Microphotonics

Dries Van Thourhout, Roel Baets (UGent)
Heidi Ottevaere (VUB)

Academic year 2017-2018

Universiteit Gent
Vakgroep Informatietechnologie (INTEC)
Technologiepark-Zwijnaarde 15
B-9052 Gent, België
Tel.: +32 9 264 3329 (RB)
+32 9 264 3438(DVT)
E-mail: roel.baets@ugent.be
dries.vanthourhout@ugent.be
WWW: www.photonics.intec.ugent.be

Vrije Universiteit Brussel
Vakgroep Toegepaste Natuurkunde
(TONA)
Pleinlaan 2
B-1050 Brussel, België
Tel.: +32 2 629 3451
E-mail: estijns@vub.ac.be
hottevaere@tona.vub.ac.be
WWW: www.tona.vub.ac.be

Contents

Chapter 1

Introduction

Contents

1.1 Microphotonics - what's in a name?

This course is a master-level course dealing essentially with propagation of light in a variety of complex structures. The understanding of the behavior of light propagation (and its interaction with materials) is then used to describe a variety of photonic components and systems. Furthermore we study how the behaviour of light propagation (and its interaction with materials) is applied in a variety of optical measurement systems. The course name "Microphotonics" may surprise a bit given this content description. The name is inspired by the fact that an increasing number of photonic systems makes use of miniaturized and/or highly integrated components in which light propagates through structures with smallest features of the order of the wavelength of light or even smaller. This evolution is reminiscent of the evolution in the field of electronics in the second half of last century, whereby components and systems were scaled down continuously and became increasingly complex, an evolution covered by the broad term microelectronics. The usage of small structures - thin films, waveguides, gratings, microlenses etc - opens up a wealth of applications in optical components and systems but it comes at a price. There is a need for a variety of models to describe the propagation of light through this variety of structures. In this course the most widely used models as well as analysis and design methods are described at various levels of abstraction. They will help the student to understand the operation principle of a multitude of photonic components and systems, as described in the later chapters of the course.

1.2 Objectives

This course builds upon the introductory course ”Photonics”. Many basic concepts are refreshed but in this course they are treated in a more rigorous way, covering a much broader range of structures, components and systems. The generic approach in the first chapters is to ”translate” a relatively complex wave propagation problem into a relatively simple analytical or semi-analytical model with a limited set of parameters. In some cases we also introduce basic numerical analysis and design methods. In the later chapters of the course the understanding gained from all these models is applied to a wide range of components, systems and applications. Both large optical systems (”macro”), as well as miniaturized systems (”micro” and even ”nano”) are covered.

This course has no experimental lab work attached to it, because the labs are part of a separate course entity ”Photonics laboratory”. The optical properties of materials are described in this course by means of a simple linear frequency-dependent refractive index (or tensor). An in depth treatment of material physics is provided in the course on ”Optical materials”. Lasers are not treated because they are part of the course on ”Lasers”. The mathematical treatments in this course are supported by the separate course ”Mathematics in photonics”.

At the end of this course the student will be able to understand the behaviour of a variety of photonic structures, components and systems on the basis of how light propagates through them. He will be able to break down a complex wave propagation problem in simpler generic problems and will have gained experience with a few design tools.

1.3 Background literature

The following books can be useful as reference book for subjects covered by the course. They can be consulted in the library of the INTEC department. They are not an explicit part of the course.

- M. Born and E. Wolf, ”Principles of Optics”, Pergamon Press
- M. Klein, T. Kurtak, ”Optics”, John Wiley
- K. D. Moller, ”Optics”,University Science Books
- J. Goodman, ”Introduction to Fourier Optics”, McGraw Hill second edition
- R. Martz , ”Integrated Optics, Designand Modeling” by, Artech House, Boston, London (ISBN 0-89006-668-X)
- Vassallo, ”Optical Wave Sciences and Technology, Part 1 Optical Waveguide Concepts”, Elsevier

Chapter 2

Matrix description of wave propagation and polarization

Contents

2.1 Electromagnetic waves

An electromagnetic wave propagating in a dielectric material with permittivity ε and permeability μ is described by Maxwell's equations in the time domain, relating the electric field **E**(**r**,t) and magnetic field **H**(**r**,t) in the absence of sources:

$$\begin{cases} \nabla \times \mathbf{E}(\mathbf{r},t) = -\mu \frac{\partial \mathbf{H}(\mathbf{r},t)}{\partial t} \\ \nabla \times \mathbf{H}(\mathbf{r},t) = \varepsilon \frac{\partial \mathbf{E}(\mathbf{r},t)}{\partial t} \end{cases} \tag{2.1}$$

In the frequency domain, the field components can be represented using the phasor notation

$$\mathbf{F}(\mathbf{r},t) = \mathrm{Re}(\tilde{\mathbf{F}}(\mathbf{r})e^{j\omega t}) = \frac{\tilde{\mathbf{F}}(\mathbf{r})e^{j\omega t} + \tilde{\mathbf{F}}^*(\mathbf{r})e^{-j\omega t}}{2} \tag{2.2}$$

in which **F** is a field vector (either **E** or **H**) and $\tilde{\mathbf{F}}(\mathbf{r})$ is a complex valued vector, representing the amplitude and phase of the sinusoidal time variation of **F** (at a frequency $f = \frac{\omega}{2\pi}$).

Using the phasor notation, we can rewrite Maxwell's equations 2.1 to get a set of differential equations relating the complex phasors representing the real field vectors.

$$\begin{cases} \nabla \times \tilde{\mathbf{E}}(r) = -j\omega\mu\tilde{\mathbf{H}}(r) \\ \nabla \times \tilde{\mathbf{H}}(r) = j\omega\varepsilon\tilde{\mathbf{E}}(r) \end{cases} \tag{2.3}$$

If an electromagnetic wave propagates in a uniform medium (ε and μ constants) equation 2.3 can be rewritten as (by applying the operator $\nabla \times \cdot$ to these equations and using $\nabla \cdot \varepsilon\tilde{\mathbf{E}} = 0$ and $\nabla \cdot \mu\tilde{\mathbf{H}} = 0$)

$$\begin{cases} \nabla^2\tilde{\mathbf{E}} + \omega^2\varepsilon\mu\tilde{\mathbf{E}} = 0 \\ \nabla^2\tilde{\mathbf{H}} + \omega^2\varepsilon\mu\tilde{\mathbf{H}} = 0 \end{cases} \tag{2.4}$$

called the *Helmholtz equations*. A general solution to these equations can be written as

$$\tilde{\mathbf{F}}(\mathbf{r}) = \tilde{\mathbf{F}}_+ e^{-j\mathbf{k}\cdot\mathbf{r}} + \tilde{\mathbf{F}}_- e^{j\mathbf{k}\cdot\mathbf{r}} \tag{2.5}$$

in which $\tilde{\mathbf{F}}(\mathbf{r})$ can be either the electric field phasor or the magnetic field phasor. The wave vector $\mathbf{k}$ can have an arbitrary orientation and is related to the frequency as

$$\mathbf{k} = \omega\sqrt{\varepsilon\mu}\mathbf{1}_\mathbf{k} \tag{2.6}$$

and defines the wavelength (in the material) λ as

$$|\mathbf{k}| = \frac{2\pi}{\lambda} \tag{2.7}$$

This is the well known plane wave solution. $\tilde{\mathbf{F}}_+$ represents a plane wave propagating in the direction $\mathbf{1}_k$, while $\tilde{\mathbf{F}}_-$ represents a plane wave propagating in the opposite direction. This can be seen by substituting the forward propagating part of equation 2.5 into equation 2.2 and assuming propagation along the z-axis ($\mathbf{k} = |\mathbf{k}|\,\mathbf{1_z}$). The time variation of the field can then be written as

$$F(z,t) = \left|\tilde{\mathbf{F}}_+\right| \cos(\omega t - kz + \arg(\tilde{\mathbf{F}}_+)) \tag{2.8}$$

This means that planes of constant phase ϕ move along the positive z-axis with a phase velocity

$$v_{ph} = \frac{\omega}{k} \tag{2.9}$$

In what follows, we will omit the $\tilde{}$ and implicitly assume that we are dealing with phasors. Inserting equation 2.5 in equation 2.3 we find that

$$\begin{cases} -j\mathbf{k} \times \mathbf{E}_+ e^{-j\mathbf{k}\cdot\mathbf{r}} + j\mathbf{k} \times \mathbf{E}_- e^{+j\mathbf{k}\cdot\mathbf{r}} = -j\omega\mu(\mathbf{H}_+ e^{-j\mathbf{k}\cdot\mathbf{r}} + \mathbf{H}_- e^{+j\mathbf{k}\cdot\mathbf{r}}) \\ -j\mathbf{k} \times \mathbf{H}_+ e^{-j\mathbf{k}\cdot\mathbf{r}} + j\mathbf{k} \times \mathbf{H}_- e^{+j\mathbf{k}\cdot\mathbf{r}} = j\omega\varepsilon(\mathbf{E}_+ e^{-j\mathbf{k}\cdot\mathbf{r}} + \mathbf{E}_- e^{+j\mathbf{k}\cdot\mathbf{r}}) \end{cases} \tag{2.10}$$

Identifying corresponding terms results in a relation between electric and magnetic field

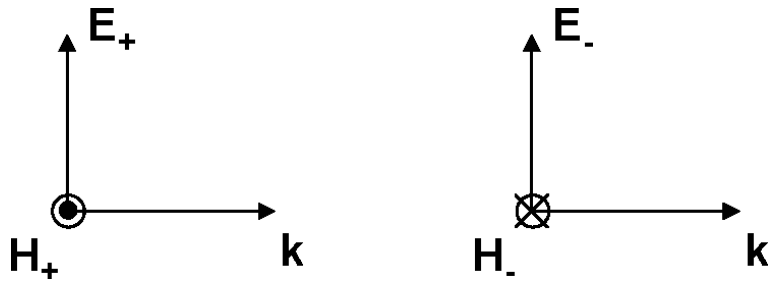


Figure 2.1: Orientation of field vectors for the plane wave solution

$$\begin{cases} \mathbf{k} \times \mathbf{E}_+ = \omega\mu\mathbf{H}_+ \\ \mathbf{k} \times \mathbf{E}_- = -\omega\mu\mathbf{H}_- \\ \mathbf{k} \times \mathbf{H}_+ = -\omega\varepsilon\mathbf{E}_+ \\ \mathbf{k} \times \mathbf{H}_- = \omega\varepsilon\mathbf{E}_- \end{cases} \tag{2.11}$$

This means that electric field vector, magnetic field vector and wave vector are all orthogonal as shown in figure 2.1.

If we now consider the case of an isotropic z-invariant dielectric waveguide described by permittivity $\varepsilon(x, y)$ and permeability $\mu(x, y)$, we can suggest a forward propagating solution of the Maxwell equations of the form

$$\begin{cases} \mathbf{E}(x, y, z) = [\mathbf{e_T}(x, y) + e_z(x, y)\mathbf{1_z}]\, e^{-j\beta z} \\ \mathbf{H}(x, y, z) = [\mathbf{h_T}(x, y) + h_z(x, y)\mathbf{1_z}]\, e^{-j\beta z} \end{cases} \tag{2.12}$$

in which $\mathbf{e_T}(x, y)$ and $\mathbf{h_T}(x, y)$ are vectors in the xy plane. Substituting equation 2.12 into equation 2.3 we can write

$$\begin{aligned} &\beta\mathbf{e_T} - j\nabla_T e_z = -\omega\mu\mathbf{1_z} \times \mathbf{h_T} \\ &\beta\mathbf{h_T} - j\nabla_T h_z = \omega\varepsilon\mathbf{1_z} \times \mathbf{e_T} \\ &\nabla_T \cdot (\mathbf{1_z} \times \mathbf{e_T}) = j\omega\mu h_z \\ &\nabla_T \cdot (\mathbf{1_z} \times \mathbf{h_T}) = -j\omega\varepsilon e_z \end{aligned} \tag{2.13}$$

in which $\nabla_T = \mathbf{1_x}\frac{\partial}{\partial x} + \mathbf{1_y}\frac{\partial}{\partial y}$ and β is still unknown. This set of equations forms an eigenvalue equation, for which there is only a solution for certain values of β. The corresponding field profiles $\{\mathbf{e_T}, e_z\mathbf{1_z}, \mathbf{h_T}, h_z\mathbf{1_z}\}$ are called *eigenmodes* or *optical modes* of the waveguide. It is easy to see that if $\{\mathbf{e_T}, e_z\mathbf{1_z}, \mathbf{h_T}, h_z\mathbf{1_z}\}\, e^{-j\beta z}$ is a solution of equation 2.13 then $\{\mathbf{e_T}, -e_z\mathbf{1_z}, -\mathbf{h_T}, h_z\mathbf{1_z}\}\, e^{j\beta z}$ is also a solution.

If β is assumed to be purely real (the mode is propagating without loss) and the waveguide is lossless (ε is real), then $\mathbf{e_T}$ and $\mathbf{h_T}$ are purely real and e_z and h_z are purely imaginary.

One can show that all waveguide modes are orthogonal or

$$\frac{1}{2}\int\int \mathbf{e}_i \times \mathbf{h}_{i'} \cdot \mathbf{1}_z dS = C\delta_{ii'} \tag{2.14}$$

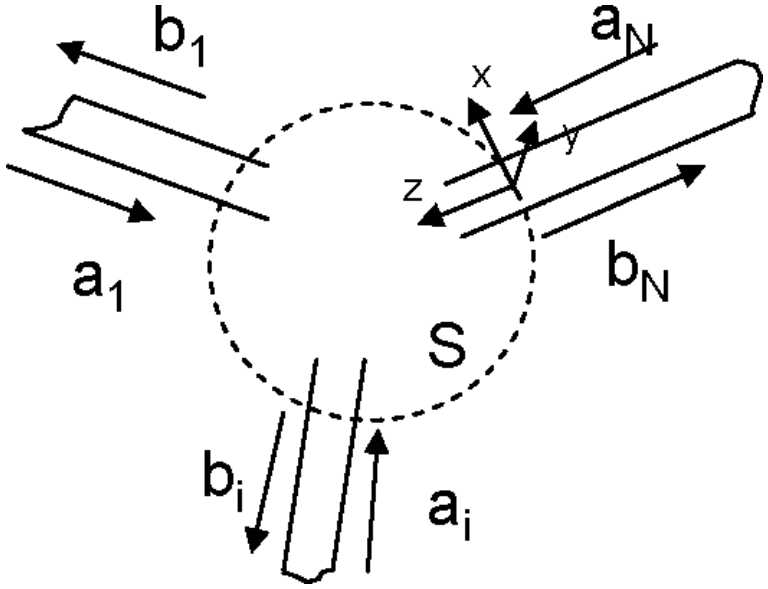


Figure 2.2: Schematics of an (optical) circuit

The fields $\mathbf{e}_i$ and $\mathbf{h}_i$ are defined as

$$\begin{cases} \mathbf{e}_i = \mathbf{e}_{\mathbf{T},\mathbf{i}} + e_{z,i}\mathbf{1}_\mathbf{z} \\ \mathbf{h}_i = \mathbf{h}_{\mathbf{T},\mathbf{i}} + h_{z,i}\mathbf{1}_\mathbf{z} \end{cases} \tag{2.15}$$

and correspond with an eigenvalue β_i.

In the case of a lossless waveguide equation 2.14 can be rewritten, due to the above mentioned properties of the fields, as

$$\frac{1}{2}\int\int \mathbf{e}_i \times \mathbf{h}^*_{i'} \cdot \mathbf{1}_z dS = C\delta_{ii'} \tag{2.16}$$

2.2 Matrix description of wave propagation in linear systems

2.2.1 Introduction

In this section we will gain insight into the global transmission properties of electromagnetic circuits. The results are applicable to microwave circuits as well as to optical circuits. We will look at the circuit as a black box, which exchanges energy with the outside through several physical outlets that can be optical waveguides or mere free space electromagnetic beams as shown in figure 2.2.

In figure 2.2 a_i and b_i represent the complex amplitude of the ingoing and outgoing normalized electromagnetic mode (carrying unit power) using the phasor notation. This means that the transversal electric and magnetic field at position z=0 (with position z=0 chosen to be at the intersection of the outlet and the surface S of the black box) of port i can be written as

$$\begin{aligned}\mathbf{E}_{T,i}(x,y,t) &= \mathrm{Re}\left[(a_i+b_i)\mathbf{e}_{T,i}(x,y)e^{j\omega t}\right] \\ \mathbf{H}_{T,i}(x,y,t) &= \mathrm{Re}\left[(a_i-b_i)\mathbf{h}_{T,i}(x,y)e^{j\omega t}\right]\end{aligned} \tag{2.17}$$

with $\mathbf{e}_{T,i}(x,y)$ and $\mathbf{h}_{T,i}(x,y)$ being the electric and magnetic transversal field profile of the normalized electromagnetic mode respectively.

The ports are considered to be lossless waveguides and the electrical field is assumed to be zero everywhere on S outside the ports. Every outlet has its own propagation axis and it is essential to assume that electromagnetic modes are well confined around these axes, so that effective outlets can be defined outside which the fields are negligible.

In figure 2.2, the outlet ports are assumed to carry only one mode. This is not a restriction as different outlet ports may physically coincide to describe multi mode ports (due to the orthogonality of the waveguide modes).

In the following matrix formalisms no attention is paid to the internal details of the circuit. The only requirements are that the system has to be passive and linear. A circuit is passive if there are no sources of energy outside the circuit.

2.2.2 Scattering matrix description

As the ports are considered to be lossless waveguides, the orthonormality relation for the fields can be written as

$$\frac{1}{2}\int\int \mathbf{e}_i \times \mathbf{h}^*_{i'} \cdot \mathbf{1}_z dS = \delta_{ii'} \tag{2.18}$$

integrated over the surface S and the unit vector $\mathbf{1}_z$ directed inwards. The power absorbed in the circuit is obtained by adding the net powers entering the various ports

$$P = \sum \left(|a_i|^2 - |b_i|^2\right) = \mathbf{A}^\dagger\mathbf{A} - \mathbf{B}^\dagger\mathbf{B} \tag{2.19}$$

with **A** being the column vector of incident field amplitudes and **B** the column vector with the amplitudes of the outgoing fields. Here † represents the Hermitian conjugate of the vector **A**, being the complex conjugate of the transpose.

Since we assumed the N-port circuit to be linear, the relation between **A** and **B** can be described by an NxN scattering matrix **S**:

$$\mathbf{B} = \mathbf{S}\mathbf{A} \tag{2.20}$$

with $\mathbf{A} = (a_1, a_2, ..., a_N)^T$ and $\mathbf{B} = (b_1, b_2, ..., b_N)^T$.

If all the terms of **A** are zero except $a_i = 1$, the output waves are given by the i-th column of **S**. The diagonal term S_{ii} is the reflection coefficient of mode i, the non-diagonal terms S_{ki} are the transmission coefficients from the mode i towards the mode k.

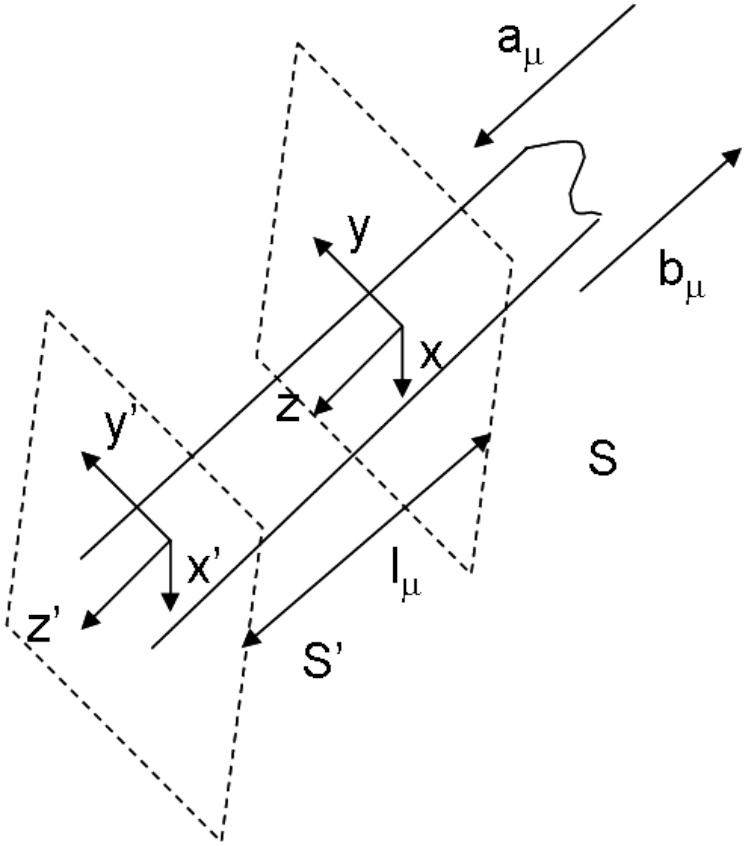


Figure 2.3: Moving the port planes

General properties of S matrices

- *Moving the port planes*

  The ports are changed when the surface S enclosing the circuit is changed. Let l_μ be the displacement of port μ counted positively if the port is moved towards the inside. The new circuit is then described by the scattering matrix $\mathbf{S}'$ related to the original scattering matrix $\mathbf{S}$ by

$$\mathbf{S}' = \mathbf{XSX} \tag{2.21}$$

  where $\mathbf{X}$ is the diagonal matrix formed with the elements $exp(jk_\mu l_\mu)$ with k_μ the propagation constant of mode μ.

- *Lossless circuits*

  Since lossless circuits absorb no power, we must have $P = 0$ in equation 2.19, hence

$$\mathbf{A}^\dagger(\mathbf{I} - \mathbf{S}^\dagger\mathbf{S})\mathbf{A} = 0 \tag{2.22}$$

  for all possible $\mathbf{A}$, with $\mathbf{I}$ the unity matrix. Therefore a sufficient condition for a circuit to be lossless is that

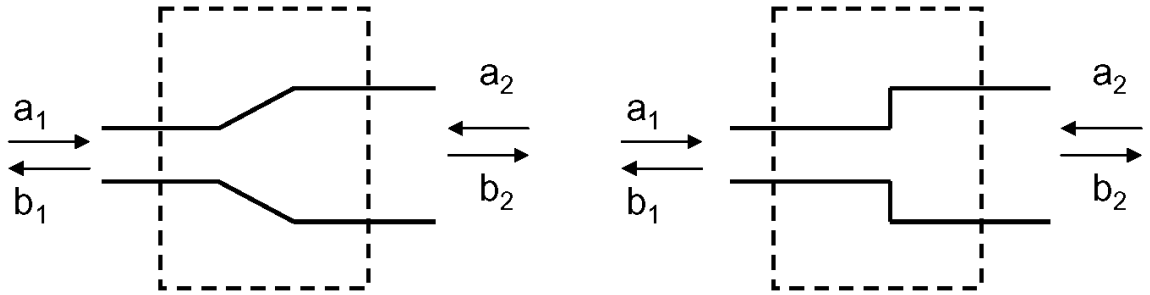


Figure 2.4: A change of waveguide width

$$S^\dagger S = \mathbf{I} \tag{2.23}$$

One can prove that this is also a necessary condition. A matrix satisfying equation 2.23 is called *unitary*.

In the case of lossy circuits we must have $P > 0$ which leads to

$$\mathbf{A}^\dagger \mathbf{S}^\dagger \mathbf{S} \mathbf{A} \leq \mathbf{A}^\dagger \mathbf{A} \tag{2.24}$$

Inserting $a_\mu = \delta_{k\mu} (\mu = 1 \cdots N)$ into equation 2.24 one obtains

$$\sum_{j=1}^{N} |S_{jk}|^2 \leq 1 \tag{2.25}$$

for $k = 1 \cdots N$.

- *Reciprocal circuits*

  If a circuit is made out of materials with symmetrical constitutive parameters ($\varepsilon^T = \varepsilon$ and $\mu^T = \mu$) the circuit is called *reciprocal*. This is the case for almost all materials, except for magnetic materials in the presence of a magnetic field. From Maxwell's equations one can deduce that the scattering matrix of a reciprocal circuit is *symmetrical*. This means that the transmission between port j and k does not depend on the propagation sense. Note that reciprocal circuits do not have to be symmetric themselves! This conclusion can have important consequences. A typical example is illustrated in Figure 2.4. Naively one might think that the transmission from the smaller left port to the larger right port will always be larger than the transmission from right to left. However, if both ports are monomodal waveguides, and taking into account that the scattering matrix for reciprocal circuits should be symmetric, we know that $S_{12} = S_{21}$ and therefore the transmission from port 1 to port 2 should be equal to the transmission from port 2 to port 1 ! (hereby we assume that all radiation is absorbed within the closed volume)

  Note that for symmetric circuits the diagonal elements of the scattering matrix should all be equal.

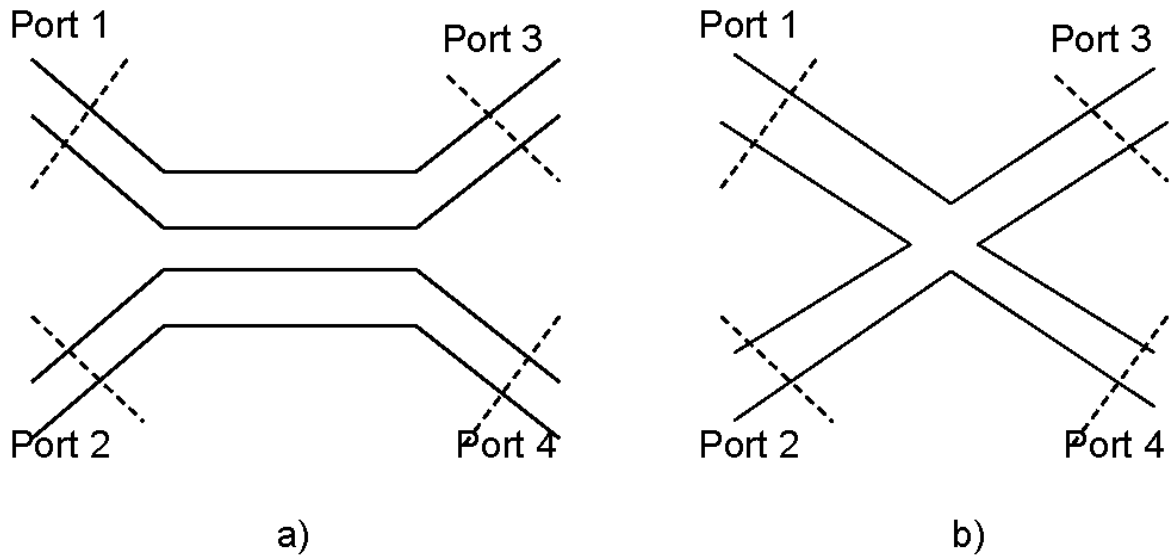


Figure 2.5: Beamsplitter in a 4 port circuit representation

Elementary applications in optics

- *Beam-splitter*

  Consider the so called directional coupler shown in figure 2.5a. This is a 4-port circuit described by a 4x4 scattering matrix. The operation of this device will be studied in detail in the chapter on dielectric waveguides but here it is sufficient to know that if two dielectric waveguides are brought closely together, light can couple between both waveguides. In addition we know that there is no transmission between port 1 and 2 and between port 3 and 4 and there are no reflections. If the system is reciprocal and lossless, the scattering matrix can therefore be written as

$$\mathbf{S} = \begin{bmatrix} 0 & 0 & a & b \\ 0 & 0 & c & d \\ a & c & 0 & 0 \\ b & d & 0 & 0 \end{bmatrix} \tag{2.26}$$

  because of reciprocity (symmetry of the scattering matrix) and the properties of the directional coupler as described above. Because the system is lossless, **S** has to be unitary, so

$$\begin{cases} |a|^2 + |b|^2 = 1 \\ |c|^2 + |d|^2 = 1 \\ |b|^2 + |d|^2 = 1 \\ |a|^2 + |c|^2 = 1 \\ ac^* + bd^* = 0 \\ ba^* + dc^* = 0 \end{cases} \tag{2.27}$$

  hence $|a| = |d|$ and $|b| = |c|$. It is not difficult to prove that the last equation of 2.27 is redundant: from all the other equations of 2.27 one can derive the last one and therefore it

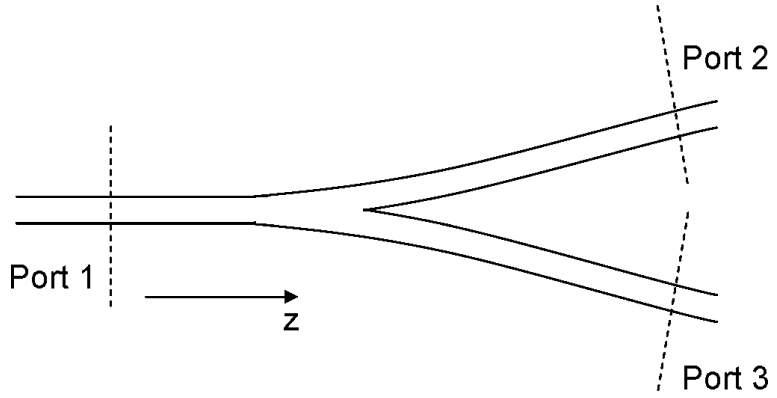


Figure 2.6: Optical power divider

can be left out. With a proper choice of the position of the ports one can use symmetry to simplify the problem. Consider a symmetry plane that images ports 1 and 2 onto 3 and 4 respectively and a symmetry point in the center of the device, which images ports 1 and 2 onto 4 and 3 respectively. In this case we get:

$$\begin{aligned} S_{13} &= S_{24} = a = d \\ S_{14} &= S_{32} = b = c \end{aligned} \tag{2.28}$$

By positioning the port planes appropriately we can also choose a and d to be real. From equation 2.27 it is clear that b and c are then purely imaginary. This means that a and c (and b and d) have a quadrature phase relationship. From this we can conclude that splitting a light beam in a lossless way always means that the resulting light beams have a quadrature phase relationship. Exactly the same reasoning can be followed to derive the scattering matrix of the beamsplitter of figure 2.5b.

- *Optical power divider*

  The simplest power divider, as shown in figure 2.6 is a 3 port circuit that divides the power entering port 1 equally into ports 2 and 3 with no reflection back to port 1. If the device is made out of dielectric optical waveguides and the circuit is slowly varying along the z-axis we can assume that there is no reflection at all when light comes from the right and no light is coupled from port 2 to 3 and vice versa. From the fact that the circuit is reciprocal and symmetrical in port 2 and 3, we can write down its scattering matrix.

$$\mathbf{S} = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & 0 \\ \frac{\sqrt{2}}{2} & 0 & 0 \end{bmatrix} \tag{2.29}$$

  This matrix is not unitary: ideal optical power dividers are lossy circuits even if they are made with lossless dielectric waveguides. The whole power can even be lost if port 2 and 3 are excited with waves having opposite phase. The reason is that the power is converted into radiating modes, which are not included in this scattering matrix description.

- *Plane waves*

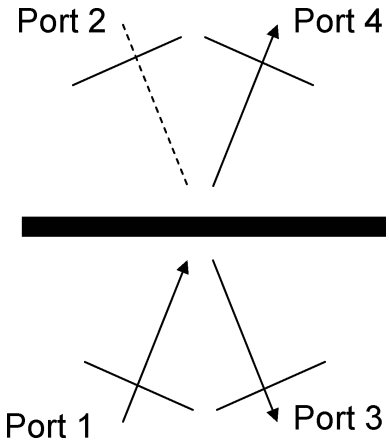


Figure 2.7: Optical power divider

Above, the "ports" of the system were eigenmodes of optical waveguides. It is possible to repeat the same reasoning taking into account a set of plane waves as the ports of a given system. Although these are not finite in extent, they are orthogonal and all properties derived above for the scattering matrix remain valid. An example is given by the beamsplitter of figure 2.7. The scattering matrix for this beamsplitter is identical to the one of the directional coupler of figure 2.5a and given by equation 2.26.

2.2.3 Transfer matrix description

The transfer matrix of an electromagnetic circuit is a matrix that describes the relation between the output quantities and the input quantities. This representation of a circuit becomes very useful when cascading different circuits, as the transfer matrix of the entire system can easily be found by multiplication of the transfer matrices of the individual circuits of the cascade. The ports of the circuit are hereby divided into ports considered as input ports of the circuit (schematically represented at the left side of the black box) and output ports (schematically represented at the right side of the black box).

Let us again consider the N-port circuit shown in 2.2, and assume that N is even. We can choose $N/2$ ports to be the left ports and $N/2$ ports to be the right ports as shown in figure 2.8. Incident modes on the left [right] ports are represented by the $N/2 \times 1$ column matrix $\mathbf{a}_l$ [$\mathbf{a}_r$] and outgoing modes are represented by $\mathbf{b}_l$ ($\mathbf{b}_r$). As the system is linear we can relate the incident and reflected waves at the input ports to the incident and reflected waves at the output port by a transfer matrix $\mathbf{T}$. Using block matrix representation we can write

$$\begin{bmatrix} \mathbf{b}_l \\ \mathbf{a}_l \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix} \tag{2.30}$$

If we now consider a cascade of p N-port networks, the transfer matrix of this cascade is the product of the transfer matrices of the individual N-ports. This can be easily understood for the case of the cascade of two N-ports (and generalized to p N-ports) as shown in 2.9. Here the $N/2$

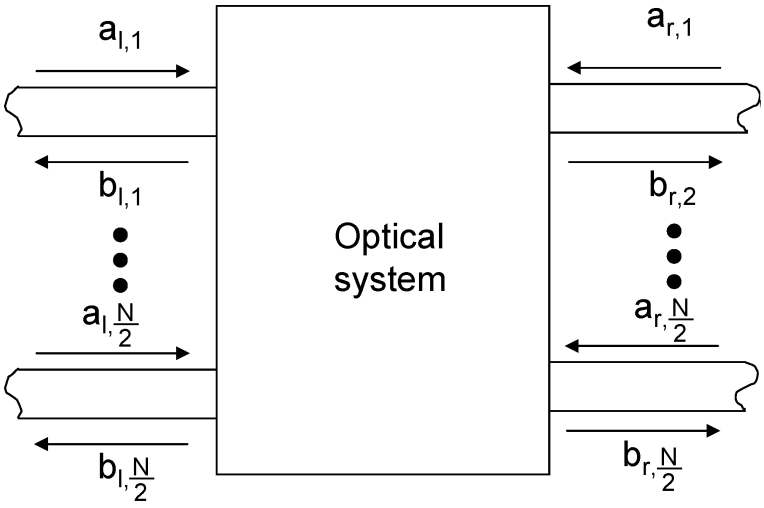


Figure 2.8: N port circuit

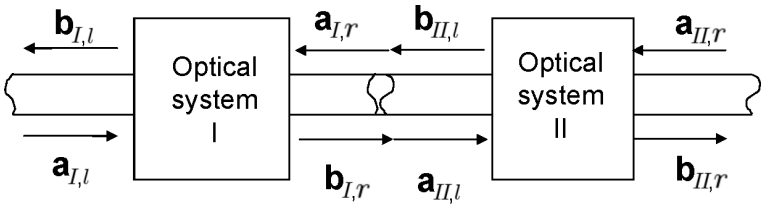


Figure 2.9: Schematic representation of the cascade of two N ports

left and $N/2$ right ports are schematically represented by two single ports described by the column vectors $\mathbf{a}_l$, $\mathbf{b}_l$,$\mathbf{a}_r$ and $\mathbf{b}_r$.

For this structure we can write

$$\begin{bmatrix} \mathbf{b}_{I,l} \\ \mathbf{a}_{I,l} \end{bmatrix} = \mathbf{T}_I \begin{bmatrix} \mathbf{a}_{I,r} \\ \mathbf{b}_{I,r} \end{bmatrix} = \mathbf{T}_I \begin{bmatrix} \mathbf{b}_{II,l} \\ \mathbf{a}_{II,l} \end{bmatrix} = \mathbf{T}_I \mathbf{T}_{II} \begin{bmatrix} \mathbf{a}_{II,r} \\ \mathbf{b}_{II,r} \end{bmatrix} = \mathbf{T}_{casc} \begin{bmatrix} \mathbf{a}_{II,r} \\ \mathbf{b}_{II,r} \end{bmatrix} \tag{2.31}$$

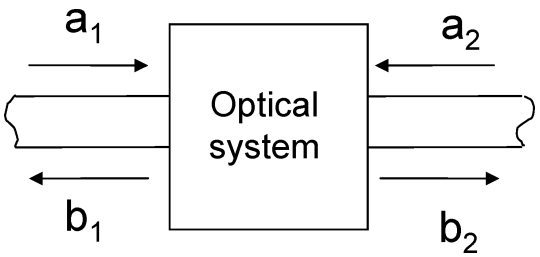


Figure 2.10: Two port circuit

Relation to the scattering matrix description

As both the scattering matrix description and the transfer matrix description relate the same parameters using a different set of input and output parameters, both descriptions are strongly related. To find the matrices $\mathbf{T}_{ij}$ from equation 2.30, we start from the scattering matrix description, relating output to input:

$$\begin{bmatrix} \mathbf{b}_l \\ \mathbf{b}_r \end{bmatrix} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{a}_l \\ \mathbf{a}_r \end{bmatrix} \tag{2.32}$$

Or, equivalently,

$$\mathbf{b}_l = \mathbf{S}_{11}\mathbf{a}_l + \mathbf{S}_{12}\mathbf{a}_r \tag{2.33}$$

$$\mathbf{b}_r = \mathbf{S}_{21}\mathbf{a}_l + \mathbf{S}_{22}\mathbf{a}_r \tag{2.34}$$

If $\mathbf{S}_{21}$ is invertible, one can rewrite 2.34 to:

$$\mathbf{a}_l = (\mathbf{S}_{21})^{-1}\left[\mathbf{b}_r - \mathbf{S}_{22}\mathbf{a}_r\right] \tag{2.35}$$

Substituting 2.35 into 2.33 and comparing to the matrix equation 2.30 yields the relationship between the the **T** and **S** matrices.

If we limit ourself to the case of a two port as shown in 2.10, the scattering matrix and transfer matrix are related by

$$\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix} = \frac{1}{S_{21}} \begin{bmatrix} -(S_{11}S_{22} - S_{12}S_{21}) & S_{11} \\ -S_{22} & 1 \end{bmatrix} \tag{2.36}$$

This relation is useful because for some circuits it is easy to derive the scattering matrix, but not its transmission matrix. Equation 2.36 relates both representations.

If the circuit is reciprocal, then $S_{12} = S_{21}$ which gives $det(\mathbf{T}) = \frac{S_{12}S_{21}}{S_{21}^2} = 1$.

2.3 Matrix description of light polarization

2.3.1 Polarized waves

Polarization refers to the time dependence of the electric field vector **E**(**r**,t) at a point **r** in space. The orientation of the electric wavevector changes in time at a certain point **r**. Once we know the polarization state of this electric field vector **E**, the polarization of the three remaining field vectors **D**, **H** and **B** can be found through Maxwell's equations. Therefore, in the following we will concentrate on the electric field vector **E**.

Assume a monochromatic wave at position **r**. The electric field vector is described by

$$\mathbf{E}(\mathbf{r}, t) = E_x(t)\mathbf{1}_x + E_y(t)\mathbf{1}_y + E_z(t)\mathbf{1}_z \tag{2.37}$$

Each component (i=x,y,z) can be written as

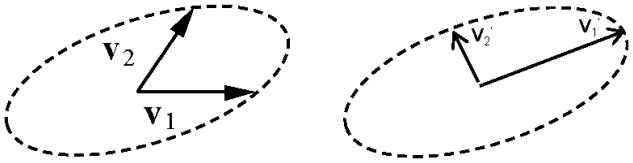


Figure 2.11: Elliptical polarization state: (a) general and (b) with a appropriately chosen origin of the time axis

$$E_i(t) = A_i \cos(\omega t + \phi_i) = (A_i \cos(\phi_i)) \cos(\omega t) - (A_i \sin(\phi_i)) \sin(\omega t) \tag{2.38}$$

using equation 2.38, equation 2.37 becomes

$$\mathbf{E}(\mathbf{r}, t) = (\sum_i A_i \cos(\phi_i)\mathbf{1}_i) \cos(\omega t) - (\sum_i A_i \sin(\phi_i)\mathbf{1}_i) \sin(\omega t) = \mathbf{v}_1 \cos(\omega t) + \mathbf{v}_2 \sin(\omega t) \tag{2.39}$$

From this equation it is clear that, for a strictly monochromatic wave, the most general polarization state is the *elliptical polarization* state for which the end-point of the electric field vector describes an ellipse in space. Such an ellipse is periodically described at a repetition rate equal to the optical frequency $f = \frac{\omega}{2\pi}$. This is visualized in figure 2.11a. By appropriately choosing the origin of the time axis t, the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ can be made to coincide with the main axes of the ellipse, as shown in figure 2.11b, where the new vectors are called $\mathbf{v}'_1$ and $\mathbf{v}'_2$. These new vectors are now orthogonal to each other. If the x-axis and y-axis are also chosen along those directions (with new axes x' and y'), the electric field can finally be written as

$$\mathbf{E}(\mathbf{r}, t) = \mathbf{v}'_1 \cos(\omega t') + \mathbf{v}'_2 \sin(\omega t') = \mathbf{E}_{x'} \cos(\omega t') + \mathbf{E}_{y'} \sin(\omega t') \tag{2.40}$$

For a complete representation of the elliptical polarization we need to know

- the orientation in space of the plane of the ellipse of polarization
- shape and size of the ellipse
- sense of traveling the ellipse and absolute temporal phase (i.e. the orientation of the electric field vector at t=0)

The shape of the ellipse is determined by

- the *azimuth* θ, being the angle between the major axis of the ellipse and the positive direction of the X axis with $-\frac{\pi}{2} \leq \theta < \frac{\pi}{2}$.
- its *ellipticity* e, being the ratio of the length of the minor axis to the major axis

$$e = \frac{b}{a} \tag{2.41}$$

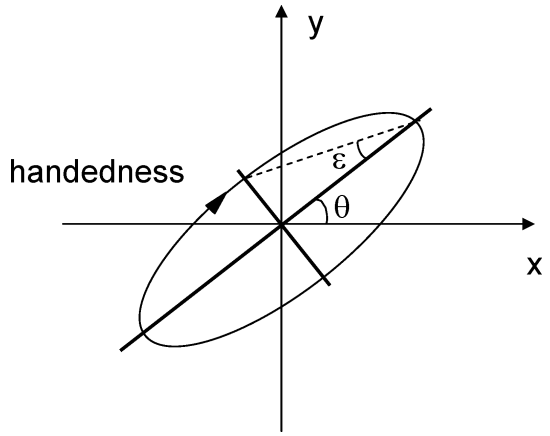


Figure 2.12: Azimuth and ellipticity angle

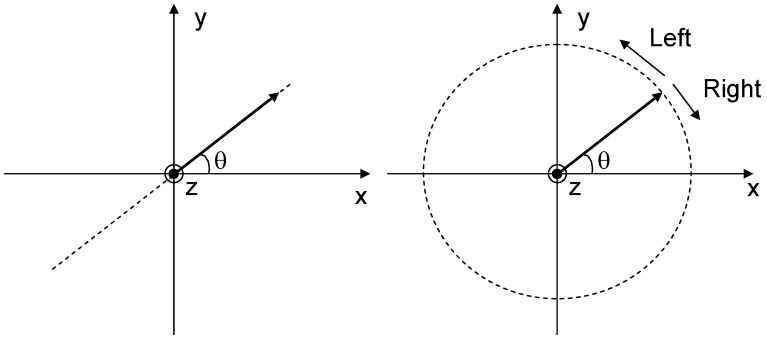


Figure 2.13: Linear and circular polarization

- its *handedness*, determining the sense in which the ellipse is described. The polarization is defined to be right handed (left handed) if the ellipse is traversed in a clockwise (counter-clockwise) sense seen when looking against the direction of propagation (into the beam).

It is convenient to incorporate the handedness in the definition of the ellipticity by allowing the ellipticity to assume positive (right-handed) and negative (left-handed) values. Therefore,

$$-1 \leq e \leq 1 \tag{2.42}$$

Together with the ellipticity an ellipticity angle ϵ is introduced $(-\frac{\pi}{4} \leq \epsilon \leq \frac{\pi}{4})$

$$\tan(\epsilon) = e \tag{2.43}$$

as shown in figure 2.12.

The circular and linear polarization state are special cases of the more general state of elliptical polarization as shown in figure 2.13, where we assume that the beam propagates in the positive z direction. For circular polarization $e = \pm 1$. For linear polarization $e = 0$. Circular polarizations have undetermined azimuth and linear polarizations have no handedness.

2.3.2 Representation of polarization

Jones vector of a plane monochromatic wave

The electric field component of a plane monochromatic wave, propagating in the positive z direction of a right handed Cartesian coordinate system xyz, is described by

$$\mathbf{E}(z,t) = E_x \cos(\omega t - \frac{2\pi}{\lambda} z + \phi_x)\mathbf{1}_x + E_y \cos(\omega t - \frac{2\pi}{\lambda} z + \phi_y)\mathbf{1}_y \tag{2.44}$$

Suppressing the fixed unit vectors $\mathbf{1}_x$ and $\mathbf{1}_y$, and the temporal information of the signal (as the frequency is known), choosing the reference plane at z=0 and using phasor notation, the plane wave can be represented by the Jones vector

$$\mathbf{E} = \begin{bmatrix} E_x e^{j\phi_x} \\ E_y e^{j\phi_y} \end{bmatrix} \tag{2.45}$$

The Jones vector is a complex vector that is a mathematical representation of a real wave. The intensity of the optical wave can be written as

$$I = |E_x|^2 + |E_y|^2 = \mathbf{E}^\dagger \mathbf{E} \tag{2.46}$$

A wave of unit intensity is said to be normalized and its Jones vector is said to be normal. Such a vector satisfies the condition

$$\mathbf{E}^\dagger \mathbf{E} = 1 \tag{2.47}$$

A normalized wave with Jones vector

$$\mathbf{E} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{2.48}$$

represents a linearly polarized wave along the x axis and vice versa for the linear polarization along the y axis. Left- and right-handed circularly polarized waves are represented by

$$\mathbf{E}_l = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -j \end{bmatrix} \tag{2.49}$$

$$\mathbf{E}_r = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ j \end{bmatrix} \tag{2.50}$$

In general, the Jones representation of a normalized elliptically polarized beam with azimuth θ and ellipticity angle ϵ is given by

$$e^{j\delta} \begin{bmatrix} \cos(\theta)\cos(\epsilon) - j\sin(\theta)\sin(\epsilon) \\ \sin(\theta)\cos(\epsilon) + j\cos(\theta)\sin(\epsilon) \end{bmatrix} \tag{2.51}$$

with $e^{j\delta}$ an arbitrary phase factor.

Two waves $\mathbf{E}_1$ and $\mathbf{E}_2$ are said to be orthogonal if

$$\mathbf{E}_1^\dagger \mathbf{E}_2 = \mathbf{E}_2^\dagger \mathbf{E}_1 = 0 \tag{2.52}$$

As every polarization state can be described as the superposition of two orthogonal (and normalized) polarization states, the same is true for the Jones vector. Every Jones vector can be described as the superposition of two orthonormal Jones vectors.

$$\mathbf{E} = A_1 \mathbf{E}_1 + A_2 \mathbf{E}_2 \tag{2.53}$$

with $\mathbf{E}_1^\dagger \mathbf{E}_2 = \mathbf{E}_2^\dagger \mathbf{E}_1 = 0$ and $\mathbf{E}_1^\dagger \mathbf{E}_1 = \mathbf{E}_2^\dagger \mathbf{E}_2 = 1$.

This means that $\mathbf{E}$ can be written as the superposition of any two linearly polarized waves with orthogonal polarization, as the superposition of two circularly polarized waves (one right-handed and one left-handed) and of any two orthogonal elliptical polarizations. Changing from one set of basis Jones vectors to another set implies a linear transformation to get the new representation of the same polarization state in the new set of basis vectors $\mathbf{E}_{old} = \mathbf{F}\mathbf{E}_{new}$. The columns of $\mathbf{F}$ represent the new basis vectors in the old coordinate system. To change from the cartesian basis vectors to the circular basis vectors, we can write

$$\begin{bmatrix} E_x \\ E_y \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -j & j \end{bmatrix} \begin{bmatrix} E_l \\ E_r \end{bmatrix} \tag{2.54}$$

Poincaré sphere

As the polarization ellipse of a monochromatic plane wave propagating in a known direction is completely determined by the azimuth θ and ellipticity angle ϵ, it is useful to represent the polarization state as a point on a sphere (with a radius equal to 1) as shown in figure 2.14. This point is completely characterized by its longitude 2θ and its latitude 2ϵ. As the equator represents polarization states with $\epsilon = 0$ these states are all linear polarizations. Above (below) the equator all polarization states are right-handed (left-handed). The north and south pole of the sphere represent the right-handed and left-handed circular polarization. Diametrically opposite points represent pairs of orthogonal polarization. This is easily verified using the Jones vector representation 2.51 and the definition of orthogonality 2.52.

2.3.3 Quasi-monochromatic waves

If we now consider a quasi-monochromatic wave, both field strength and phase become slowly varying functions in time

$$\mathbf{E}(z,t) = E_x(t)\cos(\omega t - \frac{2\pi}{\lambda}z + \phi_x(t))\mathbf{x} + E_y(t)\cos(\omega t - \frac{2\pi}{\lambda}z + \phi_y(t))\mathbf{y} \tag{2.55}$$

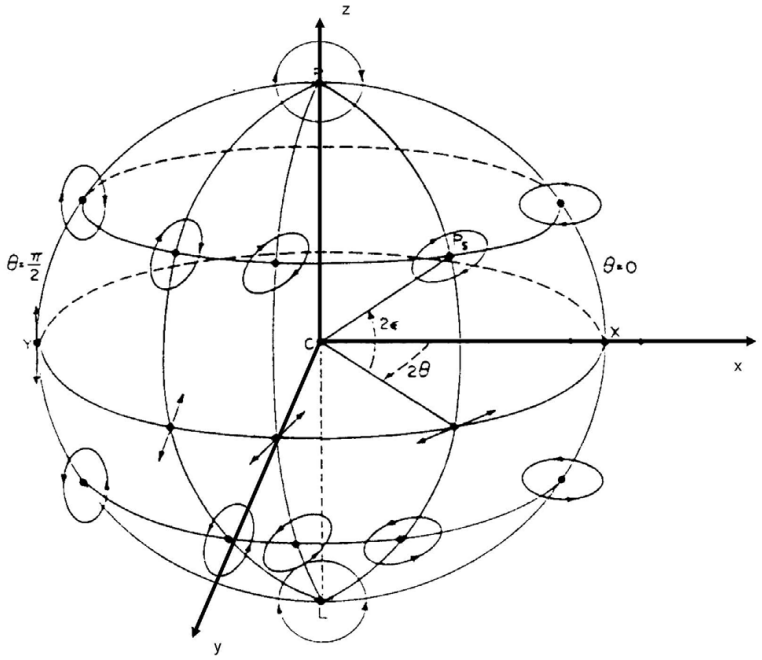


Figure 2.14: Poincaré sphere

The polarization state will still be an elliptical polarization but the shape and orientation of the ellipse will vary in time. This means that the Jones vector becomes a function of time and also the position on the Poincaré sphere will change with time.

If there is no correlation between the two components of the Jones vector, the wave is called unpolarized and the polarization state will describe a random walk on the Poincaré sphere. Partially polarized light describes a random walk around a particular point on the sphere. For completely polarized light, the point on the sphere is fixed in time. This requires

$$\frac{E_x(t)}{E_y(t)} = \text{constant} \tag{2.56}$$

and

$$\phi_x(t) - \phi_y(t) = \text{constant} \tag{2.57}$$

Quasi-monochronatic waves are mathematically treated using the Stokes parameters (S_0,S_1,S_2,S_3) defined as

$$\begin{cases} S_0 =< E_x^2(t) > + < E_y^2(t) > \\ S_1 =< E_x^2(t) > - < E_y^2(t) > \\ S_2 = 2 < E_x(t)E_y(t)\cos(\phi_x(t) - \phi_y(t)) > \\ S_3 = 2 < E_x(t)E_y(t)\sin(\phi_x(t) - \phi_y(t)) > \end{cases} \tag{2.58}$$

in which $< v >$ represents the time average of v.

S_0 represents the total intensity of the optical wave. Parameter S_1 is large (in absolute value) if the quasi-monochromatic wave has a preference for linear polarization along x or y, while S_2 represents the tendency towards linear polarization along $\pm 45^\circ$ (E_x and E_y have equal amplitudes and the phase difference between x and y direction is 0 or π). S_3 represents the preference for circular polarization (E_x and E_y have equal amplitudes and there is a phase difference $\frac{\pi}{2}$ or $\frac{3\pi}{2}$).

This definition implies that unpolarized light is represented by the Stokes parameters (S_0,0,0,0) because there is no preferential direction of polarization. Completely polarized light (still defined by its ellipticity angle ϵ and azimuth θ) is represented by (S_0,S_0cos(2ϵ)cos(2θ),S_0cos(2ϵ)sin(2θ),S_0sin(2ϵ)). This can be understood by writing down the Stokes parameters as a function of a time varying Jones vector (using equation 2.51)

$$\mathbf{E}(t) = \begin{bmatrix} E_{c,x} \\ E_{c,y} \end{bmatrix} = \begin{bmatrix} E_x(t)e^{j\phi_x(t)} \\ E_y(t)e^{j\phi_y(t)} \end{bmatrix} = A(t)e^{j\delta(t)} \begin{bmatrix} \cos(\theta)\cos(\epsilon) - j\sin(\theta)\sin(\epsilon) \\ \sin(\theta)\cos(\epsilon) + j\cos(\theta)\sin(\epsilon) \end{bmatrix} \tag{2.59}$$

as

$$\begin{cases} S_0 = < |E_{c,x}|^2 > + < |E_{c,y}|^2 > \\ S_1 = < |E_{c,x}|^2 > - < |E_{c,y}|^2 > \\ S_2 = 2 < \mathrm{Re}(E_{c,x}{}^* \, E_{c,y}) > \\ S_3 = 2 < \mathrm{Re}(-j \, E_{c,x}{}^* \, E_{c,y}) > \end{cases} \tag{2.60}$$

This representation implies that the Stokes parameters of a completely polarized beam with intensity $I = 1$ is represented by the parameters $(1, S_1, S_2, S_3)$ where (S_1, S_2, S_3) represent the cartesian coordinates of the point P on the Poincaré sphere using the coordinate system as shown in figure 2.14.

$$P(x, y, z) = (\cos(2\epsilon)\cos(2\theta), \cos(2\epsilon)\sin(2\theta), \sin(2\epsilon)) = (S_1, S_2, S_3) \tag{2.61}$$

For completely polarized light the Stokes parameters are related by

$$S_0^2 = S_1^2 + S_2^2 + S_3^2 \tag{2.62}$$

For partially or unpolarized light the inequality

$$S_0^2 \geq S_1^2 + S_2^2 + S_3^2 \tag{2.63}$$

is valid. These equations suggests that the general case of partially polarized light can be treated by splitting that wave into two components, a totally polarized component and an unpolarized component

$$\mathbf{S} = \mathbf{S}_{un} + \mathbf{S}_{tp} \tag{2.64}$$

where

$$\begin{aligned}
&\mathbf{S} = (S_0, S_1, S_2, S_3) \\
&\mathbf{S}_{un} = ([S_0 - \sqrt{(S_1^2 + S_2^2 + S_3^2)}], 0, 0, 0) \\
&\mathbf{S}_{tp} = (\sqrt{(S_1^2 + S_2^2 + S_3^2)}, S_1, S_2, S_3)
\end{aligned} \tag{2.65}$$

Note that each Stokes parameter of the original beam is obtained by adding the corresponding parameters of the component beams.

From the equations above we can define the degree of polarization $\wp$ as being the ratio of the intensity of the totally polarized component to the total intensity of the wave

$$\wp = \frac{\sqrt{(S_1^2 + S_2^2 + S_3^2)}}{S_0} \tag{2.66}$$

In terms of the total intensity I, the degree of polarization $\wp$, the azimuth θ and ellipticity angle ϵ of the ellipse of polarization of the totally polarized component, the Stokes parameters of any quasi-monochromatic wave can be cast in the form

$$\mathbf{S} = I(1, \wp\cos(2\epsilon)\cos(2\theta), \wp\cos(2\epsilon)\sin(2\theta), \wp\sin(2\epsilon)) \tag{2.67}$$

A normalized partially polarized wave ($I = 1$) is therefore represented by a point inside the Poincaré sphere with spherical coordinates ($\wp$,2θ,2ϵ).

The interpretation of the Stokes parameters suggests a simple experiment by which they may be measured for a given wave. Let I_0 denote the total intensity of the wave and let I_x, I_y, $I_{\frac{\pi}{4}}$,$I_{\frac{-\pi}{4}}$, I_l and I_r represent the intensities transmitted by an ideal variable polarizer placed in the path of the wave and adjusted to transmit the x,y,$\frac{\pi}{4}$,$\frac{-\pi}{4}$ linear polarizations and the left- and right-circular polarizations, respectively. In terms of these intensities the Stokes parameters are given by

$$\begin{cases}
S_0 = (I_x + I_y) = (I_l + I_r) = (I_{\frac{\pi}{4}} + I_{-\frac{\pi}{4}}) \\
S_1 = I_x - I_y \\
S_2 = I_{\frac{\pi}{4}} - I_{-\frac{\pi}{4}} \\
S_3 = I_r - I_l
\end{cases} \tag{2.68}$$

2.3.4 Propagation of polarized light through polarizing optical systems

The Jones-matrix formalism

Consider a uniform monochromatic plane wave that is incident onto a non-depolarizing optical system. A non-depolarizing system is defined as an optical system in which the degree of polarization of the output beam $\wp_o$ is larger than or equal to the degree of polarization of the input beam $\wp_i$. In the case where the input beam is completely polarized, this means that the output beam is also completely polarized.

As a result of the interaction between the incident wave and the optical system the polarization of the plane wave may change. The optical system including the incident optical wave and one of the outgoing waves is schematically represented in figure 2.15.

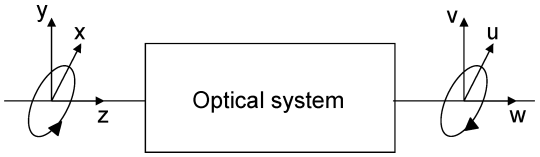


Figure 2.15: Schematic representation of a polarized beam incident to an optical system

Two cartesian coordinate systems (xyz) and (uvw) are associated with the incident and outgoing plane wave and the z and w directions are chosen parallel to the respective wavevectors **k** and **l**. The choice of the origin of both coordinate systems is arbitrary.

Using these coordinate systems we can represent the polarization state of the input beam by its Jones vector

$$\mathbf{E}_i = \begin{bmatrix} E_{ix} \\ E_{iy} \end{bmatrix} \tag{2.69}$$

and the polarization of the output beam

$$\mathbf{E}_o = \begin{bmatrix} E_{ou} \\ E_{ov} \end{bmatrix} \tag{2.70}$$

In the absence of non-linear processes we can relate the components of the Jones vector of the incident wave to that of the outgoing wave using the Jones matrix

$$\begin{bmatrix} E_{ou} \\ E_{ov} \end{bmatrix} = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} E_{ix} \\ E_{iy} \end{bmatrix} \tag{2.71}$$

Equation 2.71 expresses the law of interaction between the incident wave and the optical system as a simple linear matrix transformation of the Jones vector representation of the wave. The elements T_{ij} of the Jones matrix are in general complex. If optical systems are cascaded then the interaction with an incident beam and this combined system is represented by the matrix multipication of the individual Jones matrices.

- *Quarter wave plate*

  Let us consider an optical system that transforms a linear polarization state (azimuth angle $\theta = \frac{\pi}{4}$) to a right handed circular wave. The system also has to satisfy the requirement that linear polarization states along x and y direction remain unchanged. The Jones matrix then has to satisfy

  $$\frac{1}{\sqrt{2}} e^{-j\phi} \begin{bmatrix} 1 \\ j \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \tag{2.72}$$

  where B and C have to be zero because of the requirement for x and y polarization invariance. It's easy to see that the Jones vector representation of both x and y linear polarization are eigenvectors of the Jones matrix describing the optical system.

The Jones matrix of the optical system then has to be of the form

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} e^{-j\phi} & 0 \\ 0 & e^{-j(\phi-\frac{\pi}{2})} \end{bmatrix} \tag{2.73}$$

This means that the optical system has to retard the x polarization state and y polarization state differently (and up to a phase shift of $\frac{\pi}{2}$). This can be achieved using a plate of uniaxially anisotropic material in which x and y polarization experience a different refractive index (n_x and n_y respectively). The Jones matrix describing the polarization change through a plate of length z can be written as

$$\begin{bmatrix} e^{-jk_0 n_x z} & 0 \\ 0 & e^{-jk_0 n_y z} \end{bmatrix} \tag{2.74}$$

in which k_0 is the free space wave vector. If the thickness L of the plate satisfies

$$k_0(n_x - n_y)L = \frac{\pi}{2} \tag{2.75}$$

equation 2.74 and 2.73 are equivalent and all conditions are satisfied. A uniaxially anisotropic plate satisfying equation 2.76 is called a *quarter wave plate*. Notice that a quarter wave plate is not a quarter wavelength thick. Equation 2.76 can be rewritten as

$$n_x L - n_y L = \frac{\lambda}{4} \tag{2.76}$$

In other words: the *difference* between the optical thickness for x- and y-direction is a quarter wavelength.

- *Optical activity*

  As a second example we will consider an optical system that rotates an arbitrary linear polarization state over an angle θ. If we consider x and y polarization states we can immediately write down the Jones matrix of the optical system as

$$\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{2.77}$$

  Due to the superposition principle this optical system will also rotate an arbitrary linear polarization state over an angle θ.

  This Jones matrix can be realized using a plate of optically active material. This is a crystalline material with a helical molecular structure, so that a left-handed circular wave and a right-handed circular wave injected along the axis of the helix experiences a different phase velocity ($\frac{\omega}{k_l}$ and $\frac{\omega}{k_r}$) dependent on the rotation sense of the molecular structure. The Jones matrix of a plate with thickness z must satisfy

$$e^{-jk_r z}\begin{bmatrix} 1 \\ j \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 \\ j \end{bmatrix} \tag{2.78}$$

  and

$$e^{-jk_l z}\begin{bmatrix} 1 \\ -j \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} 1 \\ -j \end{bmatrix} \tag{2.79}$$

If we introduce the average wave number k and the specific rotary power ρ as

$$\begin{cases} k = \frac{k_r + k_l}{2} \\ \rho = \frac{k_r - k_l}{2} \end{cases} \tag{2.80}$$

then we can solve equation 2.78 and 2.79 to find the Jones matrix

$$e^{-jkz}\begin{bmatrix} \cos(\rho z) & -\sin(\rho z) \\ \sin(\rho z) & \cos(\rho z) \end{bmatrix} \tag{2.81}$$

Comparison with equation 2.77 indicates that the polarization state in a plate of optically active material, excited by a linear polarization state, is continuously rotating the linear polarization, determined by the specific rotary power.

Graphical representation of polarization change

As completely polarized light can be represented by a point on the Poincaré sphere, we can graphically track the change of polarization through an optical system by drawing the locus of polarization states throughout the optical system.

