\frac{1}{2j}(e^{j\varphi_c}\partial(\omega - \omega_c) - e^{-j\varphi_c}\partial(\omega + \omega_c))] \tag{11.9}$$

or

$$F(\omega) = \frac{1}{2j}e^{j\varphi_c}\tilde{f}(\omega - \omega_c) - \frac{1}{2j}e^{-j\varphi_c}\tilde{f}(\omega + \omega_c) \tag{11.10}$$

with $\tilde{f}(\omega)$ the Fourier transform of the function $f(t)$. This means that the spectral content of the amplitude modulated signal consists of the spectral content $\tilde{f}(\omega)$, centered around the carrier radial frequency ω_c (and $-\omega_c$). An amplitude modulated optical carrier and its Fourier transform is shown in figure 11.1 (b) . Again, since the spectral content of the time varying function $f(t)$ is much more narrow band than the carrier frequency ω_c, the spectral content of the modulated carrier is centered closely around this carrier frequency.

Square wave modulation of an optical carrier

Another important example of amplitude modulation of an optical carrier is the use of a square wave function, especially due to its use in digital communications. In this case, the field can be written as

$$\psi(t) = A(1 + ms(t))\sin(\omega_c t + \varphi_c) \tag{11.11}$$

with $s(t)$ satisfying

$$\begin{array}{rcll} s(t) & = & +1 & (iT < t < iT + T_1) \\ & = & -1 & (iT + T_1 < t < (i+1)T) \end{array} \tag{11.12}$$

in which i is an integer, T is the square wave period, $\frac{T_1}{T}$ is the duty cycle of the square wave and m is the modulation depth.

Using the formalism described in the previous section to determine the spectral content of the signal, one can find that the Fourier transform of the square wave modulated signal consists of a dirac function at the optical carrier radial frequency, surrounded by an infinite set of dirac functions spaced by a radial frequency $\frac{2\pi}{T}$ ($\omega = \omega_c \pm k\frac{2\pi}{T}$) and with a weight proportional to

$$w_k = \frac{Am}{\pi k}\left[1 - \exp(-j2k\pi\frac{T_1}{T})\right] \tag{11.13}$$

with k an integer not equal to zero. Indeed, one can write the periodic function $s(t)$ as a Fourier series

$$s(t) = \sum_{k=-\infty}^{+\infty} w_k \exp(j\frac{k2\pi t}{T}) \tag{11.14}$$

with

$$w_k = \frac{1}{T}\int_0^T s(t)\exp(-j\frac{k}{T}2\pi t)dt = \frac{1}{\pi k}\left[1 - \exp(-j2k\pi\frac{T_1}{T})\right] \tag{11.15}$$

Using equation 11.10, this directly results in the spectral content of the square wave modulated signal, as shown in figure 11.1(c).

11.4.2 Frequency and phase modulation

Besides the amplitude, also the frequency and phase can be used to imprint a signal on an optical carrier. In these cases, the modulated signal can be written as $\psi_f(t) = A\sin(\omega_c(t)t + \varphi_c)$ and $\psi_p(t) = A\sin(\omega_c t + \varphi_c(t))$, in the case of frequency and amplitude modulation respectively. As one can write

$$\sin(\omega_c(t)t + \varphi_c) = \sin((\omega_{c,0} + \tilde{\omega}(t))t + \varphi_c) = \sin(\omega_{c,0}t + (\tilde{\omega}(t)t + \varphi_c)) = \sin(\omega_{c,0}t + \varphi(t)) \tag{11.16}$$

it is clear that frequency modulation and phase modulation of an optical carrier are strongly related. If we consider the special case of sinusoidal phase modulation of an optical carrier, one can write the field as

$$\psi(t) = A\sin(\omega_c t + 2\pi m\sin(\omega_s t)) \tag{11.17}$$

or

$$\psi(t) = \frac{A}{2j}(e^{j(\omega_c t + 2\pi m\sin(\omega_s t))} - e^{-j(\omega_c t + 2\pi m\sin(\omega_s t))}) \tag{11.18}$$

with m the modulation depth of the sinusoidal phase modulation ($0 < m < 1$).

In order to find the spectral content of a sinusoidally phase modulated monochromatic carrier, the *Bessel functions of the first kind* have to be introduced. These functions can be defined as the coefficients $J_k(\beta)$ in the so-called two-sided Laurent expansion of the generating function

$$e^{\frac{1}{2}\beta(z-\frac{1}{z})} = \sum_{k=-\infty}^{+\infty} J_k(\beta)z^k \tag{11.19}$$

Replacing z by $e^{j\omega_s t}$, we obtain

$$x(t) = e^{j\beta \sin(\omega_s t)} = \sum_{k=-\infty}^{+\infty} J_k(\beta) e^{jk\omega_s t} \tag{11.20}$$

In other words, $J_k(\beta)$ is the amplitude of the kth harmonic in the Fourier series expansion of the periodic signal $x(t)$.

Multiplying equation 11.33 by $e^{j\omega_c t}$ yields

$$x(t)e^{j\omega_c t} = e^{j(\omega_c t + \beta \sin(\omega_s t))} = \sum_{k=-\infty}^{+\infty} J_k(\beta) e^{j(k\omega_s + \omega_c)t} \tag{11.21}$$

From equation 11.17 and equation 11.21, one finds that

$$A\sin(\omega_c t + 2\pi m \sin(\omega_s t)) = \frac{A}{2j}\left(\sum_{k=-\infty}^{+\infty} J_k(2\pi m) e^{j(k\omega_s+\omega_c)t} - \sum_{k=-\infty}^{+\infty} J_k^*(2\pi m) e^{-j(k\omega_s+\omega_c)t}\right) \tag{11.22}$$

One can show that for real values of β, $J_k(\beta)$ is also real. Therefore, by taking the Fourier transform of this expansion, we find that the spectral content of the sinusoidally phase modulated carrier contains frequency components at $\omega_c + k\omega_s$ and $-(\omega_c + k\omega_s)$ (with k an integer from $-\infty$ to $+\infty$) with a complex amplitude equal to $\pm\frac{A}{2j}J_k(2\pi m)$. This is shown in figure 11.1(d). The lowest order Bessel functions of the first kind are shown in figure 11.2. Negative order Bessel functions are related to these positive order functions by $J_{-n}(\beta) = (-1)^n J_n(\beta)$.

It can be seen in the figure that when β is zero, $J_0(0) = 1$ and $J_k(0) = 0$ for all k different from zero. Since $J_0(\beta)$ is the amplitude of the carrier frequency, there are no side bands when $\beta = 0$ (as there is no modulation). As β (and therefore the modulation depth m) increases, the sidebands begin to grow while the carrier term diminishes. This is how phase modulation produces an expanded bandwidth as the modulation depth is increased. As a comparison, this is not the case for a sinusoidally amplitude modulated optical carrier, where the bandwidth doesn't increase for an increased modulation depth .

11.4.3 Intensity modulation

So far, the discussed modulation schemes assumed a monochromatic optical carrier onto which information was imprinted by means of amplitude, phase or frequency modulation. This is the case when narrow line width laser light is used as an optical carrier. For short distance communication, often a low cost light emitting diode is used, of which the output intensity is modulated. As the emission spectrum of this light source is already very broadband, modulating the intensity nearly doesn't modify the spectral content of the optical signal.

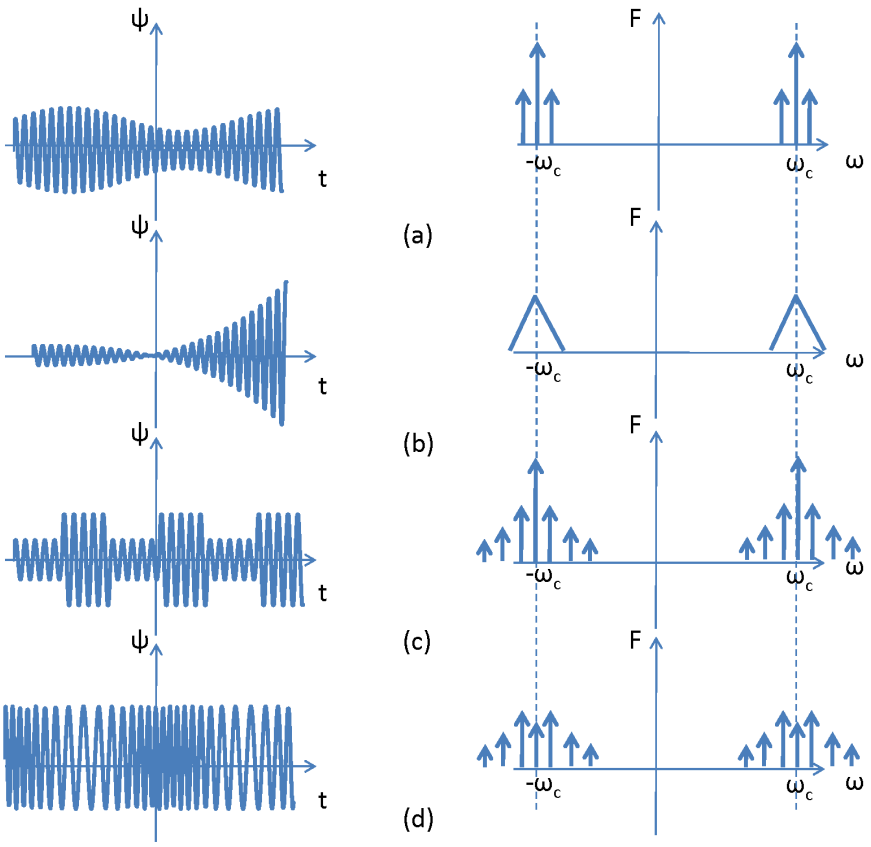

Figure 11.1: Various analog modulation formats of monochromatic carriers: amplitude modulation using a sinusoidal signal, a signal $f(t)$ and a square wave signal, and sinusoidal phase modulation. Thee figures shown the spectral content are not drawn to scale for clarity (typical optical carrier frequency $f_c = 193THz$ and a typical modulation frequency is $f_s = 10GHz$)

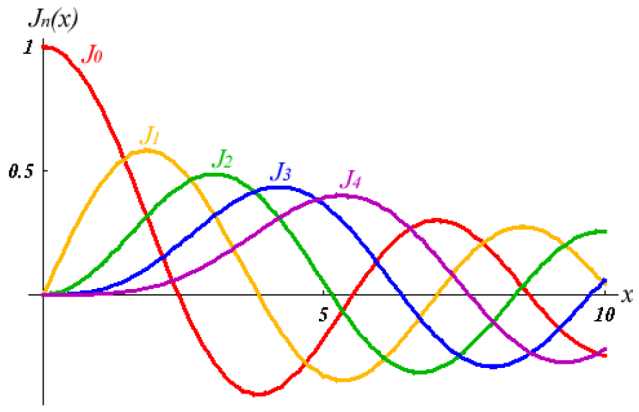

Figure 11.2: Lowest order Bessel functions of the first kind

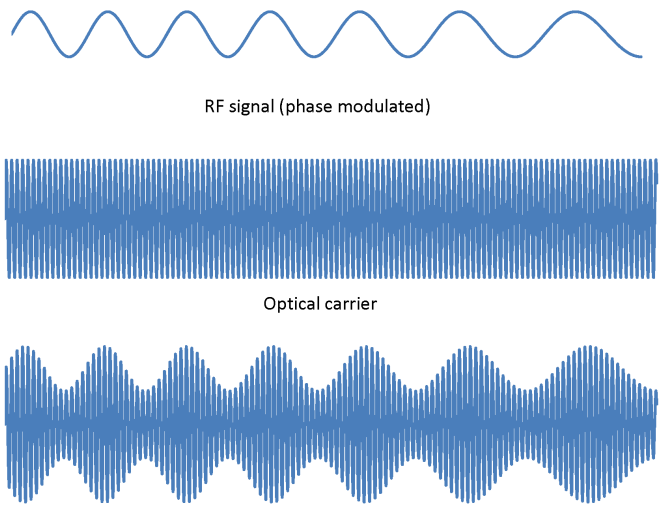

Figure 11.3: Principle of Radio-over-Fiber technology

11.4.4 Optical carrier versus radio frequency carrier

From the above discussion it is clear that amplitude, phase and frequency modulation can be realized on a monochromatic carrier signal. No assumptions have been made on the frequency of this carrier however. Therefore, this carrier could be both an optical carrier or a radio frequency (RF) carrier. One can also modulate the amplitude of the optical carrier with an RF wave, onto which the information is encoded. This technique of modulating the radio frequency subcarrier onto an optical carrier for distribution over i.e. a fiber network is known as radio-over-fiber technology (RoF). For example, for the case of a phase modulated RF subcarrier imprinted on an optical carrier by means of amplitude modulation, the field can be written as

$$\psi(t) = (1 + A\sin(\omega_{sc,RF}t + \varphi_{sc,RF}(t))\sin(\omega_{c,OP}t + \varphi_{c,OP}(t)) \tag{11.23}$$

This radio-over-fiber technology is schematically illustrated in figure 11.3.

11.5 Digital modulation

In the previous section the analog modulation of an optical carrier signal was discussed. In this section we will describe how an optical carrier can be imprinted with digital information. However, as nearly all information is analog by nature, we will first discuss how one has to transform the analog information into a digital form, by means of the sampling theorem.

11.6 Sampling theorem

Consider the analog signal $f(t)$ represented in figure 11.4. We will consider this to be a bandlimited signal. This means that the signal contains no energy at radial frequencies higher than some

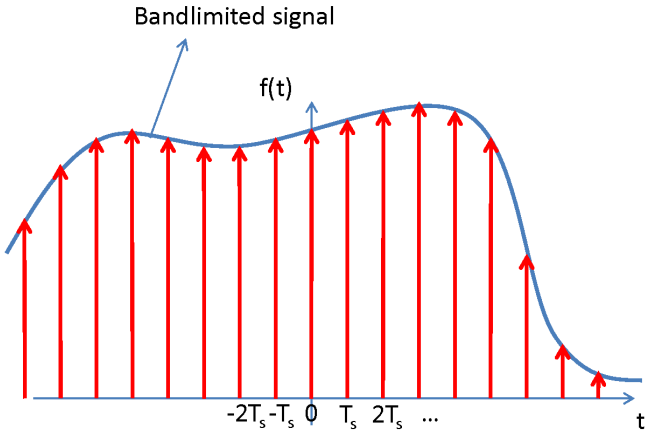

Figure 11.4: Nyquist-Shannon sampling theorem: construction of the function $x(t)$

bandwidth $2\pi B$. The sampling theorem describes how fast one should sample this analog signal (sampling is the process of converting a signal, i.e. a function of continuous time, into a numeric sequence, i.e. a function of discrete time) without losing information from the original signal. It might seem strange that there exists such a minimum sampling rate (referred to as the Nyquist rate), as all the information in between two samples is lost. However, a signal that is bandlimited (as assumed) is constrained in how rapidly it can change in time, and therefore how much detail it can convey in an interval of time. The sampling theorem asserts that the uniformly spaced discrete samples are a complete representation of the signal if the signal bandwidth B is less than half the sampling rate f_s. This is the so called Nyquist-Shannon sampling theorem. In this section we shall derive this sampling theorem.

To prove the Nyquist- Shannon theorem, we will construct a new signal $x(t)$ from the samples taken at nT_s from the original signal $f(t)$ as follows:

$$x(t) = \frac{T_s}{2\pi} \sum_{k=-\infty}^{+\infty} f(kT_s)\partial(t - nT_s) \tag{11.24}$$

This signal consists of a set of dirac functions, spaced by the sampling interval T_s and with a weight equal to the corresponding sample $f(kT_s)$, as shown in figure 11.4 . $x(t)$ can be rewritten as

$$x(t) = \frac{T_s}{2\pi} \sum_{k=-\infty}^{+\infty} f(kT_s)\partial(t - nT_s) = \frac{T_s}{2\pi} f(t) \sum_{k=-\infty}^{+\infty} \partial(t - nT_s) \tag{11.25}$$

Calculating the Fourier transform of this signal $X(\omega)$ results in

$$X(\omega) = \frac{T_s}{2\pi} \tilde{f}(\omega) * F[\sum_{k=-\infty}^{+\infty} \partial(t - kT_s)] = \frac{T_s}{2\pi} \tilde{f}(\omega) * \sum_{k=-\infty}^{+\infty} \mathrm{e}^{-j\omega kT_s} \tag{11.26}$$

with $\tilde{f}(\omega)$ the Fourier transform of the bandlimited signal $f(t)$. As

$$\frac{T_s}{2\pi}\sum_{k=-\infty}^{+\infty} \mathrm{e}^{-j\omega kT_s} = \sum_{k=-\infty}^{+\infty} \partial(\omega - \frac{2\pi k}{T_s}) \tag{11.27}$$

we can rewrite equation 11.26 as

$$X(\omega) = \tilde{f}(\omega) * \sum_{k=-\infty}^{+\infty} \partial(\omega - \frac{2\pi k}{T_s}) = \sum_{k=-\infty}^{+\infty} \tilde{f}(\omega - \frac{2\pi k}{T_s}) \tag{11.28}$$

This equation implies that the Fourier transform of the constructed signal $x(t)$ consists of an infinite number of replicas of the spectral content of the original analog signal $f(t)$, shifted by $\frac{2\pi k}{T_s}$. This is shown in figure 11.5 for two cases: in case (a), the highest radial frequency occurring in the original analog signal $2\pi B$ is larger than $\frac{\pi}{T_s}$, while in case (b), the highest radial frequency occurring in the original analog signal $2\pi B$ is smaller than $\frac{\pi}{T_s}$. It is clear that in case (a), it is no longer possible to reconstruct the original Fourier transform of the analog signal $f(t)$ from $X(\omega)$, as part of the Fourier spectrum overlaps with a shifted version of that spectrum (and adds therefore). In case (b) the Fourier spectrum doesn't overlap with a shifted version, and therefore the original Fourier spectrum (and hence the original analog signal $f(t)$) can be reconstructed, by passing the signal $x(t)$ at the receiver side through an ideal low-pass filter with a cut-off radial frequency $\frac{\pi}{T_s}$. Therefore, as soon as $\frac{2\pi}{T_s} \geq 2(2\pi B)$ or $f_s \geq 2B$, all the information of the original bandlimited analog signal is contained in the discrete samples at times $\frac{n}{f_s}$. This is the so-called Nyquist criterion.

Passing the signal $x(t)$ at the receiver side through an ideal low-pass filter (transmission $LP(\omega)$ equal to 1 for radial frequencies lower than $\frac{\pi}{T_s}$, and zero elsewhere), implies that the individual dirac pulses need to be replaced by sinc-functions. Indeed

$$X'(\omega) = LP(\omega)X(\omega) \tag{11.29}$$

or

$$x'(t) = h_{LP}(t) * x(t) = h_{LP}(t) * \frac{T_s}{2\pi}\sum_{k=-\infty}^{+\infty} f(kT_s)\partial(t - kT_s) = \frac{T_s}{2\pi}\sum_{k=-\infty}^{+\infty} f(kT_s)h_{LP}(t - kT_s) \tag{11.30}$$

with $h_{LP}(t)$ the impulse response of the filter

$$h_{LP}(t) = F^{-1}[LP(\omega)] = \int_{-\frac{\pi}{T_s}}^{+\frac{\pi}{T_s}} 1e^{+j\omega t}d\omega = \frac{\sin(\frac{\pi t}{T_s})}{\frac{\pi t}{T_s}}\frac{2\pi}{T_s} \tag{11.31}$$

$h_{LP}(t)$ is plotted in figure 11.6. Note that the sinc-function is zero for $t = nT_s$, with n an integer different from zero.

In case (a), where part of the Fourier spectrum of $f(t)$ overlaps with a shifted version of that spectrum, frequencies above half the sampling rate will be reconstructed as (and appear as) frequencies

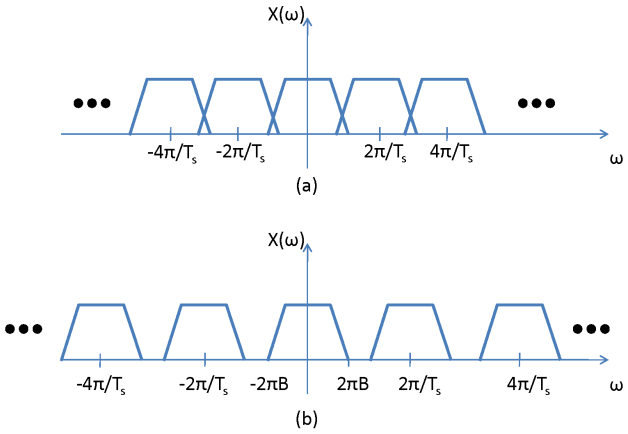

Figure 11.5: Spectral content of the signal $x(t)$

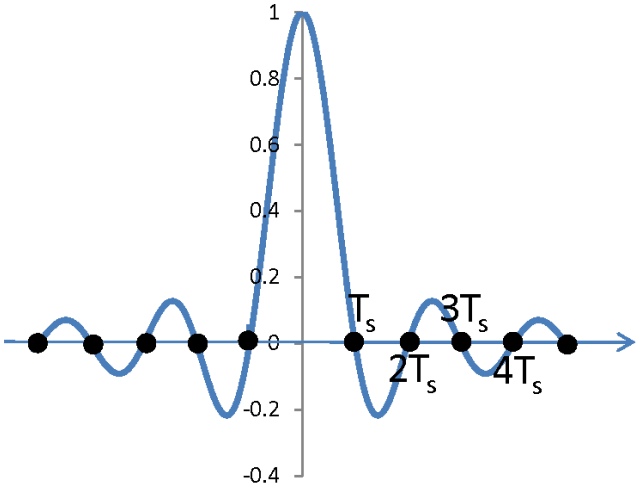

Figure 11.6: shape of the sinc-function required for the reconstruction of the analog input signal

below half the sampling rate. The resulting distortion is called aliasing; the reconstructed signal is said to be an alias of the original signal, in the sense that it has the same set of sample values.

In this analysis, we assumed that the signal to be sampled was a baseband signal $f(t)$, this is a signal with a central frequency at $\omega = 0$. In the case of a modulated carrier, the spectral content is located around the carrier frequency ω_c and $-\omega_c$. The information encoded in the signal is however typically a baseband signal before modulation at the transmitter and after detection at the receiver. Therefore, the Nyquist-Shannon criterion can also be applied to the discussed modulated carriers.

In practice, a signal $f(t)$ will never be perfectly bandlimited, neither can the reconstruction formula be precisely implemented. This would require the summing of an infinite number of points, and weighing it with a sinc-function which is infinitely extending in time. Also, an infinite resolution was assumed, which is a practical system not the case. This quantization which occurs due to the availability of only a limited number of bits, further distorts the sampling process.

11.7 Bandwidth of optical signals

While in the previous analytical expressions the radial frequency was used to calculate the bandwidth of an optical signal, in practice the corresponding frequency $f = \frac{\omega}{2\pi}$ is used. So, a 1.55μm optical carrier (carrier frequency about 193THz) which is modulated with a periodic signal with a period of 100ps, occupies an optical bandwidth of a few tens of GHz (depending on the modulation format) around the 193THz carrier frequency.

11.8 Digital modulation formats

In the section on analog modulation formats, information was imprinted directly on an optical carrier by means of amplitude modulation, phase modulation or frequency modulation. Also the concept of using RF subcarriers was introduced, which are modulated on an optical carrier. These schemes are also used for digital modulation formats. One major difference with the analog modulation is that in digital modulation formats, only a discrete set of signals are available. At the receiver side, the objective is to determine from a noise and other impairments perturbed signal which waveform from the finite set of waveforms has been sent by the transmitter. In this section we will describe amplitude shift keying (ASK), phase shift (PSK) and frequency shift keying (FSK), as well as a combination of amplitude and phase shift keying, namely quadrature amplitude modulation (QAM). Besides the modulation formats, we will also consider how the digital information is demodulated from the optical carrier / RF subcarrier at the receiver side. In order to schematically represent the different modulation formats, the constellation diagram will be introduced first.

11.8.1 Constellation diagram

A monochromatic carrier can be modulated in amplitude, phase and frequency. In case the modulation is restricted to amplitude and phase, the different "digital symbols" that can be sent by

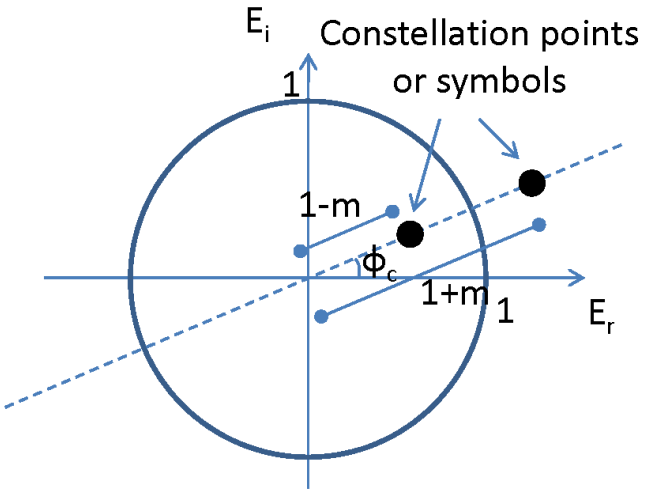

Figure 11.7: Constellation diagram

the transmitter can be represented in a two-dimensional constellation diagram, plotting the possible phasors $A_k e^{j\varphi_k}$ of the optical carrier in the complex plane, which correspond to a signal $A_k sin(\omega_c t + \varphi_k)$ in a signalling interval T. An example of such a constellation diagram, in the case of the square wave amplitude modulation of a monochromatic optical carrier, as described in the section on analog modulation formats, is shown in figure 11.7. Although in theory the constellation diagram consists only of discrete points, impairments at the transmitter, along the communication channel and at the receiver side (such as noise), will blur these constellation diagrams, thereby requiring decision circuitry to decide which symbol was originally sent. This addition of noise will be considered in more detail in the next chapter.

11.8.2 Amplitude shift keying

In amplitude shift keying (ASK), 2^N amplitude levels can be sent by the transmitter. These amplitude levels are equidistant. This results in the constellation diagram shown in figure 11.10, for the case of $N = 2$. In this case, each symbol sent by the transmitter consists of two bits of information.

In the case $N = 1$, ASK is referred to as OOK (On-Off keying). This is the simplest form of amplitude shift keying. However, more complex waveforms can be used to to represent a digital '1' or '0' in the OOK format . The most used waveforms are non-return-to-zero (NRZ), return-to-zero (RZ) and Manchester coding.

The NRZ waveform is probably the most commonly used waveform. In this case typically a binary 1 is represented by one level and a binary zero is represented by another level. There is a change in level whenever the data changes from a one to a zero or from a zero to a one.

In RZ-coding, a binary 1 is represented by a half-bit-wide pulse, and a zero is represented by the absence of a pulse.

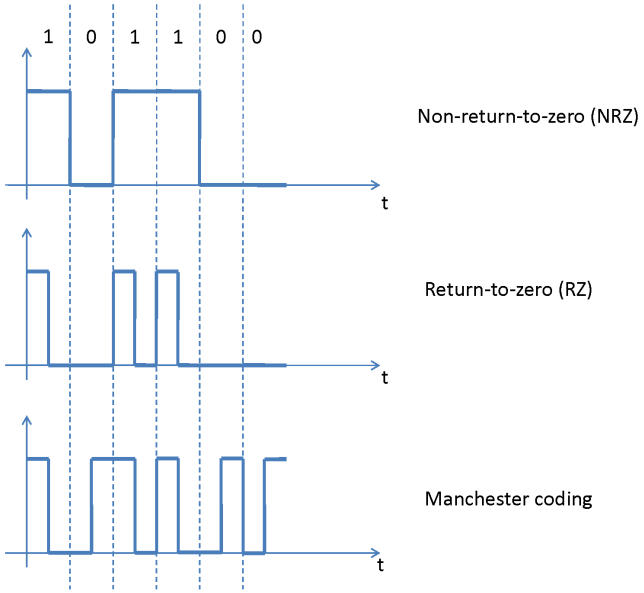

Figure 11.8: Line coding formats: non-return-to-zero, return-to-zero and Manchester coding

In Manchester coding, a binary 1 is represented by a half-bit-wide pulse positioned during the first half of the bit interval. A zero is represented by a half-bit-wide pulse positioned during the second half of the bit interval.

These three modulation formats are illustrated in figure 11.8 .

Besides these waveforms there exist a myriad of other waveform formats (or line codes). The reason for the large selection relates to the differences in performance that characterize each waveform. Some waveform formats allow for easy self-clocking (bit synchronization is for example easier in Manchester coding as there is a transition in the middle of every bit interval wether a one or a zero is sent), some waveforms are more immune to noise (for example NRZ waveforms have a better immunity to noise than RZ), and some waveforms allow a more efficient bandwidth utilization, by allowing a reduction in the required signal bandwidth for a given data rate.

11.8.3 Phase shift keying

In phase shift keying (PSK), only the phase of the optical signal is varied and can take on 2^N different values, while the amplitude stays constant. This means that the constellation points are located on a circle with radius 1 in the complex plane. The case of $N = 2$ is particularly interesting as it is used a lot in practical systems. This results in the quadrature phase shift keying (QPSK) constellation diagram shown in figure 11.10.

Differential phase shift keying (DPSK) is a format that is often used in high-speed optical communication systems. The modulating signal is not the binary code itself, but a code that records changes in the binary code.

For forming for example a DBPSK (differential binary PSK) signal from a BPSK signal (with constellation points at 1 and -1 in the complex plane, which makes it an alternative form of ASK),

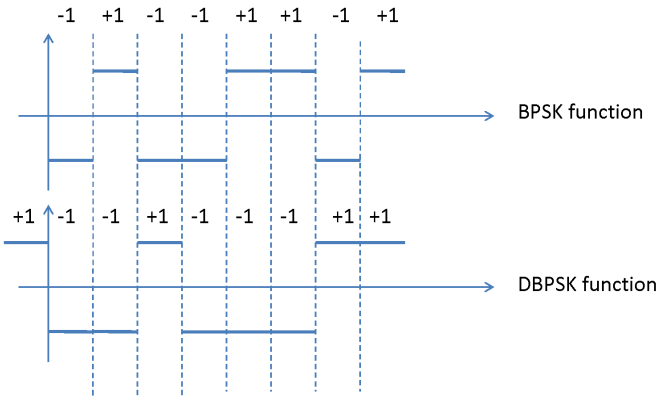

Figure 11.9: BPSK and Differential BPSK modulation format

the BPSK signal is converted to a DBPSK signal by two rules: a 1 in the BPSK signal (phase 0) is denoted by no change in the DBPSK signal and a -1 (π phase shift) in the BPSK signal is denoted by a change in the DBPSK signal. The DBPSK sequence is initialized with a leading 1. An example of corresponding patterns is shown in figure 11.9. In order to find the optical field, the optical carrier signal $sin(\omega_c t + \varphi_c)$ has to be multiplied by the respective (D)BPSK functions. In differentially-encoded QPSK (DQPSK), the phase-shifts are 0, $\frac{\pi}{2}$, π, $-\frac{\pi}{2}$ corresponding to data '00', '01', '11', '10'.

11.8.4 Quadrature amplitude modulation

A combination of Amplitude Shift Keying (ASK) and Phase Shift Keying (PSK) is used in quadrature amplitude modulation (QAM). In this case the constellation diagram consists of a square lattice of constellation points. When only four lattice points are used (QAM-4), this modulation format is the same as QPSK, as shown in figure 11.10. More complex constellation diagrams, e.g. QAM-512, are used in practice, typically in the RF subcarrier modulation format.

11.8.5 Frequency shift keying

In Frequency Shift Keying (FSK), a different frequency of the optical carrier wave is used to represent the different symbols. Typically, FSK in its standard form (wide-band FSK) consumes a lot of bandwidth, as two different carrier frequencies are used, which are not very closely spaced. When the difference between the optical frequencies of the two optical carriers is half of the data rate and the phase at each bit transition instant is continuous, a so-called minimum shift-keying (MSK) format is used. The signal $s(t)$ can be written in this case as

$$s(t) = \cos[\omega_c t + b_k(t)\frac{\pi t}{2T} + \phi_k] \tag{11.32}$$

where $b_k(t)$ takes on the value of +1 and -1, and ϕ_k is chosen such that the phase of the signal changes continuously. From this expression it is clear that the frequency shift keying is the same as phase modulation of the optical carrier using a sawtooth driving signal. MSK has the advantage

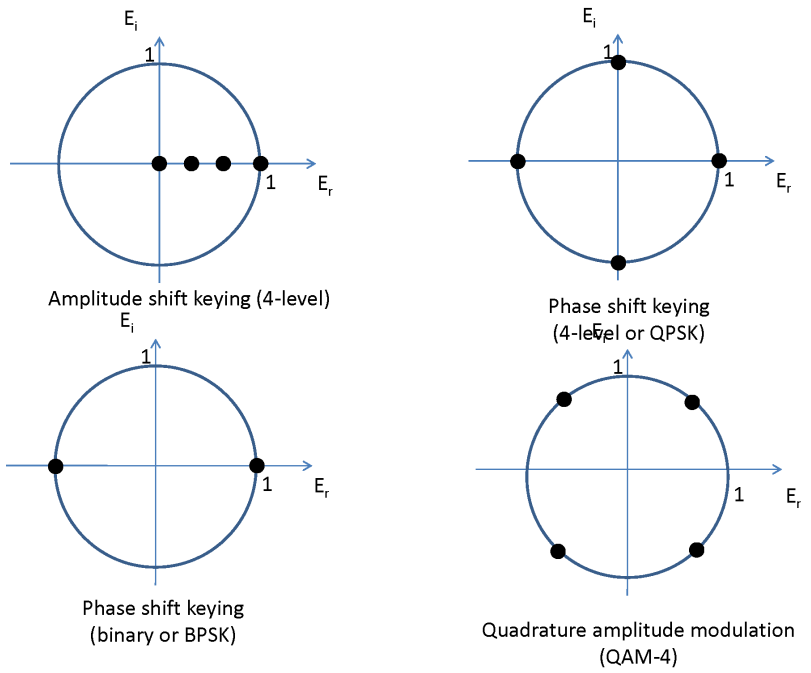

Figure 11.10: Constellation diagram of various modulation formats: ASK, BPSK, QPSK and QAM-4

of higher spectral efficiency compared to simple FSK. The demodulation is however somewhat more complicated.

11.9 Demodulation

After the information was modulated onto the carrier at the transmitter and transmitted over the channel medium, it needs to be demodulated again at the receiver side.

There are two types of demodulation which are distinguished by the need to know the phase of the carrier (or the RF subcarrier) to perform the demodulation. Demodulation schemes requiring the knowledge of the phase of the carrier are termed coherent. Those that do not need the phase are termed incoherent. Incoherent demodulation can be applied to ASK and FSK (when sending the optical signal through a bandpass filter, which filters out one carrier frequency).

In PSK, the information is demodulated at the receiver by means of coherent optical detection. Coherent optical detection implies using a local oscillator (LO) at the receiver side, of which the phase is locked with respect to the source used at the transmitter side. With coherent demodulation systems, the incoming signal is compared with a replica of the carrier wave. For example, by letting the local oscillator and the incoming field interfere on an envelope detector, we find that a signal s_k is created, the amplitude of which is directly related to the phase of the incoming data signal.

$$s_k =< (\sin(\omega_c t) + \sin(\omega_c t + \varphi_k))^2 >= 1 + \cos(\varphi_k) \tag{11.33}$$

The difficulty with coherent detection is the need to keep the phase of the replica signal, termed local oscillator, "locked" to the carrier. This is not easy to do, definitely not in optical communication (it is done frequently for RF subcarrier demodulation). Oscillators are sensitive to (among other things) temperature, and a "free-running" oscillator will gradually drift in frequency and phase.

Therefore, often *differential* phase shift keying is used. Differential PSK is actually a simple form of coding. The modulating signal is not the binary code itself, but a code that records changes in the binary code, as explained in the previous section. This way, the demodulator only needs to determine changes in the incoming signal phase. This it can do done by comparing the incoming signal with the same signal which is delayed by one bit period.

11.10 PRBS signals and eye diagrams

For testing a communication channel or individual optical signal processing components, it is interesting to have a random digital optical signal at the input and assess how the component under investigation processes this signals. In practice, a pseudo random bit sequence (PRBS) is used as a data signal. This PRBS is created by an algorithm which generates a sequence of numbers (bits) that approximate the properties of random numbers. The sequence is not truly random in that it is completely determined by a relatively small set of initial values. The maximal length of the PRBS is determined by the algorithm and is $2^N - 1$, after which the PRBS signal starts to repeat itself.

The eye diagram is an oscilloscope display of a digital signal, repetitively sampled to get a good representation of its behavior. The eye diagram is a useful tool for the qualitative analysis of signals used in digital transmission. It provides at-a-glance evaluation of system performance and can offer insight into the nature of imperfections. Careful analysis of this visual display can give the user a first-order approximation of the signal-to-noise ratio, clock timing jitter, etc.. Linear impairments (such as dispersion and the associated intersymbol interference or a lack of bandwidth in the system) and nonlinear impairments can be observed. An undistorted eye diagram of a bandlimited signal together with a "real life" eye diagram is shown in figure 11.11.

11.11 Multiplexing techniques

Multiplexing techniques are used to more efficiently make use of the bandwidth of the channel medium (i.e. optical fiber in the case of optical communication). It allows to send multiple signals, addressed to different users, over a single optical fiber. These signals therefore have to be multiplexed at the transmitter side and are demultiplexed (unraveled) at the receiver side.

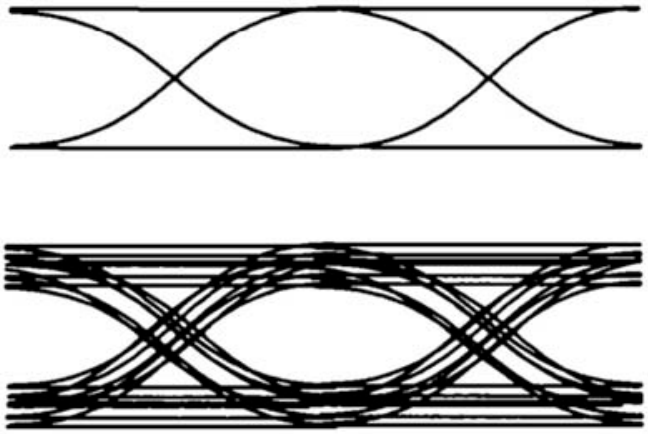

Figure 11.11: Eye diagram of a PRBS signal: undistorted eye diagram and a real life eye diagram

11.11.1 Wavelength division multiplexing

Wavelength-division multiplexing (WDM) is a technology which multiplexes multiple optical carrier signals on a single optical fiber by using different wavelengths of laser light to carry different signals. This allows for a multiplication in capacity of the fiber-optic link. Each individual optical carrier can be modulated using a different modulation format. A WDM system uses a multiplexer at the transmitter to join the signals together, and a demultiplexer at the receiver to split them apart. This technology is used both in long haul telecommunication systems (using DWDM or dense WDM, where the different optical carrier wavelengths are spaced 0.8nm or 100GHz apart) and in shorter distance communication link (using CWDM or coarse WDM, where the different optical carrier wavelengths typically are spaced 20nm apart).

11.11.2 Frequency domain multiplexing

As discussed before, one can also modulate the amplitude of the optical carrier with an RF subcarrier wave, onto which the information is encoded. This means that multiplexing can also be achieved in the RF domain, using RF subcarriers with a different frequency and imprinting these on a single optical carrier.

11.11.3 Time domain multiplexing

In the case of time domain multiplexing, the time domain is divided into several recurrent time slots of fixed length, one for each sub-channel (each user). One TDM frame consists of one time slot per sub-channel. After the last sub-channel the cycle starts all over again with a new frame. Each user is only allowed to transmit/receive information in its assigned time slot.

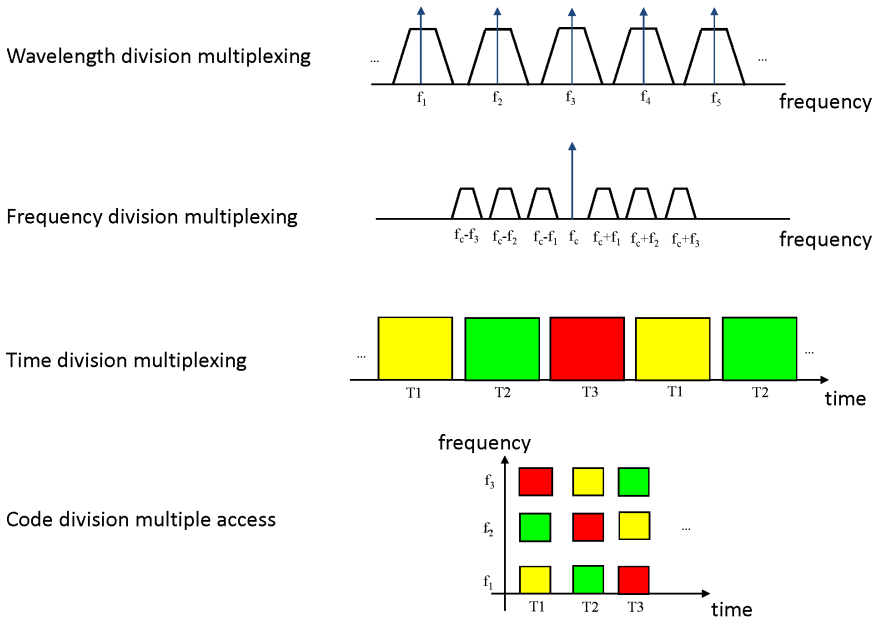

Figure 11.12: Multiplexing techniques: wavelength division multiplexing, frequency division multiplexing, time division multiplexing and code division multiple access

11.11.4 Code division multiple access

A hybrid combination of frequency domain multiplexing and time domain multiplexing is code-division multiple access or CDMA. CDMA employs a special coding scheme (where each transmitter and receiver is assigned a code) to allow multiple users to be multiplexed over the same physical channel. In this case, the time domain is divided into several time slots, and the assignment of a particular frequency band to a signal source is reordered during each time slot. This is the so-called frequency hopping CDMA. Each user receives the whole signal, but employs a code sequence, which allows him to extract the information which was meant for him, while suppressing the other channels.

CDMA is a form of "spread-spectrum" signaling, since the modulated coded signal has a much higher bandwidth than the data being communicated. Spread-spectrum techniques allow for higher security in information transport (eves dropping) and are more resistant to for example jamming.

These four multiplexing techniques are schematically illustrated in figure 11.12: for the wavelength division multiplexing different carrier frequencies are individually modulated and multiplexed on a single fiber, while in FDM a single optical carrier is imprinted with frequency multiplexed RF signals. For time domain multiplexing and CDMA, the "codes" in time space (and frequency space for CDMA) are also shown.

Chapter 12

Optical signals with stochastic modulation

Contents

12.1 Introduction

In the previous chapter, various modulation formats for optical communication were described. The analysis of these signals was however restricted to deterministic signals: the "information" which was encoded on the optical carrier wave was a periodic signal (sinusoidal wave or square wave), which contains no real information. Also the influence of noise in the communication channel was not assessed (which is in a real communication system always added, both at the transmitter side, the communication channel and at the receiver). In order to discuss these topics, stochastic signal analysis needs to be introduced in this chapter. Indeed, noise can be regarded as a stochastic process, and also the encoding of "real" information onto the optical carrier can be described by a stochastic process (although the information is not random in nature for the transmitter and receiver).

12.2 Stochastic signals

12.2.1 Stochastic variables

A statistical variable or stochastic variable (e.g. x) is an ensemble of possible values X which all have a certain probability density of occurrence $f(X)$. A stochastic process $x(t)$ is an ensemble of possible functions $X(t)$ with a certain probability density of occurrence $f(X(t))$. Alternatively, one

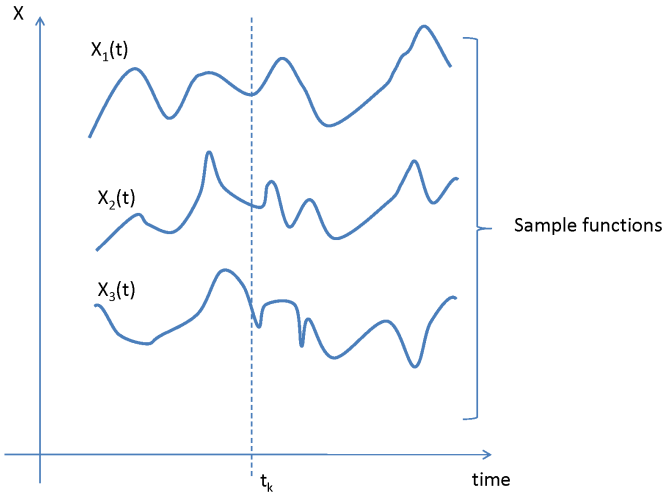

Figure 12.1: Sample functions from a stochastic process

can say that at each time instant t, $x(t)$ is a statistical variable with a certain probability density function $f(x(t))$. However the probability function can be different at different times, i.e. $f = f(x;t)$. An example of a few sample functions $X_i(t)$ from a stochastic process are shown in figure 12.1.

The function $f(x(t);t)$ doesn't completely define a stochastic process however: it only defines the first order statistical properties. In general, one also needs higher order statistical properties, e.g. $f(x(t_1), x(t_2), t_1, t_2)$, etc. to completely define a stochastic process.

12.2.2 Stationarity and ergodicity

Most (if not all) of the stochastic processes with a physical origin can be called approximately *stationary*. This means that none of their statistics are time-dependent; i.e. a time shift doesn't affect any of its statistical properties. For a stationary process, $f(x;t)$ would be independent of time, i.e. $f(x;t) = f(x)$. The same holds for the average value (or first order moment) of a stationary process: $m =< x(t) >$ with $<>$ the expectation (or statistical average) value

$$m = \int x f(x) dx \tag{12.1}$$

with m a constant, independent of time. For a stationary process, the second order moment can be written as

$$< x(t_1)x(t_2) >= f(t_2 - t_1) \tag{12.2}$$

Most physical stochastic processes are in addition *ergodic*, which means that the statistical averages over different sample functions can also be calculated as time averages of a single sample function. This ergodicity thus allows to calculate the statistical averages, even if no information about the statistics is available and it is therefore a very useful property. For ergodic processes one thus has:

$$m = \lim_{T\to\infty} \frac{1}{T} \int_{\frac{-T}{2}}^{\frac{T}{2}} x(t)dt \tag{12.3}$$

12.2.3 Autocorrelation of a stochastic process

The autocorrelation R_x of a stochastic process $x(t)$, which is stationary and ergodic, can be defined as:

$$R_x(\tau) =< x(t)x(t+\tau) >= \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} x(t)x(t+\tau)dt \tag{12.4}$$

Due to the ergodicity and stationarity, this quantity is not stochastic. The autocorrelation is a measure of the statistical correlation between the values of $x(t)$ and the values of $x(t+\tau)$. That is, if $R_x(\tau)$ decreases rapidly as τ increases, there is little correspondence between values of x and values of x, a time τ later. Or in other words, in this case we have a stochastic process $x(t)$ which fluctuates very fast in time.

12.2.4 Spectral density of a stochastic process

Since the autocorrelation $R_x(\tau)$ decreases rapidly with τ if the stochastic process contains very fast fluctuations, and vice versa, conclusions about the frequency content of $x(t)$ can be drawn from $R_x(\tau)$. This frequency content is expressed by the power spectrum of $x(t)$, which can be derived as follows. With $F_x(\omega)$ the Fourier transform of $x(t)$, we can express the average power $P =< x^2 >$ as:

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x^2(t)dt = \lim_{T\to\infty} \frac{1}{T} \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)dt \int_{-\infty}^{+\infty} F_x(\omega)e^{j\omega t}d\omega \tag{12.5}$$

or

$$P = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{+\infty} F_x(\omega)d\omega \int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)e^{j\omega t}d\omega = \lim_{T\to\infty} \frac{1}{T} \int_{-\infty}^{+\infty} |F_x(\omega)|^2 \, d\omega = \int_{-\infty}^{+\infty} S_x(\omega)d\omega \tag{12.6}$$

because $F_x(-\omega) = F_x^*(\omega)$ for a real signal $x(t)$. The power spectrum of x, $S_x(\omega)$, is therefore defined as:

$$S_x(\omega) = \lim_{T\to\infty} \frac{1}{T} |F_x(\omega)|^2 \tag{12.7}$$

Mathematically speaking, the above analysis was not rigorously correct, as a signal $x(t)$ of infinite duration doesn't necessarily have a Fourier transform, e.g. the integral from $-\infty$ to $+\infty$ might not converge or in other words, $F_x(\omega)$ could be singular. The power spectrum as defined through the limit for T on the other hand will always exist, even for infinitely long signals.

Using the definitions for $S_x(\omega)$ and $R_x(\tau)$, one can easily derive the relation between both. Indeed,

$$\begin{aligned} R_x(\tau) &= \lim_{\mathrm{T}\to\infty} \tfrac{1}{T} \int_{-T/2}^{T/2} dt \int_{-\infty}^{\infty} \mathrm{d}\omega_1 F_x(\omega_1) \exp(j\omega_1 t) \int_{-\infty}^{\infty} \mathrm{d}\omega_2 F_x(\omega_2) \exp(j\omega_2(t+\tau)) \\ &= \lim_{\mathrm{T}\to\infty} \tfrac{1}{T} \int_{-\infty}^{\infty} \mathrm{d}\omega_1 \int_{-\infty}^{\infty} \mathrm{d}\omega_2 F_x(\omega_1) F_x(\omega_2) \exp(j\omega_2\tau) \int_{-\mathrm{T}/2}^{\mathrm{T}/2} \mathrm{dt} \exp[jt(\omega_1+\omega_2)] \\ &= \lim_{\mathrm{T}\to\infty} \tfrac{1}{T} \int_{-\infty}^{\infty} \mathrm{d}\omega_1 \int_{-\infty}^{\infty} \mathrm{d}\omega_2 F_x(\omega_1) F_x(\omega_2) \exp(j\omega_2\tau) \partial(\omega_1+\omega_2) \end{aligned} \tag{12.8}$$

This expression can be rewritten as (because $F_x(-\omega) = F_x^*(\omega)$ for a real signal $x(t)$)

$$\mathrm{R_x}(\tau) = \lim_{\mathrm{T}\to\infty} \int_{-\infty}^{\infty} \mathrm{d}\omega \frac{|F_x(\omega)|^2}{T} \exp(j\omega\tau) = \int_{-\infty}^{\infty} \mathrm{d}\omega S_x(\omega) \exp(j\omega\tau) \tag{12.9}$$

In other words, the power spectrum and the autocorrelation are each others Fourier transform. This is the so-called *Wiener-Khintchine theorem*.

12.2.5 White processes and Gaussian processes

Stochastic processes for which $S_x(\omega)$ is constant are called white processes. This implies that the autocorrelation function is a dirac function or $R_x(\tau) = N\partial(\tau)$.

Stochastic processes are Gaussian if the distribution function $f(x_1, ..., x_n; t_1, ..., t_n)$ is Gaussian for all values of n and $t_1, ..., t_n$. Stationary and ergodic Gaussian stochastic processes are completely determined by their statistical average m and autocorrelation function $R_x(\tau)$ or by their statistical average m and spectral density $S_x(\omega)$.

12.3 Power spectrum of digitally modulated signals

In this section we will use the analysis of stochastic signals described above, to assess the power spectrum of various types of digitally modulated signals. We will describe non-return-to-zero (NRZ) amplitude shift keying, return-to-zero (RZ) amplitude shift keying and (differential) phase shift keying signals.

12.3.1 non-return-to-zero amplitude shift keying

In an amplitude shift keying modulation format, the modulated optical field can be written as $\psi(t) = v(t)\sin(\omega_c t)$. In the case of a non-return-to-zero on-off keying format, $v(t)$ can be written as

$$\mathrm{v(t) = x_k(kT < t < (k+1)T)} \tag{12.10}$$

with $P(x_k = 1) = 1/2$ and $P(x_k = 0) = 1/2$.

$v(t)$ is constant in each interval of duration T, with a value which is either 0 or 1 and with both values being equally probable. We write $v(t)$ as $v(t) = \frac{y(t)+1}{2}$, with $y(t)$ being

$$\mathrm{y(t) = y_k(kT < t < (k+1)T)} \tag{12.11}$$

with $P(y_k = 1) = 1/2$ and $P(y_k = -1) = 1/2$.

One can prove that the signal as described above is ergodic (and stationary), but this lies outside the scope of this course. Therefore, the statistical and time average of $y(t)$ are both zero.