When a linearly polarized wave with Jones vector

$$\begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \tag{2.82}$$

is incident on a layer of an uniaxially anisotropic material with Jones matrix

$$\begin{bmatrix} 1 & 0 \\ 0 & e^{j\Delta kz} \end{bmatrix}, \tag{2.83}$$

with $\Delta k = k_0(n_x - n_y)$ and z the thickness of the layer, the Jones vector of the transmitted wave equals

$$\begin{bmatrix} \cos(\theta) \\ \sin(\theta)e^{j\Delta kz} \end{bmatrix}. \tag{2.84}$$

and the corresponding point on the Poincaré sphere is

$$P(x, y, z) = (\cos(2\theta), \sin(2\theta)\cos(\Delta kz), \sin(2\theta)\sin(\Delta kz)). \tag{2.85}$$

From this expression it is easy to understand that a point P on the Poincaré sphere (now with a general initial polarization), will, when it propagates through an uniaxially anisotropic material,

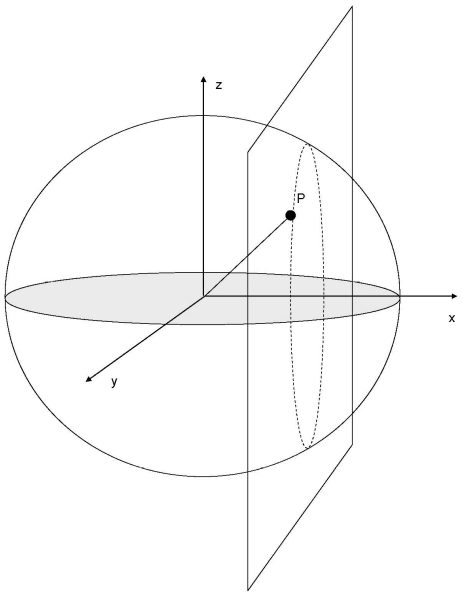


Figure 2.16: Representation of polarization change through an uniaxially anisotropic medium

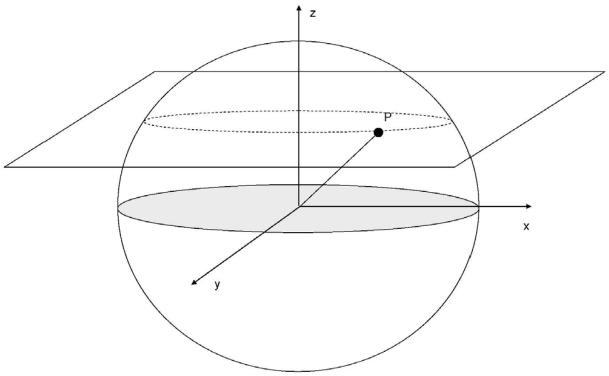


Figure 2.17: Representation of polarization change through an optically active material

describe a circle formed by the intersection of the Poincaré sphere and a plane parallel to the yz-plane (due to the invariance of S_1) that contains the initial polarization point **P** as shown in figure 2.16.

For the case of optical activity the point will describe a circle formed by the intersection of the Poincaré sphere and a plane parallel to the xy-plane that contains the initial polarization point **P** as shown in figure 2.17. We can prove this by considering a wave with the general polarization given in Equation 2.51, which is incident on an optical active material with the transfer matrix given in Equation 2.81. One can calculate that the point on the Poincaré sphere corresponding with the transmitted wave has coördinates:

$$P(x, y, z) = (\cos(2\epsilon)\cos(2(\theta + \rho z)), \cos(2\epsilon)\sin(2(\theta + \rho z)), \sin(2\epsilon)). \tag{2.86}$$

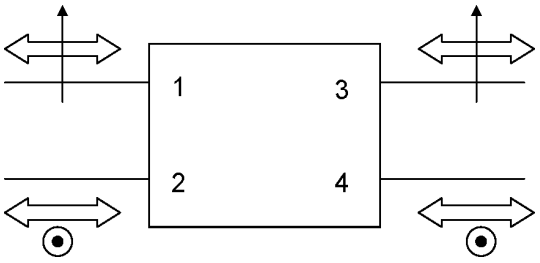


Figure 2.18: System description of a quarter wave plate. The orthogonal polarizations are parallel to the principal axes of the birefringent material.

Scattering matrix and transfer matrix description including polarization

As the scattering and transfer matrix description only assumes orthogonal modes in its ports, polarization changes through an optical system can also be described by a scattering and transfer matrix description when using orthogonal polarization states. The resulting scattering and transfer matrix will depend on the choice of orthogonal polarization states.

- *Quarter wave plate*

  Let us consider the quarter wave plate discussed in paragraph 2.3.4. If we choose the orthogonal polarization states as in figure 2.18, following the same axis convention as in equation 2.74, we can readily write the scattering matrix as

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & j \\ 1 & 0 & 0 & 0 \\ 0 & j & 0 & 0 \end{bmatrix} \tag{2.87}$$

  From the symmetry of the scattering matrix we see that a quarter wave plate is a reciprocal circuit.

- *Optical activity*

  Let us now consider an optical active material that rotates a linear polarization over an angle $\theta = \rho L$, with ρ the specific rotary power and L the thickness of the optically active plate. We choose the orthogonal polarization states to be circular and define the propagation constant of a circular wave propagating along (opposite to) the rotation sense of the molecules k_p (k_m). For the circuit represented in figure 2.19, the scattering matrix becomes

$$\begin{bmatrix} 0 & 0 & e^{-jk_pL} & 0 \\ 0 & 0 & 0 & e^{-jk_mL} \\ e^{-jk_pL} & 0 & 0 & 0 \\ 0 & e^{-jk_mL} & 0 & 0 \end{bmatrix} \tag{2.88}$$

  Because the scattering matrix is symmetric, this device is also reciprocal.

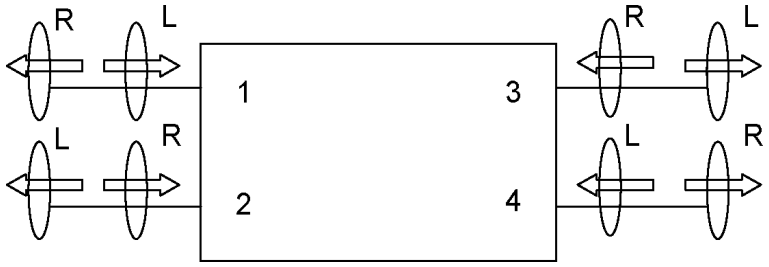


Figure 2.19: System description of an optically active material. The orthogonal polarizations are left- and right-handed circularly polarized light.

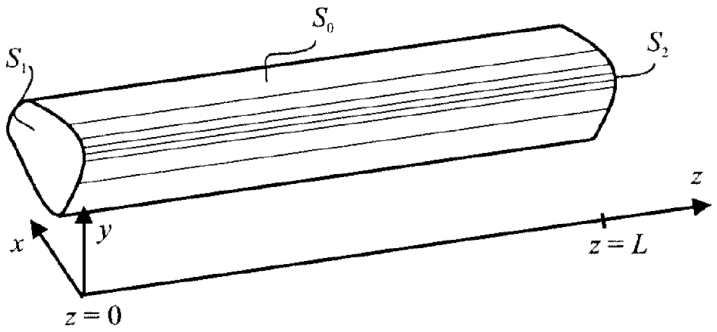


Figure 2.20: A two port enclosed by a surface S, with $S = S_0 + S_1 + S_2$.

2.A Appendix: Reciprocity

In this paragraph we will demonstrate how a reciproque component needs to have a symmetric scattering matrix. To simplify the mathematics, we will only prove this explicitly for a two port circuit, but the heuristic is the same for a general N port circuit.

Consider a linear, time-invariant optical two port system enclosed by a surface $S = S_0 + S_1 + S_2$, as depicted in figure 2.20. S_1 and S_2 are resp. the intersections at the left and right port of the system, such that the propagation direction of the eigenmodes at those ports is perpendicular to the corresponding surface. Moreover, for mathematical simplicity we will assume S_0 to be at infinity. If all materials within S have a symmetrical dielectric tensor ϵ and a symmetrical magnetic tensor μ, then the Lorentz reciprocity principle holds. When there are no optical sources within S, the Lorentz reciprocity theorem can be written as:

$$\oint_S \mathbf{e}_a \times \mathbf{h}_b \cdot \mathbf{ds} = \oint_S \mathbf{e}_b \times \mathbf{h}_a \cdot \mathbf{ds}, \tag{2.89}$$

with $(\mathbf{e}_a, \mathbf{h}_a)$ and $(\mathbf{e}_b, \mathbf{h}_b)$ two fields, caused by sources outside S. We now choose $(\mathbf{e}_a, \mathbf{h}_a)$ to be the solution for a wave incident from the left on S_1 only, and $(\mathbf{e}_b, \mathbf{h}_b)$ to be the field for a wave incident from the right on S_2.
We can then split the surface integrals over S in equation 2.89 into the surface integrals over S_0, S_1 and S_2. The surface integrals over S_0 at infinity are zero, as the modal fields decay there fast

enough (this is difficult to prove and this prove lies beyond the scope of these course notes). As the propagation direction of the eigenmodes at the ports is perpendicular to the surfaces, **ds** is parallel to $\pm\mathbf{1}_z$ at S_1 and S_2 and thus:

$$((\mathbf{e}_{T,\alpha} + e_{z,\alpha}\mathbf{1}_z) \times (\mathbf{h}_{T,\beta} + h_{z,\beta}\mathbf{1}_z)) \cdot \mathbf{ds} = (\mathbf{e}_{T,\alpha} \times \mathbf{h}_{T,\beta}) \cdot \mathbf{ds}, \tag{2.90}$$

which gives us:

$$\oint\limits_{S'=S_1+S_2} \mathbf{e}_{T,a} \times \mathbf{h}_{T,b} \cdot \mathbf{ds} = \oint\limits_{S'=S_1+S_2} \mathbf{e}_{T,b} \times \mathbf{h}_{T,a} \cdot \mathbf{ds}. \tag{2.91}$$

So we now only have to write the transverse field components of the a and b field at both S_1 and S_2 in function of the incident field with amplitude resp. c_a and c_b and the elements of the scattering matrix of this structure. We then get:

$$\begin{array}{ll}
\mathbf{e}_{T,a;1} = c_a(1 + S_{11})\mathbf{e}_{T;1} & \\
\mathbf{h}_{T,a;1} = c_a(1 - S_{11})\mathbf{h}_{T;1} & \text{at } S_1, \\
\mathbf{e}_{T,a;2} = c_a S_{12}\mathbf{e}_{T;2} & \\
\mathbf{h}_{T,a;2} = c_a S_{12}\mathbf{h}_{T;2} & \text{at } S_2, \\
\mathbf{e}_{T,b;1} = c_b S_{21}\mathbf{e}_{T;1} & \\
\mathbf{h}_{T,b;1} = -c_b S_{21}\mathbf{h}_{T;1} & \text{at } S_1, \\
\mathbf{e}_{T,b;2} = c_b(1 + S_{22})\mathbf{e}_{T;2} & \\
\mathbf{h}_{T,b;2} = -c_b(1 - S_{22})\mathbf{h}_{T;2} & \text{at } S_2.
\end{array} \tag{2.92}$$

In these formula ($\mathbf{e}_{T;i}$,$\mathbf{h}_{T;i}$) are the eigenmodes at surface S_i (with $i = 1$ or 2), while ($\mathbf{e}_{T,\alpha;i}$,$\mathbf{h}_{T,\alpha;i}$) (with $\alpha = a$ or b) are the total a/b-fields at surface S_i, caused by the incidence field, transmission through the structure and reflection on the structure. Substitution of these fields in equation 2.89 and reorganisation of the equations gives:

$$\begin{aligned}
-c_a(1 + S_{11})c_b S_{21} \oint\limits_{S_1} \mathbf{e}_T \times \mathbf{h}_T \cdot \mathbf{ds} - c_a S_{12} c_b(1 - S_{22}) \oint\limits_{S_2} \mathbf{e}_T \times \mathbf{h}_T \cdot \mathbf{ds} = \\
c_b S_{21} c_a(1 - S_{11}) \oint\limits_{S_1} \mathbf{e}_T \times \mathbf{h}_T \cdot \mathbf{ds} + c_b(1 + S_{22})c_a(S_{12}) \oint\limits_{S_2} \mathbf{e}_T \times \mathbf{h}_T \cdot \mathbf{ds}
\end{aligned} \tag{2.93}$$

$$\begin{aligned}
[c_a c_b\,(-(1 + S_{11})S_{21} - S_{21}(1 - S_{11}))] \oint\limits_{S_1} \mathbf{e}_{T;1} \times \mathbf{h}_{T;1} \cdot \mathbf{ds} = \\
[c_a c_b\,(S_{12}(1 + S_{22}) + S_{12}(1 - S_{22}))] \oint\limits_{S_2} \mathbf{e}_{T;2} \times \mathbf{h}_{T;2} \cdot \mathbf{ds}.
\end{aligned} \tag{2.94}$$

If we use the normalization of the eigenmodes ($\mathbf{e}_{T;i}$,$\mathbf{h}_{T;i}$) given in equation 2.18, assume lossless waveguides at the ports[1] (which makes it possible to choose ($\mathbf{e}_{T;i}$,$\mathbf{h}_{T;i}$) purely real), and take into account **ds** is codirectional with $-\mathbf{1}_z$ and $\mathbf{1}_z$ at resp. S_1 and S_2 we have:

$$\oint\limits_{S_1} \mathbf{e}_{T;1} \times \mathbf{h}_{T;1} \cdot \mathbf{ds} = -\oint\limits_{S_2} \mathbf{e}_{T;2} \times \mathbf{h}_{T;2} \cdot \mathbf{ds}, \tag{2.95}$$

[1]However, in principle we are not obliged to assume this. We can always normalize the fields using equation 2.14 with $C = 1$.

which gives us finally:

$$S_{12} = S_{21}. \tag{2.96}$$

The scattering matrix of this two port circuit is thus symmetric. In the same way we can prove this for a general N port circuit. The scattering matrix is symmetric if for every possible $i \neq j$ combination $S_{ij} = S_{ji}$. We thus need $N(N+1)/2$ independent equations. We can generate those equations by writing down for every possible $i \neq j$ combination:

$$\sum_k \oint_{S_k} \mathbf{e}_{T,a(i)} \times \mathbf{h}_{T,b(j)} \cdot \mathbf{ds} = \sum_k \oint_{S_k} \mathbf{e}_{T,b(j)} \times \mathbf{h}_{T,a(i)} \cdot \mathbf{ds}, \tag{2.97}$$

where the $a(i)/b(j)$-fields are now the fields caused by an incident field at resp. port i and j.

Chapter 3

Thin Films

Contents

3.1 Introduction

This chapter deals with the propagation of optical waves in and through thin films. Thin films are used for a very wide range of optical applications ranging from a thin metal film for use as a mirror to stacks of quarter-wave layers for highly reflective coatings in laser cavities. Thin films also serve as material systems for integrated photonic circuits such as polymer films for polymer waveguide circuits or SOI (silicon-on-insulator) circuits. In this chapter we will gain insight in the optical properties of thin films and provide tools to deal with thin film problems in a quantitative way.

The problem considered is schematically represented in Figure 3.1: a light source generates electromagnetic waves and illuminates a stack of layers of different media. It is clear that the propagation of electromagnetic waves generated by the light source will be heavily influenced by the layered stack of media. Along their path the travelling light waves will be reflected and transmitted multiple times by interfaces between different media. As a result, different waves will contribute to the electromagnetic field in each layer of the stack giving rise to interference phenomena. As a general result, part of the light incident on the stack of layers will be transmitted, part of it will be absorbed and part of it will be reflected. In this chapter we will introduce a systematic approach that allows to calculate the reflection and transmission of light at such a stack of layers. Moreover, this approach can be used to calculate the electromagnetic field in all layers of the stack, even for a very high number of layers. This method uses the transfer matrices for wave propagation through layers and at interfaces and will provide an efficient toolbox for a wide variety of thin film optical problems. Throughout this chapter, we consider linear, homogeneous and isotropic media. How-

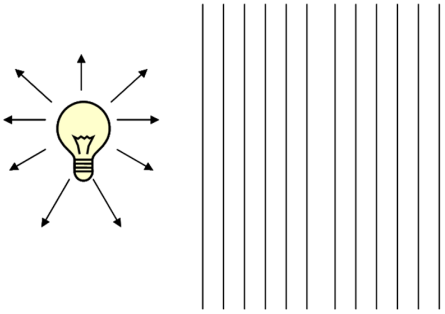


Figure 3.1: The problem of light propagation through layered media.

ever, we do not restrict ourselves to lossless media. As we will see, the transfer matrix formalism of wave propagation perfectly allows to deal with complex refractive indices.

Consider a general light source producing light with a discrete or continuous spectrum of frequencies. Given that the materials involved are linear, each frequency component of the light interacts independently with the media. Therefore it is legitimate to investigate the properties of layered media using monochromatic light, that is light with a unique frequency of oscillation, and do so for every frequency component in the spectrum. Moreover, we learned in the chapter on Fourier optics that in any plane in space each vector component of the electric (or magnetic) field $U(x, y, z)$ of a monochromatic wave can be expanded into plane waves. It follows that we can limit our investigation to the propagation of monochromatic plane waves.

The phenomenon of interference is fundamental for the understanding of the optical behaviour of thin films. This follows from the insight that multiple reflections and refractions by the different interfaces in the stack give rise to numerous field contributions that interfere with each other. Therefore, we will first treat interference before we go on with the transfer matrix formalism. We will end by investigating some examples of thin film applications.

3.2 Basics of interference

The term interference generally indicates that two (or more) phenomena are interacting with each other. Within the context of electromagnetism, this term may be a bit misleading. The reason is this: wave propagation of electromagnetic fields is described by Maxwell's equations. In linear media, these equations are linear differential equations and thus, waves will not interact with each other: the total field vector is always the sum of the individual field vectors. In other words, for a wave problem with two (or more) sources, the propagating fields do not influence each other, as the principle of superposition holds.

However, propagating waves often *seem* to interact in a non-linear way. This is because the total field depends strongly on the phase difference between the waves. In the end, this means that the total intensity is (often) not equal to the sum (or superposition) of the individual intensities. For example, if two waves with the same frequency and the same intensity come together, the total intensity can take any value between zero (destructive interference) and four times (constructive interference) the value of the intensity of the individual wave, depending on their relative phase

difference. The reason for the apparent interaction is that the quantity observed is not the amplitude of the waves but the energy density, which is a non-linear function of the electromagnetic field. Note that the concept of *phase difference* only makes sense for sources with (nearly) equal wavelengths.

3.2.1 Intensity

Any detector - be it the human eye, a photographic plate or an optical power meter - will detect the energy density $U(\mathbf{r},t)$ associated with the total electric field $\mathbf{E}(\mathbf{r},t)$. The response time of a detector is of course finite and therefore, a time-averaged energy density $\langle U(\mathbf{r},t)\rangle$ will be detected, defined as the intensity of the wave:

$$I = I(\mathbf{r},t) = \langle U(\mathbf{r},t)\rangle = \epsilon \langle \mathbf{E}(\mathbf{r},t)\cdot\mathbf{E}(\mathbf{r},t)\rangle , \tag{3.1}$$

where the time-average of the function $U(\mathbf{r},t)$ is defined as

$$\langle U(\mathbf{r},t)\rangle = \frac{1}{T}\int_0^T U(\mathbf{r},t)dt. \tag{3.2}$$

The averaging is performed over a time interval T.

If two fields $\mathbf{E}_1(\mathbf{r},t)$ and $\mathbf{E}_2(\mathbf{r},t)$ are present, the time-averaged energy density is the result of the total field $\mathbf{E}(\mathbf{r},t)$ which is the superposition of $\mathbf{E}_1(\mathbf{r},t)$ and $\mathbf{E}_2(\mathbf{r},t)$:

$$\begin{aligned} I = I(\mathbf{r},t) &= \langle U(\mathbf{r},t)\rangle \\ &= \epsilon \langle (\mathbf{E}_1(\mathbf{r},t)+\mathbf{E}_2(\mathbf{r},t))\cdot(\mathbf{E}_1(\mathbf{r},t)+\mathbf{E}_2(\mathbf{r},t))\rangle \\ &= \epsilon \langle \mathbf{E}_1(\mathbf{r},t)\cdot\mathbf{E}_1(\mathbf{r},t)+\mathbf{E}_2(\mathbf{r},t)\cdot\mathbf{E}_2(\mathbf{r},t)+2\mathbf{E}_1(\mathbf{r},t)\cdot\mathbf{E}_2(\mathbf{r},t)\rangle . \\ &= I_1+I_2+2I_{12} \end{aligned} \tag{3.3}$$

The first two terms in 3.3 are the time-averaged energy densities or intensities associated with each wave separately. The third term is a cross-term that gives rise to interference phenomena. The same holds when multiple waves interfere. In that case, multiple cross terms will add to the sum of the individual intensities. Before we start a more detailed discussion of the cross-term, we first investigate the aspect of polarization within the context of interference.

3.2.2 Polarization

The vectorial nature of the cross-term is of great importance. The cross-term is a dot-product of two vectors and is zero only when the two vectors have orthogonal directions. For the case of plane waves, this implies that interference occurs always except when the waves have orthogonal polarization states. This property will help us to simplify the mathematical treatment of interference problems. Consider for example a plane wave, propagating in the z-direction:

$$\mathbf{E}(z,t) = \Re[\mathbf{A}e^{+j(\omega t-kz)}]. \tag{3.4}$$

$\mathbf{A}$ is a complex vector that lies in the xy-plane and is defined as

$$\mathbf{A} = A_x e^{j\phi_x(t)}\mathbf{e}_x + A_y e^{j\phi_y(t)}\mathbf{e}_y, \tag{3.5}$$

or equivalently, the x and y-components of the field vector $\mathbf{E}$ are described by:

$$\begin{array}{ll} E_x & = A_x \cos(\omega t - kz + \phi_x(t)) \\ E_y & = A_y \cos(\omega t - kz + \phi_y(t)), \end{array} \tag{3.6}$$

with A_x and A_y positive numbers.

As we know from chapter 2, the phase difference $\delta(t) = \phi_x(t) - \phi_y(t)$ determines the polarization state (linear, circular, elliptic, partial, unpolarized) of the plane wave. Each of the components E_x and E_y can be regarded as linearly polarized plane waves with orthogonal directions of oscillation. And given that those waves will never interfere, they can be treated separately. In conclusion, any interference problem of waves with arbitrary polarization states can be treated by first decomposing the waves into orthogonally polarized wave components and then investigate interference between the wave components of colinear polarization (thus having parallel field vectors). This is what we will do in the next section.

3.2.3 Interference of two plane waves

Consider two monochromatic plane waves with frequencies ω_1 and ω_2, wave vectors $\mathbf{k}_1$ and $\mathbf{k}_2$ and with colinear polarisation.

$$\begin{array}{ll} \mathbf{E}_1(\mathbf{r},t) & = \mathbf{A}_1 \cos(\omega_1 t - \mathbf{k}_1 \cdot \mathbf{r} + \phi_1(t)) \\ \mathbf{E}_2(\mathbf{r},t) & = \mathbf{A}_2 \cos(\omega_2 t - \mathbf{k}_2 \cdot \mathbf{r} + \phi_2(t)). \end{array} \tag{3.7}$$

Since both fields have colinear polarization, we may discard the vectorial nature of the fields and write for each of the fields ($n = 1, 2$):

$$E_n(\mathbf{r},t) = A_n \cos(\omega_n t - \mathbf{k}_n \cdot \mathbf{r} + \phi_n(t)). \tag{3.8}$$

The intensity associated with this plane wave can be easily calculated and yields:

$$\begin{array}{ll} I & = \epsilon \left\langle |A_n|^2 \cos^2\left(\omega_n t - \mathbf{k}_n \cdot \mathbf{r} + \phi_n(t)\right) \right\rangle \\ & = \frac{\epsilon}{2} |A_n|^2 \left\langle \{1 + \cos\left(2\left[\omega_n t - \mathbf{k}_n \cdot \mathbf{r} + \phi(t)\right]\right)\} \right\rangle \\ & = \frac{\epsilon}{2} |A_n|^2 \end{array} \tag{3.9}$$

One will get the same result when working with the complex vector notation: the intensity is directly related to the modulus squared of the complex amplitude.

We will now examine the effect of the cross term in detail. As we will be calculating energy-densities involving products of field vectors, it is safe practice to use the real field vector notation.

The intensity or time-averaged energy density thus becomes:

$$
\begin{aligned}
I &= \epsilon\left\langle E_1^2\right\rangle + \epsilon\left\langle E_2^2\right\rangle + 2\epsilon\left\langle E_1 E_2\right\rangle \\
&= \epsilon\frac{A_1^2}{2} + \epsilon\frac{A_2^2}{2} + \epsilon A_1 A_2\left[\left\langle\cos\left[(\omega_1-\omega_2)t - (\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) - \phi_2(t)\right]\right\rangle\right] + \\
&\quad \epsilon A_1 A_2\left[\left\langle\cos\left[(\omega_1+\omega_2)t - (\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) + \phi_2(t)\right]\right\rangle\right] \\
&= I_1 + I_2 + 2\sqrt{I_1 I_2}\left[\left\langle\cos\left[(\omega_1-\omega_2)t - (\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) - \phi_2(t)\right]\right\rangle\right] + \\
&\quad 2\sqrt{I_1 I_2}\left[\left\langle\cos\left[(\omega_1+\omega_2)t - (\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) + \phi_2(t)\right]\right\rangle\right],
\end{aligned}
\tag{3.10}
$$

where I_1 and I_2 are the intensities of the individual waves.

The sum frequency term $\langle\cos\left[(\omega_1+\omega_2)t - (\mathbf{k}_1+\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) + \phi_2(t)\right]\rangle$ will be averaged out by every detector as $T >> \frac{1}{\omega_1+\omega_2}$, so we find:

$$
I = I_1 + I_2 + 2\sqrt{I_1 I_2}\left\langle\cos\left[(\omega_1-\omega_2)t - (\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r} + \phi_1(t) - \phi_2(t)\right]\right\rangle. \tag{3.11}
$$

If $\omega_1 \neq \omega_2$, the cross term will be non-zero only when $\frac{1}{\omega_1-\omega_2} >> T$, i.e. if the two frequencies are only slightly different. The result is a time-dependent beating of $\langle U\rangle$. However, when $\frac{1}{\omega_1-\omega_2} << T$ the last term in 3.11 will be zero.

If $\omega_1 = \omega_2$, then the phase difference between the two fields is $\mathbf{K}\cdot\mathbf{r} + \delta$ where $\mathbf{K} = \mathbf{k}_2 - \mathbf{k}_1$ and $\delta(t) = \phi_1(t) - \phi_2(t)$. If $\delta(t)$ is constant over a long enough period of time ($>> T$), then the two waves are mutually coherent and a stationary interference pattern will be observed in space. For $A_1 = A_2 = A_0$ and thus $I_1 = I_2 = I_0$ the interference pattern is described by:

$$
\begin{aligned}
I &= 2\epsilon\tfrac{A_0^2}{2}\left[1 + \langle\cos(\mathbf{K}\cdot\mathbf{r} + \delta(t))\rangle\right] \\
&= 4I_0\cos^2(\tfrac{\mathbf{K}\cdot\mathbf{r}+\delta(t)}{2}) \\
&= 4I_0\cos^2(\tfrac{\Phi}{2}),
\end{aligned}
\tag{3.12}
$$

where $\Phi = \mathbf{K}\cdot\mathbf{r} + \delta(t)$ and I_0 is the intensity of the individual waves. The intensity thus varies periodically in space in the direction of $\mathbf{K}$ and varies between 0 and 4 times the intensity of a single wave. The spatial period of the interference pattern is

$$
\Delta = \frac{2\pi}{|\mathbf{K}|} = \frac{\lambda_0}{2sin(\theta/2)}, \tag{3.13}
$$

where θ is the angle between the two wave vectors and $\lambda_0 = 2\pi c/\omega$ is the vacuum wavelength of light.

Equation (3.12) indicates a strong dependence of the total intensity on the phase difference. This dependence is plotted in Figure 3.3. Also when the two waves originate from the same source, but propagate along different paths with different optical path lengths before they come together, as illustrated in figure 3.2, the detected intensity will depend on the value of the phase $\Phi = \mathbf{K}\cdot\mathbf{r} + \delta(t)$.

Assuming a path length difference (also called delay) of d, the phase equals

$$
\Phi = \frac{2\pi d}{\lambda} = \frac{2\pi n d}{\lambda_0} \tag{3.14}
$$

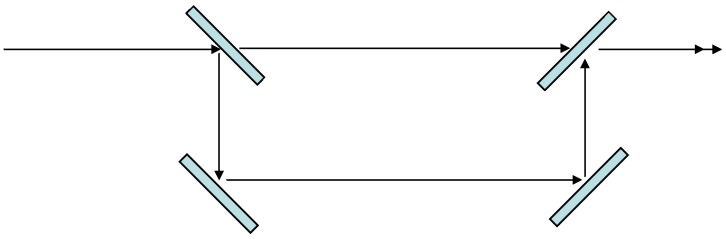


Figure 3.2: Interference of two waves originating from the same source, after having travelled a different path length

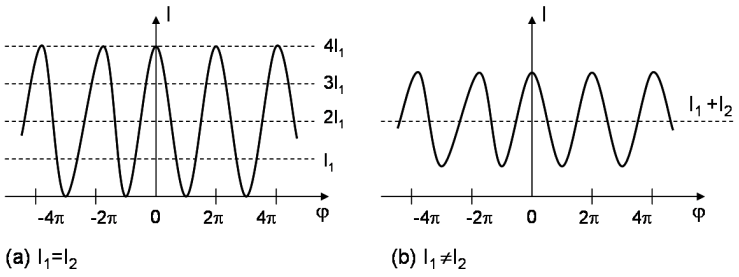


Figure 3.3: Interference of two waves, (a) $I_1 = I_2 = I_0$, (b) $I_1 \neq I_2$.

An interferometer uses the strong phase-dependence of the intensity 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. A tiny change in optical path length causes a change in Φ and thus a change in detected intensity.

3.2.4 Interference of multiple plane waves

The optical properties of thin films are mostly due to interference effects of more than two waves. When multiple waves come together, the total field is given by

$$\mathbf{E}(\mathbf{r}, t) = \sum_{n=1}^{N} \mathbf{E}_n(\mathbf{r}, t) \tag{3.15}$$

and the intensity of the total field can be calculated by definition (3.1)

Considering fields of colinear polarization,

$$\mathbf{E}_n(\mathbf{r}, t) = \mathbf{A}_n \cos(\omega_n t - \mathbf{k}_n \cdot \mathbf{r} + \phi_n(t)), \tag{3.16}$$

we may discard the vectorial nature of the fields to calculate the intensity of the total field:

$$I = \sum_{n=1}^{N} I_n + \sum_{n=1}^{N} \sum_{m=1, m\neq n}^{N} \sqrt{(I_n)}\sqrt{(I_m)} \left\langle \cos\left[(\omega_n - \omega_m)t - (\mathbf{k}_n - \mathbf{k}_m)\cdot\mathbf{r} + \phi_n(t) - \phi_m(t)\right]\right\rangle, \quad (3.17)$$

where $I_n = \epsilon\frac{|\mathbf{A}_n|^2}{2}$. The first term is the sum of the individual intensities and the second term determines the interference pattern.

In this section, we will briefly examine two special cases that are regularly encountered in thin film applications. To simplify the notations and calculations, we will from now on work with the complex notation of the field amplitudes.

Interference of M plane waves with equal amplitudes and constant phase difference

Consider M plane waves $E_m = A_m e^{j\omega t}$ with complex amplitudes defined by

$$A_m = \sqrt{I_0}e^{[j(m-1)\delta]}, \quad m = 1, 2, \ldots, M. \quad (3.18)$$

The waves have equal intensity I_0 and a constant phase difference δ. We define $h = e^{j\delta}$, so that $A_m = {I_0}^{1/2}h^{m-1}$. The complex amplitude of the total wave becomes

$$\begin{aligned} A &= \sqrt{I_0}\left(1 + h + h^2 + \ldots + h^{M-1}\right) \\ &= \sqrt{I_0}\frac{1-h^M}{1-h} \\ &= \sqrt{I_0}\frac{1-e^{jM\delta}}{1-e^{j\delta}} \end{aligned} \quad (3.19)$$

and the intensity is

$$I = |A|^2 = I_0\left|\frac{e^{-jM\delta/2} - e^{jM\delta/2}}{e^{-j\delta/2} - e^{j\delta/2}}\right|^2 \quad (3.20)$$

so that

$$I(\delta) = I_0\frac{sin^2\left(M\delta/2\right)}{sin^2\left(\delta/2.\right)} \quad (3.21)$$

This function is plotted in Figure 3.4 for different values of M. For M=2, the same interference pattern is found as described above. For higher values of M, the interference pattern exhibits peaks. Indeed, for $\delta \to 0$, we find that

$$\lim_{\delta/2\to 0} I(\delta) = I_0 M^2, \quad (3.22)$$

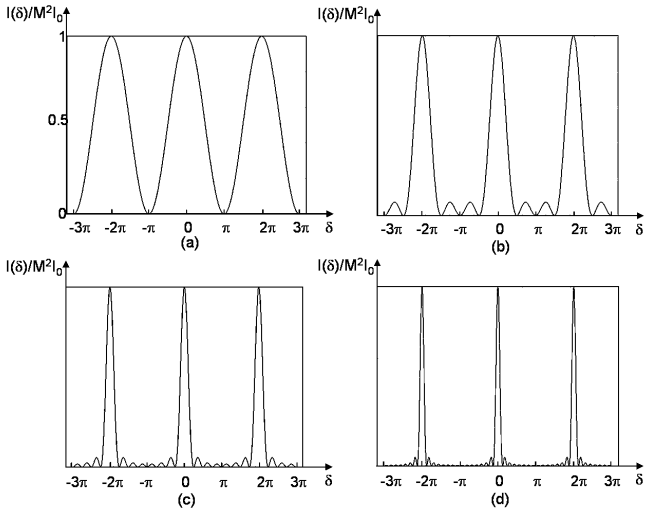


Figure 3.4: Interference of a finite number of plane waves with equal amplitudes and constant phase difference: intensity $\frac{I(\delta)}{M^2 I_0}$ in function of phase difference δ for (a) $M = 2$, (b) $M = 4$, (c) $M = 8$ and (d) $M = 16$.

indicating constructive interference of all the fields. This happens each time when $\delta = 2m\pi$. However, for certain values of δ, namely $\frac{2\pi}{M}, \frac{4\pi}{M}, ..., \frac{2(M-1)\pi}{M}$, $I(\delta)$ becomes 0. That is why the interference patterns exhibit $M - 1$ minima and $M - 2$ secondary maxima.

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 exhibits the behavior described above, in function of the angle.

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

Let us now consider an infinite number of plane waves with exponentially decreasing amplitude coming together. The complex amplitudes are given by:

$$A_1 = \sqrt{I_0}, \quad A_2 = hA_1, \quad A_3 = hA_2 = h^2 A_1, \quad \ldots \tag{3.23}$$

with now $h = |h| e^{j\delta}$ and $|h| < 1$. I_0 is again the intensity of the initial wave. The superposition of all these waves has complex amplitude

$$\begin{aligned} A &= A_1 + A_2 + A_3 + \ldots \\ &= \sqrt{I_0}(1 + h + h^2 + \ldots) \\ &= \frac{\sqrt{I_0}}{1-h} \\ &= \frac{\sqrt{I_0}}{1 - |h| e^{j\delta}}, \end{aligned} \tag{3.24}$$

which can be rewritten as

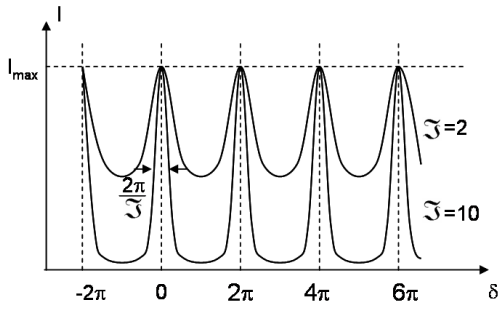


Figure 3.5: Interference of an infinite number of waves with progressively declining amplitude and equal phase difference: intensity I in function of phase difference δ.

$$I = \frac{I_0}{(1-|h|)^2 + 4\,|h|\,sin^2(\delta/2)}. \tag{3.25}$$

or

$$I = \frac{I_{max}}{1 + \left(\frac{2\mathfrak{F}}{\pi}\right)^2 sin^2\left(\frac{\delta}{2}\right)} \tag{3.26}$$

with

$$I_{max} = \frac{I_0}{(1-|h|)^2} \tag{3.27}$$

and

$$\mathfrak{F} = \frac{\pi\,|h|^{\frac{1}{2}}}{1-|h|} \tag{3.28}$$

a parameter called finesse.

As illustrated in figure 3.5 the intensity is a periodic function of δ with period 2π. It reaches the maximum I_{max} for $\delta = 2\pi q$, with q an integer. When the finesse $\mathfrak{F}$ is large (so r 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 $r = 0$. An important parameter associated with this interference pattern, is the so-called *Full Width at Half Maximum* (FWHM), equal to

$$\Delta\delta = \frac{2\pi}{\mathfrak{F}}. \tag{3.29}$$

The finesse $\mathfrak{F}$ is the ratio between the period 2π of the peaks and the FWHM of the transmission peaks.

This example is especially relevant in practice, in particular for the Fabry-Perot interferometer. We will come back to this structure in section 3.4.1.

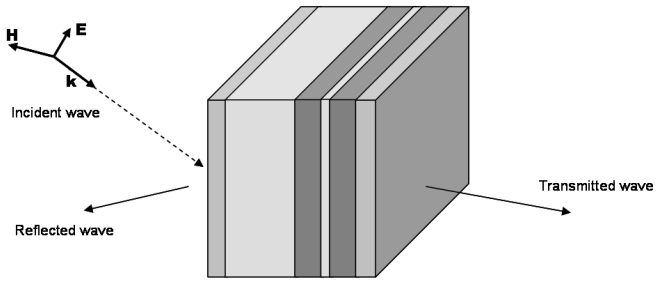


Figure 3.6: Reflection and transmission at a multilayer stack.

3.3 Transfer Matrix Formulation for Multilayer Systems

We will now investigate the general problem of wave propagation in and through a multilayer structure. Consider the dielectric structure depicted in Figure 3.6. The multilayer structure can be regarded as an optical system with one input and one output port. Thus, a transfer matrix **T** can be assigned to the system which relates the incident and reflected waves at the input port with the incident and reflected waves at the output port. If light is incident from the left, the reflected and transmitted field can thus be calculated once the transfer matrix **T** is known. In this section we will build the matrix **T** by regarding the multilayer stack as a cascaded system of interfaces and layers, each with their own transfer matrix. We get the complete transfer matrix **T** by multiplying the individual transfer matrices.

The layered medium consists of a stack of layers with thickness d_i and refractive indices n_i, separated by interface planes. A plane wave with wave vector **k** is incident on the stack, as depicted in Figure 3.6. It is convenient to choose the following coordinate system. Let the normal to the interface planes be the x-axis, then the interface planes are parallel to the (y,z)-plane and the multilayer stack (n_i, d_i) is defined as follows:

$$n(x) = \begin{array}{ll} n_0, & x < x_0, \\ n_1, & x_0 < x < x_1, \\ n_2, & x_1 < x < x_2, \\ \vdots & \\ n_N, & x_{N-1} < x, \end{array} \tag{3.30}$$

where n_i is the complex refractive index of layer i and where x_i marks the position of the planar interface between the layers i and $(i+1)$. The layers have a thickness d_i equal to $x_{i+1} - x_i$.

We now define the z-axis. The normal to the interface planes and the wave vector of the incident plane wave define a plane, which is called the incident plane. Let the z-axis be the coordinate axis in this plane orthogonal to the x-axis. As a result, the y-axis is defined and the wave vector lies entirely in the (x,z)-plane: its y-component is zero, $\mathbf{k}_y = 0$. The multilayer structure in this coordinate system is depicted in Figure 3.7 (a).

As we noted before, it is sufficient to treat the problem for two orthogonal polarizations. The easiest way is to solve the problem once for the TE-polarization (s-wave) with the electric field parallel to the interfaces of the stack and once for the TM-polarization (p-wave) with the magnetic

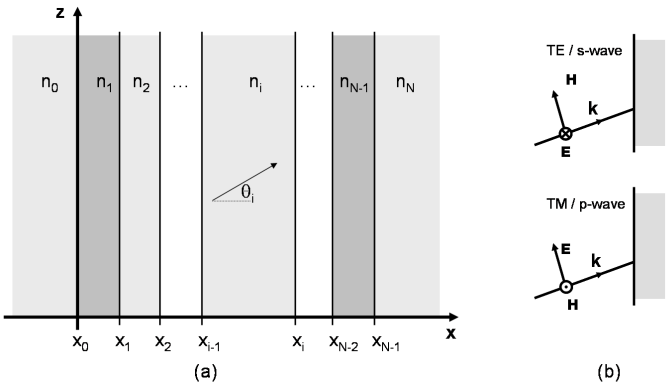


Figure 3.7: A multilayer stack of dielectric media.

field parallel to the interfaces of the stack. The orientation of both polarizations is illustrated in Figure 3.7 (b). In the following calculations, the two problems - one for each polarization - are discriminated by the polarization dependent reflection and transmission coefficients. When we consider only one polarization (TE or TM), the electric field is fully described by its amplitude $E(x, y, z)$. In each layer, all contributions to the forward propagating wave (towards increasing x) have the same direction and thus form one plane wave. The same holds for the backward propagating wave (towards decreasing x). The total field in layer i can thus be described as

$$\begin{aligned} E(x,y,z) &= A_F e^{-j(k_{x,i}x+k_z z)} + A_B e^{-j(-k_{x,i}x+k_z z)} \\ &= A_F e^{-jk_{x,i}x} e^{-jk_z z} + A_B e^{+jk_{x,i}x} e^{-jk_z z} \\ &= E_F(x) e^{-jk_z z} + E_B(x) e^{-jk_z z} \end{aligned} \tag{3.31}$$

where $k_{x,i}$ is the x-component of the wave vector $\mathbf{k}_i$ in layer i:

$$k_{x,i} = \left[(n_i k_0)^2 - k_z^2 \right]^{1/2}, i = 0, 1, 2, ..., N, \tag{3.32}$$

where $k_0 = (\omega/c)$. Due to boundary conditions that relate the field amplitudes of the incident, reflected and transmitted wave at the interfaces, k_z remains constant. When working with lossless media and in the absence of total internal reflection, $k_{x,i}$ can be related to the ray angle θ_i (see Figure 3.8) in the following way:

$$k_{x,i} = n_i k_0 \cos\theta_i. \tag{3.33}$$

However, it should be noted that $k_{x,i}$ is a complex number and that its imaginary part can be nonzero. This will be the case when $\Im(n_i)$ is nonzero or when $k_z > n_i k_0$. The former corresponds to wave propagation in a lossy medium, the latter corresponds to incident angles that are bigger than the critical angle for total internal reflection. When total internal reflection occurs, the electric field vector decreases exponentially in the x-direction and the attenuation occurs within a distance of q^{-1} where (if we choose $+j$ as the solution for $\sqrt{-1}$):

$$q_i = \left[k_z^2 - n_i k_0^2 \right]^{1/2}, i = 0, 1, 2, ..., N, \tag{3.34}$$

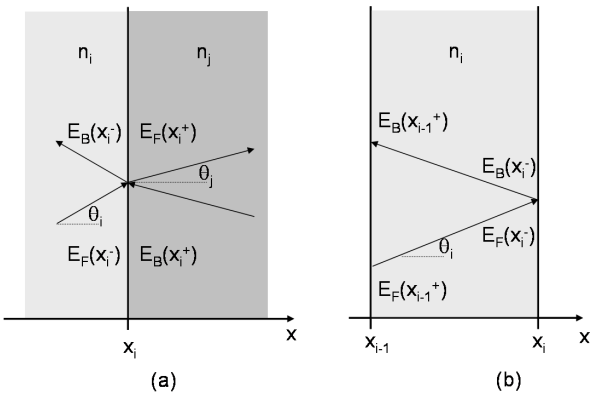


Figure 3.8: (a) Reflection and transmission at an interface. (b) Translation through a layer with thickness d_i and refractive index n_i.

In that case, the plane wave 3.31 is an evanescent wave propagating parallel to the interface surfaces in the z-direction.

For a given z, the complete transfer matrix **T** for wave propagation relates the complex amplitudes E_F and E_B just before and just behind the first and the last interface of the multilayer stack.

$$\begin{bmatrix} E_F(x_0^-) \\ E_B(x_0^-) \end{bmatrix} = \mathbf{T}_{0N} \begin{bmatrix} E_F(x_{N-1}^+) \\ E_B(x_{N-1}^+) \end{bmatrix} = \begin{bmatrix} T_{11}^{0N} & T_{12}^{0N} \\ T_{21}^{0N} & T_{22}^{0N} \end{bmatrix} \begin{bmatrix} E_F(x_{N-1}^+) \\ E_B(x_{N-1}^+) \end{bmatrix}. \tag{3.35}$$

We first build the individual transfer matrices for wave propagation through an interface and for wave propagation through a layer and will then calculate the complete transfer matrix by multiplying all individual ones. We start with the transfer matrix for wave propagation through an interface. Consider the interface between layer i and layer j (Figure 3.8 (a)). For a given z, the amplitudes of the forward and backward propagating waves just before and just behind the interface are related in the following way:

$$\begin{bmatrix} E_F(x_i^+) \\ E_B(x_i^-) \end{bmatrix} = \begin{bmatrix} t_{ij} & r_{ji} \\ r_{ij} & t_{ji} \end{bmatrix} \begin{bmatrix} E_F(x_i^-) \\ E_B(x_i^+) \end{bmatrix} \tag{3.36}$$

where we use the complex reflection and transmission coefficients (Fresnel coefficients) for the situation where only one wave is incident on the interface. This is actually the scattering matrix description of the interface between two layers.

In the case of TE-polarization, the Fresnel coefficients are given by:

$$r_{ij} = \frac{E_B(x_i^-)}{E_F(x_i^-)} = \frac{k_x^i - k_x^j}{k_x^i + k_x^j} \tag{3.37}$$

$$t_{ij} = \frac{E_F(x_i^+)}{E_F(x_i^-)} = 1 + r_{ij} = \frac{2k_x^i}{k_x^i + k_x^j} \tag{3.38}$$

and in the case of TM-polarization:

$$r_{ij} = \frac{E_B(x_i^-)}{E_F(x_i^-)} = \frac{n_i^2 k_x^j - n_j^2 k_x^i}{n_i^2 k_x^j + n_j^2 k_x^i} \tag{3.39}$$

$$t_{ij} = \frac{E_F(x_i^+)}{E_F(x_i^-)} = \frac{n_i}{n_j}(1 + r_{ij}) \tag{3.40}$$

or their equivalent as a function of the refractive indices n_i, n_j and ray angles θ_i, θ_j, for TE-polarization:

$$r_{ij} = \frac{E_B(x_i^-)}{E_F(x_i^-)} = \frac{n_i \cos\theta_i - n_j \cos\theta_j}{n_i \cos\theta_i + n_j \cos\theta_j} \tag{3.41}$$

$$t_{ij} = \frac{E_F(x_i^+)}{E_F(x_i^-)} = 1 + r_{ij} = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_j \cos\theta_j} \tag{3.42}$$

and for TM-polarization:

$$r_{ij} = \frac{E_B(x_i^-)}{E_F(x_i^-)} = \frac{n_j \cos\theta_i - n_i \cos\theta_j}{n_j \cos\theta_i + n_i \cos\theta_j} \tag{3.43}$$

$$t_{ij} = \frac{E_F(x_i^+)}{E_F(x_i^-)} = \frac{n_i}{n_j}(1 + r_{ij}) = \frac{2n_i \cos\theta_i}{n_j \cos\theta_i + n_i \cos\theta_j} \tag{3.44}$$

For perpendicular incidence, there is no difference between the TE and TM case. Equation (3.37) differs however from equation (3.39) in this case. 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 3.9 depicts the definition of the unit vectors for the E and H-field for the incident, reflected and refracted wave.

Equations (3.36) can be rewritten in such a way that a relationship is described between the forward and backward propagating wave in layer i and the forward and backward propagating wave in layer j:

$$\begin{bmatrix} E_F(x_i^-) \\ E_B(x_i^-) \end{bmatrix} = \begin{bmatrix} \frac{1}{t_{ij}} & -\frac{r_{ji}}{t_{ij}} \\ \frac{r_{ij}}{t_{ij}} & t_{ji} - \frac{r_{ij}}{t_{ij}} r_{ji} \end{bmatrix} \begin{bmatrix} E_F(x_i^+) \\ E_B(x_i^+) \end{bmatrix}. \tag{3.45}$$

We make use of the symmetry relations of the Fresnel coefficients

$$r_{ij} = -r_{ji} \tag{3.46}$$

$$t_{ij} t_{ji} - r_{ij} r_{ji} = 1 \tag{3.47}$$

to simplify this expression and get:

$$\begin{bmatrix} E_F(x_i^-) \\ E_B(x_i^-) \end{bmatrix} = \frac{1}{t_{ij}} \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix} \begin{bmatrix} E_F(x_i^+) \\ E_B(x_i^+) \end{bmatrix}. \tag{3.48}$$

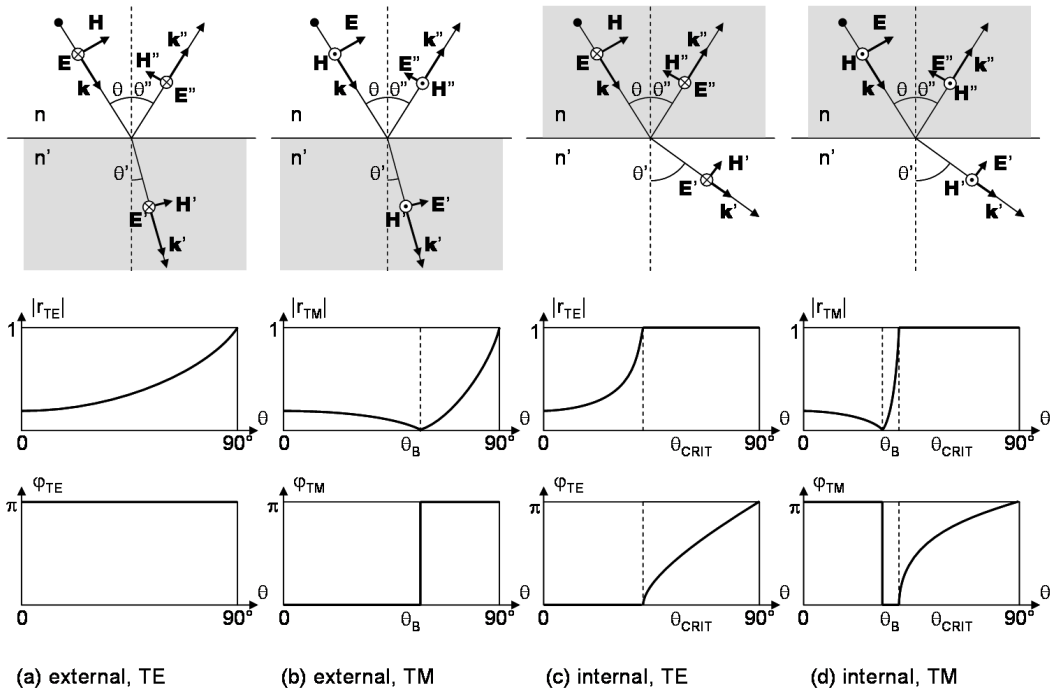


Figure 3.9: Magnitude and phase of the reflection coefficient in function of incidence angle for (a) external reflection ($n_j/n_i = 1.5$) and TE polarization, (b) external reflection ($n_j/n_i = 1.5$) and TM polarization, (c) internal reflection ($n_i/n_j = 1.5$) and TE polarization and (d) internal reflection ($n_i/n_j = 1.5$) and TM polarization.

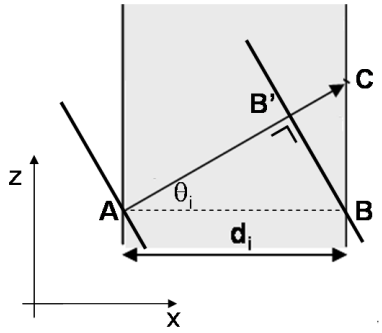


Figure 3.10: Inspection of Equation (3.52). The phase change Φ corresponds to the path length $|AB'|$ and not to the path length $|AC|$.

Thus, the transfer matrix $\mathbf{T}_{ij}$ for wave propagation through the interface between layers i and j reads:

$$\mathbf{T}_{ij} = \frac{1}{t_{ij}} \begin{bmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{bmatrix}. \tag{3.49}$$

The next step is to build the transfer matrix for wave propagation through a layer. Consider the layer i. (Figure 3.8 (b)). For a given z, the amplitudes of the forward and backward propagating waves just before and just behind the interfaces of the adjacent layers are related in the following way:

$$\begin{array}{ll} E_F(x_i^-) & = E_F(x_{i-1}^+)e^{-jk_{x,i}d_i} \\ E_B(x_{i-1}^+) & = E_B(x_i^-)e^{-jk_{x,i}d_i}, \end{array} \tag{3.50}$$

with d_i the thickness of layer i. This follows straightforwardly from Equations (3.31). Thus, the transfer matrix $\mathbf{T}_i$ for wave propagation through layer i reads:

$$\mathbf{T}_i = \begin{bmatrix} e^{j\Phi_i} & 0 \\ 0 & e^{-j\Phi_i} \end{bmatrix} \tag{3.51}$$

in which we defined $\Phi_i = k_{x,i}d_i$, which in general is a complex quantity. In a lossless medium and in the absence of total internal reflection, Φ_i is a real quantity:

$$\Phi_i = k_{x,i}d_i = \frac{2\pi}{\lambda_0} n_i d_i \cos\theta_i, \tag{3.52}$$

where $k_{x,i}$ was substituted using (3.33). Equation (3.52) might look surprising at first sight. Clearly, Φ_i represents a phase change but the phase change does not correspond to the path length $|AC|$. Careful inspection of Equation (3.52) shows that the phase change Φ corresponds to the path length $|AB'|$. The reason for this is that in the transfer matrix formalism, the transfer matrices always describe relationships between E_F and E_B at a constant z-level. Thus Φ corresponds to the distance between the two phase fronts through points A and B, equal to $|AB'|$.

The complete transfer matrix for the wave propagation through the layered medium can now be found by multiplication of all individual transfer matrices.

$$\begin{bmatrix} E_F(x_0^-) \\ E_B(x_0^-) \end{bmatrix} = \mathbf{T}_{0N} \begin{bmatrix} E_F(x_{N-1}^+) \\ E_B(x_{N-1}^+) \end{bmatrix} = \begin{bmatrix} T_{11}^{0N} & T_{12}^{0N} \\ T_{21}^{0N} & T_{22}^{0N} \end{bmatrix} \begin{bmatrix} E_F(x_{N-1}^+) \\ E_B(x_{N-1}^+) \end{bmatrix} \tag{3.53}$$

where

$$\mathbf{T}_{0N} = \mathbf{T}_{01}\mathbf{T}_1\mathbf{T}_{12}\mathbf{T}_2...\mathbf{T}_{(N-1)}\mathbf{T}_{(N-1)N}. \tag{3.54}$$

Equation (3.53) is known as the matrix formulation for wave propagation through multilayer systems. Now, this expression can be used for solving a variety of wave propagation problems. Let us first apply the transfer matrix method to our original problem: a plane wave is incident on a stack of layers. First calculate the individual transfer matrices according to the described method (equations (3.49) and (3.51)). Next, calculate the complete transfer matrix by multiplication using Equation (3.53). As there is no incident field from the right side, we have $E_B(x_{N-1}^+) = 0$. So, we are left with the following matrix formulation for wave propagation in the multilayer stack:

$$\begin{bmatrix} E_F(x_0^-) \\ E_B(x_0^-) \end{bmatrix} = \begin{bmatrix} T_{11}^{0N} & T_{12}^{0N} \\ T_{21}^{0N} & T_{22}^{0N} \end{bmatrix} \begin{bmatrix} E_F(x_{N-1}^+) \\ 0 \end{bmatrix}. \tag{3.55}$$

This leads to two equations and two unknowns, namely $E_B(x_0^-)$ and $E_F(x_{N-1}^+)$, which can now easily be solved for. The reflection and transmission coefficients follow from the ratios $E_B(x_0^-)/E_F(x_0^-)$ and $E_F(x_{N-1}^+)/E_F(x_0^-)$ respectively.

$$\begin{aligned} r &= \frac{E_B(x_0^-)}{E_F(x_0^-)} = \frac{T_{21}^{0N}}{T_{11}^{0N}} \\ t &= \frac{E_F(x_{N-1}^+)}{E_F(x_0^-)} = \frac{1}{T_{11}^{0N}} \end{aligned} \tag{3.56}$$

From the amplitude reflection and transmission coefficients, the power reflection and transmission coefficients can be easily calculated. The power density of a (forward or backward) plane wave in layer i is given by the real part of the Poynting vector:

$$\begin{aligned} \langle P(x) \rangle &= Re(\mathbf{S}(x)) \\ &= Re(\mathbf{E}(x) \times \mathbf{H}(x)) \\ &= Re(n) \frac{|\mathbf{E}(x)|^2}{2Z_0} \end{aligned} \tag{3.57}$$

with Z_0 the impedance of vacuum (= $\sqrt{\mu_0/\epsilon_0}$= 377Ω). This power expresses the power per unit area perpendicular to the propagation direction (which differs from layer to layer). In order to calculate the power reflection and transmission of the multilayer thin film, it is better to express the power density per unit of area parallel to the layers. This modified power density is given by:

$$\begin{aligned} \langle P(x_i)_{||} \rangle &= \cos\theta_i \langle P(x_i) \rangle \\ &= Re(n) \cos\theta_i \frac{|\mathbf{E}(x_i)|^2}{2Z_0} \end{aligned} \tag{3.58}$$

This allows to calculate the power reflection and transmission coefficients of the thin film:

$$\begin{aligned} R &= \frac{Re(n_0)\ \cos\theta_0\ |\mathbf{E}_B(x_0)|^2}{Re(n_0)\ \cos\theta_0\ |\mathbf{E}_F(x_0)|^2} = |r|^2 \\ T &= \frac{Re(n_N)\ \cos\theta_N\ |\mathbf{E}_F(x_N)|^2}{Re(n_0)\ \cos\theta_0\ |\mathbf{E}_F(x_0)|^2} = \frac{Re(n_N)\ \cos\theta_N}{Re(n_0)\ \cos\theta_0}\ |t|^2 . \end{aligned} \tag{3.59}$$

To conclude our discussion of the transfer matrix formalism and to show its broad applicability, let us consider a different problem. What if no light is incident, nor from the left nor from the right? This means: $E_F(x_0^-) = 0$ and $E_B(x_{N-1}^+) = 0$. In this case, we are left with the following matrix formulation for wave propagation in the multilayer stack:

$$\begin{bmatrix} 0 \\ E_B(x_0^-) \end{bmatrix} = \begin{bmatrix} T_{11}^{0N} & T_{12}^{0N} \\ T_{21}^{0N} & T_{22}^{0N} \end{bmatrix} \begin{bmatrix} E_F(x_{N-1}^+) \\ 0 \end{bmatrix}. \tag{3.60}$$

So, even if no light is incident, this system has nonzero solutions provided that the following condition is fulfilled:

$$T_{11}^{0N} = 0. \tag{3.61}$$

Let us briefly examine this condition. T_{11}^{0N} is a function of k_z, ω, indices n_i and thicknesses t_i of the layers. Given a multilayer structure (n_i, d_i) and a frequency ω, Equation (3.61) can thus be used to solve for the propagation constant k_z of all the confined modes supported by the multilayer structure. This example is an illustration of the application of the transfer matrix formalism to waveguide problems. We will come back to the transfer matrix method in a later chapter where it will be used to analyze slab waveguides.

3.4 Applications of thin films

In this section we discuss two important applications of thin film structures: the Fabry-Perot etalon and optical coatings.

3.4.1 Fabry-Perot etalon

The Fabry-Perot interferometer, or etalon, can be considered as the simplest type of optical resonator. Normally, such an instrument consists of two parallel dielectric mirrors separated at a distance l. Here, we consider a simple structure that consists of a plane-parallel plate of thickness l and refractive index n immersed in a medium of index n' (Figure 3.11). This is a simple multilayer stack and the transfer matrix formalism can be applied in a straightforward way. 3 transfer matrices need to be calculated and multiplied in order to get the complete transfer matrix of the 3-layer system. First, we calculate the transfer matrices associated with the two interfaces:

$$\mathbf{T}_{12} = \frac{1}{t_{12}} \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} = \frac{1}{t_{12}} \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}. \tag{3.62}$$

and

$$\mathbf{T}_{23} = \frac{1}{t_{23}} \begin{bmatrix} 1 & r_{23} \\ r_{23} & 1 \end{bmatrix} = \frac{1}{t_{23}} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}. \tag{3.63}$$

where $r = r_{12} = -r_{23}$. The transfer matrix associated to the layer between the two interfaces is the following:

$$\mathbf{T}_2 = \begin{bmatrix} e^{j\Phi} & 0 \\ 0 & e^{-j\Phi} \end{bmatrix} = e^{j\Phi} \begin{bmatrix} 1 & 0 \\ 0 & e^{-2j\Phi} \end{bmatrix} \tag{3.64}$$

where $\Phi = \frac{2\pi}{\lambda_0} nl \cos\theta$. Multiplying yields the complete transfer matrix:

$$\mathbf{T}_{13} = \mathbf{T} = \mathbf{T}_{12}\mathbf{T}_2\mathbf{T}_{23} = \frac{1}{t_{12}t_{23}} e^{j\Phi} \begin{bmatrix} 1 - r^2 e^{-2j\Phi} & -r(1 - e^{-2j\Phi}) \\ r(1 - e^{-2j\Phi}) & -r^2 + e^{-2j\Phi} \end{bmatrix}. \tag{3.65}$$

When such a Fabry-Perot etalon is illuminated by a light beam, the situation corresponds to one incident wave from the left of the structure. Thus, the reflection and the transmission of the structure are described by the formulas in Equation (3.55). Substitution and simplification yields for the reflection and transmission coefficients r_{FP} and t_{FP} according to 3.56:

$$\begin{aligned} r_{FP} &= \tfrac{T_{21}}{T_{11}} = \tfrac{r(1-e^{-2j\Phi})}{1-r^2e^{-2j\Phi}} \\ t_{FP} &= \tfrac{1}{T_{11}} = \tfrac{t_{12}t_{23}e^{-j\Phi}}{1-r^2e^{-2j\Phi}}, \end{aligned} \tag{3.66}$$

and for the power reflection and transmission coefficients R_{FP} and T_{FP} according to 3.59:

$$\begin{aligned} R_{FP} &= \left|\tfrac{T_{21}}{T_{11}}\right|^2 = \tfrac{4r^2\sin^2\Phi}{(1-r^2)^2+4r^2\sin^2\Phi} \\ T_{FP} &= \left|\tfrac{1}{T_{11}}\right|^2 = \tfrac{|t_{12}t_{23}|^2}{(1-r^2)^2+4r^2\sin^2\Phi}, \end{aligned} \tag{3.67}$$

describing the reflection and transmission of a Fabry-Perot etalon. Note that our basic model of the Fabry-Perot etalon does not contain any loss mechanisms, so conservation of energy requires $R_{FP} + T_{FP} = 1$, which is indeed the case. In what follows, we will work with the fraction of the intensity reflected R and the fraction of the intensity transmitted T at one interface and refer to it as the mirror's reflection and transmittance:

$$\begin{aligned} R &= r_{12}^2 = r_{23}^2 = r^2 \\ T &= t_{12}t_{23}. \end{aligned} \tag{3.68}$$

For the moment, we consider no losses, so $R+T = 1$. Let us now discuss the results of our calculations given by 3.67. First of all, the formula for the transmitted intensity (3.67) looks very similar to the formula (3.25) we found in section 3.2.4 describing the interference pattern of an infinite number of plane waves with progressively declining amplitude and equal phase difference. In fact, they are the same. This is not surprising at all. It is clear from Figure 3.11 that the transmission of a Fabry-Perot etalon is resulting from an infinite number of waves interfering with each other. If the incident intensity is taken as unity, the first transmitted contribution has intensity $|t_{12}t_{23}|$. The second contribution is decreased by a factor $|r_{12}r_{23}|$, the third contribution again, and so on. The phase difference between each contribution remains constant and equal to $\delta = 2\Phi = 2\frac{2\pi}{\lambda_0} nl\cos\theta$.