To calculate the autocorrelation function $R_y(\tau) =< y(t)y(t+\tau) >$, we first consider the cases $\tau = 0$ and $\tau > T$. It is obvious that $R_y(0) =< y^2(t) >= 1$. For $\tau > T$, $y(t)$ and $y(t+\tau)$ always are in different bit periods and so $R_y(\tau) =< y_k y_l >$, with $l > k$. Since y_k and y_l are uncorrelated, it follows that $R_y(\tau) = 0$ for $\tau > T$. For the more difficult case of $0 < \tau < T$ we make use of the ergodicity and replace the statistical average by a temporal average. I.e. we calculate

$$R_y(\tau) = \lim_{n\to\infty} \frac{1}{2nT} \int_{-nT}^{nT} y(t)y(t+\tau)dt \tag{12.12}$$

for $0 < \tau < T$, with n the number of bit periods. This integral can be rewritten as a sum over the different bit periods

$$R_y(\tau) = \lim_{n\to\infty} \frac{1}{2nT} \sum_{k=-n}^{n-1} \int_{kT}^{(k+1)T} y(t)y(t+\tau)dt \tag{12.13}$$

and in each bit interval, a fraction $(T-\tau)$ will be overlapping with the same bit, in which case $y(t)y(t+\tau) = 1$ and a fraction τ will be overlapping with the next bit, in which case $y(t)y(t+\tau)$ will be zero when averaged over a large number of bits. As a result, we find:

$$R_y(\tau) = \frac{1}{T}(T-\tau) = 1 - \frac{\tau}{T} \tag{12.14}$$

for $0 < \tau < T$.

As $R_x(-\tau) = R_x(\tau)$, we find that the spectral density can now easily be calculated using the Wiener-Khintchine theorem, indeed

$$S_y(\omega) = \int_{-\infty}^{\infty} d\tau R_y(\tau)\exp(-j\omega\tau) = \int_{-T}^{T} d\tau \left(1 - \frac{|\tau|}{T}\right)\exp(-j\omega\tau) = T\frac{\sin^2(\omega\frac{T}{2})}{\left(\omega\frac{T}{2}\right)^2} \tag{12.15}$$

From this power spectrum it is easy to calculate the power spectrum of the function $v(t)$ as $S_v(\omega) = \frac{\partial(\omega)+S_y(\omega)}{4}$

The autocorrelation function of $v(t)$ and its spectral density $S_f(\omega)$ are plotted in figure 12.2(a).

12.3.2 return-to-zero amplitude shift keying

We consider a rectangular return-to-zero signal $\psi(t) = v(t)\sin(\omega_c t)$, with $v(t) = x_k$, for $kT < t < (k+d)T$ (with $0 < d < 1$) and $v(t) = 0$ for $(k+d)T < t < (k+1)T$. x_k, which takes on the values 1 and 0, has the same statistical properties as for the NRZ case.

In order to find the power spectrum, again one has to calculate the autocorrelation function. We leave it as an exercise to show that the autocorrelation function looks like the function depicted in figure 12.2(b).

This autocorrelation function can be analytically expressed as the sum of a non-periodic function R_{nper} and a periodic function. The non-periodic function can be written as

$$R_{nper}(\tau) = \frac{d}{4} - \frac{|\tau|}{4T}\text{for } |\tau| \leq dT \tag{12.16}$$

with $R_{nper}(\tau)$ equals zero elsewhere. The periodic function can be written as

$$R_{per}(\tau) = \sum_{n=-\infty}^{+\infty} R_{nper}(\tau - nT) = R_{nper}(\tau) * \sum_{n=-\infty}^{+\infty} \partial(\tau - nT) \tag{12.17}$$

The power spectrum corresponding with the non-periodic part is given by

$$S_v(\omega) = \frac{d^2T}{4}\left[\frac{\sin(\frac{\omega dT}{2})}{\frac{\omega dT}{2}}\right]^2 \tag{12.18}$$

The spectrum corresponding with the periodic part consists of a number of dirac functions at frequencies $f = n/T = nB$, with B the bit rate (with a weight given by the corresponding non-periodic function contribution to the power spectrum).

The power spectrum for a RZ signal has the same form as the spectrum of a NRZ signal, but occupies a much larger bandwidth (inversely proportional with $1/d$ and proportional with the bit rate B). The lines at multiples of the bit rate now allow to extract a clock with frequency equal to the bit rate. This was not the case for a NRZ signal, where it can be seen that for $f = 1/T = B$, the spectral density is zero.

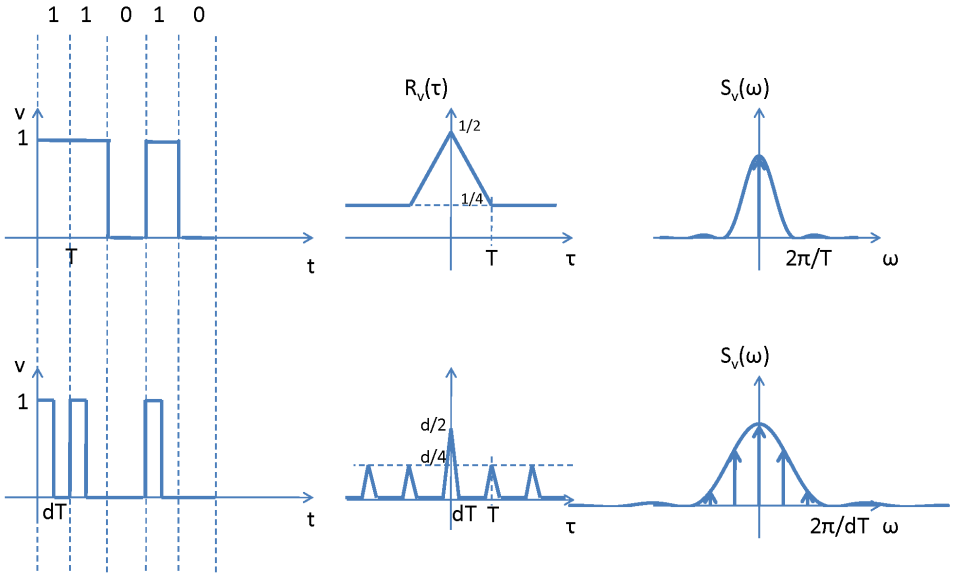

Figure 12.2: Power spectrum of a NRZ and RZ modulation

12.3.3 Phase shift keying

A binary phase shift keying (BPSK) signal is a signal with constant amplitude of which the phase changes between 0 and π for a logical '0', resp. '1'. It is thus mathematically equivalent to the signal $y(t)$ considered under section 1 and also has the same spectrum (i.e. the NRZ spectrum but without the DC component). A DBPSK signal is a signal with constant amplitude of which the phase changes with an amount π for every '1' and stays constant for every '0'. It also has the same spectrum as the signal $y(t)$ considered in the previous section.

12.4 Influence of noise in a digital communication channel

12.4.1 White Gaussian noise

The term noise refers to unwanted signals that are always present in a communication system. The presence of noise superimposed on a signal tends to obscure or mask the signal, which leads to errors in detection. There are many sources of noise, including thermal noise, relative intensity noise, phase noise, shot noise... which are differentiated by their origin and statistical properties. Often noise can be described by a zero-mean Gaussian random process. This implies that the noise term $n(t)$ is a random function, whose value n, at any arbitrary time is statistically characterized by the Gaussian probability density function $p(n)$:

$$p(n) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}\left(\frac{n}{\sigma}\right)^2) \tag{12.19}$$

where σ^2 is the variance of n.

Many types of noise can be considered to be a white process, meaning that the power spectral density is the same for all frequencies of interest. This is the case for thermal noise, and can be assumed for other types of noise, within the system bandwidth. The noise terms often can be simply added to the (noise-free) sent signal. This is often referred to as additive white Gaussian noise (AWGN).

For a white Gaussian noise process $n(t)$, one can write

$$R_n(\tau) = \sigma^2 \partial(\tau) = \frac{N_0}{2} \partial(\tau) \tag{12.20}$$

12.4.2 Sources of noise in an optical link

In an optical link, several sources of noise can be identified.

Thermal noise is the electronic noise generated by the thermal agitation of the charge carriers inside an electrical conductor at equilibrium. It is a white stochastic process, with a Gaussian distribution.

Shot noise consists of random fluctuations of the electric current due to the fact that the current is carried by discrete charges. This is also a white process, characterized by a Poisson distribution, which resembles a Gaussian distribution function for large carrier numbers.

The amplified spontaneous emission in lasers and optical amplifiers is another source of noise: electrons in the upper energy level can also decay by spontaneous emission, which occurs at random. Photons are emitted spontaneously in all directions, but a proportion of those will be emitted in a direction that falls within the aperture of the laser/amplifier waveguide and therefore add to the signal.

It speaks for itself that these contributions are the most harmful there where the signals are the weakest.

12.4.3 Detection of binary signals in Gaussian noise

In this section we will discuss the detection of binary signals in the presence of zero-mean additive white Gaussian noise (AWGN). The signal received by the receiver is represented by $r(t) = s_i(t) + n(t)$, where $s_i(t) = s_1(t)$ for a binary 1 and where $s_i(t) = s_2(t)$ for a binary zero, in the signalling interval T.

A first step in signal detection consists of reducing the received waveform $r(t)$ to a single number $z = s_i(t_s) + n(t_s)$, the sample value. In case there would be no noise, one should choose the sampling time such that z is either a_1 in the case of a binary 1 or a_2 in the case of a binary 0.

Therefore, one can write the conditional probability density functions, $p(z|s_1)$ and $p(z|s_2)$ as

$$\begin{aligned} p(z|s_1) &= \tfrac{1}{\sigma\sqrt{2\pi}} \exp(-\tfrac{1}{2}\left(\tfrac{z-a_1}{\sigma}\right)^2) \\ p(z|s_2) &= \tfrac{1}{\sigma\sqrt{2\pi}} \exp(-\tfrac{1}{2}\left(\tfrac{z-a_2}{\sigma}\right)^2) \end{aligned} \tag{12.21}$$

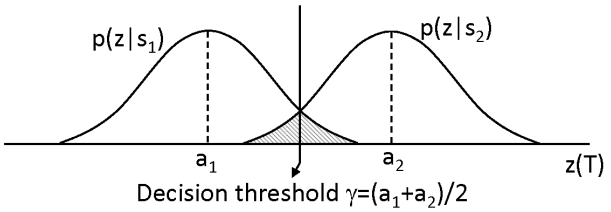

Figure 12.3: Detection of binary signals in Gaussian noise: origin of bit errors

This is schematically illustrated in figure 12.3. The second step of the signal detection process consists of comparing the test sample z, to a threshold level $\gamma = \frac{a_1+a_2}{2}$ (in the case that s_1 and s_2 are equally probable).

For this binary signal example, there are two ways in which errors can occur. An error e will occur when $s_1(t)$ was sent, and the channel noise results in a receiver output signal z, which is lower than the threshold level γ. The probability of such an occurrence is

$$P(e|s_1) = \int_{-\infty}^{\gamma} p(z|s_1)dz \tag{12.22}$$

A similar equation can be deduced for the case where $s_2(t)$ is sent and an error occurs

$$P(e|s_2) = \int_{\gamma}^{+\infty} p(z|s_2)dz \tag{12.23}$$

If both signals are equally probable, we can write the probability of a bit error as

$$P_B = \frac{1}{2}P(e|s_1) + \frac{1}{2}P(e|s_2) = \int_{\gamma}^{+\infty} p(z|s_2)dz = \int_{-\infty}^{\gamma} p(z|s_1)dz \tag{12.24}$$

or

$$P_B = \int_{\gamma}^{+\infty} \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{z-a_2}{\sigma})^2)dz = Q(\frac{a_2-a_1}{2\sigma}) \tag{12.25}$$

with $Q(x)$, the complementary error function

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{+\infty} \exp(-\frac{u^2}{2})du \tag{12.26}$$

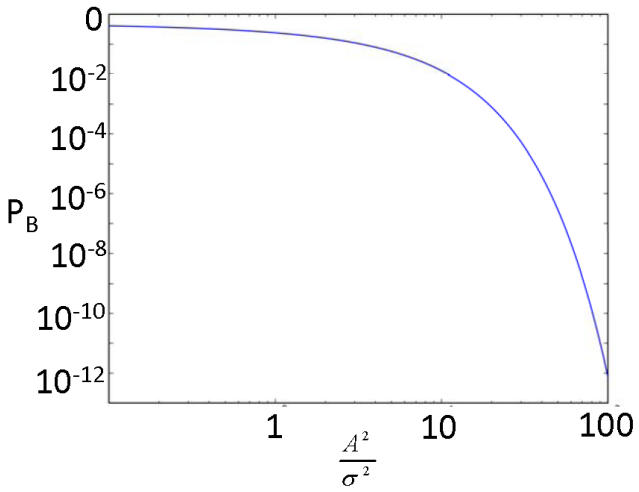

Figure 12.4: Complementary error function: bit error probability versus signal to noise ratio

Figure 12.5: Application of a bandpass filter prior to sampling to improve the signal to noise ratio

It is clear that the real signal in this case is the difference $A = a_2 - a_1$. This complementary error function is plotted in figure 12.4, as a function of $\frac{A^2}{\sigma^2}$, with $A = a_2 - a_1$. Practical systems require bit error rates in the order of 10^{-9} or better. This puts an upper limit on the amount of noise that can be tolerated in a system (although redundancy can be incorporated in the data signal to be able to detect and correct bit errors in a bit stream, e.g. forward error correction codes or FEC). The ratio $\frac{A^2}{\sigma^2}$ is therefore referred to as the signal to noise ratio (SNR).

Although this analysis was performed on the most simple digital modulation format, it can be extended to more complex formats. The addition of noise to a communication system is represented in the constellation diagram by a blurring of the individual constellation points. Also in these cases a decision circuit has to decide which constellation point is the closest to the sample. For all of these modulation formats an assessment of bit error probability can be performed, resulting in slightly differing "waterfall"-like curves as the one shown in figure 12.4.

In order to improve the signal to noise ratio, a bandpass filter can be applied before sampling the signal, as shown in figure 12.5. This bandpass filter changes the power spectrum of the noise component as

$$\begin{aligned} R_{n'}(u) &= \frac{N_0}{2} \int_{-\infty}^{+\infty} h(t)\, h(t+u)\, dt \\ S_{n'}(f) &= \frac{N_0}{2} |H(f)|^2 \\ P_{n'} = R_{n'}(0) &= \frac{N_0}{2} \int_{-\infty}^{+\infty} h^2(t)\, dt = \frac{N_0}{2} \int_{-\infty}^{+\infty} |H(f)|^2\, df \end{aligned} \tag{12.27}$$

This change in noise spectrum is also referred to as colored noise. Using an ideal bandpass filter with a bandwidth W and assuming white noise at the input this results in

$$P_{n'} = R_{n'}(0) = \frac{N_0}{2} \int_{-\infty}^{+\infty} h^2(t)\, dt = \frac{N_0}{2} \int_{-\infty}^{+\infty} |H(f)|^2\, df = N_0 W \tag{12.28}$$

The minimum bandwidth of the bandpass filter is limited by the bandwidth if the signal $s(t)$. This is determined by the spectral efficiency of the modulation format, being the number of bits that can be transmitted per second, per Hertz of the signal power spectrum.

Part V

Lasers and Optoelectronic Components

Chapter 13

Lasers

Contents

13.1 Gain medium

The most important coherent optical source is the *LASER* or "Light Amplification by Stimulated Emission of Radiation". It reveals the process of stimulated emission upon which the light amplification process is based. A laser as we know it, refers to an oscillator, in which amplification is obtained due to stimulated emission. Although Albert Einstein proved the existence of stimulated emission already in 1917, it took until 1960 to show laser oscillation at optical frequencies. Theodore Maiman made the first laser operational on 16 May 1960 at the Hughes Research Laboratory in California, by shining a high-power flash lamp on a ruby rod with silver-coated surfaces. The succeeding years saw all kinds of laser types being developed. It would take some time however for lasers to be used in a broad range of applications. The laser was described as "a solution looking for a problem".

Nowadays, the number of applications for lasers is growing fast, and there is no reason to believe that this trend will slow down. The main application fields are material processing, medical treatments, optical recording (e.g. compact disk), optical fiber communication, metrology (e.g. distance measurements), barcode readers, holography, laser induced fabrication techniques, architectural lighting, etc. In almost all of these applications, lasers are used for their high optical power and coherence. This combination makes it possible to focus laser light to an extremely small and intense spot.

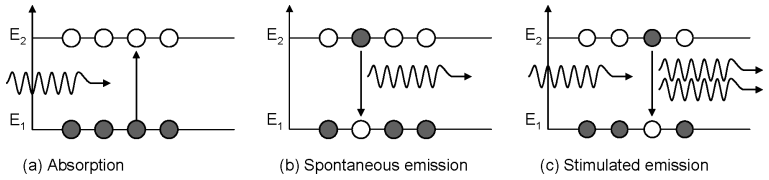

Figure 13.1: Absorption, spontaneous emission and stimulated emission.

13.1.1 Emission and absorption

We discussed the interaction of photons and atoms exhaustively in chapter 10. To refresh our minds, let us quote again the possible interaction mechanisms with the respective rate equations (see figure 13.1)

1. Absorption.

   The excitation of an atom to a higher energy level due to absorption of a photon.

$$\frac{dN_2}{dt} = -\frac{dN_1}{dt} = P_{abs} N_1 = B_{12} \rho\left(\nu_0\right) N_1 \tag{13.1}$$

2. Spontaneous emission.

   The relaxation of an atom to a lower energy level with emission of a photon.

$$\frac{dN_2}{dt} = -\frac{dN_1}{dt} = -P_{sp} N_2 = -A_{21} N_2 \tag{13.2}$$

3. Stimulated emission.

   The relaxation of an atom to a lower energy level with emission of a photon having a phase, frequency and polarization equal to that of the incident photon causing the relaxation.

$$\frac{dN_2}{dt} == -\frac{dN_1}{dt} = -P_{st} N_2 = -B_{21} \rho\left(\nu_0\right) N_2 \tag{13.3}$$

With B_{12}, A_{21} en B_{21} the Einstein coefficients given by:

$$B_{12} = B_{21} \tag{13.4}$$

$$A_{21} = B_{21} \frac{8\pi h \nu^3}{c^3} \tag{13.5}$$

The Einstein coefficients can be derived from equations (10.32) and (10.35).

13.1.2 Population inversion

First, we examine the absorption or amplification of a monochromatic radiation field in case of an arbitrary occupation of the two energy levels. To this extent, we consider a cylindrical volume with unit surface area and thickness dx (figure 13.2).

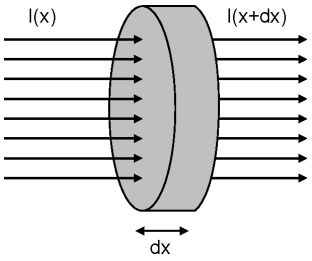

Figure 13.2: Amplification of a monochromatic field propagating through an amplifying substance.

The incident field is impinging the cylinder perpendicularly with an intensity given by:

$$I = \rho(\nu_0)\frac{c}{n} = N_f h\nu_0 \frac{c}{n} \tag{13.6}$$

with c the speed of light, n the refractive index, and N_f the number of photons per unit volume.

While the light propagates through the material, its intensity is altered due to absorption and stimulated emission of photons. Per unit time and unit volume, $N_1\rho(\nu_0)B_{12}$ photons are absorbed, and $N_2\rho(\nu_0)B_{21}$ photons are produced by stimulated emission, if the occupation of the lowest and the occupation of the highest energy level is respectively represented by N_1 and N_2 (with a degeneration factor of 1). This is translated in a change of intensity after propagating over a distance dx:

$$I + dI = I + (N_2 - N_1)\rho(\nu_0)B_{21}dx\ h\nu_0 \tag{13.7}$$

or

$$\begin{aligned}\frac{dI}{dx} &= (N_2 - N_1)\rho(\nu_0)B_{21}h\nu_0 \\ &= \frac{h\nu_0 n}{c}(N_2 - N_1)I\ B_{21}\end{aligned} \tag{13.8}$$

Integration over x gives

$$I = I_0 e^{gx} \tag{13.9}$$

with

$$g = (N_2 - N_1)\frac{h\nu_0 n}{c}B_{21} \tag{13.10}$$

g is the relative power increase per unit distance, expressed in 1/m (or 1/cm) and is called the 'gain'. The net gain is positive only if the occupation of the highest energy level is larger than the occupation of the lowest energy level. This is called population inversion. The gain is zero if both occupations are equal. It is as if the material is transparent. The intensity of the light is constant while propagating, although absorption and emission still occur.

13.1.3 Pump systems

How can we 'pump' a system to population inversion? At first sight, it seems enough to let a 2-level system (a system with 2 energy levels relevant for laser action) absorb a flood of photons in order to bring enough atoms to a higher energy level.

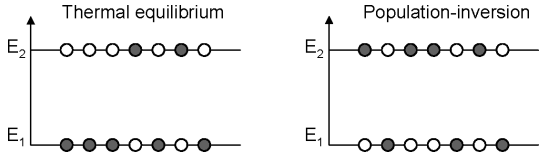

Figure 13.3: Thermal equilibrium versus population inversion.

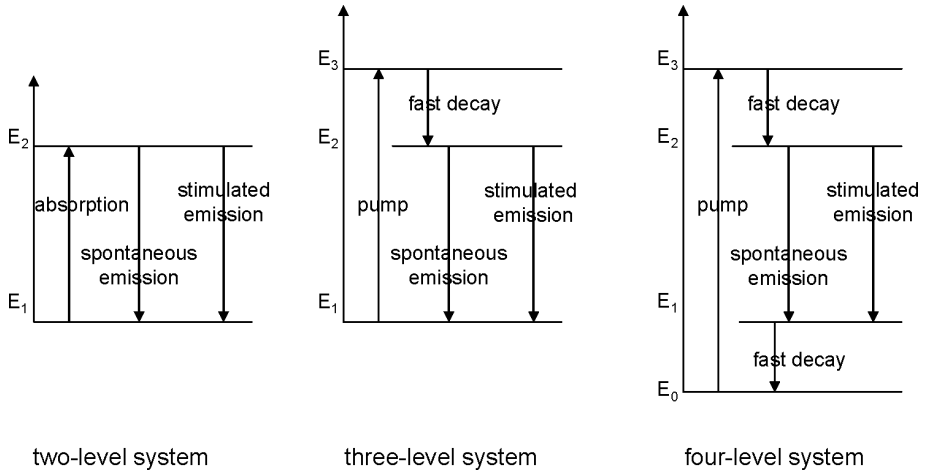

Figure 13.4: Establishing population inversion in a two-, three- and four-level system.

This is however not true in a static regime. From the moment population inversion would occur, further absorption is obstructed, and, if excited further, the occupation of the higher level would decrease due to the enhanced stimulated emission. In a 2-level system, population-inversion can not be attained. Considering a net upwards flux equal to the net downwards flux in a static regime, we have:

$$(N_1 - N_2)\,\rho\,(\nu_0)\,B_{12} - A_{21}N_2 = 0 \tag{13.11}$$

and thus

$$N_2 = \frac{N_1}{1 + \frac{A_{21}}{B_{21}N_f h\nu_0}} \tag{13.12}$$

It is clear that, by an ever increasing photon density, population inversion can be arbitrarily well approximated but never established.

However, population inversion can be established in systems with more than two energy levels, as shown in figure 13.4.

In a 3-level system, atoms are pumped to the third energy level. This can be realized by incident photons having an energy corresponding to this energy difference. Atoms at this third energy level can relax spontaneously to the second energy level. Relaxation from this second energy level to the base energy level corresponds with the laser transition. The needed population inversion for stimulated emission can only be built up if the lifetime of the atoms at level 3 (average time spent at this level) is much smaller than at level 2. If this is the case, energy level 3 will only be weakly occupied. Excitation of atoms from level 1 to level 3 by absorption of the incident photons is then unimpeded. Level 2 can be quite crowded, leading to population inversion. As the flux of atoms excited from level 1 to level 3 is proportional to the occupation degree of level 1 and as the

occupation of level 2 has to be even higher than this level 1 occupation (population inversion), it is clear that the needed pump power to reach population inversion will be pretty high.

For that reason, it is more convenient to work with a 4-level system. Laser action is established between energy levels 1 and 2, while the outer levels - level 0 and 3 - are used by the pump system. In that way, both processes are decoupled. The ideal situation is reached when relaxations from level 3 to level 2 and from level 1 to level 0 are fast in comparison with the lifetime of level 2. Level 3 is occupied weakly, a guarantee for an efficient pump process. In the meantime, level 2 is more easily filled than level 1. This means that with a relatively low pump power, the transition between level 1 and level 2 can be inverted. However, the 4-level system demands a high energy to pump the outer transition. The difference between the pump transition energy and the laser transition energy is lost irrevocably.

How the pump system works, depends on the type of laser. Optical excitation, gas ionization, electron bombardments, release of chemical energy, etc. can all be used to pump the laser. A semiconductor laser is pumped by a current injection through its junction. In this case, the energy levels are no longer discrete, but the carriers are distributed over the energy bands.

13.1.4 Homogeneous and inhomogeneous broadening

As described above, laser transitions in many systems occur from and to discrete energy levels. In this case, the spectral dependence of the absorption and gain curves is expected to be sharply defined as well, or in other words, only photons with a specific frequency ν_0 are absorbed or emitted. However, in reality, several phenomena result in a broadening of the linewidth. This makes it possible that photons with slightly deviating frequency $\nu_0 \pm \delta\nu$ are absorbed or emitted as well. The line shape function $g(\nu_0)$ of the atomic transition is no longer a discrete peak but shows a certain width. Two types of mechanisms are distinguished: homogeneous and inhomogeneous broadening.

Homogeneous broadening

Homogeneous broadening is an increase of the linewidth of an atomic transition caused by effects that equally affect different radiating or absorbing atoms. The lineshape of the individual atoms and the lineshape of total emission and absorption spectrum are identical. The atoms are indistinguishable.

- Natural broadening

  Natural broadening finds its origin in the finite lifetime of the atoms at the higher energy level. Heisenberg's uncertainty principle dictates:

  $$\delta(E_2 - E_1).\tau = \frac{h}{2\pi}, \tag{13.13}$$

  or, expressed in terms of frequency

  $$\delta\nu = \frac{1}{2\pi\tau}. \tag{13.14}$$

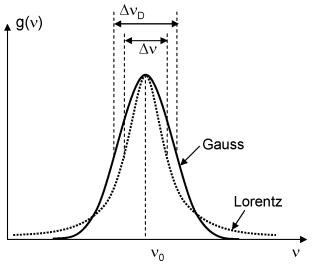

Figure 13.5: Gaussian versus Lorentzian line shape.

The electron is only for a short time in an excited state, so its energy cannot have a precise value. Since energy levels are 'fuzzy', atoms can absorb photons with slightly different energy, with the probability of absorption declining as the difference in the photon's energy from the 'true' energy of the transition increases.

- Collisional Broadening

  The energy levels of an atom are perturbed by collisions or close encounters with other atoms or ions. When molecules collide with each other or with phonons (crystal lattice vibrations), the gain and absorption curve is broadened. The broadening is enhanced with increasing temperature and pressure (in case of a gas). The gain curve is a Lorentzian lineshape (figure 13.5a):

$$g(\nu) = \frac{\Delta\nu}{2\pi\left[(\nu-\nu_0)^2 + \left(\frac{\Delta\nu}{2}\right)^2\right]} \tag{13.15}$$

  with ν_0 the central frequency and $\Delta\nu$ the 3dB bandwidth of the broadened lineshape.

Inhomogeneous broadening

Inhomogeneous broadening is an increase of the linewidth of an atomic transition caused by effects that act differently on different radiating or absorbing atoms. Inhomogeneous broadening spreads the resonance frequency ν of the individual atoms over the frequency interval $[\nu_0 - \delta\nu, \nu_0 + \delta\nu]$. This can be caused by e.g. the different velocities of the atoms of a gas or by different lattice locations of atoms in a solid medium. Light with a specific frequency will interact with a group of atoms, while light with a slightly different frequency will interact with another group of atoms. This mechanism spreads the line shape of the system as a whole without broadening the line shape of every single atom.

For example, elastic strain (at microscopic level) and defects in crystal structures result in a different local environment for the individual atoms. This influences the energy levels of the atoms and leads to inhomogeneous broadening. In a semiconductor, for example, electrons and holes are spread over energy bands, instead of linked to discrete energy levels. This can be considered as inhomogeneous broadening of one energy level. Another important example of inhomogeneous broadening is Doppler-broadening in gas lasers. Thermal agitation is the random movement of atoms. When a photon interacts with an atom that propagates in the same direction as the photon,

the atom will experience the light with a slightly different frequency due to the Doppler effect. Although all atoms show the same energy levels and transitions, there will be a certain broadening concerning interaction with light. The 3dB linewidth associated with this effect is given by:

$$\Delta\nu_{doppler} = 2\nu_0\sqrt{\frac{2kT}{Mc^2}\ln 2}, \tag{13.16}$$

with M the atom mass. Doppler broadening is most significant for light atoms at high temperatures. At room temperature for a He-Ne laser, Doppler broadening is about 1.5 GHz. Inhomogeneous broadening results in a Gaussian gain/absorption function (figure 13.5):

$$g(\nu) = \frac{2\sqrt{\ln 2}}{\sqrt{\pi}\Delta\nu_D} e^{-\left[4\ln 2\left(\frac{\nu-\nu_0}{\Delta\nu_D}\right)^2\right]}. \tag{13.17}$$

Remark

It is not always clear how to distinguish inhomogeneous and homogeneous broadening. For example, Doppler broadening is considered as a homogeneous broadening when the average time that an atom moves in a certain direction with a certain velocity is small with respect to the lifetime of the excited level. Electrons and holes will be able to relax equally rapidly in the conduction and valence band respectively, resulting in homogeneous broadening.

13.1.5 Gain saturation

If the frequency of the light incident on an inverted medium approaches the optical transition level, we expect gain. When the intensity of the incident light increases, the amount of downwards transitions will increase as well. The extent of population inversion will decrease, as well as the gain. Dependent on the extent of pumping, the gain will reside between a minimal value in the order of transparency and a maximal value in case of small optical intensities. This is called gain saturation, because, if the material is used as an optical amplifier, the intensity of the outgoing light will saturate as a function of the incident power (figure 13.6).

For the spectral gain function, saturation acts differently for homogeneous and inhomogeneous broadening of the material. In case of homogeneous broadening, all atoms are considered identical, resulting in a decrease of the spectral gain function as a whole, when the level of population inversion decreases (figure 13.6c).

In case of inhomogeneous broadening, the atoms themselves show a certain energy bandwidth. This implies that the incident light will only interact with those atoms showing a corresponding transition energy. As a result, the spectral gain function will show a local dip for the frequency of the incident light (figure 13.6d). This is called 'hole burning', as if the dip is burned into the spectral gain function. Some materials (for example semiconductors) show homogeneous broadening at low optical intensities. Replenishing of the levels is fast in comparison with the lifetime of stimulated emission. The latter will decrease with increasing optical intensities and inhomogeneous broadening then dominates, resulting in a dip in the spectral gain function at the optical frequency.

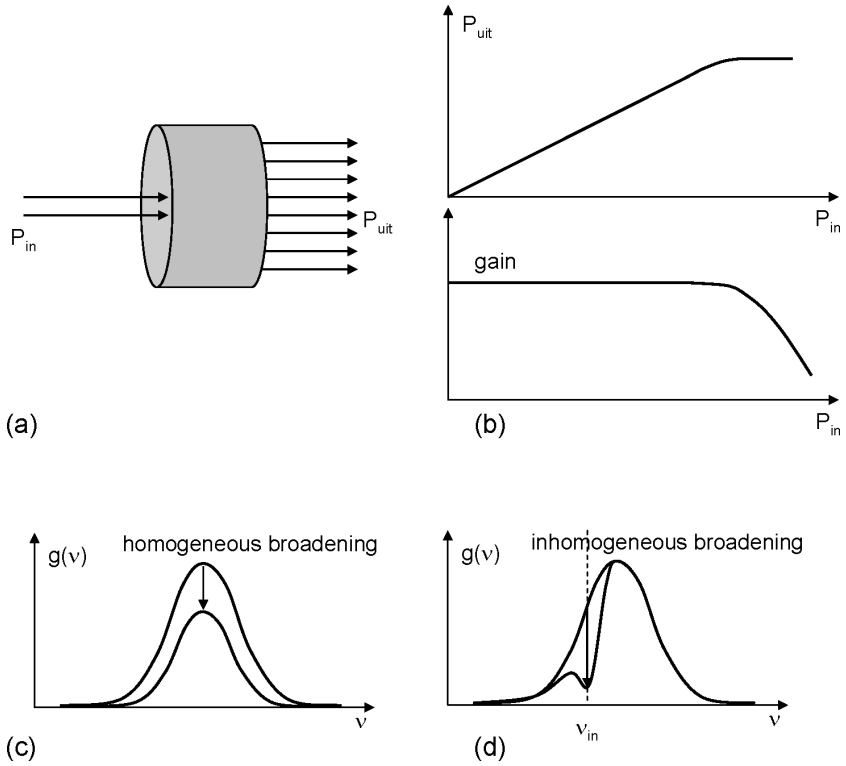

Figure 13.6: Gain saturation.

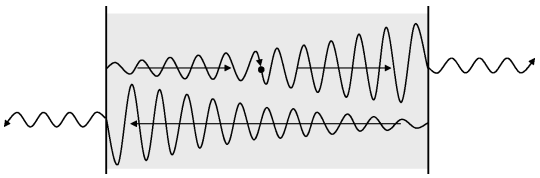

Figure 13.7: A cavity with amplification.

13.2 Laser cavities

13.2.1 Introduction

The previous section explains how to obtain amplification of light in a material. To realize lasing, we need to place this amplifying material in a resonator. A resonator consists of a cavity with two fully or partially reflecting ends. The amplifying material together with the reflecting ends form the necessary conditions for oscillation: amplification and feedback. Laser oscillators, or lasers, often show a lateral dimension larger than the transversal dimensions. This makes it possible to analyze the oscillation mechanism in a simple manner.

In this section we discuss the most elementary analyzing methods for laser cavities. We explain the principles and main concepts of modes for these lasers. To obtain oscillation, the light propagating in the cavity has to satisfy a condition for resonance. The resonance condition implies that the phase and amplitude of the field at a certain position in the laser remains the same after one

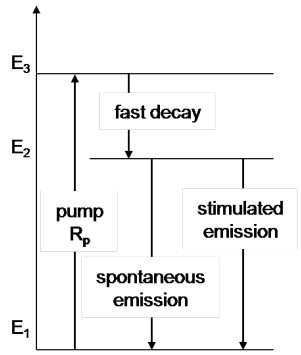

Figure 13.8: A three-level system.

round trip in the laser (figure 13.7), or in other words, the loop gain equals unity. This resonance condition is intuitively interpreted when considering the process to start the laser activity. As long as the laser is weakly pumped, there is barely any amplification. On the contrary, spontaneously emitted light will be present. When pumping gets stronger, light over a finite wavelength range is amplified. Part of the spontaneously emitted light will be amplified. Some frequencies will interfere constructively while propagating in the laser cavity, while others will interfere destructively. The amplitude of the constructively interfering light increases. This increase of amplitude will continue until a balance exists between the rate of pumping electrons to a higher energy level and the relaxation of electrons to lower energy levels due to stimulated emission pro rata of the intensity of the light in the resonator. At that time, a stable regime is created in which the loop gain is equal to one.

13.2.2 Resonance: Rate equations analysis

The simplest way to describe the oscillation mechanism is to use the rate equations. These equations describe the dynamics of the average amount of particles per unit volume in the cavity. These particles can be electrons, atoms or photons. They do not tell us anything about the phase or the frequency of the light to fulfil the resonance condition. They do tell us however the conditions for a power balance in a quantum mechanical way.

Using the notations of the previous section, a simple set of 'rate equations' for a three-level system (level 1, 2 and 3, see figure 13.8) looks like:

$$\begin{aligned}
\frac{dN_2}{dt} &= R_p - A_{21}N_2 - N_f\ h\nu\ B_{21}\left(N_2 - N_1\right) \\
\frac{dN_1}{dt} &= A_{21}N_2 + N_f\ h\nu\ B_{21}\left(N_2 - N_1\right) - R_p \\
\frac{dN_f}{dt} &= N_f\ h\nu\ B_{21}\left(N_2 - N_1\right) + \beta A_{21}N_2 - \frac{N_f}{\tau_p}
\end{aligned} \tag{13.18}$$

The first two equations describe the amount of particles at level 1 and level 2 as a function of time. The third equation describes the amount of photons N_f as a function of time. R_p is the pump rate, i.e. the amount of particles per unit of time pumped from level 1 to level 2 via level 3. β represents the fraction of spontaneously emitted photons that is coupled with the laser oscillation. The loss

term N_f/τ_p represents the amount of photons that leaves the laser cavity per unit of time (due to transmission losses at the end facets, scattering, absorption, etc.). τ_p can be considered as the photon lifetime, i.e. the average time a photon spends in the laser cavity.

A numerical analysis of this set of nonlinear equations is rather simple. For the static regime, the derivations w.r.t. time are set to zero. In case of a three-level system, the first two equations are set equal. One equation must be added to solve this set of two equations and three unknown parameters. The additional equation describes the total atom concentration in the system:

$$N_1 + N_2 = N. \tag{13.19}$$

Equations (13.18) are rewritten as:

$$\begin{aligned} \frac{dN_2}{dt} &= R_p - A_{21}N_2 - cgN_f \\ \frac{dN_f}{dt} &= cgN_f - \frac{N_f}{\tau_p} + \beta A_{21}N_2 \\ g &= B_{21}\frac{h\nu}{c}\left(2N_2 - N\right) \end{aligned} \tag{13.20}$$

For simplicity, we assumed a refractive index n equal to one. If this is not true, one needs to change c by c/n. Neglecting the rather small term $\beta A_{21}N_2$, the net relative photon amplification per unit of time is given by $cg - 1/\tau_p$. Dividing this term by the speed of light c, the net relative photon amplification per unit of distance is obtained. The loop gain in a laser with a length L is thus given by:

$$\text{loop gain} = e^{\left(g - \frac{1}{c\tau_p}\right)2L} \tag{13.21}$$

As long as the gain g is smaller than $1/c\tau_p$, the loop gain will be smaller than 1 and laser action is impossible. Solving equations (13.18) or (13.20) gives us N_2, N_1 and N_f as a function of the pump rate R. This relation is shown in figure 13.9a. As long as the pump rate is low, population inversion and thus light amplification do not occur. The small amount of light that escapes from the cavity is spontaneously emitted light. Increasing the pump rate accomplishes population inversion: the laser material becomes transparent. This however is not sufficient for resonance due to the losses in the cavity (expressed by a finite τ_p). It is needed to pump more for the laser to reach the oscillation threshold. Then the material gain compensates for cavity losses. The loop gain is one.

Increasing the pump rate even higher, the loop gain must remain one to preserve the static regime. *Therefore, the gain g must be clamped to a fixed value at and above treshold.* This implies that the values of N_2 and N_1 need to be clamped on the value they have at the oscillation threshold. This is possible if the photon density N_f increases as well. Stimulated emission will increase also, compensating the increased pump rate. The analytical solution of the static equations for N_1 and N_2 is very simple, if we assume that N_f is zero for pump levels lower than or equal to the threshold value. Above threshold we assume that spontaneous emission is small in comparison with stimulated emission. Spontaneous emission can thus be neglected above threshold. We get:

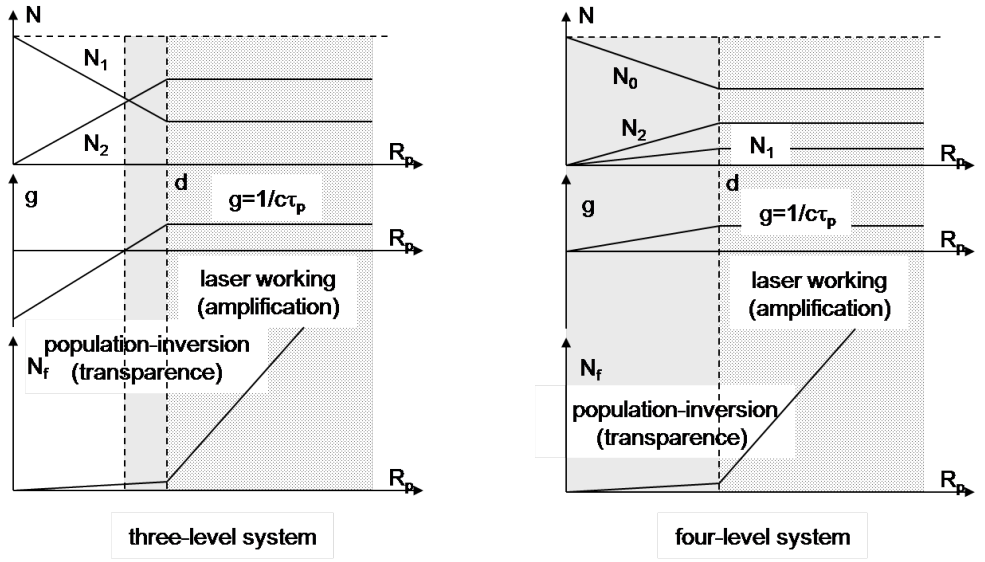

Figure 13.9: Occupation of the respective energy levels as a function of the pump rate for a three- and a four-level system.

- Below threshold:

$$N_2 - N_1 < \frac{1}{h\nu\ B_{21}\ \tau_p} \tag{13.22}$$

$$g < \frac{1}{c\tau_p} \tag{13.23}$$

$$N_2 = \frac{R_p}{A_{21}} \tag{13.24}$$

$$N_1 = N - N_2 \tag{13.25}$$

$$N_f = 0 \tag{13.26}$$

- At threshold:

$$(N_2 - N_1)_d = \frac{1}{h\nu B_{21}\tau_p}$$

$$\Longrightarrow N_{2d} = \frac{N + N_2 - N_1}{2} = \frac{1}{2}\left(N + \frac{1}{h\nu B_{21}\ \tau_p}\right) \tag{13.27}$$

$$g = \frac{1}{c\tau_p} \tag{13.28}$$

$$R_d = N_{2d}A_{21} = \frac{A_{21}}{2}\left(N + \frac{1}{h\nu\ B_{21}\ \tau_p}\right) \tag{13.29}$$

$$N_{1d} = N - N_{2d} \tag{13.30}$$

$$N_f = 0 \tag{13.31}$$

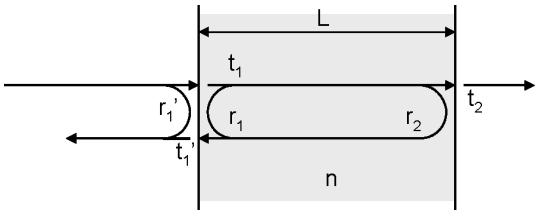

Figure 13.10: A plane Fabry-Perot cavity.

- Above threshold:

$$
\begin{aligned}
N_2 - N_1 &= (N_2 - N_1)_d && (13.32)\\
g &= \frac{1}{c\tau_p} && (13.33)\\
N_2 &= N_{2d} && (13.34)\\
N_1 &= N_{1d} && (13.35)\\
N_f &= \frac{R_p - R_d}{h\nu\, B_{21}\,(N_2 - N_1)_d} = \tau_p\,(R_p - R_d) && (13.36)
\end{aligned}
$$

These last equations tell us that the optical power increases linearly as a function of the pump rate. For a 4-level system, similar equations are found showing conceptually the same principle. This is presented in figure 13.9b.

13.2.3 Resonance: analysis with plane waves

The frequency needed to fulfill the phase resonance condition in the cavity can not be deduced using the rate equations. To calculate the frequency, it is most easy to consider a simple one-dimensional analysis of the cavity. The cavity is assumed to be transversally invariant. This allows us to treat the waves in the cavity as plane waves.

Let us consider the optical transmission of the structure shown in figure 13.10. The structure consists of two parallel semitransparent mirrors at a distance L from each other. In chapter 6 we called such a device a Fabry-Perot etalon or a Fabry-Perot interferometer. For example, a glass substrate covered on both sides with a thin metal layer, semitransparent and semireflective, can be a practical implementation of such an etalon. The transmission coefficient t for the electromagnetic field can be calculated as the sum of the contributions of successive reflections (see (4.68), interference of an infinite number of waves with progressively decreasing amplitude but identical phase shift). If the structure does not show any losses, this results in:

$$
\begin{aligned}
t &= t_1 t_2 e^{-j\phi} + t_1 t_2 r_1 r_2 e^{-j3\phi} + t_1 t_2 \,(r_1 r_2)^2\, e^{-j5\phi} + \ldots\\
&= \frac{t_1 t_2 e^{-j\phi}}{1 - r_1 r_2 e^{-j2\phi}} && (13.37)
\end{aligned}
$$

with

$$
\phi = k_0 n L. \qquad (13.38)
$$

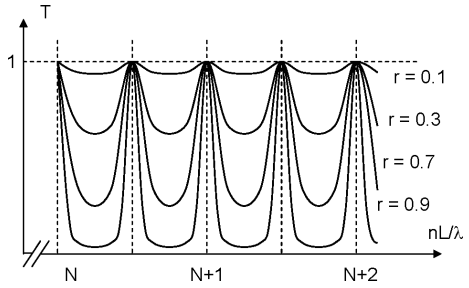

Figure 13.11: The transmission spectrum of the Fabry-Perot cavity from figure 13.10.

If r_1 and r_2 are real, the power transmission coefficient is given by:

$$T = |t|^2 = \frac{\frac{|t_1 t_2|^2}{(1-r_1 r_2)^2}}{1 + \frac{4 r_1 r_2}{(1-r_1 r_2)^2} \sin^2 \phi} = \frac{T_{\max}}{1 + F \sin^2 \phi} \tag{13.39}$$

with

$$T_{\max} = \frac{|t_1 t_2|^2}{(1 - r_1 r_2)^2} \text{ and } F = \frac{4 r_1 r_2}{(1 - r_1 r_2)^2} \tag{13.40}$$

If the structure is symmetric ($r_1 = r_2$), and the mirrors are lossless ($t_1 t_2 = (1 + r_1')(1 + r_2) = (1 - r_1)(1 + r_2)$), then T_{max} will be equal to one. This is consistent with equation (6.95) in chapter 6. The spectral transmission is shown in figure 13.11. The periodical maxima get sharper as $r_1 r_2 = r^2$ approximates 1. These maxima appear when

$$2\phi = 2m\pi, \text{ with } m \text{ integer} \tag{13.41}$$

or

$$L = m \frac{\lambda}{2n}. \tag{13.42}$$

In other words, the length of the etalon needs to be a whole number of half the wavelength of the light. The number m is in general quite large (10^3 to 10^7). It is easy to show that the spectral period or spectral distance between two adjacent peaks is given by:

$$\Delta\lambda = \frac{\lambda^2}{2nL}, \tag{13.43}$$

or

$$\Delta\nu = \frac{c}{2nL}, \tag{13.44}$$

and thus

$$\frac{1}{\Delta\nu} = \frac{2nL}{c}. \tag{13.45}$$

In words: the (temporal) period of the mode spacing equals the round trip time of the cavity. The finesse $\mathfrak{F}$ of the Fabry-Perot cavity (see chapter 4) is defined as the distance between successive maxima, divided by the 3dB width of a maximum. The larger the reflection coefficients are, the larger the finesse $\mathfrak{F}$.

A quality factor Q can be defined as well. The quality factor of an oscillator is the amount of radials an oscillator covers before its energy has decreased with a factor of $1/e$. Translating this to

a Fabry-Perot cavity, the Q factor is defined as 2π times the number of 'round trips' made by the light in the cavity before its intensity has decreased with a factor of $1/e$. Considering the material as being lossless, we can calculate Q:

$$\begin{aligned} \left(r_1^2 r_2^2\right)^{Q/2\pi} &= \frac{1}{e} \\ Q &= \frac{2\pi}{\ln\left(1/r_1^2 r_2^2\right)} \end{aligned} \tag{13.46}$$

It is clear that Q will be smaller when the cavity has higher (transmission) losses. The quality factor is a measure for the ratio of energy stored in the oscillator (or cavity) to the energy leaving the cavity periodically.

Let us consider the same structure made of an amplifying material. If g is the gain per length unit, the increase in amplitude of the optical field after one round trip is given by:

$$\sqrt{\exp\left(g2L\right)} = \exp\left(gL\right) \tag{13.47}$$

Equation (13.37) is then reformulated as:

$$t = \frac{t_1 t_2 \exp\left(gL/2\right) \exp\left(-j\phi\right)}{1 - r_1 r_2 \exp\left(gL\right) \exp\left(-2j\phi\right)} \tag{13.48}$$

Transmission will be infinitely large if:

$$r_1 r_2 \exp\left(gL\right) \exp\left(-2j\phi\right) = 1 \tag{13.49}$$

Or: for an input power equal to zero, the structure can generate a finite power. This is exactly what is meant with the resonance condition. The gain g in equation (13.49) depends on the pump rate R and on the wavelength λ. (13.49) is as such a complex equation of two real unknown quantities R and λ. This complex equation can be split up in an equation for the intensity (loop gain) and an equation for the phase shift:

$$r_1^2 r_2^2 \exp\left(2g\left(R,\lambda\right)L\right) = 1, \tag{13.50}$$

$$\frac{2\pi}{\lambda} nL = m\pi. \tag{13.51}$$

Let us now consider the spectral loop gain in figure 13.12. Besides the spectral loop gain, figure 13.12 also shows the frequencies for which the phase resonance condition is fulfilled. Two situations can occur: the width of the spectral loop gain can be either small or broad compared to the distance between the phase resonance frequencies. In the first case, the laser cavity will emit light at a single frequency. In most practical cases however, the width of the spectral loop gain is broad in comparison with the distance between the spectral resonance peaks. In these situations, the laser is able to emit at different phase resonance frequencies. The laser is said to show several axial or longitudinal modes.

The loop gain for the respective phase resonance frequencies is slightly different however. If the material is broadened homogeneously, only one mode can have a loop gain equal to one. The loop gain for the other modes will be slightly smaller than one. They will oscillate in the cavity due to the contributions of spontaneous emission, but their intensity will be smaller than the intensity

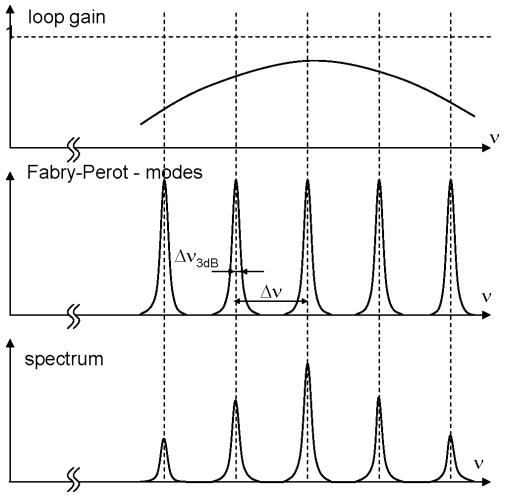

Figure 13.12: Determination of the laser spectrum.

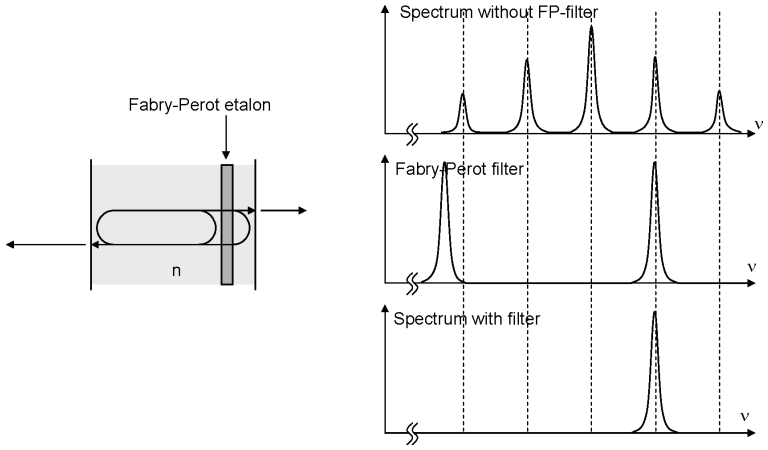

Figure 13.13: A laser cavity with a Fabry-Perot etalon as mode filter.

of the principal mode. In case of homogeneous broadening, all modes will 'eat' from the same reservoir of particles at the higher energy level.