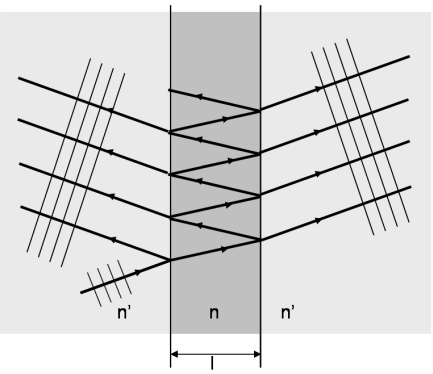


Figure 3.11: Reflection and transmission at a parallel plate structure such as the Fabry-Perot etalon.

Thus, from the viewpoint of interference, we end up with exactly the formula 3.25 for the tranmission through this structure. The transfer matrix formalism for wave propagation gives the same result, as it should be.

Let us now take a closer look at the transmission characteristics of a Fabry-Perot etalon. According to Equation (3.67), transmission is maximal and equal to unity when

$$\Phi = \frac{2\pi n}{\lambda} l \cos\theta = m\pi, m = integer, \tag{3.69}$$

where θ is the ray angle in the medium. Φ is called the resonator round trip phase. On the other hand, the transmission is minimal and equal to

$$T_{FP,min} = \frac{(1-R)^2}{(1-R)^2 + 4R} = \frac{1}{1 + \frac{4R}{(1-R)^2}} \tag{3.70}$$

when the resonator round trip phase equals

$$\Phi = \frac{2\pi n}{\lambda} l \cos\theta = (2m-1)\frac{\pi}{2}, m = integer. \tag{3.71}$$

By using $\lambda = c/\nu$, the condition for maximal transmission can also be written as

$$\nu_m = m\frac{c}{2nl\cos\theta}, m = integer, \tag{3.72}$$

where ν is the optical frequency. For a given l and n, Equation (3.72) defines the resonance frequencies of the Fabry-Perot etalon. At resonance, all transmitted contributions interfere constructively. Off resonance, the contributions do no longer interfere constructively. Two neighboring resonance frequencies are separated by the so-called free spectral range:

$$\Delta\nu = \nu_{m+1} - \nu_m = \frac{c}{2nl\cos\theta}, m = integer. \tag{3.73}$$

At this point, it is interesting to look back at Figure 3.5 depicting the intensity of an infinite number of interfering waves with declining amplitude and equal phase difference. It should be clear now that this figure also depicts the transmission of a Fabry-Perot cavity as a function of phase, see

Fig. 3.12. However, note that the Fabry-Perot peaks are at the positions $\Phi = m\pi$, which correspond to $\delta = 2m\pi$ in Fig. 3.5.

The finesse $\mathfrak{F}$ is directly related to the mirror's reflection R at each interface by

$$\mathfrak{F} = \frac{\pi\sqrt{R}}{1-R}. \tag{3.74}$$

Thus, the higher the mirror's reflection R at each interface, the higher is the finesse and the sharper are the transmission peaks:

$$\begin{array}{lll} R = 50\% & \rightarrow & \mathfrak{F} \cong 4 \\ R = 90\% & \rightarrow & \mathfrak{F} = 30 \\ R = 98\% & \rightarrow & \mathfrak{F} = 156 \end{array} \tag{3.75}$$

Another important dimensionless quantity in connection with resonances is the *quality factor* or Q-factor. It is defined as

$$Q = \frac{\omega_r}{\Delta\omega} \tag{3.76}$$

with ω_r the (angular) resonant frequency, and $\Delta\omega$ the FWHM-bandwidth (Full Width at Half Maximum). It is a measure of the sharpness of the resonance in the spectrum. We see in Fig. 3.12 e.g. that a Fabry-Perot with strongly reflecting mirrors gives sharper peaks with a larger Q, intuitively indicating that the resonance is 'better defined' or 'purer'.

Now we investigate the phase characteristics of the Fabry-Perot etalon. If we write the amplitude transmission coefficient 3.66 as

$$t = |t|\, e^{-j\angle t}, \tag{3.77}$$

and examine the phase shift $\angle t$ as a function of the cavity length l (or equivalently Φ), we note that a strong dispersion of the phases versus Φ exists at resonance. This is shown in Figure 3.12. Note that there is a link to the concept of *group delay*, defined as

$$\frac{\mathrm{d}\angle t}{\mathrm{d}\omega} \tag{3.78}$$

which measures the propagation time of a signal through the structure. We clearly note an increase of the group delay at resonance, which corresponds to the intuition that light then spends a longer time in the cavity. Remark that without the cavity structure, the phase would be a straight line ($\angle t = k_0 nd$) in Figure 3.12.

When designing a Fabry-Perot resonator with high finesse, one needs to ensure that $T_{FP,\min}$ given by Equation (3.70) is as small as possible. Therefore R needs to be close to 1. In practice, this is difficult because of the available materials. Indeed, the refractive index of optical materials is limited to about $n \approx 4$, which means for a Fabry-Perot in air a maximal reflection of only $R \approx 36\%$.

One solution to this problem is offered by applying metal coatings on the interfaces. Due to the presence of such a metal coating, an additional phase difference will occur upon reflection. So assume:

$$r = \sqrt{R}e^{+j\alpha}. \tag{3.79}$$

With this r one can calculate for the transmission

$$T_{FP} \quad = \left| \frac{T}{1 - Re^{-j2(\phi-\alpha)}} \right|^2 \tag{3.80}$$

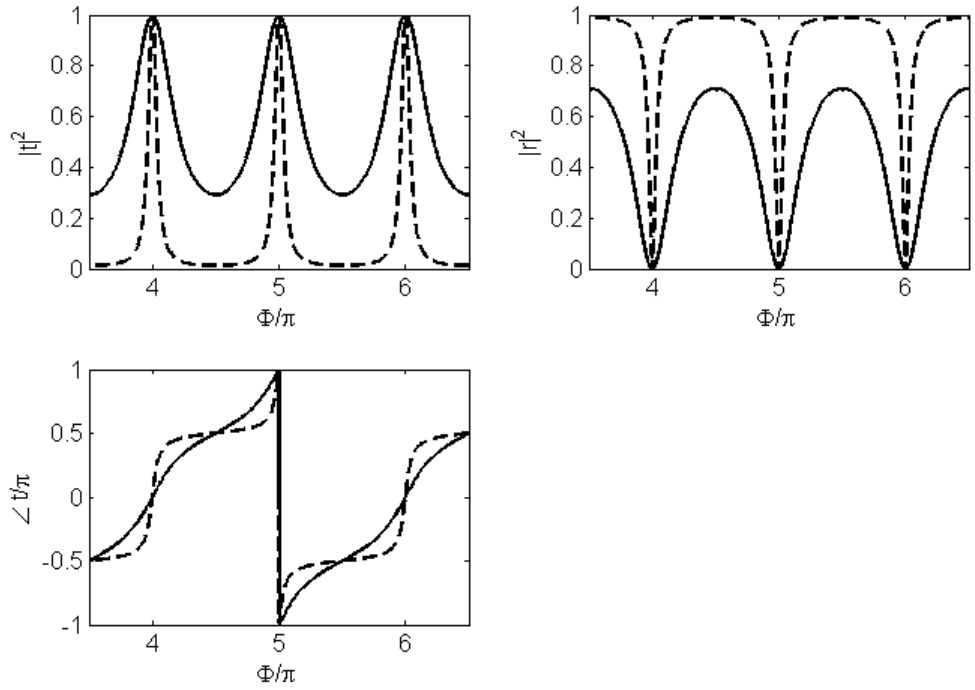


Figure 3.12: Power transmission/reflection and phase of transmission for a Fabry-Perot etalon as a function of cavity length l (or equivalently Φ). Solid line: $R = 30\%$, dashed line: $R = 80\%$.

where T is no longer equal to $1 - R$ but to $1 - R - A$ with A the intensity absorption in the metal film. As a result, the maximal transmission drops from unity to

$$T_{FP,\max} = \frac{T^2}{(1-R)^2} = \left[1 - \frac{A}{1-R}\right]^2. \tag{3.81}$$

The relation between the ratio $\frac{T_{FP,\max}}{T_{FP,\min}}$ and the mirror's reflection R is depicted in Figure 3.13.

To provide more insight into the operation of the Fabry-Perot cavities we plot the energy in and around the cavities in various situations, see Figure 3.14. We take air as surrounding medium of the cavity and change n in the cavity to get the wanted reflection R. This can be done using the formula

$$n = \frac{1+\sqrt{R}}{1-\sqrt{R}}. \tag{3.82}$$

It is clear we need higher n to get a higher reflection. We compare the energy-distribution for $R = 30\%$ ($n = 3.4$) and $R = 80\%$ ($n = 17.9$, which is a very irrealistic value), for $\Phi/\pi = 2$ (top), $\Phi/\pi = 2.5$ (middle) and $\Phi/\pi = 3$ (bottom). One notices that at resonance the energy in the cavities is enhanced, and the transmission is unity. Off resonance (middle graphs) we find a standing wave in reflection, and a small field in the cavity and also a small transmission.

We know the resonator round trip phase can be written as

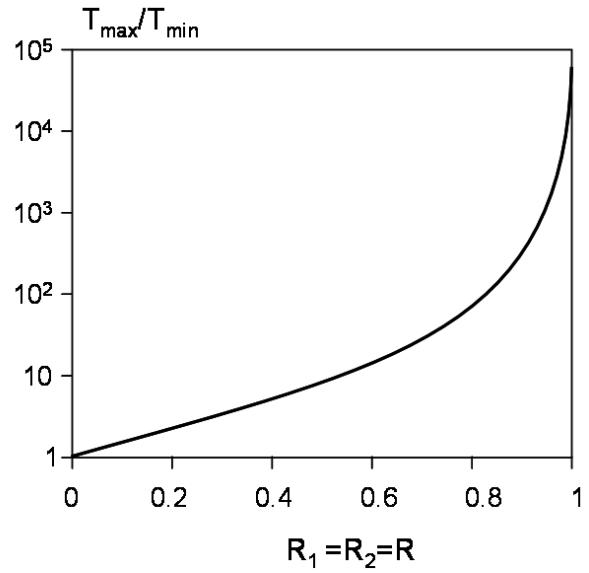


Figure 3.13: $\frac{T_{FP,\max}}{T_{FP,\min}}$ as a function of R.

$$\Phi = \frac{2\pi}{\lambda} nd. \tag{3.83}$$

So, if we use in both structures the same cavity length d and want to have the same Φ-value, we need to change λ, as the ratio $\frac{\lambda}{n}$ needs to stay constant. Thus, in the structure with $R = 80\%$, we need a longer wavelength than for the structure with $R = 30\%$. This can be seen in the change in the period of the standing wave pattern in air for the case $\Phi/\pi = 2.5$. However, the period of the standing wave pattern in the cavity stays constant for fixed d and Φ-values.

Most Fabry-Perot etalons are made of two identical mirrors. In fact, the symmetry of the mirror reflection is important to obtain high finesse. Any asymmetry in the reflection will lead to a decrease in either transmission or finesse. (It is left as an exercise to investigate the transmission properties of an asymmetric Fabry-Perot etalon in detail.)

3.4.2 Coatings

Multilayer stacks can be used as coatings that alter the optical properties of a substrate, for example increasing or decreasing reflection. Anti-reflective coatings and highly-reflective coatings are two examples that are widely used.

AR-coatings: quarter-wave layer

Minimal reflection or equivalently maximal transmission is desirable in many applications. One example is efficient coupling of light from a fibre into a waveguide. In designing an anti-reflective-coating one needs to ensure that the reflection at the front of the coating interferes destructively

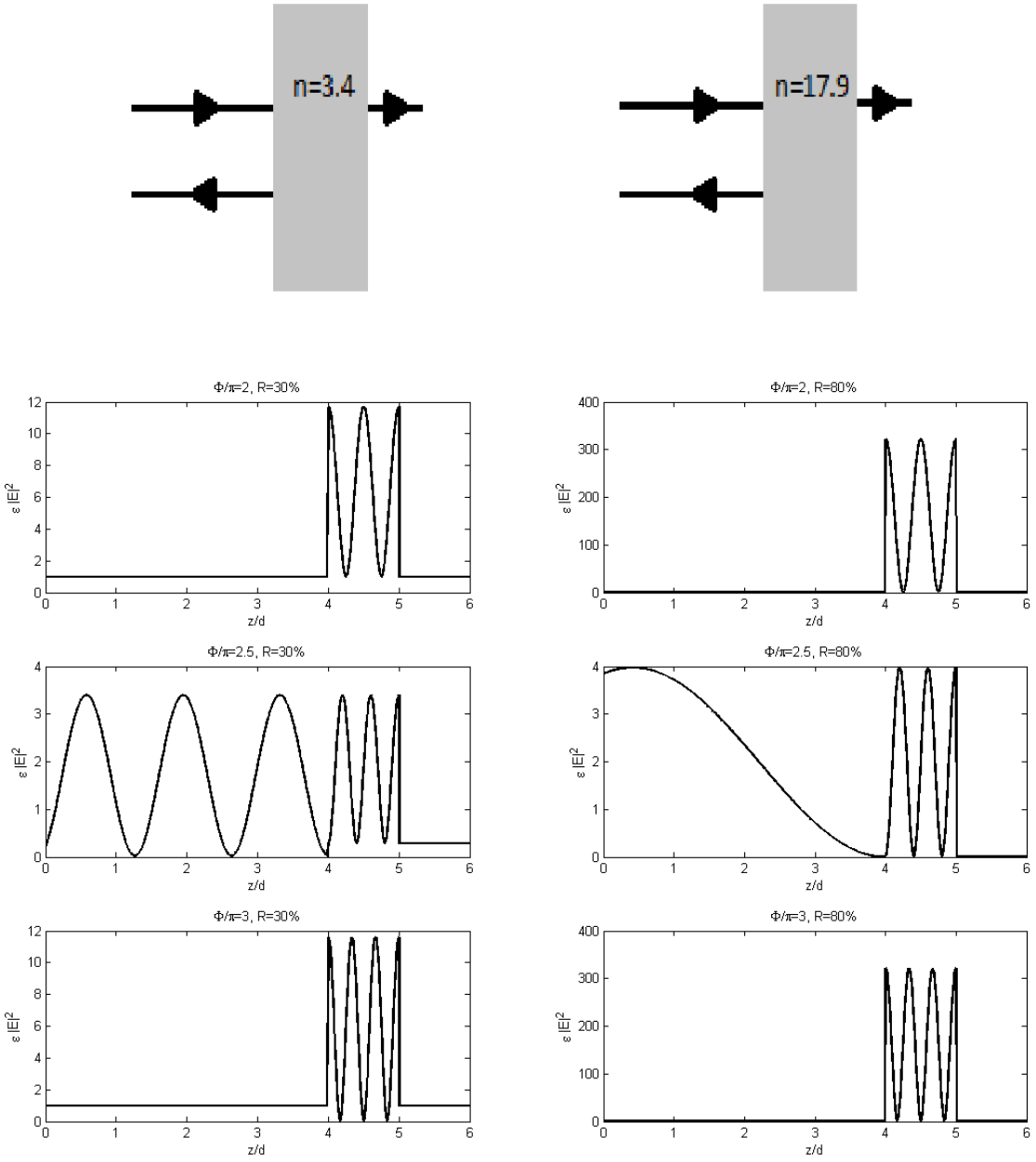


Figure 3.14: Energy ($\epsilon|E|^2$) in the Fabry-Perot cavities for $R = 30\%$ (left, with $n = 3.4$) and $R = 80\%$ (right, with $n = 17.9$ - an irrealistic value). The situation is shown for $\Phi/\pi = 2$ (top), $\Phi/\pi = 2.5$ (middle) and $\Phi/\pi = 3$ (bottom). The cavity is located between $z = 4d$ and $z = 5d$. Light is incident from the left with $|E|^2 = 1$.

with the reflection at the back of the coating. Assume a film thickness d. If $n_1 < n_2 < n_3$ with n_2 the refractive index of the coating, the two waves will interfere destructively if

$$d = \frac{1}{4}\frac{\lambda_0}{n_2}, \tag{3.84}$$

hence the name quarter-wave layer. The question remains which refractive index is needed for the coating. From the transfer matrix formalism for this 3-layer structure, one easily obtains

$$r = \frac{r_{12} + r_{23}e^{-2j\Phi}}{1 + r_{12}r_{23}e^{-2j\Phi}} = \frac{r_{12} - r_{23}}{1 - r_{12}r_{23}}, \tag{3.85}$$

where we used $\Phi = \frac{2\pi}{\lambda_0}n_2 d = \frac{\pi}{2}$. Given that $r_{ij} = \frac{n_i - n_j}{n_i + n_j}$, the reflection will be zero when

$$n_2 = \sqrt{n_1 n_3}. \tag{3.86}$$

In practice however, materials with such a refractive index may not exist. Nevertheless, using available materials with an index of refraction close to that given by Equation (3.86), a great reduction in reflection is obtained. We also notice that by using materials with an index of refraction between n_1 and n_3, a single-layer coating will always reduce the reflection, regardless of the layer thickness. Thus, the reflection of a coated surface is always lower than that of an uncoated one, provided that the index of refraction of the coating is between that of the two media. In some sense, the coating smooths dielectric discontinuity.

High-reflective coatings

High-reflective mirrors are desirable in many applications. These include high-finesse Fabry-Perot interferometers and low-loss laser resonators. Mirrors made of metallic films such as silver, aluminum or gold are generally of high reflection. For example, a silver mirror can achieve reflection approaching 99 % in the visible spectrum. Approximately 1 % of light energy penetrates the surface of the metal and gets absorbed in the bulk of the metal. These metallic mirrors cannot be used with high-power lasers because even a small fraction of absorption can cause severe heating problems. Thus there is a need to design high-reflection mirrors by using materials that have (almost) no absorption.

The dielectric layered structure that consists of alternating quarter-wave layers of two different materials is the simplest way to obtain high reflection. This is the so-called Bragg reflector. If for a certain wavelength λ_0, the thicknesses d_1, d_2 and the refractive indices n_1 and n_2 of the consecutive layers can be controlled so that:

$$n_1 d_1 = n_2 d_2 = \frac{\lambda_0}{4}, \tag{3.87}$$

then the reflected beams from the different interfaces will all interfere constructively, leading to a peak in the reflection spectrum for this wavelength.

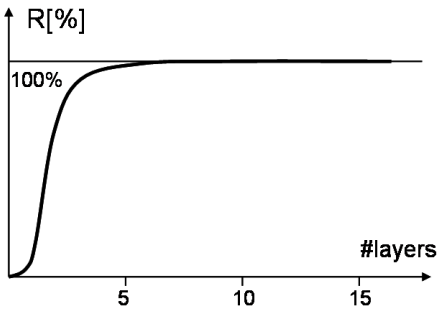


Figure 3.15: Reflection of a HR-coating for a He-Ne laser at wavelength λ_0 =633nm. n_1 =2.32 (ZnS), n_2=1.38 (MgF_2). A reflection of 98.9 % is achieved already after 13 layers.

Using the matrix method, the peak reflection $R_{HR,\max}$ can be calculated and is given by

$$R_{HR,\max} = \left(\frac{1 - \left(\frac{n_s}{n_a}\right)\left(\frac{n_2}{n_1}\right)^{2N}}{1 + \left(\frac{n_s}{n_a}\right)\left(\frac{n_2}{n_1}\right)^{2N}} \right)^2, \tag{3.88}$$

where n_s is the refractive index of the substrate, n_a that of air and N is the number of periods. $R_{HR,\max}$ converges to 1 as N increases. The convergence improves as the ratio $\frac{n_1}{n_2}$ becomes larger. This is illustrated in Figure 3.15 for the case of a high reflective coating for a He-Ne laser at wavelength $\lambda_0 = 633nm$ and using a quarter-waves stack of ZnS (n_1 =2.32) and MgF (n_2=1.38). However, Bragg reflectors can provide high reflection over any desired spectral regime of interest by properly tailoring the layer thicknesses. It can be proven that the bandwidth of these reflectors is given by

$$\frac{\Delta\lambda}{\lambda} = \frac{4}{\pi} \sin^{-1} \frac{|n_2 - n_1|}{n_2 + n_1}, \tag{3.89}$$

which can be approximately written as

$$\frac{\Delta\lambda}{\lambda} = \frac{2}{\pi} \frac{\Delta n}{n}, \tag{3.90}$$

provided that $n_1 \cong n_2 \cong n$. Thus, a high bandwidth is obtained by increasing the difference between the refractive indices of the layers. These trends are illustrated in Figure 3.16.

To obtain more insight in the Bragg reflector we plot a few field profiles in Figure 3.18, using the Bragg structure sketched in Figure 3.17. In the bandgap we expect a strong reflection, leading to an exponentially decaying field in the reflector, which is confirmed in Fig. 3.18(d). On the sides of the bandgap there are frequencies with reflection equal to zero (and thus total transmission), these correspond to Fabry-Perot like resonances, built up in between the ends of the Bragg reflector. Indeed, the field profile in Fig. 3.18(c) shows a bump in the profile in the reflector, and total transmission. In addition, the next reflection minimum is shown in Fig. 3.18(a), which indicates a higher order resonance, as it has two maxima in the field profile. In Fig. 3.18(b) the situation in between the previous resonances is plotted, at the point where the reflection has a (relative) maximum. We notice an intermediate situation, with one and a half lobes inside the cavity, leading to a significant reflection (indicated by the standing wave pattern).

Bragg reflectors can be made to reflect broad bands of light by stacking up several periodic layered media with different periods. In this case, each reflector acts as a band rejection filter for each

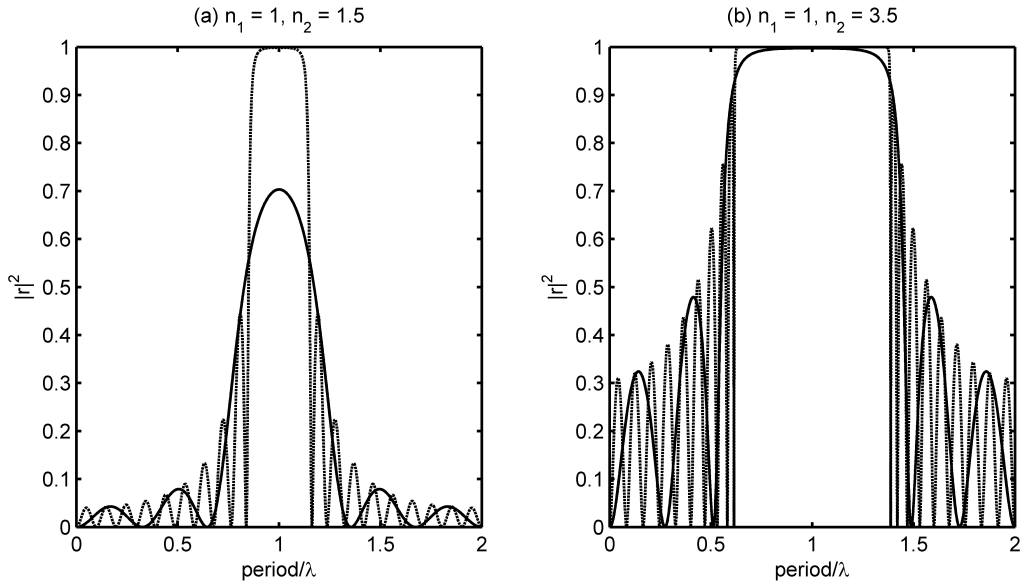


Figure 3.16: Reflection spectrum of Bragg coatings with different index contrasts Δn and number of layers N: (a) $\Delta n = 0.5$, (b) $\Delta n = 2.5$. Solid (dashed) curves correspond with $N = 3$ ($N = 10$), respectively. Note: $n = (n_1 + n_2)/2$ is the average refractive index and $\Lambda = d_1 + d_2$ is the period of the periodic layer structure

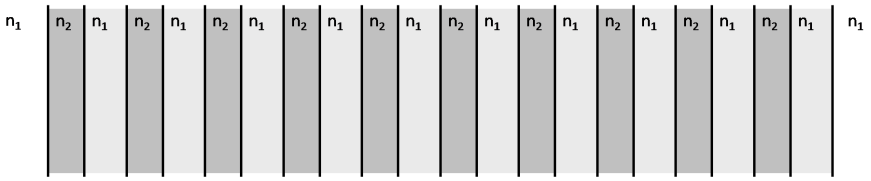


Figure 3.17: Schematic of a Bragg reflector mirror.

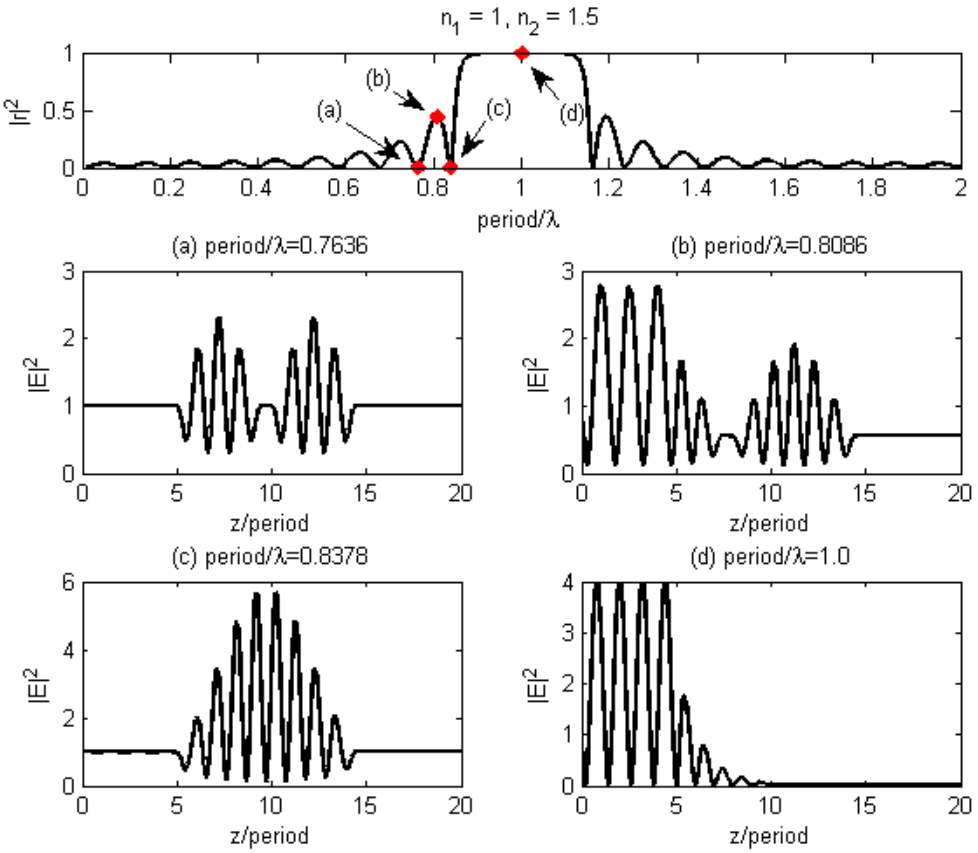


Figure 3.18: Reflection spectrum and field plots in and around the Bragg reflector from Figure 3.17. Light with $|E| = 1$ is incident from the left side. The reflector, located between $z/\text{period} = 5$ and 15, has $N = 10$ periods with $n_1 = 1, n_2 = 1.5$. The different $period/\lambda$-ratios used in the field plots are marked with big dots in the reflection spectrum.

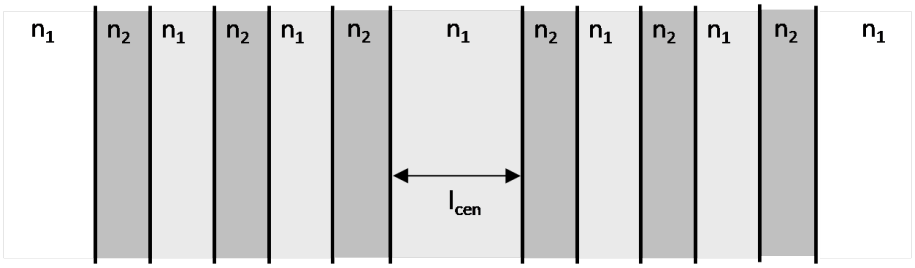


Figure 3.19: Schematic of a Fabry-Perot cavity with Bragg reflector mirrors.

wavelength. If the bandwidths are wide enough to have substantial overlap, the whole structure can reject a broad band of light.

Fabry-Perot with Bragg reflector

One can combine the concepts of the Fabry-Perot cavity with the Bragg reflector to create a very high quality resonance. This can be done by enlarging one of the layers in the middle of a Bragg reflector. In this way the sides of the Bragg reflector act as good mirrors (in the bandgap), and the 'defect' layer accommodates a Fabry-Perot type cavity mode. Thus, ideally the Bragg wavelength should correspond with the Fabry-Perot resonance wavelength.

A sketch of such a structure is shown in Fig. 3.19, with l_{cen} the thickness of the defect layer. Reflection spectra are depicted in Fig. 3.20. We see a sharp dip in reflection in the center of the bandgap in Fig. 3.20(a)(solid). The spectrum of the Bragg reflector without the defect is indicated with a dashed line. For this particular defect mode we chose $l_{cen} = \lambda_0/2n_1$, thus giving a fundamental Fabry-Perot resonance at λ_0, which is also the center of our Bragg bandgap (for $l_1 = \lambda_0/4n_1, l_2 = \lambda_0/4n_2$). A thicker defect layer can give rise to multiple Fabry-Perot type modes in one bandgap, this is illustrated in Fig 3.20(b).

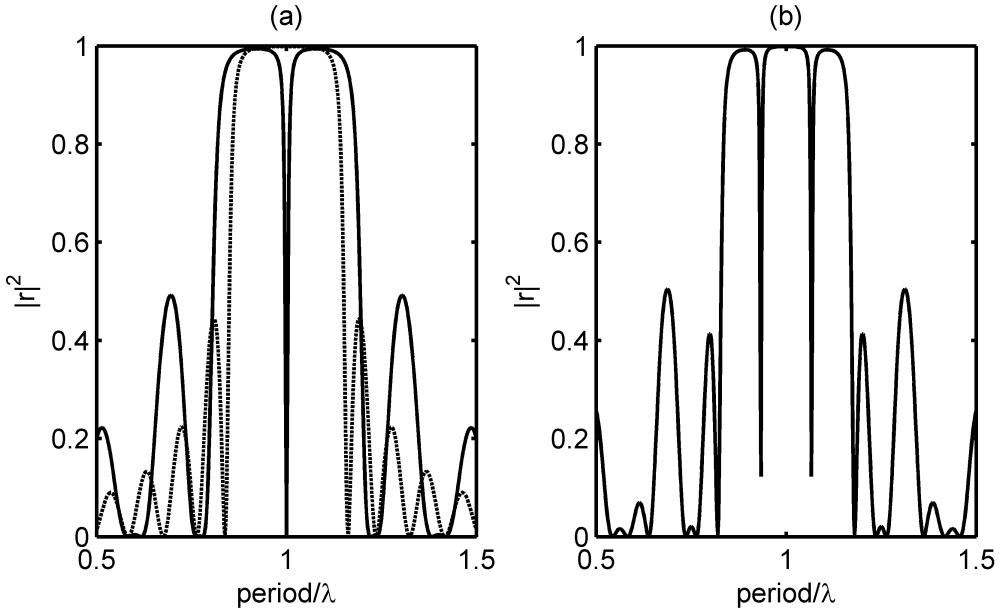


Figure 3.20: Reflection spectra of Bragg structures with $n_1 = 1, n_2 = 1.5$. (a) Solid line: n_1-defect in center with $l_{cen} = \lambda_0/2n_1$, 5 periods on left and right. Dashed line: Bragg without defect and 10 periods. (b) Thicker n_1-defect in center with $l_{cen} = 5.5\lambda_0/2n_1$, 5 periods on left and right.

Chapter 4

Fourier Optics

Contents

4.1 Basic principles of scalar diffraction theory

4.1.1 Introduction

Diffraction can be defined as "any deviation of a light ray from rectilinear propagation, which is not caused by reflection nor refraction". It was already known for centuries that light rays, passing through a small aperture in an opaque screen do not form a sharp shadow on a distant screen. That smooth transition from light to shadow could only be explained by assuming that light has a $wavelike$ character. Diffraction theory has been further developed by Huygens, Fresnel, Kirchhoff and Sommerfeld; the latter was the first to find an exact solution for the diffraction of a plane wave at a semi-infinite thin conducting plate. In this chapter we will limit ourselves to the approximation of a $scalar$ theory: only one single component of the electric or magnetic field vector is considered. This also means that we neglect the (possible) coupling between electric and magnetic fields. By comparing this approximation with exact theories, and also with experiments, it turns out that this scalar diffraction theory is good whenever:

- the diffracting aperture is large compared with the wavelength of the light;
- the diffracting field is calculated at a large distance from the aperture.

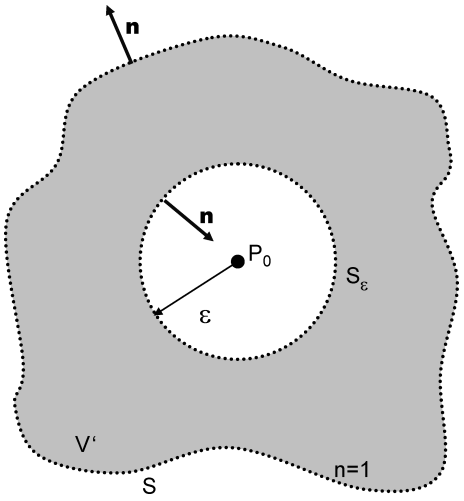


Figure 4.1: Volume V enclosed by surface S

4.1.2 Integral theorem of Helmholtz and Kirchhoff

Suppose one wants to calculate the electric field in a point of observation P_0. Consider then an arbitrary closed surface S surrounding P_0, and enclosing a volume V (figure 4.1). We also assume that the space is homogeneous, with an index of refraction $n = 1$ (free space). The theory can easily be extended to media with a real index n, simply by replacing the vacuum wavelength λ by λ/n in all equations. We moreover assume that there are no light sources nor light traps within V, and that the light is monochromatic with frequency f (or angular frequency $\omega = 2\pi f$ or wavelength $\lambda = c/f$). This implies that each field can be represented by its complex phasor, which gives the amplitude and phase. Within the volume V both the electric field E and the magnetic field H obey a Helmholtz equation:

$$\begin{aligned} \nabla^2\mathbf{E} + k^2\mathbf{E} &= 0 \\ \nabla^2\mathbf{H} + k^2\mathbf{H} &= 0 \end{aligned} \tag{4.1}$$

In the scalar approximation each component of the electric field or the magnetic field obeys the same equation. We will, from now on, represent this component by the symbol U, and we will call this the $field$. It satisfies:

$$\nabla^2 U + k^2 U = 0 \tag{4.2}$$

It is clear that $|U|^2$ is proportional to the irradiance, which gives the power density. In order to calculate the field U in the point of observation P_0 one starts from Green's theorem

$$\iiint_V \left(G\nabla^2 U - U\nabla^2 G\right) dv = \iint_S \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) ds \tag{4.3}$$

which applies whenever the field function U and Green's function G, together with their first and second derivatives are single-valued and continuous within and on S. For Green's function one chooses here a unit-amplitude spherical wave expanding about the point P_0

$$G = \frac{e^{-jk|\mathbf{r}-\mathbf{r}_0|}}{|\mathbf{r}-\mathbf{r}_0|}, \text{ with } k = \frac{2\pi}{\lambda} \tag{4.4}$$

Because this function is not continuous in P_0 we have to exclude this point from V. Therefore a small sphere with surface S_ϵ and radius ϵ around P_0 is excluded from the volume V. Green's theorem is now applied in the volume V' lying between S en S_ϵ with enclosing surface $S' = S + S_\epsilon$. It is clear that G, being a spherical wave, also obeys a Helmholtz equation

$$\nabla^2 G + k^2 G = 0 \tag{4.5}$$

Hence the left-hand side of Green's equation reduces to:

$$\int\int_{V'}\int \left(G\nabla^2 U - U\nabla^2 G\right) dv = \int\int_{V'}\int \left(GUk^2 - UGk^2\right) dv = 0 \tag{4.6}$$

and consequently

$$\int\int_{S'} \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) ds = 0 \tag{4.7}$$

or

$$\int\int_{S} \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) ds = -\int\int_{S_\varepsilon} \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) ds \tag{4.8}$$

Note that for a general point P_1 on S', one has

$$G(P_1) = \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.9}$$

$$\frac{\partial G}{\partial n}(P_1) = -\cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0)\left(jk + \frac{1}{|\mathbf{r}_1 - \mathbf{r}_0|}\right)\frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.10}$$

If P_1 lies on S_ϵ, then

$$G(P_1) = \frac{e^{-jk\varepsilon}}{\varepsilon} \tag{4.11}$$

and

$$\frac{\partial G}{\partial n}(P_1) = \left(jk + \frac{1}{\varepsilon}\right)\frac{e^{-jk\varepsilon}}{\varepsilon} \tag{4.12}$$

Letting ϵ now become arbitrary small, the continuity of U and its derivative around P_0 allows us to write:

$$\begin{aligned}\int\int_{S_\varepsilon} \left(G\frac{\partial U}{\partial n} - U\frac{\partial G}{\partial n}\right) ds &= 4\pi\varepsilon^2\left[\frac{\partial U(P_0)}{\partial n}\frac{e^{-jk\varepsilon}}{\varepsilon} - U(P_0)\frac{e^{-jk\varepsilon}}{\varepsilon}\left(jk + \frac{1}{\varepsilon}\right)\right] \\ &= -4\pi U(P_0)\end{aligned} \tag{4.13}$$

which, after substitution in (4.8) gives

$$U(P_0) = \frac{1}{4\pi}\int\int_{S} \left(\frac{\partial U}{\partial n}\frac{e^{-jk|\mathbf{r} - \mathbf{r}_0|}}{|\mathbf{r} - \mathbf{r}_0|} - U\frac{\partial}{\partial n}\frac{e^{-jk|\mathbf{r} - \mathbf{r}_0|}}{|\mathbf{r} - \mathbf{r}_0|}\right) ds \tag{4.14}$$

This formula allows for the field at an arbitrary point P_0 to be expressed in terms of its boundary values on any closed surface surrounding that point. It is known as the *integral theorem of Helmholtz and Kirchhoff*.

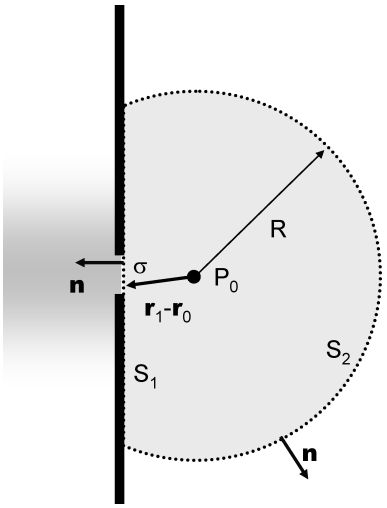


Figure 4.2: Diffraction trough an aperture in a screen

4.1.3 Diffraction through an aperture in a planar screen

Consider now the diffraction of light by an aperture in a screen (figure 4.2). The light wave is assumed to impinge from the left, and the field at P_0 is to be calculated. The previous integral theorem can be used, on condition that the surface of integration S is carefully chosen. Following Kirchhoff, we choose the surface S to consist of two parts: a plane surface S_1 lying directly behind the diffracting screen, joined and closed by a large spherical cap S_2 of radius R and centered at the observation point P_0. Applying the integral theorem of Helmholtz and Kirchhoff gives:

$$U(P_0) = \frac{1}{4\pi} \int\int_{S_1+S_2} \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds \tag{4.15}$$

where, as before

$$G = \frac{e^{-jk|\mathbf{r}-\mathbf{r}_0|}}{|\mathbf{r}-\mathbf{r}_0|} \tag{4.16}$$

Note here that Green's function is defined for the complete space, and not only in a half one. Moreover, one can prove that the function U, because it satisfies the Helmholtz equation, also satisfies the *Sommerfeld's radiation condition*:

$$\lim_{R\to\infty} R \left(\frac{\partial U}{\partial n} + jkU \right) = 0 \tag{4.17}$$

which implies that the integral (4.15) over S_2 will vanish when R becomes arbitrary large. Next we need to know the values of U and its derivative on the surface S_1. Here we follow the assumptions which Kirchhoff adopted, and which are still known as *Kirchhoff's boundary conditions*:

1. across the aperture σ, the field U and its derivative $\partial U/\partial n$ are exactly the same as they would be in the absence of a screen
2. over that portion of S_1 which differs from $sigma$, we set $U = 0$ and $\partial U/\partial n = 0$

Although these assumptions seem intuitively reasonable, they are mathematically inconsistent! Indeed: when a solution of a 3-dimensional wave equation is zero, together with its derivative, on a finite surface, then it has to be zero everywhere. Nevertheless, it turns out that Kirchhoff's boundary conditions yields results which agree very well with experiments, at least when the approximations of section 4.1.1 are satisfied. The field in the point of observation P_0 is consequently:

$$U(P_0) = \frac{1}{4\pi} \int\int_{\sigma} \left(\frac{\partial U}{\partial n} G - U \frac{\partial G}{\partial n} \right) ds \tag{4.18}$$

As we calculate the field U only in observation points P_0 at large distances from the aperture, it follows that

$$|\mathbf{r} - \mathbf{r}_0| \gg \lambda \tag{4.19}$$

$$k \gg \frac{1}{|\mathbf{r} - \mathbf{r}_0|} \tag{4.20}$$

and (4.10) becomes :

$$\frac{\partial G}{\partial n}(P_1) = -jk \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.21}$$

Substituting this in (4.18) gives:

$$U(P_0) = \frac{1}{4\pi} \int\int_{\sigma} \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \left(\frac{\partial U}{\partial n} + jkU \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \right) ds_1 \tag{4.22}$$

This formula is known as the *Fresnel-Kirchhoff diffraction formula*.

The inconsistency of Kirchhoff's boundary conditions was removed by Sommerfeld by choosing an alternative Green's function. He considered

$$G' = \frac{e^{-jk|\mathbf{r} - \mathbf{r}_0|}}{|\mathbf{r} - \mathbf{r}_0|} - \frac{e^{-jk|\mathbf{r} - \mathbf{r}'_0|}}{|\mathbf{r} - \mathbf{r}'_0|} \tag{4.23}$$

where P'_0 is the mirror image of P_0 on the opposite side of the screen (figure 4.3). The derivative now becomes:

$$\begin{aligned} \frac{\partial G'}{\partial n} = & -\cos(\mathbf{n}, \mathbf{r} - \mathbf{r}_0) \left(jk + \tfrac{1}{|\mathbf{r} - \mathbf{r}_0|} \right) \tfrac{e^{-jk|\mathbf{r} - \mathbf{r}_0|}}{|\mathbf{r} - \mathbf{r}_0|} \\ & + \cos(\mathbf{n}, \mathbf{r} - \mathbf{r}'_0) \left(jk + \tfrac{1}{|\mathbf{r} - \mathbf{r}'_0|} \right) \tfrac{e^{-jk|\mathbf{r} - \mathbf{r}'_0|}}{|\mathbf{r} - \mathbf{r}'_0|} \end{aligned} \tag{4.24}$$

For each point P_1 on the screen S_1 one has

$$|\mathbf{r}_1 - \mathbf{r}_0| = \left|\mathbf{r}_1 - \mathbf{r}'_0\right| \tag{4.25}$$

$$\cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) = -\cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}'_0) \tag{4.26}$$

hence

$$G'(P_1) = 0 \tag{4.27}$$

$$\frac{\partial G'}{\partial n}(P_1) = -2 \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \left(jk + \frac{1}{|\mathbf{r}_1 - \mathbf{r}_0|} \right) \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.28}$$

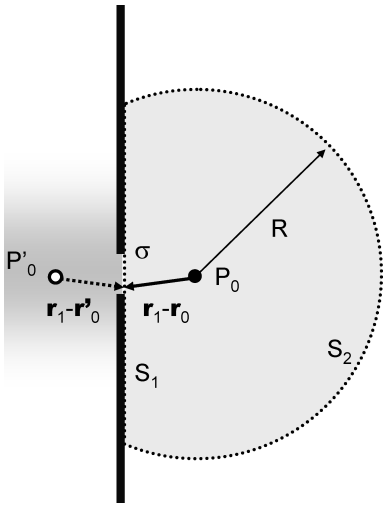


Figure 4.3: Diffraction through an aperture in a screen

Because G' is zero on the complete surface S_1, equation (4.15) reduces to

$$U(P_0) = \frac{1}{4\pi} \int\int_{S_1} \left(-U \frac{\partial G'}{\partial n} \right) ds \tag{4.29}$$

which is sometimes called the *first Rayleigh-Sommerfeld diffraction formula*. This expression also shows that it is not necessary to choose a value for the derivative of U on S_1; the knowledge of the field U suffices. One sets $U \equiv 0$ on that portion of S_1 which differs from the aperture σ, whereas across the aperture, U is exactly the same as it would be in the absence of the screen. So no boundary conditions need to be chosen for the derivative of U, and the inconsistencies of the Kirchhoff's boundary conditions have been removed. This finally leads to:

$$U(P_0) = \frac{-1}{j\lambda} \int\int_{\sigma} U(P_1) \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} ds_1 \tag{4.30}$$

or

$$U(P_0) = \int\int_{\sigma} h(P_0, P_1) U(P_1) ds_1 \tag{4.31}$$

if we set

$$h(P_0, P_1) = \frac{-1}{j\lambda} \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.32}$$

in which $h(P_0, P_1)$ is a weighting factor that is applied to the field $U(P_1)$ in order to synthesize the field in P_0 . Formulas (4.30) and (4.32) are known as the *Rayleigh-Sommerfeld diffraction formulas*. They show that the field is a superposition (=integral) of spherical waves starting from each point in the aperture, each with an appropriate amplitude and obliquity factor. This is called the *Huygens - Fresnel principle*, because it is an extension of the intuitive concept of secondary wavelets, formulated by Huygens already in 1678. Although the Sommerfeld formulation removes the inconsistencies in Kirchhoff's theory, in practical applications both formulas give essentially the same solutions, provided the aperture is much larger than the wavelength. Nevertheless one generally chooses to use the first Rayleigh Sommerfeld solution because of its simplicity.

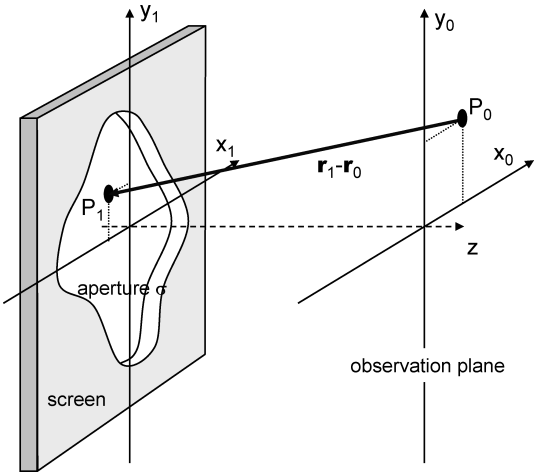


Figure 4.4: Transmission through an aperture

4.2 Fresnel and Fraunhofer diffraction

4.2.1 Fresnel diffraction formula

Assume now that the diffracting aperture lies in the (x_1, y_1) plane and is illuminated from the left by a monochromatic wave U (figure 4.4). The field is calculated in the plane of observation (x_0, y_0), parallel to the (x_1, y_1) plane, but a distance z to the right. The field in the point P_0 is given by (4.32) which we rewrite as

$$U(P_0) = \int\int_{\sigma} h(x_0, y_0, x_1, y_1) U(x_1, y_1) dx_1 dy_1 \tag{4.33}$$

with

$$h(x_0, y_0, x_1, y_1) = \frac{-1}{j\lambda} \cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \frac{e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|}}{|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.34}$$

and

$$|\mathbf{r}_1 - \mathbf{r}_0| = \sqrt{z^2 + (x_0 - x_1)^2 + (y_0 - y_1)^2} \tag{4.35}$$

Suppose now that the axial distance z is much larger than the transverse dimensions. Then

$$\cos(\mathbf{n}, \mathbf{r}_1 - \mathbf{r}_0) \cong 1 \tag{4.36}$$

The error is smaller than 5%, when the angle between $\mathbf{n}$ and $\mathbf{r}_1 - \mathbf{r}_0$ is smaller than $18°$. Also the expression $|\mathbf{r}_1 - \mathbf{r}_0|$ in the denominator of (4.34) may be replaced by z:

$$h(x_0, y_0, x_1, y_1) \cong \frac{-1}{j\lambda z} e^{-jk|\mathbf{r}_1 - \mathbf{r}_0|} \tag{4.37}$$

Furthermore one can develop the exponential in a binomial expansion, retaining only the first two terms:

$$\begin{aligned} |\mathbf{r}_1 - \mathbf{r}_0| &= \sqrt{z^2 + (x_1 - x_0)^2 + (y_1 - y_0)^2} \\ &\cong z \left[1 + \frac{1}{2} \left(\frac{x_1 - x_0}{z} \right)^2 + \tfrac{1}{2} \left(\frac{y_1 - y_0}{z} \right)^2 \right] \end{aligned} \tag{4.38}$$

This gives

$$h(x_0, y_0, x_1, y_1) \cong \frac{-e^{-jkz}}{j\lambda z} e^{-\frac{jk}{2z}\left[(x_1-x_0)^2+(y_1-y_0)^2\right]} \tag{4.39}$$

We can now replace the integration over the aperture σ by an integration over the entire plane, if we put

$$U(x_1, y_1) \equiv 0 \tag{4.40}$$

outside the aperture σ. This finally gives

$$U(x_0, y_0) = \frac{-e^{-jkz}}{j\lambda z} \int\int_{-\infty}^{+\infty} U(x_1, y_1) e^{-\frac{jk}{2z}\left[(x_1-x_0)^2+(y_1-y_0)^2\right]} dx_1 dy_1 \tag{4.41}$$

or

$$U(x_0, y_0) = \frac{-e^{-jkz}}{j\lambda z} e^{-\frac{jk}{2z}\left[x_0^2+y_0^2\right]} \int\int_{-\infty}^{+\infty} U(x_1, y_1) e^{-\frac{jk}{2z}\left[x_1^2+y_1^2\right]} e^{j\frac{2\pi}{\lambda z}[x_0 x_1 + y_0 y_1]} dx_1 dy_1 \tag{4.42}$$

This result is called the *Fresnel diffraction integral*. It clearly shows that the field $U(x_0, y_0)$ in the observation plane is the 2-dimensional Fourier transform of the field in the object plane

$$U(x_1, y_1) e^{-\frac{jk}{2z}\left[x_1^2+y_1^2\right]} \tag{4.43}$$

where the spatial frequencies are defined by:

$$f_x = -\frac{x_0}{\lambda z} \text{ en } f_y = -\frac{y_0}{\lambda z} \tag{4.44}$$

As this result is valid close to the aperture, it is called the *near-field approximation* ; one sometimes speaks of the *Fresnel diffraction regime*. This approximation is not valid too close to the aperture; however it is not easy to calculate exactly the limits of validity. A sufficient condition is that the higher-order term in the expansion be small, but this is not a necessary condition. Indeed. it suffices that they do not change the value of the integral too much *after integration*, and this also depends on the function U. In regions where the exponential varies only slowly ("stationary-phase" regime) the contributions of the higher-order terms may often be neglected, even for large values of $k/2z$. The general conclusions of deeper analyses is that the accuracy of the Fresnel approximation is extremely good to distances that are very close to the aperture.

4.2.2 Fraunhofer approximation

When the distance z between the two planes is so large that

$$z >> \frac{k\left(x_1^2 + y_1^2\right)_{\max}}{2} \tag{4.45}$$

then a further simplification is possible: the quadratic phase term can also be neglected, giving:

$$U(x_0, y_0) = \frac{-e^{-jkz}}{j\lambda z} e^{-\frac{jk}{2z}\left[x_0^2+y_0^2\right]} \int\int_{-\infty}^{+\infty} U(x_1, y_1) e^{j\frac{2\pi}{\lambda z}[x_0 x_1 + y_0 y_1]} dx_1 dy_1 \tag{4.46}$$

This shows that the field in the image plane is the Fourier transform of the field in the aperture, when the spatial frequencies are set to $f_x = -x_0/\lambda z$ en $f_y = -y_0/\lambda z$. The region where this approximation is valid is called the *far field* or the *Fraunhofer diffraction regime*. For example for a HeNe laser with a wavelength of about $6.10^{-7}m$ and an aperture of $1mm$ the far field starts at about $z > 5m$.

Important remark: In the previous sections we calculated the diffraction of a field incident upon an aperture. This aperture limited the transverse extent of the field, a condition necessary to develop the theoretical model. When, on the other hand, the incident field is smaller than the aperture itself (example: a point source, or a laser beam), then the aperture has no influence on the (diffraction of the) field. This implies that the theory remains valid for describing diffraction of any field with $finite$ transverse dimensions.

4.2.3 Examples of Fraunhofer diffraction patterns

Rectangular aperture

Consider a rectangular aperture in a screen. The amplitude transmittance of that screen can then be written as:

$$t(x_1, y_1) = \text{rect}\left(\frac{x_1}{\ell_x}\right)\text{rect}\left(\frac{y_1}{\ell_y}\right) \tag{4.47}$$

in which ℓ_x and ℓ_y are the width and height of the aperture. Suppose this aperture is illuminated by a plane monochromatic wave of unit amplitude, normally incident on the screen, then $U(x_1, y_1) = t(x_1, y_1)$. Formula (4.46) then gives the Fraunhofer diffraction pattern:

$$U(x_0, y_0) = -\frac{e^{-jkz}}{j\lambda z}e^{-\frac{jk}{2z}(x_0^2+y_0^2)}F_{2D}\left\{U(x_1, y_1)\right\} \tag{4.48}$$

with

$$f_x = -\frac{x_0}{\lambda z} \text{ and } f_y = -\frac{y_0}{\lambda z} \tag{4.49}$$

Because

$$F_{2D}\left\{U\left(x_1,y_1\right)\right\} = \ell_x\ell_y\text{sinc}\left(\pi\ell_x f_x\right)\text{sinc}\left(\pi\ell_y f_y\right) \tag{4.50}$$

with

$$\text{sinc}(x) = \frac{\sin x}{x} \tag{4.51}$$

one finds

$$U\left(x_0, y_0\right) = \frac{-e^{-jkz}}{j\lambda z}e^{-j\frac{k}{2z}(x_0^2+y_0^2)}\ell_x\ell_y\text{sinc}\left(\frac{\pi\ell_x x_0}{\lambda z}\right)\text{sinc}\left(\frac{\pi\ell_y y_0}{\lambda z}\right) \tag{4.52}$$

The irradiance I of the diffraction pattern is consequently:

$$I(x_0, y_0) = \left(\frac{\ell_x\ell_y}{\lambda z}\right)^2\text{sinc}^2(\frac{\pi\ell_x x_0}{\lambda z})\text{sinc}^2(\frac{\pi\ell_y y_0}{\lambda z}) \tag{4.53}$$

Figure 4.5 shows the pattern along the x_0 axis, figure 4.6 shows an experimental far field diffraction pattern of a rectangular aperture **[?]**. Remark: in optics text books the sinc-function is usually defined as $sinc(x) = (sin\pi x)/\pi x)$.

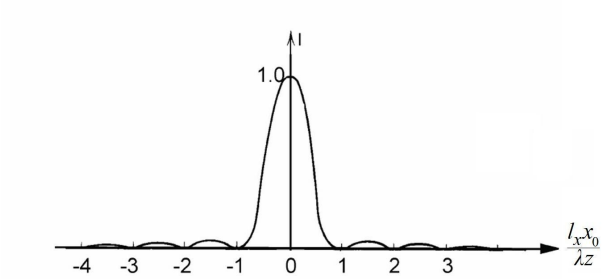


Figure 4.5: Fraunhofer diffraction pattern of a rectangular aperture

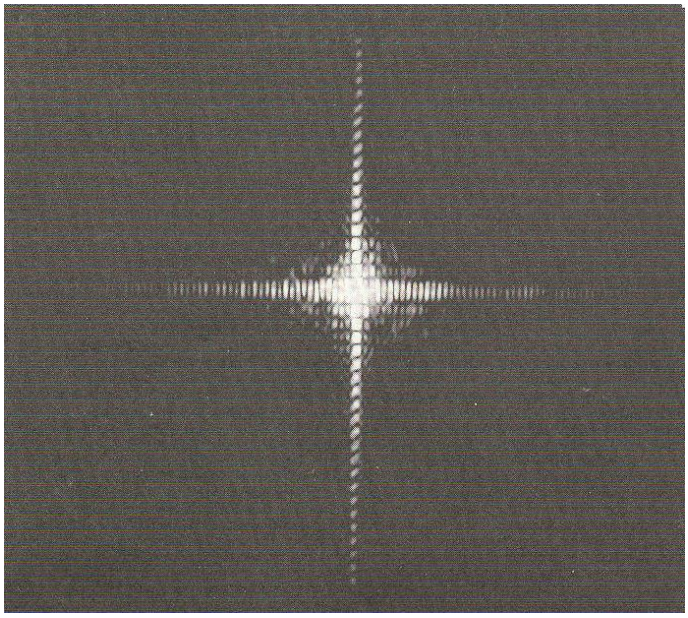


Figure 4.6: An experimental Fraunhofer diffraction pattern of a rectangular aperture (from [**?**])

x	$\left[2\frac{J_1(\pi x)}{\pi x}\right]^2$	max or min
0	1	max
1.220	0	min
1.635	0.0175	max
2.233	0	min
2.679	0.0042	max
3.238	0	min
3.699	0.0016	max

Table 4.1: Values of the Airy pattern at successive maxima and minima.

Circular aperture

The amplitude transmittance t of a circular aperture with diameter l is given by

$$t(r_1) = \text{circ}\left[\frac{r_1}{\ell/2}\right] \tag{4.54}$$

Because of the circular symmetry, the Fourier transform in formula (4.52) reduces to a Fourier-Bessel transform B:

$$B\{f\} = 2\pi \int_0^{+\infty} r f(r) J_0(2\pi f r)\, dr \tag{4.55}$$

If, once again, the illuminating wave is a plane monochromatic wave, of unit amplitude, normally incident on the screen, then $U(r_1) = t(r_1)$.

Because

$$B\left\{\text{circ}\left(\frac{r_1}{\ell/2}\right)\right\} = \left(\frac{\ell}{2}\right)^2 \frac{J_1(\pi \ell f)}{\ell f/2} \tag{4.56}$$

one finds

$$U(r_0) = \frac{e^{-jkz}}{j\lambda z} e^{-j\,\frac{kr_0^2}{2z}} \left[\left(\frac{\ell}{2}\right)^2 \frac{J_1\left(\frac{\pi \ell r_0}{\lambda z}\right)}{\frac{\ell r_0}{2\lambda z}}\right] \tag{4.57}$$

or

$$U(r_0) = e^{-jkz} e^{-j\,\frac{kr_0^2}{2z}} \frac{k\ell^2}{j8z} \left[2\frac{J_1\left(\frac{k\ell r_0}{2z}\right)}{\frac{k\ell r_0}{2z}}\right] \tag{4.58}$$

and the irradiance becomes:

$$I(r_0) = \left(\frac{k\ell^2}{8z}\right)^2 \left[2\frac{J_1\left(\frac{k\ell r_0}{2z}\right)}{\frac{k\ell r_0}{2z}}\right]^2 \tag{4.59}$$

This light distribution is called an $Airy$ pattern, after G.B. Airy, an astronomer who first derived it.

From figure 4.7 one can see that the distance r_0 to the first zero equals $r_0 = 1.22\lambda z/\ell$. This is also the radius of the circular Airy-spot. Figure 4.8 shows an experimental far field diffraction pattern of a circular aperture **[?]**.

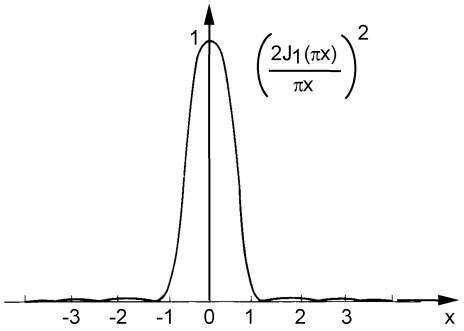


Figure 4.7: Fraunhofer diffraction pattern of a circular aperture or Airy pattern

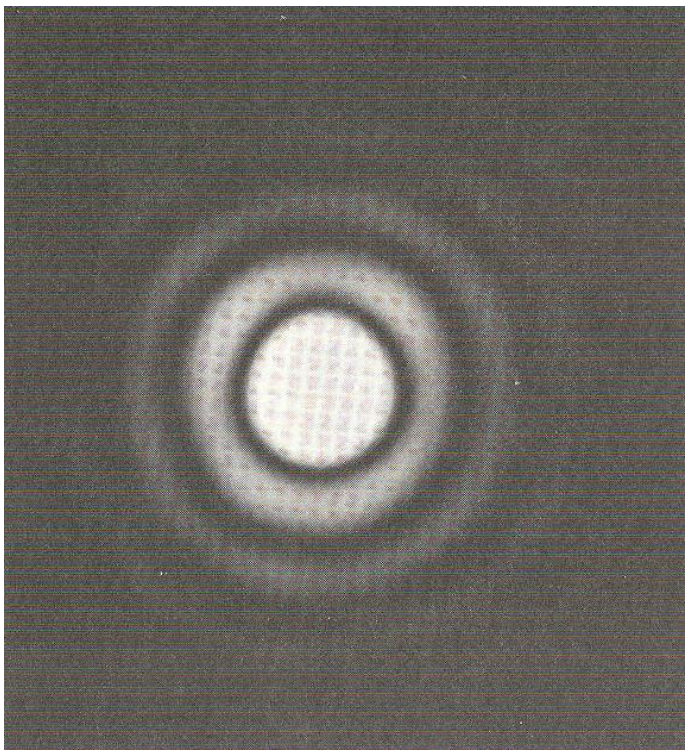


Figure 4.8: An experimental Fraunhofer diffraction pattern of a circular aperture, from [**?**]

Gaussian beam

As third example we consider a Gaussian beam with its waist in the (x_1, y_1)-plane. Hence it has a plane wavefront, and an amplitude distribution:

$$U(x_1, y_1) = e^{-\frac{x_1^2+y_1^2}{w^2}} \tag{4.60}$$

This amplitude does not go to zero at a finite distance from the axis; hence the previous formulas do not apply, at least in theory. However the exponential decreases so fast that we can still use the diffraction formulas. Because the Fourier transform of Gaussian is again a Gaussian:

$$F_{2D}\left\{e^{-\frac{x^2+y^2}{w^2}}\right\} = \pi w^2 e^{-\pi^2 w^2\left(f_x^2+f_y^2\right)} \tag{4.61}$$

we find

$$U(x_0, y_0) = -\pi\frac{w^2 e^{-jkz}}{j\lambda z} e^{\frac{-jk}{2z}\left(x_0^2+y_0^2\right)} e^{-\pi^2 w^2\left(f_x^2+f_y^2\right)} \tag{4.62}$$

with

$$f_x = -\frac{x_0}{\lambda z} \text{ and } f_y = -\frac{y_0}{\lambda z} \tag{4.63}$$

Hence in the far field a Gaussian beam behaves as a spherical wave, with half-angle:

$$\theta = \frac{\lambda}{\pi w} \tag{4.64}$$

This is exactly the same result we obtained (in another way) when studying Gaussian beams in the lectures on lasers. Condition (4.45) means that the Fraunhofer regime starts after a few times the Rayleigh range.