As an example, consider the $He-Ne$ laser emitting at 633nm. The cavity of this laser is typically 30 cm long, resulting in a mode spacing of 500 MHz or 0.0007 nm. As mentioned above, the spectral width of the loop gain is about 1.5GHz due to Doppler broadening. Thus, the spectrum of a He-Ne laser will show several longitudinal side modes. Another example is the $GaAs$ semiconductor laser. The cavity is typically small, i.e. on the order of 0.3 mm. This results in a mode spacing of 140GHz or 0.4 nm. In spite of this very large mode spacing, we will find several longitudinal modes in the emitted spectrum of the laser due to the very broad loop gain (typically 50 nm).

If we want a laser with several longitudinal side modes to lase in a single mode, we can use a passive Fabry-Perot etalon in the cavity (figure 13.13). This structure of two parallel mirrors shows a frequency selective transmission profile. In this way the spectral loop gain is forced to show a sharper maximum, suppressing the unwanted longitudinal side modes.

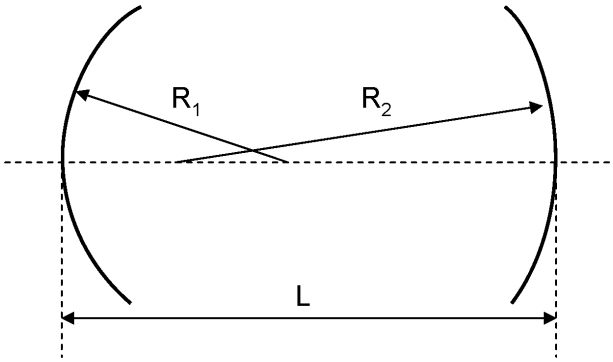

Figure 13.14: The cavity with spherical mirrors.

13.2.4 Resonance: beam theory analysis

In the previous paragraph, we assumed a transversally invariant cavity. This allows an analysis of the cavity based on plane waves. In practical situations, this is not true. A real cavity has finite transverse dimensions.

Assume we work with a cavity using plane mirrors with finite dimensions. We can intuitively guess that an oscillating electromagnetic wave loses power per round trip in the cavity due to light diffracting beside the finite mirrors. We call this an unstable resonator. An unstable resonator can still show laser activity, if the stimulated emission is strong enough to compensate for this loss of power. In most lasers, this loss will be avoided as good as possible using spherical mirrors. Spherical mirrors can transform the divergence of the light due to diffraction into a convergent propagation. Using beam theory, we can deduce the conditions for the curve of the mirrors to create a stable resonator.

Using the matrix formalism for translation (3.39) and for reflection at a spherical mirror (3.62), the system matrix for one round trip propagation in the cavity is given by:

$$\begin{aligned} \mathbf{M} &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -P_2 & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -P_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & L \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 - P_1 L & L(2 - P_1 L) \\ P_1 P_2 L - P_1 - P_2 & 1 - P_1 L - 2P_2 L + P_1 P_2 L^2 \end{bmatrix} \end{aligned} \tag{13.52}$$

with

$$P_{1,2} = \frac{2}{R_{1,2}} \tag{13.53}$$

Two succeeding periods are then characterized by:

$$\begin{bmatrix} x_{n+1} \\ \alpha_{n+1} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_n \\ \alpha_n \end{bmatrix} \tag{13.54}$$

$$\begin{bmatrix} x_{n+2} \\ \alpha_{n+2} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_{n+1} \\ \alpha_{n+1} \end{bmatrix} \tag{13.55}$$

Elimination of the angles gives us a recursive equation for the transversal position after each period:

$$x_{n+2} - (A + D)x_{n+1} + (AD - BC)x_n = 0 \tag{13.56}$$

As $AD - BC = 1$ we have:

$$x_{n+2} - (A + D)x_{n+1} + x_n = 0 \tag{13.57}$$

with

$$A + D = 2\left[1 - P_1L - P_2L + \frac{P_1P_2L^2}{2}\right] = 2\left[2\left(1 - \frac{P_1L}{2}\right)\left(1 - \frac{P_2L}{2}\right) - 1\right] \tag{13.58}$$

We propose a solution for this difference equation:

$$x_n = e^{\pm jn\theta} \tag{13.59}$$

Substituting this solution in the difference equation gives:

$$\cos\theta = \frac{1}{2}(A + D) \tag{13.60}$$

The general solutions takes the following form:

$$x_n = \rho_+ e^{jn\theta} + \rho_- e^{-jn\theta} \tag{13.61}$$

As a function of the boundary conditions x_1 and x_2, i.e. the transversal location of the incident beam and of the beam after one period, we obtain:

$$x_n = -x_1\frac{\sin(n-2)\theta}{\sin\theta} + x_2\frac{\sin(n-1)\theta}{\sin\theta}. \tag{13.62}$$

It is clear that the solution oscillates periodically around the optical axis. This is true if θ is real or:

$$-1 \leq \cos\theta \leq 1 \tag{13.63}$$

or

$$-1 \leq 2\left(1 - \frac{P_1L}{2}\right)\left(1 - \frac{P_2L}{2}\right) - 1 \leq 1 \tag{13.64}$$

or

$$0 \leq \left(1 - \frac{P_1L}{2}\right)\left(1 - \frac{P_2L}{2}\right) \leq 1 \tag{13.65}$$

or

$$0 \leq \left(1 - \frac{L}{R_1}\right)\left(1 - \frac{L}{R_2}\right) \leq 1 \tag{13.66}$$

This expression is shown graphically in figure 13.15. The gray zones in the graph represent the configurations for a stable resonator. Along the bisectors ($R_1 = R_2 = R$), the condition is:

$$R \geq L/2. \tag{13.67}$$

Possible symmetric configurations are shown in figure 13.16. Configurations with $R = \infty$, $R = L$ and $R = L/2$ are at the edge of stability. The figure presents *concentric* and *confocal* situations. In a concentric configuration, the centers of the two spherical mirrors coincide ($R_1 + R_2 = L$). In a confocal configuration, the foci of the two mirrors coincide ($R_1 + R2 = 2L$). It is clear that a simple paraxial beam theory gives us the necessary conditions for a laser to be stable, i.e. to show low losses. This condition is connected with the (amplitude) resonance condition, but does not tell us whether or not the cavity can be brought above threshold. To this end, light amplification and the transversal intensity profile of the field have to be taken into account. Wave theory becomes necessary.

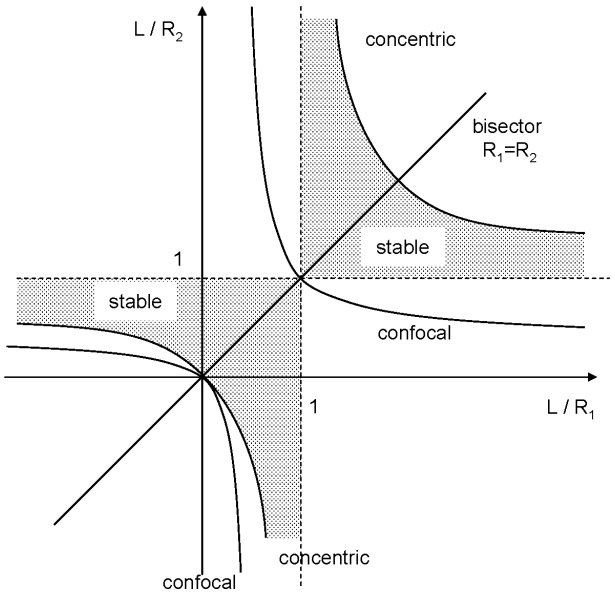

Figure 13.15: Stable resonators with spherical mirrors.

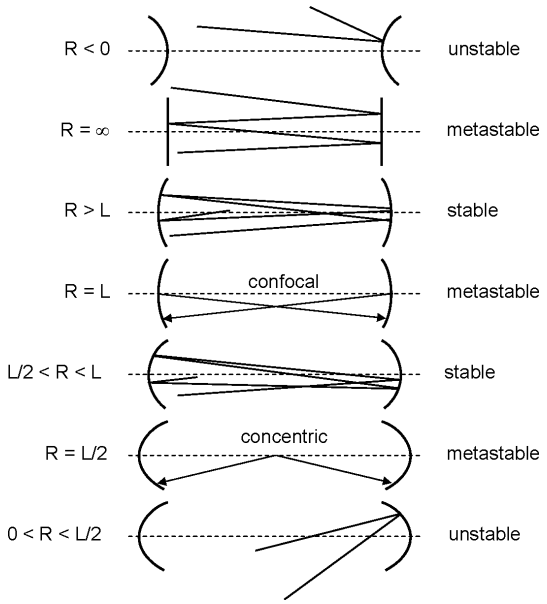

Figure 13.16: Different cases of a stable cavity with spherical mirrors.

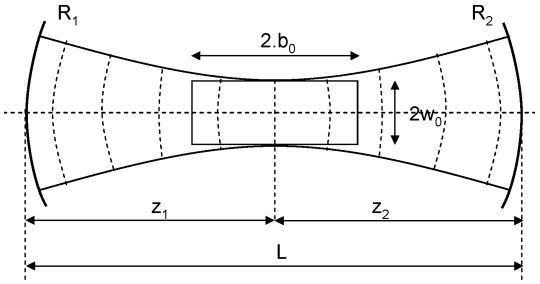

Figure 13.17: A Gaussian bundle in a cavity with spherical mirrors.

13.2.5 Resonance: Gaussian beam analysis

Finding an exact solution for the modes in the resonator is not simple. We show however that for long cavities with a length much larger than the transversal dimensions, modes show a Gaussian transversal amplitude distribution and a spherical phase front (at the end mirrors, this front coincides with the surface of these spherical mirrors). We present this using a paraxial approach, and we assume that a homogeneous medium is located between the spherical mirrors (this is true for most gas lasers).

The (amplitude) resonance condition implies that the transversal field profile is identical after one round trip in the cavity. Or $q_1 = q_2$ in equation (5.23):

$$q = \frac{Aq + B}{Cq + D} \tag{13.68}$$

with A, B, C and D the elements of the system matrix of the resonator (see previous section). This allows us to check whether or not the resonator has Gaussian solutions and to deduce their profile. Intuitively it is also possible to deduce a sufficient condition for resonance. If, as schematically presented in figure 13.17, the phase front of the Gaussian wave at the location of the two mirrors coincides with the mirror surfaces, we expect to have found a solution of the resonator cavity.

The characteristics of the Gaussian beam can be deduced from (see chapter 5):

$$R_1 = z_1 + \frac{b_0^2}{z_1} \tag{13.69}$$

$$R_2 = z_2 + \frac{b_0^2}{z_2} \tag{13.70}$$

$$L = z_1 + z_2 \tag{13.71}$$

If L, R_1 and R_2 are given, b_0, w_0, z_1 and z_2 can be found. For a configuration with $R_1 = R_2$ we have:

$$z_1 = z_2 = \frac{L}{2} \tag{13.72}$$

$$b_0^2 = \frac{L}{2}\left(R - \frac{L}{2}\right) \tag{13.73}$$

$$w_0 = \sqrt{\frac{2}{k}\sqrt{\frac{L}{2}\left(R - \frac{L}{2}\right)}} \tag{13.74}$$

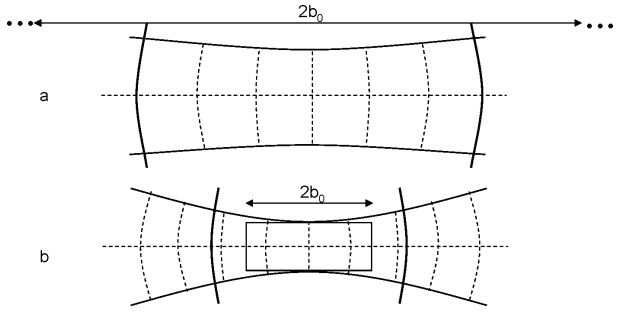

Figure 13.18: A Gaussian bundle in a cavity with spherical mirrors. (a) The cavity is shorter than the Rayleigh length. (b) The cavity is longer than the Rayleigh length.

It is clear that a solution can only be found if:

$$R \geq \frac{L}{2} \tag{13.75}$$

This is exactly the same condition as we obtained using beam theory.

By studying the beam profile as a function of the ratio of the curve of the mirrors to the length of the cavity, we can distinguish three regimes (figure 13.18): The Rayleigh length of the Gaussian beam will be larger than the cavity length if the curve of the mirrors is much larger than the cavity length. The beam can then be considered as a quasi-plane beam in the cavity. When the light escapes from the cavity, the beam will only fan open at a large distance from the output facet of the laser (figure 13.18a). If the mirrors are confocal, i.e. when the curves of the mirrors are as large as the length of the cavity, the Rayleigh length of the beam will be exactly half the cavity length. For mirror curves smaller than the cavity length, the Rayleigh length as well will be smaller than the cavity length, resulting in a fanning open of the beam inside the cavity. The beam outside the cavity can be considered as spherical (figure 13.18b).

The cavity mode described above is the lowest order transversal mode, i.e. the TEM_{00}-mode. The higher order modes, i.e. the Hermite-Gaussian beams, are also possible solutions of the cavity.

Considering the three-dimensional structure, we would find that the transversal field is the product of two Hermite-polynomials and a 2D Gaussian profile. These modes have two mode numbers (see equation (5.31)). Some Gauss-Hermite modes are depicted in figure 5.7.

We conclude that modes in a laser cavity have three independent mode numbers. Besides the two transversal mode numbers, each mode will have a longitudinal mode number related to the phase resonance condition. As mentioned above, all longitudinal modes have a different oscillation frequency. The transversal modes (with same longitudinal mode number) also have a slightly different oscillation frequency because of their possible different propagation constants (figure 13.19).

The number of longitudinal modes is predominantly determined by the spectral loop gain. The number of transversal modes depends on the transverse dimensions of the cavity. The expression for Hermite-Gaussian beams reveals that for a given $w(z)$, the higher order modes have a larger transverse width. By limiting the transversal dimension of the mirrors or the amplifying medium, or by simply inserting a diaphragm inside the cavity, the higher order modes can be suppressed.

As an example, let us consider a He-Ne laser with a typical beam diameter of 2mm, i.e. $w_0 = 1$mm. This results in $b_0 = 5m$ which is much larger than the typical cavity length of 300mm. The radius

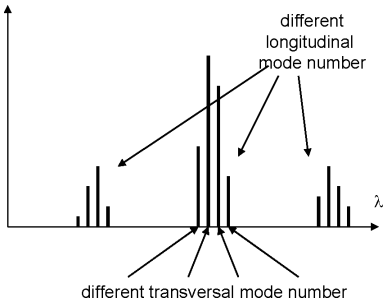

Figure 13.19: Longitudinal and transversal laser modes.

of curvature of the mirrors needs to be approximately 170m. It is clear that the laser mirrors have to be fabricated with a high accuracy. The divergence angle of the beam will be less than one arc-minute.

13.3 Characteristics of laser beams

Light generated by lasers is quite different from light generated by other sources like bulbs, TL tubes, etc. Laser light is highly monochromatic, coherent, directional and shows a high radiance. We briefly discuss these specific characteristics. The possibility to generate very short light pulses is a less fundamental but a very important characteristic of laser light. We discuss this in a separate section.

13.3.1 Monochromaticity

As opposed to conventional light sources that generate light with a broad spectral range, a laser emits light at a certain frequency. This high degree of monochromaticity is accomplished by two effects. First, only light in a small spectral range is amplified (this is determined by the width of the loop gain). Second, this amplified light oscillates in a cavity, imposing conditions for the oscillating frequency, namely the resonance frequency. The latter effect strongly reduces the width of the line function obtained for spontaneous emission.

13.3.2 Coherence

If a light source is perfectly monochromatic, the source is perfectly coherent as well. The electromagnetic field varies purely sinusoidal in time and this for all positions. This means that there exists a fixed phase relation between the lightfield at two different positions in space, and between the lightfield at two different moments at a certain position. A laser is a source that pursues this perfect coherence. However, perfect coherence is never attained, of course. Laser light is partially coherent. But how is partial coherence defined?

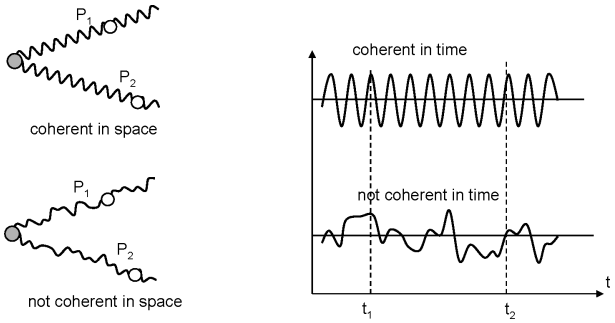

Figure 13.20: Temporal and spatial coherence.

Partial coherence is expressed by the coherence degree $\gamma_{12}(\tau)$. This is a measure for the correlation between the field at a point P_1 and the field at a point P_2 at different times t and $t+\tau$:

$$\gamma_{12}(\tau) = \frac{\langle E_1(t+\tau)E_2^*(t)\rangle}{\sqrt{\left\langle |E_1(t)|^2 \right\rangle \left\langle |E_2(t)|^2 \right\rangle}} \tag{13.76}$$

with $<>$ a temporal or statistical average (which are normally the same as the processes are in general ergodic). E_1 and E_2 are analytical signals corresponding to the real field. The absolute value of γ_{12} lies between 0 and 1.

Two different aspects of coherence are considered: temporal coherence and spatial coherence. Temporal coherence is described by $\gamma_{11}(\tau)$ telling us the measure of correlation between the field at a certain position at a certain time and the field at the same position at a time τ later. The coherence time τ_c is defined as the time for which the coherence degree is decreased to a certain amount (e.g. 0.5). A coherence length l_c is defined as well:

$$l_c = \tau_c \cdot c \tag{13.77}$$

Spatial coherence is described by $\gamma_{12}(0)$ telling us the measure of correlation between the field at a certain position at a certain time and the field at a different position at the same time. Mostly the spatial coherence is measured between two points on the same wave front. Depending on the situation, spatial coherence and temporal coherence can be related.

Temporal coherence is linked with the spectral width of the field. For purely monochromatic fields, coherence is infinite. If the field has a spectral width $\Delta\nu$, the temporal coherence τ_c will be in the order of:

$$\tau_c \approx \frac{1}{\Delta\nu} \tag{13.78}$$

Taking the He-Ne laser with its spectral width of 1.5 GHz for example, a coherence length of 20cm is achieved. The coherence length is important in interferometric applications, such as holography, where the laser beam is split in separate beams to subsequently let them interfere. The coherence length needs to be longer than the largest optical path difference between the two bundles.

13.3.3 Directionality

Incandescent lamps emit light in all directions. Using all kinds of optical systems, this light can be redirected in a certain desired direction. The cavity of a laser determines the directionality of the light propagating in the cavity and the light escaping from the cavity. In most cases, laser light can be described by Gaussian beams diverging with the smallest angle for a given beam width. In chapter 5, we defined the M^2-number (see equation (5.33)), a quality label for laser beams. For example, using a laser it is possible to create a laser spot on the moon (about 400.000km removed from the earth) with a diameter of only 800m.

13.3.4 Radiance

A low power laser beam (few milliwatts) has a radiance some orders of magnitude larger than the brightest conventional light sources. This is due to the high degree of directionality, resulting in very high intensities when focusing the laser beam (see equation (2.8) for the definition of radiance).

13.4 Pulsed Lasers

The lasers discussed above are mainly used in a continuous wave (CW) operation. However, many applications need short intense pulses periodically. Two techniques can be used to produce these short laser pulses.

13.4.1 Q-switching

The first technique is *Q-switching*. An element that changes the quality factor is put inside the cavity. For example, a rotatable mirror or an optical intensity modulator (e.g. an electro-optical or acousto-optical cell) can be used. This is shown in figure 13.21.

Initially, the cavity losses are made to be large ($t < t_0$). Pumping is started. Due to the huge losses, the threshold for lasing is high. The population inversion is pumped to a high level. At $t = t_0$, the cavity losses are lowered, corresponding to a sudden increase of Q. The number of excited particles needed for laser action is strongly decreased. Because of the high degree of population inversion, the system contains a surplus of excited elements, causing an intense stimulated emission. The number of photons in the cavity increases fast due to the strong stimulated emission, diminishing population inversion as well. The moment population inversion reaches its threshold for continuous operation, the light intensity is at its maximal value ($t = t_1$). In the meantime, the cavity losses are increased again. The photon density relaxes to zero. Using this technique, it is possible to create peak powers in the order of MWatt to GWatt. The pulses last some ns. The pulse energy can be as high as 1 Joule. Of course, it is essential that the cavity losses can be varied very fast.

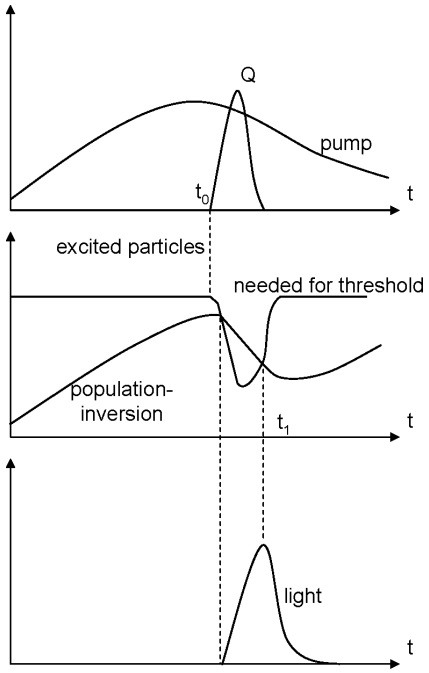

Figure 13.21: Q-switching.

13.4.2 Mode-locking

If even shorter pulses are desired, mode locking has to be used. This technique is based on the presence of several longitudinal modes. The total electrical field emitted by the laser can be written as, in complex notation:

$$E(t) = \mathrm{Re}\left[\sum_n E_n e^{j[(\omega_0 + n\Delta\omega)t + \phi_n]}\right] \tag{13.79}$$

with

$$\Delta\omega = \pi\frac{c}{L}. \tag{13.80}$$

Normally, the phase shifts ϕ_n fluctuate due to noise. The total field intensity can thus be approximated by the sum of intensities of the individual modes, increased by a certain amount of intensity noise (see figure 13.22).

However, imagine that it is possible to lock the phase shifts at constant values. The spectrum of the laser will then be very similar to the spectrum of an amplitude modulated carrier wave. As a function of time, we expect a periodically fluctuating intensity. Moreover, when all phase shifts ϕ_n are chosen in such a way that the modes are at their maximal value at the same time (and this periodically repeated), we expect strong laser pulses. This is for example the case when $\phi_n = 0$ for all modes. Let us assume for simplicity that there are N modes with the same amplitude. The

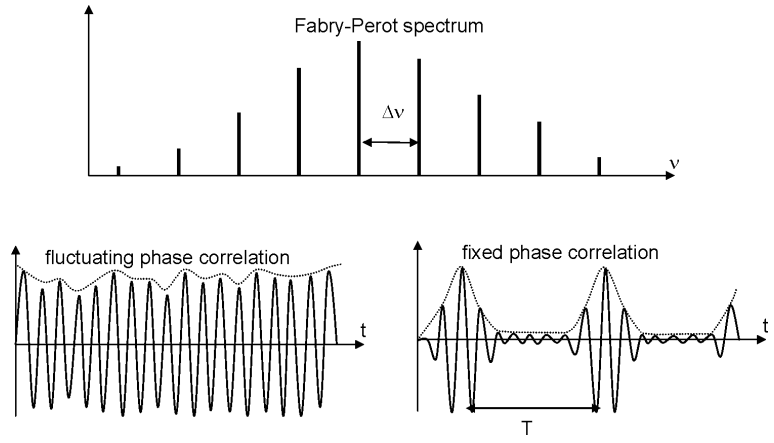

Figure 13.22: Mode locking by using a fixed phase shift between the respective modes.

total field is:

$$\begin{aligned} E(t) &= \mathrm{Re}\left[\sum_{-(N-1)/2}^{(N-1)/2} e^{j(\omega_0+n\Delta\omega)t}\right] \\ &= \mathrm{Re}\left[e^{j\omega_0 t}\right]\frac{\sin\frac{N\Delta\omega t}{2}}{\sin\frac{\Delta\omega t}{2}} \end{aligned} \tag{13.81}$$

and the intensity distribution:

$$I(t) = \frac{\sin^2\frac{N\Delta\omega t}{2}}{\sin^2\frac{\Delta\omega t}{2}} \tag{13.82}$$

It is clear that the period is given by:

$$T = \frac{2\pi}{\Delta\omega} = \frac{2L}{c}. \tag{13.83}$$

This is the time needed for one round trip in the cavity. The peak intensity is given by N times the average intensity and the pulse duration τ is about T/N, corresponding to a pulse length of $2L/N$. Thus, the pulse is short compared to the cavity length. The number of modes N is determined by the spectral line of the loop gain, setting the maximal number of modes to:

$$N_{\max} = \frac{\Delta\omega_{gain}}{\Delta\omega} \tag{13.84}$$

with $\Delta\omega_{gain}$ the spectral width of the loop gain. The duration of the pulses can thus be as short as:

$$\tau_{\min} = \frac{2\pi}{\Delta\omega_{gain}} \tag{13.85}$$

A short pulse travels back and forth in the laser cavity with a period equal to the roundtrip time. This intuitively points out a method to achieve this 'mode locking': inserting a suitably fast light modulator at the end of the laser cavity (near the mirror) which is actuated in such a way that the modulator is transparent when the pulse passes and non-transparent when the pulse is elsewhere In that way the laser has no other option than to oscillate in this pulsed fashion. The period of

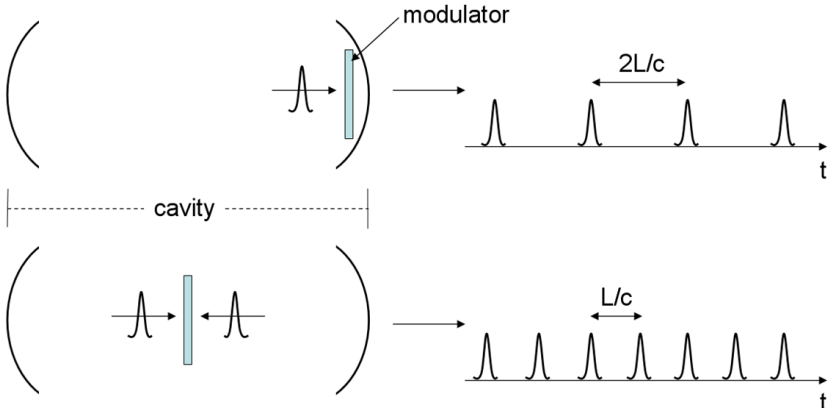

Figure 13.23: Position of the modulator in the laser cavity and resulting pulse train.

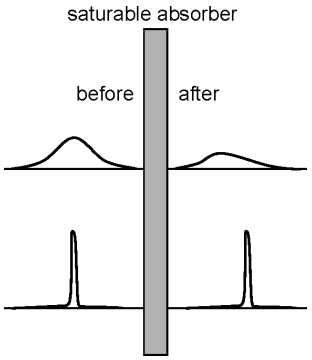

Figure 13.24: A long and a short pulse propagating through a saturable absorber.

the modulating signal should then be equal to the roundtrip time, $2L/c$, and the transmission has to be large during a time that is short in comparison with the time of one cavity roundtrip. This way only one pulse can survive in the laser. We can now see that if the modulator is placed in the center of the cavity (see figure 13.23) and if it has a pulsed periodic transmission profile with period L/c, then two pulses can exist in the cavity. They cross each other at the modulator when its transmission peaks. This results in an output pulse train with twice the repetition rate. The same reasoning can be followed for a modulator at other places in the cavity. To obtain this kind of mode locking, an external signal is needed to drive the modulator. This is why it is called 'active' mode locking.

Alternatively, a saturable absorber is put in the cavity (see figure 13.24). This nonlinear element shows a high absorption when the light intensity is low and a low absorption when the light intensity is high. It is clear that in these systems, a short intense pulse will be favored with respect to longer less intense pulses with the same total power. Mode locking is established automatically. This is called 'passive' mode locking.

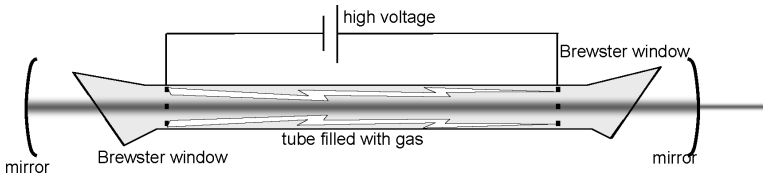

Figure 13.25: The gas laser.

13.5 Types of lasers

13.5.1 Introduction

All kinds of lasers exist. Nevertheless, the basic principles remain the same: a material emitting light at a specific wavelength range is brought to population inversion; a cavity with a high quality factor is holding this material. The difference between the laser types is in the first place found in the used laser material. This implies different methods to pump, different dimensions and different technical implementations of the cavity. We discuss the most important laser classes, grouped according to the phase of the active material: gas ($He-Ne$ laser, $Ar-$laser, CO_2-laser, etc.), liquid (dye laser), or solid state lasers ($dopedisolators, semiconductor$ lasers) or $freeelectron$ lasers consisting of an electron beam in a vacuum cavity.

13.5.2 Gas lasers

Gas lasers have been the most popular lasers for a long time, e.g. the $He-Ne$ laser, $Argon$ laser, $Krypton$ laser and CO_2 laser. Although they remain popular, other laser types are replacing the gas lasers more and more. For example, the He-Ne laser competes with the semiconductor laser. This hard competition is related to the quite large dimensions of gas lasers, their need for an expensive high voltage supply and their relatively short lifetime. Different types of gas lasers exist. The energy transitions can be electronic transitions of atoms or ions or vibrational/rotational transitions of molecules. In all three cases, pumping is due to excitations caused by electronic or molecular collisions in a gas discharge. The gas discharge is generated by a high voltage between two electrodes in the low pressure gas mixture. One gas in the gas mixture is excited by electronic collisions. Its energy is transferred to the other gas by atomic or molecular collisions.

The typical cavity of a gas laser is sketched in figure 13.25. Two windows inclined at the Brewster angle end the plasma tube. This causes maximal transmission (minimal cavity losses) for only one polarization. Thus, the laser light is polarized. The tube is placed in between two spherical mirrors. Due to the little gain in the cavity, the tube needs to be quite long, typically 30cm to 3m. The laser threshold is reached only when using a long cavity and mirrors with a very high reflectivity.

He-Ne lasers are the oldest and most popular low power gas lasers. The pressure inside the tube is a few mmHg and the gas mixture typically contains ten times more He than Ne. It is a four-level system. The laser transition happens between two levels of the Ne-atom (figure 13.26). Helium is used for pumping. It is excited by electrons and transfers its energy by atomic collisions with Neon atoms. Several laser transitions are possible in this system. The most popular is the transition at a

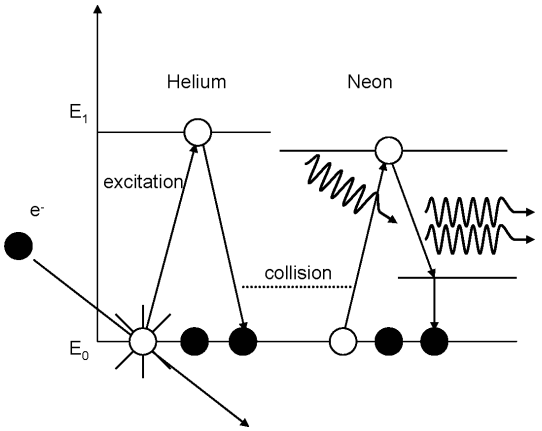

Figure 13.26: Energy levels of a He-Ne laser.

wavelength of 632.8nm; besides that one there is also 1150nm, 1520nm and 3390nm. The frequency selection is obtained by the wavelength dependent reflectivity of the mirrors.

The power of a He-Ne laser is low due to the small efficiency of about 0.01%. Most lasers emit a power of about 1 to 10mWatt. But the light can be made very pure, both concerning transversal and longitudinal modes.

A second important class of gas lasers are the ion lasers, like the *Argon* laser and *Krypton* laser. First the gas is ionized by electrons in the plasma, and further excited to even higher energy levels. The relaxation from these levels causes laser action at several frequencies (lines). An Argon laser emits at lines between 350 and 520 nm. A Krypton laser can cover the whole visual spectrum. The frequency is selected by the use of a rotatable prism in the cavity (in between the tube and the facet mirror) or by employing a coated frequency selective mirror. Although the efficiency of these lasers is not much higher than the efficiency of a He-Ne laser, the emitted power can be relatively high, typically about 20Watt. The technology used to obtain this higher output power is quite complex. Specifically, the heat dissipation is critical. High power ion lasers need a water-cooling system. Moreover, the tube of the cavity is made of special materials showing a large thermal conductivity (BeO, graphite,etc.).

Metal vapor lasers form a third class of gas lasers. The active particles are metal atoms or ions in a low pressure atmosphere. A popular example is the He-Cd laser with the most used lines at 442 nm and 325 nm.

Molecular lasers, such as the CO_2 laser, the *nitrogen* laser and the *excimer* laser are the fourth and final class of gas lasers. Vibrational and rotational modes of the CO_2 molecule are responsible for the transitions in a CO_2 laser. Another gas, typically nitrogen in combination with helium, is used for the excitation of the CO_2 atoms. The CO_2 laser is in most cases used to emit at 10.6μm (far infrared). The windows and output mirrors need to be transparent in the infrared. This restricts the possible materials to Ge, ZnSe, GaAs, diamond, etc. The high power efficiency of the CO_2 laser is remarkable. It can be as high as 30%. This high efficiency makes it possible to build lasers emitting high powers to about several kWatts in continuous wave. The main application is material processing. Nitrogen lasers emit mainly at 337.1 nm.

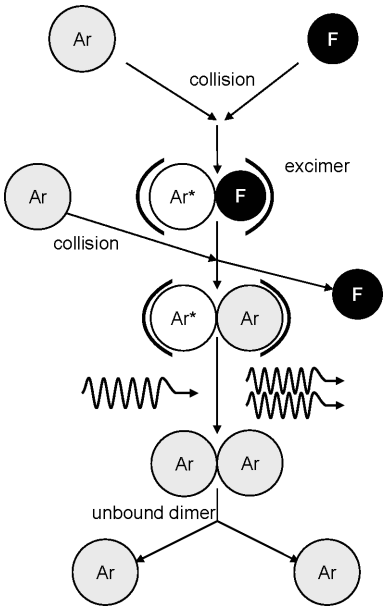

Figure 13.27: The principle of the excimer laser.

The excited particles of an *excimer* laser are excited dimers (excimer). This is a metastable molecule consisting of an excited atom/molecule compound with a not-excited atom/molecule. Mostly it concerns a halide - noble gas combination. Popular for their high efficiency are the XeF and the KrF lasers (10%).

Let us examine the ArF excimer laser (figure 13.27). A gas mixture of Argon and Fluor is heated by means of a discharge. Some Argon and Fluor atoms collide and compound as a stable excimer. Upon collision with an Argon atom, an Ar*Ar excimer can be formed. When excited, this excimer is weakly bound, as opposed to the ground state, where it is not bound. In presence of a photon field, the excimer will quickly relax to the ground state. It disintegrates with emission of a photon.

Excimer lasers emit at wavelengths between 120 and 500nm. They are the main lasers used in the UV spectrum. Two important types of applications can be considered. In the first applications the UV light is used for its short wavelength (for example high resolution imaging systems). This field is gaining importance in the deep UV lithography for the definition of advanced integrated circuits with line widths (smallest width of the patterns) of about 0.1-0.3 micrometer. The second kind of application uses the excimer lasers for its high photon energy. It permits to stimulate all kinds of chemical processes. We mention laser ablation. Laser ablation is used to 'drill' very small holes or vias in printed circuits or plastic shields.

13.5.3 Solid-state lasers: the doped isolator laser

Solid state lasers use a transparent substance (crystalline or amorphous glass, usually an oxide) as the active medium, doped with a small amount of metal ions to provide the energy states necessary for lasing. The pumping mechanism is the radiation from a powerful light source, such as a flash lamp. The first laser of this type, and the very first laser in general, is the *ruby laser*,

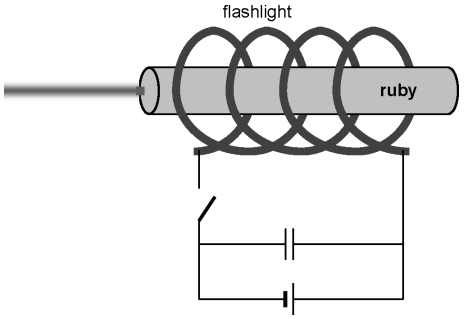

Figure 13.28: The principle of the first ruby laser.

invented in 1960 by *Maiman*. Ruby, and similarly sapphire and corundum, are crystalline Al_2O_3. The crystals are different due to the presence of impurities. These impurities define the typical color of the crystals (ruby is red, sapphire blue and corundum white or transparent). Ruby lasers uses synthetically grown Al_2O_3, doped with 0.05 volume% Cr^{3+} ions. It is a three-level system, demanding for a high pump level to reach population inversion. Continuous wave operation is as such difficult due to the high heat dissipation. The laser transition has a wavelength of 0.6943μm, dark red light, only just visible to the human eye.

Typically, a rod of 10 cm length and 1 cm diameter is used. The first lasers had polished and metalized end facets. Nowadays, external spherical mirrors are used. The crystal can be cut with an inclination angle equal to the Brewster angle (figure 13.25). An intense flashlight pumps the system. The flashlight is curled as a spiral around the crystal rod, surrounded by reflectors (see figure 13.28). The Neodymium-YAG laser is the most important solid-state laser. Its active material is Yttrium-Aluminum garnet or $Y_3Al_5O_{12}$, doped with Nd^{3+} ions. It is a four-level system, which means that less energy is needed to pump the system to the laser threshold in comparison with the ruby laser. The emitted wavelength is 1.06μm. Alternatively, a second laser transition can be used at 1.3μm, although requiring more pump energy (higher threshold).

The Nd-YAG laser works both in a continuous wave and pulsed operation. In both cases, optical pumping is used mostly (flashlight or continuous). An often used configuration is sketched in figure 13.29a. Both the laser rod and the pump light source (a rod as well), are placed at the foci of the surrounding elliptic reflector. High optical powers, on the order of 100 Watt, can be obtained. The efficiency is about several percents. Material processing is one of the main applications where Nd-YAG lasers are used. The choice to use a Nd-YAG or a CO_2 laser for material processing is based on the needed resolution (the wavelength of a YAG laser is ten times smaller) and the absorption/reflection characteristics of the material for the given wavelengths.

Recently ($\pm$1987) an alternative manner of pumping has been developed, using the light of a high power semiconductor laser (1Watt) axially coupled into the crystalline rod (figure 13.29b). The Nd-YAG light is extracted at the other side of the rod. The efficiency of this configuration is far better due to a better use of the pump light, both spatially and spectrally. With an efficiency of 20-30%, an output power in the order of 100mWatt can be obtained. These lasers can be made compact and are quite cheap. Moreover, it is possible to insert a nonlinear crystal in the cavity for frequency-doubling. Green light with a frequency of 530nm is obtainable.

Many variants on the Nd-YAG laser exist. YAG can be replaced by other crystals, like YLF (Yttrium-Lithium-Fluoride) or even amorphous glass. Amorphous glass is cheap and can be pro-

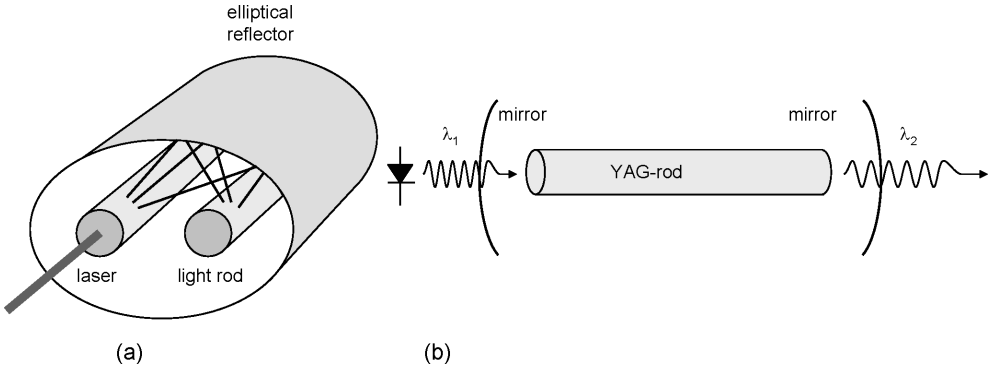

Figure 13.29: Pump mechanisms for the Nd-YAG laser: (a) with rod lamp and elliptical mirror, (b) with a laser diode.

cessed easily. However, the structure of the glass results in a broad spectral gain width and thus a higher required pump energy (only pulsed operation). Due to the higher possible concentration of Nd in the glass, the energy per pulse can be higher.

An important evolution for solid-state lasers, are the tunable wavelength solid-state lasers. The active material of these lasers shows a very broad spectral gain width. A tunable frequency selective transmission element is inserted in the cavity. Doing so, any wavelength in the spectral gain window can be chosen to be emitted. Examples of wavelength tunable solid-state lasers are Alexandrite Lasers and Titanium-Sapphire Laser (active material is Al_2O_3 doped with Titanium ions) with a tunable range between 700 and 950nm. These lasers are in general optically pumped by an axially installed gas laser.

Fiber lasers form a special class of solid-state lasers. The active gain medium is an optical fiber which is doped with rare-earth ions such as neodymium (Nd^{3+}), erbium (Er^{3+}) or ytterbium (Yb^{3+}). The pump light is usually provided by one or more fiber-coupled laser diodes and propagates in the fiber. The cavity is often formed by splicing fiber Bragg gratings to the doped fiber. These are optical fibers with a periodically varying refractive index in the direction of propagation.

Fiber lasers are an attractive alternative for the heavy, fragile and power consumptive bulk solid-state lasers. The light in the fiber is shielded from the surroundings and the laser is quite robust. Doped fibers boast a high gain efficiency and the fibers can operate with low pump powers while output powers can be as high as several kilowatts. Due to new fiber concepts and technologies, fiber lasers have made massive progress during the past few years and are ready to compete with solid-state lasers in many practical applications.

13.5.4 Semiconductor lasers

Semiconductor lasers are discussed in more detail in chapter 14.

13.5.5 Dye lasers

A *dye* laser is a laser that uses an organic dye as a lasing medium, usually a liquid solution. Organic dyes (organic molecules that strongly absorb light at some wavelength ranges of the optical

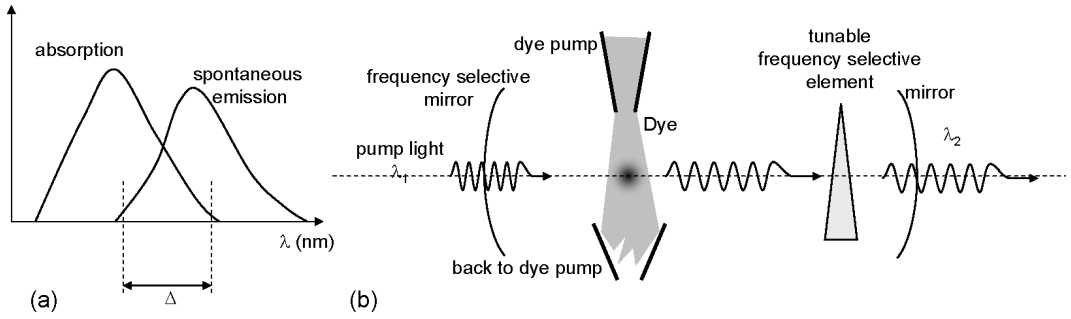

Figure 13.30: (a) Absorption and emission spectrum of a dye; (b) schematic setup of a dye laser.

spectrum) efficiently emit light when relaxing to the ground state. Due to the many available energy levels, the laser can be tuned over a broad spectral range. Figure 13.30a shows a typical absorption and emission spectrum of the Rhodamine 6G molecule. Using a tunable dispersive element (etalon, prism or diffraction grating) laser oscillation can be obtained for any wavelength in a spectral range of about 50nm. With the existing dye lasers, the whole visible spectrum and near-infrared can be covered.

A simplified scheme of a dye laser is given in figure 13.30b. One of the spherical mirrors is transparent for the pump light (often Krypton or Argon laser light), the other spherical mirror transmits the dye laser light. The dye, dissolved in water or alcohol and pumped in a closed circuit, is squirted into the laser beam. This system avoids cooling problems. Some dispersive elements inside the cavity complete the laser.

13.5.6 The free electron laser

A free electron laser transforms the kinetic energy of a relativistic electron beam into electromagnetic (EM) radiation. The interaction between the photons and the electrons is rather complex and is not described in detail here. The basic principles are the following. An accelerated electron emits radiation. E.g. synchrotron radiation in ring accelerators, which is however not a suitable candidate for lasing. The electron beam can be produced by a particle accelerator like a microtron, storage ring, etc. The transformation occurs when the beam goes through an alternating magnetic field that forces the electrons to move in an oscillatory trajectory along the axis of the system. One can prove that an electron exchanges energy with an existing electromagnetic field only if its velocity has a component parallel to the present electric field. This requirement is imposed by the conservation of energy and conservation of momentum. Considering the electromagnetic field to be amplified as a plane wave propagating along the z-axis, the electric field is oriented normal to the z-direction. Thus, the electron and the photon may not propagate in exactly the same direction.

Now, we want the electron to transfer its energy to the electromagnetic field. Or, the electromagnetic field receives energy, to which a negative work is related. (Work is defined as positive if the force transfers energy to the object and negative if the force transfers energy from the object)(figure 13.31). Power is defined as the rate at which a force does work on a body:

$$\begin{aligned} \Delta W &= \mathbf{F}.\boldsymbol{\Delta}\mathbf{r}_e \\ &= q_e.\mathbf{E}.\boldsymbol{\Delta}\mathbf{r}_e = -e\mathbf{E}.\boldsymbol{\Delta}\mathbf{r}_e \end{aligned} \tag{13.86}$$

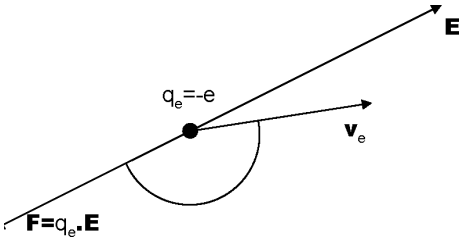

Figure 13.31: Emission of photons by accelerating electrons.

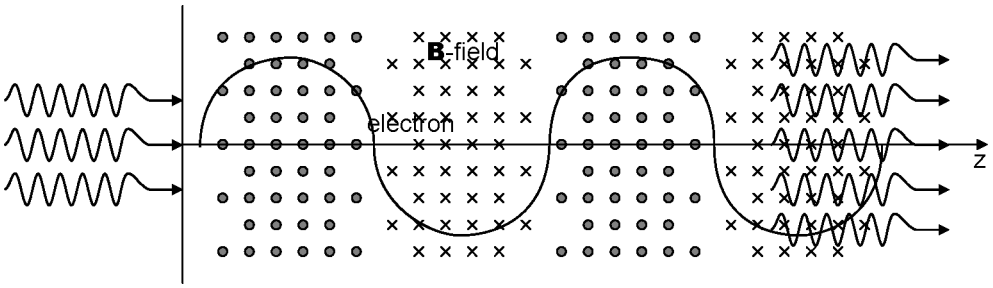

Figure 13.32: Propagation of an electron in an alternating magnetic field.

The energy that an electron transfers per time unit to the electromagnetic field is then given by:

$$\frac{dE_e}{dt} = -\frac{dW}{dt} = e\mathbf{E}.\frac{d\mathbf{r}_e}{dt} = e\mathbf{E}.\mathbf{v}_e \tag{13.87}$$

Assuming the electromagnetic field to be a plane wave, propagating in the z-direction and linearly polarized along the x-direction, the transferred energy can be written as:

$$\frac{dE_e}{dt} = eE_{0x}.v_{ex}.e^{-j(\omega t - k_z z)} \tag{13.88}$$

If the electromagnetic field and the electron propagate rectilinearly in different directions, interaction will occur, but the average energy transfer will be zero due to the oscillating character of the electromagnetic wave. The sign of the energy transfer would change every time the electron gets behind the electromagnetic field another distance $\lambda/2$.

Therefore, the electron needs to cover a periodic path instead of a rectilinear one. This can be obtained using a spatially alternating DC magnetic field oriented along the y-axis (see figure 13.32). An undulator or a Wiggler is used for this purpose (see figure 13.33). If the period P of the spatially alternating DC magnetic field is given by:

$$P = \lambda \frac{v_z}{c - v_z} \tag{13.89}$$

with c the velocity of light in the laser substance (often this is vacuum) and v_z the speed of the electron along the z-axis, E_x and v_x will be in phase and change sign simultaneously. The transferred energy is positive and net amplification is obtained. This is depicted in figure 13.34, showing the electron and the electric field at different times. The electromagnetic field passes by the electron, but the speed of the electron along the x-direction v_x inverts when E_x changes sign. This results in a product $E_x \cdot v_x$ that remains positive.

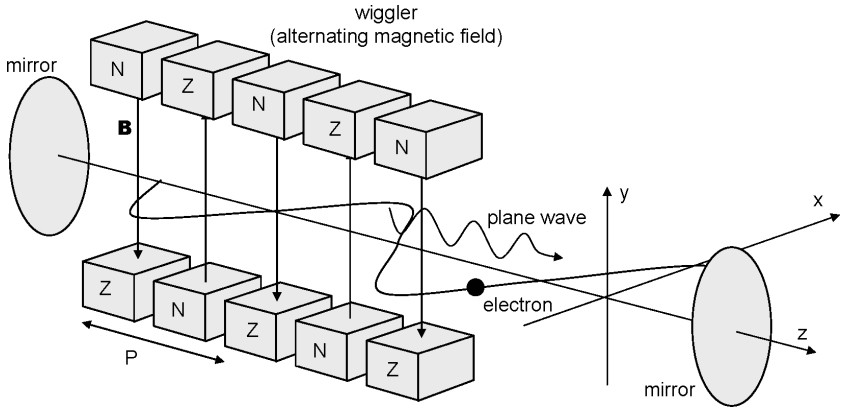

Figure 13.33: The 'Wiggler' in a free-electron laser.

To obtain laser action with this amplification mechanism, the wiggler needs to be set up in between the mirrors of the cavity. The ensemble action of the electromagnetic field and the alternating B field of the wiggler bundles the electrons in packets. These electrons amplify the electromagnetic field coherently.

Most experiments with free electron lasers are limited to wavelength ranges in microwave and infrared. The applications remain restricted. It is also possible to use this mechanism for amplification of laser light. The mirrors are removed and the laser light (e.g. from a CO_2-laser) is directed to the Wiggler and amplified.

An important disadvantage of the free electron laser is the need for an electron accelerator, and the related dimensions. However, the free electron laser is a very efficient laser source as the energy of the electrons that is not used for amplification can be recuperated. The possible broad wavelength range and high output power promises the free electron laser a wealth of applications.

Bibliography

[ST91] B.E.A. Saleh and M.V. Teich. *Fundamentals of Photonics*. John Wiley and Sons, ISBN 0-471-83965-5, New York, 1991.

[Sve98] O. Svelto. *Principles of Lasers*. Plenum Press, ISBN 0-306-45748-2, 4th edition, 1998.

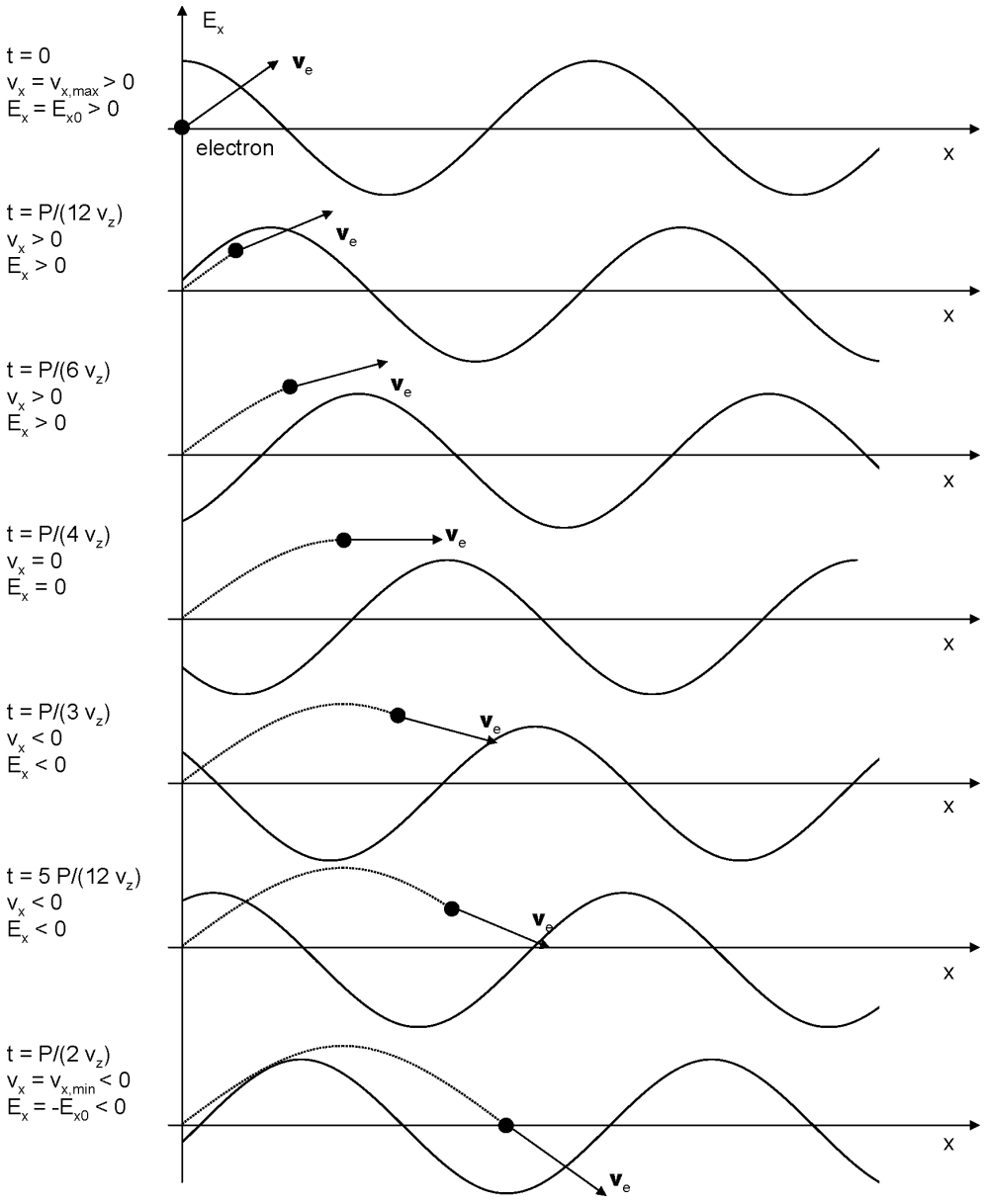

Figure 13.34: The path of an electron in a free-electron laser.

Chapter 14

Semiconductor Light Sources

Contents

Semiconductors and more specifically semiconductor diodes are of utmost importance in photonics. They are at the basis of a large number of components at the interface between electrical signals and optical signals, including light emitting diodes (LED's), laser diodes, photodiodes, photovoltaic cells etc.

In this chapter we start with a discussion of the optical properties of semiconductors and continue with a review of the basic properties of semiconductor diodes before discussing light emitting semiconductor components. In the next chapter light detecting semiconductor components are covered.

The student is assumed to be familiar with basic concepts of semiconductor physics. Hereafter a list of terms (and associated symbols) is given of which the reader is assumed to have some basic understanding:

- bandgap E_g
- valence band
- conduction band
- direct and indirect bandgap
- electrons
- holes
- electron/hole effective mass m_e^*/m_h^*
- Fermi-Dirac distribution

	IIIa	IVa	Va	VIa
	B	C	N	O
IIb	Al	Si	P	S
Zn	Ga	Ge	As	Se
Cd	In	Sn	Sb	Te

Figure 14.1: Semiconductor elements in the table of Mendeljev.

- Fermi level E_f
- density of states D_c/D_v
- intrinsic semiconductor
- intrinsic electron/hole concentration n_i
- doping
- donor/acceptor concentration N_d/N_a
- electron/hole mobility μ_n/μ_p
- electron/hole diffusion
- electron/hole diffusion coefficient D_n/D_p
- recombination
- electron/hole lifetime τ_n/τ_p

14.1 Optical properties of semiconductors

14.1.1 Types of semiconductors

There are many sorts of semiconductors. They can be classified according to the number of elements in their chemical combination. The most common and mostly used semiconductors are the elementary semiconductors, such as Silicon and Germanium, consisting of one element out of group IV of the table of Mendeljev (see figure 14.1). They have a diamond structure, in which each atom is bound covalently to four other identical atoms according to a tetrahedron structure. The crystal structure of compound semiconductors consists of different elements. In III-V semiconductors these elements are group-III elements (Al, Ga, In) and group-V elements(P, As, Sb). In II-VI semiconductors we find elements from group II (Cd, Zn, Hg...) and group VI (O, S, Se, Te...). Finally, we also have the IV-VI semiconductors in which Pb for example is the group-IV element.

In the compound semiconductors we also make a difference between binary, ternary and quaternary semiconductors. Examples of binary semiconductors are GaAs, AlAs, InP, ZnSe etc. GaAs and InP also have the diamond structure, but in this case each Ga (resp. In) atom is bound to

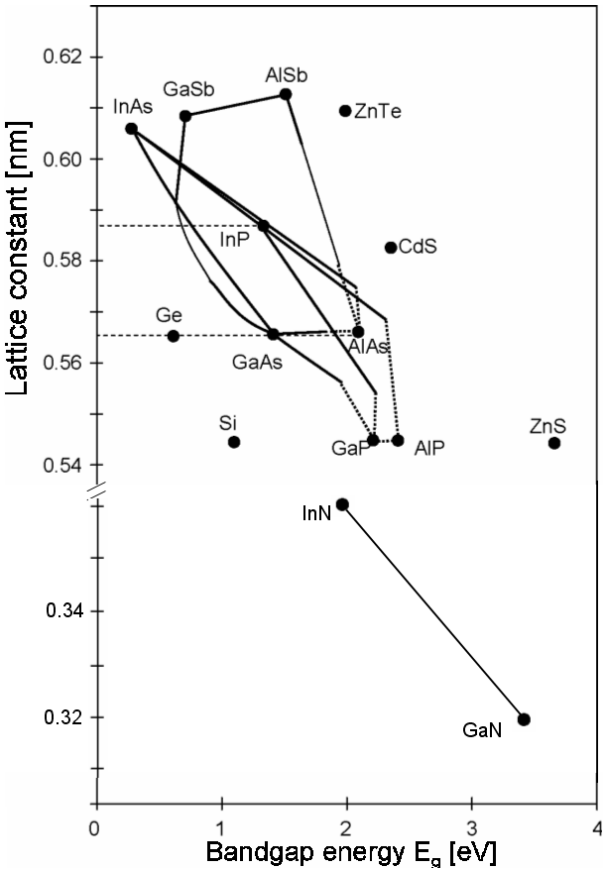

Figure 14.2: Correlation between the bandgap and the lattice constant for some important semiconductors.

four As (resp. P) atoms and vice versa. This bond is no longer purely covalent, but has a slightly ionic character. This means that there is a partial transfer of electrons from one type of atoms to the other. This gives rise to dipole momenta at an atomic level that contribute to the dielectric constant and cause a deviation between the static dielectric constant and the optical dielectric constant. Due to this slightly ionic character, these semiconductors are also called polar semiconductors.

When we mix two different binary semiconductors, we get ternary semiconductors. GaAs and AlAs can be mixed in any relation to $Al_xGa_{1-x}As$. Each As-atom is still surrounded by four group-III atoms, that can be Ga or Al, so that there is a mean fraction x Al and a fraction $1-x$ Ga. Quaternary semiconductors arise by mixing three binary semiconductors. Examples are $In_xGa_{1-x}As_yP_{1-y}$ and $In_xAl_yGa_{1-x-y}As$.

Next to the chemical structure, semiconductors can also be classified according to their band structure. For the optical properties it is of utmost importance to know whether or not a semiconductor has a direct or indirect band structure. When having a direct band structure, the minimum in the conduction band will occur for the same k-vector as the maximum of the valence band. This means that the free electrons have approximately the same k-value as the free holes, which is good for their interaction. The elementary semiconductors like Ge and Si have an indirect band structure. Many (but not all) III-V and II-VI semiconductors have a direct band structure.

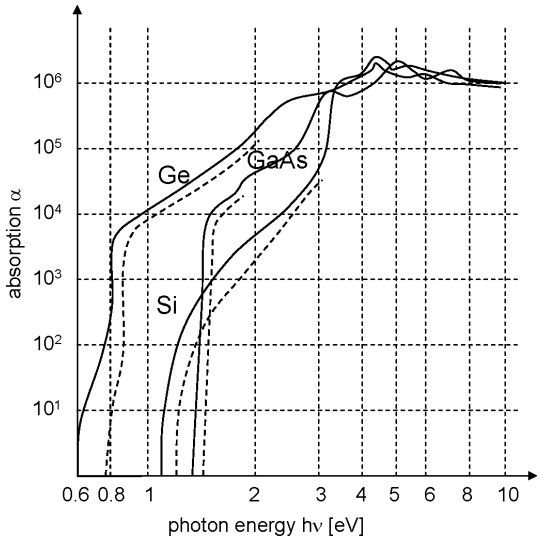

Figure 14.3: Absortion coefficient α of a number of important semiconductors.

The lattice constant and the bandgap are represented in figure 14.2 for a number of semiconductors. The lines denote ternary semiconductors. These lines combine two binary semiconductors. In practise all layers are grown on an appropriate substrate and thus al layers need to have the same lattice constant as this substrate. GaAs and InP are the mostly used substrates in the III-V semiconductors. Notice on the figure that $Al_xGa_{1-x}As$ has a lattice constant that is quasi independent of x. This is very handy cause it implies that AlGaAs-layers can be grown on GaAs and on each other with an arbitrary Al-concentration. The case of $In_xGa_{1-x}As_yP_{1-y}$ is bit more difficult. This material is usually grown on InP. One of the two degrees of freedom x and y has to be sacrificed to equal the lattice constant to the one of InP. The other degree can then be used to choose the bandgap.

The last qualification is according to doping with donors or acceptors which leads to n-type and p-type semiconductors respectively. Si and Ge are doped with group-III acceptors or group-V donors. III-V semiconductors are doped with group-II acceptors or group-VI donors (group-IV atoms can sometimes be an acceptors sometimes a donor).

14.1.2 Optical properties

Like in all other dielectric materials, optical properties of semiconductors can be described by a complex refractive index $n_C = n_R + jn_I$. The imaginary part expresses whether the material shows losses or amplification. The power absorption coefficient α is defined in this context

$$\alpha = -\frac{4\pi n_I}{\lambda} \tag{14.1}$$

A positive value for α implies losses.

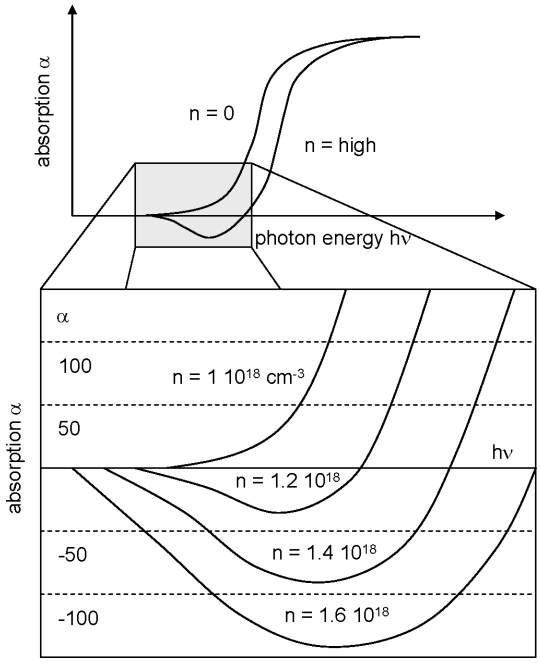

Figure 14.4: Absorption/gain-spectrum in function of the electron concentration.

Absorption and amplification of light in a semiconductor

The absorption in semiconductors shows a typical behavior in function of the wavelength. As long as the photon energy of the incident light is small (i.e. long wavelengths) compared to the bandgap of the material, only little absorption will occur. For larger photon energies, photons can cause excitation of electrons from the valence band to the conduction with transfer of energy from the photon to the electron. The k-value of the electron remains the same in this transition. This means that the absorption coefficient is low for $E < E_g$ and high for $E > E_g$. In semiconductors with a direct band structure (e.g. GaAs) this transition is abrupt, while for indirect semiconductors (e.g. Si) this transition is more gradual. The absorption coefficient of a few semiconductors is showed in figure 14.3. We note that α is larger than $10^4\,\mathrm{cm}^{-1}$ if $E > E_g$. This means that the incident light has been considerably absorbed after a distance $10^{-4}\,\mathrm{cm}$ ($= 1\,\mu\mathrm{m}$). If the material shows population inversion, which implies in semiconductors that there has to be a large concentration of electrons in the conduction band as well as holes in the valence band (this means that the material is no longer in a state of thermal equilibrium), stimulated emission will become more important than absorption. *Stimulated* emission is the process in which an electron recombines with a hole in the valence band (in other words drops back to the valence band) under the influence of an incident photon. The energy that comes free in this process is released as a new photon that has the same propagation direction and phase as the incident photon. Light amplification thus occurs. In other words, the absorption coefficient becomes negative. This phenomenon arises when the photon energy is approximately equal to the bandgap.

A typical absorption/gain spectrum is sketched in figure 14.4 for the quaternary semiconductor InGaAsP. The latter is important for optical communication. The absorption/gain at the band transition is represented as function of the photon energy for different values of the electron concentration n (that is supposed to be equal to the hole concentration). A typical gain value is $100\,\mathrm{cm}^{-1}$

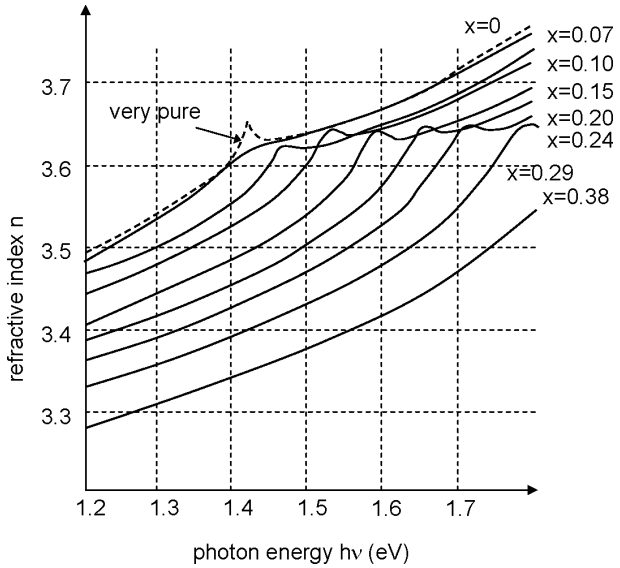

Figure 14.5: Refractive index of $Al_xGa_{1-x}As$ in function of the photon energy and x.

(much smaller than the absorption values of $10^4\,cm^{-1}$ and more for greater photon energies). Notice that the gain only occurs for photon energies near the bandgap.

Refractive index

The refractive index n of most semiconductors is pretty high. Si, Ge, GaAs, InP all have a refractive index between 3 and 4. Generally spoken, semiconductors with a large bandgap will have a rather small refractive index and conversely a small bandgap will lead to a bigger refractive index. In $Al_xGa_{1-x}As$ for example the bandgap will increase and the refractive index will decrease with increasing Al-percentage. This characteristic is of crucial importance for semiconductor lasers. Furthermore, the refractive index is dispersive (wavelength-dependent) and normally decreases with increasing wavelength. A small peak often occurs in the refractive index near the bandgap. Due to the very sudden variation of the absorption, the refractive index will show a perturbation there (because of the Kramers-Kronig relations). In figure 14.5 the refractive index of $Al_xGa_{1-x}As$ is showed in function of different x-values.

Spontaneous emission

If a semiconductor is brought out of thermal equilibrium (e.g. by absorption of light or by a current injection across a junction) the condition $np = n_i^2$ will be broken. If $np > n_i^2$ (i.e. a surplus of electrons or holes), the semiconductor will spontaneously try to restore the equilibrium. Electrons will hereby recombine with holes releasing the energy to a photon (radiant recombination) or to another electron or hole (which then gains kinetic energy) or to a phonon (crystal lattice vibration). The latter cases are called non-radiant recombination. A recombination process is often described by means of the lifetime of the electrons and the holes. The lifetime is the average time a charge carrier spends in an excited state before falling back to its ground state. The recombination process

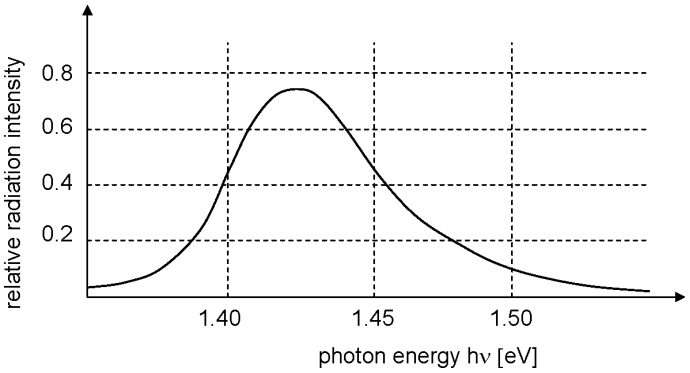

Figure 14.6: Typical spontaneous emission spectrum.

with the smallest lifetime will be dominant. In semiconductors with a indirect bandgap, the probability of radiant recombination is small as the electrons at the bottom of the conduction band have a different k-value than the holes at the top of the valence band. The difference in k-vector can no longer be compensated by a photon. Non-radiant recombination processes therefore dominate in which the difference in k-vector is usually compensated by a phonon. Radiant recombination can dominate in semiconductors with a direct bandgap, which causes a good energy transfer of the excited particles to light. Spontaneous emission is of course the strongest for photon energies near the bandgap. A typical spontaneous emission spectrum is sketched in figure 14.6.

Other phenomena

In conclusion of this paragraph, we briefly mention that the optical properties of semiconductors show a number of more complex aspects that are being used in lots of components. The complex refractive index of a semiconductor can be influenced in different ways. We mention:

- influence of the temperature on n (thermo-optic effect)
- influence of the electron and hole concentration on n (the previously mentioned effect of charge carriers on the gain/absorption as well as the plasma effect)
- influence of a static electric field on the refractive index or on the absorption (Pockels-effect, Kerr-effect, Stark-effect)
- influence of elastic tension on the refractive index (photo-elastic effect)

14.2 Diodes

14.2.1 The pn-junction

Suppose that in a certain semiconductor crystal a n-type area with doping N_d borders on a p-type area with doping N_a. Such a junction is called a *homojunction*. The concentration gradient will cause diffusion currents: electrons will move from the n-type to the p-type and leave their

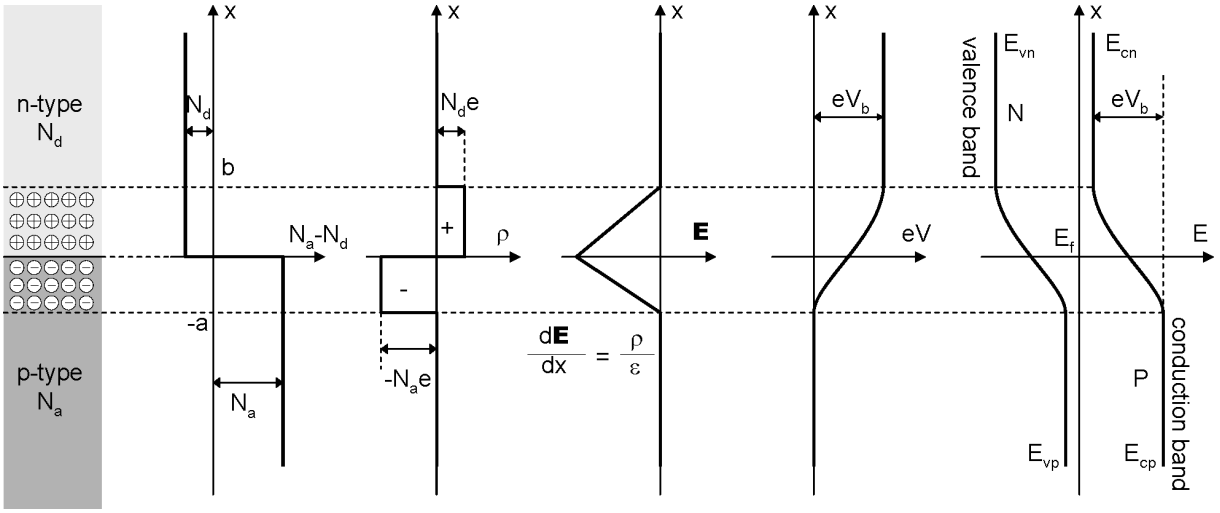

Figure 14.7: The pn-junction: doping, space charge, built-in potential and band-bending.

positively charged donor atoms behind, analogously holes will move from the p-type to the n-type and leave their negatively charged acceptor ions behind. This gives rise to a space charge and an electric field $\mathbf{E}$ that counteracts further diffusion. The situation is sketched in figure 14.7. If no external voltage V_a is applied across the junction, the built-in electric field and the corresponding internal voltage across the junction (built-in potential V_b) will be just so large that the diffusion forces are compensated by the forces caused by the electric field, so that the netto current across the junction is equal to zero. In this situation of equilibrium, the fermi-level is constant in the entire structure. The (fixed) charges to the left and the right of the junction form the *depletion region* or space charge area, in which almost no free charges are present. The depletion layer becomes thicker for decreasing doping levels. The voltage drop across the depletion region implies a drop in the energy levels of the conduction and the valence band. This is called *band-bending*.

The calculation of V_b follows from figure 14.7 with the condition that the fermi-level is constant across the entire structure. The energy of the bottom of the conduction band in the neutral n- and p-areas is denoted as E_{cn} and E_{cp}, analogously for the top of the valence band E_{vn} and E_{vp}. We get:

$$eV_b = E_{cp} - E_{cn} = E_g - (E_{cn} - E_f) - (E_f - E_{vp}) \tag{14.2}$$

Using

$$n = N_c \exp\left[-\frac{E_c - E_f}{k_B T}\right] \tag{14.3}$$

$$p = N_v \exp\left[-\frac{E_f - E_v}{k_B T}\right] \tag{14.4}$$

$$n_i = \sqrt{np} = \sqrt{N_c N_v} \exp\left[-\frac{E_g}{2k_B T}\right] \tag{14.5}$$

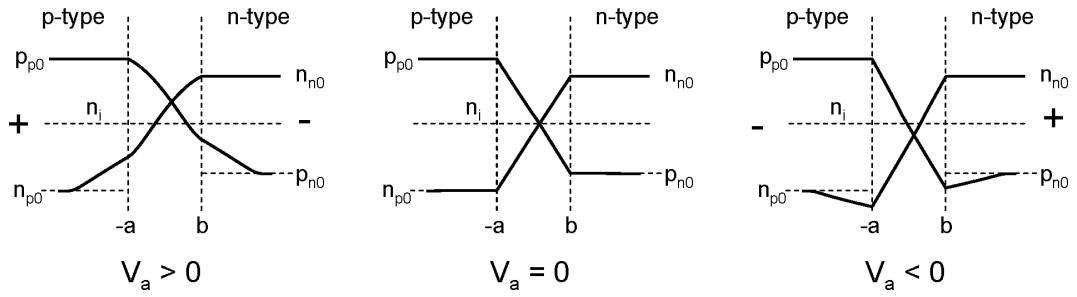

Figure 14.8: Charge carrier density in the pn-junction: forward biased, not biased, backward biased (logarithmic scale).

with

$$N_c = 2\left[\frac{2\pi m_e^* k_B T}{h^2}\right]^{3/2} \tag{14.6}$$

$$N_v = 2\left[\frac{2\pi m_h^* k_B T}{h^2}\right]^{3/2} \tag{14.7}$$

we obtain:

$$V_b = \frac{k_B T}{e} \ln\left(\frac{N_d N_a}{n_i^2}\right) \tag{14.8}$$

For the progress of n and p we get

$$n(x) = n_{n0} \exp\left[\frac{e(V(x) - V_b)}{k_B T}\right] \tag{14.9}$$

$$p(x) = p_{p0} \exp\left[-\frac{eV(x)}{k_B T}\right] \tag{14.10}$$

with n_{n0} the electron concentration in the n-area and p_{p0} the hole concentration in the p-area sufficiently far from the junction. $V(x)$ is the potential. The densities n and p are indeed very small in the depletion region compared to N_d and N_a. The space charge $\rho(x)$ in the depletion region is thus fully dertermined by the density of ionized donors and acceptors. The situation is depicted in figure 14.8 ($V_a = 0$).

The field $\mathbf{E}(x)$ can be calculated with the Poisson equation:

$$\frac{\partial \mathbf{E}(x)}{\partial x} = -\frac{\partial^2 V(x)}{\partial x^2} = \frac{\rho(x)}{\epsilon} \tag{14.11}$$

For the maximal electric field $\mathbf{E}_m$ we get

$$\mathbf{E}_m = -\frac{eN_a a}{\epsilon} = -\frac{eN_d b}{\epsilon} \tag{14.12}$$

The width W of the depletion region can be calculated as

$$W = a + b \tag{14.13}$$

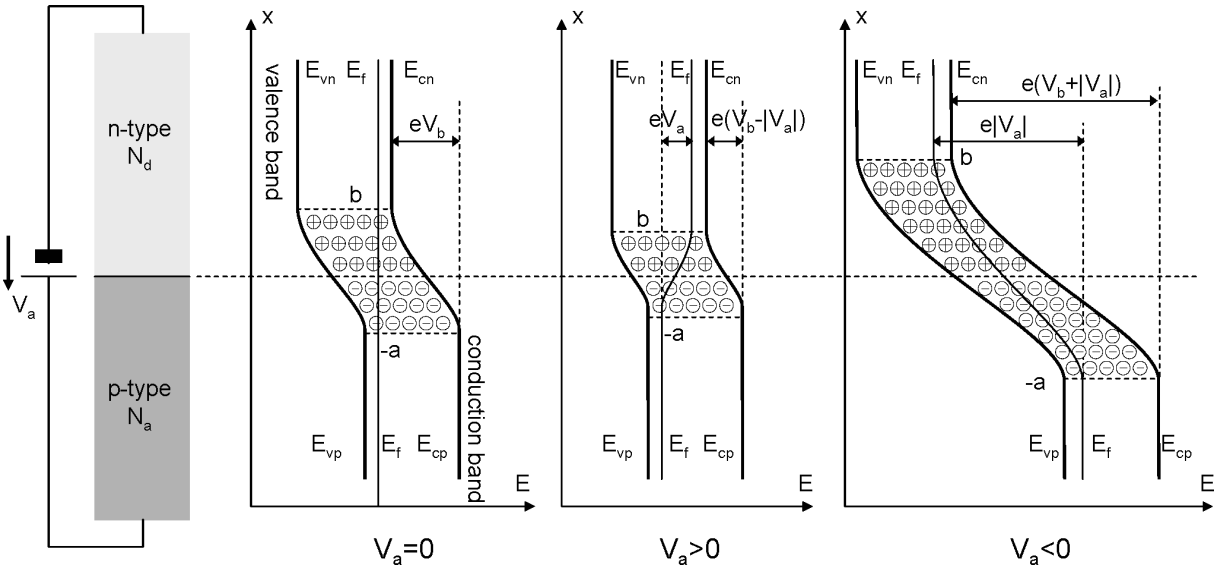

Figure 14.9: The pn-junction with an external voltage V_a.

We get

$$W = \sqrt{\frac{2\epsilon}{e}\frac{N_a + N_d}{N_a N_d} V_b} \tag{14.14}$$

EXAMPLE: For Si with $E_g = 1.12\,\text{eV}$, $E_c - E_d = 0.045\,\text{eV} = E_a - E_v$, $\epsilon = 11.9\,\epsilon_0$, $N_a = 10^{19}\,\text{cm}^{-3}$, $N_d = 10^{15}\,\text{cm}^{-3}$ we calculate $V_b \approx 0.8\,\text{V}$, $W \approx 1\,\mu\text{m}$ and $\mathbf{E}_m = 1.6 \times 10^4\,\text{V/cm}$.

The pn-junction with an external voltage

An applied voltage V_a is defined as positive if it decreases the internal potential barrier, this is if the p-area is positively biased w.r.t. the n-area. The external voltage will be mainly across the depletion region. The fermi-level is now no longer constant everywhere: the fermi-level in the p-type area will be an amount eV_a lower than the fermi-level in the n-type area. The altered band-bending is sketched in figure 14.9[1]. For the calculation of the electric field and the width of the depletion layer, if suffices to replace V_b by $V_b - V_a$ in the results of the previous paragraph.

When no voltage is applied, the netto electron current and netto hole current in the pn-junction is zero. This is obvious for the neutral areas. In the depletion area the diffusion and drift current are equal but opposite. For $V_a > 0$ the potential barrier is decreased and the diffusion current gets the upper hand. As a consequence, the holes diffuse into the neutral area past $x = b$, where they penetrate over a distance L_p. As in this area $np > n_i^2$ applies, the (minority) holes will recombine here. The same happens with the electrons: they penetrate the area $x < -a$ over a distance L_n and recombine there. The distances L_p and L_n are the diffusion lengths of the minority carriers: for holes $L_p = \sqrt{D_p \tau_p}$ applies and for electrons $L_n = \sqrt{D_n \tau_n}$ applies. This length increases when the lifetime of the minority carrier increases and the diffusion coefficient becomes larger. For $V_a < 0$ the opposite happens: minority carriers are extracted instead of being injected over a distance of

[1] In fact, the situation is a bit more complex: for semiconductors that are not in thermal equilibrium we actually have to define *two* fermi-levels: one for the electrons and one for the holes. These are also called *quasi-fermi-levels*

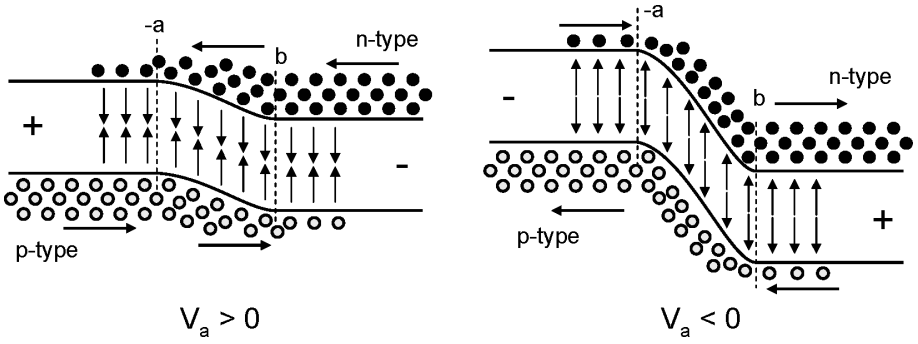

Figure 14.10: Charge carrier transport in a biased pn-junction.

approximately a diffusion length. A graphic representation of these phenomena is given in figure 14.10. The density of the charge carriers is also sketched for both cases in figure 14.8.

If we make a few assumptions, the current-voltage characteristic of the pn-junction can be calculated. For this we start from the continuity equations (one-dimensional):

For the holes in a n-type semiconductor we get:

$$\frac{\partial p}{\partial t} = G_p - \frac{p - p_{n_0}}{\tau_p} - p\,\mu_p \frac{\partial \mathbf{E}}{\partial x} - \mu_p\,\mathbf{E}\frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} \tag{14.15}$$

For the electrons in a p-type semiconductor we get:

$$\frac{\partial n}{\partial t} = G_n - \frac{n - n_{p_0}}{\tau_n} - n\,\mu_n \frac{\partial \mathbf{E}}{\partial x} - \mu_n\,\mathbf{E}\frac{\partial n}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} \tag{14.16}$$

The calculations are left behind. The final result for the current density J is:

$$J = J_S \left[\exp\left(\frac{eV_a}{k_B T}\right) - 1\right] \tag{14.17}$$

in which the *saturation current density* J_S (for backward bias) equals to:

$$J_S = e\left(\frac{D_n n_{p0}}{L_n} + \frac{D_p p_{n0}}{L_p}\right) \tag{14.18}$$

This is the famous Schockley equation (see figure 14.11). We can clearly recognize the diode characteristic. Finally we notice that the total current in both biases is determined by the magnitude of the diffusion current at the edges of the depletion region. In other words: the current in a pn-junction is diffusion-limited.

14.2.2 Heterojunctions and double heterojunctions

A *heterojunction* is a junction consisting of two semiconductors with a different composition. When the semiconducors are the same type, it is called an *isotype* heterojunction, otherwise an *anisotype* heterojunction. Let us look e.g. at the case of a P-n junction. This is a heterojunction of a p-type semiconductor with a n-type semiconductor whereby E_g is the largest in the p-area. A similar

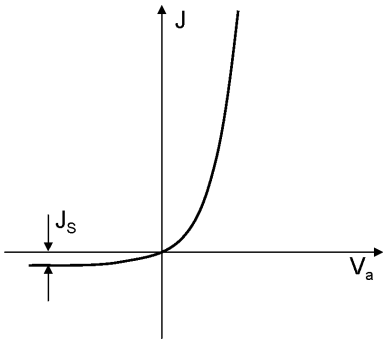

Figure 14.11: The current-voltage characteristic of a pn-junction.

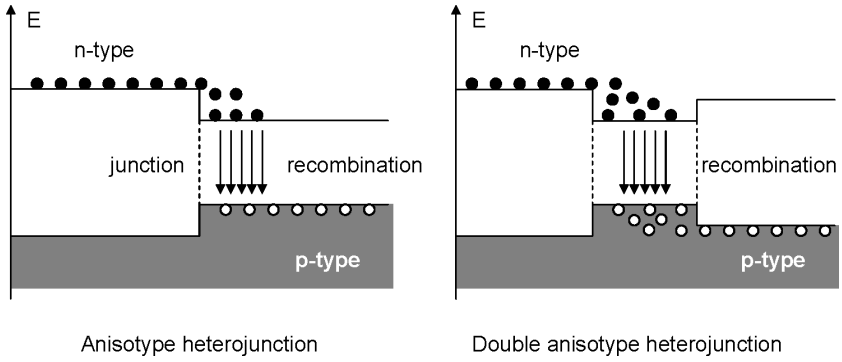

Figure 14.12: Heterojunction and double heterojunction.

band-bending occurs as in the pn-homojunction. At the boundary plane a discontinuity however occurs. The situation for forward bias is depicted in figure 14.12, where we have left the band-bending behind. A heterojunction has an extra degree of freedom when designing a component. In a regular pn-junction, forward biased, there is an electron as well as a hole current through the junction. The relative magnitude of these two currents is determined by the relative doping levels. In a heterojunction the current will mainly consist of charges coming out of the material with the highest bandgap into the one with the lowest bandgap, independent of the doping. A very important structure is the double heterojunction in which a layer with a low bandgap is placed between two layers with a large bandgap. On the right in figure 14.12 the basic band diagram of a double heterojunction is depicted. Such a structure forms a potential well for the electrons and the holes. Most of the semiconductor lasers are based on the charge confinement in such a potential well. This can for example be realized with a N-p-P heterojunction. When forward biased, electrons are brought out of the N-area and holes out of the P-area. The confinement of a large electron and hole density in the p-layer leads to population inversion and the recombination results in laser emission. Furthermore, the material with a lower bandgap usually has a higher refractive index. The structure thus acts as a waveguide in which the photons are trapped by total internal reflection.

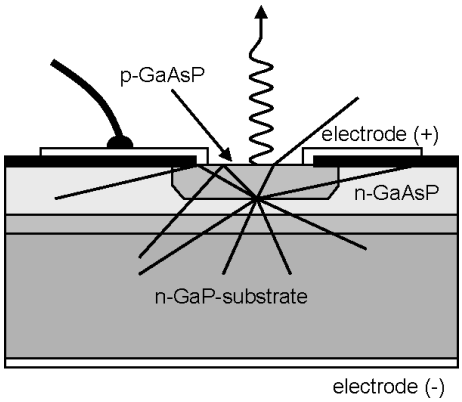

Figure 14.13: Surface-emitting LED.

14.3 Light emitting diodes

14.3.1 Electroluminescence

We already know that electron-hole recombination in a semiconductor can cause light emission. In a semiconductor at thermal equilibrium, the concentration of electrons and holes is small so that the light emission is very small too. We can however strongly increase photon emission by bringing the semiconductor out of thermal equilibrium using an external source of electron-hole pairs. This can be done e.g. by illuminating the material, but it is usually caused by forward-biasing a pn-junction. In that case the holes diffuse from the p-area to the n-area, and the electrons diffuse from the n-area to the p-area, respectively. This light source is called a *light emitting diode (LED)* and the generation of light is called *electroluminescence*.

The rate of photon emission can be calculated using the rate G by which the electron-hole pairs are injected. The photon flux generated per unit volume is proportional to G. In static conditions, $G = \Delta n/\tau$ has to apply, with Δn the surplus of electron-hole pairs and τ the lifetime. As only radiative recombinations produce photons, we have to introduce an *internal quantum efficiency* η_i: $\eta_i = U_r/U = \tau/\tau_r$. The photon flux Φ_i generated in a volume V then becomes:

$$\Phi_i = \eta_i G V = \eta_i \frac{V\Delta n}{\tau} = \frac{V\Delta n}{\tau_r} \tag{14.19}$$

The efficiency of a LED is strongly dependent on η_i. As η_i is a lot larger for direct semiconductors than for indirect semiconductors, LEDs and lasers are usually made of direct semiconductors.

14.3.2 LED-characteristics

Efficiency

A basic problem in LED's is that the generated photons are not easily extracted from the semiconductor material. Let us consider for example the case of a planar surface-emitting structure, as shown in figure 14.13. Internally the light is emitted isotropically in all directions. 50% is lost due

to emission to the substrate. For the other 50%, not radiated towards the substrate, a large part is lost due to total internal reflections at the semiconductor-air interface. Indeed, light rays can only escape to the air if the angle between the ray and the normal of the surface is smaller than the critical angle for total internal reflection, which is approximately 17^o for III-V semiconductor like GaAs. Furthermore, a part of the light is lost because of the reflection on the upper electrode (which is usually ring-shaped). The overall extraction efficiency is typically smaller than 1%. The (external) radiation characteristic of this kind of LED is approximately Lambertian. This is the result of an internally isotropic light distribution combined with refraction at a plane surface.

Thus, next to the internal quantum efficiency, we have to introduce an extraction efficiency η_e. The photon flux Φ_o leaving the LED, is:

$$\Phi_o = \eta_e \Phi_i \tag{14.20}$$

The low efficiency can be countered by curving the semiconductor-air interface so that as many incident rays as possible are approximately perpendicular to the interface. This is however very hard to realize technologically. Therefore, the LED is often integrated in another material (with a refractive index as high as possible) in which a curved interface can easily be made. This curved interface can even act as a collimating lens. Many display-LEDs are manufactured in this way.

Φ_o can also be written as

$$\Phi_o = \eta_e \eta_i \frac{i}{e} = \eta_{ex} \frac{i}{e} \tag{14.21}$$

with i the current through the pn-junction and η_{ex} the *external quantum efficiency*.

The optical power P_0 of the LED then equals to

$$P_0 = h\nu\Phi_0 = \eta_{ex} h\nu \frac{i}{e} \tag{14.22}$$

Modulation bandwidth

A last important aspect of LEDs concerns the modulation bandwidth. The modulation bandwidth can be defined by considering the sinusoidal variation of the drive current of the LED with a frequency ω around a fixed bias drive current. This results in a sinusoidal modulation of the optical output power, with an amplitude depending on the modulation frequency. For LEDs, this amplitude decreases monotonically with increasing frequency of the drive current. Typically, the modulation bandwidth is defined as the frequency of the drive current for which the amplitude of the modulated optical output has decreased by a factor of 2 compared to the case of a low frequency drive current (3dB bandwidth).

This modulation bandwidth is primarily determined by the lifetime τ of the injected minority carriers, that recombine radiatively. For a sufficiently low injection, the transformation of current variation to light variation is linear, corresponding to a first-order transfer characteristic:

$$R(f) = \frac{\Delta P}{\Delta I} = \frac{R(0)}{\sqrt{1 + 4\pi^2 f^2 \tau^2}} \tag{14.23}$$

Here f is the frequency of sinusoidal modulation of the diode current (around a static working point, so that the total current always remains positive), ΔI is the amplitude of this modulation

and ΔP is the amplitude of the resulting sinusoidal variation in optical power. Thus, the 3dB-bandwidth becomes

$$f_{3dB} = \frac{1}{2\pi\tau} \tag{14.24}$$

In III-V semiconductors the lifetime is typically a few ns, so that the bandwidth is about 50 to 100 MHz. Some LED-types have bandwidths up to 1 GHz.

14.3.3 LED-types

Besides the previous case of the isotype pn-junction, a LED can also consist of an anisotype heterojunction. In this case, recombination primarily occurs at the side with the smallest bandgap. A third kind of device is the anisotype double heterojunction. Here the middle layer (low bandgap) is filled with electrons from the n-layer and with holes from the p-layer, which causes recombination in this middle layer. At static equilibrium, the electron- and hole-concentration is just large enough so that the input of charges is compensated by recombination. This concentration is sometimes higher than the doping concentration. Due to neutrality:

$$n \approx p \tag{14.25}$$

So the electron-hole spontaneous emission recombination rate U becomes

$$U \approx Bnp \approx Bn^2 \approx Bp^2 \tag{14.26}$$

The lifetime is then inversely proportional to the concentration and therefore dependent on the current. The LED emits light at a photon energy that is approximately equal to the bandgap of the material. In order to have different colors, different semiconductors need to be used. In the following table, the most common types are given:

λ [nm]	Color	Material	Application
1000-1600	Infrared	In_xGa_{1-x} As_yP_{1-y}	Optical fiber communication
850-900	Infrared	GaAs	Idem + wireless communication
650	Red	$GaAs_{60}P_{40}$ of InGaP	Displays
620	Orange	$GaAs_{35}P_{65}$:N	Displays
590	Yellow	$GaAs_{15}P_{85}$:N	Displays
570	Green	GaP:N	Displays
400-500	Blue	SiC, II-VI SC, InGaN	Displays

$GaAs_{1-x}P_x$ is by far the most used material for visible LEDs. For $x > 45\%$ it is an indirect semiconductor however. This problem is solved by an isoelectronic doping with nitrogen (GaAsP:N). Here the nitrogen atom (also a group V atom) replaces a phosphor atom which results in new energy levels close to the conduction band, that act as centers of recombination. The internal efficiency of good infrared and red LEDs lies close to 100%. For other colors, particularly green and blue, the efficiency is a lot lower. However, progress is still being made by employing new material technologies.

Besides inorganic semiconductors, more and more organic semiconductors are being used for LEDs. These plastic LEDs are called OLEDs. Their performance is for now far below that of inorganic LEDs.

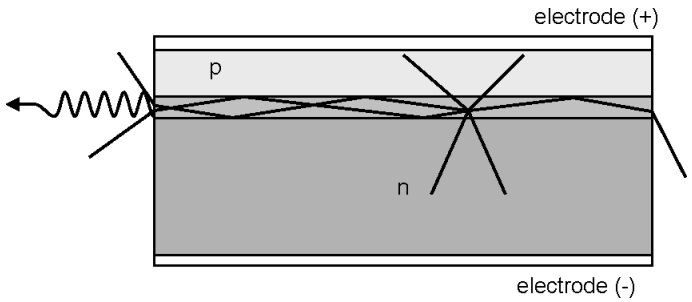

Figure 14.14: Edge-emitting LED.

A completely different kind of LED is the edge-emitting LED (see figure 14.14). Here, the LED is fitted into a waveguide structure (the double heterostructure can fulfill this function in one direction). Part of the light is guided through the waveguide to the edge of the chip, where it is emitted into the air. The structure looks a lot like the laser diode (see further), with one major difference: there may not be a cavity present. The reflections at the end-mirrors are thus suppressed. The extraction efficiency of the edge-emitting LED is equal to that of the surface-emitting LED. There is however an important difference in radiance (luminance). The emitting surface of a surface-emitting LED is usually a lot larger than for an edge-emitting LED, where it is determined by the waveguide dimensions. Because of this, the radiance of the edge-emitting LED is, at equal power, larger than the radiance of the surface-emitting LED. This is e.g. important for imaging to a little spot (e.g. for use in optical fibers). There is also a difference in the spectrum. The spectrum of a surface-emitting LED is approximately equal to the spectrum of internal spontaneous emission. In an edge-emitting LED however, light travels a certain distance through the light-generating material. Therefore shorter wavelengths in the spectrum are absorbed (and thus the spectrum is narrower). Finally, an edge-emitting LED can also be used as a superluminescent LED. For this purpose the current is chosen so large that stimulated emission becomes more important than absorption. The spontaneously emitted light is then amplified while propagating in the waveguide. This narrows the spectrum and increases the efficiency.

14.3.4 Applications

LEDs have many applications. Indicator lights on all sorts of devices are the most common application. Green, orange (amber) and red LEDs are commonly used here. Until now the use of LEDs in displays had been limited (except for the very large displays in stadia e.g.), because blue LEDs were expensive and had a low efficiency. However, the recent development of the blue LEDs has made a breakthrough because of the use of GaN.

Another application that is gaining ground is the 'LED-lights'. Nowadays, many red LEDs are already used in car lights and street signs, instead of the classic incandescent lamps, mainly because of their longer lifetime. There are also LEDs that emit white light. These are actually blue (or UV) LEDs, covered with a phosphor layer in which the highly energetic blue light is converted to white light (like in a TL-tube). Infrared LEDs are used for 'invisible lighting' (security), and especially for transmission of information. The latter application is found e.g. in remote controls and optical fiber connections (mainly short range connections). Another use is the opto-coupler

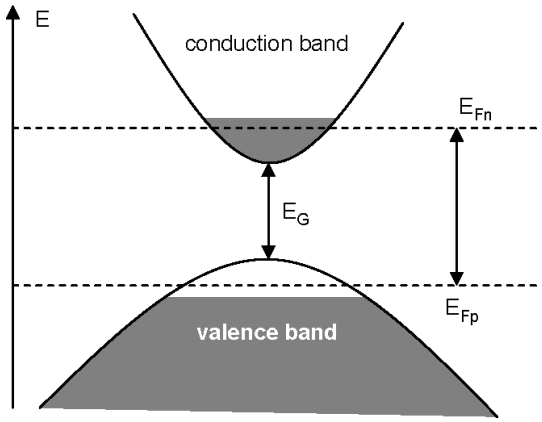

Figure 14.15: Population inversion in a semiconductor.

(or opto-isolator) in which a LED and a photodetector are joined in a closed packaging in order to have a galvanic separated information connection.

Compared to the laser diode, the LED has a lot of shortcomings: low efficiency, low power, low radiance, low modulation bandwidth and a broad spectrum. On the other hand, LEDs are less sensitive to temperature, cheaper and more reliable. In addition, the low temporal coherence is an advantage in a number of applications, as this suppresses the sensitivity to interferometric disturbances.

14.4 Laser diodes

After the demonstration of the first ruby laser in 1960, it quickly became clear that semiconductor material would also make lasing possible, with the major advantage that the 'pump' would be an electric current. In 1962, three different research groups, independent of each other, showed lasing in semiconductor diodes. Today the laser diode has become an indispensable component in lots of applications. As main applications we mention optical fiber communication and data recording on compact disks.

14.4.1 Amplification, feedback and laser oscillation

The structure of a semiconductor laser diode resembles that of the LED. In both cases photons are generated because of an electric injection in a pn-junction. The emitted light of a laser diode originates however from stimulated emission instead of spontaneous emission, as is the case in a LED. The optical feedback necessary for laser oscillation is obtained by mirrors, typically formed by cleaving the semiconductor wafer.

Amplification

When a semiconductor is pumped to population inversion, optical amplification can occur with a peak value g_p given by

$$g_p = \alpha \left(\frac{J}{J_T} - 1 \right) \tag{14.27}$$

Here J is the injected current density, J_T the current density for transparence and α the absorption when there is no current injection. In order to reach population inversion, the conduction band in the semiconductor has to be strongly occupied and the valence band relatively empty. In other words, a large concentration of free electrons and free holes are needed. It can be proven that population inversion occurs only if the energy distance between the quasi-Fermi levels are larger than the bandgap (condition of Bernard and Durrafourg). The photon energy of the gain maximum then lies between both values (see figure 14.15). In practice we have to keep in mind that the photon energy of the emitted light is approximately equal to the bandgap. Typically an electron and hole concentration of the order 2×10^{18} cm^{-3} is needed for the material to be transparent. This is a very high concentration and would be difficult to achieve in a large volume of semiconductor due to thermal reasons. The active volume in a laser is therefore made as small as possible.

Feedback

Feedback is usually obtained by two cleaved surfaces perpendicular to the plane of the junction. The reflection on the surfaces causes the active area of the pn-junction to function as an optical resonator. The power reflectance R is given by

$$R = \left(\frac{n-1}{n+1} \right)^2, \tag{14.28}$$

with n the refractive index of the semiconductor. For GaAs e.g. $n = 3.6$ so that $R = 0.32$.

Resonator losses

The partial reflection at the cleaved surfaces enables photons to leak out. These photons are emitted as usable laser light. For a resonator with length L, these loss terms can be expressed as a loss α_m per unit length in the resonator

$$\alpha_m = \frac{1}{2L} \ln \frac{1}{R_1 R_2} \tag{14.29}$$

The total losses α_r contain another term α_s due to scattering at optical irregularities. We can write

$$\alpha_r = \alpha_s + \alpha_m \tag{14.30}$$

Laser resonance

The amplitude condition for laser resonance is given by $g_p = \alpha_r$. Using (14.27) we find the threshold for the current density

$$J_{th} = \frac{\alpha_r + \alpha}{\alpha} J_T \tag{14.31}$$

The threshold current density J_{th} is $(\alpha_r + \alpha)/\alpha$ times larger than the transparency current density J_T due to resonator losses. J_{th} is an important parameter concerning the performance of a laser diode: smaller values of J_{th} mean a better laser diode.

14.4.2 Laser diode characteristics

When $J > J_{th}$ laser oscillation begins and the photon flux Φ is built up in the resonator. For an internal photon flux Φ, we can write

$$\Phi = \begin{cases} \eta_i \frac{i - i_{th}}{e}, & i > i_{th} \\ 0, & i < i_{th} \end{cases} \tag{14.32}$$

with $i = JA$ the current flowing through the junction with surface area A. The internal laser power P is then given by

$$P = \eta_i (i - i_{th}) \frac{h\nu}{e} \tag{14.33}$$

The photon flux Φ_o leaving the resonator is the product of the internal photon flux and the emission efficiency η_e

$$\Phi_o = \eta_e \eta_i \frac{i - i_{th}}{e} \tag{14.34}$$

If the light coming out of both mirrors is used then $\eta_e = \alpha_m/\alpha_r$. For mirrors with identical reflectance R we get

$$\eta_e = \frac{1}{\alpha_r L} \ln \frac{1}{R} \tag{14.35}$$

The emitted laser power is then given by

$$P_o = \eta_d (i - i_{th}) \frac{h\nu}{e} \tag{14.36}$$

in which $\eta_d = \eta_e \eta_i$ is the *external differential quantum efficiency*. The emitted laser power P_o in function of the injection current i is shown in figure 14.19.

The *differential responsivity* (in W/A) $\mathcal{R}_d$ can be defined as

$$\mathcal{R}_d = \frac{dP_o}{di} = \eta_d \frac{h\nu}{e} \tag{14.37}$$

Finally, the *overall efficiency* η is defined as the ratio of the emitted laser power to the electrical input power iV and is given by

$$\eta = \eta_d \left(1 - \frac{i_{th}}{i}\right) \frac{h\nu}{eV} \tag{14.38}$$

We can now easily deduce that mirrors with a high reflectivity cause low mirror losses and thus a low threshold current, but also a low extraction efficiency. In practice, there is no point in making the mirror losses α_m smaller than the other losses α_s.

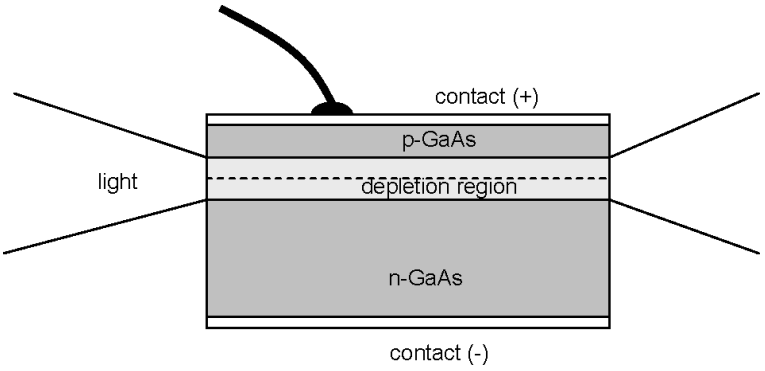

Figure 14.16: Semiconductor laser diode of the first generation.