4.2.4 Fresnel diffraction at a square aperture

Calculations of Fresnel diffraction patterns are much more complicated than Fraunhofer ones, simply because one can not use the well-known Fourier transform formulas. We illustrate this with the simple example of a square aperture with side ℓ, normally illuminated with a monochromatic plane wave of unit amplitude. The field in the object plane is then

$$U(x_1, y_1) = \text{rect}\left(\frac{x_1}{\ell}\right)\text{rect}\left(\frac{y_1}{\ell}\right) \tag{4.65}$$

Application of formula (4.41) gives

$$\begin{aligned} U(x_0, y_0) &= \frac{-e^{-jkz}}{j\lambda z}\int_{-\ell/2}^{+\ell/2}\int e^{-\frac{jk}{2z}\left[(x_1-x_0)^2+(y_1-y_0)^2\right]}dx_1 dy_1 \\ &= \frac{-e^{-jkz}}{j\lambda z}\int_{-\ell/2}^{+\ell/2} e^{-\frac{jk}{2z}(x_1-x_0)^2}dx_1 \int_{-\ell/2}^{+\ell/2} e^{-\frac{jk}{2z}(y_1-y_0)^2}dy_1 \end{aligned} \tag{4.66}$$

Introducing new variables u and v by

$$u = \sqrt{\frac{k}{\pi z}}\,(x_1 - x_0) \text{ and } v = \sqrt{\frac{k}{\pi z}}\,(y_1 - y_0) \tag{4.67}$$

gives

$$U(x_0, y_0) = \frac{-e^{-jkz}}{j\lambda z}\frac{\pi z}{k}\int_{u_1}^{u_2} e^{-\frac{j\pi u^2}{2}}du \int_{v_1}^{v_2} e^{-\frac{j\pi v^2}{2}}dv \tag{4.68}$$

with

$$\begin{aligned} u_1 &= \sqrt{\frac{k}{\pi z}}\left(-\frac{\ell}{2} - x_0\right),\ u_2 = \sqrt{\frac{k}{\pi z}}\left(+\frac{\ell}{2} - x_0\right) \\ v_1 &= \sqrt{\frac{k}{\pi z}}\left(-\frac{\ell}{2} - x_0\right),\ v_2 = \sqrt{\frac{k}{\pi z}}\left(\frac{\ell}{2} - x_0\right) \end{aligned} \tag{4.69}$$

These integrals can be expressed in terms of the Fresnel integrals:

$$C\left(\alpha\right) = \int_0^{\alpha} \cos\frac{\pi t^2}{2}dt \text{ and } S\left(\alpha\right) = \int_0^{\alpha} \sin\frac{\pi t^2}{2}dt \tag{4.70}$$

Because

$$\begin{aligned} \int_{u_1}^{u_2} e^{-\frac{j\pi u^2}{2}}du &= \int_0^{u_2}\left(\cos\frac{\pi}{2}u^2 - j\sin\frac{\pi}{2}u^2\right)du - \int_0^{u_1}\left(\cos\frac{\pi}{2}u^2 - j\sin\frac{\pi}{2}u^2\right)du \\ &= \left[C\left(u_2\right) - C\left(u_1\right)\right] - j\left[S\left(u_2\right) - S\left(u_1\right)\right] \end{aligned} \tag{4.71}$$

we find

$$\begin{aligned} U(x_0, y_0) = -\frac{e^{-jkz}}{2j} \quad & \left([C(u_2) - C(u_1)] - j\left[S(u_2) - S(u_1)\right]\right). \\ & \left([C(v_2) - C(v_1)] - j\left[S(v_2) - S(v_1)\right]\right) \end{aligned} \tag{4.72}$$

and the irradiance distribution I is:

$$I(x_0) = \frac{1}{4}\left([C(u_2) - C(u_1)]^2 + [S(u_2) - S(u_1)]^2\right)\left([C(v_2) - C(v_1)]^2 + [S(v_2) - S(v_1)]^2\right) \tag{4.73}$$

The meaning of this expression can be easily understood with the help of figure 4.9a which gives $C(u)$ on the horizontal and $S(u)$ on the vertical axis as a function of the parameter u. This graph is known as the Cornu-spiral. One can prove that the irradiance $I(x_0, y_0)$ is proportional to the length of a line segment connecting two points on the spiral. When the image point shifts, the representative points run through the spiral. Hence the irradiance oscillates strongly, as illustrated in figure 4.9b . This graph shows the Fresnel diffraction pattern of a one-dimensional slit of length l, measured at a given distance z from the slit. When the observation plane approaches the plane of the slit, the shape of the diffraction pattern approaches the shape of the slit itself. On the other hand, at larger distances, the diffraction pattern becomes much wider than the slit, ending up (at very large distances) with a Fraunhofer diffraction pattern.

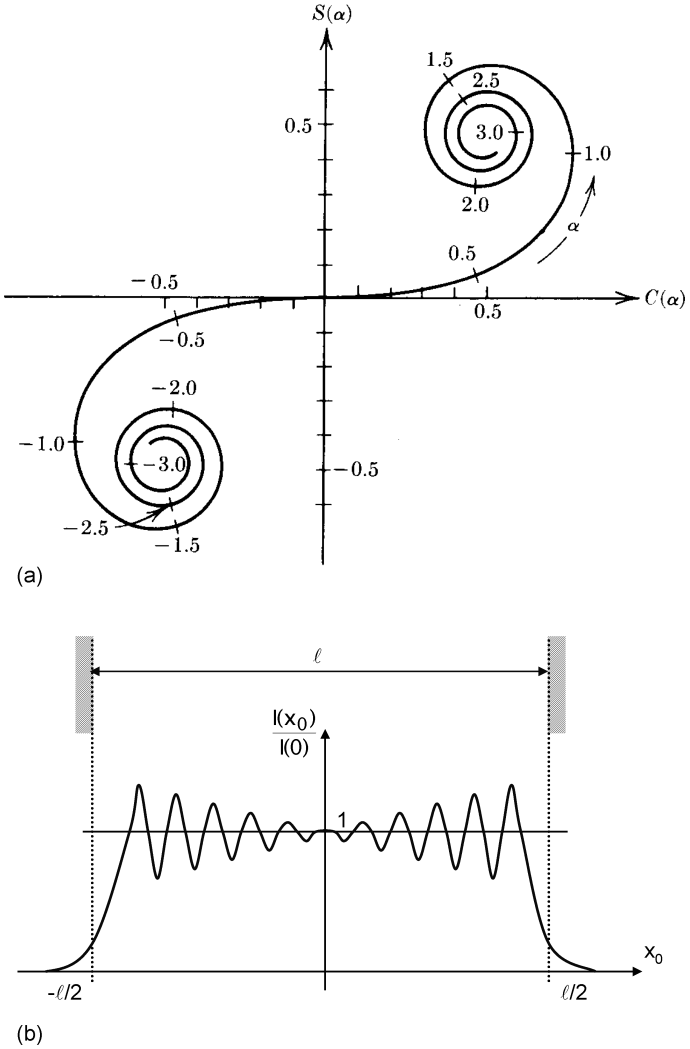


Figure 4.9: Calculation of Fresnel diffraction. (a) Cornu-spiral, (b) diffraction pattern of a slit

4.2.5 Fresnel diffraction and spatial frequencies

Formulas (4.34) and (4.41) can also be written as:

$$U(x_0, y_0) = \int\int_S h(x_0 - x_1, y_0 - y_1) U(x_1, y_1) dx_1 dy_1 \tag{4.74}$$

with

$$h(x_0 - x_1, y_0 - y_1) \cong \frac{-e^{-jkz}}{j\lambda z} e^{-\frac{jk}{2z}\left[(x_1 - x_0)^2 + (y_1 - y_0)^2\right]} \tag{4.75}$$

You immediately recognize a convolution structure:

$$f * g = \int_{-\infty}^{+\infty} f(\tau) g(t - \tau) \, d\tau \tag{4.76}$$

Hence the image $U(x_0, y_0)$ is a two-dimensional convolution between $U(x_1, y_1)$ and

$$h(x_0, y_0) = \frac{-e^{-jkz}}{j\lambda z} e^{-\frac{jk}{2z}\left[x_0^2 + y_0^2\right]} \tag{4.77}$$

If we define the following Fourier transforms:

$$\begin{array}{ll} F_0(f_x, f_y) & = F_{2D}\{U(x_0, y_0)\} \\ F_1(f_x, f_y) & = F_{2D}\{U(x_1, y_1)\} \\ H(f_x, f_y) & = F_{2D}\{h(x_0, y_0)\} \end{array} \tag{4.78}$$

then the Fourier transform of (4.74) gives

$$F_0(f_x, f_y) = F_1(f_x, f_y) \cdot H(f_x, f_y) \tag{4.79}$$

This means that $H(f_x, f_y)$ is a *transfer function* which describes the evolution of the spatial spectrum of the light within the Fresnel diffraction regime. That transfer function $H(f_x, f_y)$ is nothing else than the two-dimensional Fourier-transform of (4.77):

$$\begin{array}{ll} H(f_x, f_y) & = F_{2D}\{h(x_0, y_0)\} \\ & = -\dfrac{e^{-jkz}}{j\lambda z} F_x\left\{e^{-\frac{jk}{2z}x_0^2}\right\} F_y\left\{e^{-\frac{jk}{2z}y_0^2}\right\} \end{array} \tag{4.80}$$

With the formula:

$$F\left\{e^{j\alpha t^2}\right\} = \sqrt{\frac{\pi}{\alpha}} e^{j\frac{\pi}{4}} e^{-j\frac{(2\pi f)^2}{4\alpha}} \tag{4.81}$$

this becomes

$$\begin{array}{ll} H(f_x, f_y) & = -\dfrac{e^{-jkz}}{j\lambda z} \sqrt{\dfrac{-2\pi z}{k}} e^{j\frac{\pi}{4}} e^{-j\frac{4\pi^2 f_x^2}{-4k/2z}} \sqrt{\dfrac{-2\pi z}{k}} e^{j\frac{\pi}{4}} e^{-j\frac{4\pi^2 f_y^2}{-4k/2z}} \\ & = -\dfrac{e^{-jkz}}{j\lambda z} \dfrac{-2\pi z}{k} j e^{-j\pi\lambda z f_x^2} e^{-j\pi\lambda z f_y^2} \\ & = e^{-jkz} e^{j\pi\lambda z\left(f_x^2 + f_y^2\right)} \end{array} \tag{4.82}$$

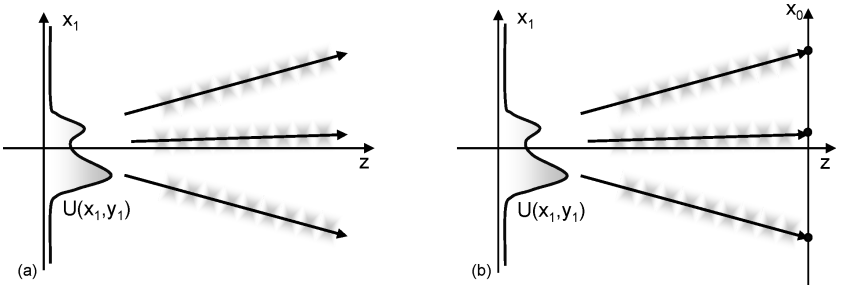


Figure 4.10: Plane-wave spectrum of an aperture

4.2.6 The angular spectrum of plane waves

In this section we will give a physical explanation of the previous conclusions. Let us rewrite (4.78):

$$F_1(f_x, f_y) = \int\int_{-\infty}^{+\infty} U(x_1, y_1)\, e^{-j2\pi(x_1 f_x + y_1 f_y)} dx_1 dy_1 \tag{4.83}$$

and its inverse

$$U(x_1, y_1) = \int\int_{-\infty}^{+\infty} F_1(f_x, f_y)\, e^{+j2\pi(x_1 f_x + y_1 f_y)} df_x df_y \tag{4.84}$$

The latter formula says that the field U in the (x_1, y_1)-plane can be considered as a superposition of fields $\exp(j2\pi(x_1 f_x + y_1 f_y))$, each with its own amplitude $F_1(f_x, f_y)$. Now you know that $\exp(j2\pi(x_1 f_x + y_1 f_y))$, is nothing else than the intersection of the (x_1, y_1)-plane, with a plane wave of which the propagation **k**-vector has components:

$$k_x = 2\pi f_x, \qquad k_y = 2\pi f_y \quad \text{en} \quad k_z = \sqrt{k^2 - k_x^2 - k_y^2} \tag{4.85}$$

In other words: the Fourier decomposition of the field $U(x_1, y_1)$ is a decomposition in plane waves. This is illustrated in figure 4.10a.

This explains why the function $F_1(f_x, f_y)$ is called the *angular spectrum of plane waves*. When propagating over a distance z in a homogeneous space, each of those plane waves acquires a phase increase of $\exp(-j2\pi f_z z)$, with

$$f_z = \sqrt{\frac{1}{\lambda^2} - f_x^2 - f_y^2} \tag{4.86}$$

The field in an arbitrary (x_0, y_0)-plane, after propagating over a distance z, is found by adding the new plane waves together:

$$U(x_0, y_0) = \int\int_{-\infty}^{\infty} F_1(f_x, f_y)\, H(f_x, f_y)\, e^{j2\pi(x_o f_x + y_o f_y)} df_x df_y \tag{4.87}$$

with

$$H(f_x, f_y) = e^{-j2\pi\sqrt{\frac{1}{\lambda^2} - f_x^2 - f_y^2}\, z} \tag{4.88}$$

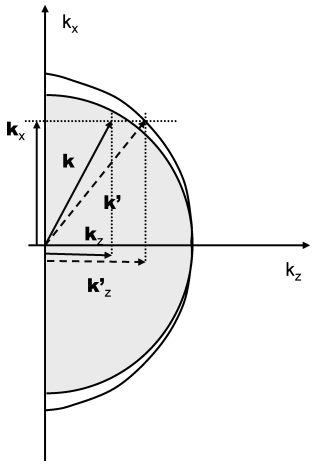


Figure 4.11: Paraboloid approximation of a spherical wavefront

For paraxial fields:

$$f_x \ll f_z \text{ en } f_y \ll f_z \tag{4.89}$$

hence

$$f_z \approx \frac{1}{\lambda}\left(1 - \frac{(\lambda f_x)^2 + (\lambda f_y)^2}{2}\right) \tag{4.90}$$

and consequently:

$$H\left(f_x, f_y\right) = e^{-jkz} e^{j\pi\lambda z\left(f_x^2 + f_y^2\right)} \tag{4.91}$$

which is exactly the same result as (4.82).

The paraxial approximation is illustrated in figure 4.11. In principle the **k**-vector lies on a circle (for a given wavelength); in the paraxial approximation it lies on a paraboloid.

We have already seen that the field at a large distance (Fraunhofer regime) is, up to a proportionality constant and a phase curvature, the Fourier transform of the original field (when the spatial frequencies f_x en f_y are expressed as function of the position x_0 en y_0). This can create some confusion, because we have now seen that this Fourier transform is in fact nothing else than a decomposition in plane waves. Those two models are not contradictory, when each of those waves are replaced by a light ray, starting at the $z = 0$ plane. The field at position (x_0, y_0, z) is then created by the ray with k-vector:

$$\begin{aligned} k_x &= 2\pi f_x = -\tfrac{2\pi}{\lambda}\tfrac{x_0}{z} \\ k_y &= 2\pi f_y = -\tfrac{2\pi}{\lambda}\tfrac{y_0}{z} \end{aligned} \tag{4.92}$$

These are exactly the same relations as in the Fraunhofer formula. This is illustrated in figure 4.10b. In reality the interpretation of the Fraunhofer formula is a little bit more complicated, because a plane wave extends to infinity, and is not a simple ray. In the next section we show mathematically how the Fresnel-field (= superposition of plane waves at each position) transforms into the Fraunhofer-field (= one single plane wave at each position). This will show that the ray model can indeed be used.

4.2.7 Transition from Fresnel to Fraunhofer regime

The transition from the Fresnel to the Fraunhofer diffraction regime can best be understood with the angular-spectrum description. We start with the angular spectrum (= decomposition in plane waves) in the $z = 0$ plane. In order to calculate the field at position z we use the propagator (4.82):

$$H(f_x, f_y) = e^{-jkz} e^{j\pi\lambda z\left(f_x^2+f_y^2\right)} \tag{4.93}$$

For simplicity we will limit ourselves here to *two* dimensions. The field at position z is then :

$$\begin{aligned} U(x,z) &= e^{-jkz} \int_{-\infty}^{+\infty} F_1(f_x)\ e^{j\pi\lambda z f_x^2}\ e^{j2\pi f_x x}\ df_x \\ &= e^{-jkz} \int_{-\infty}^{+\infty} F_1(f_x)\ e^{j\pi\left(\sqrt{\lambda z}\, f_x + \frac{x}{\sqrt{\lambda z}}\right)^2} e^{-j\pi\frac{x^2}{\lambda z}}\ df_x \end{aligned} \tag{4.94}$$

For obtaining the Fraunhofer regime, z has to be large enough. The leading term in the integrand is the exponential one:

$$e^{j\pi\left(\sqrt{\lambda z}\, f_x + \frac{x}{\sqrt{\lambda z}}\right)^2} \tag{4.95}$$

For large z values, this function oscillates very fast in f_x, except when:

$$\sqrt{\lambda z} f_x \approx -\frac{x}{\sqrt{\lambda z}} \tag{4.96}$$

The fast oscillating part does not contribute to the integral, hence:

$$\begin{aligned} U(x,z) &= e^{-jkz}\ e^{-j\pi\frac{x^2}{\lambda z}}\ F_1\left(f_x = \frac{-x}{\lambda z}\right) \int_{-\infty}^{+\infty} e^{j\pi\lambda z f_x^2}\ df_x \\ &= e^{-jkz}\ e^{-j\pi\frac{x^2}{\lambda z}}\ F_1\left(f_x = \frac{-x}{\lambda z}\right)\ \frac{2}{\sqrt{2\lambda z}} \int_0^{+\infty} e^{j\pi\frac{t^2}{2}} dt \quad \text{with} \quad t = \sqrt{2\lambda z}\ f_x \end{aligned} \tag{4.97}$$

Because:

$$\int_0^{+\infty} e^{j\pi\frac{t^2}{2}} dt = C(\infty) + jS(\infty) = \frac{1}{2}(1+j) \tag{4.98}$$

the field at position z (z large) is then:

$$U(x,z) = \frac{e^{-jkz}}{\sqrt{2\lambda z}}\ (1+j)\ F_1\left(f_x = \frac{-x}{\lambda z}\right)\ e^{-j\pi\frac{x^2}{\lambda z}} \tag{4.99}$$

In a similar way one can prove the general expression:

$$U(x,y,z) = \frac{e^{-jkz}}{2\lambda z}\ (1+j)^2\ F_1\left(f_x = \frac{-x}{\lambda z}, f_y = \frac{-y}{\lambda z}\right)\ e^{\frac{-j\pi}{\lambda z}\left(x^2+y^2\right)} \tag{4.100}$$

which is similar to (4.46).

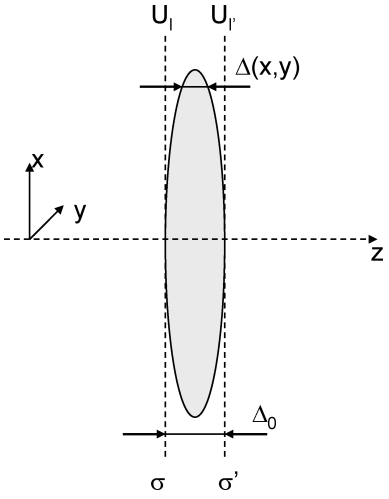


Figure 4.12: A thin lens

4.2.8 Coherent versus incoherent fields

Up to now we always considered all fields to be monochromatic: in each point in space the field oscillates sinusoidal, with a well-defined amplitude and phase. The phase *difference* between two different points in space and time is then constant. We call this kind of fields *coherent* ones. For these coherent fields the Fourier transform can be interpreted as a decomposition in plane waves, and each spatial frequency corresponds with one single direction of propagation. But light sources can also be very broadband, as for instance when white light is used. When the field is not monochromatic, things are completely different. Now there is no fixed phase relation anymore between two different points, and the fields are said to be *incoherent*. Of course it is still possible to Fourier transform that field, but this has not anymore the meaning of a decomposition in plane waves. One can consider incoherent light to be a superposition of monochromatic, coherent contributions. This can be illustrated by calculating the far field light distribution of a rectangular aperture. When illuminated with coherent light (eg with a laser) the far field is a sinc-distribution, as we have calculated in section 2.3. If this same aperture is now illuminated with white light, the far field is a very diffuse spot. This can however be seen as an incoherent superposition of many sinc-contributions, each one with its own width.

4.3 Fourier transforming properties of optical systems

4.3.1 Phase transformation by a lens

When light passes through a lens, it undergoes a phase transform (figure 4.12); in this section we will calculate it for a monochromatic plane wave. A lens is composed of material (e.g. glass) in which light travels *slower* than in air; this is described by the index of refraction n. Here we assume that the lens is a *thin* lens, which implies that a light ray will leave the tangent plane of the lens T at the same transverse position (x, y) as on entering. The only effect of the lens is a retardation over a time delay which is proportional to the local thickness $\Delta(x, y)$ of the lens. If we write Δ_0 for the maximal thickness of the lens (the thickness in the middle), then the phase retardation between

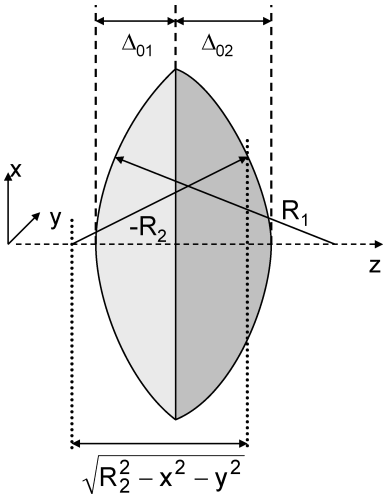


Figure 4.13: Parameters of a thin lens

both tangent planes σ and σ' is given by

$$\phi(x,y) = \underbrace{kn\Delta(x,y)}_{\downarrow\, lens} + \underbrace{k\left[\Delta_0 - \Delta(x,y)\right]}_{\downarrow\, air} \tag{4.101}$$

Let us write $U_l(x,y)$ for the field incident on the first tangent plane; the field leaving the lens is then:

$$U_{l'}(x,y) = t_l(x,y)U_l(x,y) \tag{4.102}$$

with

$$t_l(x,y) = e^{-jk\Delta_0}e^{-jk(n-1)\Delta(x,y)} \tag{4.103}$$

For calculating $\Delta(x,y)$ we split the lens in two parts (figure 4.13) such that

$$\Delta(x,y) = \Delta_1(x,y) + \Delta_2(x,y) \tag{4.104}$$

The radius of curvature R is positive for a concave surface, and negative for a convex one. This gives

$$\begin{aligned}\Delta_1(x,y) &= \Delta_{01} - \left(R_1 - \sqrt{R_1^2 - x^2 - y^2}\right) \\ &= \Delta_{01} - R_1\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_1^2}}\right)\end{aligned} \tag{4.105}$$

and

$$\begin{aligned}\Delta_2(x,y) &= \Delta_{02} - \left(-R_2 - \sqrt{R_2^2 - x^2 - y^2}\right) \\ &= \Delta_{02} + R_2\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_2^2}}\right)\end{aligned} \tag{4.106}$$

Consequently

$$\Delta(x,y) = \Delta_0 - R_1\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_1^2}}\right) + R_2\left(1 - \sqrt{1 - \frac{x^2+y^2}{R_2^2}}\right) \tag{4.107}$$

with

$$\Delta_0 = \Delta_{01} + \Delta_{02} \tag{4.108}$$

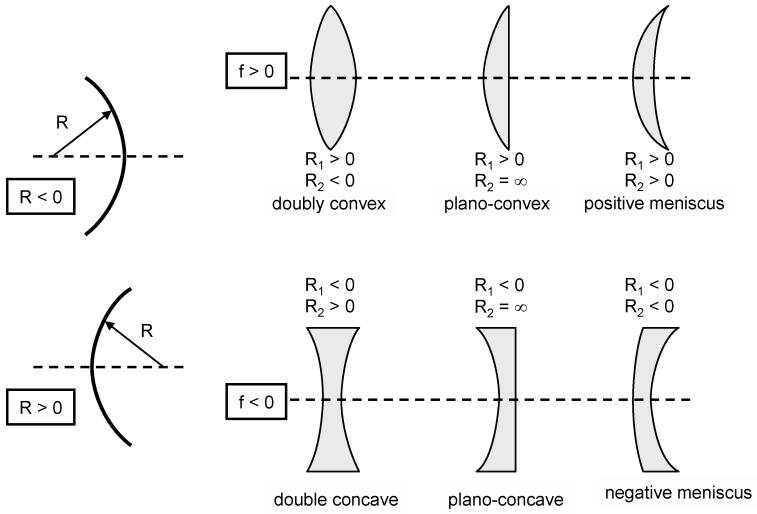


Figure 4.14: Convex and concave lenses

4.3.2 The paraxial approximation

The expression for $\Delta(x, y)$ can be simplified in the paraxial approximation; then :

$$\begin{aligned} \sqrt{1 - \frac{x^2 + y^2}{R_1^2}} &\cong 1 - \frac{x^2 + y^2}{2R_1^2} \\ \sqrt{1 - \frac{x^2 + y^2}{R_2^2}} &\cong 1 - \frac{x^2 + y^2}{2R_2^2} \end{aligned} \tag{4.109}$$

hence

$$\Delta(x, y) = \Delta_0 - \frac{x^2 + y^2}{2} \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \tag{4.110}$$

Substitution of (4.110) in (4.103) gives

$$t_l(x, y) = e^{-jkn\Delta_0} e^{jk(n-1)\frac{x^2+y^2}{2}\left[\frac{1}{R_1} - \frac{1}{R_2}\right]} \tag{4.111}$$

We note that the physical parameters of the lens n, R_1 and R_2 can be combined in one single parameter f (which we call the "focal distance")

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \tag{4.112}$$

The phase transformation of the lens now becomes:

$$t_l(x, y) = e^{-jkn\Delta_0} e^{j\frac{k}{2f}\left(x^2+y^2\right)} \tag{4.113}$$

The sign conventions we have adopted for the radii of curvature R_1 and R_2 for a double convex lens in figure 4.15 can also be used for other types of lenses, see figure 4.14. You can control that for the upper row in figure 4.14 the foci f are positive, whereas for the lower row they are negative.

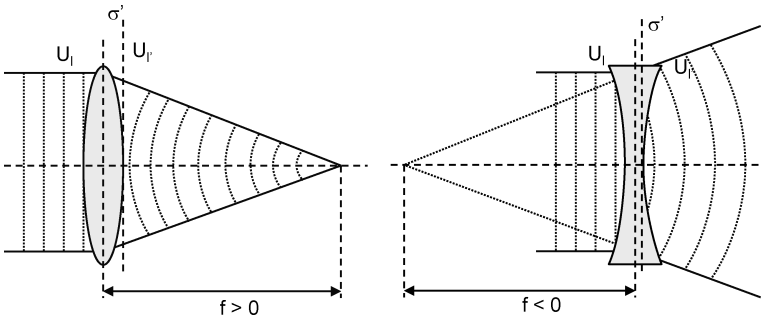


Figure 4.15: Wavefront at convex and concave lenses

When a plane unit-amplitude wave is incident perpendicular on the lens, the field leaving it is given by:

$$U_{l'}(x,y) = e^{-jkn\Delta_0} e^{j\frac{k}{2f}\left(x^2+y^2\right)} \tag{4.114}$$

The first part gives simply a constant phase retardation, whereas the second part describes a curvature of the wavefront converging towards a point on the z-axis at a distance f from the lens (figure 4.15a).

The field in the plane σ' is indeed given by:

$$e^{jk\sqrt{f^2+x^2+y^2}} = e^{jkf\sqrt{1+\frac{x^2+y^2}{f^2}}} \tag{4.115}$$

or in the paraxial approximation:

$$e^{jkf\sqrt{1+\frac{x^2+y^2}{f^2}}} \cong e^{jkf} e^{j\frac{k}{2f}\left(x^2+y^2\right)} \tag{4.116}$$

and this is the same as in (4.114).

If the focal distance is negative, then the wave is a divergent spherical wave, the origin of which lies a distance f in front of the lens. Note that these conclusions are only valid in the paraxial approximation. If this approximation is not valid, then the wave leaving the lens is not a spherical one, and all kinds of aberrations will show up.

4.3.3 Fourier-transforming properties of a mask placed against a lens

Suppose now we position a (gray-scale) mask with amplitude transmittance $t_0(x,y)$ in a plane σ just in front of a lens (figure 4.16) which is supposed to be larger than the mask. This is now uniformly illuminated by a normally incident, monochromatic plane wave of amplitude A, propagating along the $+z$ axis.

The field U incident on the lens is then

$$U_l\left(x,y\right) = At_0\left(x,y\right) \tag{4.117}$$

Formula (4.113) gives the field immediately behind the lens:

$$U_{l'}(x,y) = U_l(x,y) e^{-jkn\Delta_0} e^{j\frac{k}{2f}\left(x^2+y^2\right)} \tag{4.118}$$

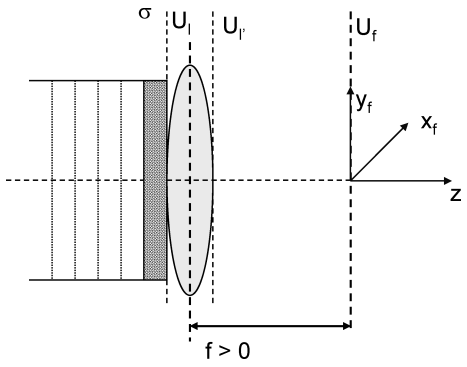


Figure 4.16: A mask placed against a lens

This field propagates further along the z-axis. The field at a distance $z = f$ (i.e. in the back focal plane) can be calculated with the Fresnel diffraction formula (4.42).

$$U_f(x_f, y_f) = -\frac{e^{-jkf}}{j\lambda f} e^{-j\frac{k}{2f}\left(x_f^2+y_f^2\right)} \int_{-\infty}^{+\infty}\int U_{l'}(x,y) e^{-j\frac{k}{2f}\left(x^2+y^2\right)} e^{j\frac{2\pi}{\lambda f}\left(xx_f+yy_f\right)} dxdy \tag{4.119}$$

After substituting U_l one finds

$$U_f(x_f, y_f) = -\frac{e^{-jkf}}{j\lambda f} e^{-j\frac{k}{2f}\left(x_f^2+y_f^2\right)} e^{-jkn\Delta_0} A \int_{-\infty}^{+\infty}\int t_0(x,y) e^{j\frac{2\pi}{\lambda f}\left(xx_f+yy_f\right)} dxdy \tag{4.120}$$

This implies that the field in the back focal plane is proportional to the two-dimensional Fourier transform of the transmittance function $t_0(x, y)$ of the object/mask (or, in general, the field incident on the lens), in which the spatial frequencies (f_x, f_y) and the positions in the focal plane (x_f, y_f) are related by:

$$\begin{aligned} f_x &= -\frac{x_f}{\lambda f} \\ f_y &= -\frac{y_f}{\lambda f} \end{aligned} \tag{4.121}$$

This Fourier-transform relation is not an exact one, because of the quadratic phase factor

$$e^{-j\frac{k}{2f}\left(x_f^2+y_f^2\right)} \tag{4.122}$$

which is not constant in the focal plane. However, this phase factor disappears when calculating the power distribution in the focal plane. Consequently the power distribution in the focal plane is exactly given by the *power spectrum* of the mask. This means that without a lens the far field (i.e. the plane wave decomposition) is found at large distance, but with the lens the far field is found in the lens focal plane (obviously with a different scaling factor).

4.3.4 Fourier-transform properties of a mask placed a distance in front of a lens

The same mask is now placed a distance d_0 in front of the lens and illuminated in the same way (figure 4.17).

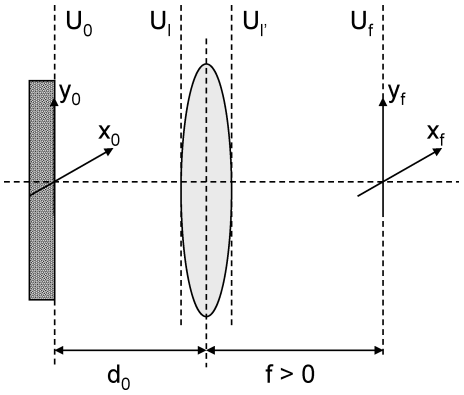


Figure 4.17: A mask at a distance d_o in front of a lens

We assume that d_o is large enough, so that we may use the Fresnel diffraction formula. This means that we can apply formulas (4.79) and (4.82).

Let us call

$$F_0(f_x, f_y) = F_{2D}\{At_0\} \tag{4.123}$$

and

$$F_l(f_x, f_y) = F\{U_l\} \tag{4.124}$$

then

$$F_l(f_x, f_y) = F_0(f_x, f_y) e^{j\pi\lambda d_0\left(f_x^2+f_y^2\right)} e^{-jkd_0} \tag{4.125}$$

Expression (4.120) can now be rewritten as

$$U_f(x_f, y_f) = -\frac{e^{-jkf}}{j\lambda f} e^{-jkn\Delta_0} e^{-j\frac{k}{2f}\left(x_f^2+y_f^2\right)} F_l\left(-\frac{x_f}{\lambda f}, -\frac{y_f}{\lambda f}\right) \tag{4.126}$$

Substituting (4.125) in (4.126) gives, after deleting a constant phase factor:

$$U_f(x_f, y_f) = -\frac{1}{j\lambda f} e^{-j\frac{k}{2f}\left(x_f^2+y_f^2\right)} e^{j\frac{\pi\lambda d_0}{\lambda^2 f^2}\left(x_f^2+y_f^2\right)} F_0\left(-\frac{x_f}{\lambda f}, -\frac{y_f}{\lambda f}\right) \tag{4.127}$$

or

$$U_f(x_f, y_f) = -\frac{A}{j\lambda f} e^{-j\frac{k}{2f}\left(1-\frac{d_0}{f}\right)\left(x_f^2+y_f^2\right)} \int_{-\infty}^{+\infty}\int t_0(x, y).e^{+j\frac{2\pi}{\lambda f}(xx_f+yy_f)} dxdy \tag{4.128}$$

By choosing $d_0 = f$ the exponential in front of the integral disappears, and we end up with an exact Fourier transform between the transmittance t_0 and the field in the focal plane. In these calculations we neglected the finite transverse dimensions of the lens; this is allowed whenever the mask is small as compared to the lens.

4.3.5 Optical convolution processor

It is possible to realize, in an optical set up, a convolution between two functions

$$\left| \int_{-\infty}^{+\infty} g(\xi, \eta)\, h(x-\xi, y-\eta)\, d\xi d\eta \right| \tag{4.129}$$

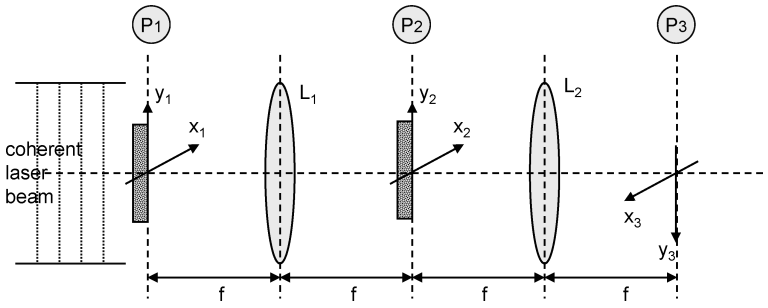


Figure 4.18: The optical convolution processor

The set up is shown in figure 4.18. The "input"-function g is realized as a mask with transmittance function $g(x_1, y_1)$ and placed in plane P_1, which is the first focal plane of a lens L_1. It is normally illuminated with a monochromatic plane wave. In the second focal plane P_2 of this lens one obtains the Fourier transform $k_1 G(-x_2/\lambda f, -y_2/\lambda f)$ of g (in which $k_1 =$ complex constant). In this plane P_2 one puts a second mask with transmittance function

$$t(x_2, y_2) = k_2 H\left(\frac{-x_2}{\lambda f}, \frac{-y_2}{\lambda f}\right) \tag{4.130}$$

with

$$H\left(f_x, f_y\right) = F_{2D}\left\{h\left(x, y\right)\right\} \tag{4.131}$$

Behind this mask, the field is proportional to $G.H$; hence in the focal plane P_3 of the second lens L_2 one finds the following irradiance:

$$I\left(x_3, y_3\right) = K\left|\int_{-\infty}^{+\infty} g\left(\xi, \eta\right) h\left(x_3 - \xi, y_3 - \eta\right) d\xi d\eta\right|^2 \tag{4.132}$$

This kind of convolution processor can be practically realized by inserting in the planes P_1 and P_2 a LC-SLM (Liquid Crystal Spatial Light Modulator). This is a two-dimensional transmission screen, with which one can realize an arbitrary two-dimensional function with the help of a computer. In plane P_3 one puts a CCD (Charged Coupled Device) Camera which transforms the optical information in an electrical one. An advantage of this processor is the fact that the "calculations" are done immediately (the speed is only limited by the input and output speed of the LC-SLM and CCD). Disadvantage (as compared to the calculations with a digital computer) is the analog character of the calculations, with its inaccuracy, the non-linearity in the LC-SLM and CCD, and also the aberrations in the lenses. Moreover there is also a technological problem: the function H is usually a complex function. This implies that in plane P_2 one needs an SLM in which not only the amplitude, but also the phase transmittance should be simulated electronically; and this is a very difficult technological problem.

4.4 Resolving power of an optical system

4.4.1 Introduction

In the previous section we always assumed that the lenses were unlimited in size; moreover we only calculated the field in the $focal$ plane. In this section we look for the field or image in an $image$ plane; moreover we take the finite dimension of the lens into account. A perfect imaging system has an infinite resolving power. In real systems, the resolving power is limited, mainly because of two reasons. First: geometrical aberrations limit the sharpness of the image, and consequently fine details get lost. In principle those aberrations can be reduced, almost as much as one wishes. The second reason is the presence of *diffraction*. Indeed, as every optical system has only finite transverse dimensions, one always finds a diffraction spot in the image. This spot can be reduced by increasing the diameter of the optical components, but this invariably increases the geometrical aberrations; hence it is not clear whether the overall quality of the image will improve. In this section we will focus on the resolution of a diffraction-limited system, i.e. a system free of abberations.

The resolving power is usually defined by considering two points in the object, and calculating the image of them, to see whether or not those images overlap. This method is similar to the analysis of linear systems with an impulse response technique: indeed, each object can be considered as being a collection of points; the image is then the collection of all image points or spots.

It turns out that the resolving power of an optical system differs whether the object is coherently or incoherently illuminated, because diffraction depends on the degree of coherence. For perfect coherent illumination, the image is build up by adding the *complex* amplitudes of all the diffraction patterns of all points of the object. With incoherent illumination on the other hand, one has to add all the irradiances of the image-points. It should be clear that the latter always gives smoother images: a complex amplitude can indeed be negative, giving ripples due to destructive interference. But it is less clear what influence this has on the resolving power of the optical system. It turns out that this depends on the actual optical system: depending on the phase relations in the object field, the coherent image can either look sharper or less sharp than the incoherent one! This is for example important in microscopy, where a correct choice of the illumination can dramatically increase the quality of the final image.

4.4.2 The point spread function of a diffraction-limited system

We first start with a simple set up: consider an object composed of one single point, and an optical system with one single aberration-free thin lens. What is the image? The image is described with an index i ("image": coordinates x_i, y_i); in the object plane we have an index o ("object": coordinates x_o, y_o). Due to the linearity of the wave-propagation phenomenon,we can always write:

$$U_i(x_i, y_i) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x_i, y_i; x_o, y_o)\, U_o(x_o, y_o)\, dx_o dy_o \tag{4.133}$$

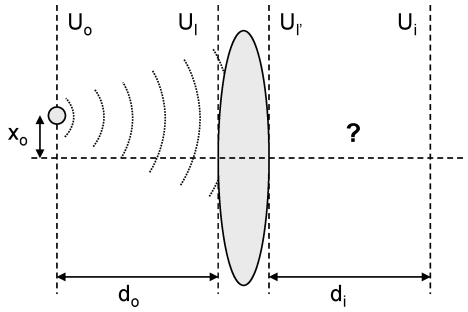


Figure 4.19: Image of a point source through a thin lens

for perfect coherent illumination, and

$$|U_{\,i}\,(x_i,y_i)|^2 = \kappa \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |h\,(x_i,y_i;x_o,y_o)|^2\,|U_o\,(x_o,y_o)|^2\,dx_o dy_o \tag{4.134}$$

for an incoherent object, with

$$\kappa = \frac{1}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} |h(0,0;x,y)|^2\,dxdy} \tag{4.135}$$

The function $h(x_i,y_i;x_o,y_o)$ is called the *point spread function* or PSF; it is the image of a unit-amplitude point source at position (x_o,y_o). This is the function we are looking for. The incident wave on the lens is a spherical wave, paraxially described by :

$$U_l\,(x_l,y_l;x_o,y_o) = \frac{-e^{-jkd_o}}{j\lambda d_o}\exp\left(-j\frac{k}{2d_o}\left[(x_l-x_o)^2+(y_l-y_o)^2\right]\right) \tag{4.136}$$

Behind the lens we have (apart from a constant phase factor)

$$U_{l'}\,(x_{l'},y_{l'};x_o,y_o) = U_l\,(x_{l'},y_{l'};x_o,y_o)\,P\,(x_{l'},y_{l'})\exp\left(j\frac{k}{2f}\left[x_{l'}^2+y_{l'}^2\right]\right) \tag{4.137}$$

The function $P(x,y)$ is the pupil function (it is 1 inside the lens, 0 outside).

Application of the Fresnel diffraction formula gives:

$$\begin{aligned} &h\,(x_i,y_i;x_o,y_o) = \\ &\quad\frac{-e^{-jkd_i}}{j\lambda d_i}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U_{l'}\,(x_{l'},y_{l'};x_o,y_o)\exp\left(-j\tfrac{k}{2d_i}\left[\left(x_i-x_{l'}\right)^2+\left(y_i-y_{l'}\right)^2\right]\right)dx_{l'}dy_{l'} \end{aligned} \tag{4.138}$$

Combining those expressions we find :

$$\begin{aligned} h\,(x_i,y_i;x_o,y_o) = \;& \frac{-e^{-jk(d_o+d_i)}}{\lambda^2 d_o d_i}\exp\left(-j\frac{k}{2d_i}\left[x_i^2+y_i^2\right]\right)\exp\left(-j\frac{k}{2d_o}\left[x_o^2+y_o^2\right]\right) \\ & \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P\,(x,y)\exp\left(-j\frac{k}{2}\left(\frac{1}{d_o}+\frac{1}{d_i}-\frac{1}{f}\right)\left(x^2+y^2\right)\right) \\ & \exp\left(+jk\left[\left(\frac{x_o}{d_o}+\frac{x_i}{d_i}\right)x+\left(\frac{y_o}{d_o}+\frac{y_i}{d_i}\right)y\right]\right)\,dxdy \end{aligned} \tag{4.139}$$

The quadratic phase terms in front of the integral can be neglected for object and image locations close to the axis and will be omitted hereafter; moreover we know that, in imaging:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f} \tag{4.140}$$

So we find:

$$\begin{aligned} &h\left(x_i, y_i; x_o, y_o\right) = \\ &\quad \frac{-e^{-jk(d_o+d_i)}}{\lambda^2 d_o d_i} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P\left(x,y\right) \exp\left(+jk\left[\left(\frac{x_o}{d_o}+\frac{x_i}{d_i}\right)x + \left(\frac{y_o}{d_o}+\frac{y_i}{d_i}\right)y\right]\right) dxdy \end{aligned} \tag{4.141}$$

Because the lateral magnification M of a thin lens equals $-d_i/d_o$, we can write:

$$\begin{aligned} &h\left(x_i, y_i; x_o, y_o\right) = \\ &\quad \frac{-e^{-jk(d_o+d_i)}}{\lambda^2 d_o d_i} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P\left(x,y\right) \exp\left(+j\frac{2\pi}{\lambda d_i}\left[\left(x_i - Mx_o\right)x + \left(y_i - My_o\right)y\right]\right) dxdy \end{aligned} \tag{4.142}$$

This implies that the point spread function is nothing else than the Fraunhofer diffraction pattern (= Fourier transform) of the exit pupil $P(x,y)$, centered on the image coordinates $x_i = Mx_o$ and $y_i = My_o$. Further more we see that the point spread function only depends on the combined variables $x_i - Mx_o$ and $y_i - My_o$. In other words, they depend on the relative position with respect to the ideal geometrical image point. Or in other words: if the point source in the object plane moves around, the resulting image will only shift and will not change its intensity distribution. If we now introduce new variables:

$$x_o^{'} = Mx_o \quad \text{and} \quad y_o^{'} = My_o \tag{4.143}$$

and

$$x' = \frac{x}{\lambda d_i} \quad \text{and} \quad y' = \frac{y}{\lambda d_i} \tag{4.144}$$

then we finally find

$$\begin{aligned} &h\left(x_i, y_i; x_o^{'}, y_o^{'}\right) = h\left(x_i - x_o^{'}, y_i - y_o^{'}; 0, 0\right) \triangleq h\left(x_i - x_o^{'}, y_i - y_o^{'}\right) \\ &\quad -e^{-jk(d_o+d_i)}\frac{d_i}{d_o}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P\left(\lambda d_i x^{'}, \lambda d_i y^{'}\right) \exp\left(+j2\pi\left[\left(x_i - x_o^{'}\right)x^{'} + \left(y_i - y_o^{'}\right)y^{'}\right]\right) dx^{'}dy^{'} \end{aligned} \tag{4.145}$$

Because this point spread function is only dependent on $x_i - x_o^{'}$ and $y_i - y_o^{'}$, the object-to-image transfer is space-invariant and the integral in (4.133) turns out to be a convolution:

$$U_i\left(x_i, y_i\right) = \frac{1}{M^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h\left(x_i - x_o^{'}, y_i - y_o^{'}\right) U_o\left(\frac{x_o^{'}}{M}, \frac{y_o^{'}}{M}\right) dx_o^{'}dy_o^{'} \tag{4.146}$$

If we now define:

$$h' = \frac{1}{M^2}h \tag{4.147}$$

then we obtain:

$$U_i\left(x_i, y_i\right) = h'\left(x_i, y_i\right) * U_o\left(\frac{x_i}{M}, \frac{y_i}{M}\right) \tag{4.148}$$

The function $U_o\left(\frac{x_i}{M}, \frac{y_i}{M}\right)$ is the *perfect* image, as found in paraxial geometrical optics. The actual image is the convolution of this perfect geometrical image with the function h', given by:

$$h'(x_i, y_i) = e^{-jk(d_o+d_i)} \frac{1}{M} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} P\left(\lambda d_i x', \lambda d_i y'\right) \exp\left(j2\pi \left[x_i x' + y_i y'\right]\right) dx' dy' \tag{4.149}$$

This is in fact the inverse Fourier transform of the pupil function. The factor $1/M$ in this formula shows that the image, which is M times larger than the object (a consequence of formula 4.148), is also M times weaker; the power concentration in two dimensions is consequently M^2 times weaker. This is of course consistent with the law of conservation of energy.

In the previous calculations we considered a very simple optical system: namely one single thin lens. More complicated optical systems can always be reduced to a single thick lens, at least in the paraxial approximation, when they are aberration-free. It can then be shown that the main conclusions remain valid, on condition that the aperture is chosen to be the exit pupil of the system. When the pupil $P(x, y)$ is a slit, then h' is a sinc-function. Mostly, however, the aperture is circular; then $h'(x, y)$ is the 2-dimensional Fourier transform of this function, which gives the well known Airy-function for $|h'|^2$. Consequently:

$$\left|h'(r)\right|^2 = \left(\frac{kD^2}{8d_i}\right)^2 \left[2\frac{J_1\left(kDr/2d_i\right)}{kDr/2d_i}\right]^2 \tag{4.150}$$

In this formula D is the diameter of the exit pupil, d_i is the distance between this pupil and the image plane, k is the wavenumber and r is the radial coordinate in the image plane. The first zero of the Airy pattern lies at a distance

$$r = 1.22\frac{\lambda d_i}{D} \tag{4.151}$$

There exist different conventions for defining the resolving power of an optical system. One of the possibilities is the so called *Rayleigh criterion*, according to which two points are just resolved when the center of one of the Airy patterns coincides with the first zero of the other one. This implies that the expression above gives the distance in the image plane between two just-resolved points. The distance in object plane can then be calculated using the transverse magnification M of the optical system.

One also knows that the NA (= Numerical Aperture) of a lens, as seen from object space, is approximated by

$$NA_o = \frac{D}{2d_o} \tag{4.152}$$

The corresponding NA in image space is then

$$NA_i = \frac{D}{2d_i} \tag{4.153}$$

and consequently:

$$\begin{aligned} \text{resolution in object plane} &= 0.61\frac{\lambda}{NA_o} \approx \frac{\lambda}{NA_o} \\ \text{resolution in image plane} &= 0.61\frac{\lambda}{NA_i} \approx \frac{\lambda}{NA_i} \end{aligned} \tag{4.154}$$

For a telescope (in which the object lies at $-\infty$), it is better to work with the $angular$ resolution in object space. It is:

$$\text{angular resolution} = 1.22\frac{\lambda}{D} \approx \frac{\lambda}{D} \tag{4.155}$$

All those formulas clearly show that a good resolution is only possible when using optical systems with a large diameter D.

4.4.3 Frequency response of a diffraction-limited system

By taking advantage of the convolution theorem, it is also possible to express the resolution by means of a transfer function concept in the spatial frequency domain. Indeed, from

$$U_i(x_i, y_i) = h'(x_i, y_i) * U_o\left(\frac{x_i}{M}, \frac{y_i}{M}\right) \text{ (coherent)} \tag{4.156}$$

and

$$|U_i(x_i, y_i)|^2 = \kappa \left|h'(x_i, y_i)\right|^2 * \left|U_o\left(\frac{x_i}{M}, \frac{y_i}{M}\right)\right|^2 \text{ (incoherent)} \tag{4.157}$$

one concludes, after Fourier transformation:

$$U_i^F(f_x, f_y) = M^2 H(f_x, f_y)\, U_o^F(Mf_x, Mf_y) \text{ (coherent)} \tag{4.158}$$

and

$$I_i^F(f_x, f_y) = M^2 O(f_x, f_y)\, I_o^F(Mf_x, Mf_y) \text{ (incoherent)} \tag{4.159}$$

In these formulas the functions U_o^F and H are, in the coherent case, the Fourier transforms of the functions U_o and $h^{'}$; whereas in the incoherent case I_o^F and O are the Fourier transforms of $|U_o|^2$ and $\kappa|h^{'}|^2$ respectively. $H(f_x)$ is called the coherent transfer function, and $O(f_x, f_y)$ is called the OTF or the optical transfer function. The absolute value of $O(f_x, f_y)$ is the *modulation transfer function* (MTF), and its phase is the phase transfer function. One can prove the following general relations:

$$\begin{aligned} &|O(f_x, f_y)| \leq |O(0,0)| = 1 \\ &O(f_x, f_y) = O(-f_x, -f_y) \end{aligned} \tag{4.160}$$

With the autocorrelation theory one can further prove that:

$$O(f_x, f_y) = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} H(f_x^{'}, f_y^{'})\, H^*(f_x^{'} - f_x, f_y^{'} - f_y) df_x^{'} df_y^{'}}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left|H(f_x^{'}, f_y^{'})\right|^2 df_x^{'} df_y^{'}} \tag{4.161}$$

in which H is the Fourier transform of $h^{'}$. A simple change in variables finally gives:

$$O(f_x, f_y) = \frac{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} H(f_x^{'} + f_x/2, f_y^{'} + f_y/2)\, H^*(f_x^{'} - f_x/2, f_y^{'} - f_y/2) df_x^{'} df_y^{'}}{\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \left|H(f_x^{'}, f_y^{'})\right|^2 df_x^{'} df_y^{'}} \tag{4.162}$$

This formula shows the relation between the coherent transfer function $H(f_x, f_y)$ and the incoherent optical transfer function $O(f_x, f_y)$.

There is a simple relation between the coherent transfer function H and the pupil function P. This relation can be found by comparing the inverse fourier transform of $h^{'}$

$$h'(x_i, y_i) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} H(f_x, f_y) \exp\left(j2\pi\left[x_i f_x + y_i f_y\right]\right) df_x df_y \tag{4.163}$$

with formula 4.149 to find that (the constant phase term can be neglected.) :

$$H(f_x, f_y) = \frac{1}{M} P(\lambda d_i f_x, \lambda d_i f_y) \tag{4.164}$$

Consequently the coherent transfer function is proportional to the pupil function itself (after an appropriate change of variables). The incoherent transfer function is then

$$O(f_x, f_y) = \frac{\int_{-\infty}^{+\infty} P\left(x' + \frac{\lambda d_i f_x}{2}, y' + \frac{\lambda d_i f_y}{2}\right) P\left(x' - \frac{\lambda d_i f_x}{2}, y' - \frac{\lambda d_i f_y}{2}\right) dx'dy'}{\int_{-\infty}^{+\infty} |P(x', y')|^2 dx'dy'} \tag{4.165}$$

The pupil function equals 1 inside and 0 outside the aperture. Consequently the denominator in the expression above is nothing else than the area of the exit pupil. The numerator is the area of overlap between the aperture and its own shifted image, after shifting over $\lambda d_i f_x$ in the x- and $\lambda d_i f_y$ in the y-direction. This is shown in figure 4.20. From this geometrical interpretation it follows that the OTF of an aberration-free system is always real and non-negative (and hence the OTF equals the MTF). The function is not necessarily monotonically decreasing.

Let us now have a closer look at the coherent and incoherent transfer function for a diffraction-limited optical system with a square resp. circular exit pupil. The square pupil has a pupil function:

$$P(x, y) = \mathrm{rect}\left(\frac{x}{D}\right) \mathrm{rect}\left(\frac{y}{D}\right) \tag{4.166}$$

Which gives a coherent transfer function:

$$H(f_x, f_y) = \mathrm{rect}\left(\frac{\lambda d_i f_x}{D}\right) \mathrm{rect}\left(\frac{\lambda d_i f_y}{D}\right) \tag{4.167}$$

This has cutoff frequencies in x- and y-direction given by:

$$f_c = \frac{D}{2\lambda d_i} \tag{4.168}$$

The incoherent transfer function is easy to calculate. One obtains:

$$O(f_x, f_y) = \Lambda\left(\frac{f_x}{2f_c}\right) \Lambda\left(\frac{f_y}{2f_c}\right) \tag{4.169}$$

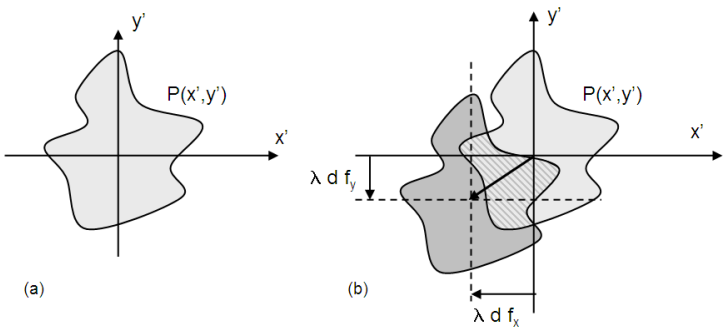


Figure 4.20: Calculation of the overlap of an aperture with itself

Λ is the triangle function (with a value of 1 when the argument equals 0, 0 when the argument is larger than or equal to 1, and linear in between).

The circular aperture gives:

$$\begin{aligned} P(x,y) &= \text{circ}\left(\frac{\sqrt{x^2+y^2}}{D/2}\right) \\ H(f_x,f_y) &= \text{circ}\left(\frac{2\lambda d_i\sqrt{f_x^2+f_y^2}}{D}\right) \\ f_c &= \frac{D}{2\lambda d_i} \end{aligned} \tag{4.170}$$

The OTF is, once again, easy to present in a graph, but rather difficult to calculate analytically. One finds:

$$\begin{aligned} O(f) &= \frac{2}{\pi}\left[\arccos\left(\tfrac{f}{2f_c}\right) - \tfrac{f}{2f_c}\sqrt{1-\left(\tfrac{f}{2f_c}\right)^2}\right] \quad \text{when } f \le 2f_c \\ O(f) &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \text{when } f \ge 2f_c \end{aligned} \tag{4.171}$$

with $f = \sqrt{f_x^2+f_y^2}$.

From 4.169 and 4.171 one can see that both in the case of a square and of a circular exit pupil, the MTF is zero for spatial frequencies beyond $2f_c$. This means that the intensity distribution in the image plane cannot contain spatial frequency components beyond $2f_c$. In other words if the intensity distribution in the object plane contains spatial frequency components beyond $2f_c/M$, these rapid spatial variations will not be resolved in the image plane. This is another way to phrase the diffraction limit. It is worth comparing the Rayleigh criterion with the limit caused by the cut-off frequency. Equation 4.151 states that two object points can be resolved if they are spaced by more than $1.22\lambda d_o/D$ (the corresponding imagepoints are spaced by $1.22\lambda d_i/D$). The highest spatial frequency that is resolved is given by $\frac{D}{\lambda d_o}$ in the object plane and $\frac{D}{\lambda d_i}$ in the image plane. The corresponding periods are $\frac{\lambda d_o}{D}$ and $\frac{\lambda d_i}{D}$. Clearly the Rayleigh criterion and the cut-off frequency limit are consistent with each other.

Figure 4.23 shows both OTF's as function of the f-coordinate. It is interesting to compare the maximum frequency that can be resolved by the system in the coherent and incoherent case. In the coherent case (equation (4.167)), frequencies up to f_c can be resolved. In the incoherent case

(equation (4.169)), the maximum resolvable frequency equals $2f_c$! To understand this intuitively, consider figure 4.21. As explained in 4.2.8, fields can be decomposed into plane waves, where each spatial frequency corresponds to one single direction of propagation. Since the lens has finite

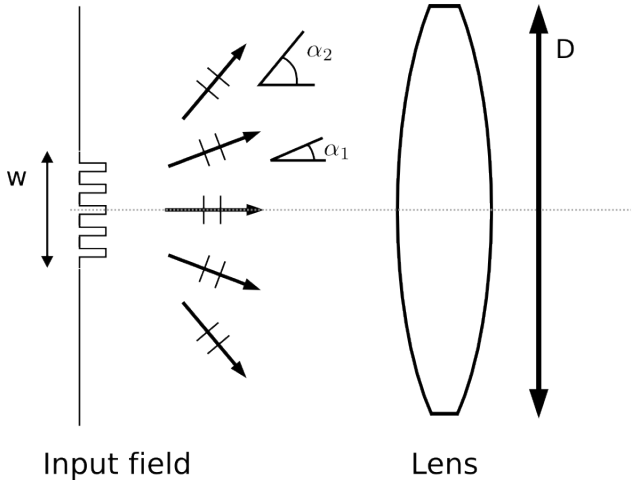


Figure 4.21: Decomposition of a periodic image into plane waves. The lens has a finite diameter D.

dimensions, it cannot capture all incoming angles (in figure 4.21 for example, α_2 is not imaged anymore). When even the angle α_1 is no more captured by the lens, we have reached the cut-off frequency, and only the zeroth-order component is captured. We also saw that incoherent light can be considered as a superposition of coherent contributions, whereby all possible phase relationships between different parts of the field should be taken into account. If we now look at the intensity distribution depicted in 4.22, and consider two specific field distributions -as shown in the figure- from the ensemble of all possible field distributions with the same intensity distribution, we see that, from a coherent point of view, the frequency of the second field is only half the frequency of the original intensity profile. If the frequency is halved, the angles are halved, and as such more frequencies are captured by the lens. In other words, an intensity distribution with frequency f can be considered to be the result of the superposition of an ensemble of field distributions, including a field distribution with frequency f and another with frequency f/2. We showed the two most extreme cases, all other contributions lie somewhere in between.

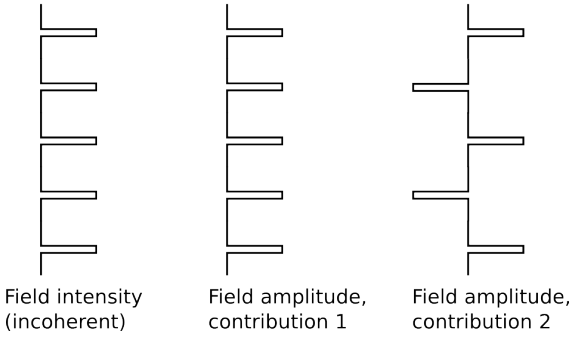


Figure 4.22: The incoherent field can be seen as a sum of coherent contributions. The two pictures on the right show two such contributions. In the second case, the odd rows are phase-shifted by π.

4.4.4 The effect of abberations

In the discussion above we completely neglected possible geometrical aberrations in the optical system. In principle it is not allowed to use the concept of transfer function when there are aberrations, simply because the optical system is not space invariant anymore; at best, one could use a "local" transfer function. Nevertheless, in practical situations, one still continues to use them. The MTF of an optical system with aberrations is always worse than in a diffraction-limited system without aberrations: for each spatial frequency the MTF is lower than in an aberration-free system. This is because the wavefront leaving the optical system is no longer a spherical one, but shows phase aberrations, which implies that the OTF now becomes complex. We can model this by assuming the pupil is still illuminated by a perfect spherical wave, but that a plate inside the aperture deforms the phase. The plate induces an effective path-length error of $kW(x,y)$. The pupil function now becomes complex, and is referred to as the generalized pupil function:

$$P_{gen}(x,y) = P(x,y)e^{jkW(x,y)} \tag{4.172}$$

To illustrate this, we consider the simple case of a defocussing error ϵ:

$$\frac{1}{d_i} + \frac{1}{d_o} - \frac{1}{f} = \epsilon, \tag{4.173}$$

If we look back to (4.139), and keeping in mind the new generalized pupil function, W(x,y) can easily be found to be:

$$W(x,y) = \frac{\epsilon\left(x^2+y^2\right)}{2} \tag{4.174}$$

After some math, one finds the OTF for the system with defocussing ϵ and a rectangular aperture of width D:

$$\begin{aligned} O\left(f_x,f_y\right) &= \Lambda\left(\frac{f_x}{2f_c}\right)\Lambda\left(\frac{f_y}{2f_c}\right) \\ &\quad \operatorname{sinc}\left[\frac{\epsilon D^2}{\lambda}\left(\frac{f_x}{2f_c}\right)\left(1-\frac{|f_x|}{2f_c}\right)\right]\operatorname{sinc}\left[\frac{\epsilon D^2}{\lambda}\left(\frac{f_y}{2f_c}\right)\left(1-\frac{|f_y|}{2f_c}\right)\right] \end{aligned} \tag{4.175}$$

For ϵ=0, this equation simplifies to (4.169). Note that the OTF is real. More in general, if $W(x,y)$ is centrosymmetric, the OTF will be real (this can easily be seen from the definition of the OTF).

As one can see, for some frequencies, the MTF can become zero or even negative. Those frequencies give a good resolving power, but the image is reversed (meaning: light and dark are inverted). At the zeros of the MTF there is no resolving power. This is illustrated in figure 4.24 and figure 4.25.

Figure 4.24 shows on the left a test pattern which is presented to the optical system; on the right its image. One clearly sees the part with a reversed image.

Figure 4.25 shows the MTF of an optical system with rectangular aperture for different degrees of aberrations; these aberrations were realized by displacing the image plane slightly with respect to the correct position. Again one can see the zeros and the negative parts of the MTF.

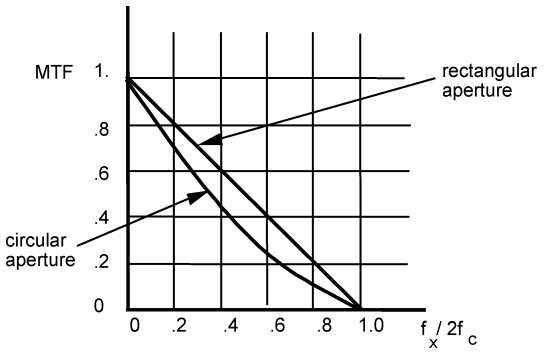


Figure 4.23: MTF of a rectangular and circular aperture.