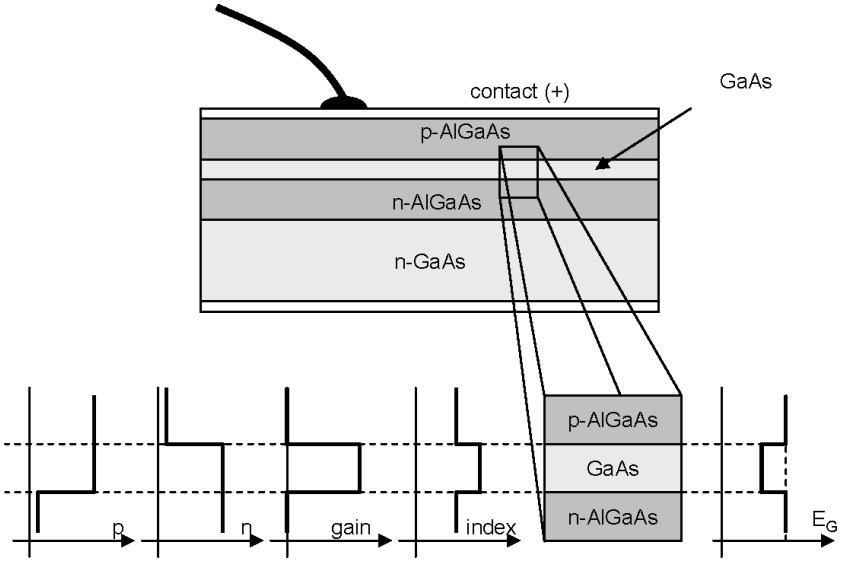

Figure 14.17: Double heterostructure laser.

14.4.3 Laser diode types

The first semiconductor lasers consisted of a simple forward-biased pn-junction (see figure 14.16). The threshold current density J_{th} was very high however ($> 10\,\mathrm{kA/\,cm^2}$) and the efficiency was low so that only pulsed operation was possible. We mention several reasons. First, the minority carriers spread out on both sides of the junction because of diffusion, so that a very large current density is needed to obtain a sufficiently high charge carrier concentration. Second, light amplification is only obtained over a small area with a thickness of a few μm. The light generated in this amplification layer will therefore diffract quickly and leave this layer. In addition, because one uses the plane walls of the cleaved crystal as mirrors, the diverging light does not bend into convergence. All this means that the cavity shows high losses, and the material has to be pumped far above transparence.

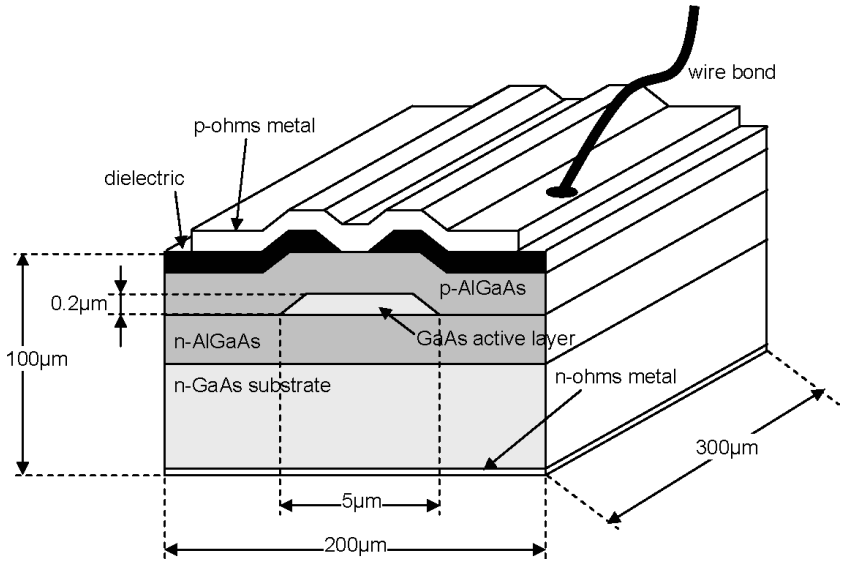

Figure 14.18: Typical dimensions of a laser diode.

The double heterostructure laser

The double heterostructure laser (see figure 14.17) brought a solution to all these problems. Here a thin active layer of GaAs e.g. (typically $0.2\,\mu$m) is surrounded by other layers ('cladding' layers) consisting of another material (e.g. AlGaAs) with a larger bandgap. One side is p-type doped and the other side n-type. When the junction is forward-biased, high concentrations of electrons as well as holes are created in this thin middle layer. The charge carriers can not spread out because of diffusion, as they are trapped between the potential barrier of the higher bandgap on both sides of the active layer. Due to this charge confinement it is possible to achieve transparency at a much lower current density (typically $500\,\mathrm{A/cm^2}$). Optically the situation is also very different. Usually semiconductors with a higher bandgap have a lower refractive index. Thus, the active layer is confined between two layers with a lower refractive index. This forms a waveguide. So the photons generated by stimulated emission do not spread out because of diffraction, but they are 'locked inside' the waveguide due to total internal reflection (optical confinement). This strongly reduces the cavity losses.

Until now we have only discussed what happens in the transversal direction (perpendicular to the double heterostructure) and the longitudinal cavity direction. In the third direction (lateral direction) we also try to confine the electrons and holes as well as the photons, just as in the transversal direction. To this end different techniques are used that we will not discuss here. However, it is important to keep in mind that a typical laser with active layer dimensions of $0.2\,\mu$m thick $\times$ $5\,\mu$m wide $\times$ $300\,\mu$m long has a threshold current of approximately 10 to 20 mA. A three-dimensional sketch of a laser diode is depicted in figure 14.18.

The light propagating in the cavity can no longer be described with Gaussian beams (chapter 5). There, free diffraction in a uniform medium was assumed, while here we are using a waveguide. When solving Maxwell's equations, similar results are obtained however, i.e. a number of modi with a longitudinal, lateral and transversal mode number (see chapter 7). Only one lateral/transversal mode is desired in practice, and this is obtained by correctly choosing the dimen-

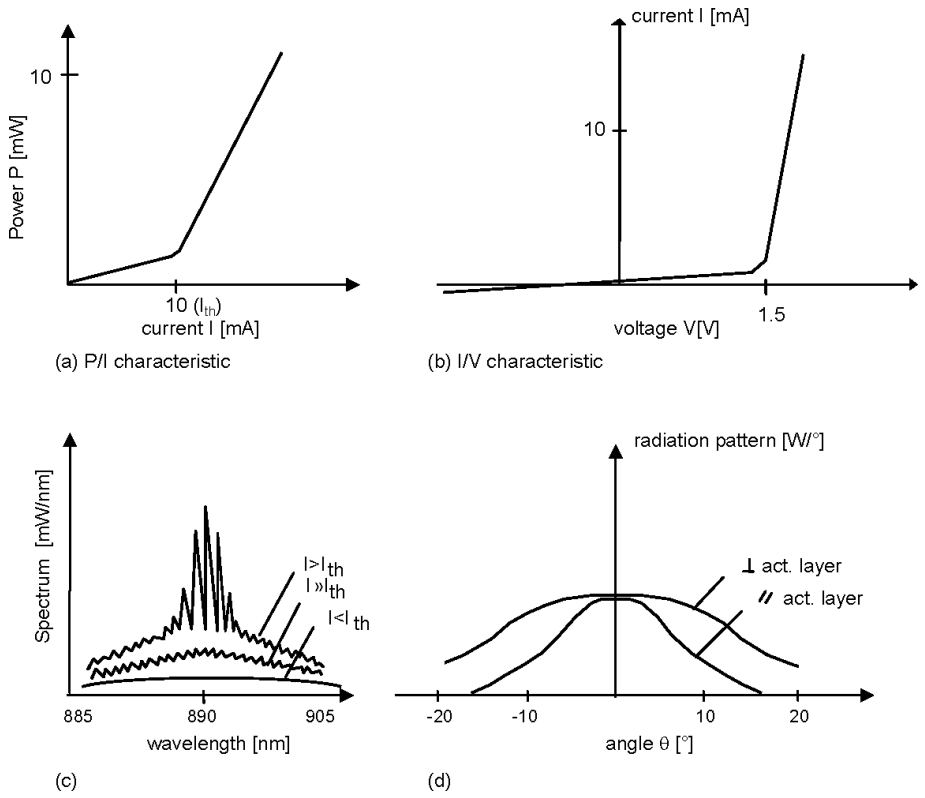

Figure 14.19: Characteristics of a laser diode.

sions and differences between the refractive indices. The suppression of multiple longitudinal modes is more difficult however. The gain spectrum of the material is broad as the semiconductor has a band structure instead of discrete levels. Despite the large mode spacing because of the short cavity (typically 0.3 mm long), multiple modes arise easily. Special structures have to be used in order to suppress the side modes. To this extent a strongly filtering object is brought inside the cavity. If this object is tunable, the laser can emit light at each wavelength in the gain spectrum. In figure 14.19 a number of typical characteristics are shown of a GaAs-AlGaAs laser diode. The far-field radiation pattern of a laser diode is usually strongly divergent. If a laser has only one lateral/transversal mode, the field profile is clock-shaped and behaves roughly as a Gaussian beam. Thus, the divergence angle is inversely proportional to the bundle width. As the bundle usually has an elliptical shape (due to the rectangular shape of the active layer cross section), the far field will also be elliptical. In the direction in which the rectangular waveguide cross section is the narrowest the far field pattern will be the broadest (and vice versa).

In this example we used a GaAs/AlGaAs combination. The III-V semiconductors (consisting on the one hand of one or more elements of group III (Al, Ga, In) and on the other hand of one or more elements of group V (P, As, Sb)) offer a variety of possibilities. This is shown in figure 14.2 in, where the bandgap and crystal lattice constant are denoted for most of the combinations. The points denote binary combinations, the lines ternary combinations (these are actually certain mixtures of two binary combinations) and the planes between the lines denote quaternary combinations. In order to make a laser with a certain emission wavelength, we have to use the right semiconductor with the proper bandgap (=photon energy) for the active layer. Furthermore, for

the cladding layers a semiconductor has to be used with a larger bandgap (and thus a smaller refractive index).

There is however a technological restriction. In order to fabricate layers with proper crystalline quality, all layers and the substrate need to have the same lattice constant. The used materials therefore have to be located along a line on the diagram. Therefore the number of appropriate substrates is limited. The substrates always consist of a binary combination, with GaAs and InP the most important examples. We can easily deduce from the figure that with a GaAs substrate, it is possible to create lasers with emission wavelengths between 700 and 900 nm. With an InP substrate the emission wavelengths lie between 900 and 1600 nm. The first are used ubiquitously for optical data recording (e.g. compact disk) and optical fiber communication at short range, while the latter is very important for optical fiber communication at long range (especially at 1.3 and 1.55 μm).

It took a while to fabricate semiconductor lasers that emit visible light. Since 1988 however, red laser diodes are commercially available. These lasers have an active layer of InGaP and cladding layers of InAlGaP on a GaAs substrate. The wavelength is approximately 650 nm. Applications are for example replacement of the He-Ne lasers, barcode readers, new generations of CD systems, etc.

14.4.4 Comparison laser diodes and other lasers

Let us finally list the important differences between semiconductor lasers and most of the other lasers:

- energy bands instead of discrete levels
- cavity with waveguide and plane mirrors instead of free diffraction and spherical mirrors
- very small dimensions
- emitted light bundle can be diffraction limited (this means: good spatial coherence), but still have a large divergence angle (e.g. 20^o) due to the small dimensions of the field in the waveguide. A lens is required in order to obtain a collimated bundle.
- the spectrum can contain different longitudinal modes (this means bad temporal coherence)

The main advantages of the laser diodes are:

- very compact packaging, comparable with electronic components
- simple pump system with low voltages and currents
- possibility of modulation (due to current variations) with a large bandwidth (a few GHz)
- high efficiency (10 to 50 %)
- broad range of usable wavelengths (naturally not in the same component)
- tunable in wavelength by varying the temperature or by integrating tunable filters inside the cavity

Bibliography

Chapter 15

Semiconductor Detectors

Contents

15.1 Introduction

In a photodetector optical power is converted into something measurable, usually an electric current. There are generally two types of photodetectors: thermal detectors and photoelectric detectors.

Thermal detectors (bolometers)

In thermal detectors photons are converted into heat, and the resulting change in temperature is detected by measuring the resistance of a temperature sensitive resistor. Most of the thermal detectors are inefficient and relatively slow because of the large time constant when there is a change in temperature. Therefore, they are not suitable for the majority of applications.

Photoelectric detectors

The principle of photoelectric detectors is based on the photoeffect or photoelectric effect. Absorption of a photon in certain materials results in the generation of mobile charge carriers. When an electric field is applied, they cause a measurable electric current.

In this chapter we will discuss photoelectric detectors more thoroughly.

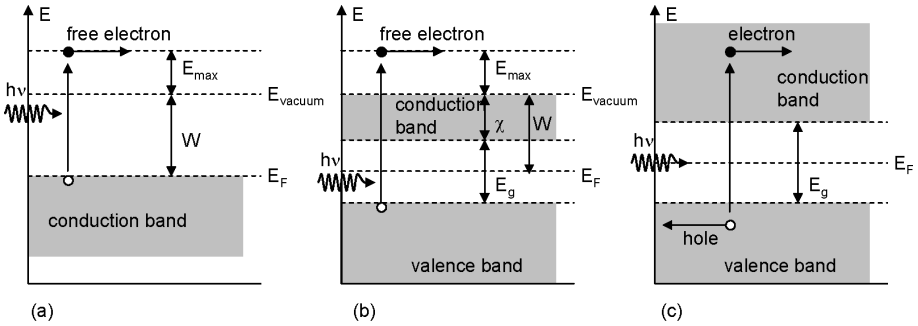

Figure 15.1: The photoeffect: (a) External photoeffect in metals. (b) External photoeffect in semiconductors. (c) Internal photoeffect in semiconductors.

15.1.1 The photoeffect

There are two kinds of the photoeffect: the *external* and the *internal* photoeffect. In the external photoeffect the generated electrons escape from the material and are called free electrons. This is also called *photoelectron emission*. In the internal photoeffect the generated free charge carriers remain inside the material and they increase the conductivity. This process is also known as *photoconductivity* and occurs in nearly all semiconductors.

External photoeffect

The principle is depicted in figure 15.1. A photon with energy $h\nu$ incident on the metal releases an electron from a half-filled conduction band (figure 15.1a). Due to the conservation of energy, the maximal energy of the free electron is

$$E_{max} = h\nu - W \tag{15.1}$$

where W is the energy difference between the vacuum level and the Fermi level of the metal. W is also called the *work function* of the metal. Free electrons originating from levels below the Fermi level have a lower energy. The lowest work function of a metal is approximately 2 eV, so that photoemission detectors based on metals are only applicable in the visual and ultraviolet spectrum. Photoelectric emission is even possible in semiconductors (figure 15.1b). In that case the free electrons mainly originate from the valence band and have a maximal energy

$$E_{max} = h\nu - (E_g + \chi) \tag{15.2}$$

with χ the electron affinity of the material ($\chi = E_{vac} - E_c$) and E_g the bandgap. $E_g + \chi$ lies minimally around 1.4 eV, so that photoemission detectors based on semiconductors are also applicable in the near-infrared.

Photodetectors based on photoelectric emission are usually built in the form of vacuum tubes, also called photomultiplier tubes (see figure 15.2). Here electrons are emitted from the surface of the cathode and move towards the anode, which is kept at a higher electric potential. Because of this, an electric current arises proportional to the photon flux. The emitted electrons gain kinetic energy as they travel through the electric field and can impact metals or semiconductors placed in the tube, the so-called *dynodes*, resulting in the release of multiple secondary electrons. This causes an amplification of the generated electric current.

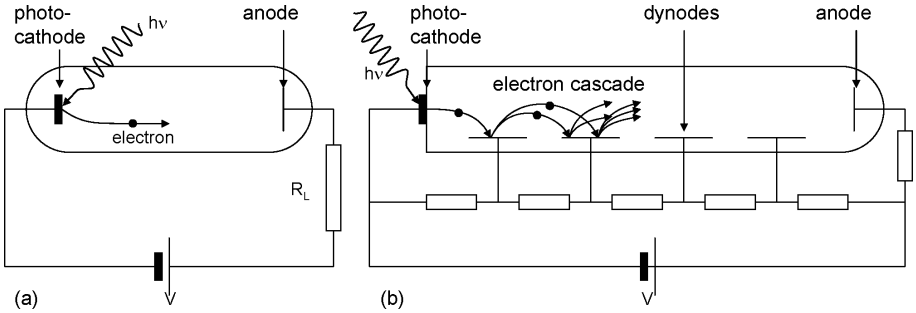

Figure 15.2: Schematic representation of a photodetector based on photoelectric emission: the photomultiplier tube.

Internal photoeffect

When absorbing a photon with an energy $h\nu$ an electron in the conduction band and a hole in the valence band arises (figure 15.1c). When an electric field is applied, the electron and the hole move in opposite directions through the semiconductor, which causes an electric current in the electric circuit of the detector.

The photodiode

The *photodiode* consists of a pn-junction and is based on the internal photoeffect. Charge carriers are created by photons that are absorbed in the depletion layer of the junction. These charge carriers are subjected to the local electric field, which causes an electric current to flow. Some of the photodiodes have internal gain mechanisms that physically amplify the current in the semiconductor to improve detection. If the electric field in the depletion layer becomes sufficiently high, electrons and holes can acquire enough energy to create other electrons and holes due to impact ionization. This situation can be obtained by applying a sufficiently large reverse bias across the junction. This type of photodiodes are called *avalanche photodiodes* (APD).

In summary, the following processes can be distinguished in a semiconductor photodiode:

- Generation: absorbed photons generate free charge carriers.
- Transport: an applied electric field drains the electrons and holes away, causing an electric current.
- Gain: In APDs internal gain occurs because of impact ionization.

15.1.2 Quantum efficiency

The *quantum efficiency* η of a photodetector is defined as the probability that a single photon incident on the detector creates an electron-hole pair that contributes to the electric detector current, as not every incident photon contributes to this current. A part of the photon flux is reflected at the surface of the detector. Furthermore, light intensity decreases exponentially inside the semiconductor. This means that not every photon will be absorbed in a photodetector with a limited

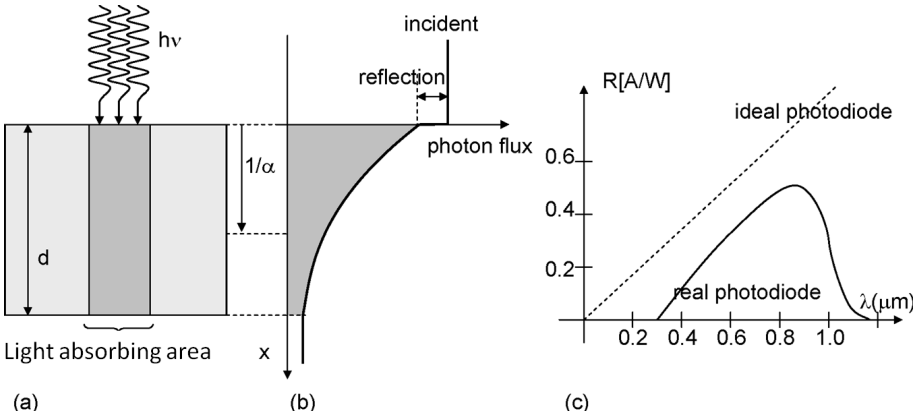

Figure 15.3: (a) and (b) Efficiency and (c) responsivity of a photodetector

thickness d (see figure 15.3a and 15.3b). Finally, recombination can occur at the surface of the photodetector caused by a high concentration of recombination centers. These charge carriers will also not contribute to the photoelectric current.

Therefore η ($0 \leq \eta \leq 1$) is written as:

$$\eta = (1 - R)\zeta[1 - \exp(-\alpha d)] \tag{15.3}$$

with R the optical power reflectance, ζ the fraction of electron-hole pairs contributing to the detector current and α the absorption coefficient. R can be reduced by covering the surface with an antireflection coating. ζ can be optimized by careful material growth, and the exponential factor can be reduced by making the photodiode sufficiently thick. η will also be a function of the wavelength as α is wavelength-dependent. For large wavelengths, when $h\nu = hc/\lambda < E_g$, η becomes very small because of the very low absorption. However, for sufficiently short wavelengths, most of the light is absorbed near the surface of the photodetector, but then recombination gets the upper hand so η decreases.

15.1.3 Responsivity

The responsivity $\mathcal{R}$ is the ratio of the detector current to the incident optical power. If each incident photon would produce a photoelectron, a photon flux Φ would create an electron flux Φ. In a closed circuit, this results in an electric current $i_f = e\Phi$. An optical power $P = h\nu\Phi$ would then result in a current $i_f = eP/h\nu$. Because only a fraction η of the incident photons contributes to the electron current, we get

$$i_f = \eta e\Phi = \frac{\eta eP}{h\nu} = \mathcal{R}P \tag{15.4}$$

The responsivity $\mathcal{R}$ can thus be written as

$$\mathcal{R} = \frac{\eta e}{h\nu} = \eta\frac{\lambda}{1.24} \tag{15.5}$$

This relation is shown schematically in figure 15.3c. If one does not take the wavelength dependence of η into consideration, $\mathcal{R}$ is a linear function of the wavelength. This is easily understood because of the very large photon energy at small wavelengths. When absorption of such a highly

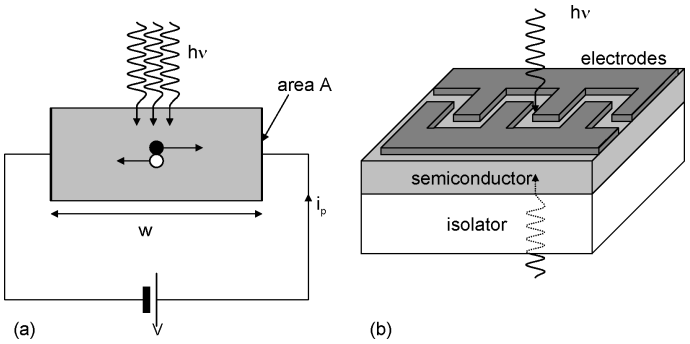

Figure 15.4: Working principle and schematic representation of a photoconductor detector.

energetic photon occurs, the electron is excited from the valence band to a higher energy level in the conduction band, where it will relax to the bottom of the conduction band. The released energy is lost.

In detectors with a gain G, the responsivity has to be multiplied with the factor G. In this case the quantum efficiency can be larger than 1.

15.2 The photoconductor

In a photoconductor detector the photon flux is determined by measuring the photoconductivity. When an external electric field is applied to an illuminated semiconductor, mobile charge carriers create an electric current in the detector circuit. Photoconductor detectors detect either the photocurrent, which is proportional to the photon flux, or the voltage drop across a load resistor in the circuit. The detector consists of a layer of semiconductor material, and usually the cathode as well as the anode is attached to the same side of the surface. The distance between the cathode and the anode has to be optimized in order to maximize light transmission on the one hand and minimize the transit time of the charge carriers on the other hand.

The increase in conductivity caused by a photon flux Φ (number of photons per second incident on a volume wA, see figure 15.4a) is calculated in the following way. A fraction η of the photon flux is absorbed and generates electron-hole pairs. The pair-production rate G_L (per unit volume) is thus $G_L = \eta\Phi/wA$. If τ is the lifetime of these additional charge carriers, recombination will take place at a rate $U = \Delta n/\tau$, with Δn the photoelectron concentration. In static conditions we obtain

$$\Delta n = \frac{\eta\tau\Phi}{wA} \tag{15.6}$$

This results in an increase of the conductivity of

$$\Delta\sigma = e\Delta n(\mu_e + \mu_h) = \frac{e\eta\tau(\mu_e + \mu_h)}{wA}\Phi \tag{15.7}$$

This increase is in fact proportional to the photon flux.

As the current density is given by $J_f = \Delta\sigma E$ and $v_e = \mu_e E$ and $v_h = \mu_h E$, with E the electric field, we can write

$$J_f = \frac{e\eta\tau(v_e + v_h)}{wA}\Phi \tag{15.8}$$

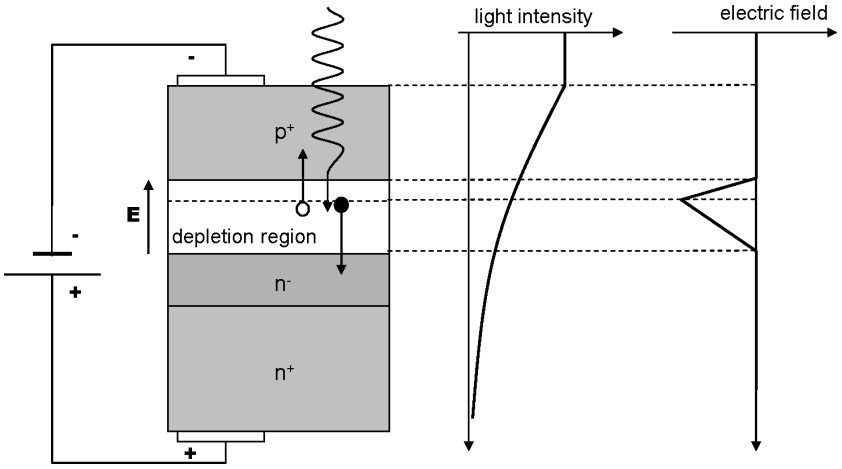

Figure 15.5: Working principle of a photodiode.

and

$$i_f = AJ_f = \frac{e\eta\tau(v_e + v_h)}{w}\Phi \tag{15.9}$$

As usually $v_h \ll v_e$, this becomes

$$i_f = e\eta\frac{\tau}{\tau_e}\Phi \tag{15.10}$$

with $\tau_e = w/v_e$.

Gain

If we compare (15.10) to (15.4), we notice an internal gain mechanism $G = \tau/\tau_e$. This gain is caused by the difference in recombination lifetime and transit time. Assume for example that the electrons are more mobile than the holes and that the lifetime is very long. The mobile electron will in that case reach the edge of the conductor a lot faster than the hole, which travels to the opposite edge. The continuity condition of the electric current forces the external circuit to provide another electron immediately. This electron is then injected at the opposite side of the detector. The new electron travels to the right again, faster than the hole travels to the left, and this process repeats itself until recombination occurs. The number of passages of an electron per photon is thus τ/τ_e, which is the gain factor. If $\tau < \tau_e$ only a fraction of the electron-hole pairs will contribute to the current. τ_e is determined by the dimensions of the detector and the applied voltage. Typical values are $w = 1\,\text{mm}$ and $v_e = 10^7\,\text{cm/s}$ so that $\tau_e \approx 10^{-8}\,\text{s}$. Then the recombination time varies from $10^{-13}\,\text{s}$ to several seconds, depending on the material and the doping.

15.3 The photodiode

15.3.1 Working principle

A photodiode is mainly the opposite of a light emitting diode. It suffices to reverse the applied electric voltage to change a LED into a photodiode. The working principle is depicted in fig-

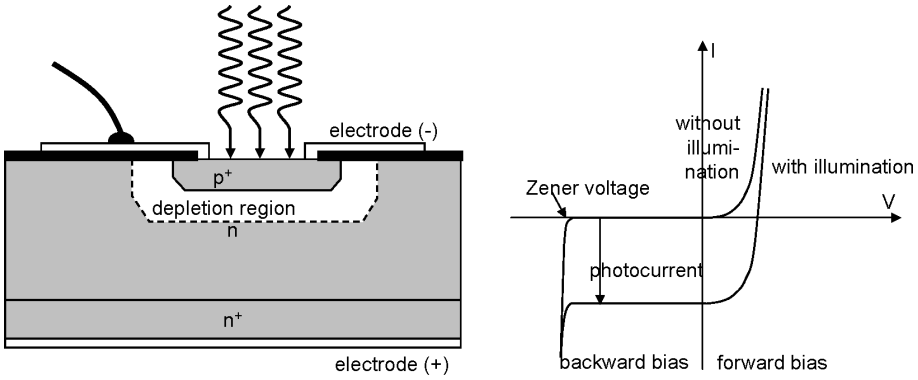

Figure 15.6: Schematic representation and IV-characteristic of a photodiode.

ure 15.5. Light is incident on a semiconductor diode. The bandgap of the semiconductor is chosen in such a way that the light is strongly absorbed. The light intensity thus decreases exponentially and rapidly in the semiconductor.

Electron-hole pairs are created because of the absorption of light. If these pairs are created in a neutral area of the semiconductor, they will quickly recombine (and may cause light emission). However, if the electron-hole pair is created in the depletion layer of the pn-junction, then the electron is led away to the n-area and the hole to the p-area because of the present electric field.

If an external resistor is attached to the diode, a current is able to flow. This current is called the *photocurrent*. The IV-characteristic of a diode with and without illumination is depicted in figure 15.6. In the third and fourth quadrant we find an inverse current that increases proportional to the incident light intensity. The third quadrant is the normal photodiode region: the current is almost not dependent on the applied voltage and nearly zero if there is no lighting ('dark current'). In the fourth quadrant, the diode produces electric power. This is the photovoltaic or solar cell working principle.

For photovoltaics, both crystalline, poly-crystalline and amorphous semiconductors can be used. For consumer applications, silicon solar cells are used, since they show the best efficiency to cost ratio. A cross-section of a silicon solar cell is shown in figure 15.7, showing the p-n junction and a texturized and AR-coated surface to improve the efficiency of the solar cell.

The pin photodiode

The regular pn-photodiode has the disadvantage that the depletion area is relatively thin compared to the distance across which absorption occurs. For this reason, a pin structure is often used (see figure 15.8). In a pin structure a weakly doped (almost intrinsic, i-) area is placed between the p- and n-area. When applying a reverse bias the i-area is completely cleared and there is a weak electric field. This i-area is chosen to be a lot larger than the depletion layer in a pn-junction. A great part of incident light is thus absorbed in the area with the electric field. In this way, responsivity is increased.

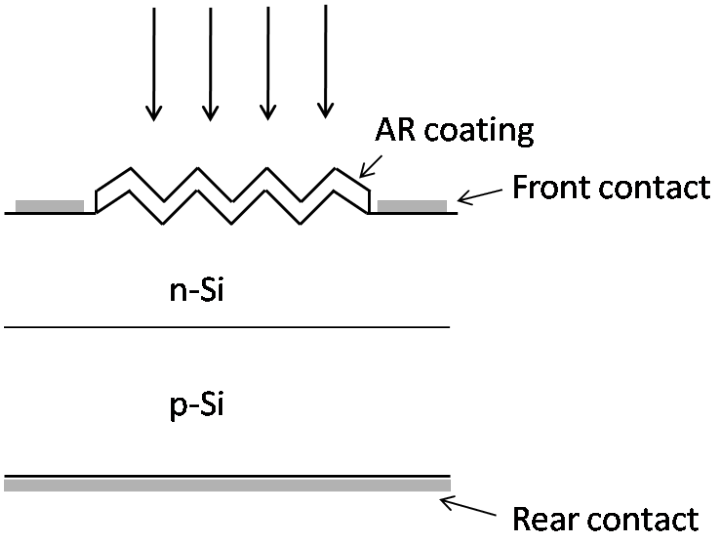

Figure 15.7: Cross-section of a solar cell

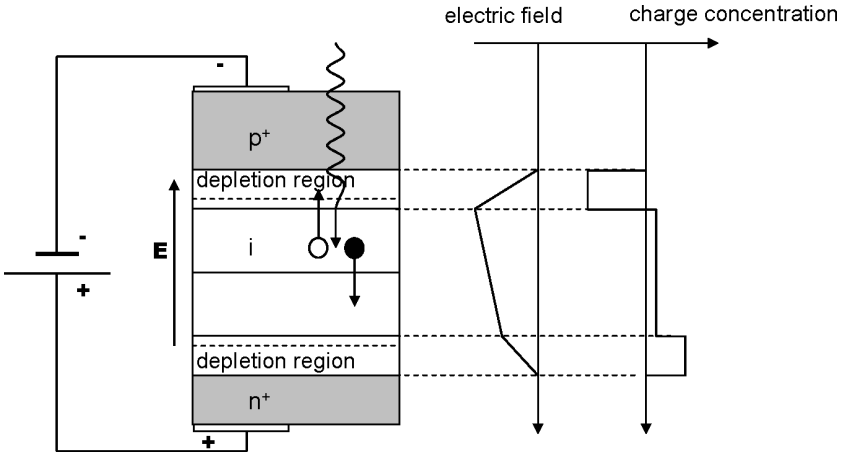

Figure 15.8: Working principle of a pin-photodiode.

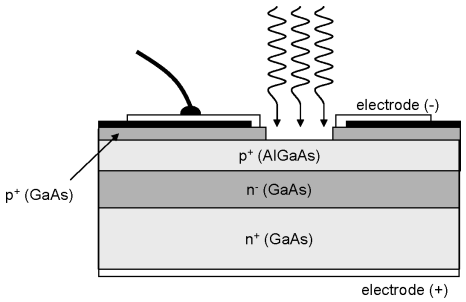

Figure 15.9: Schematic representation of a heterostructure photodiode.

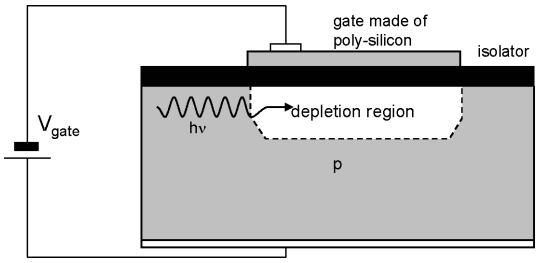

Figure 15.10: Collection of charges in a MOS-capacitance.

The heterostructure photodiode

In a pin structure, the absorption near the surface remains a problem. A solution is given by the heterostructure photodiode (see figure 15.9). Here a p^+n^-n-structure is used. For a certain wavelength range, light is not absorbed by the p-layer with a large bandgap, but it will be absorbed by the n-layers with a smaller bandgap. Quantum efficiency may approach 100 % in this structure.

15.3.2 Modulation bandwidth

The modulation bandwidth of photodiodes is determined by two factors. First, the charges need a certain time to travel in the area in which an electric field is present. This time is usually of the order of a few tens to hundred ps. Second, and more constraining, the diode forms a capacitance that, together with the load resistor, has a RC time constant. For this reason it is important to choose the surface area of the diode as small as possible.

15.4 Semiconductor image recorders

When multiple photodetectors are brought together in a matrix, it becomes possible to register the photon flux as a function of place and time. In that way, an electronic version of an optical image can be obtained. Three functions have to be fulfilled in these image recorders

- collection of charges
- transfer of charges
- measurement of charges

The (historically) most important family of semiconductor image recorders is the CCD sensor (charged coupled device), although recently ($\pm$10 years) for some applications a better alternative can be found in the CMOS-sensor.

The CCD camera

In a CCD camera charges are collected by means of a metal-oxide-semiconductor capacitance (MOS) (see figure 15.10). When a positive voltage is applied at the gate electrode, the holes are

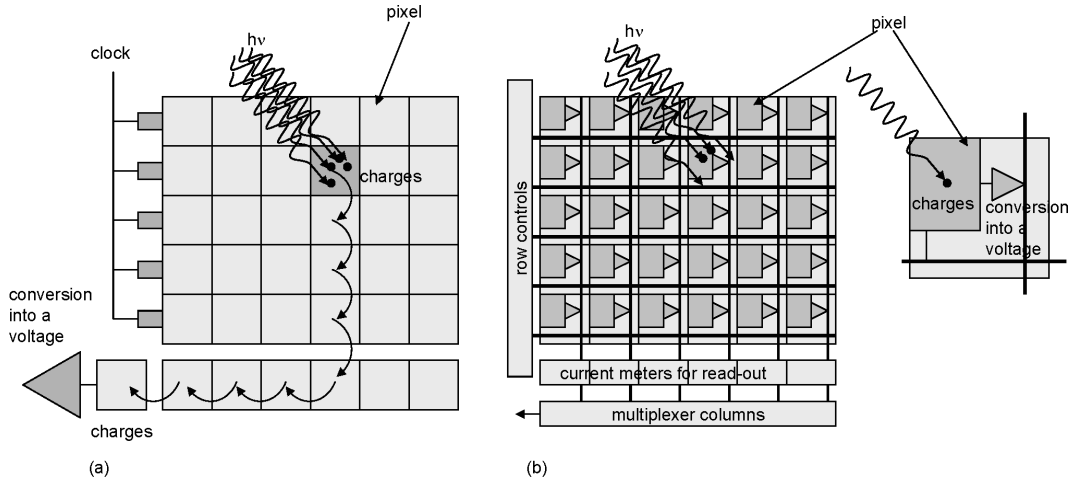

Figure 15.11: Working principle of (a) a CCD camera and (b) a CMOS camera.

driven away and a depletion area is created. The absorption of a photon in the silicon layer gives rise to an electron-hole pair. Subsequently the electron remains captured in the depletion area, which acts as a potential well. The number of electrons that can be captured ("well capacity") depends on the applied voltage, the thickness of the silicon and the surface area of the gate electrode. The collected charge is proportional to the incident photon flux, unless saturation occurs.

Once the electrons are captured in the MOS capacitance, the charge can be transported form one gate to another, by applying the right voltages to the matrix of gate electrodes. Charges are in that way drained to the read-out structure, where they are converted into a voltage (see figure 15.11a).

The CMOS camera

The collection of charges in a CMOS camera is the same as in a CCD camera. However, the charge to voltage conversion now happens in each pixel itself (see figure 15.11b). CMOS cameras can be integrated with analog and digital circuits onto the same chip ("Camera on a chip"). For the time being CCD cameras are superior to CMOS cameras concerning image quality, but the CMOS technology is evolving rapidly. Due to integration higher frame rates can be obtained and they can be produced relatively cheap. CMOS cameras are therefore ideal for low-cost applications like webcams.

Color cameras

Apart from the wavelength-dependent sensitivity because of material absorption, no mechanism is present in a CCD or CMOS to select a specific wavelength for detection. Colors can only be detected by adding some kind of color filter.

Bibliography

Chapter 16

Technology of Optoelectronic Semiconductor Components

Contents

The realization of optoelectronic semiconductor components used for research, telecommunications and optical information processing demands a number of specific materials, technological processes and special facilities. In this chapter we will give a realistic view of the fabrication of optoelectronic components like lasers, LEDs and detectors. However, a lot of technological details are left out.

We start with a brief description of a few fundamental steps that occur in the production process of e.g. semiconductor lasers:

- epitaxial growth: the growth of the layer structure
- photolithography: defining structures into the photoresist
- etching processes: the removal of material
- plasma deposition and plasma etching
- metallization: placing contacts and connections

In the last paragraph, we describe the concrete production of laser diodes in GaAs/AlGaAs.

The so-called III-V semiconductors are used for nearly all the applications in optoelectronics. The two most important substrate materials are GaAs and InP, on which layers of InGaAs, AlGaAs and InGaAsP can be grown epitaxially. In contrast with Si, which is by far the most used semiconductor material, III-V materials have a direct bandgap. This enables the efficient generation (light emitting diodes and laser diodes) and detection (optical detectors, solar cells...) of light. A lot of the technological process steps used in the fabrication of III-V components, are derived from production techniques for integrated circuits in silicon. A modification of each process is required however as we are dealing with other materials and want to make other components.

16.1 Crystal growth

The basic material for the fabrication of semiconductor components is a wafer with a thickness of 0.4 to 1 mm that is being cut off of a perfect monocrystal and polished afterwards. For the fabrication of Si-circuits, wafers are used with a diameter between 4"[1] and 300 mm. For III-V materials, 2" to 4" wafers are employed usually. The wafers are smaller as it is more difficult to obtain a uniform composition of the material during growth, for the basic material (e.g. the same number of Ga-atoms and As-atoms) as well as for the doping elements. This wafer is only used as a substrate, because an optoelectronic component consists of a number of layers with different optical and electrical properties, in other words layers that consist of different materials.

16.2 Epitaxial growth

Using epitaxial growth techniques, monocrystalline semiconductor layers are deposited on the monocrystalline substrate. In this way the crystal lattice is continued from the substrate to the deposited layer, although they have a different chemical composition (e.g. a AlGaAs layer onto a GaAs substrate) or doping (e.g. a n-doped layer on a semi-isolating substrate). Naturally we have to make sure that the lattice constants are the same for the different materials. A lot of optoelectronic components are based on such a piling up of layers. Different growing techniques exist for the fabrication of the layer structures. These are however always derived from one of the three basic techniques: *Liquid Phase Epitaxy (LPE)*, *Metal Organic Chemical Vapour Deposition (MOCVD)* or also called *Metal Organic Vapour Phase Epitaxy (MOVPE)* and *Molecular Beam Epitaxy (MBE)*. The most popular technique is MOCVD, of which a schematic representation is given in figure 16.1.

When using MOCVD an epitaxial layer is grown from the gas phase. When growing III-V materials, we start from metalorganic materials (group III) and hydrides (group V). The carrier gas is hydrogen gas. By controlling the flow and temperature of the different gas components, the ratio of the different gases in the reactor is varied. These gases are sent through the reactor by a valve system and the remaining gases are led to the exhaust pipe via a bypass.

The substrate is located onto a so-called susceptor in the reactor which is electrically heated, with IR-lamps or RF-induction. The high temperature causes a reaction above the substrate resulting in the deposition of epitaxial layers.

[1] 1" $\simeq$ 25.4 mm

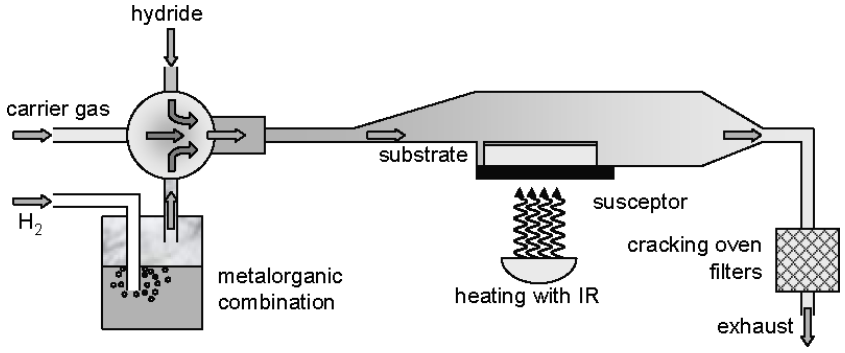

Figure 16.1: Schematic representation of a MOCVD device.

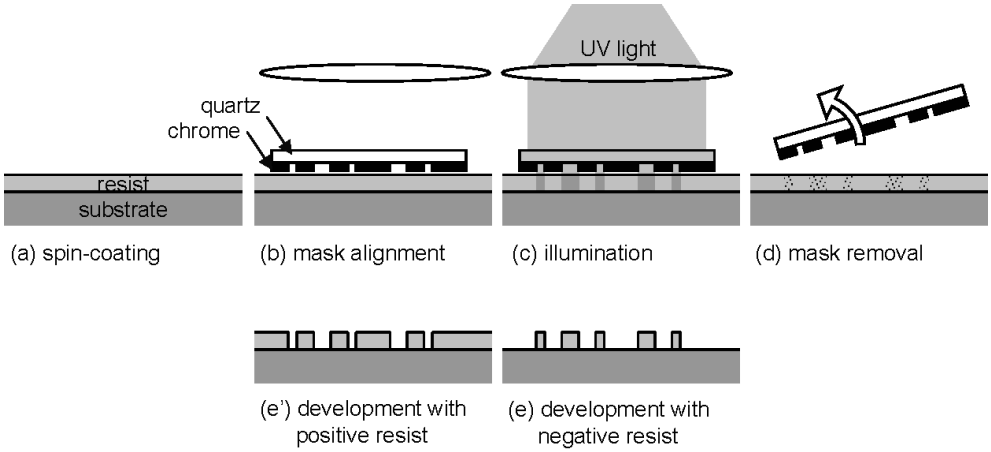

Figure 16.2: Successive steps in a photolithographic process.

16.3 Photolithography

For the fabrication of laser diodes numerous other steps, besides the growth of crystalline layers on the substrate, have to be followed, such as the etching of material and metallization. Only certain parts of the substrate need etching and the metallization has to occur in specific patterns. To accomplish this a masking layer is needed in order to protect certain parts of the substrate.

The most commonly used materials for this purpose are UV-sensitive polymers or the so-called *photoresist*. The pattern of a mask is first transferred photolithographically in this substance and then this resist pattern is used as a mask for the final process step. There are numerous photosensitive polymers on the market, each with their own specific spectral sensitivity and range of layer thickness. The needed resist is chosen as a function of the application. UV-lithography ($\lambda \sim 300 - 400\,\text{nm}$) is by far the most used technique, though we continuously strive after smaller dimensions, so other light sources are used like deep-UV, X-rays... The minimal detail size W depends on the wavelength: for contact lithography (see further) there is a rule of thumb $W \sim \sqrt{\lambda g}$, with g the distance between the mask and the bottom of the resist layer. For projection photolithography $W \sim k\lambda/\text{NA}$ applies, with NA the numerical aperture of the projection system and k a correction factor.

Photolithography consists of the sequence of a number of constituent processes (see figure 16.2):

1. cleaning of the substrates
2. putting on the resist
3. baking of the resist
4. alignment of the substrate w.r.t. the mask
5. lighting of the resist
6. development of the resist

The first step can be considered as one of the most important steps in the entire process. The cleaning includes a degreasing with solvents. Afterwards the samples are rinsed with deionized water and heated up for a long time to make the surface completely free of moisture.

Then the photoresist layer is applied, which can generally be done in two different ways. When using *dip-coating* the substrate is submerged in the photopolymer and then slowly pulled up. When using *spin-coating* the substrate is covered with resist and then spinned around at high speed (3000 to 5000 rpm). Uniform and well reproducible layers are obtained in this way. After putting on the resist, the polymer is baked in a nitrogen environment according to a carefully controlled procedure.

In the next step the desired pattern is transferred via UV-illumination in a *mask-aligner* from the mask to the resist layer. This mask is a glass plate on which the designed pattern is present in a thin metal film (usually chrome). Using a microscope and micrometer screws, the substrate is aligned w.r.t. the pattern on the mask. This alignment is critical as different process steps (in a mask design these are called different levels) have to be carried out after each other in order to get a full component. When the substrate is well positioned, the resist layer is lit through the transparent zones in the chrome layer. In that way the mask pattern is copied into the resist layer.

INTEC uses *contact lithography* as lighting technique. In this method the substrate is pressed against the mask by creating a vacuum in the space between the mask and the substrate. When illuminated, a 1 to 1 image of the mask is obtained. The simple optics is an advantage of this technique but the mask can be damaged due to the mechanical contact between the chip and the mask. Alternatively, when using *projection lithography*, a lens system is needed between the mask and the substrate. An enlarged version of the pattern is put on the mask. This is then projected reduced onto the substrate. Afterwards the substrate is shifted and projected again. In that way the whole substrate can be projected with the same chip design. Such a device is called a *stepper*. Projection lithography is often used in production environments.

After illumination, the polymer layer is developed in a basic solution. We can choose between positive or negative photosensitive substances, where either the illuminated or the non-illuminated parts are developed and removed.

16.4 Wet etching

There are two types of etching processes: *wet etching* and *dry etching*. The term wet etching denotes the fact that the chemicals are used in their liquid phase. In a dry etching process on the other hand, chemically reactive gases are used. Dry etching is discussed in the next section.

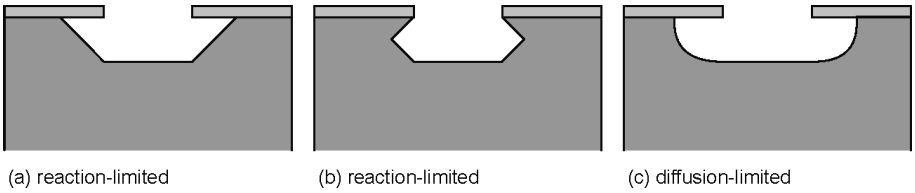

Figure 16.3: Reaction-limited wet etching process versus diffusion-limited wet etching process.

When using wet etching of III-V materials, we have to keep in mind that these materials are built from different elements. E.g. if we want to etch GaAs, the etch mixture has to react with Ga as well as with As. The situation becomes even more complex when working with ternary or quaternary combinations. This problem is handled in the same way for nearly all materials. In the first step the surface is oxidized. Afterwards the oxides are dissolved in an acid or basic environment. In the classic etch mixture $H_2SO_4/H_2O_2/H_2O$, that is used in the processing of the laserdiode to etch the mesa-stripe, H_2O_2 oxidizes the surface and H_2SO_4 dissolves the oxide. Both constituent processes are in competition with each other. For certain ratios of the products, the oxidation takes place quickly and the dissolution (determined by the diffusion process) will be speed-limiting so we can call it a diffusion-limited etching process. Using other ratios the oxidation process (determined by the chemical reaction) will be a lot slower than the diffusion process, and we obtain a reaction-limited etching process.

The obtained etching profile depends on the type of process that occurs (see figure 16.3). With a diffusion-limited process, the diffusion time, and thus the distance between the semiconductor surface and the etching mixture, will determine the etching speed. Because of this, a circular profile arises at the edge of the mask. When using a reaction-limited process, the etching speed will be determined by the chemical reactivity and as this is dependent of the crystallographic direction, lattice planes will become visible (in the case of GaAs these are typically (111)-planes).

By choosing the right etching mixture, we can ensure that certain materials are etched while others remain untouched. These are called *selective* etching mixtures.

The greatest limitation of wet etching processes is the dependence of the etching speed on the opening in the mask and the impossibility to etch fine structures because of underetching. As a rule of thumb we can assume that structures smaller than three times the etching depth will cause problems when using wet etching.

16.5 Plasma deposition and plasma etching

The metal patterns that connect the underlying components and form the electric contacts with the environment have to be isolated from each other and from the underlying layer structure. To this end, one uses polyimide and dielectrics such as SiO_2 or Si_3N_4. Plasma activated processes are being used to create these isolating layers and etch the patterns: *Plasma Enhanced Chemical Vapour Deposition (PECVD)* and *Plasma Etching (PE)*, with the following variants *Reactive Ion Etching (RIE)* and *Inductive Coupled Plasma Reactive Ion Etching (ICP-RIE)*.

During the deposition (PECVD), the choice of gases determines the composition of the layers: SiO_2, Si_3N_4 or a combination in between (oxynitrides) can be obtained. For Si_3N_4 the gas mixture

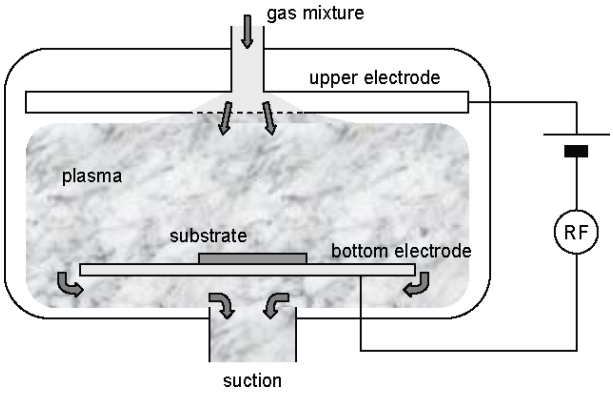

Figure 16.4: Schematic representation of a PECVD-device.

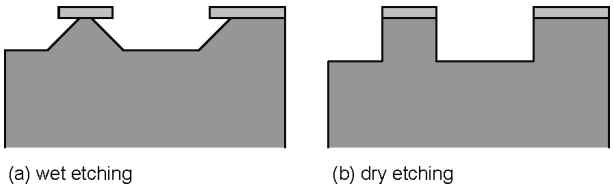

Figure 16.5: Differences between the etching profile when using dry and wet etching.

SiH_4-NH_3-N_2 is used, while for SiO_2 the mixture SiH_4-N_2O is used. The gas mixture is brought in the reactor through small holes in the upper electrode. By applying an AC-voltage between the electrodes, the gas mixture will be ionized. During the reaction, radicals and ions from this plasma create a uniform film on the substrate with the desired composition and quality. Eventual reaction products and the rest of the gases are sucked out of the room to the pump installation through an opening at the bottom electrode (see figure 16.4).

The most important advantage of this dry deposition technique is the low reaction temperature, enabled by the use of extremely reactive chemical radicals (e.g. monoatomic N). The formation of high quality Si_3N_4 can normally only take place at temperatures above 700 °C, but these temperatures have to be avoided in the III-V technology. This deposition process is furthermore omnidirectional which enables a good coverage of non-planar structures (this is hard to get with sputter processes).

The principle of plasma etching is analogous. Radicals and ions react with the material that has to be etched by choosing an appropriate gas mixture. These materials can be dielectrics, semiconductors as well as metals. The volatile reaction products are then sucked away by the pump system.

The major difference between wet and dry etching processes becomes clear when looking at a cross section of the etched material (see figure 16.5). When using a wet etching process, nothing stops the chemicals from etching the material vertically as well as horizontally, and thus under the mask. Thus, in the case of a wet etching process, we always have a profile in which a clear underetching is observed. When using dry etching we can accelerate the ions in a certain direction, by adapting the structure of the reactor and the process itself, and make sure that there is only vertical etching. The resulting profiles are steeper and there is almost no underetching.

Due to the high investment costs and the complexity of the process, dry etching is used only if wet etching is not possible. Typical applications are the etching of narrow, deep structures, independent of the crystal orientation.

16.6 Metallization

The application of metal patterns is of course an indispensable step in the production of each electronic component, as they are the contacts of the device with the environment. The quality of the metal-semiconductor contact can greatly influence the performance of a component.

The most important part of the installation for the deposition by evaporation is the vacuum tube in which the required pressure is obtained by a combination of vacuum pumps and turbomolecular pumps. The lower the final pressure, the better the contact will be. Typically a pressure lower than 10^{-5} mbar is desired as the gas molecules cause contamination.

There are two techniques to deposit a metal film: thermal deposition by evaporation and sputtering. When using *thermal deposition by evaporation* we can further distinguish Joule evaporators and electron beam evaporators.

- In Joule evaporators, a large current (> 100 A) is sent through a crucible in which an amount of the material that has to be evaporated is present. This causes the metal to evaporate and deposit itself on the edges of the vacuum tube and on the substrates. The speed of the evaporation depends on the temperature of the crucible (determined by the current) and can be as large as a few tens of nm per minute.
- When evaporating with an electron beam device, the metal is melted and evaporated by directing a highly energetic electron bundle with magnets onto the material holder. The kinetic energy of the electrons is hereby converted into the necessary heat. The obtained temperatures are considerable so that nearly all materials can be evaporated in this way.

When using *sputtering*, the material that has to be put onto the semiconductor is present as a massive block, serving as an electrode, with on top a second electrode between which an electric voltage is applied. A plasma is created in the present argon gas. The Ar-ions are drawn to the electrode, which is made of the material that has to be deposited. Metal atoms are then released because of heavy collisions with the Ar-ions. These metal particles fill the vacuum tube and are also deposited onto the substrate that is placed above the source. The speed of sputtering in this process is typically a few nm per minute and is determined by the applied voltage, the gas pressure and the material.

In each of these methods, the metal forms a continuous film on the whole substrate. To create the desired patterns, we can use an etching process or the *lift-off* technique, as illustrated in figure 16.6.

In the etching process, the lithographic pattern is defined in the already deposited metal. In the lift-off technique, the lithographic pattern is first applied and afterwards the metal layer. The polymer pattern is then dissolved, causing the upper metal parts to be removed as well. The obtained pattern is thus the inverse of the resist pattern. From figure 16.6 it is clear that such a lift-off process is only possible if the metal film does not completely cover the resist profile. With a positive profile, the resist can no longer be dissolved in the solvent and the lift-off will fail. When

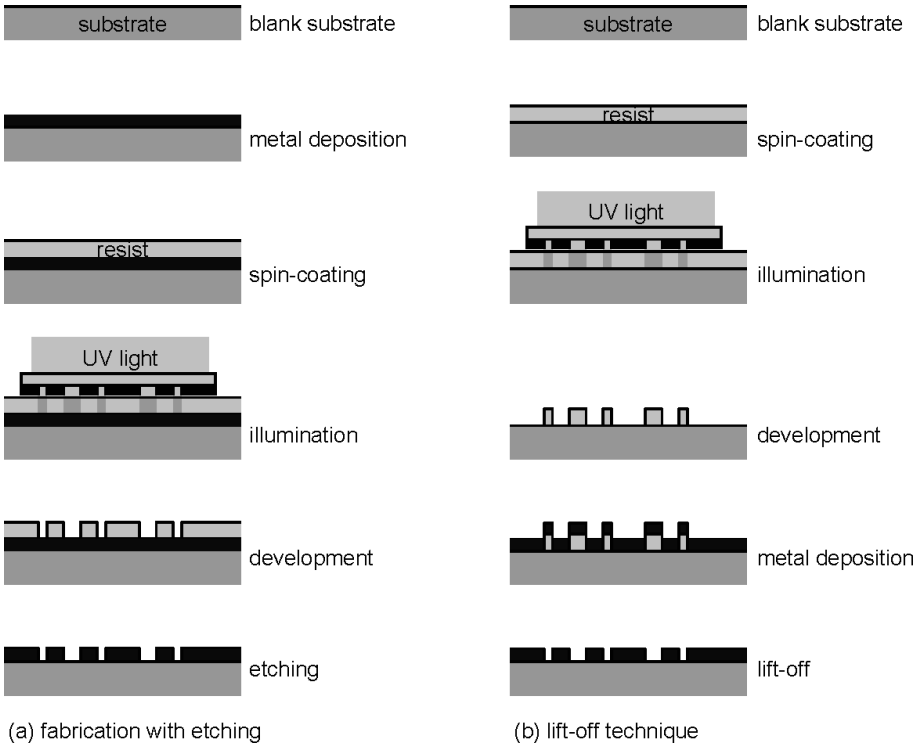

Figure 16.6: Deposition of metal layers and definition of patterns: the etching technique versus the lift-off technique.

having a straight profile, the resist can be removed but the raised edges may remain. A negative profile is ideal for the lift-off technique. This leaning profile can be obtained by *image-reversal* during the photolithography, in which advanced kinds of polymer are used as photoresist.

16.7 Packaging

Mounting and packaging of a component is required as the component itself, as fabricated during the processing, is hard to handle and subject to all sorts of influences from the environment. Packaging of components is done in different ways and the quality of the packaging usually determines the lifetime of the component.

The packaging of electric components ensures protection against the outside world and electric connections have to be made when the component is mounted. *Die-bonding* ensures the fixing of the chip inside the packaging, while *wire-bonding* takes care of the wires between the chip and the pins of the packaging. The placement of the chip inside the package is not so critical, as the wires can counter a wrong positioning. More important is that the thermal aspects are taken into account, in order to counter temperature rises.

A number of extra aspects occur for optical and optoelectronic components. Usually an extra protection is provided for the exit facets and mirrors of the component. Furthermore, we have to be aware of the temperature as a lot of the characteristics of the components strongly depend on it. We also have to take care of the optical entries and exits. This means that the component has to

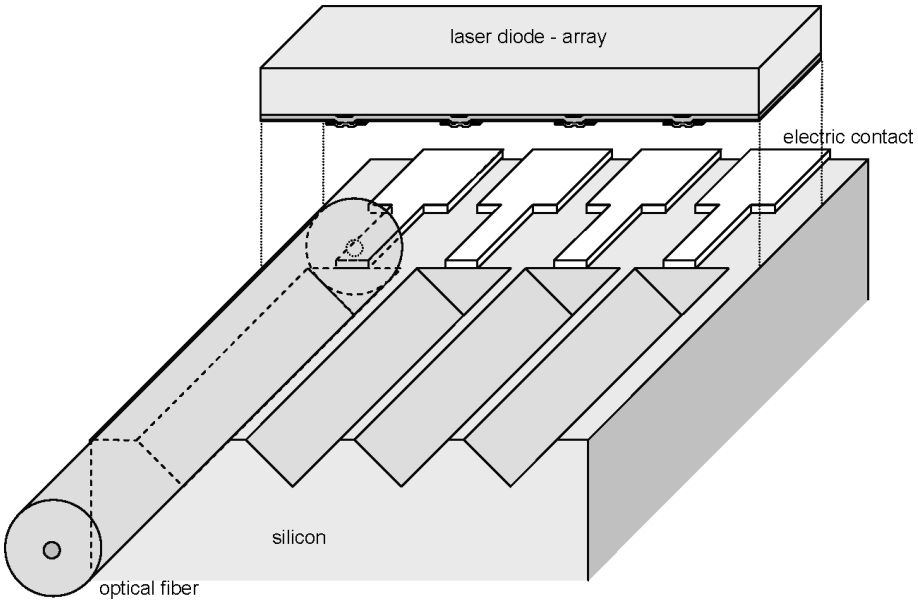

Figure 16.7: Alignment of laser diodes and optical fibers using V-grooves.

be aligned w.r.t. the micro-optics or optical fibers. Packaging is therefore a large part of the costs of a component. For a device connected with an optical fiber for example, the packaging embraces more than 60 % of the cost price. To reduce this cost price, active[2] alignment has to be avoided especially. This is e.g. realized by designing structures in which passive alignment occurs. Let us examine the alignment of a number of optical fibers (see figure 16.7). A Si-carrier is equipped with a number of contact planes and grooves. Optical fibers are placed in these grooves and the laser row is mounted onto the contact planes via a *flip-chip*[3] technology. The optical fibers are thus aligned and positioned in the V-grooves. These V-grooves are lithographically defined and aligned w.r.t. the contact planes for the laser diode. The laser is then placed via flip-chip on these contact planes, and thus perfect alignment is obtained.

The packaging costs are reduced further by integrating multiple components in one single package or even in one single chip. In that way, the alignment problem is avoided. In the present world of optoelectronics, there is a large tendency towards smaller components with more integration, better performance and lower costs.

16.8 Example: fabrication of a laser diode

The production of a typical laser diode (see chapter 14) consists schematically of 8 process steps:

1. growth of the appropriate crystalline layer structure

[2]In active alignment, the alignment is optimized by measuring and maximizing the usable exit power during the process.

[3]In this flip-chip technology, a chip is provided with contact planes over its entire surface on which an extra metal layer is put afterwards (bumps). The same happens on the carrier and the chip is then mounted upside down onto the carrier. In this way, a two-dimensional contact can be realized

2. lithography and etching of the mesa
3. deposition, lithography and etching of an isolation layer
4. lithography and application of the upper contact
5. lithography and application of an extra contact-metallization
6. thinning of the substrate by polishing it
7. placing of the bottom contact
8. cleaving the laser mirrors

We emphasize that the production scheme of these components changes continuously, evolving to more reliable and reproducible processes. Meanwhile, the scheme currently used at INTEC has also evolved compared to the one described here, but this one is chosen as it is a more traditional, commonly known and used procedure.

The layer structure needed for the fabrication of a laser diode depends strongly on the wanted characteristic of the final component. For a simple double heterostructure laser diode we start from a structure as depicted in figure 16.8 (a). As it is technologically difficult to put a good electric contact on AlGaAs, a highly doped GaAs-layer is put on top of the laser structure.

Subsequently a *mesa*[4] is created in the layer structure with etching processes (see figure 16.8 (b)-(e)). This mesa has two functions. First, the mesa acts as an optical waveguide that, combined with the laser mirrors, forms an optical cavity. Second, when a current is injected, the mesa prevents the current from leaking away laterally into the layer structure. This improves the operation of the laser as a high current concentration is obtained in the active layer, where the optical field is at its maximum.

After this step, the metal contact can be placed. However, this has to be put on top of the mesa, as only there a sufficiently good contact is ensured. On the other hand, the dimensions of the contact have to be sufficiently large in order to allow connections with other components or measuring probes. In such a large contact the current leakage that may be injected next to the laser stripe can then no longer be neglected and can interrupt the proper operation of the laser. An isolating layer is therefore applied to avoid contact of the metal film with the layer next to the mesa.

The contact between the metal film and the semiconductor will usually show a diode current-voltage characteristic. By alloying at high temperatures (350-450 ^{o}C), where the metal atoms diffuse in the GaAs, the diode behavior is transformed into a low-resistive contact and a linear I-V characteristic. An example of such an ohms contact is a n-type AuGe/Ni contact for GaAs laser diodes.

In our example, the isolating layer consists of 300 to 500 nm Si_3N_4, deposited over the entire surface. After this plasma deposition, the Si_3N_4-layer above the mesa and the electric contacts are lithographically etched away with a dry etching technique. The result is shown in figure 16.8 (h).

After a new lithographic step and by using a lift-off process, the upper contact can be applied with different metal layers (Ti/Au is usually employed for a p-type contact and AuGe/Ni is used

[4] A mesa is a plateau with steep edges, commonly found in the southwest landscape of the USA.

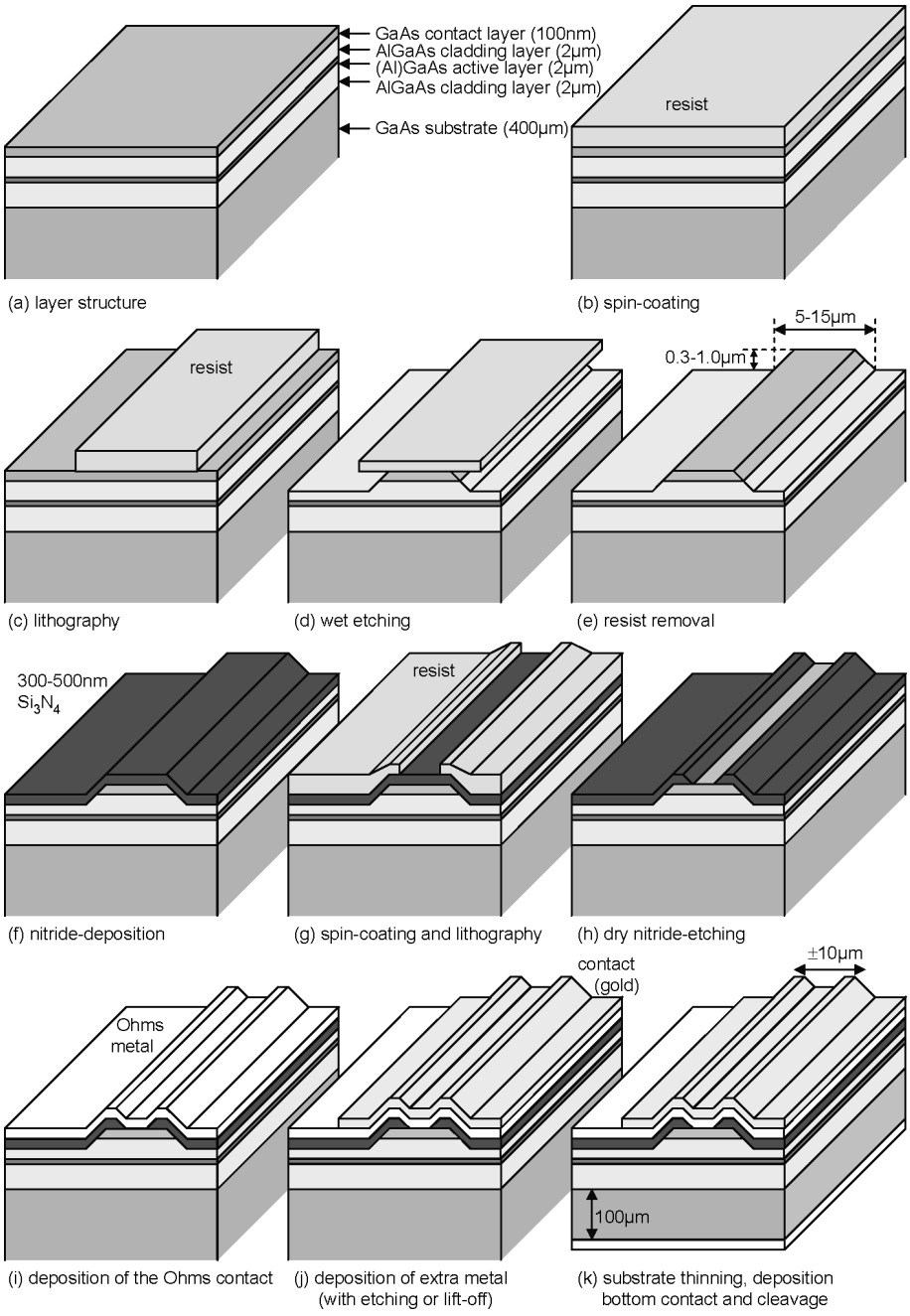

Figure 16.8: Process flow for the fabrication of a laser diode.

for a n-type contact). In case of a laser diode with a n-type substrate, the upper contact has to be p-type and vice versa. Using the same process, an extra metallization (TiW/Au) is often put on the contact plane at the upper side of the laser to further reduce the serial resistance of the electric contacts. If an even lower serial resistance is required, a thick (2 tot 4 μm) Au-layer is added electrochemically (Au-plating).

The substrate on which the whole laserdiode is fabricated, has an initial depth of approximately 400 μm. As the thermal conductivity of GaAs is pretty low, this often causes a problem for the stable functioning of the component. Therefore, in nearly all cases the substrate is thinned to approximately 100 μm by polishing techniques. In this way, the cleaving of the individual components is also simplified.

The second electric contact is put on the back side of the substrate (Ti/Au or AuGe/Ni, depending on the type of substrate). This metal film is deposited by evaporation over the entire back side of the sample and no lithographic step is thus required. As was the case for the upper contact, this one needs to be alloyed in order to get an ohms contact.

The final process step is the creation of the laser mirrors. The planes obtained by cleaving the component according to the desired dimensions, are perfect crystal planes and therefore extremely suitable as mirrors. The finished laserdiode with some typical dimensions is shown in figure 16.8 (k).

Bibliography

Chapter 17

Lighting

Contents

We are often confronted with light and lighting as the eye is one of the most important senses to mankind. Therefore an engineer has to have some knowledge of the proper use of light in numerous circumstances. We mention for example lighting at work, at home, in the streets, etc.

The following paragraph deals with a few basic concepts of lighting. We pay attention to the measurement of light emitted by a lighting device and lighting calculations. Subsequently the most common types of lamps are discussed.

17.1 Lighting calculations

There are numerous methods to calculate the light of a lighting installation. Generally they can be divided into two classes: the point-by-point method and the integral method. In the point-by-point method the illuminance caused by directly incident light as well as the light reflected onto the walls and such (figure 17.1) is calculated. Such a calculation is only possible if the luminous flux of all the light sources as well as the polar luminous intensity curves are known. Furthermore we have to know all the characteristics of the reflecting walls if we want to take these contributions into account. The point-by-point method is very accurate, but a lot of calculations are needed, especially if reflections are taken into account. However, software packages are available that can execute these calculations.

The integral method can be divided into two methods: method of the lighting efficiency (also called: method of the coefficient of efficiency) and the British Zonal('BZ') method. In these methods, the average illuminance, caused by the directly incident light on the work plane as well as

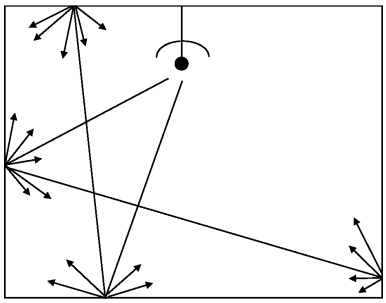

Figure 17.1: Lighting calculations with the point-by-point method.

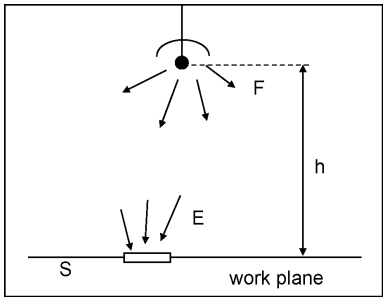

Figure 17.2: Lighting calculations with the integral method.

the reflected light (from walls, ceiling and floor: figure 17.2) is immediately calculated. Thus, this method does not allow to calculate the illuminance in a specific point, but is very simple and quick.

The method of the lighting efficiency uses a quantity η that is called the coefficient of efficiency). This quantity η is defined as the fraction of the luminous flux, produced by the lamps, that reach the floor (or an imaginary work plane 1 meter above the floor). Thus, we can write:

$$\eta = \frac{F_{work\ plane}}{F_{total}} \tag{17.1}$$

The average illuminance $E_{average}$ on the work plane is given by:

$$E_{average} = \frac{F_{work\ plane}}{S} \tag{17.2}$$

with S the surface area of the floor of the room. We get:

$$E_{average} = \eta \frac{F_{total}}{S} \tag{17.3}$$

With this relation, we can calculate the (average) illuminance when the total luminous flux produced by the lamps is known. Naturally η has to be known.

η is tabled for several lighting devices in function of the reflection coefficient of the walls, ceiling and floor and also the shape index k of the room, defined as

$$k = \frac{lw}{h(l+w)}, \tag{17.4}$$

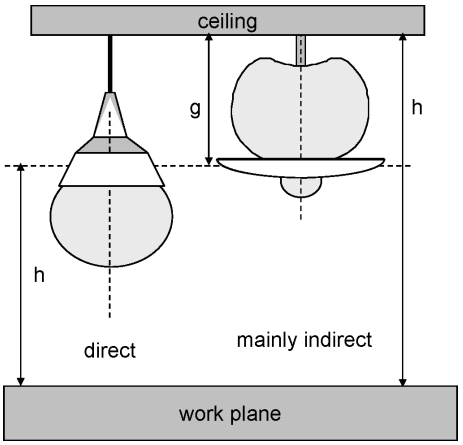

Figure 17.3: Direct and indirect lighting.

Incandescent lamps			Lighting efficiency								
			$\rho_p = 0.7$			$\rho_p = 0.5$			$\rho_p = 0.3$		
System	η	k	$\rho_m =$			$\rho_m =$			$\rho_m =$		
	[%]		0.5	0.3	0.1	0.5	0.3	0.1	0.5	0.3	0.1
Direct		0.5	0.28	0.21	0.17	0.27	0.21	0.17	0.26	0.21	0.17
	0										
	↑	1	0.47	0.41	0.37	0.46	0.41	0.36	0.45	0.39	0.36
	80										
	↓	2	0.64	0.58	0.55	0.62	0.58	0.54	0.60	0.56	0.53
	80										
		5	0.76	0.74	0.71	0.74	0.72	0.70	0.73	0.70	0.68
Mainly		0.5	0.24	0.19	0.15	0.18	0.15	0.12	0.13	0.11	0.09
indirect	69										
	↑	1	0.40	0.35	0.32	0.32	0.28	0.25	0.22	0.20	0.18
	89										
	↓	2	0.54	0.50	0.46	0.42	0.39	0.37	0.31	0.29	0.27
	20										
		5	0.65	0.63	0.60	0.50	0.49	0.48	0.37	0.36	0.35

Table 17.1: Lighting efficiency as function of the type of lighting and the reflection of the walls and ceiling.

with l and w respectively the length and width of the room. h is the distance between the work plane and the lighting device in case of mainly direct lighting, or the distance between the work plane and the ceiling in case of indirect lighting via the ceiling.

The lighting efficiency for two types of settings - a mainly direct and a mainly indirect - is represented in table 17.1 (figure 17.3). The first column denotes the amount of light, produced by the lamp, that leaves the fitting and the percentage radiated upwards respectively downwards. The other columns represent the lighting efficiency as function of the shape factor k and the reflection coefficient of the ceiling ρ_p and the walls ρ_m.

In the case of mainly direct lighting, we see that ρ_p only has a small influence on the efficiency. The efficiency also decreases and depends more on ρ_m as the room gets higher.

This method can only be used if the following conditions are approximately fulfilled:

Lighting method	Max. distance between the lighting devices
direct lighting	$1.35h$
mixed lighting with diffusers	$1.70h$
mixed lighting with TL-tubes	$1.50h$
indirect lighting	$3g$

Table 17.2: Guidelines for the maximal distance between light sources in different types of lighting.

- the room is closed and rectangularly shaped,
- the walls have a uniform and well known reflection coefficient,
- enough identical lighting devices are placed so that a certain uniformity of the illuminance is guaranteed.

Concerning this last fact, a number of rules of thumb for the maximal distance between the lighting devices can be deduced from table 17.2. A main distinction can be made between the case of direct lighting and indirect lighting.

Hereby h is the height of the light sources above the work plane, and g the distance between the fittings and the ceiling.

The method of the coefficient of efficiency is very good if the η-tables are known for the used lighting device. However, if this is not the case, the reference table that best fits the lighting device has to be used. Large margins of error can be introduced in that way. Therefore the British Zonal method has been developed. This is just an extension of the method of the coefficient of efficiency, in which a more systematic method is used to determine the reference device with which the given device corresponds the most.

17.2 Light color

The color of the illumination happens to be very important. When it is necessary to easily distinguish colors, a source that emits light as white as possible has to be used, approaching the characteristics of daylight. Furthermore, the color of artificial light plays a significant subjective role. Bright light is preferably as white as possible, while a softer 'warmer' hue is usually chosen for weaker light. Lighting with a uniform color is mostly favored to heterogeneous lighting (unless special effects are desired). The latter because the eye adapts itself to the color of the light source and will barely notice, in uniform lighting, that the objects do not look the way they do in daylight.

17.3 Characterization of light sources

17.3.1 Measurement of the illuminance and calculation of the luminous flux

The luminous intensity $I(\theta, \phi)$ of a light source in function of the angles θ and ϕ can be measured by rotating this light source in a horizontal and vertical plane w.r.t. a photocell. For small light

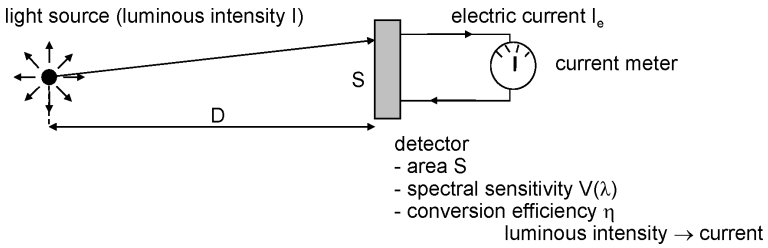

Figure 17.4: Measurement of a light source.

sources like incandescent lamps, there are no problems. However, when the sources have large dimensions, e.g. a device for fluorescent tube lamps, we have to pay attention to the fact that the luminous intensity needs to be measured at a large distance compared to the dimensions of the lighting device. Once the luminous intensity $I(\theta, \phi)$ is determined, the total luminous flux F can be calculated by numerical integration.

$$F = \int\int I(\theta, \phi)\, d\Omega = \int\int I(\theta, \phi) \sin\theta d\theta d\phi. \tag{17.5}$$

If an analytical expression is known for the luminous intensity, the integral can be calculated. The light flux of a radiator satisfying Lambert's law, is thus given by:

$$F = \int\int I_0 \cos\theta d\Omega = \pi I_0 \tag{17.6}$$

17.3.2 Direct measurement of the total luminous flux

If we want to measure the total luminous flux of a light source, we can avoid performing a number of measurements of the illuminance by using an integrating sphere photometer (also called a sphere of Ulbricht). This device immediately gives us the total luminous flux of the considered source. The integrating sphere or sphere of Ulbright (figure 17.5) consists of a hollow sphere, with a diameter much larger than the dimensions of the light source. The inner side of the sphere is painted in a white mat paint with reflection coefficient ρ that scatters the light following Lambert's law.

The paint reflects a fraction ρ of the optical power (and absorbs the rest). The light source is placed in the middle of the sphere. A photodetector is placed in the surface of the sphere and a small screen shields the detector from direct incident light. We now prove that the response of the photodetector, thus the illuminance E on this detector, is proportional to the total luminous flux F of the light source. A part dF of the luminous flux of the source illuminates the surface area dS around a point A and scatters there according to Lambert's law:

$$I = I_m \cos\theta \text{ with } I_m = \frac{\rho dF}{\pi} \text{ so that } I = \frac{\rho dF}{\pi} \cos\theta. \tag{17.7}$$

The luminous flux d^2F' originating from dS_1 that reaches a surface area dS_2 around an arbitrary point B is:

$$d^2F' = \frac{\rho dF}{\pi} \cos\theta \frac{dS_2 \cos\theta}{|AB|^2} \tag{17.8}$$

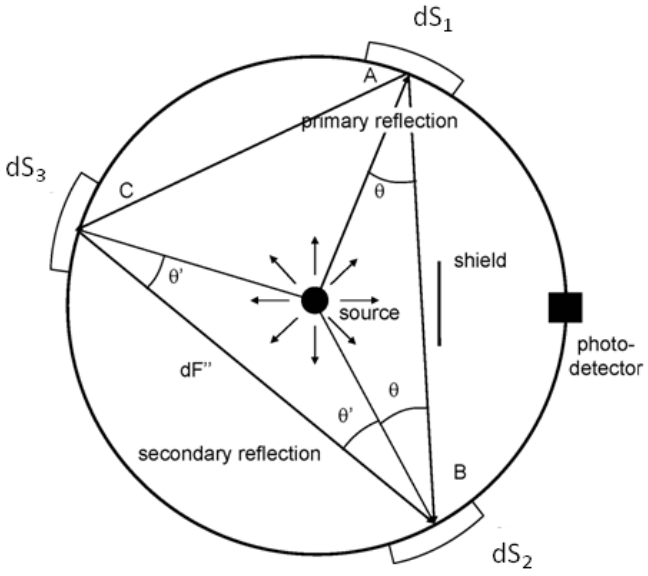

Figure 17.5: Direct measurement of a light source with an integrating sphere photometer.

with $|AB| = 2R\cos\theta$. Consequently:

$$d^2F' = \frac{\rho dF dS_2}{4\pi R^2} \tag{17.9}$$

Thus, the illuminance dE' in B, due to this first reflection of dF, is:

$$dE' = \frac{d^2F'}{dS_2} = \frac{\rho dF}{4\pi R^2}. \tag{17.10}$$

This does not depend on θ so that, considering all first reflections, a constant illuminance E' is obtained over the entire surface of the sphere:

$$E' = \frac{\rho F}{4\pi R^2}. \tag{17.11}$$

Let us now calculate the illuminance E'' due to the 2nd reflection. The surface area dS_3 now acts as a secondary radiator emitting a luminous flux

$$\frac{\rho^2 F}{4\pi R^2} dS_3 \tag{17.12}$$

according to Lambert's law. On dS_2 this contribution is:

$$d^2F'' = \frac{\rho^2 dF}{4\pi R^2}\frac{\cos\theta'}{\pi}\frac{dS_2\cos\theta'}{|CB|^2} \tag{17.13}$$

$$= \frac{\rho^2 F dS_2}{4\pi R^2}\frac{dS_3}{4\pi R^2}. \tag{17.14}$$

Thus, the illuminance dE'' in the point B, due to the 2nd reflection on the surface area dS_3, is:

$$dE'' = \frac{\rho^2 F}{4\pi R^2} \frac{dS_3}{4\pi R^2} \tag{17.15}$$

This is again independent of θ, so that the illuminance E'' due to all secondary sources - this is the entire surface of the sphere - becomes:

$$E'' = \frac{\rho^2 F}{4\pi R^2} \tag{17.16}$$

Thus, the total illuminance E_t in an arbitrary point on the inner surface of the sphere, due to all reflections, is

$$E_t = \frac{\rho F}{4\pi R^2} \left(1 + \rho + \rho^2 + \ldots\right) \tag{17.17}$$

or

$$E_t = \frac{F}{4\pi R^2} \frac{\rho}{1-\rho}. \tag{17.18}$$

We notice that the illuminance on the detector is directly proportional to the total luminous flux of the light source. The following factors limit the accuracy of the integrating sphere:

- the paint does not scatter the light according to Lambert's law,
- the reflection coefficient depends on the wavelength,
- the source is not a point and thus impedes further reflections.

The relationship above can also be deduced in another manner. As the illuminance E' after one reflection is the same on the entire inner surface (see above), the total illuminance E_t (apart from direct illumination) also has to be the same on the entire surface. The following power balance can then be formulated. The total luminous flux responsible for this illuminance is ρF. The luminous flux that is continually being absorbed by the surface (apart from direct illumination) is $4\pi R^2(1-\rho)E_t$. These two quantities have to be the same at static equilibrium from which the relationship between E_t and F follows.

17.3.3 Measurement of luminance

The relationship derived in chapter 2 between the luminance of a source and the luminous intensity on the retina is also used to measure the luminance of a light source. Instead of the eye, we consider a system with a lens and a photodetector. The latter is mounted in the plane where an image of the light source is formed by the lens (figure 17.6). The relationship $E = Ld\Omega''$ between the illuminance E on the detector and the luminance L applies again. The detector then gives an electric signal proportional to the total luminous flux on its surface (we also have to take care of the fact that the spectral sensitivity of the detector has to be the same than that of the eye).

This luminous flux is proportional to the average luminance of that part of the light source that is imaged onto the detector. The size of the detector thus determines the spatial resolution by which the luminance can be measured (resolution on the radiant surface that has to be measured). Analogously, the size of the lens determines the angular resolution of the (direction-dependent) luminance.

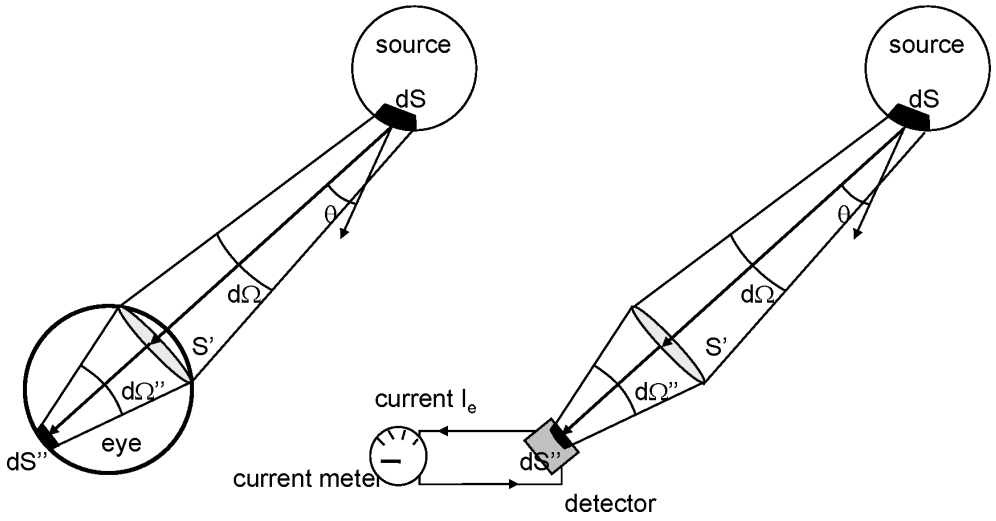

Figure 17.6: Measurement of the luminance with the eye or with a detector.

17.4 Thermal (blackbody) radiators

17.4.1 The blackbody radiator

It is a common fact that hot objects emit electromagnetic radiation. Hot metal coming out of a blast-furnace, has a yellowish white color. This radiation is caused by thermal agitation of the particles (atoms, molecules, electrons) inside the hot material as moving particles will emit radiation according to Maxwell's laws.

A blackbody radiator is an object of which the emitted radiation is solely determined by the temperature of that object. Therefore, a blackbody radiator will not reflect any radiation. It is a perfect absorber. Planck was the first to calculate the blackbody radiation spectrum (figure 17.7):

$$M_S^e(\lambda) = \frac{8\pi hc}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1} \tag{17.19}$$

This spectrum shows a maximum determined by *Wien's law*:

$$\lambda_{\max} \approx \frac{2.9\,[\mu m]}{T\,[1000K]} \tag{17.20}$$

The 'color' of a blackbody radiator is thus solely determined by its temperature. It does not matter how this temperature is achieved, for example by absorbing external radiation or by internal energy production. The sun is a blackbody radiator at a temperature of approximately 6000 K.

Real objects typically are "grey body radiators". How close a real object can resemble a blackbody radiator is determined by its emissivity ϵ. It is a measure of a material's ability to radiate absorbed energy. Absorptivity describes how easily a material absorbs the incident radiation. Kirchhoff's law states that absorptivity equals emissivity. This means that the more reflective a material is, the lower its emissivity.

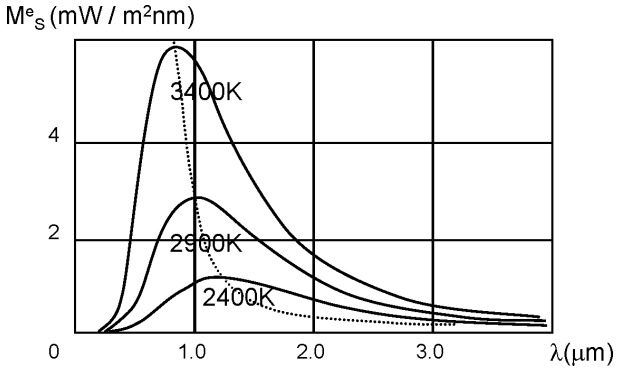

Figure 17.7: The blackbody radiation spectrum.

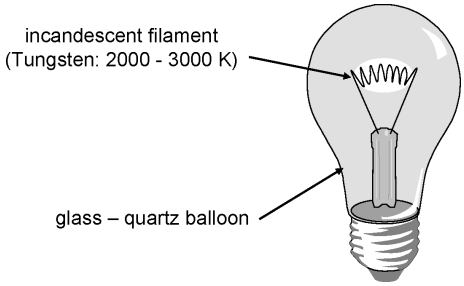

Figure 17.8: The incandescent lamp.

The total radiant exitance is simply written as the law of Stefan-Boltzmann:

$$M^e = \sigma T^4 \tag{17.21}$$

with $\sigma = 5.67.10^{-8} \frac{W}{m^2 K^4}$ the constant of Stefan-Boltzmann.

17.4.2 Incandescent lamps

An incandescent lamp consists of a balloon of glass or quartz (which is vacuum or filled with an inert gas) in which an incandescent filament is placed that is heated to high temperatures by the Joule effect (figure 17.8). To a first approximation the filament can be considered as a blackbody radiator. Wien's law then states that the higher the temperature is, the greater the part of radiation will be that is located in the visible range. The sun has a temperature of 6000K and will thus emit nearly perfectly white light.

In incandescent lamps, the filament usually consists of Tungsten heated to a temperature of 2000 to 3000 K. Higher temperatures would evaporate the filament too quickly. The radiation maximum still lies in the infrared. The light output is typically 20 lumen per Watt (electric power). Remember that a 100% efficient lamp delivers 680 lumen per Watt. The lifetime of an ordinary incandescent lamp is a few 1000 hours. In high power lamps, the balloon is always filled with a noble gas. This reduces the evaporation of the filament but causes another problem: heat losses due to conduction in the gas. This is partially solved by winding the filament into a spiral.

In halogen lamps, the balloon is filled with a halogen, usually iodine (figure 17.9). The evaporated tungsten atoms, together with this iodine, then form tungsteniodide (WI_2) in the parts of the lamp

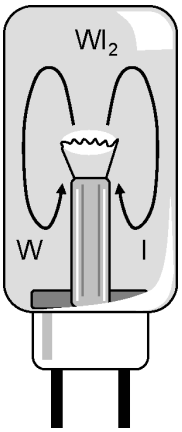

Figure 17.9: The halogen lamp.

where the temperature is below $500°$C. In the vicinity of the filament, having a temperature of 3000 K, the WI_2 will dissociate again causing an increase of the concentration of tungsten atoms near the filament, which counteracts the evaporation of the filament. The lifetime or the luminous flux per Watt is therefore 25% larger than for a normal incandescent lamp. The high temperature of the surface of the balloon, in order to have a good operation, is a practical problem.

Halogen lamps are often very compact and usually operate on low voltages (at least for low powers). The light is very white because of the high temperatures. All this makes the halogen lamp an appropriate candidate for decoration purposes. Halogen lamps are often dimmed although it is useful to know that dimming the lamps can effect the lifetime negatively. Halogen lamps are also omnipresent in the headlights of cars.

17.5 Gas discharge lamps

A gas discharge consists of an electric current through a gas or metal vapour. The light emission in a gas discharge is caused by the spontaneous transition of an atom in an excited state to a lower energy level (figure 17.10). The released energy is then emitted as an electromagnetic energy quantum $h\nu = E_1 - E_0$. The frequency of the emitted light is thus given by:

$$\nu = \frac{E_1 - E_0}{h} \tag{17.22}$$

The advantages of gas discharge lamps compared to incandescent lamps are amongst others:

- larger efficiency
- longer lifetime (10000 hours and more)
- lower temperature

Because the VI characteristic of a gas discharge lamp displays a negative resistance characteristic, a stabilization-resistor or self-induction (in case of alternating current) has to be provided ('ballast').

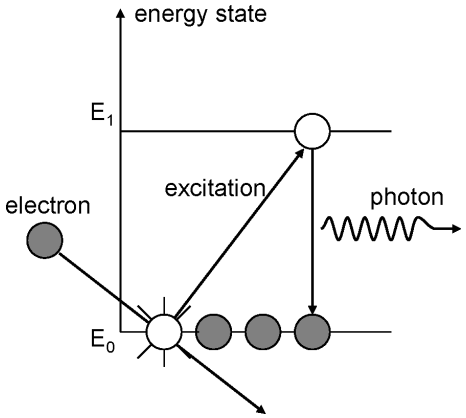

Figure 17.10: Principle of the gas discharge lamp.

The ignition of a discharge lamp can happen by a combination of the following elements:

- short but large voltage pulse by:
  - disruption of the current in an inductive circuit.
  - resonance in a tuned circuit.
- addition of a noble gas (neon or argon) in metal vapours.
- heating of the electrodes until thermal electron emission occurs.

Figure 17.11 shows a typical circuit diagram that is often used in TL lamps. Initially the entire voltage is across the glow starter. This starter is a small discharge lamp, usually filled with helium and hydrogen. This causes a gas discharge and heat is thus generated. This heat will curve the bimetallic electrode of the starter until contact is obtained with the other electrode. A current then flows through the resistance electrodes of the main discharge lamp. These electrodes warm up and ionize the surrounding gas. The bimetallic cools down after a few seconds and the circuit is disrupted, causing a large voltage across the lamp (due to the inductive ballast). A full gas discharge now arises in the lamp, causing a continuous flowing current. The final operating voltage across the lamp and the starter is too small to ignite this latter one. The capacitance C_1 improves the work factor ($\cos\phi$) of the whole circuit. Different types of discharge lamps are available according to the nature of the gas.

17.5.1 Low pressure Sodium lamps

The electric energy is converted in two resonance radiations of Na: one at 589.0nm and one at 589.6nm. The optimal conditions are 5μbar for the pressure and $270°$C for the temperature. The lamp usually consists of an U-shaped tube filled with sodium and neon (to start the discharge). The applied voltage is big enough to get a discharge of the neon in the cold lamp. This will cause the sodium to evaporate and to participate in the current conduction. First the lamp burns red and afterwards yellow. Thermal isolation is important. Therefore the U-tube is placed vacuum in a second tube. High efficiency: 140 lumen/Watt. Application: monochromatic, orange-yellow light, mainly used in traffic lighting.

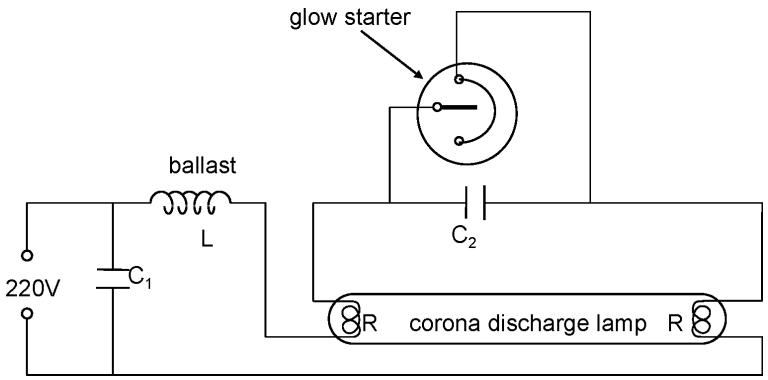

Figure 17.11: Ignition circuit of a TL lamp.

17.5.2 High pressure Sodium lamps

High pressure sodium lamps contain a mixture of sodium and mercury, with a small amount of xenon. The lamp is ignited with a short voltage pulse that causes a discharge in the xenon gas, having a bright white-blue color. After a few minutes, the warming up of the discharge has evaporated the sodium and the mercury. These will carry the largest part of the discharge current. The lamp then emits an orange-white light. Color-corrected lamps have recently come onto the market, strongly approaching daylight. The efficiency is a lot lower however. High efficiency: 80-120 lumen/Watt. Application: street lighting, outside lighting.

17.5.3 High pressure Mercury lamps

The mercury vapour is at a pressure of 1 to 20 bar. At 1 bar the light consists of a few powerful spectral lines, while a continuous spectrum joins these lines at high pressure. The temperature of the discharge is 6000K and the emitted light is whitish. The temperature is typically $800°$C at the outer surface and the whole is therefore placed in a balloon filled with nitrogen. The output is 30 to 50 lumen/Watt. At low pressure (< 1 bar) UV-light is created especially. The balloon is often made fluorescent by putting Ba-silicates on the surface. Then the UV-light is converted into visible light. Application: street lighting, lighting of large, high spaces.

17.5.4 Fluorescent lamps

(TL-lamps) These lamps consist of a glass tube filled with mercury vapour at very low pressure (7μbar). Mainly one line is excited: 253.7nm (non-visible: UV). A fluorescent layer, put on the inner surface of the tube, converts this UV-light into visible light. The composition of this layer (Zn-silicate, Cd-silicate) determines the color of the emitted light. The spectrum of a fluorescent lamp consists of a continuous spectrum increased with a few lines. The output is typically 70 lumen/Watt. The fluorescent lamp exists in tube-shape, but also in a more compact shape (energy-saving lamps), which makes it compatible w.r.t. dimensions with incandescent lamps. In recent years, a new variant of the fluorescent lamp has been developed: the induction lamp. A gas discharge is aroused with RF magnetic induction by means of a coil outside the lamp. No electrodes are thus present and there are barely any signs of wear in the lamp itself (100000 hours).

17.5.5 Xenon lamp

This lamp is filled with pure Xenon at high pressure. The electrodes are brought close to each other and an extremely intense blue-white spark arises. The luminance of this spark can be higher than the luminance of the sun. This compact lamp is more and more used in headlights of cars (replacing the less efficient halogen lamps).

17.5.6 Metal Halide lamp

This is a low pressure mercury lamp to which halides of metals like Thallium, Indium or Sodium are added. The lamp produces an intense white light that is very close to sunlight. This lamp is therefore very suitable for work places.

17.6 Light emitting diodes (LED)

LEDs are semiconductor components (usually made out of III-V semiconductors like GaAs) that have an efficient radiative recombination process. Although LEDs were, until now, mainly used for telecom applications or indicators in electronic devices, the first lighting applications of LEDs have already come onto the market. Lighting with LEDs is often called Solid State Lighting (SSL). White LEDs are realized by putting a phosphor layer on blue or UV LEDs, as the spectrum of LEDs is relatively limited. These kind of LEDs are already found in flashlights, car headlights, etc. In certain applications LEDs have started to replace the traditional lamps. Especially when durability, compactness and efficiency are an issue, LEDs are introduced.

Besides light emitting diodes based on semiconductor materials, organic light emitting diodes for lighting are a booming research topic. First commercial applications are already appearing (e.g. 40 inch OLED television screens). The advantage of organic LEDs is that they can be produced on large surfaces at a low cost. The biggest technical problem for OLEDs however is the limited lifetime of the organic materials.

Bibliography

Chapter 18

Displays

Contents

An important application of photonics and one with which we are confronted daily, are displays. This term denotes the technology that enables us to visualize information in a dynamic way. To this end there is a great variety of techniques. In this chapter we first briefly discuss the (visual) human perception, and especially the ability to see colors. Afterwards we discuss the different display technologies.

18.1 The human vision

18.1.1 The eye and the retina

The eye is the most commonly used sense of people to get an idea of their environment, or to perceive information from their surroundings. Figure 18.1 shows a schematic representation of the human eye. An image of the light coming from our surroundings is made by the eye lens onto the light sensitive retina. We already discussed the eye as an optical image system in chapter 3.

The retina is the light sensitive element of the eye. It converts light intensities in electrochemical impulses in the eye nerve that can be interpreted by the visual cortex. A cross section of the eye is depicted in figure 18.2. The retina contains two types of light sensitive cells: the ca. 120 million *rods* perceive an intensity image (grey image) of the environment (scotopic sight). There exist three types of the ca. 6 million *cones*, mainly concentrated around the yellow spot: red, green and blue. These are responsible for the perception of color (photopic sight). The retina consists of different layers of cells. In contrast with our intuition, the light sensitive cells are not on top, but buried under a number of supporting cells. The nerves that transport the signal to the eye nerve (*blind spot*) are on top.

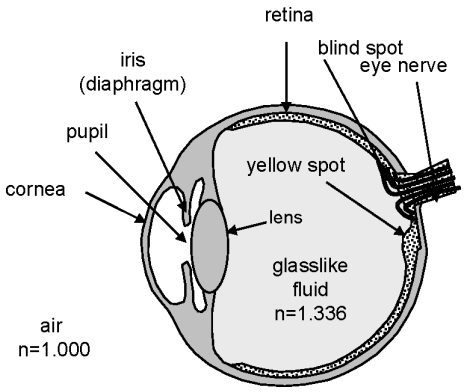

Figure 18.1: The human eye. The yellow spot is the most sensitive area of the retina.

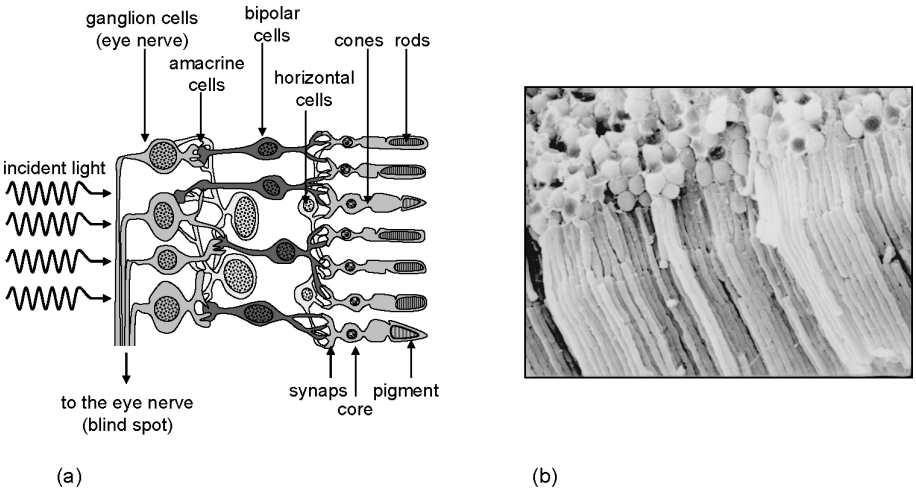

Figure 18.2: Structure of the retina. (a) The different layers of cells. (b) Rods.

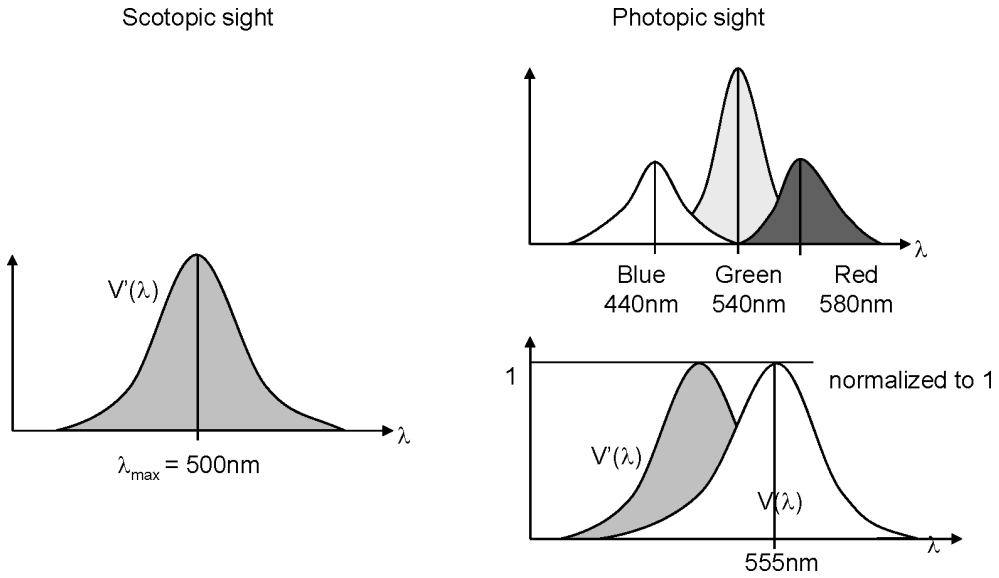

Figure 18.3: Spectral eye sensitivity curve of the rods and the cones in the human eye.