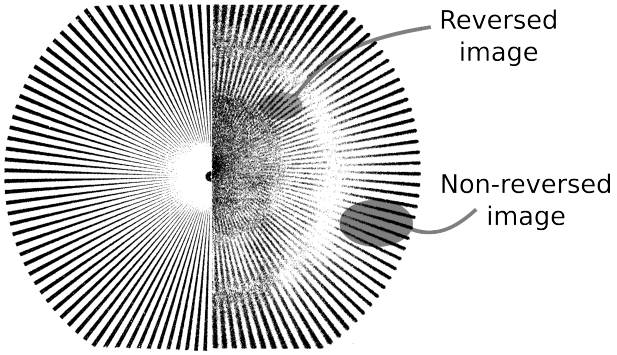


Figure 4.24: Radial test pattern (left) and image through an optical system (right)

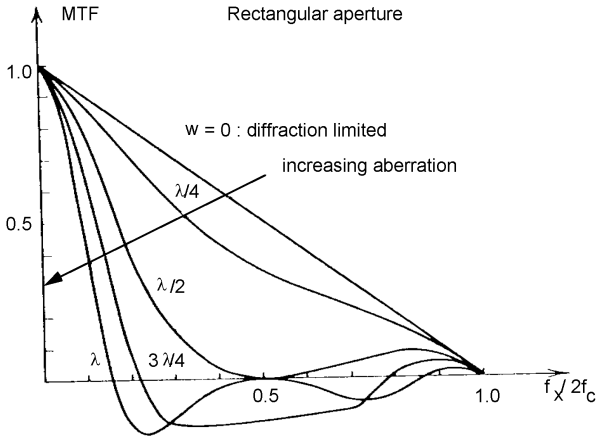


Figure 4.25: MTF of a rectangular aperture with aberrations

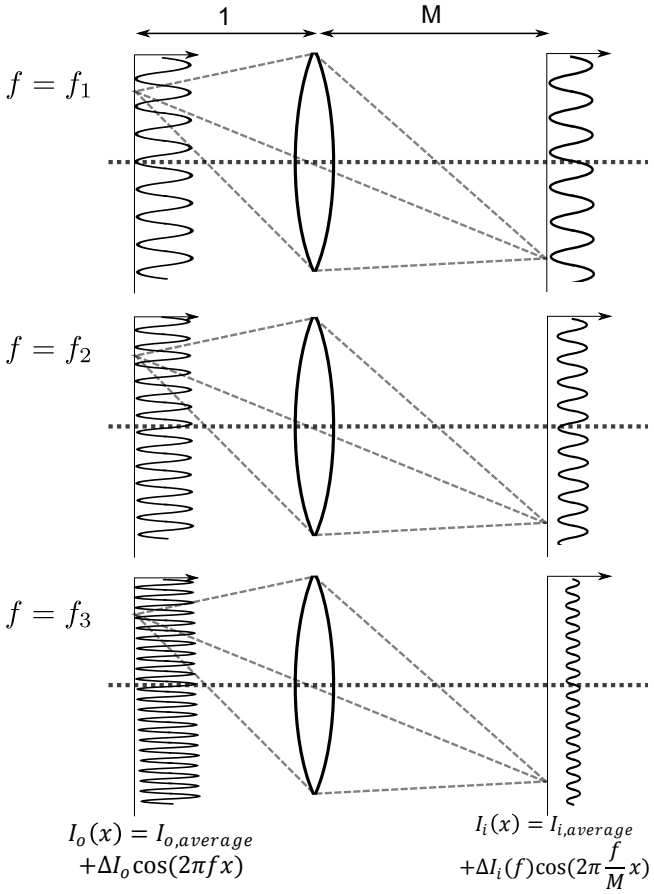


Figure 4.26: Measuring the MTF. As the spatial frequency increases, the transmission decreases.

4.4.5 Measuring the MTF of a lens system

The MTF of an optical system can be measured by incoherently illuminating a mask on which a sinusoidal transmittance pattern is written. This pattern has only one single spatial frequency (the function $I_o^F(f_x, f_y)$ is a Dirac-distribution). One then measures the amplitude of the sinusoidal image; this gives already 1 point of the MTF. Repeating this procedure for different spatial frequencies finally gives the complete MTF.

More specifically consider figure 4.26. In the object plane one places a mask with a sinusoidal transparency with spatial frequency f_1. This mask is illuminated with incoherent light. The lens system (to be characterized) images this mask into the image plane. The image will be a simusoidal intensity pattern with spatial frequency f_1/M. The MTF for this frequency is then given by the ratio of the (relative) modulation depth of the intensity pattern in the image plane divided by that of the object plane. This procedure can be repeated for other masks with other spatial frequencies. Together all the data form the MTF as a function of frequency:

$$MTF(f) = \frac{\Delta I_i(f)/I_{i,average}}{\Delta I_o/I_{o,average}} \tag{4.176}$$

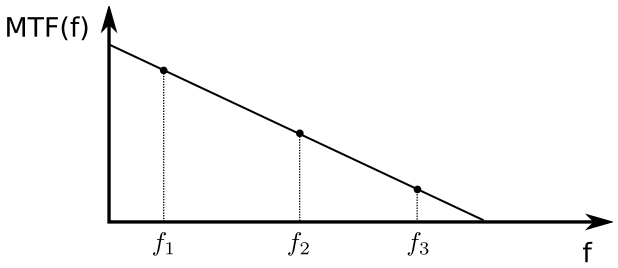


Figure 4.27: Measuring the MTF. As the spatial frequency increases, the transmission decreases.

Chapter 5

Dielectric waveguides

Contents

5.1 Introduction

During the last decade, the application of complex optical systems has enormously grown. These systems are applied in such different areas as optical fiber telecommunication, optical datacom, optical data storage (CD-ROM,DVD, Blue-Ray), sensors, printers and many more. Classical optical systems consist of a collection of separate optical components (lenses, mirrors, diffractive elements, light sources, light detectors), which are carefully assembled together. Typically all components need to be aligned very accurately with respect to each other, which makes these systems large, less robust and expensive. In the same way as electronics evolved from discrete components on printed circuit boards to monolithic integrated circuits, also in optics a miniaturization and integration process is ongoing however. The idea of integrated optics was introduced in the late 60s and comprises the integration of different optical functionalities onto a single substrate. To route the light through the components, optical waveguides are used instead of free space propagation. In the early days of integrated optics, most activities were focused on the development of single components, both passive (couplers, filters) and active (lasers, detectors). Later focus shifted towards bringing these different functions together on a single chip. This led to state-of-the-art components consisting of a combination of different complicated subcomponents. A typical example demonstrating the strength of optical integration is shown in figure 5.1. The left picture shows a fiber-based 4-channel 2x2 cross-connect module while the right picture shows an integrated device with the same functionality. It is obvious that the latter has a size several orders

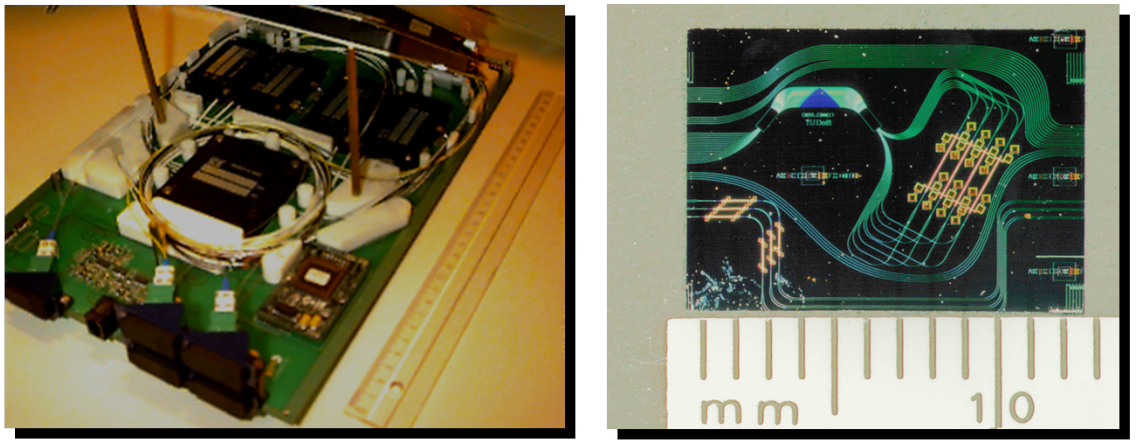


Figure 5.1: Left: compact fiber-based 4-channel 2x2 cross-connect module (Telefonica I&D, Madrid). Right: Photonic integrated 4-channel 2x2 cross-connect module (Cobra Institute, TU/e)

of magnitude smaller than the discrete device. This type of complex integrated components is currently in a research phase but there is a rapid evolution towards commercial applications.

The fundamental idea behind integrated optics is the manipulation of light by waveguides and not by free space optical components like lenses and mirrors. The optical field is guided by dielectric waveguide structures, which is possible because light prefers to be concentrated in the area with the highest refractive index. Figure 5.2 represents different types of waveguides that are used in integrated optics. The optical field will always be located in the area with the highest average refractive index.

The depicted waveguide structures can be realized in different material systems. Each material system has its own advantages and drawbacks. So for every specific application one will have to make a well-considered decision for a given material system. Below the properties of some important material systems are listed.

- InGaAsP/InP

  This crystalline semiconductor material allows for monolithic integration, being the integration of laser diodes and photodiodes together with passive components. This can be done in the $1.3\mu m$ and $1.5\mu m$ wavelength range where the optical fiber has the lowest loss. Waveguides are formed by epitaxial layer growth on an InP substrate and by etching. Due to the high refractive index and the high index contrasts, waveguide structures typically are small (order 1-$2\mu m$), which can lead to problems when coupling light from an optical fiber into the waveguide.

- AlGaAs/GaAs

  This crystalline semiconductor material allows for monolithic integration in the $0.8\mu m$ wavelength range. The corresponding components are mostly used for short distance communi-

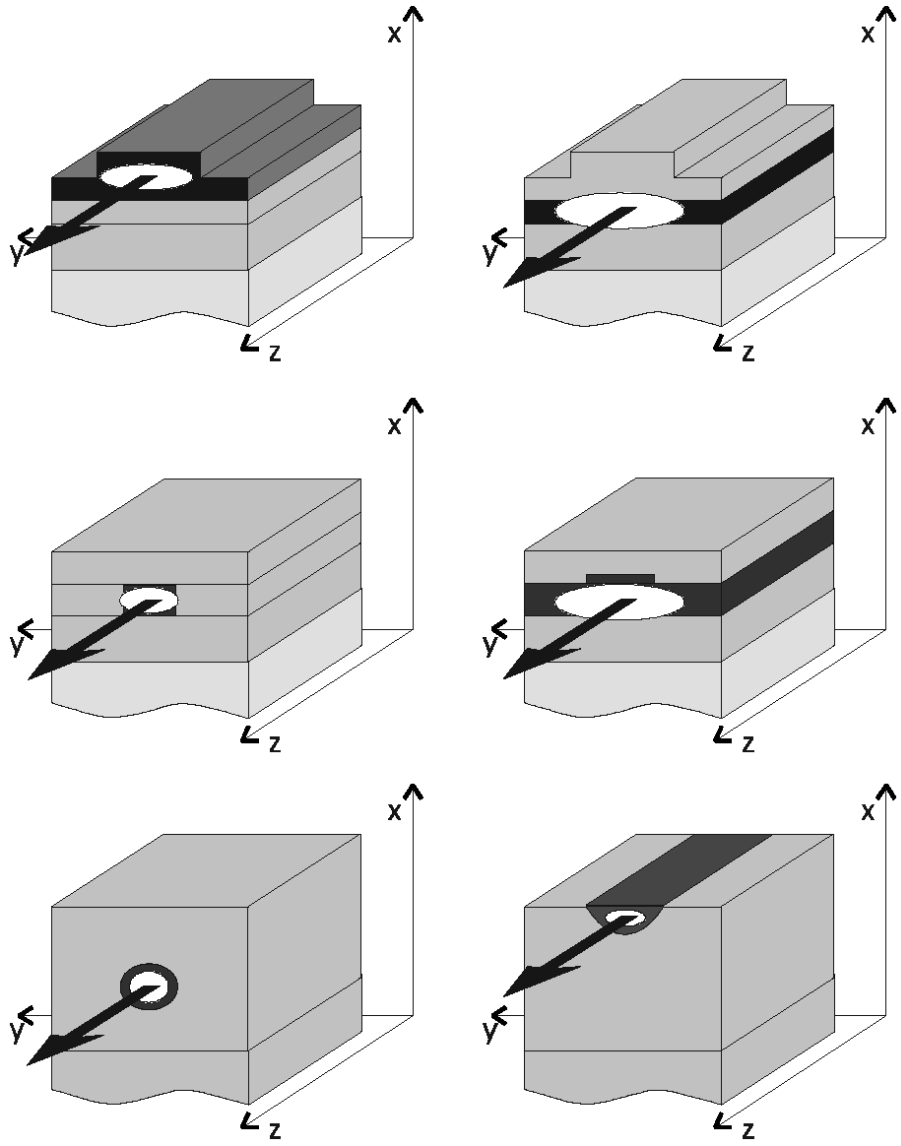


Figure 5.2: Different waveguide structures

cation, in scanners and in CD-players. The same remarks as in the InP/InGaAsP material system concerning coupling to an optical fiber hold.

- Glass

  Waveguides in glass are fabricated using two different types of technologies: diffusion and deposition. The first technique is based on the ion exchange process in special types of glasses by pure thermal or field assisted diffusion. The second technique is based on the deposition of glass (Silica or SiO2) by means of chemical vapour deposition (CVD) or flame hydrolysis (FHD) on a substrate, mostly Silicon. The large waveguide cross-section allows easy coupling with optical fibers. These are currently the preferred methods for fabricating passive optical integrated circuits. Because of the low index contrast, devices are relatively big however.

- $LiNbO_3$

  Lithium niobate is an anisotropic crystalline material with strong electro-optical and acousto-optical properties. Waveguides are fabricated using a diffusion process. Due to the anisotropy components always show a large polarization dependence. The electro-optical effect is used to realize efficient optical switches.

- Polymers

  Polymers represent a broad set of materials. For purely passive applications polycarbonate (also used for compact disks) and PMMA are the preferred materials. Other types of polymers show a large electro-optical or non-linear coefficient but this is mostly at the expense of a reduced long term stability.

- Silicon-on-Insulator

  Waveguides are fabricated in a Silicon layer (high refractive index), which is bonded on a Silicon wafer via an intermediary silica layer (low refractive index). SOI waveguides exist in two different varieties with very different properties: "fiber matched" (typical dimensions around $7\mu m$) and high-contrast (typical dimensions 500nm). The advantage of using Silicon-on-Insulator is that standard CMOS technology can be used to fabricate the waveguides. Even CMOS-electronics and optical circuits can be combined on the same substrate. The high-contrast waveguides open the way to very large scale integrated optics.

For a long time, III-V technologies were thought to push aside other technologies because they were the only ones allowing for monolithic integration. But due to the complicated fabrication process, the high cost, the high fiber to chip coupling losses and the high propagation losses of III-V waveguides, different material systems tend to co-exist, each used for their specific application. The chart below gives an overview of the most important properties of the materials discussed here (T: Telecom / I: Interconnect).

	InP	GaAs	Glass	LiNbO3	Polymers	SOI
Transmitters/Receivers	Yes	Yes	No	No	No	No
Passive optics	Yes	Yes	Yes	Yes	Yes	Yes
Wavelength range	T	I	T,I	T,I	T,I	T
Fiber to chip coupling loss	>2	>2	0.4	<1	<0.5	1
Propagation losses (dB/cm)	2	2	<0.1	<0.3	0.1-1.5	2
$\mathrm{d}n/\mathrm{d}T$ [10^{-4}/K]	1.8	2.5	<0.1	<0.1	0.1-1.5	1.7
$\lvert n_{\mathrm{TE}} - n_{\mathrm{TM}} \rvert$ [10^{-4}]	0.1-10	2-10	0.1-0.5	400	2-50	0.1-...
long term stability	+	+	+	+	?	+
wafer dimensions	2-3 inch	3 inch	$\geq$ 4 inch	3 inch	$\geq$ 8 inch	8-12 inch

Fiber to chip coupling losses can be improved by the use of tapers.

5.2 Modes of Optical Waveguides

5.2.1 Introduction

The starting point for the theoretical treatment of the interaction of light and dielectric structures are Maxwells equations. We only assume isotropic and non-magnetic media and a harmonic time dependence of $e^{j\omega t}$. In this case, and in the absence of sources and currents, we can write Maxwells equations as (see also chapter 2):

$$\begin{aligned} \nabla \times \mathbf{E}(x,y,z) &= -j\omega\mu_0 \mathbf{H}(x,y,z) \\ \nabla \times \mathbf{H}(x,y,z) &= j\omega\varepsilon(x,y,z)\mathbf{E}(x,y,z) \\ \nabla \cdot (\varepsilon(x,y,z)\mathbf{E}(x,y,z)) &= 0 \\ \nabla \cdot \mathbf{H}(x,y,z) &= 0 \end{aligned} \tag{5.1}$$

In these equations the constitutive relations

$$\begin{aligned} \mathbf{D} &= \varepsilon\mathbf{E} = \varepsilon_0 n^2 \mathbf{E} \\ \mathbf{B} &= \mu_0 \mathbf{H} \end{aligned} \tag{5.2}$$

are implicitly assumed. The real part of the refractive index profile $n(x,y,z,\omega)$ relates the wavelength inside the medium $\lambda = \frac{\lambda_0}{\mathrm{Re}(n)}$ and the vacuum wavelength λ_0. The imaginary part of the refractive index describes the absorption (or gain) of the optical field.

Based on the Maxwells curl equations, vectorial wave equations for the electric field **E** and magnetic field **H** can be written as

$$\begin{aligned} \nabla^2\mathbf{E}(\mathbf{r}) + \nabla\left(\tfrac{\nabla n^2(\mathbf{r})}{n^2(\mathbf{r})}\mathbf{E}(\mathbf{r})\right) + k_0^2 n^2(\mathbf{r})\mathbf{E}(\mathbf{r}) &= 0 \\ \nabla^2\mathbf{H}(\mathbf{r}) + \tfrac{\nabla n^2(\mathbf{r})}{n^2(\mathbf{r})} \times (\nabla \times \mathbf{H}(\mathbf{r})) + k_0^2 n^2(\mathbf{r})\mathbf{H}(\mathbf{r}) &= 0 \end{aligned} \tag{5.3}$$

In these equations the gradient of the refractive index occurs, which couples the three components of the field vector. When the refractive index is piecewise constant however, or if the variation of the refractive index is small, we can neglect these gradients so both vectorial equations (5.3)

decouple and reduce to the Helmholtz equation for every component of the field vector (both electric and magnetic field)

$$\nabla^2\Psi(\mathbf{r}) + k_0^2 n^2(\mathbf{r})\Psi(\mathbf{r}) = 0 \tag{5.4}$$

For the further study of optical waveguides, the boundary conditions at the interface between two isotropic materials with dielectric constants ε_1 and ε_2 are important. These are:

$$\begin{aligned} &n \times (\mathbf{E}_1 - \mathbf{E}_2) = 0 \\ &n \times (\mathbf{H}_1 - \mathbf{H}_2) = 0 \\ &n \cdot (\varepsilon_1\mathbf{E}_1 - \varepsilon_2\mathbf{E}_2) = 0 \\ &n \cdot (\mathbf{H}_1 - \mathbf{H}_2) = 0 \end{aligned} \tag{5.5}$$

, which means that the tangential components of the electric and the magnetic field and the normal component of the magnetic field are continuous at an interface. The normal component of the electric field is discontinuous at an interface.

By applying these boundary conditions the field components will in general be related (although they seemed to be uncoupled by neglecting the refractive index gradient terms in equation (5.3)). Note that it is often sufficient to apply the boundary conditions for the tangential components because then automatically the boundary conditions for the normal components are met.

The general solution of Maxwells equations for an arbitrary dielectric structure $\varepsilon(x, y, z)$ requires the solution of a complex set of partial differential equations and requires a lot of computation power. Therefore, in the early days of integrated optics, a lot of effort was put in the development of acceptable approximated calculation methods. A typical example is the effective index method. As more powerful computers became available, numerical methods like finite differences and finite elements methods were used for the analysis of waveguide structures. Nevertheless, approximated solutions in general and the effective index method in particular remain very important design and modeling tools. Some of the methods will be described later in this chapter.

5.2.2 Modes of longitudinally invariant dielectric waveguide structures

In this section we will consider waveguide structures that are invariant in the propagation direction of the optical power. A typical example is shown in figure 5.3. When we choose the z-direction as the propagation direction we can write the refractive index profile as $n(\mathbf{r}) = n(x, y)$

An eigenmode of the waveguide structure is a propagating or evanescent wave of which the transversal shape does not change during propagation. An eigenmode propagating in the positive z-direction is represented by

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= \mathbf{e}(x, y)e^{-j\beta z} \\ \mathbf{H}(\mathbf{r}) &= \mathbf{h}(x, y)e^{-j\beta z} \end{aligned} \tag{5.6}$$

Three different parameters can be used to describe the propagation characteristics of the eigenmode. The first parameter is the propagation constant β, the second is the effective refractive index

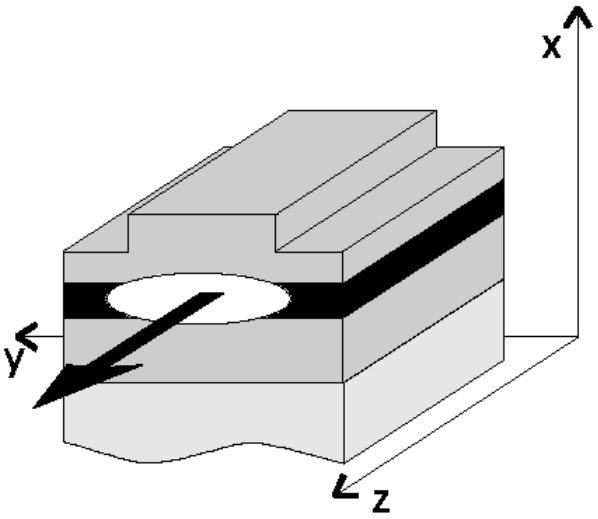


Figure 5.3: Longitudinally invariant waveguide

$$n_{eff} = \frac{\beta}{k_0} \tag{5.7}$$

and the third parameter is the effective dielectric constant

$$\varepsilon_{eff} = n_{eff}^2 \tag{5.8}$$

In the following section we will show that this is the eigenvalue of the eigenvalue equation resulting from Maxwells equations and for which the eigenmodes are the solutions we are looking for. Before we continue with a detailed analysis of the eigenvalue problem, we will first list some important properties of lossless optical waveguides, namely waveguides for which

$$\mathrm{Im}(\varepsilon(x, y)) = 0 \tag{5.9}$$

Figure 5.4 shows the dielectric profile of a hypothetic waveguide together with different eigenmodes of the structure. Theoretically one can show that

1. There are no eigenmodes with an eigenvalue larger than the maximum of the dielectric function.

$$\varepsilon_{eff} < \max(\varepsilon(\mathbf{r}_t)) \tag{5.10}$$

2. Guided modes belong to a discrete set of eigenvalues. These are in the range

$$\varepsilon_{\max} > \varepsilon_{eff} > \max(\varepsilon_{clad}) \tag{5.11}$$

   For these modes

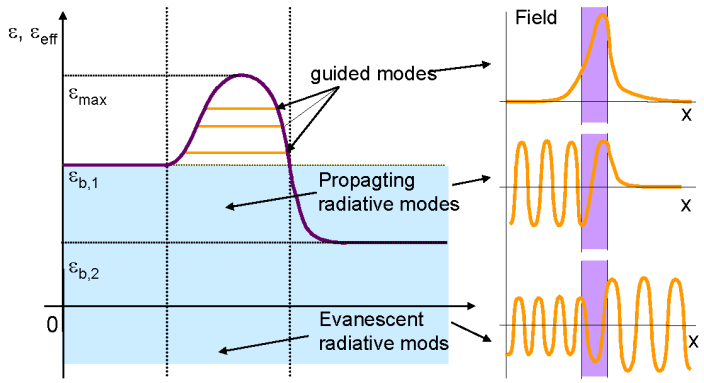


Figure 5.4: Eigenmodes in an optical waveguide

$$\lim_{|\mathbf{r}_t|\to\infty} \Psi(\mathbf{r}_t) = 0 \tag{5.12}$$

Note that there are waveguide structures that do not support a guided mode.

3. The continuous part of the spectrum is formed by the radiating modes for which the eigenvalues

$$\varepsilon_{eff} < \max(\varepsilon_{clad}) \tag{5.13}$$

Radiating modes show an oscillating behaviour along at least one side of the waveguide structure. Depending on their effective refractive index they are classified as propagating or evanescent radiating modes. In the last case the effective refractive index is purely imaginary.

4. Guided and radiating modes form a complete set of functions. This means that every field inside the waveguide can be represented by a sum of these modes:

$$\mathbf{E}(x,y,z) = \sum_m a_m \mathbf{e}_m(x,y) e^{-j\beta_m z} + \int a(k)\mathbf{e}_k(x,y)e^{-jkz}dk \tag{5.14}$$

In this equation we can see the discrete sum of the guided modes and the continuous spectrum of radiating modes.

Note that a radiating mode can be associated with a plane wave incident from the side (figure 5.5).

5.2.3 The slab waveguide

A further simplification to analyze waveguide structures is to consider waveguide structures that are not only invariant in the propagation direction but also in the direction perpendicular to the

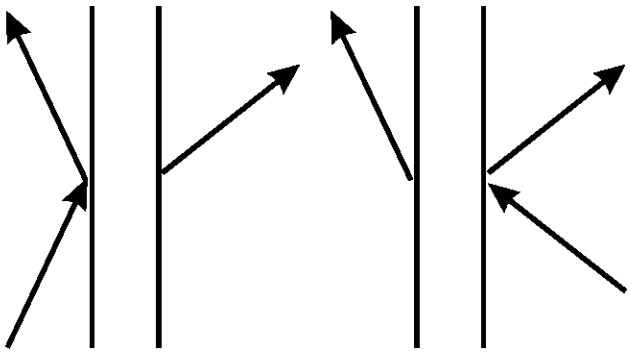


Figure 5.5: Radiating mode

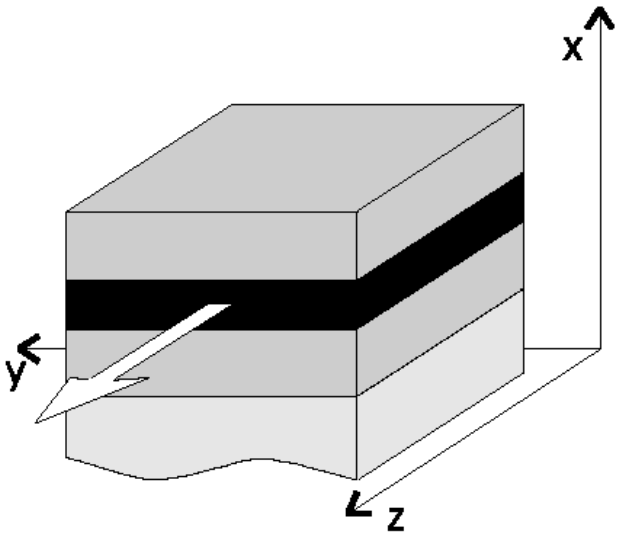


Figure 5.6: Slab waveguide

propagation direction as shown in figure 5.6. These slab waveguides are in practice hardly used but their analysis is the starting point of many approximated theories and so called slab solvers are the basis of many simulation tools for optical waveguides.

We can rewrite equation (5.6) as

$$\begin{aligned} \mathbf{E}(\mathbf{r}) &= \mathbf{e}(x)e^{-j\beta z} \\ \mathbf{H}(\mathbf{r}) &= \mathbf{h}(x)e^{-j\beta z} \end{aligned} \tag{5.15}$$

Substituting these equations in Maxwells curl equations leads to two sets of equations for the so called transverse electric (TE) and transverse magnetic eigenmodes, characterized by the field components $e_y(x), h_x(x), h_z(x)$ and $h_y(x), e_x(x), e_z(x)$ respectively:

$$
\begin{aligned}
TE &\begin{cases} \beta e_y(x) = -\omega\mu_0 h_x(x) \\ \frac{de_y(x)}{dx} = -j\omega\mu_0 h_z(x) \\ \omega_0\varepsilon_0 n^2(x) e_y(x) = -\beta h_x(x) + j\frac{dh_z(x)}{dx} \end{cases} \\
TM &\begin{cases} \beta h_y(x) = \omega\varepsilon_0 n^2(x) e_x(x) \\ \frac{dh_y(x)}{dx} = j\omega\varepsilon_0 n^2(x) e_z(x) \\ \omega\mu_0 h_y(x) = \beta e_x(x) - j\frac{de_z(x)}{dx} \end{cases}
\end{aligned}
\tag{5.16}
$$

in which $\varepsilon_r(x)$ is replaced by $n^2(x)$. By eliminating the x and z field components we can derive a second order differential equation for $e_y(x)$ and $h_y(x)$ respectively:

$$
\begin{aligned}
&\frac{d^2 e_y(x)}{dx^2} + k_0^2 n^2(x) e_y(x) = \beta^2 e_y(x) \\
&\frac{d}{dx}\left(\frac{1}{k_0^2 n^2(x)}\frac{dh_y(x)}{dx}\right) + h_y(x) = \frac{\beta^2}{k_0^2 n^2(x)} h_y(x)
\end{aligned}
\tag{5.17}
$$

An important type of slab waveguide is the so called multi layer slab waveguide. This waveguide consists of a number of layers with refractive index n_i. Because the refractive index is piecewise constant in the multi layer slab waveguide, we can rewrite equations 5.17 as

$$
\begin{aligned}
&\frac{d^2 e_{y,i}(x)}{dx^2} + k_0^2 n_i^2 e_{y,i}(x) = \beta^2 e_{y,i}(x) \\
&\frac{d}{dx}\left(\frac{1}{k_0^2 n_i^2}\frac{dh_{y,i}(x)}{dx}\right) + h_{y,i}(x) = \frac{\beta^2}{k_0^2 n_i^2} h_{y,i}(x)
\end{aligned}
\tag{5.18}
$$

and the TE and TM equations become identical. Solutions will be different however because of the different boundary conditions for TE and TM field components at the interfaces between the different layers. We will continue to work with the TE equation, in order not to overload the notation. The analysis for the TM equation is similar.

The general solution to equation 5.18 can be written as

$$
e_{y,i} = A_i e^{jk_{x,i}(x-a_i)} + B_i e^{-jk_{x,i}(x-a_i)} \tag{5.19}
$$

with

$$
k_{x,i} = \sqrt{k_0^2 n_i^2 - \beta^2} \tag{5.20}
$$

The used notations are clarified in figure 5.7.

Based on equations 5.5 we can derive the following boundary conditions for the interface between two layers:

$$
\begin{cases} e_{y,i}(a_i) = e_{y,i+1}(a_i) \\ \frac{de_{y,i}(a_i)}{dx} = \frac{de_{y,i+1}(a_i)}{dx} \end{cases} \tag{5.21}
$$

apply. Using equations 5.21 and 5.19 we can derive the following matrix relation:

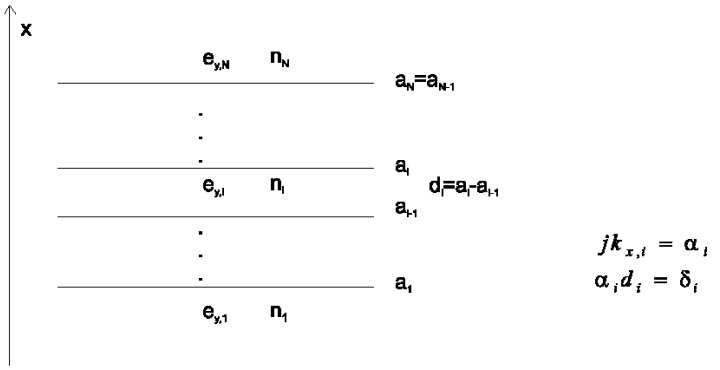


Figure 5.7: Multi layer slab waveguides

$$\begin{bmatrix} A_i \\ B_i \end{bmatrix} = \frac{1}{2\alpha_i} \begin{bmatrix} (\alpha_i + \alpha_{i+1})e^{-\delta_{i+1}} & (\alpha_i - \alpha_{i+1})e^{\delta_{i+1}} \\ (\alpha_i - \alpha_{i+1})e^{-\delta_{i+1}} & (\alpha_i + \alpha_{i+1})e^{\delta_{i+1}} \end{bmatrix} \begin{bmatrix} A_{i+1} \\ B_{i+1} \end{bmatrix} \tag{5.22}$$

By repeating this procedure for all layers, the following matrix equation can be derived:

$$\begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = \begin{bmatrix} t_{11}(\beta^2) & t_{12}(\beta^2) \\ t_{21}(\beta^2) & t_{22}(\beta^2) \end{bmatrix} \begin{bmatrix} A_N \\ B_N \end{bmatrix} \tag{5.23}$$

For guided modes

$$\lim_{x \to \pm\infty} e_y(x) = 0 \tag{5.24}$$

and because $\beta > k_0 n_N$ and $\beta > k_0 n_1$ we can write that $A_1 = B_N = 0$ (if we choose $+j$ as the solution to $\sqrt{-1}$). From equation 5.23 we can see that this can only be fulfilled if $t_{11}(\beta^2) = 0$. The solutions to this dispersion equation yield the guided modes of the structure. Note that this matrix description actually is identical to the analysis for the reflection and transmission of a plane wave at a layer stack discussed in the previous chapter.

Finding the solutions to the dispersion equation has to be done numerically. Only in the case of a 3-layer slab waveguide some interesting properties can be derived analytically.

In the case of a 3-layer slab waveguide, equation 5.19 can be written as (with the notations as depicted in figure 5.8)

$$\begin{cases} e_y = Ae^{-\delta x} & (x \geq 0) \\ e_y = A\cos(\kappa x) + B\sin(\kappa x) & (-d \leq x \leq 0) \\ e_y = (A\cos(\kappa d) - B\sin(\kappa d))e^{\gamma(x+d)} & (x \leq -d) \end{cases} \tag{5.25}$$

With

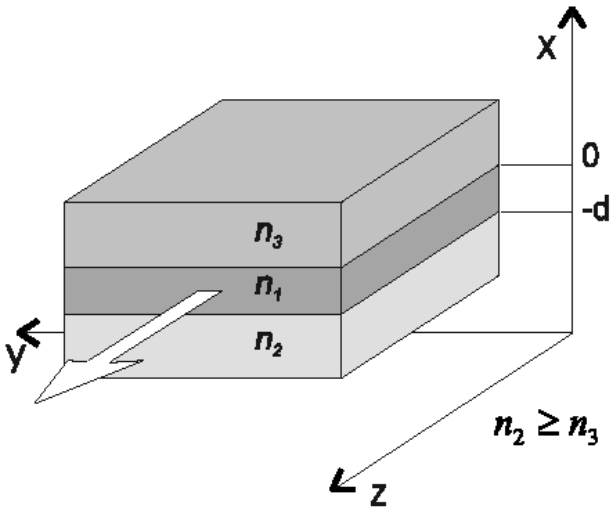


Figure 5.8: Three layer slab waveguide

$$\begin{aligned} \delta &= \sqrt{\beta^2 - n_3^2 k_0^2} \\ \kappa &= \sqrt{n_1^2 k_0^2 - \beta^2} \\ \gamma &= \sqrt{\beta^2 - n_2^2 k_0^2} \end{aligned} \tag{5.26}$$

Hereby we already used the first of the boundary conditions (5.21). To determine A and B we also have to apply the second boundary condition of (5.21). In this way we find following transcendental equation (for TE slab modes):

$$\tan(\kappa d) = \frac{\kappa(\gamma + \delta)}{\kappa^2 - \gamma\delta} \tag{5.27}$$

This is an eigenvalue equation for β with discrete solutions.

For TM slab modes one can find that this transcendental equation can be written as

$$\tan(\kappa d) = \frac{\kappa(\gamma \frac{n_1^2}{n_2^2} + \delta \frac{n_1^2}{n_3^2})}{\kappa^2 - \gamma \frac{n_1^2}{n_2^2} \delta \frac{n_1^2}{n_3^2}} \tag{5.28}$$

With each eigenvalue β an eigenmode can be associated. Often an ω-β diagram is used, being the graphical representation of the dispersion relation for the different eigenmodes. To get a graph (i.e. for the TE polarized slab modes), which is generic, a number of normalized units are used: the normalized frequency ν, the relative effective index b and the asymmetry factor a_{TE}. These are defined as:

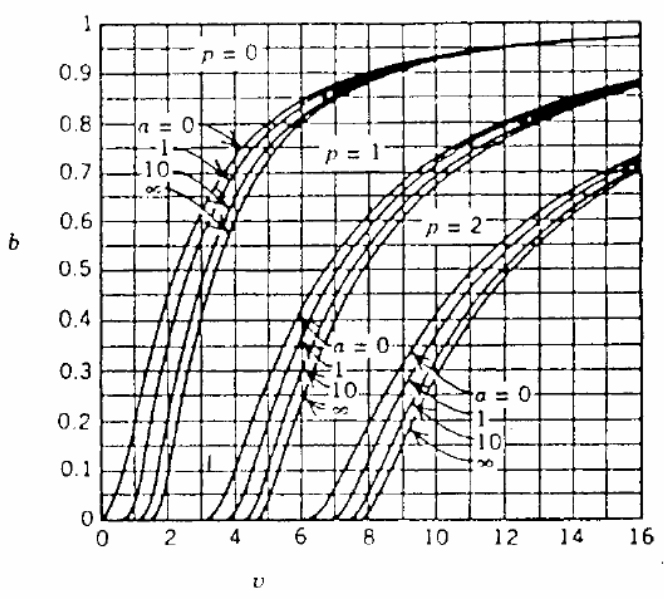


Figure 5.9: Waveguide dispersion curves

$$\begin{aligned} \nu &= k_0 d\sqrt{n_1^2 - n_2^2} \\ b &= \frac{n_{eff}^2 - n_2^2}{n_1^2 - n_2^2} \\ a_{TE} &= \frac{n_2^2 - n_3^2}{n_1^2 - n_2^2} \end{aligned} \tag{5.29}$$

These definitions result in the dispersion curves depicted in figure 5.9 for TE eigenmodes (for different values of the asymmetry factor a_{TE}.

The number of guided TE modes can be calculated from equation 5.25 to 5.27:

$$M = 1 + \mathrm{Int}[\frac{1}{\pi}(\nu - \mathrm{Arctan}(\sqrt{a}_{TE})] \tag{5.30}$$

In this equation Int[...] means the integer part of the argument. For symmetrical waveguides this formula is simplified to $M = 1 + \mathrm{Int}[\frac{\nu}{\pi}]$. So there is at least 1 guided mode. In the case of a symmetrical waveguide knowing the normalized frequency ν is sufficient to determine the amount of guided modes. Therefore this number is frequently used when describing optical waveguides.

Note that the effective index can be considered as some kind of average refractive index felt by the guided mode. In this context we can also define the confinement factor Γ. It is defined (for TE polarization) as

$$\Gamma_i^{TE} = \frac{\int_i E_y^2 dx}{\int_{-\infty}^{+\infty} E_y^2 dx} \tag{5.31}$$

and is a measure for the confinement of the eigenmode inside layer i. Obviously

$$\sum_{i=1}^{N} \Gamma_i^{TE} = 1 \tag{5.32}$$

The confinement factor is often used in the theory of laser operation, to denote which part of the optical power is located inside the active layer, where there is gain.

5.2.4 The effective index method

Lets return to the two-dimensional waveguide. For many waveguide types the lateral dimensions are larger than the transversal dimensions. Moreover, the vertical index contrast often is very low. In that case, the modes of the waveguide often will show a quasi-TE or quasi-TM behaviour and can be approximately described by the scalar Helmholtz equation. The effective index method gives an approximate solution to this equation.

As said, the starting point for the effective index method is the scalar Helmholtz equation:

$$\nabla^2 \Psi(x,y,z) + k_0^2 n^2(x,y)\Psi(x,y,z) = 0 \tag{5.33}$$

In this equation Ψ can be replaced by any of the field components. Because we are still considering longitudinally invariant waveguide structures, the z dependence of $\Psi(x,y,z)$ is given by

$$\Psi(x,y,z) = \psi(x,y)e^{-j\beta z} \tag{5.34}$$

This way, equation (5.33) becomes

$$\nabla_{xy}^2 \psi(x,y) + (k_0^2 n^2(x,y) - \beta^2)\psi(x,y) = 0 \tag{5.35}$$

This equation can be solved in a number of ways. A first method is the effective index method which is an approximate solution. We start from the assumption that we can write

$$\psi(x,y) = F(x,y)G(y) \tag{5.36}$$

in which $F(x,y)$ is a slowly varying function of y, so that we can write that

$$\frac{\partial F}{\partial y} = 0 \tag{5.37}$$

Substitution of equation 5.36 in equation 5.35 leads to

$$F(x,y)\frac{d^2G(y)}{dy^2} + 2\frac{\partial F(x,y)}{\partial y}\frac{dG(y)}{dy} + G(y)\left[\frac{\partial^2 F(x,y)}{\partial x^2} + \frac{\partial^2 F(x,y)}{\partial y^2}\right] + (n^2k_0^2 - \beta^2)F(x,y)G(y) = 0 \tag{5.38}$$

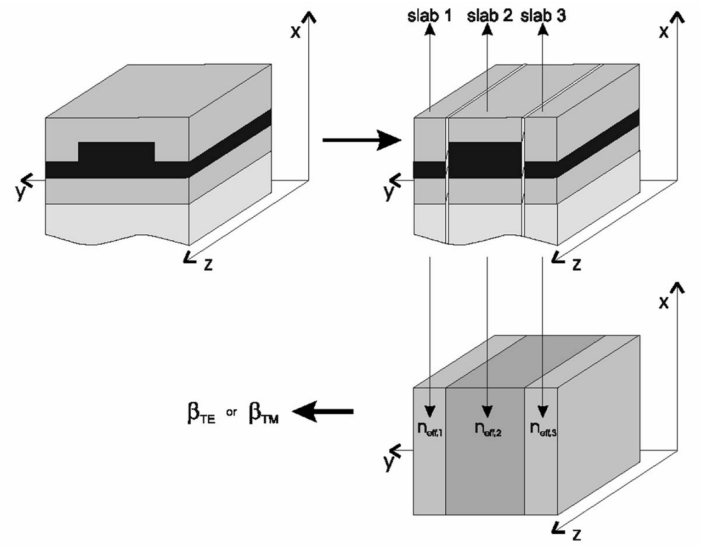


Figure 5.10: Effective index method

Using equation 5.37 this becomes

$$\frac{1}{G}\frac{d^2G}{dy^2} + \frac{1}{F}\frac{\partial^2 F}{\partial x^2} + (k_0^2 n^2(x,y) - \beta^2) = 0 \tag{5.39}$$

We now apply a technique closely resembling the classical technique of separation of variables. The only difference is that $F(x,y)$ shows a weak dependence on y and that we have to introduce an y-dependent separation variable $n_{eff}(y)$. This way we find

$$\begin{array}{l} \frac{1}{F}\frac{\partial^2 F}{\partial x^2} + k_0^2 n^2(x,y) = k_0^2 n_{eff}^2(y) \\ \frac{1}{G}\frac{d^2 G}{dy^2} - \beta^2 = -k_0^2 n_{eff}^2(y) \end{array} \tag{5.40}$$

These are the fundamental equations of the effective index method. Lets consider the first equation. We divide the two dimensional waveguide in slices for which we can assume the refractive index profile as being independent of y. This is very easy for piecewise constant refractive index profiles but is also possible for a continuously varying index profile by applying a staircase approximation (figure 5.10).

Equation (5.40) then reduces to

$$\frac{\partial^2 F_i}{\partial x^2} + k_0^2 n_i^2(x) F_i = k_0^2 n_{eff,i}^2 F_i \tag{5.41}$$

If we compare this equation with equation (5.17a) then it is clear that $n_{eff,i}$ is the effective index of a one dimensional slab waveguide with a refractive index profile $n_i(x)$. The corresponding mode profile determines $F_i(x)$. Solving the first equation of (5.40) leads to an effective index distribution $n_{eff}(y)$. Using this function we can solve the second equation of (5.40). This equation can be rewritten as

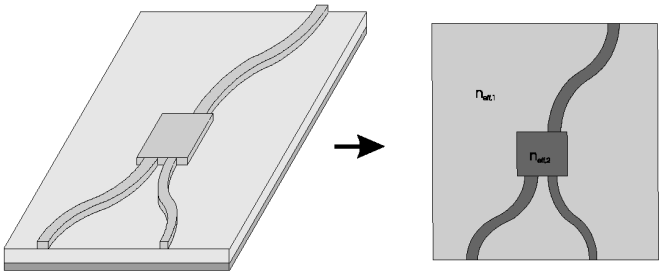


Figure 5.11: one dimensional equivalent

$$\frac{d^2G}{dy^2} + (k_0^2 n_{eff}^2(y) - \beta^2)G = 0 \tag{5.42}$$

Again, this is the equation for a one dimensional slab waveguide with refractive index profile $n_{eff}(y)$. Solving this equation results in the propagation constant β and the mode profile $G(y)$.

Although this method is in principle a scalar method, it still allows to take the polarization of the mode $\psi(x, y)$ into account. Say we are interested in the TE eigenmode of the two dimensional waveguide. The electric field vector of the mode is pointing along the y-axis, just like for the TE slab modes $F_i(x)$. To keep the same polarization state for the second equation this equation has to be solved for TM polarization. To calculate the TM eigenmode of the two dimensional waveguide, the appropriate boundary conditions need to be applied.

The errors made using the effective index method are due to the fact that equation (5.37) does not apply. This is the case in the vicinity of vertical dielectric interfaces. Generally speaking the effective index method will overestimate the propagation constants of the waveguide modes.

The effective index method is not only used to calculate two dimensional mode profiles and its corresponding propagation constants. More often it is used to simplify a two dimensional transversal waveguide structure to a one dimensional structure which can serve as a starting point for further analysis methods like the mode expansion method or the beam propagation method (figure 5.11).

5.2.5 Numerical methods

Numerical methods like finite difference or finite element methods start of from the vectorial equations or the scalar Helmholtz equation. In a finite difference method, the first step always is the discretization of the refractive index profile. To do this the waveguide is put inside a box with dimensions that are sufficiently large to suppose that the fields are zero on the edges of the box. Next, the box is divided into basic cells in which the refractive index is constant (meshing). Depending on the meshing algorithm, the mesh will or will not be equidistant.

In a second step the field equations are discretized by replacing the derivatives by their finite difference representation. In this way the set of partial differential equations is replaced by a linear set of equations which can be solved using standard algebraic methods.

The errors due to the finite difference method have two causes. There is the finite difference approximation of the derivatives and on the other side the approximation that the fields are zero

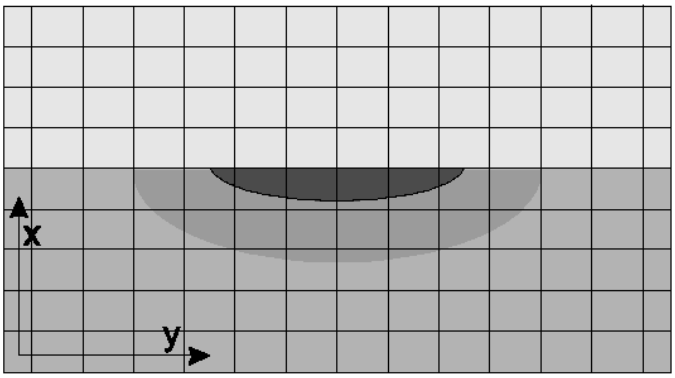


Figure 5.12: Discretization for numerical methods

on the edges of the box. To reduce these errors the discretization of the mesh can be refined and the dimensions of the box can be increased.

Note that there is a fundamental difference between the errors due to an approximate method like the effective index method and a numerical method like the finite difference method. With numerical methods we can start of with the exact Maxwell equations and the error can always be reduced by refining the discretization parameter, at the expense of larger calculation times. In approximated methods like the effective index methods, the equations are solved rigorously, but the equations are only approximations of Maxwells equations.

5.2.6 Modes of metal-dielectric surface plasmon waveguide structures

In this section we will discuss the waveguiding properties of an interface between a semi-infinite metal with a complex permittivity $\varepsilon_m = \varepsilon'_m + j\varepsilon''_m$ and a semi-infinite dielectric with permittivity $\varepsilon_d = \varepsilon'_d + j\varepsilon''_d$, as shown in figure 5.2.6.

While this layer structure consists only of two semi-infinite materials, one can still look for guided modes with a propagation constant β. If these guided modes exist, their propagation constants can be found by solving the eigenvalue equation 5.27 and 5.28 for TE and TM polarization respectively, by letting d approach to zero, as this two-layer waveguide can be treated as a limiting case of the three-layer waveguide stack discussed in the previous section. For TE polarization, this results in an eigenvalue equation

$$\gamma = -\delta \tag{5.43}$$

while for TM polarization the eigenvalue equation becomes

$$\frac{\gamma}{\varepsilon_m} + \frac{\delta}{\varepsilon_d} = 0 \tag{5.44}$$

with

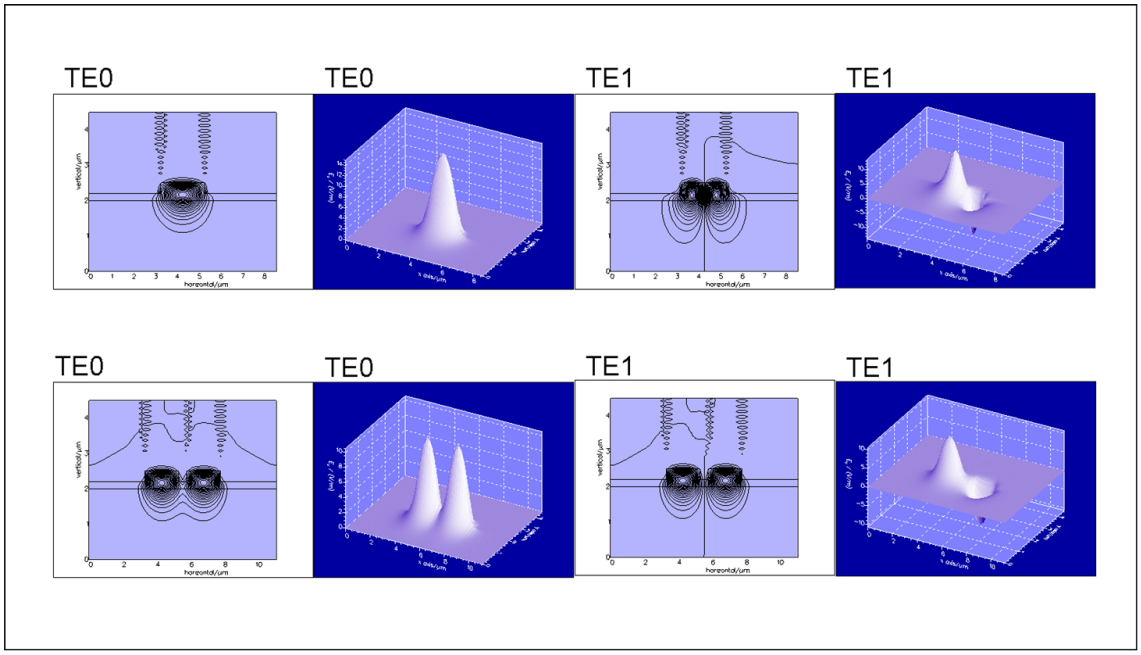


Figure 5.13: Example of 2D mode calculation using a numerical method. The upper row of plots shows the 0th and 1st order TE mode of a rib waveguide. The second row of plots shows the symmetric and antisymmetric TE mode in two coupled waveguides of the same type as the first row.

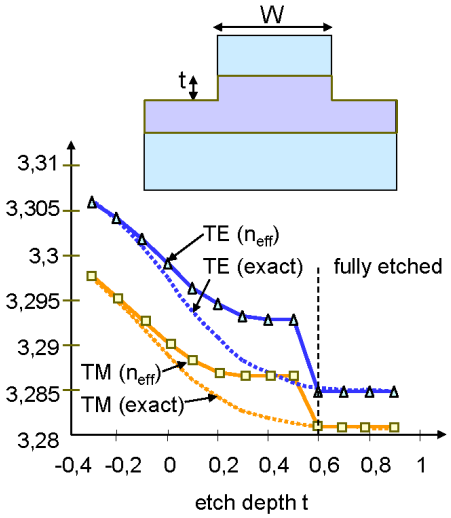


Figure 5.14: In this graph we show the effective index of the 0th order mode (TE and TM) of the depicted rib waveguide, calculated using a numerical mode solver (exact) and the effective index method. The guiding layer of the rib waveguide is 0.6μm thick. In the low index contrast case (not etched in the guided layer ($t < 0$)) the results of the effective index method closely resemble the exact results. Also when the waveguide is etched completely through the guiding layer ($t > 0.6\mu m$) this is the case. Only in the intermediate area the approximation is less good.

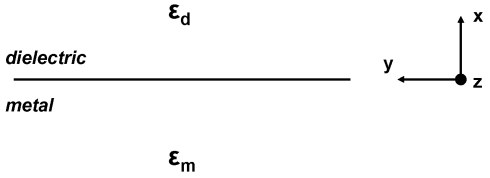


Figure 5.15: interface between a semi-infinite dielectric and a semi-infinite metal

$$\gamma = \sqrt{\beta^2 - \omega^2\mu_0\varepsilon_0\varepsilon_m} \text{ and } \delta = \sqrt{\beta^2 - \omega^2\mu_0\varepsilon_0\varepsilon_d} \tag{5.45}$$

The eigenvalue equation for TE polarization yields no solution: assume that there would be a solution β_{TE}, inserting this solution in 5.43 would result in the requirement that $\varepsilon_m = \varepsilon_d$ (and $\beta_{TE} = \omega\mu_0\varepsilon_0\varepsilon_d$), which represents the case of a plane wave solution in a uniform medium. For TM polarization, we can rewrite the eigenvalue equation as:

$$\beta = \frac{\omega}{c}\sqrt{\frac{\varepsilon_d\varepsilon_m}{\varepsilon_d + \varepsilon_m}} \tag{5.46}$$

For a lossless metal and dielectric ($\varepsilon''_m = \varepsilon''_d = 0$), we can write

$$\beta = \frac{\omega}{c}\sqrt{\frac{\varepsilon'_d\varepsilon'_m}{\varepsilon'_d + \varepsilon'_m}} \tag{5.47}$$

A guided mode exists if β is real (as an imaginary β would result in an exponentially decaying field without oscillations, which carries no power). There are two possibilities to satisfy this requirement (assuming a positive ϵ'_d):

$$\varepsilon'_m > 0 \tag{5.48}$$

or

$$\varepsilon'_m < 0 \text{ and } \left|\varepsilon'_m\right| > \varepsilon'_d \tag{5.49}$$

Substituting equation 5.47 in equation 5.45, one finds an expression for γ and δ, which describe the field profile of the electromagnetic mode as

$$\begin{cases} h_y = Ae^{-\delta x} & (x \geq 0) \\ h_y = Ae^{\gamma x} & (x \leq 0) \end{cases} \tag{5.50}$$

in analogy with the mode profile for TE polarized waveguide modes for dielectric waveguides in equation 5.50 (letting d go to zero and also taking into account the continuity of h_y at the interface).

In this equation, γ and δ are give by

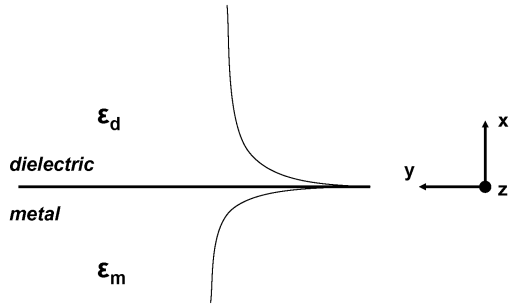


Figure 5.16: h_y-field of a surface plasmon propagating at the interface between a semi-infinite dielectric and a semi-infinite metal

$$\gamma = \frac{\omega}{c}\sqrt{\frac{-\varepsilon_m'^2}{\varepsilon_m' + \varepsilon_d'}} \tag{5.51}$$

and

$$\delta = \frac{\omega}{c}\sqrt{\frac{-\varepsilon_d'^2}{\varepsilon_m' + \varepsilon_d'}} \tag{5.52}$$

When $\varepsilon_m' < 0$ and $|\varepsilon_m'| > \varepsilon_d'$ (condition 5.49), both γ and δ are real and positive, implying that the h_y field component of the mode guided at the metal-dielectric interface consists of two exponentially decaying functions, as shown in figure 5.2.6. This mode, which exists at the interface between a semi-infinite dielectric and (perfect) metal is referred to as a surface plasmon mode. The e_x and e_z field components can easily be calculated from equation 5.16. These are also exponentially decaying (with e_z being continuous at the interface, while there is a discontinuity in the e_x component). The fields typically penetrate much deeper into the dielectric than into the metal. The fact that the field peaks at the interface, makes the propagation properties of the surface plasmon mode very sensitive to possible variations in the refractive index at the metal-dielectric interface and these modes can therefore be used to sense the presence of a very thin layer at the metal surface (e.g. the detection of a monolayer of proteins at a chemically activated gold surface).

When $\varepsilon_m' > 0$, γ and δ will be imaginary, meaning that the light will not be bound to the interface, and radiate into the two semi-infinite media (radiative surface plasmon mode).

While we assumed a perfect metal in the above discussion, the theory can be extended for the case of a lossy metal. The above conclusions remain valid, while the propagation constant obtains an imaginary part which describes the loss the surface plasmon mode experiences when travelling along the metal-dielectric interface. Propagation lengths are typically in the order of 10 to 100μm, depending on the metal used and the wavelength.

The real part of the dielectric function of a metal can be described by (according to the free electron model of Drude)

$$\varepsilon_1'(\omega) = 1 - \frac{\omega_p^2}{\omega^2} \tag{5.53}$$

with ω_p the plasma-frequency. Therefore, a non-radiating surface plasmon ($\varepsilon'_m < 0$ and $|\varepsilon'_m| > \varepsilon'_d$) can only exist when

$$\omega < \frac{\omega_p}{\sqrt{1+\varepsilon'_d}} \tag{5.54}$$

while radiating surface plasmons ($\varepsilon'_m > 0$) occur when $\omega > \omega_p$.

5.3 Propagation through dielectric waveguide structures

In this section we will discuss the propagation of light through waveguide structures for which the transversal cross section varies along the propagation direction. In some cases approximated semi-analytical calculation methods can be used (mode expansion, coupled mode theory, supermode theory). In most cases a complete numerical treatment will be necessary.

5.3.1 Mode expansion and propagation method

We assume a one dimensional z invariant waveguide that is excited by a field distribution $\Psi(x,z)$ and we want to see how this field looks like after propagating through the waveguide over a distance L.

The eigenmodes of the waveguide form a complete set and are orthonormal, so we can write

$$\Psi(x,z=0) = \sum_{i=0}^{N-1} a_i \psi_i(x) \quad \text{with} \int_{-\infty}^{+\infty} \psi_i(x)\psi_j(x)dx = \delta_{ij} \tag{5.55}$$

In this equation ψ_i is the i-th eigenmode of the waveguide. This sum includes all eigenmodes (both the guided and radiating eigenmodes, for which the sum actually has to be replaced by an integral). The expansion coefficients a_i can easily be calculated. By multiplying equation (5.55) with ψ_j and integrating this equation from $-\infty$ to $+\infty$ we find

$$a_j = \int_{-\infty}^{+\infty} \Psi(x,z=0)\psi_j(x)dx \tag{5.56}$$

The propagation of the individual eigenmodes through the waveguide is trivial (multiplying each eigenmode by its propagation factor $e^{-j\beta_j z}$). The field $\Psi(x,z)$after propagation through the waveguide over a distance L is then given by

$$\Psi(x,z=L) = \sum_{i=0}^{N-1} \left[\int_{-\infty}^{+\infty} \Psi(x,z=0)\psi_i(x)dx \right] \psi_i(x) e^{-j\beta_i L} \tag{5.57}$$

Suppose this waveguide is coupled to another, also z-invariant waveguide, with eigenmodes $\phi_i(x)$. If we assume that there are no reflections (which is an approximation), the output field of the first waveguide becomes the input field of the second waveguide. Just like in equation (5.55) we can decompose this field in the eigenmodes of the second waveguide

$$\Psi(x, z = L) = \sum_{i=0}^{N-1} a_i \psi_i(x) e^{-j\beta_i L} = \sum_{j=0}^{M-1} b_j \phi_j(x) \tag{5.58}$$

Because also the eigenfunctions $\phi_i(x)$ are orthonormal, we can find an expression for b_k:

$$b_k = \int_{-\infty}^{+\infty} \Psi(x, z = L)\phi_k(x)dx = \sum_{i=0}^{N-1} a_i \int_{-\infty}^{+\infty} \psi_i(x)\phi_k(x)dx e^{-j\beta_i L} \tag{5.59}$$

We can now propagate the field (5.58) through the second waveguide:

$$\Psi(x, z = L + L') = \sum_{j=0}^{M-1} b_j \phi_j e^{-j\gamma_j L'} \tag{5.60}$$

in which γ_j are the propagation constants of the eigenmodes in the second waveguide. This procedure can be repeated as often as required and can be written down easily in a matrix notation.

Note that at every vertical waveguide discontinuity a reflected field originates. So, not only the coupling of the eigenmodes of the first waveguide to the eigenmodes of the second waveguide needs to be taken into account, but also the coupling back to the eigenmodes of the first waveguide needs to be considered. When there is only a weak discontinuity these reflections can often be neglected and in such a case the unidirectional mode expansion propagation method described above applies. In some cases it is necessary however to include the reflections and then a bidirectional eigenmode expansion propagation method has to be used.

Figure 5.17a shows how the propagation of a field through a gradually broadening waveguide can be calculated by means of the mode expansion method. Here, the discontinuities are small and we can neglect the reflected field. For the codirectional coupler of figure 5.17b this is not the case. In this case reflections play an important role and a bidirectional method has to be used.

5.3.2 Coupled mode theory

In a regular z-independent waveguide, eigenmodes are orthonormal and propagate in an independent way. The complete field can be written as a linear combination of the eigenmodes. In some situations the waveguide structure can be seen as a small perturbation of the simple z-independent waveguide structure. In that case, the field can still be written in terms of the modes of the simple waveguide structure, but due to the perturbation these modes will no longer be decoupled. In the case two modes are dominant, the field can be written as:

$$\Psi(x, z) = C_1(z)e^{-j\beta_1 z}\psi_1(x) + C_2(z)e^{-j\beta_2 z}\psi_2(x) \tag{5.61}$$

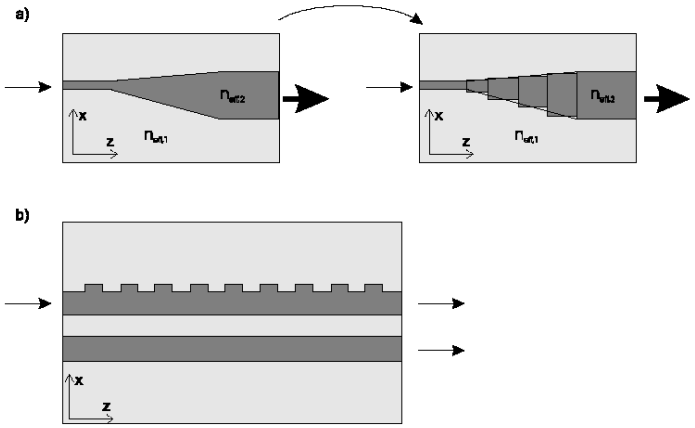


Figure 5.17: Waveguide structures

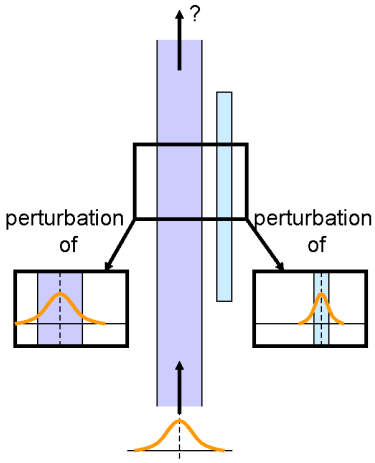


Figure 5.18: Coupled mode theory

Fast longitudinal variations are taken into account by the propagation factor $e^{-j\beta_i z}$ while the z-dependent coefficients $C_1(z)$ and $C_2(z)$ describe the coupling between the two modes ψ_1 and ψ_2. An alternative formulation takes all longitudinal variations together in one term $X_i(z)$:

$$\Psi(x,z) = X_1(z)\psi_1(x) + X_2(z)\psi_2(x) \tag{5.62}$$

We will continue to use the formulation used in equation (5.62). Depending on the envisioned structure an appropriate choice of ψ_1 and ψ_2 has to be made. In figure 5.18 a typical example is shown of a waveguide structure where coupled mode theory can be applied. Is this case we choose ψ_1 and ψ_2 to be the waveguide modes of the unperturbed waveguides. Note that we can consider this system to be double perturbed.