In chapter 2 we discussed the photometric quantities as well as the standardized eye sensitivity curves for the rods and the cones. These are once more depicted in figure 18.3. We notice that the maximal sensitivity of the cones and rods lie at different wavelengths.

18.1.2 Responsivity of the retina

The human sight has a certain slowness. This is caused by chemical processes in the retina itself as well as the actual processing in the brain. A perceived view 'stays' a finite time in the brain.

This slowness is very important for viewing applications. The brain can interpret a sufficiently quick succession of static images as movement. Joseph Plateau already comprehended this principle (see chapter 1). The threshold lies around 16 frames per second.

Display applications that have to produce a moving image, need a sufficiently high so-called *refresh rate*. In cinema, this rate is 24Hz, for television in Europe 25Hz (America: 30Hz) and most computer screens have a refresh rate of over 60Hz.

18.1.3 Depth of sight and parallax

Humans have got two eyes, separated approximately 8 cm from each other. This enables us to deduce depth information from a two-dimensional image on the retina. Let us look at the example in figure 18.4. The tree is further away from the observer than the girl. Because both eyes see the scene from a slightly different angle, the image of the girl is positioned differently for both eyes compared to the image of the tree. This phenomenon is called parallax.

This shift of both images enables us to calculate the distances between the different objects using simple trigonometry. The brain fulfills this task very efficiently.

Figure 18.4: Depth of sight using parallax. The left eye (b) and the right eye (c) perceive a different image of the scene (a).

18.2 Colorimetry

18.2.1 Primary colors

The eye is sensitive to colors. However, it is not able to decompose the light in its spectral colors. As mentioned, the light sensitivity arises because the eye contains three different types of cones, each with its own spectral sensitivity (figure 18.5). These three receptors are excited differently according to the spectrum of the incident light. We can say that the maximal sensitivity of the receptors lies roughly at blue, green and red, respectively. Each combination of receptor stimuli causes a certain color impression. Each color impression corresponds to one point in the three-dimensional space formed by the three receptor intensities. This means that the eye strongly reduces the amount of information in the spectrum of the incident light to only three quantities. This also implies that two different spectra can cause the same color impression, as long as they excite the three types of receptors the same (the two spectra then form a metameric pair).

If we have three light sources, each mainly exciting one type of receptor, we can synthetically generate every color impression (figure 18.6a). This is called additive mixing of the three ground colors or primary colors. Red, green and blue are used for this. E.g. illuminating a reflective screen equally with red and green light, gives the impression of yellow light. A color television screen is also based on additive mixing. The different types of light dots on the screen that are placed next to each other, generate the impression of a uniform color as the resolution of the eye is too small to distinguish the dots separately. Notice that complementary colors are colors that give white light after mixing them (or better: the impression of white light).

An alternative to additive mixing is subtractive mixing (figure 18.6). Then different light sources are not added together, like in additive mixing, but we begin with white light of which parts are removed. To obtain every color with subtractive mixing, other primary colors are used, namely colors that excite two of the three types of receptors but not the third one. The subtractive ground colors are thus simply obtained by mixing two regular ground colors:

- blue + green gives cyan
- green + red gives yellow

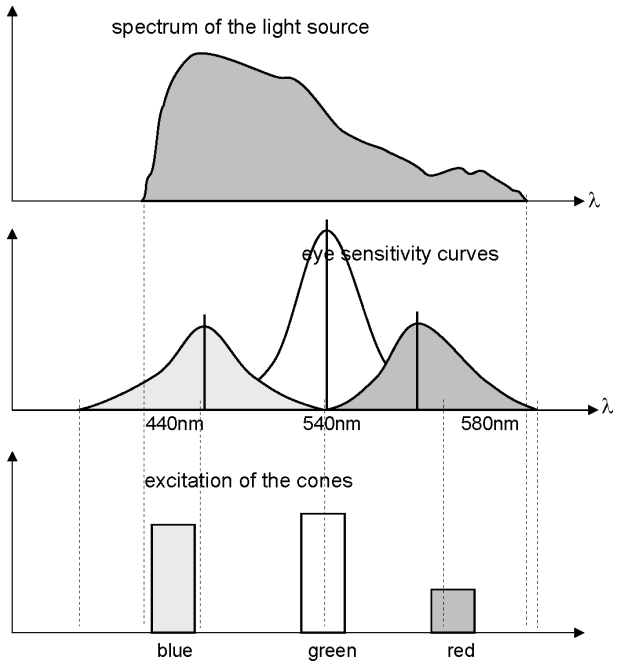

Figure 18.5: Conversion of the spectrum of a light source to color values perceived by the eye.

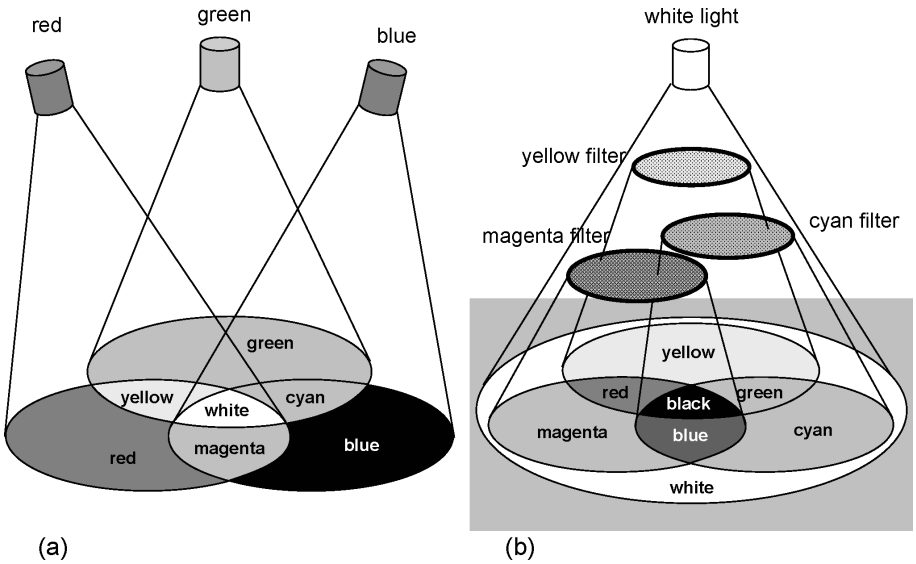

Figure 18.6: Mixing of the primary colors in order to get different hues. (a) Additive color mixing, (b) Subtractive color mixing.

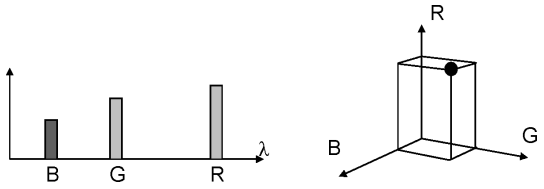

Figure 18.7: Color coordinates.

- blue + red gives magenta

A material having such a subtractive ground color, e.g. cyan, in fact absorbs the third ground color (red). If objects that are illuminated with a white light source appear colored, it is caused by a subtractive process: certain parts of the spectrum are absorbed by the object. Subtractive processes are amongst others:

- looking at a white light source through a number of color filters in series (each filter absorbs a part of the spectrum)
- mixing of paint (each color pigment absorbs a part of the spectrum)
- color prints (several prints are made with cyan, yellow and magenta above each other, black is eventually added for contrast)

18.2.2 Colorimetry

As the eye is very color-critical, one has studied methods to quantify color impressions. As discussed in the previous paragraph, this can be done by choosing three light sources with primary colors and subsequently determining the needed intensity of each to mimic the color impression of a given spectrum by adding the colors additively. So three color coordinates are obtained that give a color impression. The question now is: what is the best choice for these three basic colors? Before going in to this, it is important to notice that such color coordinates may be added linearly (this is a physiological observation). This means that when looking at two spectra and determining the color coordinates of each, additive mixing of these two spectra will have a set of color coordinates that is the sum of the two original sets. Let us now look at a color coordinate system based on spectral (monochromatic) ground colors: a red, a green and a blue spectral line (figure 18.7). We can now ask ourselves whether or not we can copy each spectral color, starting from these spectral ground colors. If I_r, I_g and I_b represent the intensities of the three ground colors and I_λ the intensity of the other spectral color with wavelength λ, it seems to be that always one of the following metameric pairs can be formed:

$$I_\lambda + I_r = I_b + I_g \tag{18.1}$$
$$I_\lambda + I_g = I_r + I_b \tag{18.2}$$
$$I_\lambda + I_b = I_r + I_g \tag{18.3}$$

The equal sign denotes metameric equivalence, while the plus sign denotes additive mixing. The following question now imposes itself: can we create the new spectral color using additive mixing

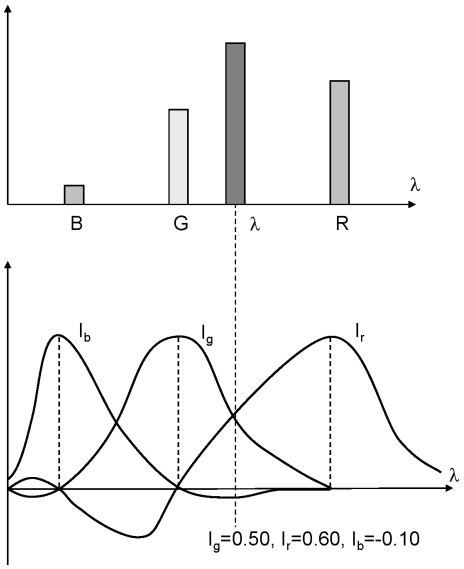

Figure 18.8: Color coordinates.

of the three basic colors, in other words,

$$I_\lambda = I_r + I_g + I_b \tag{18.4}$$

The metameric equivalences above state that this would be possible if one of the three intensities may be negative. This means of course that additive mixing is physically not possible, but mathematically each spectral color can be represented by a set of coordinates of the three spectral colors. These three coordinates are represented in figure 18.8 (for a certain choice of red, green and blue).

Because the choice of spectral colors leads to negative coordinates for a large number of colors, a new set of ground colors has been defined in 1931 by the *Commission International de l'Éclairage* (CIE) that always results in positive coordinates. These coordinates are denoted as X, Y and Z. X is the coordinate for the new red ground color, Y for the green and Z for the blue color. These coordinates are linearly related with the coordinates based on the spectral ground colors. Furthermore, the green ground color has been chosen in a way so that its spectral sensitivity approaches the one of the human eye. The Y-coordinate is then also a measure for the luminous flux in lumen. We can also normalize the (X, Y, Z) coordinates:

$$x = \frac{X}{X+Y+Z} \tag{18.5}$$

$$y = \frac{Y}{X+Y+Z} \tag{18.6}$$

$$z = \frac{Z}{X+Y+Z}. \tag{18.7}$$

x, y and z thus denote the relative contribution of the three ground colors. Naturally, two of these coordinates are enough to define a color. This has the advantage that the colors can be represented in a two-dimensional plane (usually the xy-plane): the 'CIE chromaticity diagram'. This is depicted in figure 18.9. We can indicate the spectral colors in this figure. They form a horseshoe-shaped line. Due to the definition of the CIE XYZ-system, all these spectral colors have positive

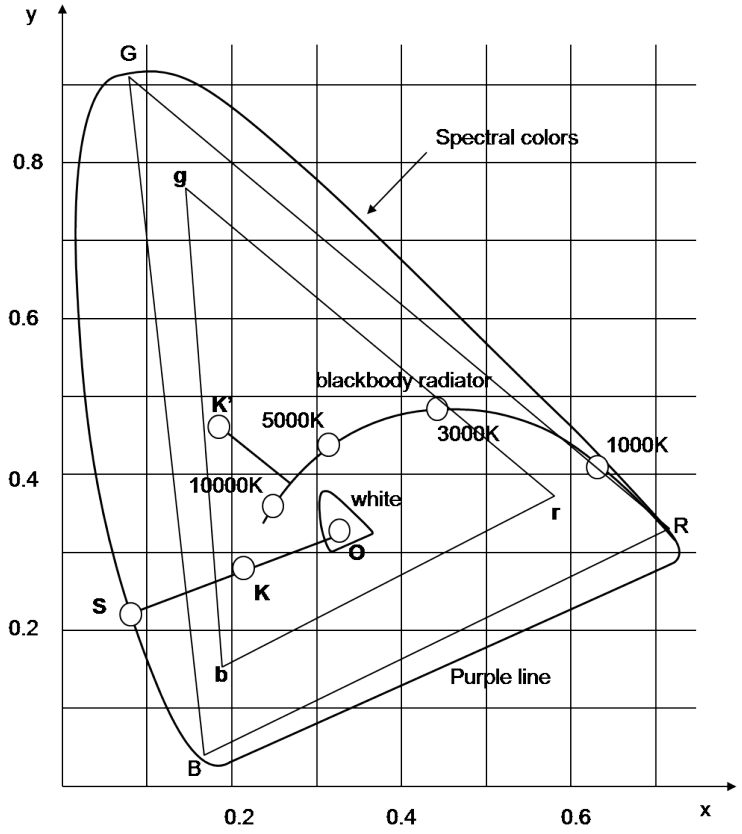

Figure 18.9: CIE chromaticity diagram, or the Yxy-coordinate system.

coordinates. This implies immediately that all existing colors (formed by an additive mixing of spectral colors) are located *inside* the horseshoe. It implies furthermore that the basic colors red, green and blue, that form the base of this diagram (the unit vectors) do not exist physically. The line that connects the two end points of the horseshoe is called the purple line. When two colors are chosen in this xy-plane, we can form by additive mixing each color that lies on the line between these two points. When we take three ground colors, we can reach each point inside the triangle formed by the three ground colors. Such a triangle for the spectral colors is shown in figure 18.9. We see that most of the other spectral colors lie outside this triangle, which confirms the previous: we can not copy the color of other spectral colors with three spectral ground colors. The color range we can obtain with a certain set of basic colors is called the 'gamut'.

The 'hue' and 'chroma' of an arbitrary color is also defined. If K is an arbitrary color (see figure 18.9), then the dominant hue of this color is the spectral color S. The chroma is given by the ratio of the lines $|OK|/|OS|$, with O the point that represents white light (approximately located in the point $x = y = z$). Next to the hue and chroma, an arbitrary color is also characterized by its 'value' of light or dark, determined by the quantity Y.

The color of a blackbody radiator as a function of its temperature is also represented in the same figure. We notice how the color evolves from red to white and finally to blue. Thus we can define a color temperature for broadband light sources. This is the temperature a blackbody would need

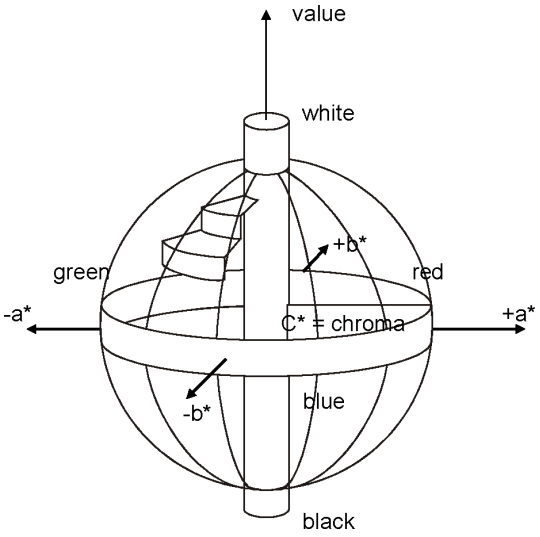

Figure 18.10: $L*a*b*$ color coordinates.

to create a similar color effect as the light source itself. Such a definition of course only makes sense if the spectrum of the light source is approximately equal to the one of a blackbody.

The Yxy-system is not the only used coordinate system. A disadvantage of this system is the fact that the 'distance' between two colors $\sqrt{x^2+y^2}$ is not at all a good measure for the physiologically sensed color difference. Alternative color systems are the $Yu'v'$ and the $L*a*b*$-system. The $Yu'v'$ system (standardized by CIE in 1976) is simply related to the Yxy-system as follows:

$$u' = \frac{4x}{-2x+12y+3} \tag{18.8}$$

$$v' = \frac{9y}{-2x+12y+3} \tag{18.9}$$

$$\tag{18.10}$$

The $L*a*b*$-system (figure 18.10) is strongly based on the concepts 'hue', 'chroma' and 'value'. The $L*$-coordinate is a measure for the 'value'. $a*$ and $b*$ together represent the 'hue' and 'chroma'. The 'chroma' is given by the quantity $C* = \sqrt{a*^2 + b*^2}$.

$a*$ and $b*$ may be positive or negative. The $-a*/+a*$-axis runs from green to grey and then to red. The $-b*/b*$-axis runs from blue to grey and then to yellow. The conversion formulas between $L*a*b*$ and XYZ are given by the following expressions:

$$L* = 116\sqrt[3]{\frac{Y}{Y_0}} - 16 \tag{18.11}$$

$$a* = 500\left[\sqrt[3]{\frac{X}{X_0}} - \sqrt[3]{\frac{Y}{Y_0}}\right] \tag{18.12}$$

$$b* = 200\left[\sqrt[3]{\frac{Y}{Y_0}} - \sqrt[3]{\frac{Z}{Z_0}}\right], \tag{18.13}$$

$$\tag{18.14}$$

in which X_0, Y_0 and Z_0 are the coordinates of the light source that illuminates the object that has to be characterized. The $L*a*b*$-coordinates are in that way characteristic for the object and rather independent of the illumination. An important advantage of the $L*a*b*$-system is the fact that the distance between two colors, defined as $\sqrt{L*^2+a*^2+b*^2}$ is a good measure for the sensed color difference.

18.2.3 Color rendering index

The *color rendering index* (CRI) (often denoted as R_a) is a quantitative measure of the ability of a light source to reproduce the colors of various objects being lit by the source. The best possible rendition of colors is specified by a CRI of 100, while the very poorest rendition is specified by a CRI of zero. For a source like a low-pressure sodium vapor lamp, which is monochromatic, the CRI is nearly zero, but for a source like an incandescent light bulb, which emits essentially black-body radiation, it is nearly a hundred. The CRI is measured by comparing the color rendering of the test source to that of a 'perfect' source which is generally a black body radiator. The precise definition is beyond the scope of this course.

18.3 Display technologies

Displays have become an important instrument in the present day information society. Therefore it is a field that undergoes a rapid technological evolution. In this section we briefly describe the different technologies, after explaining a few important concepts.

18.3.1 Important aspects of a display

Resolution

The term *resolution* is used in optics to denote the detail size by which something can be observed. In the display technology, this term is used to denote the number of separate image points or *pixels* (picture element) that can be represented. The larger the number of pixels, the better one is able to represent fine details. Sometimes resolution is expressed in image lines.

Refresh rate

This gives us the number of times per second (unit: Hertz) that the image is regenerated. This is important because the refresh rate has to be sufficiently high for the eye to observe a continuous moving image. Television images are transmitted at 25Hz. However, to obtain a more stable image, the transmission is interlaced: first the even image lines are represented, then the uneven lines. Because of this, the refresh rate seems doubled.

Gamut

The color range that a display can represent depends strongly on the used display technology. An accurate color reproduction is especially important in the graphic world.

Scanning

A lot of display technologies use *scanning* to create an image. The pixels are hereby sequentially updated (very short pulses). The image is formed serially. This is then continuously repeated. Phosphors e.g. are used to maintain the image long enough until the pixel is updated again.

Active matrix

In an *active matrix* display the pixel actively maintains its own state until it is updated again. Transistors that are incorporated in the pixel, are usually employed for this.

18.3.2 Photography and cinema

The first techniques to reproduce images accurately used an irreversible chemical reaction to capture an image. This principle is employed for over a century now in photography and cinema.

To create 'moving' images, like in most movie theaters, a *stop-and-go* mechanism is used: the image is brought before the lens, the film is then stopped, the image is represented and afterwards the film moves one frame further. A diaphragm is used during the movement so that the eye only gets to see a succession of stationary images. This process repeats itself 24 times per second for classic cinema. It is obvious that the fraction of time that the image is visible has to be sufficiently large.

The major disadvantage of classic cinema projection is the fact that it is a mechanic process. This does not only cause wear to the equipment, but considerable forces act on the pellicule during the stop-and-go process. Therefore this technique is avoided more and more in favor of digital projection.

18.3.3 The cathode ray tube

The *cathode ray tube* (CRT) has been the most wide-spread display technology until recently and was mainly used for television sets and computer screens. Because of the development of new technologies, the market share of CRT screens is decreasing drastically.

General principle

A schematic principle of a cathode ray tube is depicted in figure 18.11. A high voltage field is applied in a vacuum tube between a cathode and an anode, that also acts as the screen. The cathode is heated which causes electrons to escape. These electrons are drawn to the anode. Steering

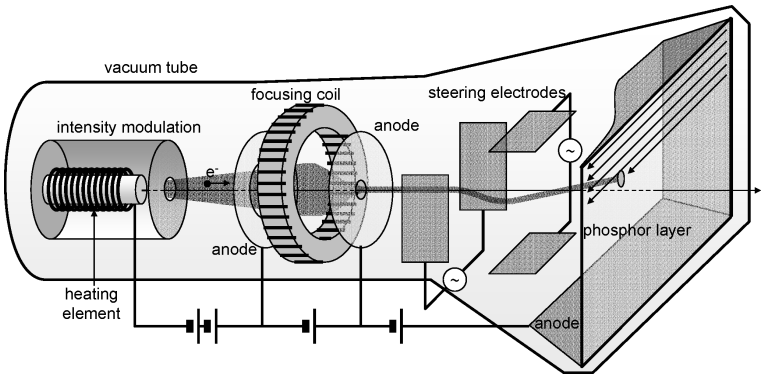

Figure 18.11: The cathode ray tube.

electrodes apply a lateral electric field to direct the electron bundle onto a specific place on the anode. In this way the electron beam *scans* the anode. The latter contains a phosphor layer that illuminates when the electron bundle impinges. By modulating the intensity of the electron bundle, synchronous with the steering electrodes, an image can be created on the screen.

The phosphor layer plays an important role. Although the electron beam scans the image very quickly (each phosphor point is illuminated during 1/15000 of a second), the phosphor will emit light during a longer time. Therefore the image remains visible before it is being refreshed

Color

To represent color on a CRT-screen, a green, red and blue image is projected simultaneously. Three different layers of phosphor are therefore put on the anode. The pattern depends on the used technique.

Originally a combination of a triangular pattern and a shadow mask (figure 18.12a) was used for color reproduction. Three electron guns were therefore steered simultaneously. Because of their different initial positions, the electron bundles will reach the anode under a different angle. A shadow mask is located there. This consists of one hole for each group of RGB color pixels in the phosphor pattern. This hole will create a different 'shadow' on the phosphor screen for each electron bundle, just at the right place of the right color phosphor.

The shadow mask is a very simple method to create color images. This method is however not very efficient as the greater part of the electron bundle is blocked by the shadow mask. This not only requires a larger current for the same light intensity, but also strongly heats up the shadow mask. The mask is therefore always made out of INVAR, a material with a very low thermal expansion coefficient.

Nowadays, an *aperture grille* is used more and more instead of a shadow mask. This is illustrated in figure 18.12b. Very fine metal wires are hereby tightened vertically. The red, green and blue phosphor pixels are now grouped horizontally. This was originally developed by Sony under the name *Chromatron*. Only one electron bundle is used and a voltage is applied between two neighbouring wires to sequentially direct the bundle to the red, green and blue pixels. This steering mechanism is however complex and susceptible to disturbances.

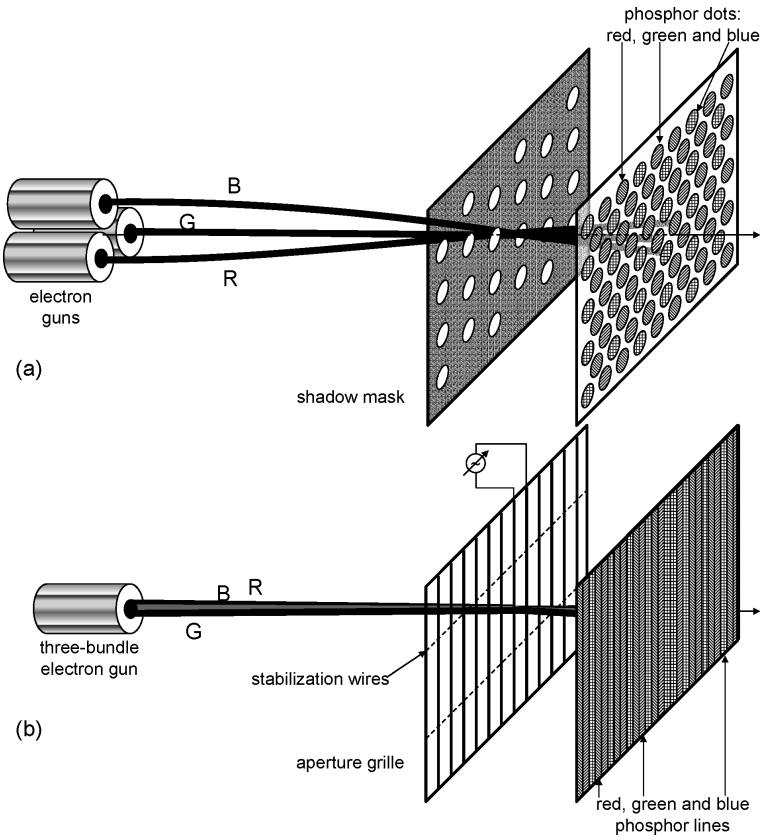

Figure 18.12: Color cathode ray tube. (a) Shadow mask. (b) Aperture grille.

For the time being, such an aperture grille is still used, but as a shadow mask (especially known under the Sony brand name *Trinitron*). Three electron guns are then again used (or one gun with three bundles). The advantage compared to a shadow mask, is that a large fraction of the electrons reaches the phosphor and that the thermal expansion is compensated by the tension in the wires. Vibrations are however insufficiently dampened due to the suspension in a vacuum. Two or more horizontal stabilization wires are therefore placed, dependent on the size of the screen. These wires are visible as fine horizontal lines on the screen.

Conclusion

Although the market share of CRT-screens is decreasing, they still give the best color reproduction because of the high quality of the phosphors.

Cathode ray tubes are also used for projectors. Three different cathode ray tubes are then used for RGB. The different images are then projected with lenses on a screen. The three colors have got to be well aligned of course. Although such projectors are clearing the path for alternative technologies, they are still used for large events, in which high powers, high light intensity and a high resolution is needed. Furthermore deep black can be obtained, a difficulty in LCD- and DLP-based projectors (see further).

18.3.4 Field emission displays

An important disadvantage of classic CRT-screens is the depth of the electron ray tube. Although improvements are continuously being made, the fact remains that the length of the tube increases proportional to the width of the screen as the electron bundle has to be able to reach the outer corners of the screen. With the common tendency for larger displays, CRT screens become unmanageably large.

Instead of scanning all the pixels with one single cathode ray tube, we can provide each pixel with its own 'electron gun'. Then the bundle does not have to scan, and the screen can be made less deep. Of course ten thousands of electron emitters have to be provided that together require approximately the same power as the original electron gun.

A possible solution was sought in field emission. When applying a sufficiently large electric field, electrons can be 'pulled out' of a material. A strong electric field needs to be created that enables the electrons to gain enough energy from the electric field $\mathbf{E}$ to overcome the work function W of the anode (see also chapter 15).

$$e\nabla\mathbf{E} > W. \tag{18.15}$$

In a classic cathode ray tube, this happens by applying a high voltage and warming up the cathode.

The first (experimental) *Field Emission Displays* (FED) are based on a cathode with a very sharp tip. At the tip, a lot of field lines will come together on a small surface, so that no high voltages are required to obtain sufficiently high field strengths. This principle is represented in figure 18.13a. A ring-shaped secondary anode is used to obtain a strong field concentration around the tip. The electrons are then accelerated towards the primary anode, that is provided with a phosphor layer.

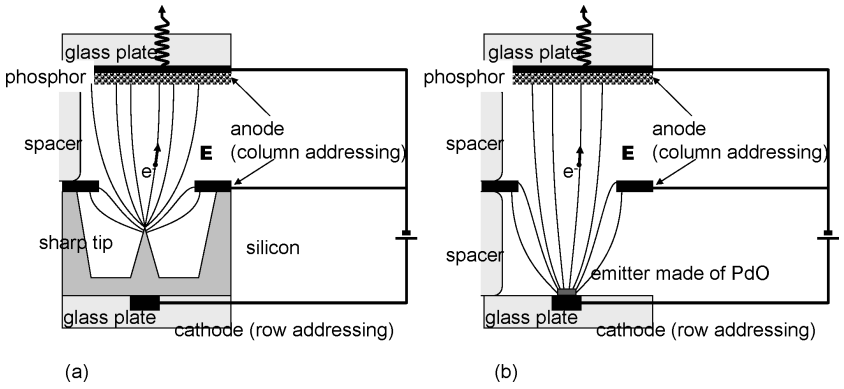

Figure 18.13: Principle of a field emission display. (a) Field emission at a sharp tip. (b) Surface emission.

These structures can be fabricated with lithography and etching processes. The sharp tip is usually made of silicon.

This technology had to deal with two important difficulties. The lower voltages that are used in a FED cause the electrons to impact the phosphors with lesser energy than in a CRT-screen. There has been a lot of research to develop phosphors that also had a good light efficiency at these lower energies. It appears on the other hand to be very difficult to fabricate a sharp tip that can withstand electric currents. The point often becomes blunt after an unacceptably short time and the field emission effect is then lost. The latter problem prevented the field emission displays from reaching a commercial stage.

Recently new life was brought to the idea under the name of *Surface Emission Displays* (SED). A sharp tip is no longer used, but a material with a very low work function, palladiumoxide (PdO). The difficulty of making a sharp tip is now gone. Displays based on this principle would be ready for the market in 2006 and could compete with LCD displays and plasma screens. An important advantage of this technology is that they can produce strongly saturated colors and deep black like CRT-screens.

18.3.5 Plasma screens

An alternative for CRT-screens is the plasma screen. Every pixel is controlled individually like in a FED. The phosphors are however not excited by an electron bundle, but by UV-radiation from a plasma discharge.

A pixel of a plasma screen is depicted in figure 18.14. A pixel consists of a small chamber filled with gas of which the walls are covered with red, green or blue phosphors. By applying a voltage between the two electrodes, the gas in the chamber is ionized and a plasma is thus created that emits UV-radiation. The UV-radiation is converted by the phosphor into visible light. This discharge can take place thousands of times per second. By increasing that frequency, the intensity of the pixels can be regulated.

Plasma screens have a very good image quality. They are however costly to produce, and therefore occur in the more expensive market of home cinema systems. They also require a lot of power compared to other technologies.

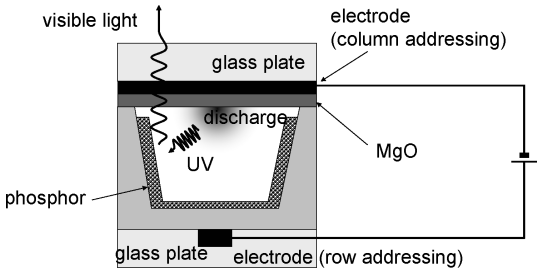

Figure 18.14: Plasma screen: the UV-light of a gas discharge is converted by the phosphors into visible light. A pixel consists of such a cell for red, green and blue.

18.3.6 Liquid Crystal Displays

At present cathode ray tubes make place for flat screens based on liquid crystals (LCD: *Liquid Crystal Display*).

Liquid Crystals

Liquid crystals are a group of materials with a number of special properties. As the name already mentions, they have properties of a liquid as well as of a crystalline material.

Liquid crystals consist of long stretched molecules, that have the tendency to align themselves in a regular way. Dependent on the type of crystal, the molecules align themselves in the same direction (a so-called *nematic* liquid crystal), and sometimes even in a regular position in space. At a molecular level, a liquid crystal acts as a regular material, in other words as a crystal. At a macroscopic level, a liquid crystal is a liquid, that can be poured from one recipient to another, or that can be 'sucked' between two plates because of capillarity.

The preferential direction of the liquid crystal molecules is influenced by external factors. In figure 18.15a, we see how a liquid crystal aligns itself with a plate with a grooved pattern. When we bring an amount of liquid crystals between two plates with grooves that are perpendicular to each other (figure 18.15), the molecules at the edges will align themselves with the grooves and the molecules in the bulk will gradually change direction to enable a transition between the boundary conditions.

Liquid crystals also react to an external electric field. When we use the contact plates as electrodes and apply a voltage, the molecules try to align themselves with the electric field.

Liquid crystals have interesting optical properties. Because of molecular anisotropy, a liquid crystal has a different refractive index for different polarizations (so-called double refraction). If polarized light is incident on a layer of liquid crystals, the polarization of the outgoing light depends on the orientation of the liquid crystal molecules.

LCD

Liquid crystal displays are based on the rotation of the polarization in a layer of liquid crystals. The principle is depicted in figure 18.16. Light coming from a white light source is sent through a

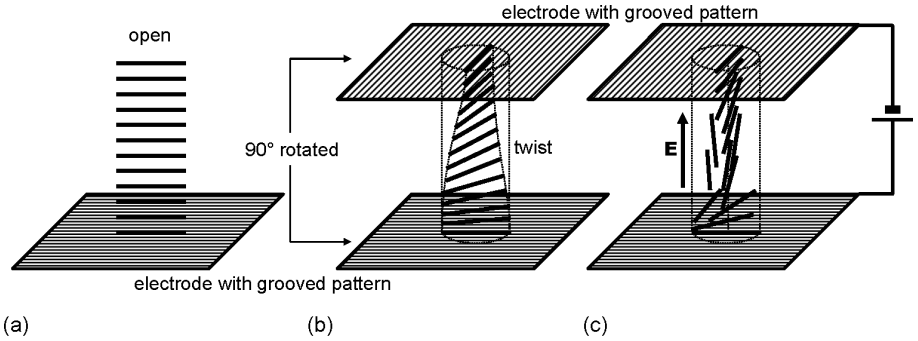

Figure 18.15: Nematic liquid crystal between electrodes with a grooved pattern. (a) If only a single plate is present, the crystal aligns itself with the grooves through the entire material. (b) When placing a second electrode and without applying a voltage, the molecules near the electrodes align themselves with the grooves. The molecules in the bulk turn continuously. (c) When applying an electric field, the molecules in the bulk align themselves with the electric field.

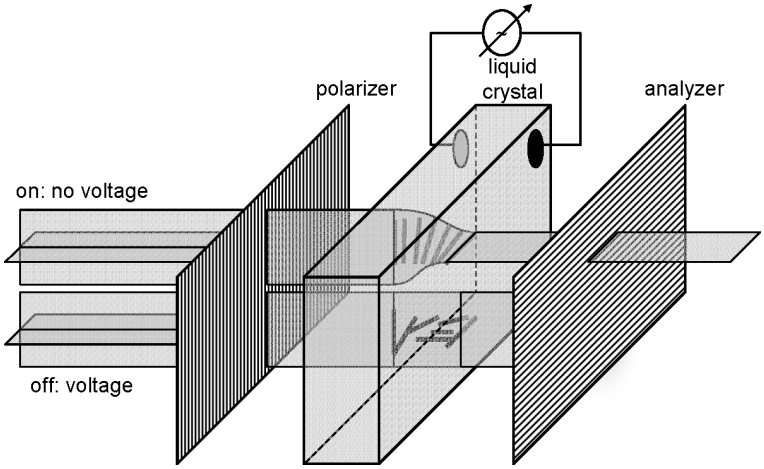

Figure 18.16: Display based on liquid crystals and polarizers.

polarizer that is linearly polarized and through a liquid crystal layer divided in pixels that can be separately controlled electrically. Afterwards the light passes through a second polarizer, perpendicular to the first one that is called the analyzer.

The thickness of the liquid crystal layer is chosen so that the polarization of the light is rotated 90 degrees if no voltage is applied. When applying the maximum voltage, the original polarization is preserved. Intermediate voltage levels bring about a partial rotation of the polarization. The analyzer lets the rotated polarization completely through and blocks the non-rotated polarization. Colors are represented by means of red, green and blue color filters on the pixels.

The control of the pixels can happen in several ways. Originally, two rows of crossed electrodes were used. Each pixel could then be controlled by using the right combination of the electrodes. Nowadays however, each pixel is provided with its own (transparent) transistors, located near by the pixel, the so-called *thin-film* transistors or TFT.

A liquid crystal cell emits no light on its own and has to be provided with an external light source. The first LCDs used regular daylight and worked with reflections. This made it very difficult to

represent colors properly. Nowadays LCDs are provided with background lighting. This lighting is usually a fluorescence tube or LED that emits white light located on the edge of the screen. The light is brought to the pixels using a light pipe (this is a glass plate in which the light is trapped by total internal reflection). By providing this light pipe with the proper roughness, we can take care that at each pixel the light pipe emits the same amount of light, resulting in a homogeneous background lighting.

A liquid crystal cell can of course also be used in projection, in which a good white lamp and a projection system is used. LCD-projectors become cheaper every day, although they have to compete with projectors based on *digital light processors* (see further).

Liquid Crystals on Silicon (LCoS)

Instead of using a liquid crystal cell in transmission, they can also be used in reflection. One of the electrodes then acts as a mirror. The principle remains the same apart from the fact that the light now has to pass two times through the crystal. The polarizer now also acts as analyzer.

The advantage of this technique is that the steering logic no longer has to be transparent, which enables us to use standard CMOS-circuits. Because of this, very small displays can be made with a large number of pixels. Such a *Liquid Crystal on Silicon* (LCoS) is used in high-performance television sets.

Disadvantages of liquid crystals

However liquid crystals also have a number of disadvantages compared to CRT-screens. First of all, they operate by blocking light selectively. This blocking is not always that selective, which causes a black screen to emit some light. Also the color production (gamut) is not as good as in CRTs. LCDs are therefore not popular in the graphics world.

Liquid crystals do not switch that fast. While CRTs have no problems with refresh rates over 100Hz, the best LCDs can go to 50Hz maximally. They are therefore less suitable to represent quick movements.

18.3.7 MEMS, Digital Light Processors

This last decade, a strong competitor for liquid crystals has shown up in the projection market. The so-called *Digital Light Processor* (DLP) or *Digital Micromirror Device* (DMD) consists of a large number of minuscule mirrors that can be placed very fast in different positions by applying an electric field (a so-called MEMS: *Micro Electromechanical System*). These chips can be made in advanced silicon technology. Each pixel is controlled separately by its own circuit.

A DLP is illuminated with an external light source. Dependent on the position of the mirrors, the light is projected on a screen or it is lost. The mirrors can be switched 'on' and 'off' up to a thousand times per second. The fraction of time that the mirrors project on the screen determines the intensity of the image, as our eye is too slow to see the fast switching.

Projectors based on DLPs have become competitive with liquid crystals the last few years. Advanced realizations are also used in digital cinema, to drive away the use of pellicule.

Although DLPs can switch more quickly than a LCD, they also have got problems with the representation of deep black, as the turned away mirrors still scatter some light.

18.3.8 Projectors

Digital projection has recently become common in the corporate world, the living room as well as the cinema. Liquid crystals and DLPs are still fighting a battle for market domination. The different technologies separate themselves especially in the representation of color images.

Transmission screens based on liquid crystals are the simplest. In these we can choose to use a liquid crystal cell that has individual color filters for the different pixels. This liquid crystal cell can be used in transmission or reflection.

The situation becomes more difficult for monochromatic liquid crystal cells or DLPs. Then we can choose between the use of one single chip for the three colors or the use of a separate chip for red, green and blue. Both techniques are depicted in figure 18.17. In the first case, the colors will be represented sequentially: a red image is first projected, then a green image and finally a blue image. A rotating color filter is used for this. Although we only need one single chip, we have to switch three times faster to represent different images. This technique is therefore mainly used in combination with DLPs.

We can also use a different chip for each color. The white light bundle is then split in three different bundles with a different color using a cube, consisting of different types of glass that show a strong material dispersion. Because of this, total internal reflection occurs for certain wavelengths, while other wavelengths are let through. Each bundle is reflected by its own chip after which the resulting bundles are brought together with a similar component.

18.3.9 Laser projection

An alternative projection method consists of directly projecting laser light. The principle is depicted in figure 18.18. A different laser is used for red, green and blue. The laser light is pointed on to a column of switchable diffractive elements (a so-called *grating light valve* or GLV). Dependent on the state of the element, the light is either reflected or diffracted in a different direction. The diffracted bundles of the different colors are then brought together and projected onto the screen. Each bundle will project an entire column of pixels (there is a GLV for each pixel). The lines are then scanned by a rotating mirror. Just like in a DLP, a GLV switches very fast and the intensity of each pixel is determined by the fraction of time the GLV is in its 'on'-state.

Laser projection can give us very bright images on a large surface. Until recently, the technique had to deal with the lack of an efficient blue semiconductor laser.

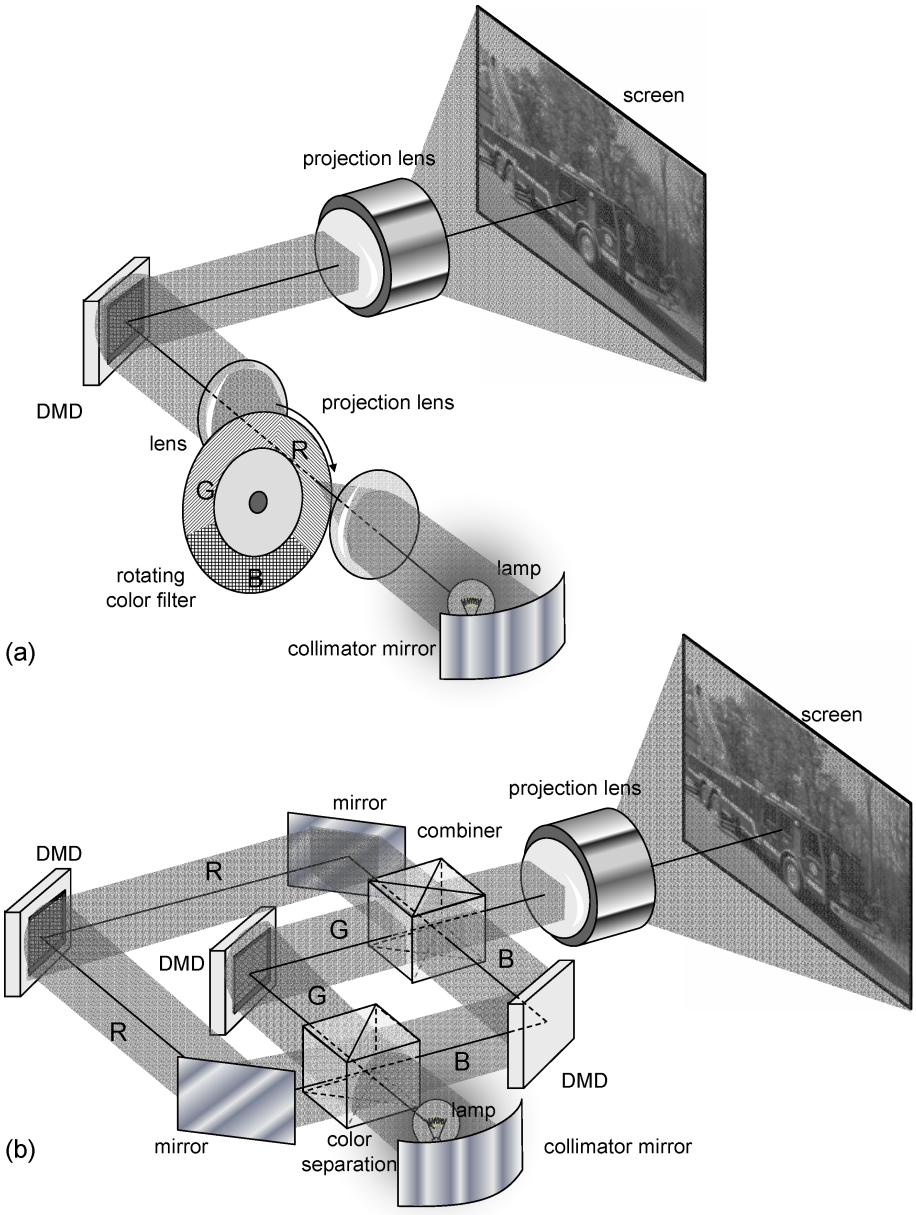

Figure 18.17: Principle of a projection screen. (a) Projection screen based on a single DMD and a rotating color filter. (b) Projector based on 3 different DMDs or (reflective) liquid crystal elements and bundle splitters.

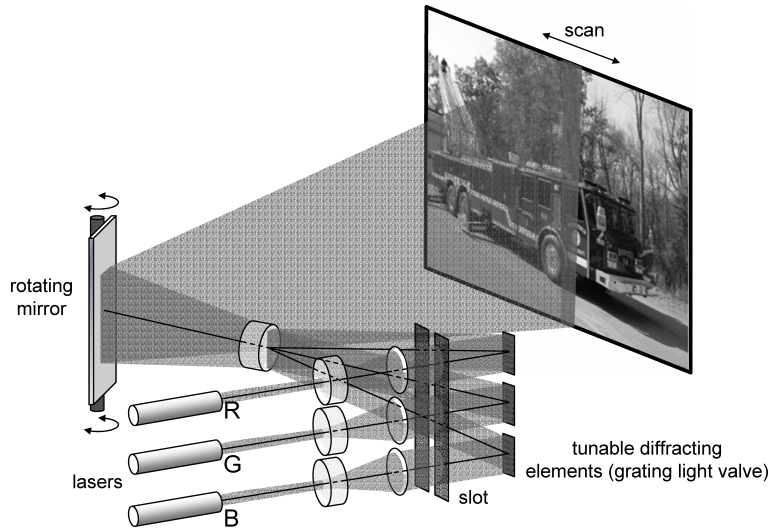

Figure 18.18: Direct laser projection

18.3.10 LED screens

Displays based on light emitting diodes (LEDs) have already been used for a long time in consumer electronics, going from the first electronic calculator to stereo sets. However, most of these screens were monochromatic and limited to textual information.

Since the introduction of the blue LEDs in the late 90's, LEDs can also be used for displays. The most striking application are the very large displays, like the Sony Jumbotron, used in stadia for large events. The red, green and blue LEDs are packed individually and the pixels are thus very large.

Recently, the first LED displays have turned up in small devices such as mobile phones and digital cameras. These displays do not work with the classic LEDs based on semiconductors, but with organic molecules, the so-called organic LEDs or OLEDs. OLED screens are very new and are still dealing with some problems like stability and bleaching (the gradually decreasing color emission as the LED ages).

Displays based on LEDs undoubtedly have a bright future, as LEDs convert electric energy into light very efficiently. Furthermore, LEDs have no need of an external light source like LCDs or DLPs.

18.4 3-D imaging

Until now we discussed displays that render a plane image. When we want to represent three-dimensional images, we have to use artificial tricks to bring parallax in the images.

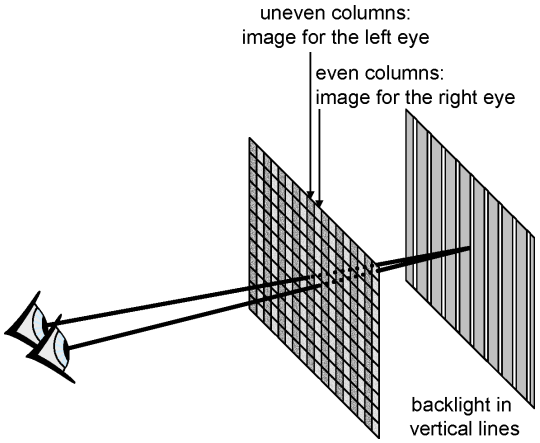

Figure 18.19: LCD-screen with a 3-D view: The background lighting is no longer homogeneous, but consists of a number of thin lines. Therefore the different pixel columns are projected either on the left eye or on the right eye.

18.4.1 3-D glasses

The simplest manner to represent three-dimensional images is giving each eye separate information. In the early stage of cinema, the technique of projecting a red and green image over each other had already been developed. With the proper glasses, each eye would see the right image. This however results in a grey image.

To make it possible to view color images, glasses are used of which the sides transmit perpendicular polarizations. Different images are then projected with a different polarization. The disadvantage of this technique is that it is only applicable in cinemas and not on television screens.

With the introduction of computers and 3-D games, a new possibility was introduced. Instead of projecting both images through each other, they are projected separately. Glasses with a liquid crystal cell then synchronously shields the unwanted image of the proper eye. The final refresh rate is halved however. This makes the use of such glasses tiring for the eyes.

18.4.2 3-D LCD screen

Recently a technique has been introduced that enables us to generate a stereoscopic image without special glasses. Instead of using a homogeneous background lighting, a lighting consisting of vertical lines is used (figure 18.19). Each background line is provided with two columns of pixels. Because of the slightly different angles under which the eye sees the image, the one eye will only see the even lines, while the other will only see the uneven lines.

A disadvantage of this technique is that a good image is obtained only when sitting straight in front of the screen. Furthermore, it works in a limited depth range. On the other hand, the technique is simple to implement, and we can change from 2-D to 3-D by switching the striped background lighting on or off.

18.4.3 Holography

All technologies discussed thus far create an image by modulating the intensity of the emitted light. In order to create full 3-D images, not only the intensities but also the phases of the wave fronts have to be correct.

In holography, the phase front of a coherent illuminated object is saved in a light sensitive material (e.g. a photographic plate) using interference. If the plate is then illuminated again with coherent light, the original phase fronts arise again and the original image becomes visible from all angles.

Holography is discussed in further detail in the course *Microphotonics*.

Bibliography

Part VI

Appendices

Appendix A

Basis van de Halfgeleiderfysica

Contents

In dit hoofdstuk behandelen we enkele basisbegrippen van de halfgeleiderfysica. We stellen eerst de dispersierelatie voor elektronen in een halfgeleiderkristal op. Aan de hand van de bandentheorie stellen we een model op voor elektrische geleiding en elektronbeweging in halfgeleiders. We onderzoeken de bezetting van de elektrontoestanden in het halfgeleiderkristal. Hierbij vatten we een ontbrekend elektron met succes op als een tweede type ladingsdrager, en duiden het aan als een *holte*. We bekijken vervolgens de invloed van onzuiverheden in het halfgeleiderkristal en komen zo tot de begrippen *n-type* en *p-type* halfgeleider. We behandelen ook kort de juncties tussen verschillende types. Tot slot bekijken we de optische eigenschappen van halfgeleiders.

A.1 Bandentheorie

A.1.1 Vrij elektron

Een elektron in een ééndimensionale, tijdsonafhankelijke potentiaal $V(x)$ kan beschreven worden aan de hand van de tijdsonafhankelijke Schrödingervergelijking:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\,\psi = E\,\psi \tag{A.1}$$

Hierin is $\hbar$ de gereduceerde constante van Planck en m de elektronmassa. Wanneer we de potentiële energie verwaarlozen, krijgen we oplossingen voor ψ van de vorm:

$$\psi = A\,e^{jkx}, \tag{A.2}$$

waarbij moet gelden

$$E = \frac{\hbar^2 k^2}{2m}. \tag{A.3}$$

Dit is de parabolische relatie tussen de (kinetische) energie E van het elektron en het golfgetal k, ook wel de *dispersierelatie* genoemd. Voor een gegeven waarde van E ligt k vast. Of omgekeerd: voor een gegeven waarde van k ligt E vast. Men ziet dat de massa m van het elektron een belangrijke rol speelt in deze relatie. Aangezien in de klassieke mechanica geldt dat de kinetische energie gerelateerd is tot het moment p als

$$E = \frac{p^2}{2m} \tag{A.4}$$

volgt hieruit nog de eenvoudige relatie tussen moment p en golfvector k:

$$p = \hbar\, k. \tag{A.5}$$

A.1.2 Elektron in een periodieke potentiaal

Een elektron in een (ééndimensionaal) kristalrooster met roosterconstante a is onderhevig aan een potentiaal $V(x)$ met dezelfde periodiciteit, veroorzaakt door de atomen in het kristal. Er geldt dus: $V(x) = V(x + a)$. *Bloch* heeft aangetoond dat de oplossingen ψ van (A.1) voor een dergelijke potentiaal de volgende gedaante hebben:

$$\psi_k(x) = u_k(x)\, e^{jkx}, \tag{A.6}$$

met $u_k(x) = u_k(x + a)$, een functie met dezelfde periodiciteit als $V(x)$ en gekenmerkt door het label k. De oplossingen (A.6) worden Blochfuncties genoemd en hebben de gedaante van een vlakke golf gemoduleerd met een functie met dezelfde periodiciteit als de potentiaal. De specifieke vorm van u_k wordt bepaald door de vorm van de potentiaal en het label k. Om de expliciete oplossingen u_k analytisch te berekenen, moeten we enkele veronderstellingen maken. Het Kronig-Penneymodel is hier de meest gevolgde aanpak. In dit model worden de functies u_k en de dispersierelatie $E(k)$ uitgerekend voor een periodieke opeenvolging van rechthoekige potentiaalputten met welbepaalde breedte en diepte (zie A.3.1). Het typisch $E(k)$ verloop wordt geschetst in figuur A.1. De fijne lijn geeft het parabolisch verband voor vrije elektronen weer en de volle curven zijn het resultaat van het Kronig-Penney model. Het globale verloop volgt dat van de vrije elektronen maar de afwijkingen nemen toe telkens k een veelvoud van π/a nadert. Voor $k = m\pi/a$, met m geheel, treden discontinuïteiten op met twee waarden voor de energie E, waartussen een gebied ligt met ontbrekende E-waarden.