The uncoupled modes ψ_1 and ψ_2 satisfy following propagation equations:

$$\begin{array}{l} \frac{dX_1}{dz} = -j\beta_1 X_1 \\ \frac{dX_2}{dz} = -j\beta_2 X_2 \end{array} \tag{5.63}$$

with the obvious solution

$$\begin{array}{l} X_1(z) = e^{-j\beta_1 z} \\ X_2(z) = e^{-j\beta_2 z} \end{array} \tag{5.64}$$

Coupled mode theory postulates that linear coupling terms need to be added to this equation to describe the perturbed system

$$\begin{array}{l} \frac{dX_1(z)}{dz} = -j\beta_1 X_1(z) - j(\kappa_{11} X_1 + \kappa_{12} X_2) \\ \frac{dX_2(z)}{dz} = -j\beta_2 X_2(z) - j(\kappa_{21} X_1 + \kappa_{22} X_2) \end{array} \tag{5.65}$$

We will assume that the coupling coefficients are z-independent and that both modes travel in the same direction (uniform codirectional coupling). Normally, some other assumptions are also made:

- The two modes are normalized
- The total power flux in the system can be calculated by the sum of the power carried by every mode seperately:

  $$P(z) = |X_1|^2 + |X_2|^2 \tag{5.66}$$

  Modes have to be power independent for this.
- The complete system is lossless ($\frac{dP(z)}{dz} = 0$). If we calculate $\frac{dP(z)}{dz}$ by means of equation (5.66) and (5.65) we find that

  $$\begin{cases} \beta_1, \beta_2, \kappa_{11}, \kappa_{22} \text{ are real} \\ \kappa_{21} = \kappa_{12}* \end{cases} \tag{5.67}$$

- Based on reciprocity and symmetry around a plane parallel to z=0 we also find (both for lossless and lossy systems) that

  $$\kappa_{21} = \kappa_{12} \tag{5.68}$$

  This means that in the lossless case all coupling constants are real.

If we choose initial conditions to be $X_1 = 1$ and $X_2 = 0$, which means that only one of the modes is excited, integrating equation (5.65) leads to the following solution (see appendix).

$$\begin{array}{l} X_1(z) = e^{-j\beta z}\left[\cos(\delta z) - j\frac{\Delta}{\delta}\sin(\delta z)\right] \\ X_2(z) = e^{-j\beta z}\left[-j\frac{\kappa_{21}}{\delta}\sin(\delta z)\right] \end{array} \tag{5.69}$$

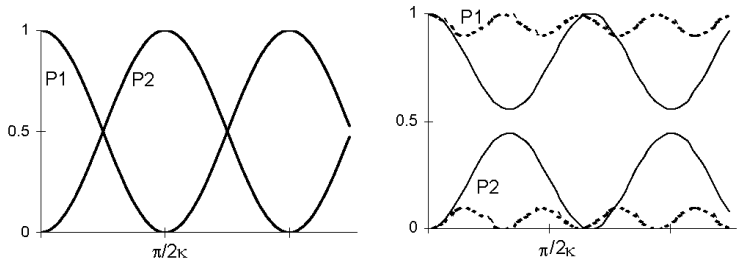


Figure 5.19: Variation of the power in the two modes ψ_1 and ψ_2 as a function of length. In the left graph $\Delta = 0$ (no phase mismatch). In the right graph the power exchange in the case of $\Delta = 1.113\kappa$ and $\Delta = 3.0\kappa$ (dashed line) is shown.

with

$$\begin{aligned} \beta &= (\beta_1 + \beta_2 + \kappa_{11} + \kappa_{22})/2 \\ \kappa &= \sqrt{\kappa_{12}\kappa_{21}} \\ \Delta &= (\beta_1 - \beta_2 + \kappa_{11} - \kappa_{22})/2 \\ \delta &= \sqrt{\Delta^2 + \kappa^2} \end{aligned} \tag{5.70}$$

with β the average propagation constant of both modes and Δ expressing the phase matching per length unit between both modes (note that in most cases κ_{11} and κ_{22} are small compared to β_1 and β_2, so $\Delta \approx \frac{\beta_1 - \beta_2}{2}$).

In figure 5.19 some typical variations of the power $P_1 = |X_1|^2$ and $P_2 = |X_2|^2$ are plotted as a function of z in the lossless case ($\kappa = \kappa_{12} = \kappa_{21}$). We can clearly see the periodic power exchange between the modes, for which the period and the amount of power exchanged depend on the coupling coefficient κ and the phase mismatch Δ.

- When there is no phase mismatch ($\Delta = 0$) complete power exchange occurs. This is the case when both modes have the same propagation constant. This case is also called synchronous coupling. The coupling length is

$$L_c = \frac{\pi}{2\kappa} \tag{5.71}$$

- As long as Δ remains smaller than κ, still an important power exchange occurs. With increasing phase mismatch, both the power exchange level and power exchange period decrease. The coupling length becomes

$$L_c = \frac{\pi}{2\delta} = \frac{\pi}{2\sqrt{\delta^2 + \kappa^2}} \tag{5.72}$$

  The maximum power that can be coupled from one mode to another is given by

$$\frac{\kappa^2}{\kappa^2 + \Delta^2} \tag{5.73}$$

- For a large phase mismatch ($\Delta >> \kappa$) power exchange is negligible. The modes are no longer coupled.

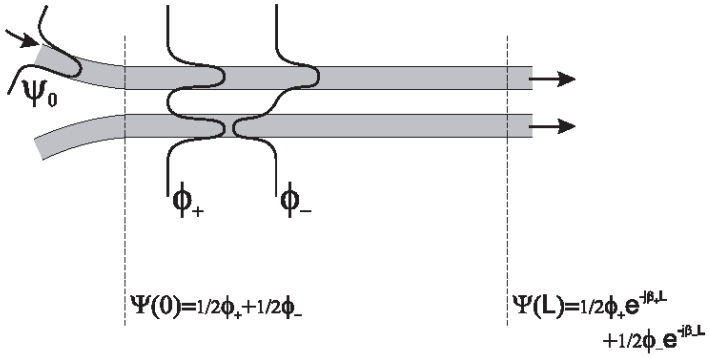


Figure 5.20: Directional coupler structure

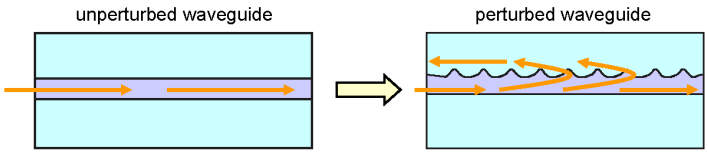


Figure 5.21: Contradirectional grating coupler

We will now consider two typical examples in which the coupled mode theory can be applied.

- Directional coupler

  In the case of a directional coupler, consisting of two waveguides running parallel to each other, the coupling coefficients can easily be calculated using perturbation theory (see Appendix). We find that

$$\begin{aligned} \kappa_{12} &= \frac{k_0^2}{2} \int (n_{12}^2 - {n_1}^2)\psi_1\psi_2 dx \\ \kappa_{21} &= \frac{k_0^2}{2} \int (n_{12}^2 - {n_2}^2)\psi_1\psi_2 dx \end{aligned} \tag{5.74}$$

- Contradirectional grating coupler

  In the case of a contradirectional grating coupler one chooses the modes ψ_1 and ψ_2 to be identical but propagating in the opposite direction. The theory described above can then be repeated, however we have to change equation (5.66) for the z-dependent power flux to

$$P(z) = |X_1|^2 - |X_2|^2 \tag{5.75}$$

  In the chapter on periodic structures this theory will be elaborated. Applications of the contradirectional coupler are DBR and DFB lasers.

5.3.3 Supermodes

An alternative calculation method uses the theory of supermodes [1]. Lets consider again a directional coupler consisting of two identical monomodal waveguides, both with fundamental mode ψ_0. We can also look at the structure as a whole as being a single waveguide. In this waveguide structure a symmetrical and antisymmetrical mode ϕ_+ and ϕ_- with propagation constants β_+ and β_- can propagate (see also figure 5.13, 5.20).

When we excite one of the waveguides with its modal field ψ_0 then we can write

$$\Psi(x, z = 0) = \psi_0(x) \approx c_+\phi_+ + c_-\phi_- \tag{5.76}$$

with $c_+ = c_- = 1/2$. After propagation over a distance L this becomes

$$\Psi(x, z = L) = c_+\phi_+ e^{-j\beta_+ L} + c_-\phi_- e^{-j\beta_- L} \tag{5.77}$$

or

$$\Psi(x, z = L) = c_+ e^{-j\beta_+ L}\left[\phi_+ + \phi_- e^{+j(\beta_+ - \beta_-)L}\right] \tag{5.78}$$

When $e^{+j(\beta_+ - \beta_-)L} = -1$ all power will be concentrated in the second waveguide. This way we find the following relation between the coupling coefficient κ and the difference $\beta_+ - \beta_-$:

$$\beta_+ - \beta_- = 2\kappa \tag{5.79}$$

Let's now consider an asymmetrical directional coupler. The modes will no longer be perfectly symmetrical and antisymmetrical but tend to look more like the modes of the individual waveguides. Equation (5.76) is still valid, but c_+ will no longer equal c_-.

After propagation over a distance L, we find

$$\Psi(x, z = L) = c_+ e^{-j\beta_+ L}\left[\phi_+ + \frac{c_-}{c_+}\phi_- e^{+j(\beta_+ - \beta_-)L}\right] \tag{5.80}$$

The period of power exchange still is given by

$$L_c = \frac{\pi}{\beta_+ - \beta_-} \tag{5.81}$$

but the power exchange is no longer complete. The amount of power exchange is given by $\frac{c_-}{c_+}$. So we can write, by equating equation 5.81 to equation 5.72

$$(\beta_+ - \beta_-) = 2\delta = 2\sqrt{\Delta^2 + \kappa^2} \tag{5.82}$$

[1] Actually this is a simplified form of the general mode expansion theory described in 5.3.1

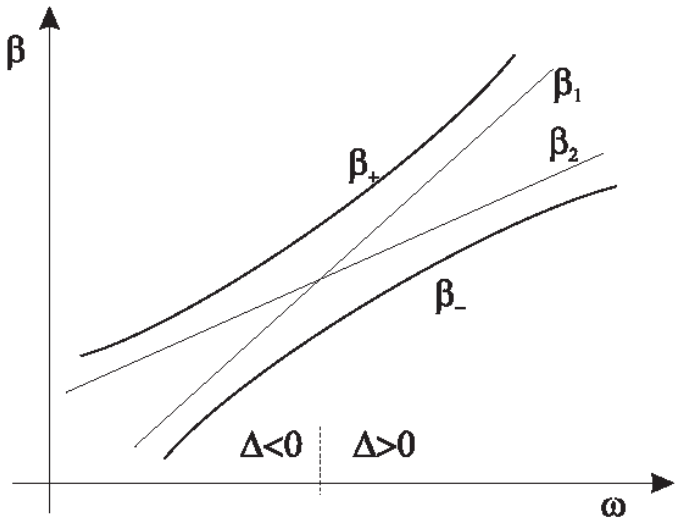


Figure 5.22: Dispersion curves of uncoupled modes and super modes

and if $|\Delta| >> \kappa$

$$(\beta_+ - \beta_-) \approx 2\,|\Delta| \approx |\beta_1 - \beta_2| \tag{5.83}$$

In certain cases it occurs that the dispersion curves $\beta_1(\omega)$ and $\beta_2(\omega)$ of the uncoupled waveguides intersect at a certain frequency. This means that the directional coupler operates in phase matched conditions for this frequency (so there is strong coupling) and for other frequencies the coupling is weak (asynchronous coupling). So we get the picture as shown in figure 5.22 in which $\beta_1, \beta_2, \beta_+$ and β_- are shown as a function of frequency. From this picture it is clear that at low frequency the supermodes resemble the propagation constant (and field profile) of one of the unperturbed modes. At the phase matching frequency this behaviour is broken and $\beta_+ - \beta_- = 2\kappa$. At higher frequencies this behaviour changes and each of the supermodes will resemble the other unperturbed waveguide mode.

5.3.4 Beam propagation method

1. Introduction

   The methods described in the previous sections only are applicable for a number of simple waveguide structures. In most cases there will be some coupling between guided and radiating modes. Because of this coupling, part of the optical power will be lost. The most important technique that takes into account this radiation loss is the beam propagation method (BPM). BPM allows to calculate the propagation of the optical field over a large distance (compared to the wavelength) and this for very complex structures. The required computer code is fairly simple and can easily be extended to waveguides with gain and loss, to waveguides with discrete longitudinal reflections or for polarization sensitive waveguide configurations like in the anisotropic $LiNbO_3$ crystal.

The BPM-method can be used to model three dimensional waveguide structures. The simulation can be done in two ways. One can use a full three dimensional version of the BPM code. This leads to the most accurate results but is very time consuming. A less time consuming method is the combination of the effective index method and a two dimensional BPM. This eliminates the large index contrast with air and reduces the simulation time substantially. It is only useful however if the radiation towards the upper and lower half space is negligible. In real structures this is often the case. Therefore, we will restrict ourself to the two dimensional BPM method. First we will discuss the classical FFT-BPM method.

2. FFT-BPM

Starting point of the discussion is the Helmholtz scalar wave equation in two dimensions

$$\nabla^2\Psi(x,z) + k_0^2 n^2(x,z)\Psi(x,z) = 0 \tag{5.84}$$

The square of the refractive index profile can be written as

$$n^2(x,z) = n_0^2 + \Delta n^2(x,z) \tag{5.85}$$

in which n_0 is a well chosen constant refractive index such that

$$\Delta n << n_0 \tag{5.86}$$

We propose a solution to equation (5.84) of the form:

$$\Psi(x,z) = \psi(x,z)e^{-jk_0n_0z} \tag{5.87}$$

The fast z-variations in $\Psi(x,z)$ are covered by the propagation factor $e^{-jk_0n_0z}$. We can assume that $\psi(x,z)$ will only be weakly z-dependent, such that

$$\left|\frac{\partial^2\psi}{\partial z^2}\right| << \left|2k_0n_0\frac{\partial\psi}{\partial z}\right| \tag{5.88}$$

This means that the amplitude function $\psi(x,z)$ varies slowly on the scale of the material wavelength $\frac{2\pi}{k_0n_0}$ (the so called paraxial approximation).

Substituting equation (5.87) into equation (5.84) we find that (taking equation (5.88) into account)

$$\frac{\partial\psi(x,z)}{\partial z} = -\frac{j}{2k_0n_0}\left(\frac{\partial^2\psi(x,z)}{\partial x^2} + k_0^2\Delta n^2(x,z)\psi(x,z)\right) \tag{5.89}$$

This is the scalar Fresnel equation or parabolic wave equation. We can write down equation (5.89) like

$$\frac{\partial\psi(x,z)}{\partial z} = (\hat{T}+\hat{N})\psi(x,z) \tag{5.90}$$

in which

$$\begin{array}{l}\hat{T} = -\frac{j}{2k_0 n_0}\frac{\partial^2}{\partial x^2} \\ \hat{N} = \frac{-jk_0}{2n_0}\Delta n^2\end{array} \tag{5.91}$$

Integration of equation (5.89) can formally be determined to be

$$\psi(x, z+\Delta z) = e^{(\hat{T}+\hat{N})\Delta z}\psi(x,z) = e^{\hat{T}\Delta z}\left[e^{\hat{N}\Delta z}\psi(x,z)\right] \tag{5.92}$$

In this solution the step Δz has to be chosen sufficiently small. On the one hand because then $\Delta n(x,z)$ can be assumed z-independent over this step, on the other hand because stating that

$$e^{(\hat{T}+\hat{N})\Delta z}\psi(x,z) = e^{\hat{T}\Delta z}\left[e^{\hat{N}\Delta z}\psi(x,z)\right] \tag{5.93}$$

which means that we apply the operators $\hat{T}$ and $\hat{N}$ sequentially and not together, is only valid for sufficiently small Δz.

This way the field is propagated stepwise through the complete waveguide structure. We will now study the operator $e^{(\hat{T}+\hat{N})\Delta z}$. Calculating the term $\left[e^{\hat{N}\Delta z}\psi(x,z)\right]$ is straightforward. The field $\psi(x,z)$ is multiplied by an x-dependent phase term. The meaning of the operator $e^{\hat{T}\Delta z}$ is less obvious. Therefore we will first study the influence of the operator on a plane wave $\psi = e^{-jk_x x}$. Developing the operator in a Taylor expansion, we can write:

$$\begin{array}{rcl} e^{\hat{T}\Delta z}e^{-jk_x x} & = & e^{-\frac{j\Delta z}{2k_0 n_0}\frac{\partial^2}{\partial x^2}}e^{-jk_x x} \\ & = & \left(1 + (-\frac{j\Delta z}{2k_0 n_0}\frac{\partial^2}{\partial x^2}) + \frac{1}{2}(-\frac{j\Delta z}{2k_0 n_0}\frac{\partial^2}{\partial x^2})^2 + ...\right)e^{-jk_x x} \\ & = & (1 + \frac{j}{2k_0 n_0}\Delta z k_x^2 + \frac{1}{2}\left(\frac{j}{2k_0 n_0}\Delta z k_x^2\right)^2 + ...)e^{-jk_x x} \\ & = & e^{\frac{j}{2k_0 n_0}\Delta z k_x^2}e^{-jk_x x} \end{array} \tag{5.94}$$

Therefore the influence of the operator on a plane wave can easily be calculated. Therefore, to calculate the influence of the operator on an arbitrary field $\psi(x,z)$ we first take the fourier transform of the field. On this plane wave expansion we can apply the operator $e^{\hat{T}\Delta z}$. By doing an inverse fourier transform we find the result of $e^{\hat{T}\Delta z}\psi(x,z)$. Note that this is equivalent to the propagation of an arbitrary field through a homogenous medium with refractive index n_0.

The calculation of the propagated field $\psi(x, z=\Delta z)$ out of the original field $\psi(x,z)$ is done as follows. First a phase correction, due to the index perturbation, is applied (operator $e^{\hat{N}\Delta z}$). Then the field is decomposed into its plane wave components, these are individually propagated through a homogenous medium with refractive index n_0 and then recomposed to the complete field distribution (operator $e^{\hat{T}\Delta z}$).

Note: Calculating $e^{(\hat{T}+\hat{N})\Delta z}\psi(x,z)$ as $e^{\hat{T}\Delta z}\left[e^{\hat{N}\Delta z}\psi(x,z)\right]$ is only an approximation. This is because the operators do not commute. A better approximation is to write

$$e^{(\hat{T}+\hat{N})\Delta z}\psi(x,z) = e^{(\frac{\hat{T}}{2})\Delta z}e^{\hat{N}\Delta z}e^{(\frac{\hat{T}}{2})\Delta z}\psi(x,z) + O(\hat{T}^3, \hat{N}^3) \tag{5.95}$$

The analysis is completely analogous to the previous one.

3. Expanding the theory

   FFT-BPM method is not often used any more. In most commercial modeling tools a finite difference BPM algorithm is used. In a finite difference BPM the paraxial wave equation is directly discretized, both longitudinally and transversally. Other extensions to the BPM algorithm are related to

   - the waveguide structure which is not paraxial (i.e. equation (5.88) is no longer valid)
   - vectorial calculations
   - reflections
   - 3D waveguides
   - numerical accuracy and efficiency
   - high contrast waveguides
   - improved boundary conditions

5.4 Optical components

5.4.1 Loss in straight waveguides

In the previous section the straight waveguide was already extensively analyzed. In this section we will therefore restrict ourself to an aspect that has not yet been dealt with: propagation losses in waveguides. There are different causes for this loss. The interaction of light and matter result in absorption and non perfect guiding results in scattering and radiation losses.

When the origin of the loss is uniformly spread over the waveguide length, the guided optical power will decrease exponentially with the propagation distance.

$$P(z) = P_0 e^{-\alpha z} \tag{5.96}$$

α is called the power attenuation coefficient. Due to the typical dimensions of an integrated optical circuit (cm scale) α typically has to be below 0.1 to 1.0 cm^{-1}, which corresponds to a loss between 0.5dB/cm to 5dB/cm. We will now discuss the different loss mechanisms which cause these losses.

- Loss through absorption

  In semiconductor materials with a direct bandgap, the easiest absorption mechanism is the electron-hole pair creation by a photon with a photon energy larger than the bandgap E_g. Sometimes this is wanted (photodetectors) and if this is not the case it can easily be prevented by choosing a material composition with a bandgap sufficiently large compared to the photon energy.

  Also free carriers play a role in the absorption process due to inter and intraband transitions. For non-intentionally doped semiconductors with a carrier concentration of about $10^{16} cm^{-3}$ this absorption typically remains below 0.1dB/cm. Also in non semiconductor material losses occur due to electronic and molecular transitions.

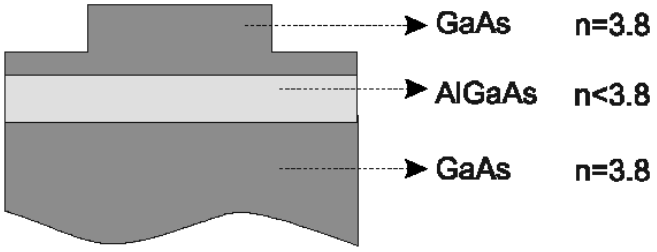


Figure 5.23: GaAs waveguide structure with radiation to the substrate

- Loss through scattering

  Scattering losses are caused by spatial fluctuations of the refractive index (volume scattering) or by the roughness at the sidewalls of the waveguide (surface roughness scattering). These can be both etched waveguide boundaries that determine the waveguide or the interface of two layers which are grown on top of each other. In practice mostly surface roughness scattering seems to be a problem. Based on some simple assumptions, following approximated equation for the scattering loss at boundaries can be obtained

$$\alpha = \alpha_{scat} \frac{(\Delta n)^2 E_s^2}{P} \tag{5.97}$$

  Δn is the index contrast at the interface, P is the power in the optical field and E_s is the field strength at the edges of the waveguide. The constant α_{scat} can only be empirically determined and depends on the etching process used.

  One type of waveguides where special care has to be taken to limit the scattering loss are the deeply etched waveguides because the refractive index contrast between air and semiconductor is very high.

- Loss through radiation

  Loss through radiation is due to the non perfect guiding of the waveguide. The simplest example is a waveguide core that is not positioned between two layers with a lower refractive index. In that case there will be no total internal reflection and power will leak out of the waveguide.

  Another possibility is the case in which the waveguide core is positioned in between two layers with lower refractive index, but where there is however an area with higher refractive index nearby, to which optical power can leak. This process is comparable to the quantum mechanical tunneling of charged particles through a potential barrier. This situation frequently occurs in GaAs/AlGaAs waveguides existing of a GaAs waveguide core cladded by AlGaAs material and grown on a GaAs substrate. The AlGaAs material compositions that can be grown lattice matched on a GaAs substrate have a refractive index lower than that of GaAs, so power will leak from the waveguide core to the substrate. An example of such a waveguide is shown in figure 5.23. By increasing the thickness of the AlGaAs cladding layer, radiation losses can be limited.

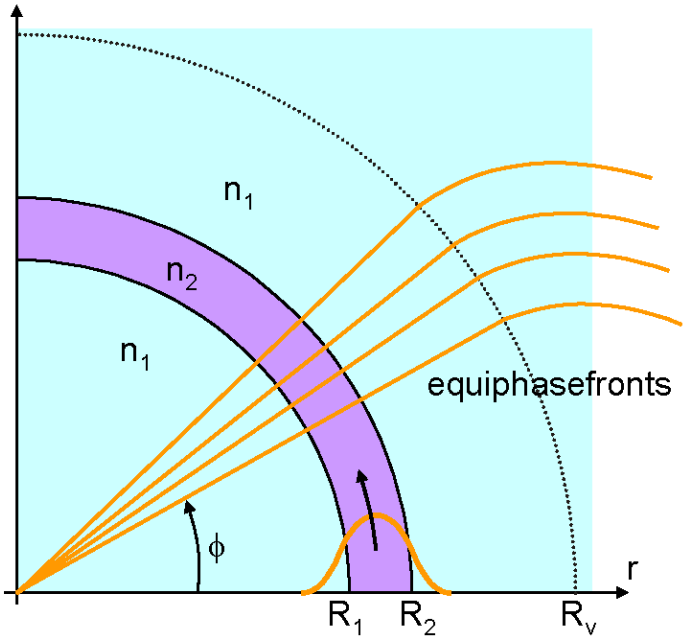


Figure 5.24: Bent waveguide

5.4.2 Bent waveguides

Bent waveguides show a fundamental radiation loss. In straight waveguides the tendency for light to diffract is compensated by the higher refractive index of the waveguide core and the waveguide mode has flat phase fronts. In bent waveguides the phase front is rotating around a rotation center. Because the group velocity of the phase fronts can not exceed the local speed of light (c/n), there is a point where the phase front bends and where radiation occurs (figure 5.24).

Bent waveguides place designers in a dilemma: radiation losses increase nearly exponentially with decreasing bend radius. The integration of multiple components on a semiconductor wafer (InP: maximum 2 inch wafers , Silicon: typically 8 inch wafers) requires waveguides that can change propagation direction on a short distance and without too many losses.

For the theoretical treatment we start off again from a two dimensional waveguide, possibly obtained by applying the effective index method. To calculate the waveguide modes we change to a cylindrical coordinate system r, ϕ. The edges of the bent waveguide will coincide with the coordinate planes $r = R_1$ and $r = R_2$. In this coordinate system the Helmholtz equation can be written as

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2} + k_0^2 n^2(r) \right) \Psi(r,\varphi) = 0 \tag{5.98}$$

with $n(r)$ the refractive index profile. We propose a solution

$$\Psi(r,\varphi) = \psi(r)\Phi(\varphi) \tag{5.99}$$

$\psi(r)$ describes the bend mode profile, while $\Phi(\phi)$ determines the propagation. Substituting this equation in the Helmholtz equation results in

$$\left(\frac{r^2}{\psi(r)}\frac{\partial^2}{\partial r^2}+\frac{r}{\psi(r)}\frac{\partial}{\partial r}+\frac{r^2}{\psi(r)}k_0^2n^2(r)\right)\psi(r)=-\frac{1}{\Phi(\varphi)}\frac{\partial^2\Phi(\varphi)}{\partial\varphi^2} \tag{5.100}$$

While the left side of equation (5.100) only depends on r and the right side only depends on ϕ we can equate both sides with a constant β_ϕ^2. Therefore we can write

$$\frac{1}{\Phi(\varphi)}\frac{\partial^2\Phi(\varphi)}{\partial\varphi^2}+\beta_\phi^2=0 \tag{5.101}$$

The general solution to this equation is

$$\Phi(\varphi)=Ce^{\pm j\beta_\phi\phi} \tag{5.102}$$

The solution is analogous to the solution for propagation in a straight waveguide. The phase fronts coincide with $\phi = cte$ planes. So they turn around the bend. β_ϕ is called the angular propagation constant (dimension rad^{-1}).

The left side of equation (5.100) can be solved directly by using Bessel functions. This requires the calculation of Bessel functions with large and complex indices, which leads to a lot of numerical problems. Another way of solving this equation is the so called conformal transformation. By substituting $r = R_te^{\frac{u}{R_t}}$ we can write equation (5.100) like

$$\left[\frac{\partial^2}{\partial u^2}+(k_0^2n_t^2(u)-\beta_t^2)\right]\psi(u)=0 \tag{5.103}$$

in which $n_t(u) = n(u)e^{\frac{u}{R_t}}$ and $\beta_t = \frac{\beta_\phi}{R_t}$.

This means that in the (u, ϕ) coordinate system the Helmholtz equation has exactly the same shape as in a cartesian coordinate system (x, z) when we replace the refractive index profile $n(u)$ by the transformed index profile $n_t(u)$. Modes and propagation constants in bent waveguides can therefore be calculated by a mode solver developed for solving straight waveguide modes, by introducing the transformed index profile $n_t(u)$. This is schematically depicted in figure 5.25.

Figure 5.26 gives an example of a refractive index profile and the corresponding bend mode profile for different values of the radius of curvature. Based on the transformed refractive index profile it is easy to understand what happens in a bend:

- From the point where

$$\mathrm{Re}(\beta_t)=n_t(u)k_0 \tag{5.104}$$

  radiation will occur and the phase front will bend backwards. That way propagation losses occur.

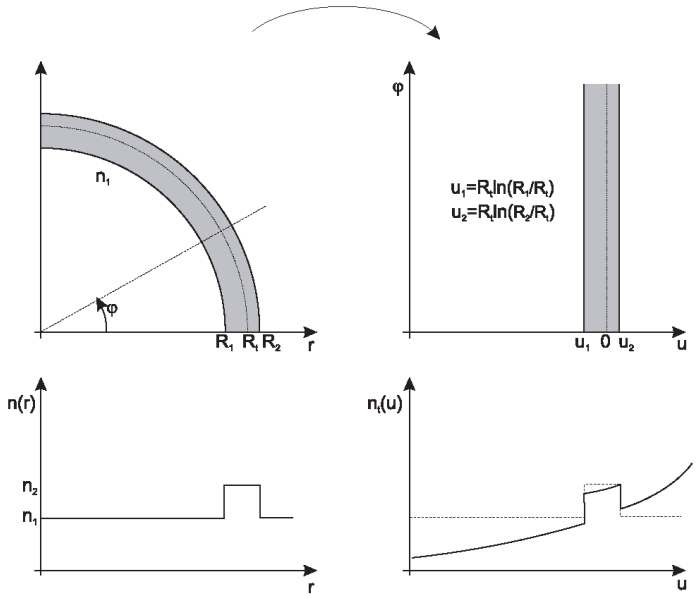
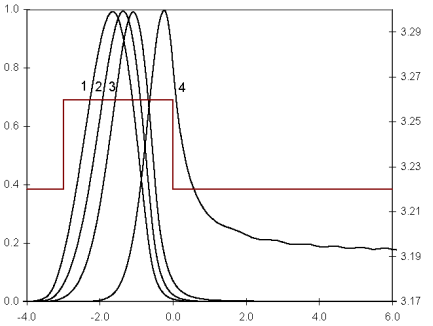


Figure 5.25: Conformal transformation



	R(μm)	α(dB/90)	offset(μm)
1	1000	0	0.38
2	500	0.01	0.63
3	300	0.19	0.85
4	100	18.35	1.25

Figure 5.26: Bend mode profiles

- The mode is most strongly guided in the area with the highest refractive index. Therefore, the mode profile will shift towards the outer rim of the bent waveguide. At the transition from straight to bent waveguide mode adaptation losses will occur.
- By decreasing the ring radius, the mode will completely move towards the outer edge of the waveguide so that the inner edge of the waveguide does not longer contribute to the guiding of the mode. These modes are called whispering gallery modes. The width of the waveguide is no longer an issue in the whispering gallery regime.
- By the increased field strength at the outer edge of the waveguide scattering losses will also increase.

These effects are all clearly visible in figure 5.26.

- Radiation losses

  Because the refractive index will be larger than the effective index of the waveguide mode at a certain distance from the bend, the effective index of a waveguide bend mode will be complex. We can write the angular propagation constant as

  $$\beta_\phi = \beta'_\phi + j\alpha_\phi \tag{5.105}$$

  and

  $$\alpha_\phi = \alpha_t R_t = -\mathrm{Im}(n_{eff,t}) k_0 R_t \tag{5.106}$$

  Figure 5.27 shows the attenuation coefficient as a function of the radius and the refractive index contrast as a parameter. It is clear that the loss increases rapidly below a certain critical bend radius.

- Mode adaptation losses

  In the bend, the mode profile is shifted outwards and is narrower than in the straight waveguide. The adaptation losses between straight and bent waveguide can be reduced by reducing the width of the straight waveguide and giving this waveguide an offset with respect to the bend, such that both waveguide profiles correspond better. The adaptation losses can be calculated by calculating the overlap integral of the bend mode and the mode in the straight waveguide. An adapted straight waveguide / bend interface is shown in figure 5.28.

5.4.3 Tapers

1. General description

   A taper is a smooth transition between two waveguides of different widths or height and is used to couple two components with different waveguide geometry. Changes in the waveguide structure result in mode conversion. This means that power is exchanged between the different modes of the waveguide. When power is coupled to radiating modes, loss occurs, but also in the case of multimodal waveguides it is often unwanted that the power is distributed over different waveguide modes. Mode conversion can be suppressed when the

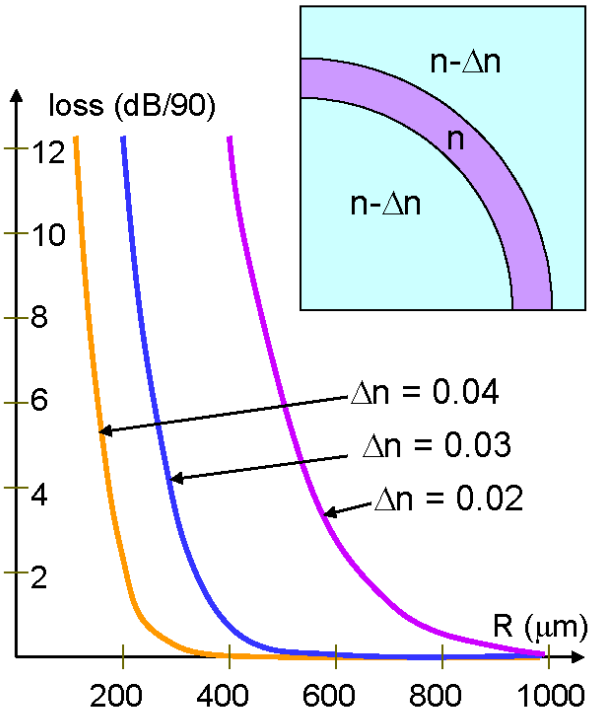


Figure 5.27: Bend mode loss (dB/90 degrees) as a function of radius and refractive index contrast

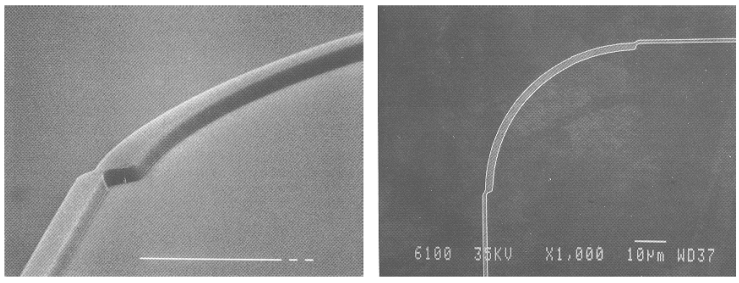


Figure 5.28: Adaptations at a straight waveguide / bend waveguide interface

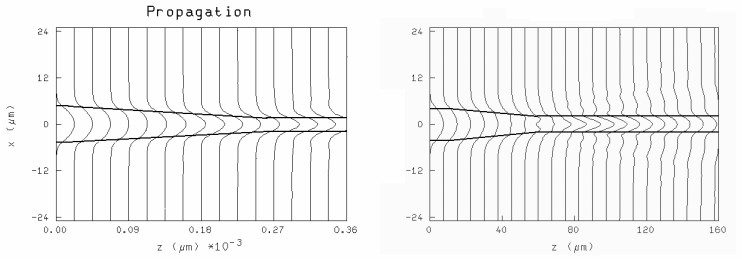


Figure 5.29: BPM calculation of an adiabatic (250um long) and a non-adiabatic (50um long) taper. Both tapers have a straight input waveguide of 10um long and a straight output waveguide of 100um.

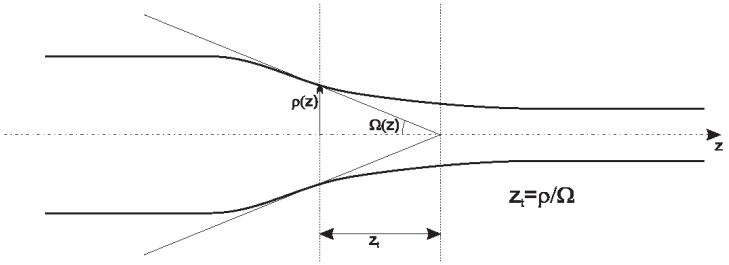


Figure 5.30: Definition of the local taper length

change in waveguide structure is very gentle. In the case we call the adaptation of the mode profile adiabatic. An adiabatic transition between two waveguide structures is a transition where the mode of the system adapts to the changing geometry without loosing power by conversion to other modes. To analyze taper structures numerical methods need to be used: BPM method or mode expansion and propagation method applied to a staircase approximation of the taper profile.

Based on some intuitive considerations it still is possible to derive a criterion to calculate the maximum adiabatic taper angle, which allows an adiabatic transition. As said before, the fundamental mode will transform adiabatically when no power coupling to higher order modes occurs. We can assume that especially the coupling to the second order mode is dominant (the first order mode is antisymmetrical). Based on this assumption, we propose the following criterion for adiabatic transitions: the taper will behave adiabatically when the local taper length is larger than the local coupling length between the fundamental mode and the next symmetrical mode. The local taper length z_t is defined as in figure 5.30.

$$z_t = \frac{\rho(z)}{\tan(\Omega(z))} \approx \frac{\rho(z)}{\Omega(z)} \tag{5.107}$$

The coupling length is given by

$$z_b = \frac{\pi}{\beta_0 - \beta_2} \tag{5.108}$$

The criterion then becomes

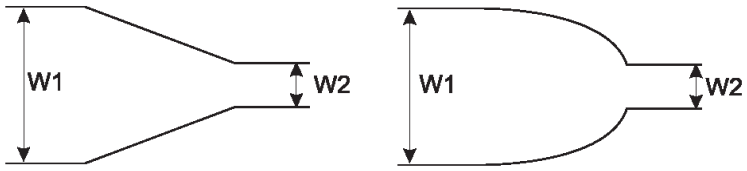


Figure 5.31: The linear and parabolic taper

$$z_t > z_b \text{ or } \Omega < \left[\frac{\beta_0 - \beta_2}{\pi}\right] \rho \tag{5.109}$$

Because the difference $\beta_0 - \beta_2$ is proportional to ρ^{-2} (this will become clear in the discussion on the multimode interference coupler), the maximum taper angle will be larger when the waveguide is narrower. Therefore, the ideal taper design is parabolic.

2. Improving coupling efficiency to an optical fiber

   A typical monomodal optical fiber has a core diameter of 9 μm, has a small refractive index contrast between core and cladding layers ($\Delta n \approx 1\%$) and the mode is circular. The waveguides fabricated in a III-V semiconductor material system typically are a few micrometer wide, have higher refractive index contrast and have non circular mode profiles.

   The fundamental modes in the two systems strongly differ, therefore the coupling losses from fiber to waveguide will be high. A possible solution is the use of a taper structure. By narrowing the integrated waveguide (this can be done both transversally or vertically) the mode will expand and a better mode matching can be obtained, thereby reducing the coupling losses.

5.4.4 Directional coupler

The directional coupler was already treated in previous sections on coupled mode theory and supermode theory. In these cases always a longitudinally invariant waveguide structure was assumed. In reality, a directional coupler consists of a central section (in which the actual coupling takes place) and an input and output section. In many cases part of the optical power will already be exchanged between the waveguides in the input and output section where the waveguides are already in close proximity. This needs to be analyzed using the beam propagation method. The most important application of the directional coupler is the use as a 3dB coupler.

5.4.5 Multimode interference coupler

An alternative to the directional coupler is the use of a multimode interference coupler (MMI). The central part of an MMI is a broad multimode waveguide. Light can be injected in and coupled out of the multimode section through a number of input and output waveguides (figure 5.34). The operation principle of the device is based on the self imaging principle. This is a property of multimodal waveguides where an input field is reproduced in single and multiple images at periodic intervals along the propagation direction of the waveguide. In this way 1xN couplers can be realized, but also cross couplers and even couplers with an arbitrary coupling ratio.

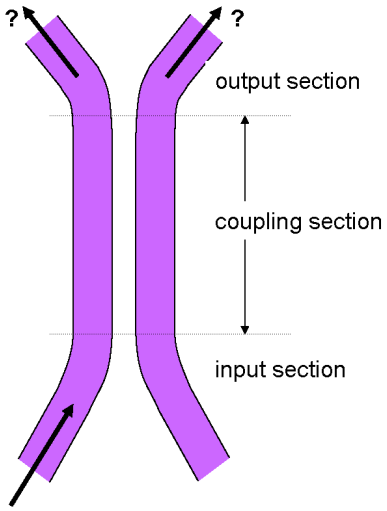


Figure 5.32: Sections of a directional coupler

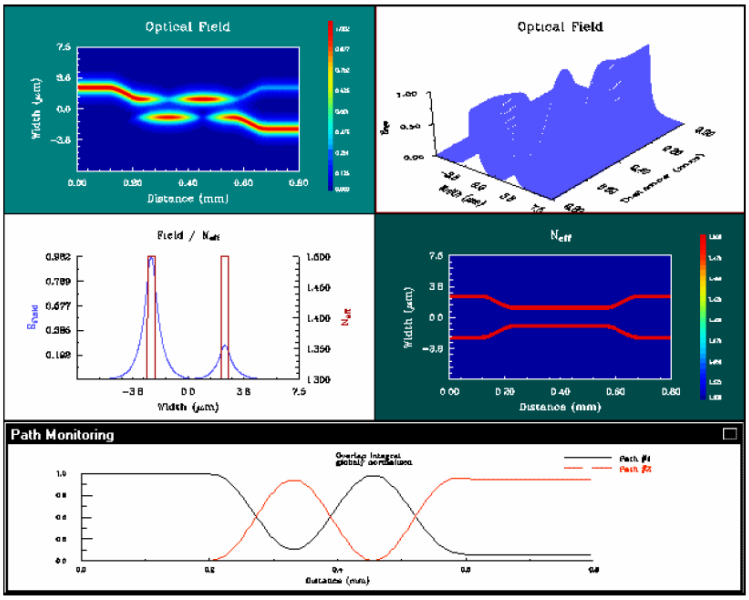


Figure 5.33: BPM simulation of a directional coupler. In the upper plots, the intensity profile and the electrical field profile is shown. The middle graphs show the field distribution at the exit of the directional coupler and the refractive index profile of the simulated directional coupler. In the bottom graph the power in both arms of the coupler is shown. The periodic power exchange is clear.

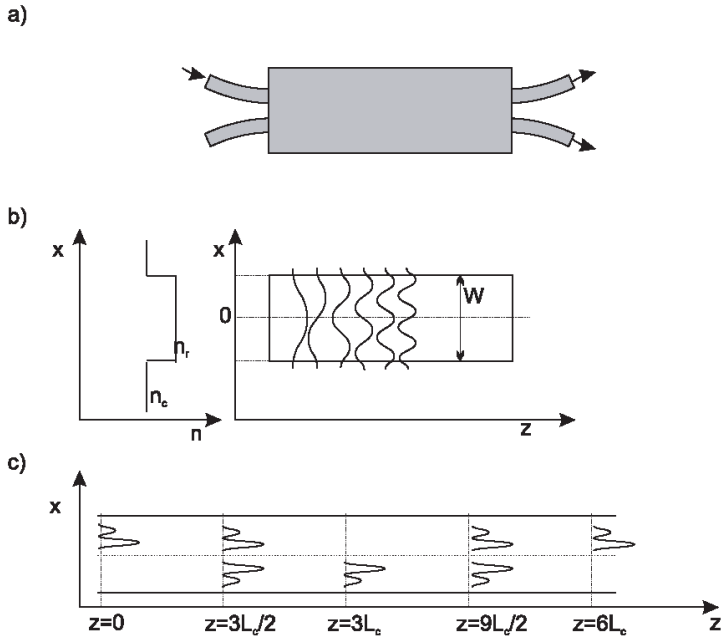


Figure 5.34: multimode interference coupler

To gain insight in the operation principle of the MMI, we will use the mode expansion and propagation theory. Lets consider the multimodal waveguide of width W which is depicted in figure 5.34b. An arbitrary input field $\Psi(x,0)$ of the MMI can be decomposed in the orthonormal eigenmodes ψ_i of the MMI.

$$\Psi(x,0) = \sum_{i=0}^{N-1} c_i \psi_i(x) \text{ with } c_i = \int \Psi(x,0)\psi_i(x)dx \tag{5.110}$$

In the expansion radiation modes have been neglected, which is acceptable as long as the width of the incident field $\Psi(x,0)$ is sufficiently small compared to the width of the MMI. After propagation over a distance z along the MMI the field becomes

$$\Psi(x,z) = \sum_{i=0}^{N-1} c_i \psi_i(x) e^{-j\beta_i z} = e^{-j\beta_0 z} \sum_{i=0}^{N-1} c_i \psi_i(x) e^{j(\beta_0 - \beta_i)z} \tag{5.111}$$

For the modes, the following approximation is made: we assume the shape and propagation constant of each mode to be given by these of the modes of a waveguide with an infinitely high refractive index contrast (field is zero at the edges of the waveguide)

$$\begin{aligned} &\psi_i(x) = \cos(k_{x,i}x) \text{ when } i \text{ is even} \\ &\psi_i(x) = \sin(k_{x,i}x) \text{ when } i \text{ is odd} \\ &k_{x,i} = \tfrac{(i+1)\pi}{W} \end{aligned} \tag{5.112}$$

For real waveguide structures this is only an approximation but in the case of high index contrast waveguides it is a very useful approximation.

As the lateral wave number $k_{x,i}$ and the propagation constant β_i are related through

$$k_{x,i}^2 + \beta_i^2 = k_0^2 n_r^2 \tag{5.113}$$

we find

$$\beta_i \approx k_0 n_r - \frac{(i+1)^2 \pi \lambda_0}{4 n_r W^2} \tag{5.114}$$

Therefore

$$\beta_0 - \beta_i = \frac{i(i+2)\pi\lambda_0}{4 n_r W^2} = \frac{i(i+2)\pi}{3 L_\pi} \tag{5.115}$$

with

$$L_\pi = \frac{\pi}{\beta_0 - \beta_1} = \frac{4 n_r W^2}{3\lambda_0} \tag{5.116}$$

If we substitute this equation into equation (5.111) then we can write down the field in the MMI after a propagation distance L as (neglecting the common phase term)

$$\Psi(x, L) = \sum_{i=0}^{N-1} c_i \psi_i(x) e^{j \frac{i(i+2)\pi}{3 L_\pi} L} \tag{5.117}$$

From this equation we can deduce some interesting properties of the MMI.

- $L = 6L_\pi$

  Then equation (5.117) becomes

  $$\Psi(x, 6L_\pi) = \Psi(x, 0) \tag{5.118}$$

  The image at this distance equals the input field

- $L = 3L_\pi$

  Because

  $$\begin{array}{l} \psi_i(-x) = \psi_i(x) \text{ when } i \text{ is even} \\ \psi_i(-x) = -\psi_i(x) \text{ when } i \text{ is odd} \end{array} \tag{5.119}$$

  we find

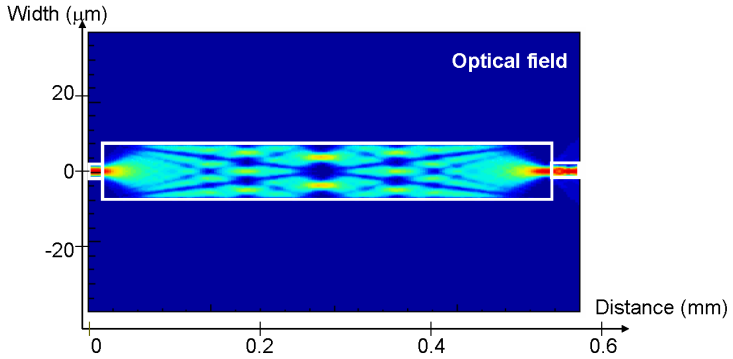


Figure 5.35: BPM simulation of a 1x1 MMI. The dimensions are 15x520 μm^2. On the figure the places where two folded (L/2), three folded (L/3,2L/3) and four folded images(L/4,3L/4) are formed, are clearly visible.

$$\begin{aligned}
\Psi(x,3L_\pi) &= \sum_{i=0}^{N-1} c_i\psi_i(x)e^{j[i(i+2)]\pi} \\
&= \left[\sum_{i=0}^{N-1} c_i\psi_i(x)e^{j[i(i+2)]\pi}\right]_{i=even} + \left[\sum_{i=0}^{N-1} c_i\psi_i(x)e^{j[i(i+2)]\pi}\right]_{i=odd} \\
&= \left[\sum_{i=0}^{N-1} c_i\psi_i(x)\right]_{i=even} + \left[\sum_{i=0}^{N-1} -c_i\psi_i(x)\right]_{i=odd} \\
&= \left[\sum_{i=0}^{N-1} c_i\psi_i(-x)\right]_{i=even} + \left[\sum_{i=0}^{N-1} c_i\psi_i(-x)\right]_{i=odd} \\
&= \sum_{i=0}^{N-1} c_i\psi_i(-x)
\end{aligned} \tag{5.120}$$

This means that at a distance $3L_\pi$ the image is the input field, mirrored around the plane $x = 0$.

- $L = 3\frac{p}{2}L_\pi$ with p odd

  After some manipulation we find

$$\begin{aligned}
\Psi(x,\tfrac{p}{2}3L_\pi) &= \sum_{i=0}^{N-1} c_i\psi_i(x)e^{j[i(i+2)]p\frac{\pi}{2}} \\
&= \left[\sum_{i=0}^{N-1} c_i\psi_i(x)\right]_{i=even} + \left[\sum_{i=0}^{N-1} (-j)^p c_i\psi_i(x)\right]_{i=odd} \\
&= \tfrac{1+(-j)^p}{2}\Psi(x,0) + \tfrac{1-(-j)^p}{2}\Psi(-x,0)
\end{aligned} \tag{5.121}$$

At this distance we find two images, of which one mirrored, both with amplitude $1/\sqrt{2}$. So an MMI with this length can be used as a 2x2 3dB coupler.

More general one can show that at the intermediate distances $L = 3\frac{p}{N}L_\pi$ (with p and N integer numbers without common divider) N multiple images are formed, with an amplitude of $\frac{1}{\sqrt{N}}$. This is illustrated in figure 5.34 and figure 5.35.

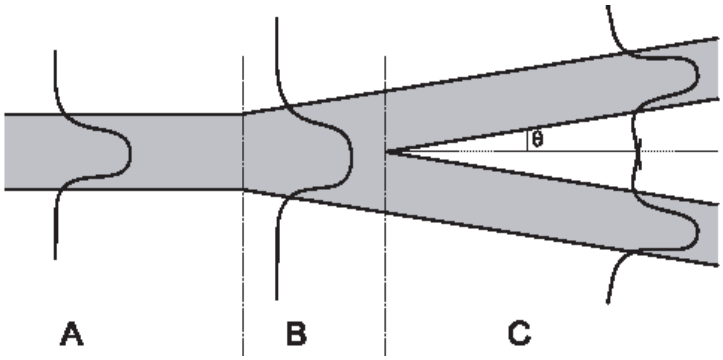


Figure 5.36: Y-junction

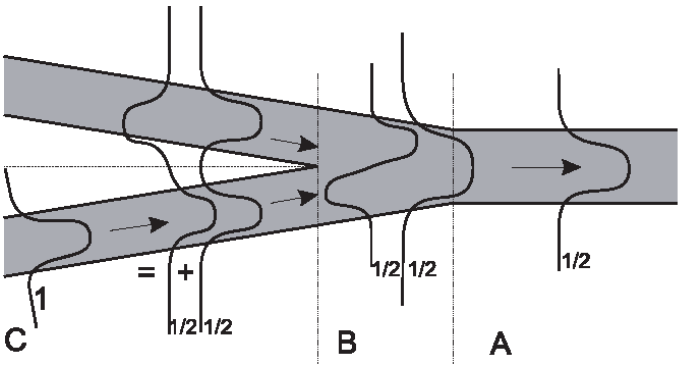


Figure 5.37: Y-junction used as a combiner

5.4.6 Y-junction

The Y-junction is composed of a single waveguide (section A), a taper (section B) and two branched waveguides (section C). An important parameter for the Y-junction is the junction angle θ. When the angle is sufficiently small, the fundamental mode will propagate without great losses in both exit waveguides in section C. This is called an adiabatic Y-junction. The standard Y-junction is designed to equally split the input power over both exit waveguides but it is also possible to design Y-junctions for different splitting ratios.

When we use the Y-junction in the other direction and excite only one branch of the Y-junction with the fundamental mode which propagates towards the junction, we can analyze what happens using the theory of supermodes.

The mode propagating in the lower branch (with amplitude 1) of the Y-junction can be decomposed into the sum of the local symmetrical and anti-symmetrical supermode each carrying half of the power (amplitude $\frac{1}{\sqrt{2}}$). The symmetrical supermode will adiabatically transform to the fundamental mode of the exit waveguide, while the antisymmetric mode will be transformed into the first order mode. However, when the waveguides are monomodal, this part of the power will radiate. As both super modes carry half of the power, this half is lost. This property can also be explained based on reciprocity and symmetry of the Y-junction.

In conclusion, we can say that when we use an adiabatically designed Y-junction as a splitter, no losses occur. When used as a combiner no losses occur when both inputs are excited equally (both in phase and amplitude). This is the reciprocal situation of the use as a splitter. When only one of the inputs is excited, half of the power will be lost and when both inputs are excited equally but in antiphase all power will be lost (for monomodal waveguides).

5.4.7 Diffraction grating

The diffraction grating is analyzed in detail in the chapter on periodic structures. In the integrated version, the diffraction grating is mostly used as an dispersive element. In this way wavelength demultiplexers can be fabricated. Another application is the longitudinal mode selection in a laser cavity. By rotating the diffraction grating a wavelength tunable laser is made.

5.4.8 Phase modulator

Phase modulation in integrated waveguides can be achieved using different physical effects that influence the optical parameters of the material. The most important effects are the electro-optical effect, the thermo-optical effect and the influence of free carriers. All these effects create a change in the refractive index of the material which leads to a phase modulation given by

$$\Delta\phi = \frac{2\pi\Delta n_{eff}L}{\lambda} \tag{5.122}$$

in which L is the length of the waveguide and λ is the operation wavelength.

5.4.9 Amplifiers

An optical amplifier can be used to boost a signal in a waveguide. The amplifier can for example be used in front of a photodetector to reduce the required sensitivity of the photodetector or it can be used as an in line amplifier for long propagation distances. For this application both a doped optical fiber amplifier (EDFA: Erbium Doped Fiber Amplifier) or an SOA (Semiconductor Optical Amplifier) can be used. Sometimes the amplifier is used to amplify different signals (all at a different wavelength) at the same time. An important issue here is the crosstalk between different wavelength channel by saturation of the amplifier. A possible solution to this problem is the so called gain clamped amplifier, where the gain in the amplifier is kept constant.

5.5 Characterization of optical waveguides

Determining the performance of a passive optical waveguide circuit often consists of measuring the power transmission of the components. A typical measurement setup is shown in figure 5.38.

A light beam from a tunable laser (1) is focused on the right facet of the chip (5) using an optical fiber (2). The laser spot can be aligned to the waveguides using a micro translation stage (7,8,9) which is often actuated piezo-electrically (11). Part of the laser light will be coupled into the

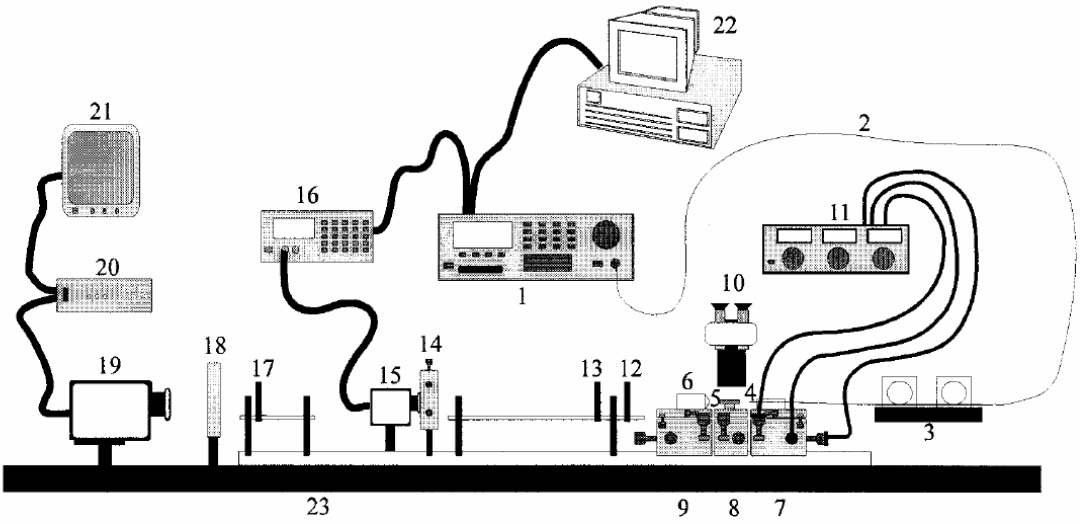


Figure 5.38: Measurement setup

waveguides. The coupling efficiency is determined by the overlap integral of the laser spot and the waveguide modes. The coupling efficiency can be increased by using a lensed fiber (4). After transmission through the waveguide light will be coupled out through the left facet and collected using an objective lens (6) of which the focus is aligned with the left facet. Using an extra lens (13) this light is focused onto a power meter (15,16). A diafragm (14) makes sure only the light that comes from the waveguide is incident on the power meter. When we turn away the power meter it is possible to get a view of the spot with a camera (19,20,21). Polarization can be controlled by polarization control wheels at the input (3) and a polarization filter (12) at the output. Two other techniques to couple light in and out of a waveguide are prism coupling and grating assisted coupling. Grating assisted coupling uses a diffraction grating which is positioned on top of the waveguide to couple light into the chip.

When we assume that the coupling efficiency into the various waveguides on a chip is equal, we can this way do relative loss measurements, where different components are compared. Often the loss of the components are measured with the loss of a straight waveguide as a reference.

It is possible to eliminate the coupling efficiency by doing cut-back loss measurements. After a series of measurements the chip is cleaved in half and one half is used to redo the measurements. The difference between both measurements is the loss in the remaining part. It is clear that this measurement technique only works when the coupling efficiencies are equal for both parts.

An alternative way to measure the waveguide losses is the so called Fabry-Perot measurement technique. When we couple laser light into a waveguide, reflections occur at the facets which can be seen as partly reflecting mirrors. By adding the complex wave amplitudes of the subsequent reflections we can calculate the total transmitted power as

$$T = \frac{(1-R)^2\tau^2}{(1-\tau^2 R)^2 + 4\tau^2 R\sin^2(\beta L)} \text{ with } \tau^2 = e^{-\alpha L} \tag{5.123}$$

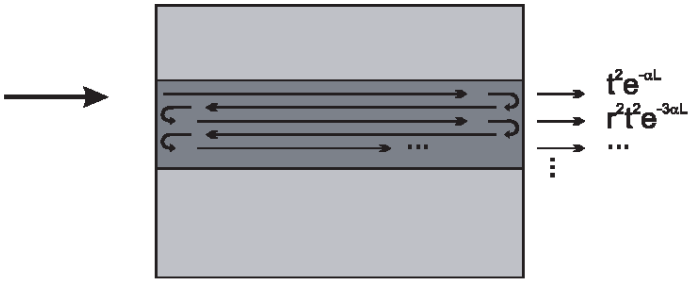


Figure 5.39: Fabry-Perot resonances in a waveguide

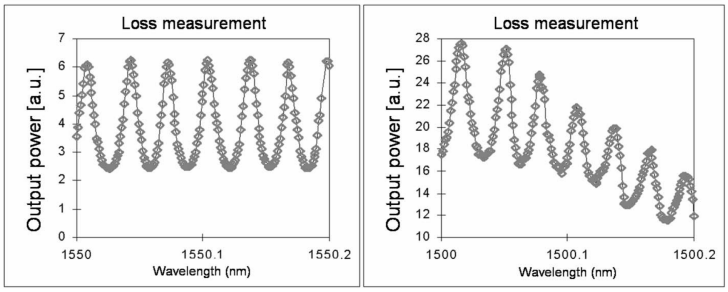


Figure 5.40: Fabry-Perot loss measurements

This transmission function is wavelength dependent (through the propagation constant β). When we calculate the ratio $\frac{T_{\max}}{T_{\min}}$ we find that

$$S = \frac{T_{\max}}{T_{\min}} = \frac{(1+\tau^2 R)^2}{(1-\tau^2 R)^2} \tag{5.124}$$

This way we can calculate the attenuation as

$$\alpha_{dB} = \frac{10}{L}(\log(R) - \log(\frac{\sqrt{S}-1}{\sqrt{S}+1}))[\mathrm{dB/cm}] \tag{5.125}$$

By measuring the transmitted power as a function of wavelength we can easily determine the ratio $\frac{T_{\max}}{T_{\min}}$. When we know the reflection coefficient R (typically 0.32 for waveguide in InP) we can determine the attenuation coefficient of the waveguide. For a correct measurement it is necessary however that only one mode is excited. In figure 5.40 two examples of this type of measurements are given. For the first measurement the determination of the ratio S is simple. For the second measurement an unambiguous measurement is impossible. There are many possible causes for this problem: parasitic reflections in the measurement setup, vibrations, multimodal excitation, wavelength dependence of the focal distance between facet and objective lens...

5.6 Appendix

5.6.1 Solving the coupled mode equations

In this appendix we will solve the coupled mode equations 5.65, here repeated:

$$\frac{dX_1(z)}{dz} = -j\beta_1 X_1(z) - j(\kappa_{11}X_1 + \kappa_{12}X_2) \tag{5.126}$$
$$\frac{dX_2(z)}{dz} = -j\beta_2 X_2(z) - j(\kappa_{21}X_1 + \kappa_{22}X_2) \tag{5.127}$$

We will rewrite these equations such that we can solve for $X_2(z)$. We start by differentiating equation (5.127).

$$\frac{d^2X_2(z)}{dz^2} = -j(\beta_2 + \kappa_{22})\frac{dX_2(z)}{dz} - j\kappa_{21}\frac{dX_1(z)}{dz} \tag{5.128}$$

The derivative $\frac{dX_1(z)}{dz}$ is defined in equation (5.126), after which we again incur (5.127) to rewrite $X_1(z)$.

$$\begin{aligned}
&\tfrac{d^2X_2(z)}{dz^2} = -j(\beta_2 + \kappa_{22})\tfrac{dX_2(z)}{dz} - j\kappa_{21}\left[-j\kappa_{12}X_2(z) - j(\beta_1 + \kappa_{11})X_1(z)\right] \\
&\tfrac{d^2X_2(z)}{dz^2} = -j(\beta_2 + \kappa_{22})\tfrac{dX_2(z)}{dz} - \kappa_{21}\kappa_{12}X_2(z) - (\beta_1 + \kappa_{11})\left(j\tfrac{dX_2(z)}{dz} - (\beta_2 + \kappa_{22})X_2(z)\right) \\
&\tfrac{d^2X_2(z)}{dz^2} + j(\beta_1 + \kappa_{11} + \beta_2 + \kappa_{22})\tfrac{dX_2(z)}{dz} + \left[\kappa_{21}\kappa_{12} - (\beta_1 + \kappa_{11})(\beta_2 + \kappa_{22})\right]X_2(z) = 0
\end{aligned} \tag{5.129}$$

As indicated in equations (5.70), it is wise to go to new parameters. We find

$$\begin{aligned}
\Delta^2 &= \frac{1}{4}\left(\beta_1^2 + \beta_2^2 + \kappa_{11}^2 + \kappa_{22}^2 - 2\beta_1\beta_2 + 2\beta_1\kappa_{11} - 2\beta_1\kappa_{22} - 2\beta_2\kappa_{11} + 2\beta_2\kappa_{22} - 2\kappa_{11}\kappa_{22}\right) \\
&= \beta^2 - (\beta_1\beta_2 + \beta_1\kappa_{22} + \beta_2\kappa_{11} + \kappa_{11}\kappa_{22}) \\
&= \delta^2 - \kappa^2 = \beta^2 - (\beta_1 + \kappa_{11})(\beta_2 + \kappa_{22})
\end{aligned} \tag{5.130}$$

Hence, the last equation from (5.129), can be rewritten as

$$\frac{d^2X_2}{dz^2} + 2j\beta\frac{dX_2}{dz} + (\delta^2 - \beta^2)X_2 = 0 \tag{5.131}$$

A second order differential equation always has a solution of the form e^{rz}. In this case r has to satisfy the equation $r^2 + 2j\beta r + \delta^2 - \beta^2 = 0$. Thus, we can find a solution for $X_2(z)$:

$$X_2(z) = c_1 e^{-j\beta z}e^{-j\delta z} + c_2 e^{-j\beta z}e^{j\delta z} \tag{5.132}$$

We impose that only one of the modes is excited. $X_1(0) = 1$ and $X_2(0) = 0$

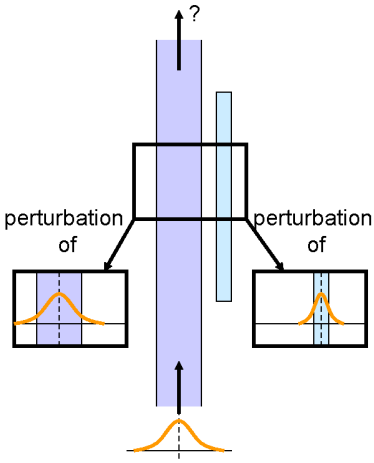


Figure 5.41: Double perturbed waveguide system

$$X_2(0) = c_1 + c_2 = 0 \tag{5.133}$$

With this information we can rewrite equation (5.132)

$$X_2(z) = 2jc_2e^{-j\beta z}\sin(\delta z) \tag{5.134}$$

The expression for $X_1(z)$ immediately follows from equation (5.127). We also impose initial conditions here.

$$\begin{array}{l} X_1(z) = \frac{1}{\kappa_{21}}\left(-2\delta c_2 e^{-j\beta z}\cos(\delta z) + 2j\left[\beta - (\beta_2 + \kappa_{22})\right]c_2 e^{-j\beta z}\sin(\delta z)\right) \\ X_1(0) = -\frac{2\delta c_2}{\kappa_{21}} = 1 \\ c_2 = -\frac{\kappa_{21}}{2\delta} \end{array} \tag{5.135}$$

From this we find the solution as depicted in equation (5.69)

$$\begin{array}{l} X_1(z) = e^{-j\beta z}\left[\cos(\delta z) - j\frac{\Delta}{\delta}\sin(\delta z)\right] \\ X_2(z) = e^{-j\beta z}\left[-j\frac{\kappa_{21}}{\delta}\sin(\delta z)\right] \end{array} \tag{5.136}$$

5.6.2 Calculation of the coupling coefficients $\kappa_{i,j}$ for a directional coupler

For the uncoupled waveguides with respective index profile $n_1(x)$ and $n_2(x)$ we can write

$$\begin{array}{l} \frac{\partial^2\varphi_1(x)}{\partial x^2} + (k_0^2 n_1^2(x) - \beta_1^2)\varphi_1(x) = 0 \\ \frac{\partial^2\varphi_2(x)}{\partial x^2} + (k_0^2 n_2^2(x) - \beta_2^2)\varphi_2(x) = 0 \end{array} \tag{5.137}$$

For the field in the coupled waveguides we can write

$$\frac{\partial^2\Psi(x,z)}{\partial x^2}+\frac{\partial^2\Psi(x,z)}{\partial z^2}+k_0^2n_{12}^2(x)\Psi(x,z)=0 \tag{5.138}$$

If we propose following solution for $\Psi(x,z)$

$$\Psi(x,z)=C_1(z)\varphi_1(x)e^{-j\beta_1 z}+C_2(z)\varphi_2(x)e^{-j\beta_2 z} \tag{5.139}$$

in which C_1 and C_2 slowly vary with z so that we can assume

$$\left|\frac{\partial^2 C_i}{\partial z^2}\right| << \left|-j\beta_i\frac{\partial C_i}{\partial z}\right| \tag{5.140}$$

Substituting equation (5.139) into equation (5.138) and taking equation (5.140) into account, we find that

$$\begin{array}{c}\left[\frac{\partial^2\varphi_1(x)}{\partial x^2}+(k_0^2n_{12}^2(x)-\beta_1^2)\varphi_1(x)\right]C_1e^{-j\beta_1 z}-2j\beta_1\varphi_1(x)\frac{dC_1}{dz}e^{-j\beta_1 z}\\ +\left[\frac{\partial^2\varphi_2(x)}{\partial x^2}+(k_0^2n_{12}^2(x)-\beta_2^2)\varphi_2(x)\right]C_2e^{-j\beta_2 z}-2j\beta_2\varphi_2(x)\frac{dC_2}{dz}e^{-j\beta_2 z}=0\end{array} \tag{5.141}$$

When we use equation (5.137), this becomes

$$\begin{array}{c}k_0^2\left[n_{12}^2(x)-n_1^2\right]\varphi_1(x)C_1e^{-j\beta_1 z}-j2\beta_1\varphi_1(x)\frac{dC_1}{dz}e^{-j\beta_1 z}\\ +k_0^2\left[n_{12}^2(x)-n_2^2\right]\varphi_2(x)C_2e^{-j\beta_2 z}-j2\beta_2\varphi_2(x)\frac{dC_2}{dz}e^{-j\beta_2 z}=0\end{array} \tag{5.142}$$

When the eigenmodes are orthonormal

$$\beta_i\int\varphi_i(x)\varphi_j(x)dx=\delta_{ij} \tag{5.143}$$

we find that (by multiplying equation (5.142) with φ_1 and integrating over x):

$$\frac{dC_1}{dz}e^{-j\beta_1 z}=-j\kappa_{11}C_1e^{-j\beta_1 z}-j\kappa_{12}C_2e^{-j\beta_2 z} \tag{5.144}$$

and by multiplying equation (5.142) with φ_2 and integrating over x:

$$\frac{dC_2}{dz}e^{-j\beta_2 z}=-j\kappa_{22}C_2e^{-j\beta_2 z}-j\kappa_{21}C_1e^{-j\beta_1 z} \tag{5.145}$$

with

$$\begin{array}{l}\kappa_{11}=\frac{1}{2}k_0^2\int(n_{12}^2-n_1^2)\psi_1^2dx\\ \kappa_{12}=\frac{1}{2}k_0^2\int(n_{12}^2-n_1^2)\psi_1\psi_2dx\\ \kappa_{21}=\frac{1}{2}k_0^2\int(n_{12}^2-n_2^2)\psi_1\psi_2dx\\ \kappa_{22}=\frac{1}{2}k_0^2\int(n_{12}^2-n_2^2)\psi_2^2dx\end{array} \tag{5.146}$$

When

$$\begin{aligned} X_1(z) &= C_1(z)e^{-j\beta_1 z} \\ X_2(z) &= C_2(z)e^{-j\beta_2 z} \end{aligned} \tag{5.147}$$

equation (5.144) and (5.145) become

$$\begin{aligned} \frac{dX_1}{dz} &= -j\beta_1 X_1 - j(\kappa_{11}X_1 + \kappa_{12}X_2) \\ \frac{dX_2}{dz} &= -j\beta_2 X_2 - j(\kappa_{21}X_1 + \kappa_{22}X_2) \end{aligned} \tag{5.148}$$

These are the equations proposed by the coupled mode theory.

Chapter 6

Periodic Structures

Contents

6.1 Introduction

Periodic structures have several applications in optics. Highly reflective mirrors, grating couplers, diffraction gratings for optical filters, monochromators and spectrum analyzers are only a few. Especially because of the introduction of wavelength division multiplexing (WDM) in optical fiber communication, gratings are becoming indispensable for various filter functions. We will start by analyzing diffraction at surface gratings, based on the ”thin lens” approximation and Fourier optics in section 2. In section 3 and 4 some general properties of periodic structures (Floquet-Bloch theorem and the Bragg condition) are deduced. In section 5 coupled wave theory for periodically layered media based on a perturbation analysis is described, while in section 6 the realization of a periodic structure using an acoustical wave and its applications are presented.

The periodic nature of the structures that are described in the chapter, more in particular the periodic variation of the refractive index around a mean value, implies the interference of a large number of scattered waves. Therefore, the optical effects are often very selective in wavelength, propagation direction and polarization.

Different classes of periodic structures exist. They are classified according to the refractive index contrast, the volume over which the periodicity occurs, the ratio of the period to the wavelength of the light etc. Because the transition between the different classes of periodic structures is often

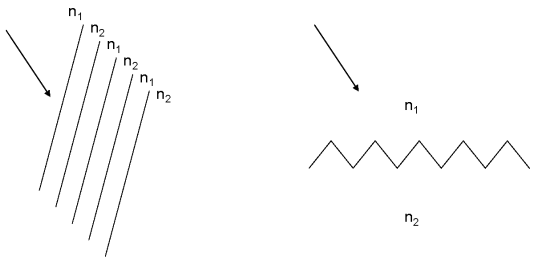


Figure 6.1: Classification of gratings: volume periodic grating and surface periodic grating

vague, these classifications are more qualitative than quantitative. The classification that we will apply here, is the one of the volume of the periodicity. In a volume periodic structure the interaction between an incident field and the periodic structure does not occur at the interface of two media. In a surface periodic structure only the interface between two media is corrugated.

In a volume periodic structure the refractive index is periodically modulated while in a surface periodic structure an interface between two media is periodically modulated. The shape of the modulation is in principal arbitrary but is in practice determined by technological limitations. Some applications of periodic structures are: filters, monochromators, DBR and DFB lasers, in- and out-coupling gratings, diffractive lenses, modulators, holography and X-ray analysis of crystals.

6.2 Diffraction at surface gratings

6.2.1 Approximate transmission theory for thin surface gratings

As light is an electromagnetic wave, one has to solve the vectorial Maxwell equations with the correct boundary conditions when trying to solve a diffraction problem at a periodic medium. Due to the complexity, approximate theories were developed with a limited applicability, but which lead to a solution in a faster and easier way. It is however often unclear where the applied approximate theory is no longer valid. Therefore, in the case of doubt the rigorous correct solution to the Maxwell equations needs to be found.

An approximate theory which we will use in this section is the so called transmission theory. This theory assumes that the scalar field immediately behind the grating can be obtained by simply multiplying the incident field with a transmission function. This means that the transmission theory relates the incident and transmitted field locally, opposed to the integral relations of Fresnel and Fraunhofer diffraction. Transmission theory therefore only applies when the thickness of the periodic media is sufficiently small (which is the same assumption as for the analysis of a thin lens). Based on physical arguments one then still has to find and appropriate shape for the transmission function $t(x)$.

Another approximation which is often made is to neglect the vectorial nature of light as in the above mentioned Fresnel and Fraunhofer diffraction theories.

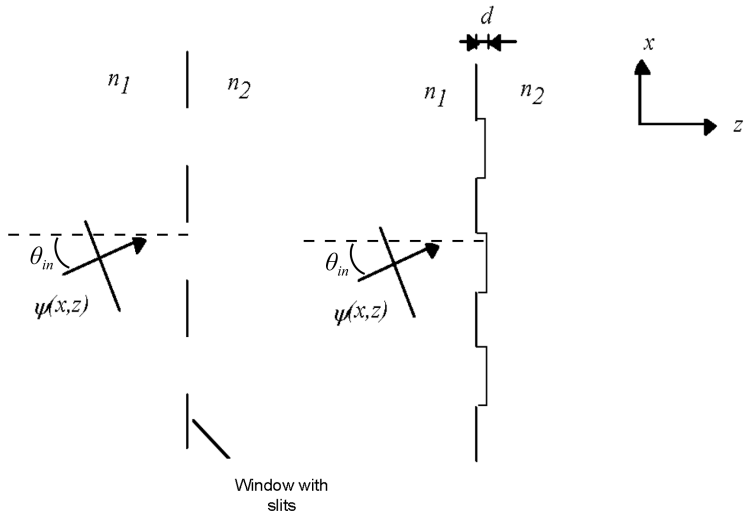


Figure 6.2: Plane wave incident on a surface grating: window with slits and dielectric grating

We will restrict ourself in this section to surface gratings that can be described with transmission theory. This implies that the thickness of the grating (as a separation layer between two homogenous media) is assumed to be small. The surface grating is presented as a periodic arrangment of slits or a very shallow binary grating (figure 6.2).

We ask ourself how an incident (plane) wave is diffracted, in reflection or in transmission. Say we work in transmission, then we can write down the following relation between the incident and the transmitted field (grating between $z = 0-$ and $z = 0+$):

$$\psi\left(x, 0^{+}\right) = t\left(x\right) \psi\left(x, 0^{-}\right) \tag{6.1}$$

with $\psi\left(x, 0^{-}\right)$ the field incident on the grating, $\psi\left(x, 0^{+}\right)$ the transmitted field after the grating and $t(x)$ the transmission function of the grating.

This equation relates the plane wave decomposition of the transmitted field to the plan wave decomposition of the incident field.

The transmission function is assumed to be zero outside the grating. $t(x)$ can be written as

$$\sum_{n=1}^{N} t_1\left(x - x_n\right) \qquad x_n = \left(n - 1\right) \Lambda \tag{6.2}$$

with Λ the period of the grating and N the number of grating periods. In this equation $t_1(x)$ $(x \in [0, \Lambda])$ describes the transmission within one grating period. Fourier analysis of $\psi\left(x, 0^{+}\right)$, which means decomposition in plane waves of the transmitted field, shows that

$$F\left(\psi\left(x, 0^{+}\right)\right) = F\left(t\left(x\right)\right) * F\left(\psi\left(x, 0^{-}\right)\right) \tag{6.3}$$

We will write the Fourier transform of the transmission function $t(x)$ as $T(f_x)$. This function is proportional to $T_1(f_x)$, being the Fourier transform of the transmission function within one grating period $t_1(x)$:

$$\begin{aligned}
T(f_x) &\triangleq F(t(x)) \\
&= F\left(\sum_{n=1}^{N} t_1(x - x_n)\right) \\
&= \int_{-\infty}^{+\infty} dx e^{-j2\pi f_x x} \left(\sum_{n=1}^{N} t_1(x - x_n)\right) \\
&= \sum_{n=1}^{N} e^{-j2\pi f_x x_n} \underbrace{\int_{-\infty}^{+\infty} d(x - x_n)\, e^{-j2\pi f_x (x - x_n)} t_1(x - x_n)}_{T_1(f_x)} \\
&= T_1(f_x) . \sum_{m=0}^{N-1} e^{jm\delta} \quad ; \ \delta = -2\pi f_x \Lambda \\
&= T_1(f_x) . e^{j(N-1)\delta/2} . \frac{\sin N\delta/2}{\sin \delta/2}
\end{aligned} \tag{6.4}$$

In this way, we can relate the incident and transmitted field via the Fourier transform. $T(f_x)$ contains two effects:

- the effect of the detailed structure of each period: $T_1(f_x)$
- the effect of the periodic (but finite) nature of the grating: $\frac{\sin N\delta/2}{\sin \delta/2}$

For a finite number of slits the Fourier transform of the scalar field behind the grating is shown in figure 6.3 when a plane incident wave is assumed.

The function $T(f_x)$ consists of a slowly varying envelope $T_1(f_x)$, multiplied with the periodic function $\sin(N\delta/2)/\sin(\delta/2)$. The latter has sharp periodic peaks with weak side lobes in between. The higher the number of grating periods, the sharper the peaks and the weaker the side lobes. Each peak is called a *diffraction order*. The peak at $f_x = 0$ is the *zeroth order* diffraction. Then follow the first and minus first order diffraction orders etc. If N is infinite, the incident plane wave is rigorously diffracted into a discrete set of diffraction orders. For finite N, each discrete diffraction order broadens and forms a beam with a finite cross-section and a diverging far-field.

If the incident beam is a plane wave incident normal to the grating ($\theta_{in} = 0$), it is obvious that:

$$F\left(\psi(x, 0^+)\right) = T(f_x) \tag{6.5}$$

Hence one finds that the diffracted field consists of diffraction orders for which:

$$f_x = \pm \frac{m}{\Lambda} \tag{6.6}$$

Since $f_x = \sin(\theta)/\lambda$ this leads to the following expression for the angle of the m-th diffraction order:

$$\sin(\theta_m) = \pm m \frac{\lambda}{\Lambda} \tag{6.7}$$

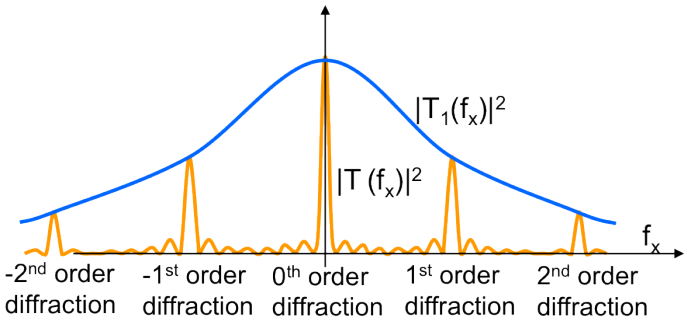


Figure 6.3: Transmission theory applied to a grating consisting of a finite number of slits

If the incident field hits the grating with a general angle θ_{in}, this equation generalizes to:

$$\boxed{\sin(\theta_m) - \sin(\theta_{in}) = \pm m\frac{\lambda}{\Lambda}} \tag{6.8}$$

This equation is called the grating equation.

As a special case, consider the blazed grating configuration from figure 6.4. A plane wave is incident on the surface grating which has a linear profile. The transmission function inside one period is

$$t_1(x) = e^{+j\frac{2\pi}{\lambda}(n_2-n_1)xd/\Lambda}e^{-j\frac{2\pi d}{\lambda}n_2} \tag{6.9}$$

and the Fourier transform

$$\begin{aligned} T_1(f_x) &= e^{-j2\pi dn_2/\lambda}\int_0^{\Lambda} e^{-j2\pi\left(f_x-\frac{d}{\lambda\Lambda}(n_2-n_1)\right)x}dx \\ &= e^{-j2\pi dn_2/\lambda}e^{-j\pi\left(f_x-\frac{d}{\lambda\Lambda}(n_2-n_1)\right)\Lambda}\frac{\sin(\pi(f_x-f_o)\Lambda)}{\pi(f_x-f_o)} \end{aligned} \tag{6.10}$$

with $f_o = \frac{d}{\lambda\Lambda}(n_2 - n_1)$.

So

$$|T(f_x)|^2 = \frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)}.\frac{\sin^2(\pi(f_x-f_o)\Lambda)}{(\pi(f_x-f_o))^2}, \quad \delta = -2\pi f_x\Lambda \tag{6.11}$$

The term

$$\frac{\sin^2(N\delta/2)}{\sin^2(\delta/2)} \tag{6.12}$$

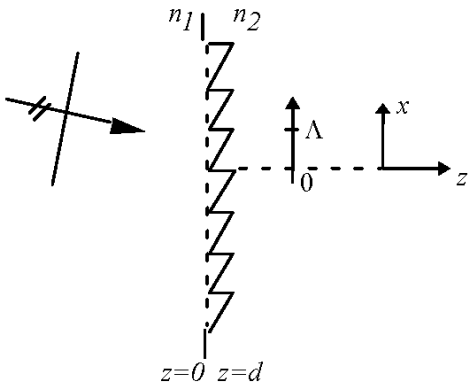


Figure 6.4: Blazed grating

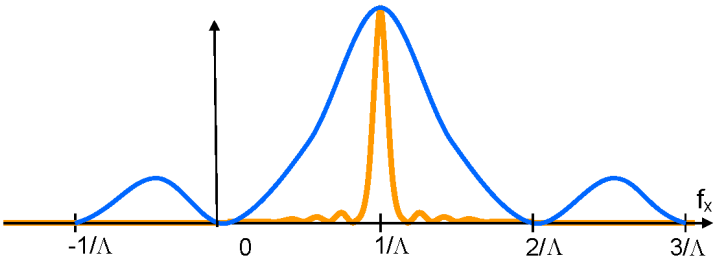


Figure 6.5: Transmission of a blazed grating

is maximal when $\delta = m2\pi$ with m an integer or when $f_x = \frac{m}{\Lambda}$. Outside these maxima this term is very small. The $sinc^2$ function is zero when $\pi(f_x - f_0)\Lambda = k\pi$ for k not equal to zero. When k is zero the $sinc^2$ function is maximal. So when we choose $f_0 = \frac{1}{\Lambda}$, then both terms are maximal for $f_x = f_0$ while outside the zeros of the sinc coincide with the maxima of the other term. When $f_0 = \frac{1}{\Lambda}$, we find that $|T(f_x)|^2$ is only significant when f_x lies around $\frac{1}{\Lambda}$ (figure 6.5). So this is diffraction to only 1 diffraction order. This configuration is called a blazed grating.

This structure strongly resembles a Fresnel lens. Both Fresnel lens and blazed grating have a sawtooth structure, but in the case of Fresnel lenses the period is much larger. In the case of a Fresnel lens the direction of the transmitted light is determined by Snell's law, but in a blazed grating it is determined by Bragg diffraction. One can easily show however that if $f_0 = \frac{1}{\Lambda}$ both directions are identical and both structures show a very similar behaviour.

6.2.2 Application: spectrometer

Consider a surface grating consisting of slits such that

$$t_1(x) = \begin{cases} 1 & 0 \leq x \leq l \\ 0 & l \leq x \leq \Lambda \end{cases} \tag{6.13}$$

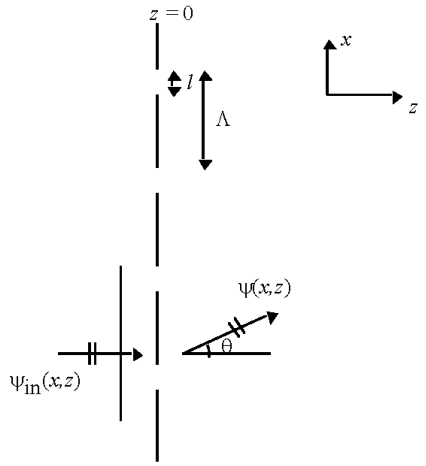


Figure 6.6: Grating spectrometer

So that

$$T_1(f_x) = \int_0^l e^{-j2\pi f_x x} dx = e^{-j\pi f_x l} \frac{\sin(\pi f_x l)}{\pi f_x} \tag{6.14}$$

This means that

$$T(f_x) = e^{-j\pi f_x l} e^{-j(N-1)\pi\Lambda f_x} \frac{\sin(\pi f_x \Lambda N)}{\sin(\pi f_x \Lambda)} \frac{\sin(\pi f_x l)}{\pi f_x} \tag{6.15}$$

Assume a plane wave is incident on the surface grating along the z-axis (figure 6.6) with amplitude 1:

$$\psi\left(x, 0^-\right) = 1 \text{ and so } F\left(\psi\left(x, 0^-\right)\right) = \delta(f_x) \tag{6.16}$$

The transmitted field then becomes

$$F\left(\psi\left(x, 0^+\right)\right) = \int_{-\infty}^{+\infty} \delta\left(f_x'\right) T\left(f_x - f_x'\right) df_x' = T(f_x) \;\; and \;\; f_x = \frac{\sin\theta}{\lambda} \tag{6.17}$$

Preferably the function $T(f_x)$ will be a sharply peaked function of f_x. This means that in the transmitted optical field there will be a well defined relation between the angle θ and the wavelengths. This property can be used to spatially separate a beam into its constituent wavelength components. Of course one will make sure that the peak in $T(f_x)$ will not occur at $f_x = 0$, because this won't introduce any wavelength selectivity ($sin(\theta) = 0$) for all wavelengths.

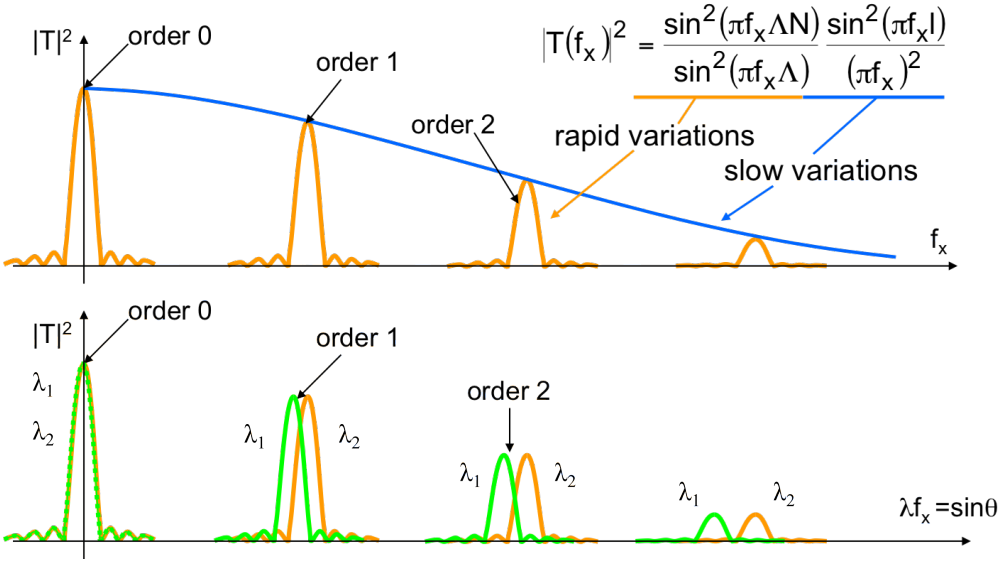


Figure 6.7: Grating spectrometer: decomposition in wavelength components

The angle dependence of the higher (first) order diffraction is used in grating spectrometers. Two important properties of a spectrometer are its resolution and its free spectral range. The free spectral range expresses the maximum wavelength span over which the spectrometer works unambiguously. The resolution expresses what is the minimal $\Delta\lambda$ that can be detected and is mostly defined by the Rayleigh criterion.

Assume the spectrometer works for m-th order diffraction. This implies that the Rayleigh criterion says that $\Delta\lambda$ is determined by stating that the maximum of the m-th order diffraction for λ_1 coincides with the first minimum in the m-th order diffraction of λ_2. Determining these two wavelengths λ_1 and λ_2 determines the resolution of the spectrometer, $\Delta\lambda = |\lambda_2 - \lambda_1|$.

So

$$|T(f_x)|^2 = \frac{\sin^2(\pi f_x \Lambda N)}{\sin^2(\pi f_x \Lambda)} \frac{\sin^2(\pi f_x l)}{(\pi f_x)^2} \tag{6.18}$$

The first factor in this equation is a rapidly varying while the second one varies slowly. For $\Delta\lambda = |\lambda_2 - \lambda_1|$ small, we will assume this second factor to be identical for λ_1 and λ_2. When the first factor has to be maximal for λ_1 and zero for λ_2 for the m-th order diffraction, then

$$[N f_x \pi \Lambda]_{\lambda_1} - [N f_x \pi \Lambda]_{\lambda_2} = \pi \quad \text{and} \quad [f_x]_{\lambda_1} - [f_x]_{\lambda_2} = \frac{1}{N\Lambda} \tag{6.19}$$

For an m-th order diffraction the light is diffracted in such a way that

$$\frac{2\pi}{\lambda} \sin\theta = m \frac{2\pi}{\Lambda} \tag{6.20}$$

so that

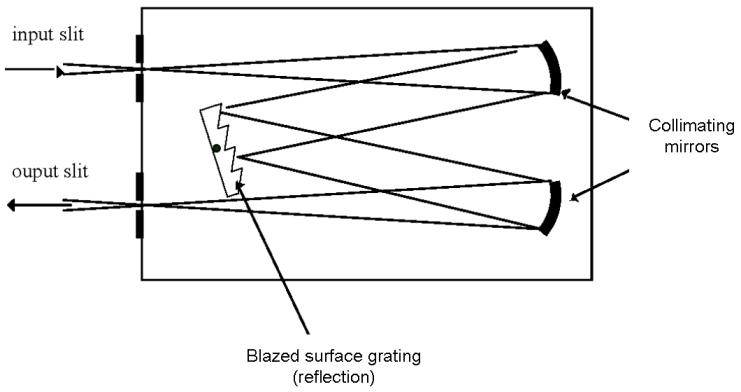


Figure 6.8: Czerny-Turner monochromator

$$f_x = \frac{m}{\Lambda} \tag{6.21}$$

or

$$\frac{f_x}{\Delta f_x} = \frac{\lambda}{\Delta\lambda} = mN \tag{6.22}$$

with N the number of slits in the grating.

The resolution can therefore be very large when a large grating with a lot of periods is used, but is however not as large as in a Fabry-Perot etalon. The free spectral range is however much larger in a grating spectrometer and is determined by the condition that for a certain angle θ the m-th order diffraction of λ_1 coincides with the $(m+1)$-th order diffraction of λ_2:

$$\left.\begin{array}{l} \frac{\sin\theta}{\lambda_1} = \frac{m}{\Lambda} \\ \frac{\sin\theta}{\lambda_2} = \frac{m+1}{\Lambda} \end{array}\right\} \Rightarrow \frac{\lambda_1 - \lambda_2}{\lambda_2} = \frac{1}{m} \quad (\sin\theta \text{ constant}) \tag{6.23}$$

6.2.3 Application: Czerny-Turner monochromator

Light enters the monochromator via the input slits and is incident to a mirror which collimates the light onto a blazed surface grating that works in reflection. This grating will diffract the different wavelength components under different angles towards the second mirror. This mirror translates these angle variations into spatial separations at the exit slit of the monochromator. By tilting the diffraction grating another wavelength component is exactly focused in the exit slit. The orientation angle α of the grating corresponds with one wavelength in the exit slit. Detection of the power at the exit slit as a function of the angle α results in the spectral decomposition of the incident light.

For a good operation it is mandatory to illuminate a grating which is as large as possible and to increase the distance between the mirrors and the grating to increase the spatial resolution (and

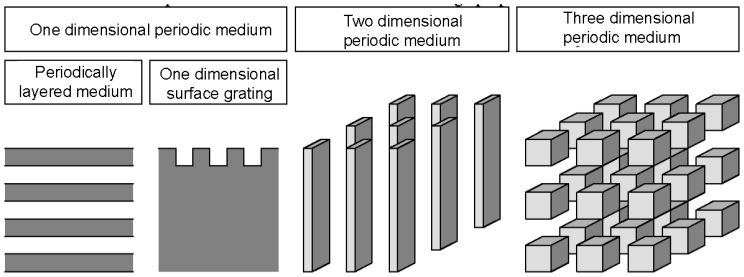


Figure 6.9: Classification of periodic structures

therefore also the wavelength resolution). A resolution of 0.1nm requires a box of 0.5m to 1m length.

6.3 Bragg condition and k-vector diagram

6.3.1 Periodicity and reciprocal lattice

The term periodicity is only defined for completely translation-invariant structures and is thus only applicable for infinitely extending structures. These translations always occur over a finite distance larger than zero, in other words a direction in which the structure is invariant, is not defined as a direction of periodicity. One says that the structure is one, two or three dimensionally periodic, when one can find one, two or three linearly independent translation vectors $\mathbf{r_i}$ for which a translation over these vectors leaves the structure invariant. These independent directions of periodicity are not unique: an infinite stacking of squares in a plane have different sets of two independent directions of periodicity. In general, when $\mathbf{r_i}$ are translation vectors, also the vector $\mathbf{r} = \mathbf{lr_1} + \mathbf{mr_2} + \mathbf{nr_3}$ will be a translation vector, with l,m and n integers.

With a direction of periodicity we can associate a $\mathbf{K}$-vector with a length $|\mathbf{K}| = 2\pi/\Lambda$ in which Λ is the period in that direction. Out of the one, two or three independent $\mathbf{K}$-vectors one can generate al possible $\mathbf{K}$-vectors of a periodic structure, associated with the lattice points of the reciprocal lattice. For a three dimensional structure with periods Λ_1,Λ_2 and Λ_3, one can write $\mathbf{K} = \mathbf{pK_1} + \mathbf{qK_2} + \mathbf{rK_3}$ with $|\mathbf{K}_i| = 2\pi/\Lambda_i$. In the appendix, we show that the relation between the translation vectors and the $\mathbf{K}$-vectors can be written as $\mathbf{r_i} \cdot \mathbf{K_j} = \mathbf{2}\pi\delta_{\mathbf{ij}}$.

The unit cell of the periodic structure is defined as the smallest part you have to stack to get the periodic structure. The concept of a unit cell actually is the multi dimensional equivalent of the one dimensional concept of a period.

In the three spatial dimensions the refractive index can be periodic, constant or random. When the refractive index is constant in all directions, this is called a homogenous space. A configuration showing periodicity in one direction is called a one dimensional periodic medium. When this medium is made out of a periodic stacking of layers, this is called a periodic layered medium. When there are two (three) directions of periodicity, this is called a two (three) dimensional periodic medium. Although perfectly periodic structures extend to infinity, in reality structures will

be finite. Therefore, in practice the properties of periodic media can slightly deviate from the theoretical predictions for infinitely extending media.

6.3.2 Bragg condition

The propagation of an arbitrary wave in a uniform medium can be considered as a set of plane waves which propagate independently (without power exchange). When the medium is periodic however the plane waves will be scattered. The periodic nature of the medium will give rise to very specific coupling between certain plane waves of the set (or between specific directions). This is described by the so called Bragg condition. One says that the plane waves are diffracted to different propagation directions, which are called diffraction orders.

Actually, the Bragg condition is a generalization of Snells law. It relates the **k**-vector of the incident plane wave and the **k**-vectors of the diffraction orders. Just like in the case of refraction at homogenous interfaces this relation between the **k**-vectors can be easily graphically depicted (**k**-vector diagram). The main goal of this section is to deduce this Bragg condition from the scalar wave equation. We will use a perturbation method which will be subsequently applied to an arbitrary (not necessarily periodic) medium, a three dimensional volume periodic medium and a two dimensional surface periodic medium.

Deduction of the field equation by means of perturbation theory

Consider an arbitrary medium, in which the scalar field (given by $\psi(\mathbf{r})$) satisfies the scalar field equation

$$\nabla^2\psi\left(\mathbf{r}\right) + k_0^2 n^2\left(\mathbf{r}\right)\psi\left(\mathbf{r}\right) = 0 \tag{6.24}$$

in which k_0 is the free space wave vector ($k_0 = \frac{2\pi}{\lambda_0}$). We assume that the medium only differs slightly from a homogenous space, such that we can write $n^2(\mathbf{r})$ as

$$n^2\left(\mathbf{r}\right) = n_0^2 + \Delta n^2\left(\mathbf{r}\right) \tag{6.25}$$

In this equation n_0 is the refractive index of the homogenous background medium and Δn the perturbation. This perturbation is small ($\Delta n^2 << n_0^2$). The scalar field can therefore be written as the sum of two contributions, namely the field ψ_0 in absence of the perturbation Δn and the correction ψ_1 due to the perturbation Δn

$$\psi = \psi_0 + \psi_1 \tag{6.26}$$

If we substitute this equation into the scalar field equation we get

$$\nabla^2\psi_0 + \nabla^2\psi_1 + k_0^2 n_0^2\psi_0 + k_0^2 n_0^2\psi_1 + k_0^2\Delta n^2\psi_0 + k_0^2\Delta n^2\psi_1 = 0 \tag{6.27}$$

Because ψ_0 satisfies the scalar wave equation in the absence of a perturbation ($\nabla^2\psi_0 + k_0^2 n_0^2\psi_0 = 0$) and because the term $k_0^2\Delta n^2\psi_1$ can be neglected, we find

$$\nabla^2\psi_1 + k_0^2 n_0^2 \psi_1 = -k_0^2 \Delta n^2 \psi_0 \tag{6.28}$$

In this equation for ψ_1, $k_0^2 \Delta n^2 \psi_0$ is considered a source term, which actually implies that multiple interference between the fields ψ_0 and ψ_1 is neglected (due to neglecting the second order term).

Arbitrary medium

As a field in the homogenous medium we consider a plane wave

$$\psi_0(\mathbf{r}) = e^{-j\mathbf{k_{in}}\cdot\mathbf{r}} \tag{6.29}$$

Both the perturbation Δn^2 as the perturbation field ψ_1 can be represented by their respective Fourier transforms (in three dimensions)

$$\Delta n^2(\mathbf{r}) = \int\int\int A(\mathbf{k}) e^{-j\mathbf{k}\cdot\mathbf{r}} dk_x dk_y dk_z \tag{6.30}$$

$$\psi_1(\mathbf{r}) = \int\int\int B(\mathbf{k}) e^{-j\mathbf{k}\cdot\mathbf{r}} dk_x dk_y dk_z \tag{6.31}$$

When we substitute this equation in equation 6.28 (which actually means doing a 3D Fourier transform of this equation) we find

$$B(\mathbf{k}) = \frac{-k_0^2 A(\mathbf{k} - \mathbf{k_{in}})}{k_0^2 n_0^2 - |\mathbf{k}|^2} \tag{6.32}$$

To obtain this result the orthogonality of the functions $e^{-j\mathbf{k}.\mathbf{r}}$ was used to obtain an equation for each individual Fourier component.

The following important considerations concerning the perturbation field ψ_1 can be made:

- When the refractive index perturbation Δn contains a Fourier component with spatial frequency $\mathbf{k}$, then (and only then) the total field will contain a spatial component $\mathbf{k} + \mathbf{k_{in}}$
- The total field contains predominantly plane waves of which the $\mathbf{k}$-vectors nearly satisfy the dispersion relation of the unperturbed medium ($|\mathbf{k}| = k_0 n_0$) as for other $\mathbf{k}$-vectors $B(\mathbf{k})$ becomes very small.
- Equation 6.32 suggests that $B(\mathbf{k})$ goes to infinity for $|\mathbf{k}| = k_0 n_0$. This non-physical behavior is caused by the perturbation approximation (and by the infinite extension of the medium).

Three dimensional volume periodic medium

In the case of a three dimensional periodic medium, the Fourier decomposition for Δn^2 becomes (note that this is a special case of the more general discussion in the previous section)

$$\Delta n^2 (\mathbf{r}) = \sum_{m,n,l} A_{mnl} e^{-j\mathbf{K_{mnl}}\mathbf{r}} \tag{6.33}$$

in which $\mathbf{K_{mnl}}$ are the **k**-vectors of the lattice. For a lattice with orthogonal base vectors $\mathbf{r_i}$ we can write

$$\mathbf{K_{mnl}} = m\mathbf{K_x} + n\mathbf{K_y} + l\mathbf{K_z} \qquad m, n, l \text{integer} \tag{6.34}$$

with $\mathbf{K_i} = \frac{\mathbf{2\pi r_i}}{|\mathbf{r_i}|^\mathbf{2}}$

Taking into account the remarks from the previous section we can say that the perturbation field ψ_1 only contains spatial components

$$\mathbf{k}'_{\mathbf{mnl}} = \mathbf{K_{mnl}} + \mathbf{k_{in}} \tag{6.35}$$

$$\psi_1 (\mathbf{r}) = \sum_{m,n,l} B_{mnl} e^{-j\mathbf{k}'_{\mathbf{mnl}}\cdot\mathbf{r}} \tag{6.36}$$

This condition is called the Bragg condition and shows the relation between the direction of the exciting and diffracted waves in the periodic medium. Applying the results from the previous section immediately gives the expansion coefficients of the perturbation field

$$B_{mnl} = \frac{-k_0^2 A_{mnl}}{+k_0^2 n_0^2 - \left|\mathbf{k}'_{mnl}\right|^2} \tag{6.37}$$

We find discrete directions of plane waves (from the Bragg condition). Again we see that there will be only important contributions to the field, for which $|\mathbf{k}'_{mnl}|^2 \approx k_0^2 n_0^2$ (meaning that these contributions nearly satisfy the dispersion relation of the unperturbed medium).

These two properties can be graphically represented as follows: Around $\mathbf{k_{in}}$ we draw a sphere with radius $k_0 n_0$. The **k**-vectors of the diffraction orders are found by adding integer multiples of $\mathbf{K_x}$, $\mathbf{K_y}$, and $\mathbf{K_z}$, until again the surface of the sphere with radius $k_0 n_0$ is reached. Note that the coupling between different plane waves is determined by the variation of Δn^2. The harmonics that build Δn^2 determine the **k**-vectors of the diffraction orders. When we consider a sinusoidally varying grating, there is only one diffraction order (at least within the assumption of a perturbation analysis, i.e. when Δn is small). The Bragg condition is valid both in reflection and in transmission. All **k**-vectors originate from the same point.

The Bragg condition and dispersion relation imply that diffraction at a three dimensional lattice leads to discrete diffraction orders. Moreover, these orders only exist for certain directions of incidence. This situation is well known in the diffraction of X-rays at crystals.

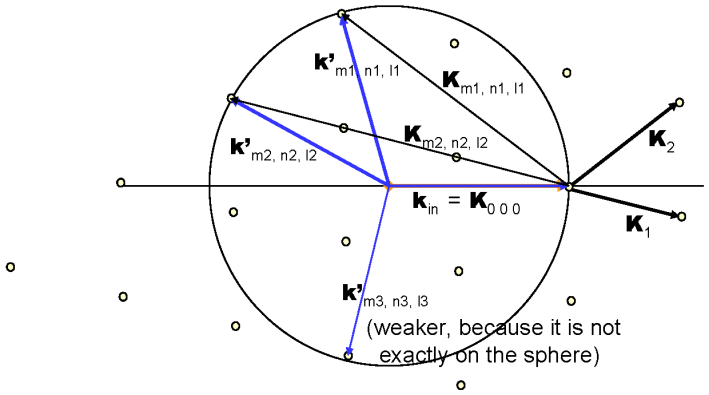


Figure 6.10: k-vector diagram for a three dimensional periodic medium

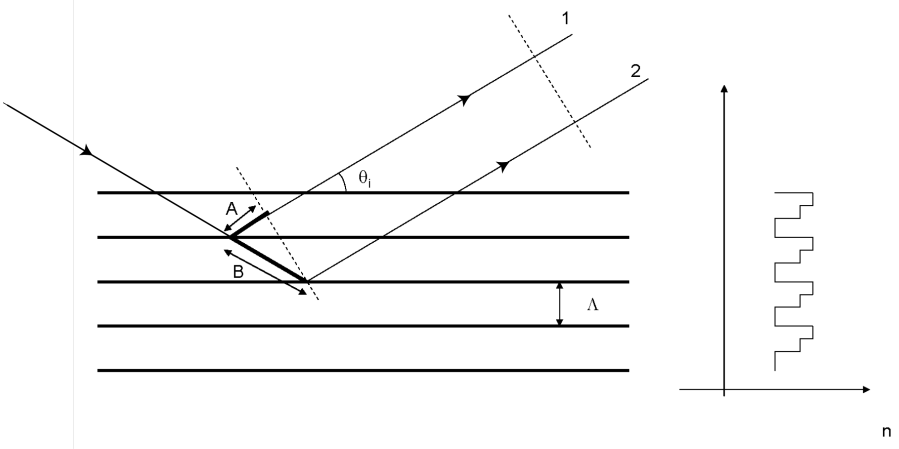


Figure 6.11: Diffraction at a layered medium: relation between interfering waves and Bragg condition

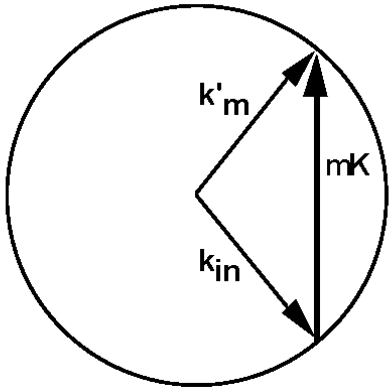


Figure 6.12: k-vector diagram for a one dimensional periodic layered medium

Let's consider the special case of a one dimensional periodic layered medium. We assume a periodic stacking of layers with period Λ. The angles under which the reflected and transmitted wave leave a layer are the same as the angle of incidence of the exciting wave. We calculate the phase difference between wave 1 and wave 2 in figure 6.11. This phase difference $\Delta\phi$ becomes

$$\Delta\phi = \frac{2\pi n_0}{\lambda}\left|B - A\right| = \frac{2\pi n_0}{\lambda} B\left(1 - \cos 2\theta_i\right) = \frac{2\pi n_0}{\lambda}\frac{\Lambda}{\sin\theta_i}\left(1 - \cos 2\theta_i\right) = \frac{4\pi n_0}{\lambda}\Lambda\sin\theta_i \tag{6.38}$$

A strong reflection will occur when both waves (and thereby also the contributions from the subsequent periods) are in phase. This happens when $|\Delta\phi| = 2m\pi$ and therefore

$$\Lambda\sin\theta_i = \frac{m\lambda}{2n_0} \tag{6.39}$$

If we consider for the same geometry the z-components of the **k**-vectors of the incident and reflected waves, then we can write

$$\begin{array}{l} k_z = \frac{2\pi}{\lambda} n_0 \sin\theta_i \\ k'_z = -\frac{2\pi}{\lambda} n_0 \sin\theta_i \end{array} \tag{6.40}$$

When both waves need to be in phase, $|k_z - k'_z|$ is given by

$$\begin{array}{l} |k_z - k'_z| = 2\sin\theta_i \frac{2\pi}{\lambda} n_0 \\ = m\frac{2\pi}{\Lambda} \end{array} \tag{6.41}$$

This exactly is the Bragg condition. This condition therefore expresses that the different reflections have to be in phase. The **k**-vector diagram for this situation is depicted in figure 6.12.

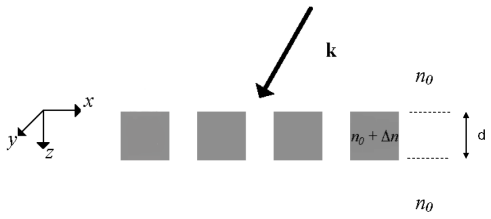


Figure 6.13: Perturbation analysis for 1D and 2D periodic media: configuration

Surface periodic medium (1D or 2D)

Now we will consider media in which there is no periodicity in the third dimension and where the periodic medium is located in a thin layer. A wave is incident to this thin layer. The layer is one or two dimensionally periodically modulated. Here, we will assume a two dimensional periodicity of the refractive index profile.

This situation can be analyzed in two ways:

- via a three dimensional Fourier transform of the refractive index profile $\Delta n^2(x, y, z)$ (as a special case of the arbitrary medium in a previous section)
- via a two dimensional Fourier transform of the refractive index profile $\Delta n^2(x, y, z)$.

We will use the three dimensional Fourier transform. As $\Delta n^2(x, y, z)$ is periodic in x and y it contains discrete Fourier components in these directions while in the third direction a continuous spectrum of spatial frequencies exists. Therefore, we propose as a spectral decomposition of $\Delta n^2(x, y, z)$ following equation

$$\Delta n^2\left(x, y, z\right) = \int_{-\infty}^{+\infty} dk_z \sum_{l,m} A_{lm}\left(k_z\right) e^{-j\mathbf{K}_{lmz}\cdot\mathbf{r}} \tag{6.42}$$

with

$$\mathbf{K_{lmz}} = lK_x\mathbf{u_x} + mK_y\mathbf{u_y} + k_z\mathbf{u_z} \qquad m, l \text{ integer} \tag{6.43}$$

in which K_x and K_y are the magnitude of the lattice vectors in x and y directions, k_z varies continuously and $\mathbf{u_x}$,$\mathbf{u_y}$ and $\mathbf{u_z}$ are unit vectors in x,y and z directions. The perturbation field ψ_1 becomes

$$\psi_1 = \int_{-\infty}^{+\infty} dk'_z \sum_{l,m} B_{lm}\left(k'_z\right) e^{-j\mathbf{k}'_{lmz}\cdot\mathbf{r}} \tag{6.44}$$

From the previous sections we know that a spatial component $\mathbf{k}$ in the refractive index profile results in a spatial component $\mathbf{k} + \mathbf{k_{in}}$ in the perturbation field ψ_1, so

$$\mathbf{k}'_{\mathbf{lmz}} = \mathbf{K}_{\mathbf{lmz}} + \mathbf{k}_{\mathbf{in}} \tag{6.45}$$

and the expansion coefficients of the perturbation field become

$$B_{lm}\left(k'_z\right) = \frac{-k_0^2 A_{lm}\left(k_z\right)}{+k_0^2 n_0^2 - \left|\mathbf{k}'_{lmz}\right|^2} \tag{6.46}$$

Note that here, the Bragg condition does not impose strict conditions to the allowed k_z-values of the scattered waves. As there is no periodicity in the z direction, the z-components of $\mathbf{K}_{\mathbf{lmz}}$ and $\mathbf{k}'_{\mathbf{lmz}}$ can vary continuously.

Again we see that $B_{lm}\left(k'_z\right)$ only has an important contribution to ψ_1 when

$$\left|\mathbf{k}'_{lmz}\right| = k_0 n_0 = |\mathbf{k}| \tag{6.47}$$

The presence of the semi infinite spaces with refractive index n_0 rigourously determine the dispersion relation in these areas. Together with the Bragg condition in the x and y direction (in the z direction the Bragg condition does not impose discrete k_z values), this dispersion relation determines the direction of allowed waves in these media.

Furthermore, $A_{lm}\left(k_z\right)$ plays an important role. Say for example that $\Delta n^2(x,y,z)$ has the following z dependence

$$\Delta n^2\left(x,y,z\right) = \begin{cases} \Delta n^2\left(x,y\right) & 0 \leq z \leq d \\ 0 & z < 0, z > d \end{cases} \tag{6.48}$$

When we represent the Fourier components of the function $\Delta n^2(x,y)$ by a_{lm}, $A_{lm}\left(k_z\right)$ becomes

$$A_{lm}\left(k_z\right) = \int_0^d a_{lm} e^{jk_z.z} dz = a_{lm} \frac{e^{jk_z.d} - 1}{jk_z} = -a_{lm} e^{jk_z d/2} d\mathrm{sinc}\left(k_z d/2\right) \tag{6.49}$$

We see that $A_{lm}\left(k_z\right)$ is a continuous function of $k_z = \left(\mathbf{k}_{\mathbf{in}} - \mathbf{k}'_{\mathbf{lmz}}\right) \cdot \mathbf{u}_{\mathbf{z}}$. This means that when the incident wave changes direction, we will always have diffraction to a number of diffraction orders which satisfy the Bragg condition. The amplitude of the diffraction orders are determined by $A_{lm}\left(k_z\right)$. Ideally the sinc-factor is 1 (this means that the k_z-component of the diffracted wave is the same as that of the incident wave).

Changing the direction of the incident wave results in diffraction to a discrete number of diffraction orders (which also change direction), opposite to diffraction in a three dimensional periodic medium where we only get diffraction for certain angles of incidence.

Via the three dimensional Fourier transform we find that the Bragg condition only imposes a condition for the x and y projection of the $\mathbf{k}$-vectors of the diffracted waves (more in particular, the Bragg condition implies that only a discrete set of k_x and k_y values are allowed for a certain angle of incidence). The z-component of the waves in the homogenous media is found by applying

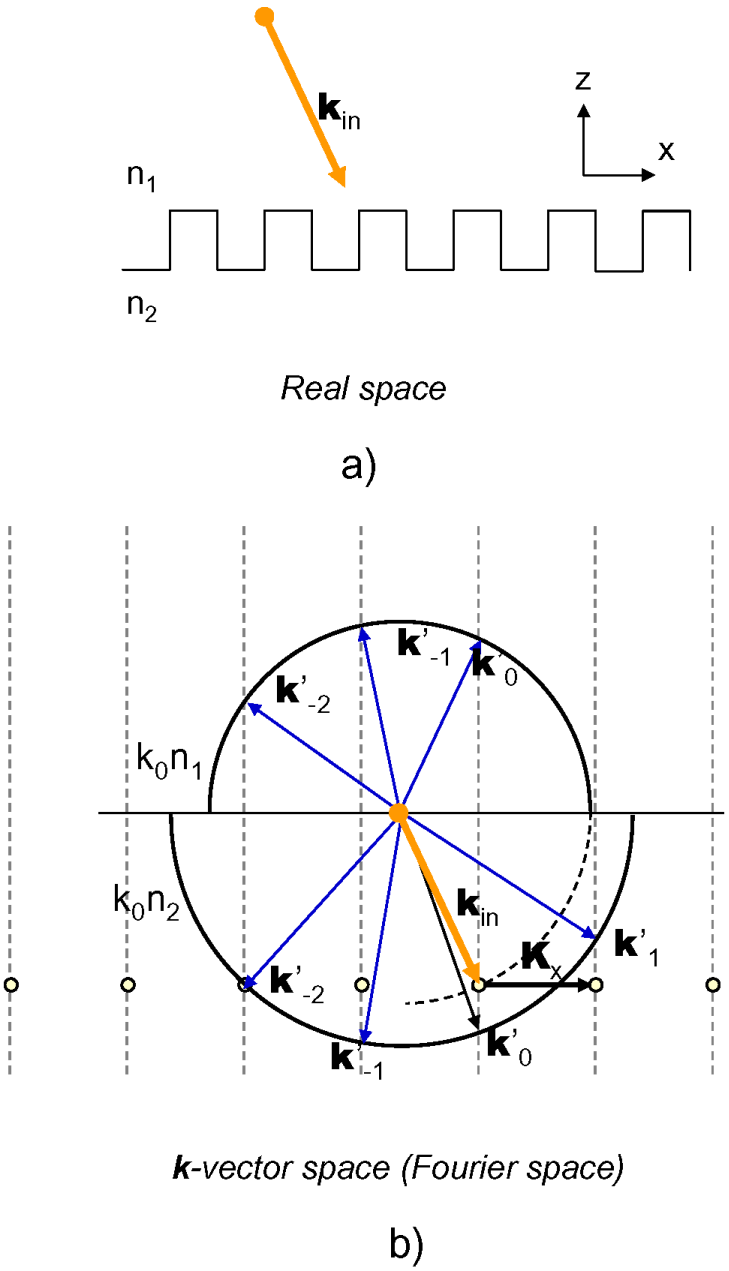


Figure 6.14: k-vector diagram for a 1D periodic medium. The lattice (periodic in the x-direction) is at the interface of medium 1 and medium 2

the dispersion relation. The k_z-components of the incident and diffracted waves do not have to be the same. This is graphically depicted in the **k**-vector diagram of figure 6.14b for the situation of figure 6.14a.

The two half circles in figure 6.14b represent the homogenous media above and below the grating. All plane waves traveling in the upper medium (both the incident wave and the upward reflected diffraction orders) are represented by **k**-vectors with their end point on a circle with radius $k_0 n_1$. All plane waves traveling in the lower medium (the downward transmitted diffraction orders) are represented by **k**-vectors with their end point on a circle with radius $k_0 n_2$.
All diffracted k-vectors can easily be found by first adding an integer multiple of $\mathbf{K_x}$ to $\mathbf{K_{in}}$ resulting in the dots and the dashed lines in figure 6.14b. These dashed lines show the regions in k-space where the Bragg condition is fulfilled, because there aren't any restrictions for the z direction. Subsequently the intersection of the dashed lines and the half circles are the points where formula 7.60 is met, constituting the end points of the **k**-vectors of the different diffraction orders.

It is easy to relate the Bragg-condition and k-vector diagram to the grating equation (6.8) derived earlier.

For the reflected diffraction orders the Bragg condition can be written as:

$$k_{in} \sin(\theta_{in}) + mK = k \sin(\theta_{m,refl}) \tag{6.50}$$

from which one finds:

$$n_1 \sin(\theta_{m,refl}) - n_1 \sin(\theta_{in}) = m\frac{\lambda_0}{\Lambda} \tag{6.51}$$

For the transmitted diffraction orders, one finds:

$$n_2 \sin(\theta_{m,trans}) - n_1 \sin(\theta_{in}) = m\frac{\lambda_0}{\Lambda} \tag{6.52}$$

These equations are the grating equation for reflection and transmission respectively. For $m = 0$ it brings us back to Snell's law.

Surface gratings and waveguides

We consider the situation where a grating is in close proximity of a waveguide as shown in figure 6.15.

As long as the tail of the optical modes is not yet zero at the grating, the modes are able to couple to each other or to plane waves through diffraction. Again the projected Bragg condition in the x-y plane needs to be satisfied.

Determine which period the grating needs to have in the following two situations:

- A guided mode couples to the same but in the opposite direction propagating guided mode
- A guided mode couples to a plane wave which propagates upwards and downwards.

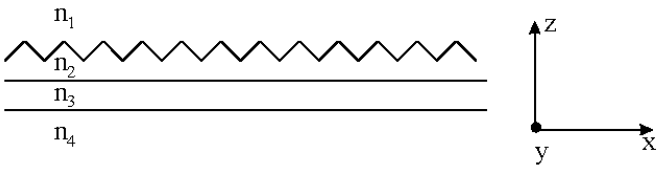


Figure 6.15: Waveguide grating

6.4 Floquet-Bloch theorem and Photonic bandgap

6.4.1 Floquet-Bloch theorem

To investigate the propagation behavior in periodic structures, in this section we will consider the modes of these structures. Therefore, we will determine the solutions of Maxwells equations in this periodic medium (no source terms). A general solution to this problem is given by the Floquet-Bloch theorem and the modes of the periodic medium are called Floquet-Bloch modes.

For the sake of simplicity, we will first consider a one dimensional problem in which the refractive index $n(x)$ is periodic in the x-direction with period Λ. $\psi(x)$ is the complex representation of the field $\psi(x,t)$ (meaning $\psi(x,t) = \mathrm{Re}\left[\psi(x)\, e^{j\omega t}\right]$) and satisfies following wave equation

$$\frac{d^2\psi}{dx^2} + k_0^2 n^2(x)\,\psi(x) = 0 \quad \text{with} \quad k_0 = \frac{\omega}{c} = \frac{2\pi}{\lambda_0} \tag{6.53}$$

in which k_0 and λ_0 respectively are the free space wave vector and free space wavelength.

This wave equation can be rephrased as an eigenfunction problem. So, (6.53) can be rewritten as:

$$-\frac{1}{n^2(x)}\frac{d^2\psi}{dx^2} = k_0^2\psi \tag{6.54}$$

or

$$A\psi = k_0^2\psi \tag{6.55}$$

with A a linear operator. The wave function $\psi(x)$ then is the eigenfunction of this operator and k_0^2 its eigenvalue.

For a constant refractive index $n(x) = n$, the solutions to these wave equations are plane waves propagating in the +x or -x direction

$$\psi(x) = e^{-jkx} \quad \text{with} \quad k = \pm n k_0 = \pm n\frac{\omega}{c} \tag{6.56}$$

When the refractive index is periodically modulated, the Floquet-Bloch theorem states that a general solution of Maxwell's equation is

$$\psi(x) = e^{-jkx} u_k(x) \tag{6.57}$$

k-value	ω**-value**	$\mathbf{u_k}$-function
k	ω	$u_k(x)$
-k	ω	$u_k^*(x)$
k+mK (m integer and K=$\frac{2\pi}{\Lambda}$)	ω	$e^{-jmKx}u_k(x)$
-k+mK	ω	$e^{-jmKx}u_k^*(x)$

Table 6.1: Relation between k-value and eigenvalue/function

in which k is no longer given by $k = nk_0$ and in which $u_k(x)$ is a periodic function with the same period as $n(x)$.

The mathematical proof of this theorem is beyond the scope of this course.

The function $\psi(x)$ has interesting symmetry and periodicity properties as a function of k, as indicated in the table.

These properties have important consequences.They show that the values k,$-k$, $k+mK$ and $-k+mK$ all have the same eigenvalues and that their eigenfunctions are related to eachother. This means we only have to look for solutions in the k-interval between 0 and $\frac{K}{2}$. This is called the first Brillouin zone. One can also say that the solutions for k,$-k$, $k+mK$ and $-k+mK$ are in a way coupled and actually respresent 1 solution. This is no surprise as the periodic function $u_k(x)$ can be written as a Fourier series:

$$u_k(x) = \sum_{l=-\infty}^{\infty} c_l e^{-jlKx} \tag{6.58}$$

This means that

$$\begin{aligned} \psi(x) &= e^{-jkx}\left(\sum_{l=-\infty}^{\infty} c_l e^{-jlKx}\right) \\ &= \sum_{l=-\infty}^{\infty} c_l e^{-j(k+lK)x} \end{aligned} \tag{6.59}$$

This expression clearly shows that the solution for a certain k-value also contains components for $k+mK$ through the periodicity of $u_k(x)$. The relation between k and $-k$ can be understood by saying that at a given frequency there is always a forward propagating and backward propagating solution with the same $|k|$-value. The terms forward and backward do not have the same meaning as in the case of a uniform medium: the coupling between k and $k+mK$ in the periodic medium actually means that a forward propagating wave always is coupled with a number of backward propagating waves (m such that $k+mK<0$) and vice versa.

In figure 6.16 the typical relation between k and ω is shown for a periodic medium. In the vicinity of k=0, we find a solution ω which only slightly differs from that of a homogeneous medium. Close to $k = \frac{K}{2}$ larger deviations occur. At $k = K/2$ a distinct behavior occurs. A forbidden zone occurs in which over a finite ω interval the eigenfunction has a k-value with a constant real part and an imaginary part different from zero. This means that at these frequencies there is no propagating eigenmode, but that the eigenmode is evanescent. This evanescent nature is due to the strong coupling at this k-value between the forward and backward wave (this will become

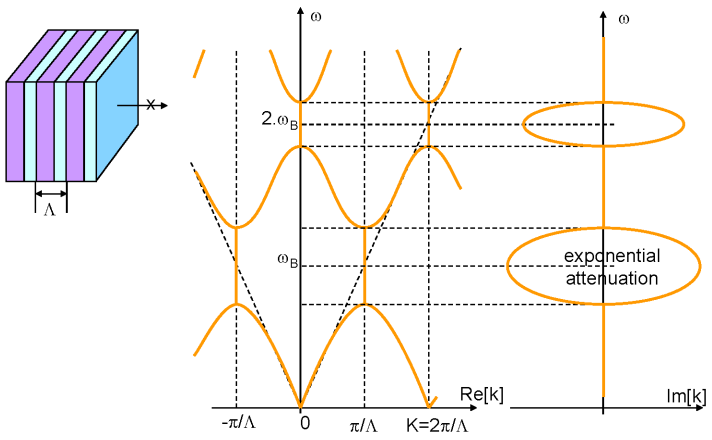


Figure 6.16: Relation between k and frequency for a periodic lossless medium. The left part shows the real part of k, the right part shows the imaginary part of k

more clear later on in this chapter). Note that the sign of the imaginary part of k is always such that the wave is decaying in the direction of its propagation.

Due to the symmetry behavior, we can see from figure 6.16, that for a given k-value, different ω are found. This structure is called a band structure with multiple bands. The forbidden zone is called the bandgap. This method of folding the dispersion relation curve around the edges of the Brillouin zone is called band folding and originates from the periodic nature of the structure.

The described behavior is mathematically very closely related to solving the Schrodinger equation in a periodic potential, which leads to the band structure of crystalline semiconductors. Due to the analogy these periodic structures are also called "photonic crystals".

Finally, we will look at what happens when the periodic medium contains a source term with a given frequency (for example a point source). This source will excite the eigenmodes belonging to this frequency. Although the total field that is generated is the sum of the propagating and evanescent waves only power is transferred to propagating modes. As there are only evanescent waves inside the forbidden zone, a source with a frequency inside the forbidden zone will not emit electromagnetic power!

This analysis can be extended to the vectorial and three dimensional case. The theorem becomes:

$$\psi_{\mathbf{k}} = e^{-j\mathbf{k}.\mathbf{r}} u_{\mathbf{k}}(\mathbf{r}) \tag{6.60}$$

in which $u_{\mathbf{k}}(\mathbf{r})$ shows the same periodicity as $n(\mathbf{r})$.