Gereduceerde k-ruimte

Een interessante eigenschap van Blochfuncties is de volgende. We beschouwen $\psi_k(x)$ overeenkomend met een bepaalde k en eigenwaarde E. Stellen we nu $k^{'} = k + 2m\pi/a$, dan vinden we:

$$\psi_k(x) = u_k(x)\, e^{jkx} = [u_k(x)\, e^{-j2m\pi x/a}]\, e^{jk^{'}x} \tag{A.7}$$

De uitdrukking tussen vierkante haken heeft dezelfde periodiciteit als $u_k(x)$. Dit betekent dat $\psi_k(x)$ ook geschreven kan worden als een Blochtoestand met golfvector $k^{'}$, voor dezelfde waarde van E. Voor een gegeven E is k dus slechts bepaald op een waarde $K = 2m\pi/a$ na. De $E(k)$ curven herhalen zich dus periodiek met periode K. Deze K noemen we de *primitieve vector* van

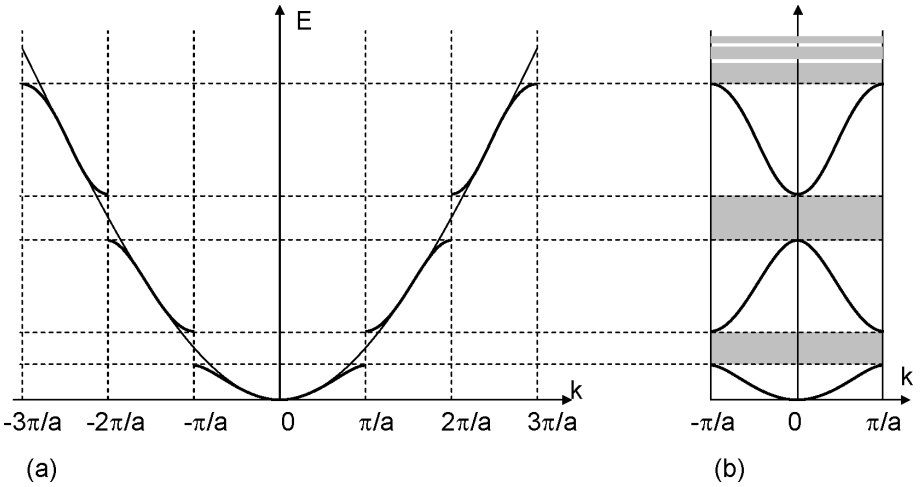

Figure A.1: (a) Dispersierelatie $E(k)$ voor een elektron in een periodieke potentiaal. (b) Gereduceerd bandenschema.

het *reciproke* rooster. Dit laat ons toe om het gehele bandenschema te reduceren tot het gebied tussen $-\pi/a$ en π/a, dat de (eerste) *Brillouinzone* genoemd wordt (zie figuur A.1 (b)). In de Brillouinzone vinden we dus verscheidene continue energiebanden $E_n(k)$ terug, gescheiden door verboden zones.

Meerdere elektronen in een kristal

De oplossing van de Schrödingervergelijking (A.1) voor één elektron in een periodieke potentiaal resulteerde in eigenfuncties

$$\psi_k(x) = u_k(x)\, e^{jkx}, \tag{A.8}$$

met geassocieerde energie $E(k)$. Als we de interactie tussen elektronen verwaarlozen, kunnen we een kristal met meerdere elektronen beschrijven door met elke toegelaten k-waarde één elektron te laten overeenkomen, totdat alle elektronen opgebruikt zijn.

Tot nu toe hebben we enkel oneindig uitgestrekte kristallen beschouwd. Wegens deze oneindige afmetingen verkregen we in elke energieband een continuüm van niveaus. Een reëel kristal bevat uiteraard slechts een beperkt aantal atomen N, en heeft een beperkte lengte $L = Na$. De verbreking van de periodiciteit geeft aanleiding tot ingewikkelde randeffecten die we hier niet zullen behandelen. Om de invloed van de beperkte afmetingen op de volumetoestanden te bespreken dienen we randvoorwaarden in te voeren. Bij een dergelijk probleem is het gebruikelijk om cyclische randvoorwaarden op te leggen, men stelt: $\psi(x) = \psi(x+L)$. Voor een Blochfunctie $\psi_k(x)$ geldt dan:

$$u_k(x+L)\, e^{jk(x+L)} = u_k(x)\, e^{jkx}. \tag{A.9}$$

Aangezien $L = Na$ geldt $u_k(x+L) = u_k(x)$ zodat

$$e^{jkL} = 1. \tag{A.10}$$

De toegelaten k-waarden zijn dan:

$$k = 0,\ \pm\frac{2\pi}{L},\ \pm\frac{4\pi}{L}, \cdots, +\frac{N\pi}{L} = \frac{\pi}{a}. \tag{A.11}$$

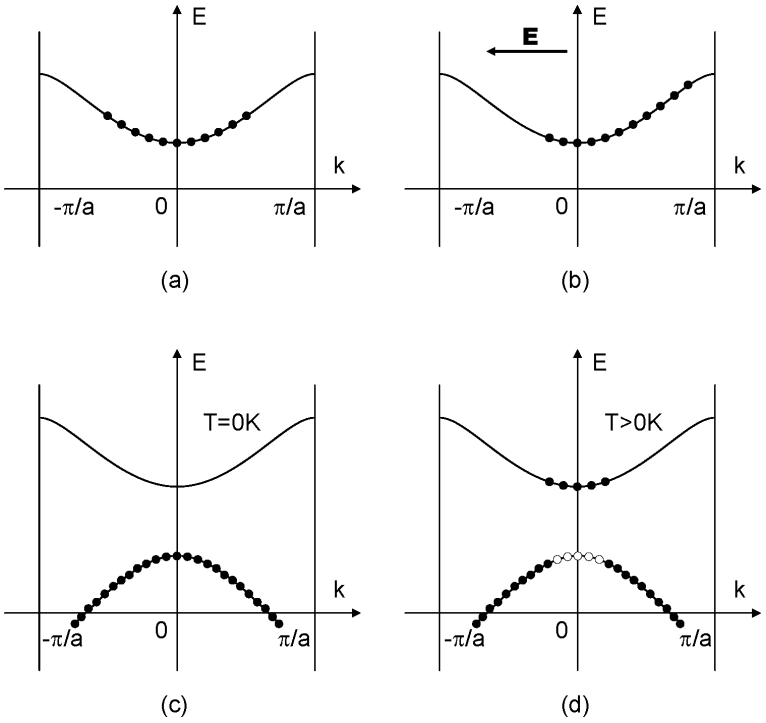

Figure A.2: Mechanisme voor elektrische geleiding door elektronen in een energieband. (a) en (b) Geleiding in metalen. (c) en (d) Geleiding in halfgeleiders.

Hierbij hebben we de rij hebben afgebroken bij $N\pi/L$, de rand van de Brillouinzone. Het aantal toegelaten k-punten in de Brillouinzone bedraagt dus precies N[1]. Houden we nu nog rekening met de twee onafhankelijke spintoestanden per toegelaten k-waarde, dan besluiten we dat er binnen de eerste Brillouinzone $2N$ beschikbare elektrontoestanden zijn per energieband. Het aantal valentie-elektronen per eenheidscel bepaalt dan of we al dan niet met volledig gevulde banden te maken hebben. Algemeen geeft een even aantal elektronen per eenheidscel aanleiding tot volledig gevulde banden. Bij een oneven aantal is de bovenste band half gevuld.

Elektrische geleiding in kristallen

Het mechanisme waarbij elektronen in een energieband bijdragen tot de elektrische geleiding is schematisch als volgt. In evenwicht (elektrisch veld[2] $\mathbf{E} = 0$) bevinden zich in de band evenveel elektronen met positieve als met negatieve k en is de snelheid van elektronen die naar links of naar rechts lopen gelijk verdeeld. Bij aanleggen van een elektrisch veld worden de elektronen versneld, waarbij in de één-elektron benadering k verandert volgens:

$$\frac{dk}{dt} = -\frac{1}{\hbar}\, e\, \mathbf{E}. \tag{A.12}$$

In deze formule wordt impliciet verondersteld dat er geen beperkingen zijn voor de toename van k, m.a.w. dat het elektron zich naar onbezette k-waarden kan verplaatsen. Wanneer er in de band

[1] Het punt $-N\pi/L = -\pi/a$ is in deze rij identiek aan π/a

[2] We noteren $\mathbf{E}$ voor een elektrisch veld en E voor een energie. Hoewel $\mathbf{E}$ algemeen een vector is, behandelen we in dit hoofdstuk voornamelijk ééndimensionale problemen waarin we $\mathbf{E}$ kunnen opvatten als een scalaire grootheid.

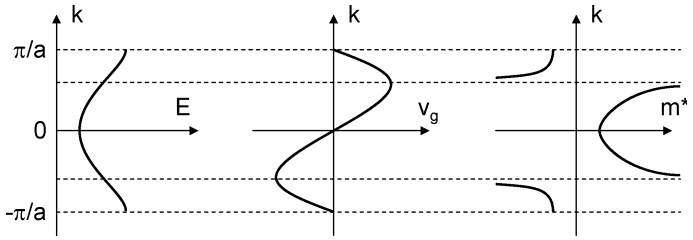

Figure A.3: Groepssnelheid v_g en effectieve massa m^* van een bandelektron.

meerdere elektronen aanwezig zijn zal men dus rekening moeten houden met de bezetting. De situatie na een zekere tijd wordt voor een gedeeltelijk gevulde band weergegeven in figuur A.2. Men ziet dat er meer elektronen naar rechts lopen dan naar links, wat resulteert in elektrische geleiding. In een reëel kristal houdt deze versnelling niet onbeperkt aan, maar wordt verbroken door botsingen (aan defecten of roostertrillingen). Er stelt zich een stationair regime in waaruit begrippen als mobiliteit, geleidingsvermogen en driftsnelheid worden bepaald. Kristallen met een gedeeltelijk gevulde band noemen we *metalen*. De resistiviteit ligt bij normale temperaturen rond $10^{-4} - 10^{-6}\,\Omega$cm.

In kristallen met een even aantal valentie-elektronen per eenheidscel wordt de volledig gevulde band meestal aangeduid als de *valentieband* en de volgende hoger gelegen band als de *conductieband*, gescheiden door de *bandkloof* met breedte E_g. De situatie waarbij de valentieband volledig gevuld is en de conductieband volledig leeg, is enkel strikt geldig bij de nulpuntstemperatuur $T = 0$. In deze situatie moet elke verandering in k gepaard gaan met een verandering in tegengestelde zin (uitwisseling van k-toestanden). Het netto-effect op de snelheid is dan nul en geleiding is niet mogelijk.

Bij hogere temperaturen zijn de elektronen verdeeld volgens de Fermi-Dirac distributie (zie A.2.1) waardoor er ook enkele elektronen in de conductieband gevonden kunnen worden. Elk elektron in de conductieband laat een holte achter in de valentieband. Bij aanleg van een elektrisch veld is er dus elektrische geleiding mogelijk, aangezien de energiebanden niet volledig gevuld of volledig leeg zijn. Bij kristallen met een voldoend kleine E_g is er een behoorlijke excitatie van elektronen over de bandkloof mogelijk. Deze kristallen worden *halfgeleiders* genoemd. Het geleidingsvermogen is afhankelijk van de densiteit van elektronen in de conductieband en van holten in de valentieband. In het vervolg van het hoofdstuk duiden we deze grootheden aan met n respectievelijk p, beide uitgedrukt in cm^{-3}. Zoals we verder zullen zien, kunnen n en p sterk beïnvloed worden door onzuiverheden toe te voegen aan de halfgeleider. Enkele beter bekende en veel gebruikte halfgeleiders zijn Si ($E_g = 1.1\,$eV) en GaAs ($E_g = 1.45\,$eV). De resistiviteit ligt in het gebied $10^{-2} - 10^{9}\,\Omega$cm. Voor kristallen met hogere E_g-waarden is de excitatie veel moeilijker en het geleidingsvermogen veel lager. Deze kristallen noemen we *isolatoren*. De resistiviteit is van de orde $10^{14} - 10^{22}\,\Omega$cm

Elektronbeweging

Om een bewegend elektron te lokaliseren als een deeltje, worden Blochgolven met licht verschillende k samengesteld tot golfpakketten. De snelheid waarmee het golfpakket zich voortplant

wordt gegeven door de groepssnelheid:

$$v_g = \frac{d\omega}{dk} = \frac{1}{\hbar}\frac{dE}{dk} \tag{A.13}$$

Voor een vrij elektron vinden we zoals verwacht $v_g = \hbar k/m$. Voor een elektron in een kristal wordt het verloop $v_g(k)$ weergegeven in figuur A.3. Uit de periodiciteit van $E(k)$ en uit symmetrieoverwegingen volgt dat bij bandextrema $v_g = 0$.

We onderzoeken nu de bewegingsvergelijking van een vrij elektron onder invloed van een uitwendige kracht, bv. bij aanleggen van een elektrisch veld $\mathbf{E}$. De energiewinst in een tijdsinterval δt bedraagt

$$\delta E = -e\,\mathbf{E}\,v_g\delta t = \frac{-e\,\mathbf{E}}{\hbar}\,\frac{dE}{dk}\delta t = \frac{dE}{dk}\,\delta k, \tag{A.14}$$

waaruit

$$\delta k = \frac{-e\,\mathbf{E}}{\hbar}\,\delta t \tag{A.15}$$

$$\hbar\,\frac{dk}{dt} = -e\,\mathbf{E}. \tag{A.16}$$

Algemener kunnen we dit schrijven als:

$$\hbar\,\frac{dk}{dt} = F, \tag{A.17}$$

met F de uitwendige kracht. Voor een vrij elektron is dit niets meer dan de tweede wet van Newton:

$$\frac{d(mv)}{dt} = F. \tag{A.18}$$

Voor het geval van een kristalelektron gaan we als volgt te werk:

$$\frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d^2E}{dk\,dt} = \frac{1}{\hbar}\frac{d^2E}{dk^2}\frac{dk}{dt} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2}F \tag{A.19}$$

$$\left[\frac{\hbar^2}{\frac{d^2E}{dk^2}}\right]\frac{dv_g}{dt} = F \tag{A.20}$$

Als we de uitdrukking tussen de rechte haken identificeren met een massa dan vinden we het verband (A.18) terug. We noemen dit de *effectieve massa* m^*, gedefinieerd door:

$$\frac{1}{m^*} = \frac{1}{\hbar^2}\frac{d^2E}{dk^2} \tag{A.21}$$

Door gebruik te maken van de effectieve massa kunnen we de beweging van een elektron in een kristal beschrijven aan de hand van de bekende wetten van de fysica. Dit betekent niet dat in een kristal de massa van een elektron zou veranderen of dat de tweede wet van Newton geweld wordt aangedaan. Als men alle krachten die op het elektron inwerken in rekening brengt, nl. zowel krachten afkomstig van het kristal als uitwendige krachten, dan blijft deze wet onveranderd gelden. De effectieve massa is k-afhankelijk. Deze afhankelijkheid is weergegeven voor een eenvoudige energieband in figuur A.3. Bij buigpunten van $E(k)$ bereikt m^* oneindig grote waarden. Er bestaan ook gebieden met met negatieve m^*. Voor halfgeleiders zijn enkel de toestanden

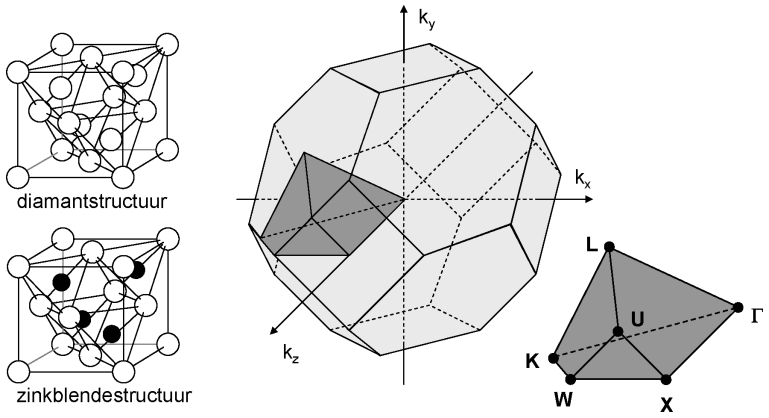

Figure A.4: Kristalstructuur en Brillouinzone van Si, Ge en GaAs.

in de buurt van de bandextrema van belang. Daar kan men de dispersierelatie benaderen door een parabolisch verloop van de vorm:

$$E = E_0 + A(k - k_0)^2, \tag{A.22}$$

wat overeenkomt met een constante waarde van m^* over een beperkt bereik.

Bandstructuur van enkele belangrijke halfgeleiders

In werkelijkheid is de bandstructuur van een halfgeleider vaak ingewikkelder dan tot hier toe werd geschetst. Een reële halfgeleider heeft vanzelfsprekend een driedimensionaal kristalrooster en dus ook een driedimensionale Brillouinzone. Bekijken we het geval van Ge, Si en GaAs. Deze halfgeleiders bezitten de diamantstructuur of zinkblendestructuur, waarvan de kubische eenheidscel geschetst is in figuur A.4. De overeenkomstige Brillouinzone is weergegeven in dezelfde figuur. De dispersierelatie langsheen de assen ΓX en ΓL zijn geschetst in figuur A.5. Een belangrijke vaststelling is dat voor zowel Si als Ge het minimum van de conductieband en het maximum van de valentieband bij verschillende k-waarde zijn gelegen. Men zegt dat ze een *indirecte* verboden zone hebben. GaAs heeft een *directe* verboden zone, het minimum van de conductieband en het maximum van de valentieband liggen bij dezelfde k-waarde. We komen hier nog op terug in 14.1.1 en 14.1.2.

A.2 Elektronen en holten in halfgeleiders

In A.1 is duidelijk gebleken dat de densiteit van ladingsdragers n en p een belangrijke rol speelt in de geleidingseigenschappen van een halfgeleider. In deze paragraaf zullen we beide grootheden berekenen rekening houdend met de thermische excitatie en met onzuiverheden (dotering).

De bezetting van het systeem wordt algemeen bepaald door het product van de dichtheid van beschikbare niveaus en de waarschijnlijkheid van bezetting: $D(E)P(E)$. Bij de berekening van n en p in een halfgeleider kan men zich beperken tot de bovenste niveaus van de valentieband en de onderste niveaus van de conductieband, en de niveaus in de verboden zone die afkomstig zijn van onzuiverheden.

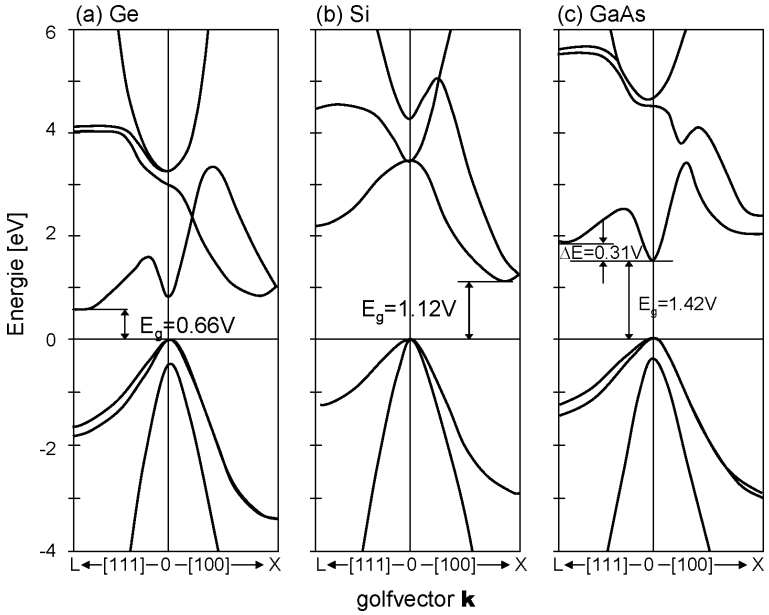

Figure A.5: Bandstructuur van enkele belangrijke halfgeleiders.

A.2.1 Bezettingswaarschijnlijkheid

De waarschijnlijkheid dat een elektrontoestand met energie E bezet wordt door een elektron wordt gegeven door de *Fermi-Dirac distributie*:

$$P_e(E) = f(E) = \frac{1}{\exp[(E - E_f)/k_B T] + 1} \tag{A.23}$$

Hierin is E_f de energie van het *ferminiveau*. Voor $T = 0$ zijn alle niveaus met $E < E_f$ bezet en alle niveaus met $E > E_f$ leeg. Voor $T > 0$ varieert $f(E)$ geleidelijk, met een bezettingswaarschijnlijkheid van $1/2$ bij E_f. De waarschijnlijkheid dat een niveau *niet* bezet is, wordt gegeven door:

$$P_h(E) = 1 - P_e(E) = \frac{1}{\exp[(E_f - E)/k_B T] + 1} \tag{A.24}$$

$P_h(E)$ geeft de waarschijnlijkheid voor bezetting met een holte. De ontbrekende elektronen aan de top van de valentieband kunnen inderdaad beschouwd worden als positief geladen deeltjes en worden holten[3] genoemd, waarbij men volgende relaties heeft tussen de eigenschappen van het ontbrekende elektron en de holte:

$$\begin{array}{rc} \text{k - vector:} & \mathbf{k}_h = -\mathbf{k}_e \\ \text{lading:} & q_h = -q_e > 0 \\ \text{energie:} & E_h = -E_e \\ \text{massa:} & m_h^* = -m_e^* \end{array} \tag{A.25}$$

$P_e(E)$ en $P_h(E)$ worden grafisch voorgesteld in figuur A.6.

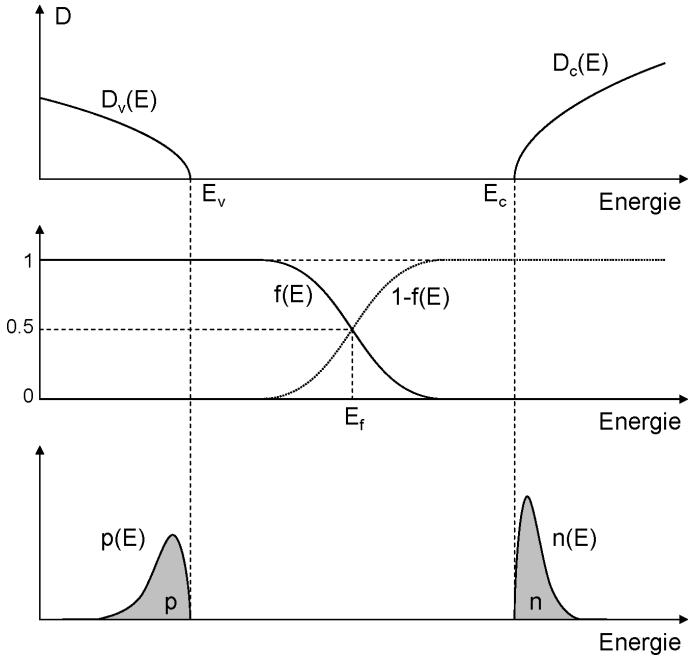

Figure A.6: Toestandsdichtheid en bezettingswaarschijnlijkheid bij intrinsieke halfgeleiders.

A.2.2 Toestandsdichtheid

De toestandsdichtheid $D(E)$ wordt gedefinieerd als het aantal toestanden per volume-eenheid dat beschikbaar is in het energie-interval $E + dE$. Gebruik makende van de parabolische benadering van de banden nabij het minimum van de conductieband en het maximum van de valentieband kunnen we $D(E)$ afleiden. Voor de conductieband vinden we:

$$D_c(E)dE = 8\pi\sqrt{2}(m_e^*)^{3/2}h^{-3}(E - E_c)^{1/2}dE \tag{A.26}$$

En analoog voor de valentieband:

$$D_v(E)dE = 8\pi\sqrt{2}(m_h^*)^{3/2}h^{-3}(E_v - E)^{1/2}dE \tag{A.27}$$

Beide toestandsdichtheden zijn geschetst in figuur A.6.

A.2.3 Intrinsieke halfgeleiders

De dichtheid van elektronen in de conductieband wordt gegeven door:

$$n(E)dE = D_c(E)P_e(E)dE = D_c(E)f(E)dE \tag{A.28}$$

In figuur A.6 wordt deze functie geschetst. Hier wordt ook duidelijk waarom we enkel toestanden nabij E_c moeten beschouwen. Onder de aanname dat het ferminiveau ongeveer halverwege de

[3]Holten worden ook wel gaten of holes (Engels) genoemd.

verboden zone ligt, neemt $f(E)$ immers snel af voor $E > E_c$. De totale densiteit van elektronen in de conductieband (n) wordt dan gegeven door:

$$n = \int_{CB} n(E)dE = \int_{E_c}^{\infty} D_c(E)f(E)dE. \tag{A.29}$$

Na uitwerking volgt:

$$n = N_c \exp\left[-\frac{E_c - E_f}{k_B T}\right] \tag{A.30}$$

waarin

$$N_c = 2\left[\frac{2\pi m_e^* k_B T}{h^2}\right]^{3/2} \tag{A.31}$$

Voor de dichtheid van holten in de valentieband geldt er:

$$p(E)dE = D_v(E)P_h(E)dE = D_v(E)[1 - f(E)]dE \tag{A.32}$$

Op analoge wijze verkrijgen we:

$$p = N_v \exp\left[-\frac{E_f - E_v}{k_B T}\right] \tag{A.33}$$

waarin

$$N_v = 2\left[\frac{2\pi m_h^* k_B T}{h^2}\right]^{3/2} \tag{A.34}$$

Een *intrinsieke* halfgeleider heeft geen onzuiverheidsniveaus in de verboden zone. Elk elektron in de conductieband laat dus een holte achter in de valentieband waaruit volgt dat $n = p$. Deze voorwaarde laat ons toe om de ligging van het ferminiveau E_f te berekenen:

$$E_f = \frac{1}{2}(E_v + E_c) + \frac{3}{4}k_B T \ln\left(\frac{m_h^*}{m_e^*}\right) \tag{A.35}$$

Als $m_e^* = m_h^*$ ligt E_f precies halfweg in de verboden zone, wat in overeenstemming is met de eerder gemaakte veronderstelling. Voor de intrinsieke densiteit van ladingsdragers $n_i = n = p$ vinden we:

$$n_i = \sqrt{np} = \sqrt{N_c N_v} \exp\left[-\frac{E_g}{2k_B T}\right] \tag{A.36}$$

We merken op dat n_i niet afhangt van de ligging van het ferminiveau, maar voor een bepaalde halfgeleider enkel van E_g en T. De situatie voor intrinsieke halfgeleiders is geschetst in figuur A.6.

VOORBEELD: Voor GaAs bij $T = 300\,\mathrm{K}$ en met $m_e^* = 0.07\,\mathrm{m}$, $m_h^* = 0.6\,\mathrm{m}$, $E_g = 1.43\,\mathrm{eV}$ krijgen we $n_i = 2.246 \times 10^6\,\mathrm{cm}^{-3}$.

A.2.4 Dotering

Men kan aan halfgeleiders zeer kleine concentraties (ppm-niveau) van onzuiverheidsatomen toevoegen (dotering). Deze introduceren nieuwe energieniveaus die tussen conductie- en valentieband kunnen liggen. *Donoren* zijn onzuiverheden waarvan de elektronen een energietoestand

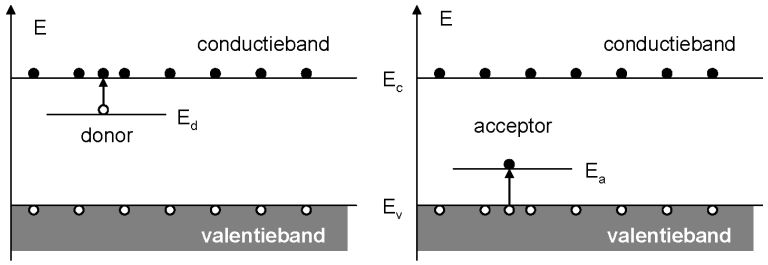

Figure A.7: Donor- en acceptorniveaus in gedoteerde halfgeleiders.

hebben vlak onder de bodem van de conductieband, zoals weergegeven in figuur A.7. Deze elektronen worden gemakkelijk thermisch geëxciteerd naar de conductieband, waarbij positieve ionen achterblijven. Voor *acceptoren* geldt een volledig duale situatie. Zij hebben onbezette elektronniveaus vlak boven de valentieband. Elektronen kunnen er dus gemakkelijk naar toe geëxciteerd worden, waarbij holten ontstaan in de valentieband.

Bekijken we eerst het geval van een halfgeleider gedoteerd met N_d donoren per cm^3. Het donorniveau is gegeven door E_d. Indien $E_c - E_f \gg k_B T$ geldt er nog steeds[4]:

$$n = N_c \exp\left[-\frac{E_c - E_f}{k_B T}\right] \tag{A.37}$$

Hierin wordt E_f echter niet langer gegeven door (A.35) maar wordt ook bepaald door de donoren. Het aantal elektronen in de conductieband is nu afkomstig van zowel het intrinsieke proces en als van de ionisatie van donoren, waarbij we veronderstellen dat alle donoren daadwerkelijk geïoniseerd worden. Bij al deze processen moet het aantal gecreëerde negatieve ladingen gelijk zijn aan het aantal positieve, of nog:

$$n = N_d + p \tag{A.38}$$

Deze relatie bepaalt ondubbelzinnig de ligging van het ferminiveau E_f, voor elke waarde van T en N_d:

$$N_c \exp\left[-\frac{E_c - E_f}{k_B T}\right] = N_d + N_v \exp\left[-\frac{E_f - E_v}{k_B T}\right] \tag{A.39}$$

Veronderstellen we dat $N_d \gg n_i$, dan krijgen we, gebruik makend van $np = n_i^2$ en (A.38):

$$n \simeq N_d \tag{A.40}$$

$$n \gg p \tag{A.41}$$

$$E_f \simeq E_c - k_B T \ln[\frac{N_c}{N_d}] \tag{A.42}$$

We hebben te maken met een extrinsieke, *n-type* halfgeleider. Het aantal elektronen in de conductieband overtreft ruimschoots het aantal holten in de valentieband. We noemen de elektronen dan ook de *majoritaire* ladingsdragers en de holten de *minoritaire* ladingsdragers. De dichtheid van conductie-elektronen is vrijwel gelijk aan de donordichtheid N_d. Het ferminiveau E_f ligt

[4]Aangezien we in de afleiding voor n en p in intrinsieke halfgeleiders enkel de veronderstelling $E - E_f \gg k_B T$ hebben gemaakt, zijn de verkregen resultaten voor n en p ook geldig voor een extrinsieke (gedoteerde) halfgeleider met $n \neq p$, evenals het resultaat $n_i^2 = np$. De ligging van het ferminiveau wordt uiteraard wel beïnvloed door de aanwezigheid van onzuiverheidsniveaus.

dichter bij E_c dan bij E_v. Typische getalwaarden voor N_d zijn 10^{15} cm^{-3} voor een zwak gedoteerde halfgeleider en 10^{19} cm^{-3} voor een sterk gedoteerde halfgeleider. Merk op dat deze getalwaarden meerdere grootteorden hoger liggen dan de intrinsieke densiteit n_i.

De duale situatie voor acceptoren leidt tot de volgende resultaten voor een extrinsieke, *p-type* halfgeleider:

$$p \simeq N_a \tag{A.43}$$

$$p \gg n \tag{A.44}$$

$$E_f \simeq E_v + k_B T \ln[\frac{N_v}{N_a}] \tag{A.45}$$

Hier zijn de elektronen minoritair en de holten majoritair, en ligt E_f dichter bij E_v dan bij E_c.

A.2.5 Geleidbaarheid

Zoals we in A.1 hebben gezien, kunnen we de beweging van een kristalelektron onder invloed van een elektrisch veld $\mathbf{E}$ beschrijven door:

$$m_e^* \frac{dv_g}{dt} = -e\mathbf{E} \tag{A.46}$$

In een geïdealiseerd perfect en stijf kristalrooster wordt de beweging van ladingsdragers niet verstoord en zal de elektrische geleidbaarheid oneindig groot worden. In een reëel kristal wordt de beweging van elektronen echter continue onderbroken door botsingen aan roosterimperfecties zoals onzuiverheidsatomen en roostertrillingen. Deze botsingen zijn vrijwel elastisch, zodat de energie van het elektron (of holte) vrijwel constant blijft. De bewegingsrichting van het elektron kan wel drastisch veranderen. De weg die de ladingsdragers gemiddeld afleggen tussen twee botsingen, noemt men de gemiddelde vrije-weglengte l.

$$l = v_T \, \tau, \tag{A.47}$$

met τ de gemiddelde tijd tussen twee botsingen en v_T de thermische snelheid[5]:

$$v_T = \sqrt{\frac{2k_B T}{m_e^*}} \tag{A.48}$$

Bij het aanleggen van een elektrisch veld $\mathbf{E}$ zullen de elektronen tussen twee botsingen versneld worden volgens (A.46), en een gemiddelde verplaatsing ondergaan tegengesteld aan $\mathbf{E}$. De resulterende *driftsnelheid* v_e is dan

$$v_e = -\frac{e\mathbf{E}}{m_e^*}\tau_n = -\mu_n \mathbf{E} \tag{A.49}$$

$$\mu_n = \frac{e\tau_n}{m_e^*} \tag{A.50}$$

[5]Aangezien de effecten van het periodiek kristalveld begrepen zitten in de effectieve massa, kunnen we de energie van het gas van ladingsdragers beschrijven als louter kinetisch energie $\frac{1}{2}m^* v^2$, waarop de Maxwell-Boltzmann statistiek van toepassing is. Hieruit kunnen we de gemiddelde, thermische snelheid van de ladingsdragers berekenen.

De constante μ_n noemen we de *driftmobiliteit* voor elektronen. Analoog krijgen we voor holten:

$$v_h = -\mu_p \mathbf{E} \tag{A.51}$$

$$\mu_p = \frac{e\tau_p}{m_h^*}. \tag{A.52}$$

De totale stroom door een eenheidsoppervlak is dan

$$J = -n\,e\,v_e + p\,e\,v_h = (n\,e\,\mu_n + p\,e\,\mu_p)\mathbf{E} \tag{A.53}$$

met n en p respectievelijk de elektron- en holtendichtheid. Gebruik makende van de relatie $J = \sigma\mathbf{E}$ krijgen we dan voor de geleidbaarheid σ:

$$\sigma = n\,e\,\mu_n + p\,e\,\mu_p = \frac{ne^2\tau_n}{m_e^*} + \frac{pe^2\tau_p}{m_h^*} \tag{A.54}$$

Voor hoge **E**-velden satureert de driftsnelheid wat resulteert in een afwijking van het Ohms gedrag.

A.2.6 Diffusie en recombinatie

In de vorige paragrafen hebben we de concentratie van ladingsdragers berekend voor een homogene halfgeleider in thermisch evenwicht ($np = n_i^2$). In een halfgeleidercomponent treden er echter afwijkingen op ten opzichte van deze ideale situatie. We onderscheiden twee verschillende gevallen:

- niet-homogene verdeling van n en p, met *diffusie* als gevolg
- afwijking van thermisch evenwicht: $np \neq n_i^2$, wat *recombinatie* veroorzaakt

Diffusie

In een systeem van beweegbare deeltjes treden diffusiestromen op wanneer er een niet-uniforme concentratieverdeling bestaat. Dit geldt eveneens voor de vrije elektronen en holten in een halfgeleider. Kijken we bv. naar de balans van ladingsdragers die door een zeker vlak op positie x_0 en met oppervlakte A stromen. Het aantal deeltjes dat van links naar rechts loopt gedurende een interbotsingstijd τ is

$$N_{L\to R} = \frac{lA}{2} n(x_0 - \frac{l}{2}) \tag{A.55}$$

terwijl het aantal dat van rechts naar links loopt gelijk is aan

$$N_{R\to L} = \frac{lA}{2} n(x_0 + \frac{l}{2}) \tag{A.56}$$

De factor $1/2$ komt van het feit dat gemiddeld gezien slechts de helft van de deeltjes naar het vlak toelopen. Het netto-aantal deeltjes dat dan van links naar rechts loopt is dan:

$$N_{L\to R} - N_{R\to L} = [n(x_0 - \frac{l}{2}) - n(x_0 + \frac{l}{2})]\frac{lA}{2} \simeq -\frac{1}{2}\frac{\partial n}{\partial x} l^2 A \tag{A.57}$$

Als j de deeltjesflux per eenheidsoppervlak is, kunnen we schrijven:

$$j = -\frac{l^2}{2\tau}\frac{\partial n}{\partial x} = -D\frac{\partial n}{\partial x} \tag{A.58}$$

waarin D de *diffusiecoëfficiënt* is, die gegeven wordt door

$$D = \frac{l^2}{2\tau} \tag{A.59}$$

We zullen de diffusiecoëfficiënt voor elektronen aanduiden als D_n en die voor holten als D_p.

Aangezien het bij diffusie opnieuw draait om netto-verplaatsing van botsende deeltjes, kunnen we een verband verwachten tussen de driftmobiliteit en de diffusiecoëfficiënt. Inderdaad, gebruik makende van (A.47) en (A.48) komen we tot de relatie

$$D = \frac{k_B T}{e}\mu \tag{A.60}$$

die ook wel de *Einsteinbetrekking* genoemd wordt.

In aanwezigheid van zowel een elektrisch veld als een concentratiegradiënt, zal de stroomdichtheid dus bestaan uit een drift- en een diffusiecomponent. Voor lage velden geldt dan:

$$\mathbf{J}_n = n\,e\,\mu_n\,\mathbf{E} + e\,D_n\,\nabla n \tag{A.61}$$

$$\mathbf{J}_p = p\,e\,\mu_p\,\mathbf{E} - e\,D_p\,\nabla p \tag{A.62}$$

$$\mathbf{J} = \mathbf{J}_n + \mathbf{J}_p = e(n\,\mu_n + p\,e\,\mu_p)\mathbf{E} + e(D_n\,\nabla n - D_p\,\nabla p) \tag{A.63}$$

Recombinatie

Wanneer een externe oorzaak extra elektron-holte paren injecteert in een halfgeleider is $np > n_i^2$. Deze situatie treedt bv. op bij optische excitatie G_L met licht waarvan de fotonenergie groter is dan de bandkloof E_g. In deze situatie worden ladingsdragers dus gecreëerd door zowel optische als thermische excitatie. Recombinatie van ladingsdragers treedt op wanneer een elektron een holte "tegenkomt". Samengevat hebben we dus drie bijdragen: optische generatie aan een tempo G_L, thermische generatie G_{th} en recombinatie R. We kunnen ook het netto-recombinatietempo definiëren als $U = R - G_{th}$. In thermisch evenwicht (zonder belichting) is $G_L = 0$, $U = 0$ en $R = G_{th}$. Algemeen kan U geschreven worden als

$$U = B(np - n_i^2) \tag{A.64}$$

met B een evenredigheidsconstante. Aangezien $n = n_0 + \Delta n$ en $p = p_0 + \Delta p$ wordt dit

$$U = B(n_0\Delta p + p_0\Delta n) \tag{A.65}$$

Bekijken we het geval van een n-type halfgeleider, waarvoor $n_n = n_{n0} + \Delta n$ en $p_n = p_{n0} + \Delta p$. Aangezien bij kleine injectie $p_{n0} \ll n_{n0} = N_d$ en $\Delta n = \Delta p$ krijgen we

$$U \approx BN_d\Delta p = \frac{p_n - p_{n_0}}{\tau_p} \tag{A.66}$$

Hierin is $\tau_p = 1/BN_d$ de *levensduur* van de minoritairen. U is dus evenredig met het overschot aan minoritairen.

De nettoverandering van de holtedichtheid p_n is dan

$$\frac{dp_n}{dt} = G_L + G_{th} - R = G_L - U, \tag{A.67}$$

We vinden dus de volgende tempovergelijking voor holten in n-type, bij kleine injectie:

$$\frac{dp_n}{dt} = G_L - \frac{p_n - p_{n_0}}{\tau_p}. \tag{A.68}$$

Voor een p-type halfgeleider kunnen we volledig analoge resultaten afleiden. We vinden zo ook een netto-recombinatietempo

$$U = \frac{n_p - n_{p_0}}{\tau_n} \tag{A.69}$$

met $\tau_n = 1/BN_a$. De grootheden τ_p en τ_n hebben de betekenis van gemiddelde levensduur. Ze zijn sterk afhankelijk van de dotering: hoe hoger de dotering, des te sneller de recombinatie van de minoritaire ladingsdragers. De recombinaties kunnen stralend of niet-stralend zijn. De totale recombinatiesnelheid is de som van de recombinatiesnelheden van de individuele processen:

$$R = R_r + R_{nr} \tag{A.70}$$

met R_r de stralende recombinatiesnelheid en R_{nr} de niet-stralende.

Continuïteitsvergelijkingen

We kunnen nu de continuïteitsvergelijkingen opstellen voor de elektronendichtheid $n(x,t)$ of de holtendichtheid $p(x,t)$, waarbij we rekening houden met de creatie en vernietiging van ladingsdragers als gevolg van generatie- en recombinatieprocessen, evenals met drift en diffusie van ladingsdragers. De toename van het aantal deeltjes binnen een volume dV per tijdseenheid is immers gelijk aan de netto-instroom vermeerdert met het nettotempo waarmee deeltjes worden gegenereerd binnen dV.

Voor holten in een n-type halfgeleider vinden we

$$\frac{\partial p}{\partial t} = G_p - \frac{p - p_{n_0}}{\tau_p} - \frac{1}{e}\nabla \cdot \mathbf{J}_p \tag{A.71}$$

Gebruik makende van (A.61) wordt dit in één dimensie

$$\frac{\partial p}{\partial t} = G_p - \frac{p - p_{n_0}}{\tau_p} - p\,\mu_p \frac{\partial \mathbf{E}}{\partial x} - \mu_p\,\mathbf{E}\frac{\partial p}{\partial x} + D_p \frac{\partial^2 p}{\partial x^2} \tag{A.72}$$

Voor elektronen in een p-type halfgeleider vinden we

$$\frac{\partial n}{\partial t} = G_n - \frac{n - n_{p_0}}{\tau_n} - n\,\mu_n \frac{\partial \mathbf{E}}{\partial x} - \mu_n\,\mathbf{E}\frac{\partial n}{\partial x} + D_n \frac{\partial^2 n}{\partial x^2} \tag{A.73}$$

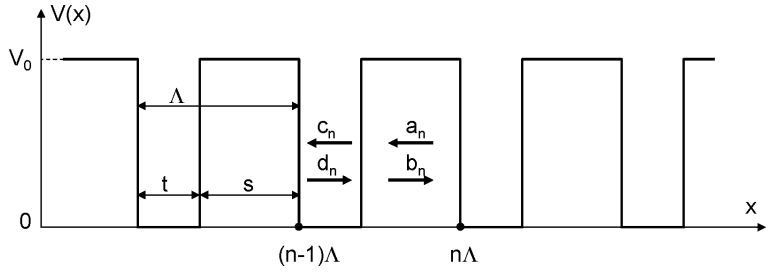

Figure A.8: Periodieke potentiaal in het Kronig-Penneymodel.

A.3 Berekeningen

A.3.1 Het Kronig-Penneymodel

In het *Kronig-Penney*model wordt de periodieke potentiaal afkomstig van het kristalrooster benaderd door een periodieke opeenvolging van rechthoekige potentiaalputten met breedte t en diepte V_0, gescheiden door potentiaalbarrières met breedte s (zie figuur A.8). De kristalperiode is Λ. We veronderstellen dat $V_0 > E$.

We zijn op zoek naar Blochoplossingen $\psi_k(x) = u_k(x)e^{jkx}$ voor de Schrödingervergelijking

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = (E - V(x))\,\psi \tag{A.74}$$

We definiëren

$$\begin{array}{rl} k_1^2 = & 2m(V_0 - E)/\hbar^2 \\ k_2^2 = & 2mE/\hbar^2 \end{array} \tag{A.75}$$

Voor de potentiaalbarrières waar $V(x) = V_0$ vinden we

$$\psi(x) = a_n e^{k_1(x-n\Lambda)} + b_n e^{-k_1(x-n\Lambda)} \tag{A.76}$$

Voor de potentiaalputten waar $V(x) = 0$ krijgen we

$$\psi(x) = c_n e^{jk_2(x-n\Lambda)} + d_n e^{-jk_2(x-n\Lambda)} \tag{A.77}$$

Om de coëfficiënten a_n, b_n, c_n en d_n te vinden dienen we randvoorwaarden in te voeren. De voorwaarde dat $\psi(x)$ en $d\psi(x)/dx$ continu zijn op $x = n\Lambda$ levert op:

$$a_n + b_n = c_{n+1}e^{-jk_2\Lambda} + d_{n+1}e^{jk_2\Lambda} \tag{A.78}$$

$$k_1 a_n - k_1 b_n = jk_2 c_{n+1}e^{-jk_2\Lambda} - jk_2 d_{n+1}e^{jk_2\Lambda} \tag{A.79}$$

Dezelfde randvoorwaarden gelden op $x = n\Lambda + t$:

$$c_{n+1}e^{-jk_2 s} + d_{n+1}e^{jk_2 s} = a_{n+1}e^{-k_1 s} + b_{n+1}e^{k_1 s} \tag{A.80}$$

$$jk_2 c_{n+1}e^{-jk_2 s} - jk_2 d_{n+1}e^{jk_2 s} = k_1 a_{n+1}e^{-k_1 s} - k_1 b_{n+1}e^{k_1 s} \tag{A.81}$$

Uit deze vier vergelijkingen kunnen we het verband tussen (a_n, b_n) en (a_{n+1}, b_{n+1}) halen:

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} \tag{A.82}$$

Hierin is A, B, C en D gegeven door

$$\begin{aligned} A = &\quad e^{-k_1 s}\left[\cos k_2 t + \tfrac{1}{2}\left(\tfrac{k_2}{k_1} - \tfrac{k_1}{k_2}\right)\sin k_2 t\right] \\ B = &\quad e^{k_1 s}\left[\tfrac{1}{2}\left(\tfrac{k_2}{k_1} + \tfrac{k_1}{k_2}\right)\sin k_2 t\right] \\ C = &\quad e^{-k_1 s}\left[-\tfrac{1}{2}\left(\tfrac{k_2}{k_1} + \tfrac{k_1}{k_2}\right)\sin k_2 t\right] \\ D = &\quad e^{k_1 s}\left[\cos k_2 t - \tfrac{1}{2}\left(\tfrac{k_2}{k_1} - \tfrac{k_1}{k_2}\right)\sin k_2 t\right] \end{aligned} \tag{A.83}$$

Met behulp van deze relatie kunnen we (a_n, b_n) in elke eenheidscel berekenen van zodra hun waarde in een bepaalde cel bekend is. We kennen dan ook (c_n, d_n).

Bekijken we de oplossing $\psi(x)$ in de n-de eenheidscel ter hoogte van de potentiaalbarrière:

$$\psi(x) = a_n e^{k_1(x-n\Lambda)} + b_n e^{-k_1(x-n\Lambda)} \tag{A.84}$$

zodat

$$\begin{aligned} \psi(x+\Lambda) &= a_{n+1} e^{k_1(x+\Lambda-(n+1)\Lambda)} + b_{n+1} e^{-k_1(x+\Lambda-(n+1)\Lambda)} \\ &= a_{n+1} e^{k_1(x-n\Lambda)} + b_{n+1} e^{-k_1(x-n\Lambda)} \end{aligned} \tag{A.85}$$

Aangezien $\psi(x)$ een Blochfunctie is, kunnen we schrijven

$$\psi(x+\Lambda) = \psi(x) e^{jk\Lambda} \tag{A.86}$$

Deze laatste twee vormen van $\psi(x)$ kunnen met elkaar verzoend worden als er geldt

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} = \begin{bmatrix} a_{n+1} \\ b_{n+1} \end{bmatrix} \mathrm{e}^{-jk\Lambda} \tag{A.87}$$

Gebruik makende van (A.82) vinden we

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} a_n \\ b_n \end{bmatrix} = e^{-jk\Lambda} \begin{bmatrix} a_n \\ b_n \end{bmatrix} \tag{A.88}$$

Dit eigenwaardeprobleem kan ook in de volgende vorm geschreven worden

$$\begin{bmatrix} A - e^{-jk\Lambda} & B \\ C & D - e^{-jk\Lambda} \end{bmatrix} \begin{bmatrix} a_n \\ b_n \end{bmatrix} = 0 \tag{A.89}$$

De voorwaarde dat er niet-triviale oplossingen zijn is dat de determinant van de matrix nul is. We vinden:

$$e^{-jk_{1,2}\Lambda} = \frac{1}{2}(A+D) \pm j\sqrt{1 - \left[\frac{1}{2}(A+D)\right]^2} \tag{A.90}$$

waarin de subscripts $1, 2$ in het linkerlid verwijzen naar de + en - in het rechterlid.

Hieruit kan de propagatieconstante k kan verkregen worden:

$$\cos(k\Lambda) = \frac{k_1^2 - k_2^2}{2k_1 k_2}\sin(k_2 t)\sinh(k_1 s) + \cos(k_2 t)\cosh(k_1 s) \tag{A.91}$$

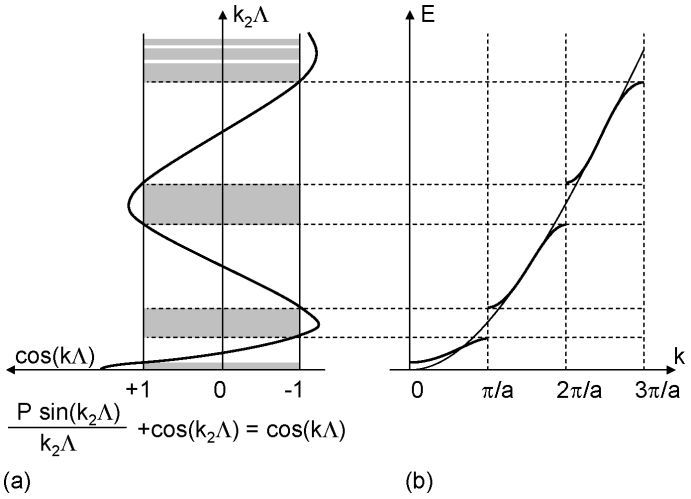

Figure A.9: Reële oplossingen in het Kronig-Penneymodel.

Om deze vergelijking te vereenvoudigen veronderstelden *Kronig* en *Penney* dat de barrières als deltafuncties opgevat kunnen worden. We laten V_0 naar oneindig lopen en s naar nul waarbij sV_0 constant blijft. (A.91) wordt dan

$$\frac{P\sin(k_2\Lambda)}{k_2\Lambda} + \cos(k_2\Lambda) = \cos(k\Lambda) \tag{A.92}$$

waarin

$$P = \frac{mV_0\, st}{\hbar^2} \tag{A.93}$$

Voor $P = 0$ vinden we $k_2\Lambda = k\Lambda$. Dit is identiek aan de situatie bij een vrij elektron. In de limiet $P \to \infty$ vinden we de situatie van een elektron in een geïsoleerde potentiaalput. Voor $0 < P < \infty$ zijn de elektronen in min of meerdere mate gebonden aan de individuele potentiaalputten. Aangezien $\cos(k\Lambda)$ tussen -1 en 1 ligt, legt (A.92) restricties op voor toegelaten k_2-waarden. Aangezien k_2 direct gerelateerd is aan de elektronenergie via (A.75), zullen er energiegebieden zijn waar geen reële k-waarden kunnen gevonden worden. In deze verboden zones is $k = m\pi/\Lambda + jk_I$ en wordt de factor $\exp(jkx)$ uit de Blochfunctie

$$\exp(jkx) = e^{j(m\pi/\Lambda)x}\exp(-|k_I|x) \tag{A.94}$$

$\psi_k(x)$ is dus een exponentieel evanescente functie. De situatie is geschetst in figuur A.9.

Index