6.4.2 Photonic Bandgap

Photonic bands

As discussed in the previous section, a 1 dimensional periodically layered structure always shows a range of forbidden frequencies for which no propagation is allowed in the medium. This became

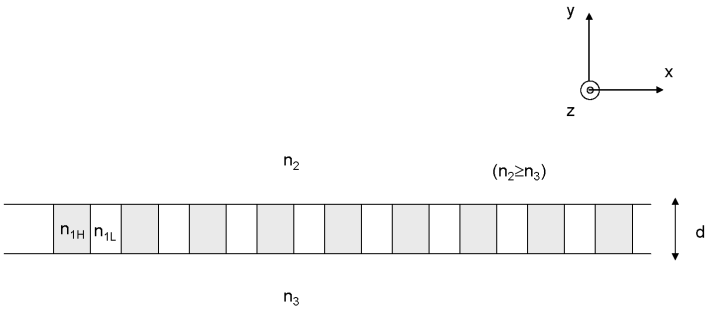


Figure 6.17: Vertical confinement of the Floquet-Bloch modes by total internal reflection

clear in the dispersion relation of the Floquet-Bloch modes, where over a range of frequencies the k-value of the mode became complex. Such a gap in the dispersion relation is called a photonic bandgap. The width of the photonic bandgap depends strongly on the refractive index contrast in the periodic medium (the larger the index contrast, the larger the photonic bandgap).

Besides the occurrence of a photonic bandgap, another interesting feature is the particular shape of the bands. As can be seen in figure 6.16, the dispersion relation is nearly linear for small k. In this range of frequencies the Floquet-Bloch mode practically obeys the dispersion relation of a homogeneous medium. When we approach the band edges, the bands no longer show a linear behavior. At the Brillouin zone edge the first derivative of the dispersion relation is zero, in other words: the group velocity

$$v_g = \frac{d\omega}{dk} \tag{6.61}$$

goes to zero, while the phase velocity $\frac{\omega}{k}$ is non-zero. This means that light slows down when approaching the band edge and finally is no longer able to transport energy at the band edge. By adequate engineering of the periodic structure, light can be slowed down to a significant fraction of the speed of light in vacuum (and hence is called slow light).

The sign of v_g also indicates the sense of propagation (in energy terms) of a Floquet-Bloch mode. If the sign (and hence the slope of $\omega(k)$ is positive, the mode has energy flow in the positive x-direction, and vice-versa. Again, one can see that sense of propagation does not depend on sign of k, but on sign of $d\omega/dk$.

Let us now consider what happens when we change the one dimensional periodically layered structure of figure 6.16 to the periodic structure shown in figure 6.17 in which one lateral dimension of the periodic structure is restricted (while still infinitely extending in the z-direction).

Restricting the dimensions of the layered medium in one transverse dimension and thereby making a waveguide structure in the y-direction imposes that the Floquet-Bloch modes of this structure will contain a k_y component, just as in the case ofa simple dielectric waveguide.

Just as in the case of a simple slab waveguide, the imposed boundary conditions at the interface between core and cladding layer, makes that only a discrete number of k_y (and therefor k_y-values) are allowed, which define discrete bands of the Floquet-Bloch modes. This is shown in figure 6.18.

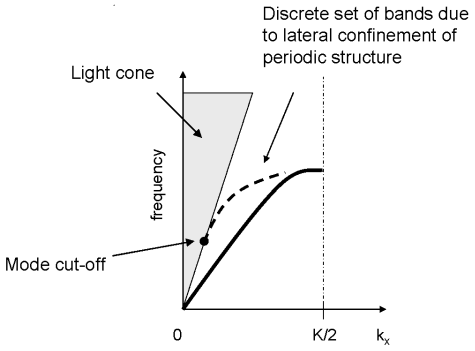


Figure 6.18: Band diagram with light cone of a laterally restricted periodic structure: illustration of the occurrence of (vertically) higher order Floquet-Bloch modes and the light cone, indicating the (ω, k)-region where only non-guided modes can occur. Photonics bands can cross light cone, which means that they change from radiative to guided and vice versa.

When the k_y component of the **k**-vector in the cladding materials is imaginary, the Floquet-Bloch mode is guided, otherwise the mode can radiate into upper or lower cladding layer (or both). This implies that we can define an area in the dispersion diagram for which modes will be radiating, this will be the case when

$$k_x < \frac{\omega}{c} n_2 \tag{6.62}$$

due to the dispersion relation of the medium n_2

$$k_x^2 + k_y^2 = (\frac{\omega}{c} n_2)^2 \tag{6.63}$$

This area is called the light cone of the structure, meaning that Floquet-Bloch modes which lie in this light cone are not guided. The light cone is indicated in figure 6.18 by the shaded area. Bands can intersect the boundary of the light cone, meaning that they change from radiative to guided and vice versa.

Two dimensional photonic crystals

Besides the discussed structure in which the medium was periodic in 1 dimension, one can also imagine structures which are periodic in two dimensions. For example we will look at the case of a square lattice of dielectric columns with radius r and dielectric constant ϵ as shown in figure 6.19.

As is discussed in the appendix, the reciprocal lattice of a square real lattice is also a square lattice with period $\frac{2\pi}{\Lambda}$. This reciprocal lattice and its unit cell (the first Brillouin zone) is shown in figure 6.20. Due to the symmetry of the structure, it is sufficient to look for solutions to Maxwells equations in the triangle formed by the Γ, M and X-point, as all other points inside the Brillouin zone can be related to a point inside this triangle. This area is called the irreducible Brillouin zone.

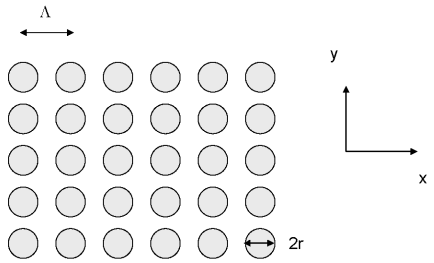


Figure 6.19: Square lattice of dielectric columns

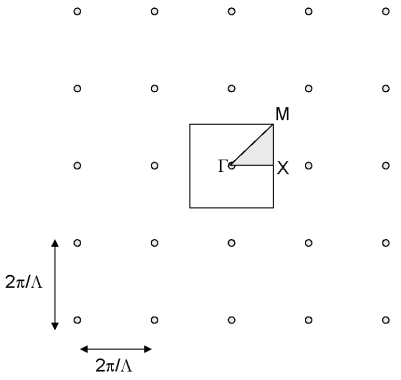


Figure 6.20: Reciprocal lattice of the square real lattice and its first Brillouin zone (together with an indication of the irreducible Brillouin zone

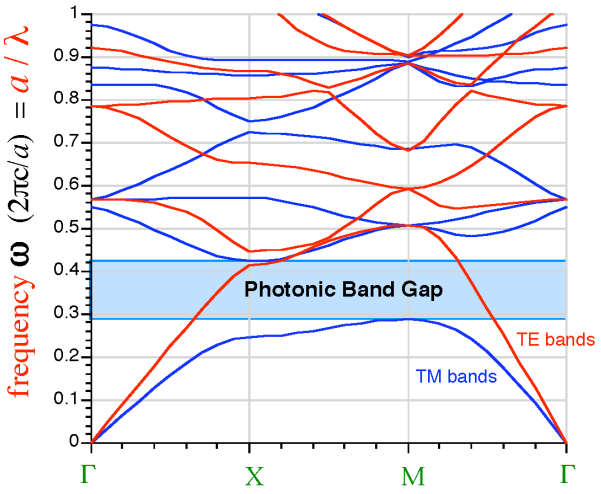


Figure 6.21: Calculated bands for TE and TM polarization for the case of a square lattice of dielectric columns in air

To fully characterize the propagation of light through this periodic medium one has to look for frequencies ω which are a solution to Maxwells equations for every point inside the irreducible Brillouin zone. One can proof however that this (ω,k) surface will always have its extreme values on the edge of the Brillouin zone (due to the symmetry of the structure). Therefore, when looking for photonic bandgaps, it is sufficient to plot the dispersion relation on the edges of the irreducible Brillouin zone.

As was discussed in the chapter on dielectric waveguides, for two dimensional problems Maxwells equations can be decomposed in two sets of equations: one for transverse electric polarization and one for transverse magnetic polarization. This is still the case for these two-dimensional photonic crystal. In this case the transverse electric polarization has its electrical field vector in the (x,y)-plane and lies the magnetical field component along the axis of the columns. In the case of the transverse magnetic polarization, the magnetic field lies in the (x,y)-plane and the electrical field lies along the axis of the columns. As there are two polarization states (opposed to the one dimensional case in which both polarization states are degenerate) two sets of bands need to be calculated. This means that a photonic bandgap may occur for one polarization state and not for the other or that both polarization states show a bandgap, which only partially overlap (or do not overlap).

Calculating the bands of these structures require advanced numerical techniques, which lie outside the scope of this course. The result of these types of calculations are shown in figure 6.21 for the square lattice of dielectric columns ($\epsilon = 12$ and $r = 0.2\Lambda$).

From this figure it is clear that a photonic bandgap exists for TM-polarization while there is no bandgap for TE polarized light.

Although very difficult to fabricate, also three dimensional photonic crystals exist. The band structure and its calculation becomes more complex and we can no longer distinguish between TE and TM polarization.

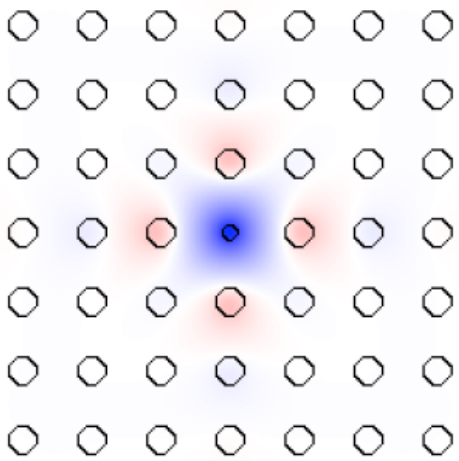


Figure 6.22: Point defect mode in a square lattice

Light confinement by defect engineering

Previously we found that two dimensional photonic crystals can show bandgaps for in-plane propagation. No modes are allowed for frequencies inside the gap. The density of states, being the number of possible modes per unit of frequency increment, is zero within the photonic bandgap. By perturbing a single lattice site, we can permit a single localized mode or a set of closely spaced modes that have frequencies inside the gap. For example, we could remove a single column from the square dielectric lattice from the previous example or we could change its dielectric constant or radius.

This defect introduces a peak in the crystals density of states. When this peak happens to be located in the photonic bandgap of the bulk crystal, then the defect induced state must be evanescent in the crystal. The light is therefore localized around the defect.

An example of such a localized mode is shown in figure 6.22 where we reduced the size of one dielectric column (TM polarization). We can clearly see that the mode is localized around the point defect

This can be applied to fabricate a threshold-less laser. In a conventional laser, spontaneous emission couples to radiation modes, even if there is only one cavity mode. Therefore, nearly all spontaneous emission leaves the cavity in these lasers. This means that the photon density inside these cavity mainly consists of photons created by stimulated emission. That is why the optical power versus current curve shows a threshold. When one succeeds in keeping all spontaneous emission inside the cavity, which is possible through these photonic bandgap materials and light localization through defects, the photon density can also be significantly influenced by spontaneous emission. In this situation population inversion is achieved much faster which can lead (in theory) to a threshold-less laser.

This effect can also be used to change the spontaneous emission characteristics of a material. When we surround an area in which spontaneous emission occurs by a photonic bandgap material (in which the bandgap is aligned with the emission peak of the material), this radiation is sent back

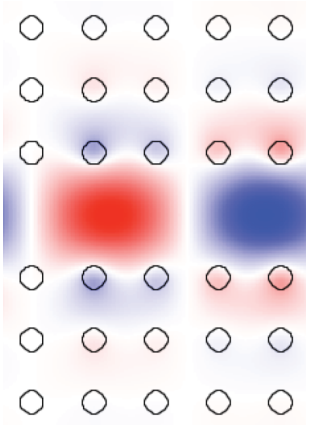


Figure 6.23: Light propagating along a line defect in a square lattice of dielectric columns

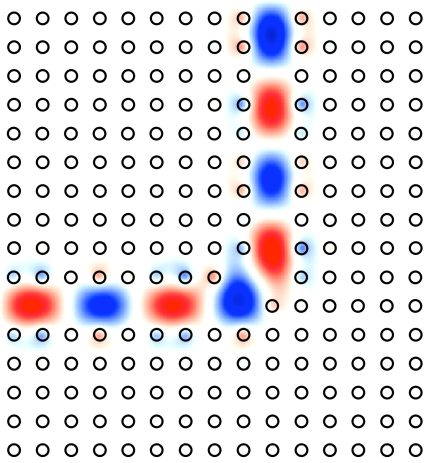


Figure 6.24: 90 degree bend in a square dielectric lattice

to the emitted area. A three dimensional cavity is formed which (if the cavity is sufficiently small, i.e. smaller than the coherence length of the spontaneous emission) shows a discrete number of cavity modes. By allowing some directions of propagation we can make sure that all spontaneous emission is coupled to these allowed energy/direction combinations. The micro-cavity LED is based on this effect.

Photonic crystals can also be used to guide light. When we replace the point defect by a linear defect, light can propagate along this linear defect. An example is shown in figure 6.23, in which one row of dielectric columns was removed. Note that in this case the light is guided in air, instead of in a high refractive index material (which is the case for waveguides based on total internal reflection).

These type of waveguides can be used for different applications. First, they show a particular dispersive behavior (larger than in the case of a waveguide based on total internal reflection) which can be used in wavelength selective devices. Secondly, very sharp bends can be fabricated in these photonic crystals as radiation of light in the bulk crystal is prohibited. A simulation example of a 90 degree bend is shown in figure 6.24.

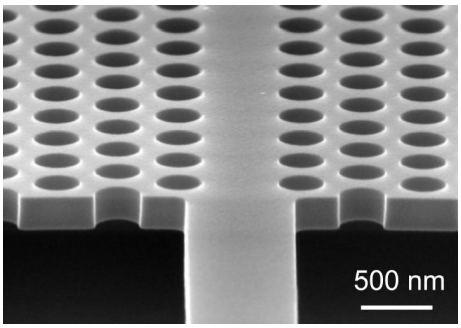


Figure 6.25: Fabricated line defect in a photonic crystal

Although we discussed two dimensional photonic crystals, real structures can't be of infinite height and light has to be confined in the vertical direction by total internal reflection. An example of a fabricated two dimensional photonic crystal is shown in figure 6.25. Here the concept of the light cone can be reintroduced.

Although we only discussed a square lattice, also more complex lattice shapes can be used like a triangular lattice, graphite lattices etc.

6.5 Periodically layered media

In the chapter on Thin Films, the layered medium was treated using the transfer matrix method. This method starts from the consideration that when a_i and b_i are the amplitude of the incident and returning plane wave in layer i, the following relation can be written based on the boundary conditions and Maxwells equations

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} a_{i+1} \\ b_{i+1} \end{pmatrix} \tag{6.64}$$

The description of a layered medium is done by simply multiplying the matrices related to what happens at the interfaces and the matrices which describe the propagation inside the layers.

$$\begin{pmatrix} a_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \cdots \begin{pmatrix} A_{N-1} & B_{N-1} \\ C_{N-1} & D_{N-1} \end{pmatrix} \begin{pmatrix} a_N \\ b_N \end{pmatrix} \tag{6.65}$$

When we state that $b_N = 0$ (because in the transmission medium no wave propagates in the $-z$-direction), we can write b_1 and a_N as a function of a_1, which results in the the reflection and transmission coefficients of the layer structure. Numerically speaking the transfer matrix method is very efficient. However, little physical insight is obtained. Therefore, we will use two alternative methods to study the periodic layered structure: the coupled wave theory and the Floquet-Bloch theory.

6.5.1 Coupled wave theory

Although the coupled wave theory can be applied to media which show a transversal structure, we shall limit ourself to media with a one dimensional periodicity that are invariant in both transversal directions. We will start from the one dimensional wave equation for the scalar field ψ

$$\frac{d^2\psi}{dz^2} + k_0^2 n^2(z)\,\psi = 0 \tag{6.66}$$

in which $n^2(z)$ is periodic in the propagation direction (z-direction) and independent of x and y. When the refractive index is constant, the total field consists of a forward propagating and backward propagating plane wave. In the coupled wave theory we assume that this field can still be written as the sum of a forward and backward propagating wave in the case of a periodic refractive index modulation, but for which the amplitudes and phases of these waves are modulated by the functions $A(z)$ and $B(z)$. These functions only contain slow z variations of the scalar field, while the fast variations are in the plane wave parts. Therefore we propose a solution of the form

$$\psi(z) = A(z)\,e^{-j\beta z} + B(z)\,e^{j\beta z} \tag{6.67}$$

in which the plan waves $e^{\pm j\beta z}$ satisfy the wave equation in the homogenous material where $n^2(z) = n_0^2$ and $\beta = k_0 n_0$.

Substituting this expression in the wave equation 6.66, we find

$$\begin{array}{l}\left\{-\beta^2 A(z) + \frac{d^2A(z)}{dz^2} - 2j\beta\frac{dA(z)}{dz}\right\}e^{-j\beta z} + \left\{-\beta^2 B(z) + \frac{d^2B(z)}{dz^2} + 2j\beta\frac{dB(z)}{dz}\right\}e^{+j\beta z} \\ +k_0^2 n^2(z)\left\{A(z)\,e^{-j\beta z} + B(z)\,e^{j\beta z}\right\} = 0\end{array} \tag{6.68}$$

When neglecting the second order derivatives of $A(z)$ and $B(z)$, this equation becomes

$$\begin{array}{l}k_0^2\left(n^2(z) - n_0^2\right)e^{-j\beta z}A(z) + k_0^2\left(n^2(z) - n_0^2\right)e^{+j\beta z}B(z) \\ -2j\beta e^{-j\beta z}\frac{dA(z)}{dz} + 2j\beta e^{+j\beta z}\frac{dB(z)}{dz} = 0\end{array} \tag{6.69}$$

For a periodic medium we can write

$$n^2(z) - n_0^2 = \sum_{m\neq 0} a_m e^{jmKz} \text{ with } K = 2\pi/\Lambda \tag{6.70}$$

When we substitute this in the previous equation, we get, after multiplying with $e^{-j\beta z}$:

$$\begin{array}{l}k_0^2 A(z)\sum\limits_{m\neq 0} a_m e^{-j(2\beta - mK)z} + k_0^2 B(z)\sum\limits_{m\neq 0} a_m e^{jmKz} \\ -2j\beta e^{-2j\beta z}\frac{dA}{dz} + 2j\beta\frac{dB}{dz} = 0\end{array} \tag{6.71}$$

To satisfy this equation, all components with the same phase velocity have to add up to 0. When for a certain m-value $2\beta \approx mK$ then the condition for the slowly varying term becomes

$$k_0^2 a_m e^{-j(2\beta - mK)z} A(z) + 2j\beta \frac{dB(z)}{dz} = 0 \tag{6.72}$$

By multiplying with $e^{+j\beta z}$ and only retaining the slowly varying terms we obtain

$$k_0^2 a_{-m} e^{j(2\beta - mK)z} B(z) - 2j\beta \frac{dA(z)}{dz} = 0 \tag{6.73}$$

Note that $2\beta \approx mK$ is nothing else but the Bragg condition for the m-th order diffraction.

We introduce

$$\Delta\beta = \frac{2\beta - mK}{2} \tag{6.74}$$

which is called the Bragg deviation. As β is a measure for the frequency ($\beta = \frac{n_0 \omega}{c}$), $\Delta\beta$ is a measure for a frequency band around β. We can rewrite equation 6.72 as

$$\frac{dB}{dz} = j\kappa_{ba} e^{-j2\Delta\beta z} A(z) \text{ with } \kappa_{ba} = \frac{k_0^2 a_m}{2\beta} \tag{6.75}$$

and equation 6.73 as

$$\frac{dA}{dz} = -j\kappa_{ab} e^{j2\Delta\beta z} B(z) \text{ with } \kappa_{ab} = \frac{k_0^2 a_{-m}}{2\beta} \tag{6.76}$$

We define κ as $\kappa = \sqrt{\kappa_{ab}\kappa_{ba}}$. The first equation can be rewritten as

$$e^{j2\Delta\beta z} \frac{dB}{dz} = j\kappa_{ba} A(z) \tag{6.77}$$

Deriving this equation to z and substituting this expression in equation 6.76, we get

$$\frac{d^2 B}{dz^2} + j2\Delta\beta \frac{dB}{dz} - \kappa^2 B(z) = 0 \tag{6.78}$$

When we suppose that $B(z) = Ce^{j\gamma z}$ (with C an arbitrary constant), we find:

$$\gamma = -\Delta\beta \pm \sqrt{\Delta\beta^2 - \kappa^2} \stackrel{\Delta}{=} -\Delta\beta \pm \delta \tag{6.79}$$

This leads to

$$B(z) = P_1 e^{-j\Delta\beta z} e^{-j\delta z} + Q_1 e^{-j\Delta\beta z} e^{j\delta z} \tag{6.80}$$

and

$$A(z) = -\frac{2\beta}{k_0^2 a_m}\left((\Delta\beta+\delta)P_1 e^{j(\Delta\beta-\delta)z} + (\Delta\beta-\delta)Q_1 e^{j(\Delta\beta+\delta)z}\right) \tag{6.81}$$

The two unknown P_1 and Q_1 then can be derived from the boundary conditions at $z = 0$ and $z = L$ (the boundaries of the periodic layered medium).

When we take $A(0) = 1$ and $B(L) = 0$, we describe an incident wave which is reflected and transmitted through the layered structure. The field reflection and transmission coefficient then becomes

$$r = \frac{B(0)}{A(0)} \tag{6.82}$$

and

$$t = \frac{\psi(L)}{A(0)} = \frac{A(L)\,e^{-j\beta L}}{A(0)} \tag{6.83}$$

After some calculations we find for P_1 and Q_1:

$$\begin{aligned} Q_1 &= \tfrac{k_0 a_m}{4n_0}\tfrac{e^{-j\delta L}}{j\Delta\beta\sin(\delta L)+\delta\cos(\delta L)} \\ P_1 &= \tfrac{-k_0 a_m}{4n_0}\tfrac{e^{j\delta L}}{j\Delta\beta\sin(\delta L)+\delta\cos(\delta L)} \end{aligned} \tag{6.84}$$

and for r and t:

$$\begin{aligned} r &= -\kappa_{ab}\tfrac{j\sin(\delta L)}{j\Delta\beta\sin(\delta L)+\delta\cos(\delta L)} \\ t &= e^{-j\beta L}e^{j\Delta\beta L}\tfrac{\delta}{j\Delta\beta\sin(\delta L)+\delta\cos(\delta L)} \end{aligned} \tag{6.85}$$

One easily verifies using equation 6.79 that $|r|^2 + |t|^2 = 1$, such that power is conserved.

To κ we can associate the meaning of a coupling per length unit, which in this case comes down to a reflection per unit length. For a DBR mirror, made out of a stack of $\frac{\lambda}{4}$ layers of a material with a high (n_H) and a low (n_L) refractive index, we can find the following relation between κ and Δn:

$$\kappa = \frac{k_0^2}{2\beta}\frac{n_H^2 - n_L^2}{\pi} \approx \frac{2\Delta n}{\lambda} \tag{6.86}$$

in which we used $n_H^2 - n_L^2 \approx 2n_0\Delta n$ and $\beta = 2\pi n_0/\lambda$.

We also find the peak reflectivity from equation 6.85. For $\Delta\beta = 0$ (maximal interaction with the lattice), $\delta = j\kappa$ and one gets

$$R_{\max} = |r|^2_{\max} = \tanh^2\kappa L \tag{6.87}$$

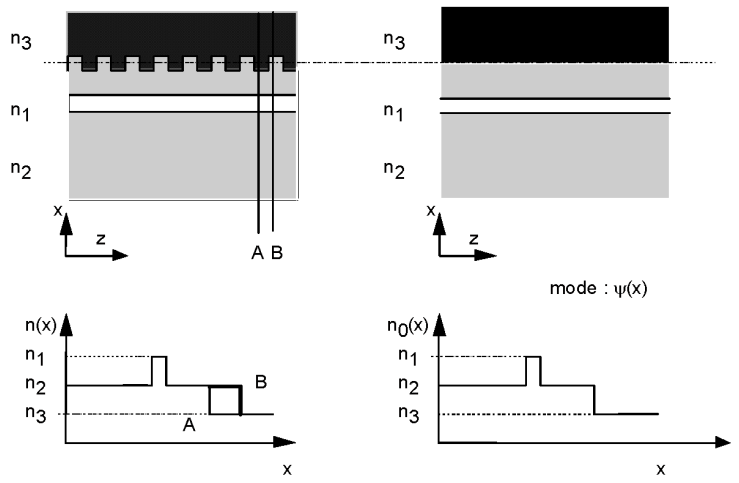


Figure 6.26: Slab waveguide in the presence of a grating

On the other side we can easily derive the wavelength span over which the DBR structure strongly reflects the incident light (the so called stop band). It it sufficient to note that as soon as δ in equation 6.79 becomes imaginary, $A(z)$ represent an exponentially damped wave, such that all propagating power is reflected. This is the case when $|\Delta\beta| < \kappa$. Therefore, we can write

$$2\frac{\Delta\beta}{\beta} = \frac{\Delta\lambda}{\lambda} = \frac{\kappa\lambda}{\pi n_0} = \frac{2}{\pi}\frac{\Delta n}{n_0} \tag{6.88}$$

in which $\Delta\lambda$ is the full wavelength span over which the DBR will reflect strongly.

This analysis is only valid for sufficiently low refractive index contrasts, because otherwise the assumption that $A(z)$ and $B(z)$ vary sufficiently slow would no longer be valid.

Furthermore, this analysis is not limited to periodically layered media, but it is also valid for periodic structures of which the layers are also transversally non homogeneous. As an example, we will discuss here the case of a slab waveguide with the presence of a grating.

The structure as shown in figure 6.26a can be seen as a perturbation of the pure waveguide structure in figure 6.26b. The refractive index profiles of the perturbed and unperturbed problem are also shown. The pure waveguide structure has a guided mode with a mode profile $\psi(x)$ and a propagation constant β. In the Fourier decomposition of the refractive index we will have to take into account that all refractive indices will be weighted with the intensity of the mode profile. The Fourier decomposition can then be written as

$$\int_{-\infty}^{+\infty} \left(n^2(x,z) - n_0^2(x)\right)\psi^2(x)\,dx = \sum_{m\neq 0} a_m e^{+jmKz} \text{ with } K = 2\pi/\Lambda \tag{6.89}$$

Again, one finds the coupled wave equations with identical expressions for κ, but in which a_m is calculated from 6.89. One finds expressions for κ_{ab} and κ_{ba} which strongly resemble the expressions for κ_{12} and κ_{21} for the case of a codirectional coupler (see the Chapter on dielectric waveguides).

All previous conclusions and expressions (peak reflection, stopband etc.) for the case of a transversally invariant structure remain valid, provided a correct interpretation of κ.

6.5.2 Floquet-Bloch theory

In this section we will deal with the problem of the previous section using the Floquet-Bloch theorem. Again, the scalar field needs to satisfy the one dimensional wave equation

$$\frac{d^2\psi}{dz^2} + n^2(z)\, k_0^2 \psi = 0 \qquad (k_0 = \tfrac{\omega}{c}) \tag{6.90}$$

in which $n^2(z)$ is periodic and has the following form

$$\begin{array}{l} n^2(z) = \left\{ \begin{array}{ll} n_1^2 & 0 < z \leq l_1 \\ n_2^2 & l_1 < z \leq \Lambda \end{array} \right. \\ n^2(z + m\Lambda) = n^2(z) \text{ with m integer} \end{array} \tag{6.91}$$

with the additional boundary condition that ψ and $\frac{d\psi}{dx}$ are continuous at each interface. The Floquet-Bloch theorem poses the following representation of ψ:

$$\psi(z) = e^{-jkz} u_k(z) \tag{6.92}$$

in which $u_k(z)$ is a periodic function in z (period Λ). If we substitute this in the wave equation we find

$$\frac{d^2 u_k}{dz^2} - 2jk\frac{du_k}{dz} + \left(n^2(z)\, k_0^2 - k^2\right) u_k(z) = 0 \tag{6.93}$$

with following boundary conditions

$$\left\{ \begin{array}{l} u_k(z) \text{ en } \frac{du_k}{dz} \text{ are continuous at } x{=}l_1 \\ u_k(0) = u_k(\Lambda) \text{ en } \left.\frac{du_k}{dz}\right|_{z=0} = \left.\frac{du_k}{dz}\right|_{z=\Lambda} \end{array} \right. \tag{6.94}$$

due to the periodicity of $u_k(z)$. Alternatively we can solve the equation in $\psi(z)$ directly. Although this equation is simpler, the boundary conditions are more complex.

The general solution for $u_k(z)$ can therefore be written as

$$u_k(z) = \left\{ \begin{array}{ll} Ae^{j(n_1 k_0 + k)z} + Be^{-j(n_1 k_0 - k)z} & 0 \leq z \leq l_1 \\ Ce^{j(n_2 k_0 + k)z} + De^{-j(n_2 k_0 - k)z} & l_1 < z \leq \Lambda = l_1 + l_2 \end{array} \right. \tag{6.95}$$

Note that for the scalar field $\psi(z)$ we find forward and backward propagating plane waves with the length of the k-vector proportional to the local refractive index. Four unknown coefficients remain (i.e. A,B,C and D) and we have four boundary conditions to apply. This results in a homogenous set of equations for these four unknowns. To get a non trivial solution, the determinant

of this set of equations has to equal 0. After some mathematics we find the following dispersion relation in n_{eff}:

$$\cos\left(k_0 n_{eff}\left(l_1+l_2\right)\right)=\cos\left(n_1 l_1 k_0\right)\cos\left(n_2 l_2 k_0\right)-\frac{n_1^2+n_2^2}{2n_1 n_2}\sin\left(n_1 l_1 k_0\right)\sin\left(n_2 l_2 k_0\right) \tag{6.96}$$

with $k = k_0 n_{eff}$.

The dispersion relation determines for given λ, n_1, n_2, l_1 and l_2 the allowed values for n_{eff}. When one doesn't find a value for n_{eff} for a certain combination of parameters, this means that that particular wavelength can't propagate in the perfectly periodic medium (infinitely extending). This can be the case. For example if

$$\begin{cases} n_1 l_1 k_0 = \pi/2 \\ n_2 l_2 k_0 = \pi/2 \\ n_1 \neq n_2 \end{cases} \tag{6.97}$$

This is the case for a stacking of layers of $\frac{\lambda}{4}$ thickness.

The dispersion relation becomes

$$\cos\left(\frac{2\pi}{\lambda}n_{eff}\left(l_1+l_2\right)\right)=-\frac{n_1^2+n_2^2}{2n_1 n_2}=\rho \tag{6.98}$$

There is no real solution for n_{eff} as

$$\left(n_1-n_2\right)^2=n_1^2+n_2^2-2n_1 n_2>0 \tag{6.99}$$

or

$$|\rho|=\frac{n_1^2+n_2^2}{2n_1 n_2}>1 \tag{6.100}$$

We can find a complex n_{eff}. This means that the wave is exponentially damped as stated before. Note that the considered configuration is an infinitely thick Bragg reflector. A Bragg reflector is in reality always of finite thickness , meaning that in a practical case the wave is not completely forbidden at the Bragg wavelength, but it will be strongly reflected. The thicker we make the mirror, the better the structure will resemble that of an infinitely extending mirror and the larger the reflection coefficient will be. It is clear that when ρ becomes larger, the window of forbidden frequencies (and high reflectivity) will increase.

Let's now study the behavior of the dispersion relation around the Bragg wavelength, which means that we have to use the general dispersion relation, assuming a complex n_{eff}. We represent the right hand part of the equation by $\rho(n_1, l_1, n_2, l_2, k_0)$. So that

$$\cos\left(k_0\left(n_{eff}^r+jn_{eff}^i\right)\left(l_1+l_2\right)\right)=\rho\left(n_1,l_1,n_2,l_2,k_0\right) \tag{6.101}$$

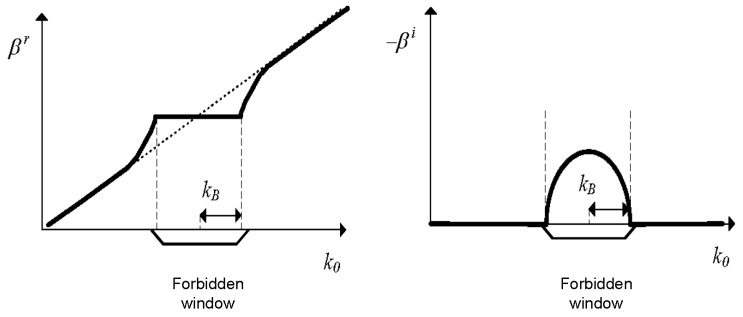


Figure 6.27: Graphical representation of the dispersion relation in a DBR structure

such that

$$\begin{cases} \cosh\left(\frac{2\pi}{\lambda}\left(l_1+l_2\right)n_{eff}^i\right)\cos\left(\frac{2\pi}{\lambda}\left(l_1+l_2\right)n_{eff}^r\right)=\rho\left(n_1,l_1,n_2,l_2,k_0\right) \\ \sinh\left(\frac{2\pi}{\lambda}\left(l_1+l_2\right)n_{eff}^i\right)\sin\left(\frac{2\pi}{\lambda}\left(l_1+l_2\right)n_{eff}^r\right)=0 \end{cases} \tag{6.102}$$

As we proposed that

$$n_{eff}^i \neq 0 \tag{6.103}$$

we get

$$\begin{array}{l} \beta^r \stackrel{\Delta}{=} \frac{2\pi}{\lambda} n_{eff}^r = m\frac{\pi}{\Lambda}, \quad \Lambda = l_1+l_2 \text{ and m integer} \\ \beta^i \stackrel{\Delta}{=} \frac{2\pi}{\lambda} n_{eff}^i = \frac{1}{\Lambda}\,\mathrm{a\cosh}\,\left(\pm\rho\left(n_1,l_1,n_2,l_2,\lambda\right)\right) \end{array} \tag{6.104}$$

The argument of the $acosh$-function is positive if

$$\begin{array}{l} \rho < 0 \text{ and m odd} \\ \rho > 0 \text{ and m even} \end{array} \tag{6.105}$$

Note that β_r does not depend on the wavelength in the forbidden frequency window. As waves can not increase exponentially we demand $\beta_i < 0$ (for waves propagating in the positive z-direction). Figure 6.27 qualitatively represents β as a function of the k-vector of the exciting wave.

The half width of the forbidden frequency window (where complex β occur) is given by

$$k_B = \frac{\pi}{4n_1l_1} - \frac{\pi}{4n_2l_2} \tag{6.106}$$

6.5.3 Example

We now are able to model the stacking of homogenous layers in three ways:

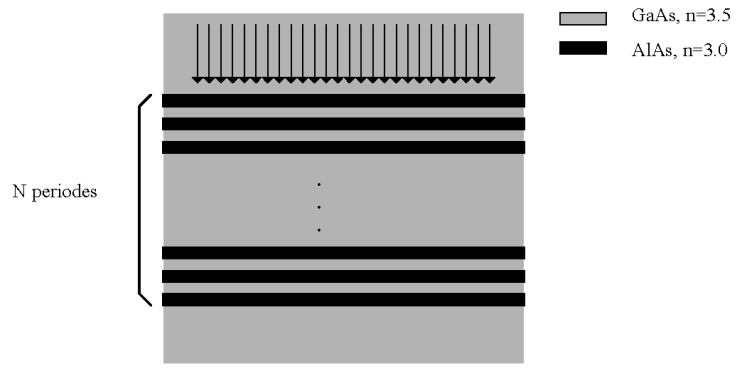


Figure 6.28: Reflection at a periodically layered medium

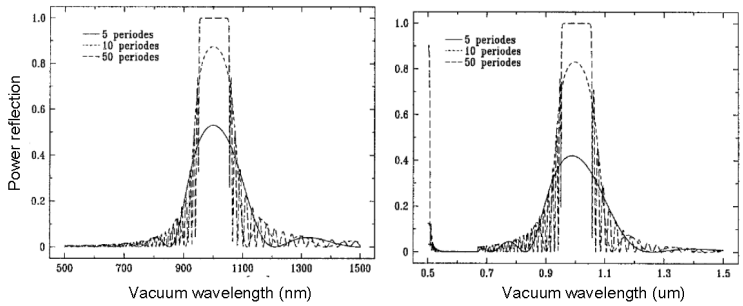


Figure 6.29: Calculation example: power reflection as a function of wavelength calculated through the transfer matrix formalism and the coupled wave theory

- through the transfer matrix formalism: this formalism allows to calculate the reflection and transmission in an exact way of an arbitrary stacking of homogenous layers (so in particular a periodic stacking). The number of periods is arbitrary.
- through coupled wave theory: this is an approximate analysis, in which the number of periods is arbitrary.
- through Floquet-Bloch: this is an exact solution. There are however an infinite number of periods.

The example that we will elaborate on here is the periodic stacking of GaAs/AlAs layers, where the layers have a thickness of $\frac{\lambda}{4}$ (wavelength in the material!). The thicknesses were chosen in such a way that the DBR mirror works for a central wavelength of 1 μm. This configuration is depicted in figure 6.28.

The modeled structure indeed shows strong reflections around a wavelength of 1 μm (for a large number of periods this reflection becomes 1). We note the good correspondence between the results from the transfer matrix formalism and the coupled wave theory (figure 6.29a and figure 6.29b). The relation with the Floquet-Bloch theorem is also clear (figure 6.30a and 6.30b): around $1\mu m$ there is a forbidden zone of the periodic medium (which results in a constant real part of n_{eff} and an imaginary part different from zero).

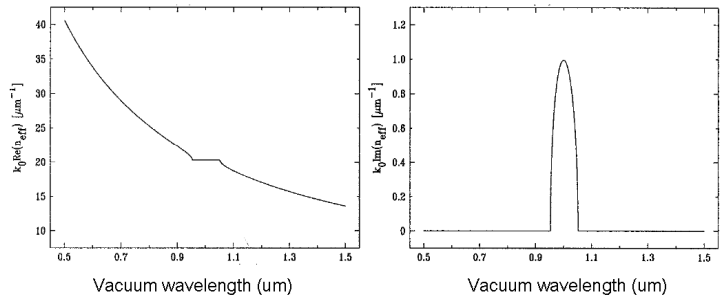


Figure 6.30: Calculation example: real and imaginary part of the effective refractive index as a function of wavelength calculated through Floquet-Bloch theory

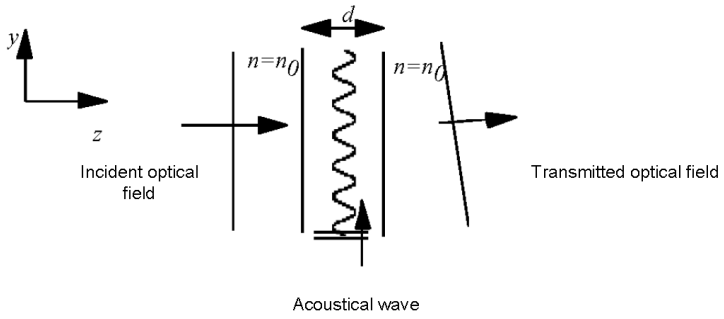


Figure 6.31: Acousto-optical diffraction

6.6 Acousto-optical diffraction

6.6.1 Theory

When one launches an acoustical wave through a solid, one induces compression of the matter which induces a refractive index variation. When there is an optical beam present, this optical beam will feel the acoustical wave by its effect on the refractive index of the material. This will change the wavefront of the optical wave. This is called acousto-optical interaction. When the result is diffraction of the optical wave, this is called acousto-optical diffraction. The grating is induced only by the presence of the acoustical wave.

We assume that the amplitude of the acoustical wave is sufficiently small such that everything can be described in a linear way (so we assume that no non-linear effects occur in the refractive index modulation). Furthermore, we assume the acoustical wave to be monochromatic with frequency Ω and propagating at a speed v. The refractive index modulation traveling in the y-direction then is:

$$\Delta n\left(y\right) = \Delta n \sin\left(\Omega\left(t - \frac{y}{\mathrm{v}}\right)\right) \tag{6.107}$$

When the thickness d of the material is sufficiently small, we can apply the transmission theory and say that

$$\psi\left(y,d\right)=t\left(y\right).\psi\left(y,0\right) \tag{6.108}$$

with

$$t\left(y\right)=e^{-jk_{o}n_{o}d}\;e^{-jk_{o}\Delta n\sin\left(\Omega\left(t-\frac{y}{\mathrm{v}}\right)\right)d} \tag{6.109}$$

in which n_0 is the refractive index of the material without acoustical wave present. When the above assumptions apply, this is called the Raman-Nath regime.

Say

$$\psi\left(y,0,t\right)=\left.Ae^{-jk_{o}zn_{o}}\;e^{j\omega t}\right]_{z=0}=Ae^{j\omega t} \tag{6.110}$$

And therefore

$$\psi\left(y,d,t\right)=Ae^{-jk_{o}n_{o}d}\;e^{-jk_{o}\Delta n\sin\left(\Omega\left(t-\frac{y}{\mathrm{v}}\right)\right)d}\;e^{j\omega t} \tag{6.111}$$

or

$$\psi\left(y,z,t\right)\;=A\;e^{-jk_{0}n_{0}z}\;e^{-jk_{o}\Delta n\sin\left(\Omega\left(t-\frac{y}{\mathrm{v}}\right)\right)d}\;e^{j\omega t}=A\;e^{-jk_{o}n_{o}z}\;e^{-jk_{o}d\Delta n\sin\Omega\tau}\;e^{j\omega t} \tag{6.112}$$

with

$$\tau=t-y/\mathrm{v} \tag{6.113}$$

Now, it is an identity that

$$e^{j\alpha\sin\theta}=\sum_{n=-\infty}^{+\infty}e^{jn\theta}J_{n}\left(\alpha\right) \tag{6.114}$$

and

$$J_{-n}\left(\alpha\right)=\left(-1\right)^{n}J_{n}\left(\alpha\right) \tag{6.115}$$

with J_n a Bessel function of the first type.

This leads to

$$\psi(y,z,t) = A\, e^{-jk_o n_o z} \sum_{N=-\infty}^{+\infty} e^{-jN\Omega\tau}\, e^{j\omega t}\, J_N(k_0 d\Delta n)$$
$$= A\, e^{-jk_o n_o z}\, e^{j\omega t} J_o(k_o d\Delta n) + A\, e^{-jk_o n_o z} \sum_{N=1}^{\infty} \left\{ e^{j(\omega t - N\Omega\tau)} J_N(k_o d\Delta n) + (-1)^N e^{j(\omega t + N\Omega\tau)} J_N(k_o d\Delta n) \right\} \tag{6.116}$$

Note that

- The zeroth order diffraction is the incident wave with a small loss of amplitude.

$$J_o(k_o \Delta n d) < 1 \quad if \quad \Delta n d \neq 0 \tag{6.117}$$

- Higher order diffraction orders have a slightly different frequency, given by $\omega \pm N\Omega$ and are diffracted under an angle that is related to the ratio of the acoustical and optical wavelength, as

$$\sin|\theta| = \frac{|k_y|}{k_z} = \frac{N\Omega/v}{k_o n_o} = N\frac{\lambda_{opt}}{\lambda_{acoust}} \tag{6.118}$$

  with λ_{opt} the wavelength in the material with refractive index n_0. This is the same condition as the projected Bragg condition.

This analysis is valid in the Raman-Nath regime, when d is sufficiently small. When d becomes large, the diffraction inside the material itself can no longer be neglected. In this case we have to look at the material as was it composed of layers with thickness Δz such that in these layers the Raman-Nath regime applies. For these layers we can redo the analysis with an incident field for layer i being the transmitted field of layer $(i-1)$.

For these thick materials, we are in the Bragg regime. We can find an upper boundary for which the Raman-Nath regime still applies, by saying that the two waves that diffract at the two boundaries can't have a large phase difference. This is no longer the case for thick layers and is related to the validity of the transmission theory.

The phase difference between ρ_1 and ρ_2 is given by

$$\Delta\phi = \phi_{\rho_2} - \phi_{\rho_1} = \frac{2\pi}{\lambda} n_o d\,(1 - \cos\theta) \tag{6.119}$$

For the N-th order diffraction

$$tg\theta \cong \sin\theta \cong \theta = N\frac{\lambda_{opt}}{\lambda_{acoust}} \tag{6.120}$$

and

$$\phi_{\rho_1} - \phi_{\rho_2} = \frac{2\pi}{\lambda} n_o d \frac{\theta^2}{2} \tag{6.121}$$

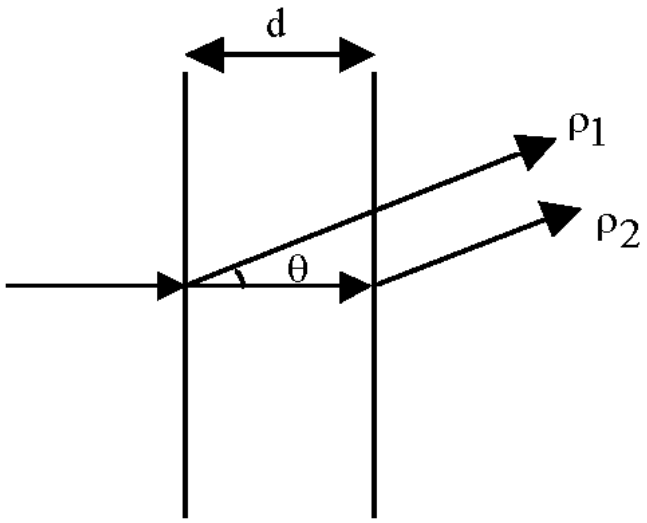


Figure 6.32: Acousto-optical diffraction: Bragg regime and Raman-Nath regime

or

$$\phi_{\rho_1} - \phi_{\rho_2} = \pi n_o d \frac{N^2 \lambda_{opt}}{\lambda^2_{acoust}} \tag{6.122}$$

When the phase difference between ρ_1 and ρ_2 equals π, these two waves can no longer interfere constructively. Therefore, the thickness corresponding to this phase shift is the critical thickness. When the acousto-optical element is thicker than this critical thickness, the Raman-Nath regime is no longer valid, or

$$\pi n_o d_{crit} \frac{N^2 \lambda_{opt}}{\lambda^2_{acoust}} = \pi \text{ and } d_{crit} = \frac{\lambda^2_{acoust}}{n_o N^2 \lambda_{opt}} \tag{6.123}$$

with

$$\begin{array}{l} d >> d_{crit} \text{ : Bragg regime} \\ d << d_{crit} \text{ : Raman - Nath regime} \end{array} \tag{6.124}$$

When working in the Bragg regime, a more rigorous modeling is necessary. The direction of the first diffraction order is given by

$$\frac{2\pi n_o}{\lambda} \sin\theta_i = \frac{2\pi n_o}{\lambda} \sin\theta_d - m\frac{2\pi}{\Lambda} \tag{6.125}$$

This is the Bragg condition along the surface. The direction θ_d is determined by the projected Bragg condition in combination with the dispersion relation. Again the optimal interaction between the incident and transmitted field will occur when the vectorial Bragg condition applies, when $\theta_i = \theta_d$.

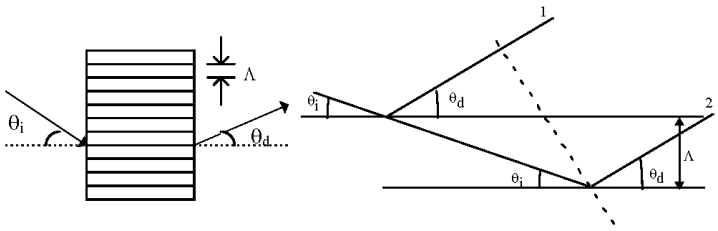


Figure 6.33: Acousto-optical diffraction: Bragg regime

This Bragg condition shows that waves 1 and 2 interfere constructively, meaning that they have a $2m\pi$ phase shift at discrete reflections at interfaces spaced by the period Λ. This image is more suited to understand the diffraction of X-rays at crystals, which have a discrete structure.

In the Bragg regime, the grating is that thick, that only one diffraction order remains. To understand why the thickness of the grating leads to the fact that only one order remains, we go back to the section in which the Bragg condition was deduced using a three dimensional Fourier transform of the refractive index profile of a one or two dimensional periodic medium. As a function of the thickness of the grating, we obtained a sinc-profile for the amplitude of the waves that satisfy the Bragg condition. For an acousto-optical element that works in the Bragg regime this sinc-function becomes very small.

6.6.2 Applications

In applications for acousto-optical diffraction, most often one works in the Bragg regime with piezo-electrical materials such that the acoustical wave can be induced by a high frequency sinusoidal signal across electrodes attached to the piezo- electrical material.

Modulation

By switching the high frequency electrical signal on and off, the optical beam is or is not diffracted and is therefore blocked by or transmitted through the aperture in the screen (figure 6.34).

Beam bending

Changing the frequency of the electrical signal leads to a change of the acoustical frequency and therefore also to a change in the angle of the beam. Often it is the case that the electrical signal contains multiple frequency components such that there are multiple diffracted beams (figure 6.35).

Frequency change

Acousto-optical diffraction slightly changes the optical frequency of the higher order diffraction. That way, the frequency of the transmitted light beam can be slightly changed compared to that of the incident light beam (figure 6.36).

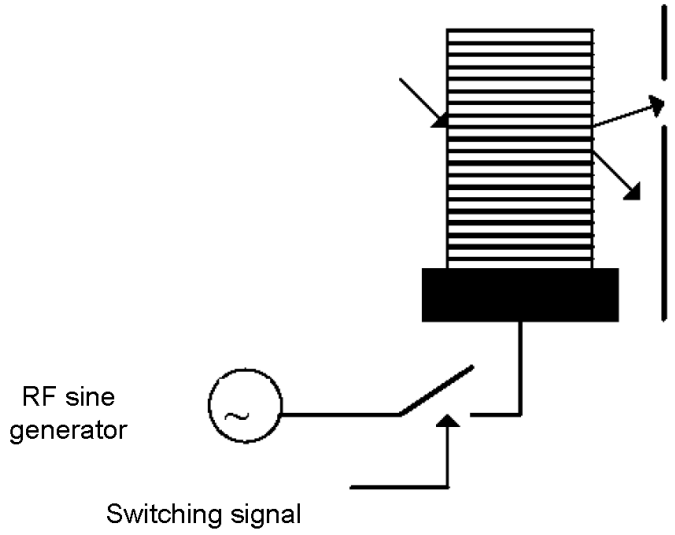


Figure 6.34: High frequency modulation of an optical beam

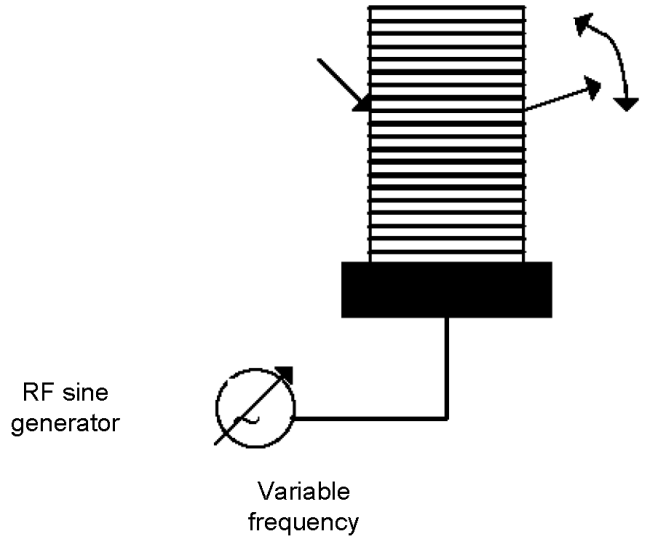


Figure 6.35: Bending of an optical beam

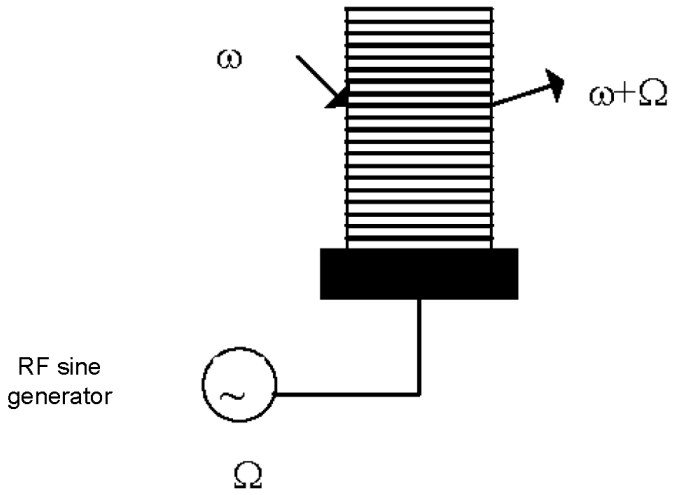


Figure 6.36: Frequency change of an optical beam

Spectrum analyzer

From the previous application it became clear that the diffraction angle is determined both by the optical and acoustical frequency of the waves. In this case we use a monochromatic wave (ω constant), meaning that the diffraction angle only becomes a function of the frequency of the acoustical wave Ω. This acoustical frequency translates in a certain diffraction angle of the optical beam (figure 6.37).

When the Bragg-cell is modulated by an electrical signal, all frequency components of the electrical signal are converted to acoustical frequencies and translated in diffraction angles. The intensity of a beam with a certain angle represents a spectral component of the electrical signal. Therefore, the spectral content of the electrical signal is therefore translated to the angle content of the optical beam.

When we place a lens behind the Bragg cell, we can find in the focal plane the intensities which correspond to the various angles. As these correspond with the spectral content of the electrical signal, we find a spatial representation of the frequency content of the signal in the focal plane. Detecting this signal using a detector array in the focal plane allows a quasi immediate spectral decomposition of the electrical signal.

The function of the RF carrier coming from the sine generator, with which the signal to be analyzed is mixed, can be understood as following. The global system can analyze a signal in its spectral components, with a relative bandwidth which is limited, meaning, when the signal frequency content is centered around f_0 and a bandwidth of Δf, the ratio of $\frac{\Delta f}{f_0}$ is limited. To still be able to analyze signals with a large bandwidth, f_0 is increased such that Δf rises proportionally.

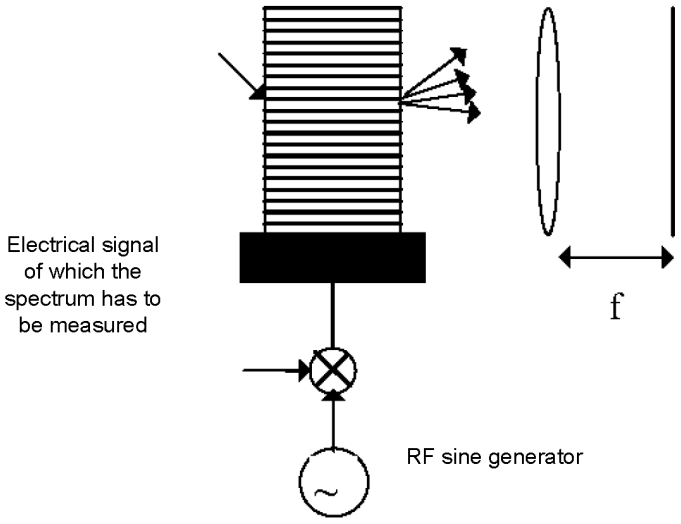


Figure 6.37: Spectrum analyzer

6.7 Holography

6.7.1 Introduction and history

In conventional image recording techniques like photography, a three-dimensional scene is imaged onto a light sensitive surface by means of a lens system, in such a way that one plane of the scene is imaged sharply, while the other planes aren't. The photograph consists of a pattern corresponding to the spatial intensity distribution in this imaging plane. The phase information of the light waves coming from the object are thereby lost. Holography or wavefront reconstruction has the characteristic property that both amplitude and phase information of the light waves scattered by an object are registered. Because most light-sensitive materials are only sensitive for the intensity of the light that impinges on the material, the phase information will have to be transformed into an intensity variation. This is accomplished by illuminating the object coherently and by letting the scattered waves interfere with a reference wave at a photo-sensitive plate. The phase and amplitude information of the scattered light waves are therefore encoded (by means of the interference pattern) into the photosensitive material. The original waves can be regenerated in a second step: the reconstruction, in which the hologram is again illuminated by the reference wave. This reference wave is diffracted by the hologram in such a way that for an observer, the reconstructed waves cannot be distinguished from the original waves scattered from the object. He can see a three-dimensional image which shows effects like perspective and depth of focus. When the observer moves around, he will see different scenes generated by the hologram illuminated by the reference beam, just like when the observer would move around the original object.

The invention of the principle of wavefront reconstruction is attributed to Dennis Gabor (1948). In the early days, the applications of optical holography were limited due to the lack of sufficiently powerful and coherent light sources. A major break through occurred in the sixties with the invention of the laser and the off-axis setup of Leith and Upatnieks. In the same period, Denisyuk made another important contribution: the volume reflection hologram that allows reconstruction

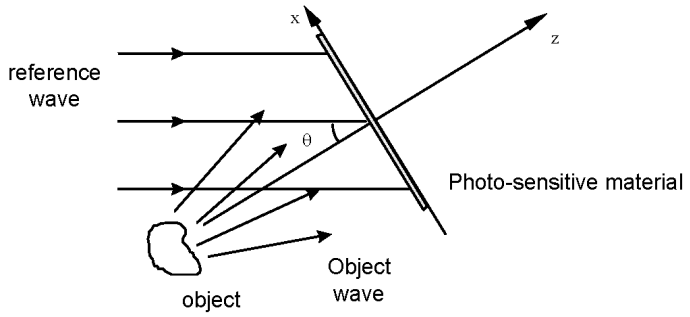


Figure 6.38: Recording of a transmission hologram

with a white light source. In 1969, Benton developed the rainbow hologram, which also allowed to use white light for the reconstruction and showed a higher luminance. Meanwhile, holography has found different applications, like in the field of non-destructive displacement analysis, using holographic interferometry. Holographic optical elements and holographic gratings have gained a lot of importance in optical information storage, image reconstruction, optical processing and telecommunication.

6.7.2 Theoretical base of the wavefront reconstruction process

In order to describe the holographic process we will restrict ourself to an amplitude transmission hologram. In these holograms the fringes originating from the interference of the object wave and the reference wave are recorded by a locally increased absorption of a thin layer of photo-sensitive material. We assume that the reference wave is a plane wave with a uniform intensity distribution, of which the direction perpendicular to the wavefronts forms an angle θ with the normal of the photo-sensitive material (see figure 6.38). We also assume that the object and reference wave are at the same wavelength and have a fixed phase relation. We will show that the absorption pattern in the photo-sensitive material allows to reconstruct the waves originating from the object.

The complex amplitudes of the object and reference wave at the photo-sensitive material can be written as

$$\begin{aligned} o(x,y) &= |o(x,y)|\, e^{-j\varphi(x,y)} \\ r(x,y) &= re^{jk\sin(\theta)x} \end{aligned} \tag{6.126}$$

with $k = \frac{2\pi}{\lambda}$.

The total intensity on the photosensitive material is therefore (disregarding some constant terms)

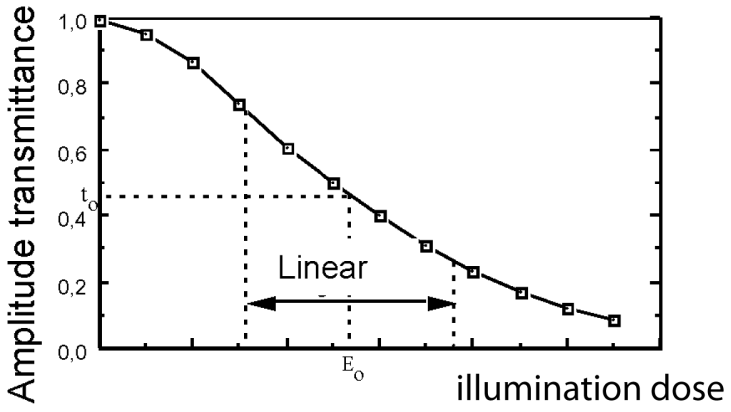


Figure 6.39: Typical t-E curve

$$\begin{aligned} I(x,y) &= [o(x,y)+r(x,y)][o(x,y)+r(x,y)]^* \\ &= o(x,y)o(x,y)^* + r(x,y)r(x,y)^* + o(x,y)r(x,y)^* + o(x,y)^*r(x,y) \\ &= |o(x,y)|^2 + r^2 + r\,|o(x,y)|\,e^{-j\phi(x,y)}e^{-jk\sin(\theta)x} + r\,|o(x,y)|\,e^{+j\phi(x,y)}e^{+jk\sin(\theta)x} \\ &= |o(x,y)|^2 + r^2 + 2r\,|o(x,y)|\cos(\phi(x,y)+k\sin(\theta)x) \\ &= I_0 + 2r\,|o(x,y)|\cos(k\sin(\theta)x+\phi(x,y)) + |o(x,y)|^2 \end{aligned} \tag{6.127}$$

with I_0 the intensity of the reference wave. This intensity is in most cases much stronger than that of the object wave.

From this equation we see that the amplitude and phase information of the object wave is coded under the form of the respective amplitude and phase modulation of a periodic carrier with spatial frequency $\frac{sin(\theta)}{\lambda}$. The illumination therefore contains a quasi-periodic fringe pattern, originating from the interference of the reference wave and the object wave, of which phase and amplitude are modulated by the amplitude and phase of the object wave.

The illumination dose E of the photo-sensitive plate is defined as the product of the illumination time T and the illumination intensity I. Assume that the amplitude transmissivity $t(x,y)$ of the photo-sensitive material after illumination and development as a function of the illumination is given by the graph in figure 6.39 and that we have chosen the illumination in such a way that we are using the linear area of the curve, then we can write

$$t(x,y) = t_0 + \beta(E - E_0) \tag{6.128}$$

typically with $\beta < 0$ or

$$t(x,y) = t_0 + \beta T[|o(x,y)|^2 + r\,|o(x,y)|\,e^{-j\phi(x,y)}e^{-jk\sin(\theta)x} + r\,|o(x,y)|\,e^{+j\phi(x,y)}e^{+jk\sin(\theta)x}] \tag{6.129}$$

In this equation t_0 represents a constant background transmissivity which depends on the photo-sensitive material and on the intensity of the reference wave. One typically chooses this intensity

in such a way that one can operate in the linear regime of the t-E curve, such that equation 6.128 is valid. The last two terms (which are a consequence of the a quasi-periodic fringe pattern) imply that the hologram is a kind of diffraction grating.

As the hologram has the character of a quasi-periodic grating, diffraction effects will occur in transmission. For reconstructing the object wave, the developed hologram is illuminated with the reference wave again, but now the object is no longer there. The complex amplitude $u(x, y)$ of the transmitted wave can be written as the sum of four terms

$$\begin{aligned} u(x,y) &= r(x,y)t(x,y) \\ &= u_1(x,y) + u_2(x,y) + u_3(x,y) + u_4(x,y) \end{aligned} \tag{6.130}$$

with

$$\begin{aligned} u_1(x,y) &= t_0 r e^{jk\sin(\theta)x} \\ u_2(x,y) &= \beta T r \left|o(x,y)\right|^2 e^{jk\sin(\theta)x} \\ u_3(x,y) &= \beta T r^2 \left|o(x,y)\right| e^{-j\phi(x,y)} \\ &= \beta T r^2 o(x,y) \\ u_4(x,y) &= \beta T r^2 o^*(x,y) e^{j2k\sin(\theta)x} \end{aligned} \tag{6.131}$$

The first term represents a plane wave that is transmitted through the hologram, which is an attenuated replica of the reference wave. This wave is surrounded by a halo due to the second term, which leads to a wave with the same propagation direction as the reference wave. The term $u_2(x, y)$ varies in space, but is mostly negligible with respect to the first term because typically one will use r much larger than $|o(x, y)|$, because of the linearity of the of the photo-sensitive medium. The third term and actually the most important term is, disregarding a constant term, the complex amplitude of the object wave and therefore forms a virtual image of the object on its original place behind the hologram. In other words, when looking into the hologram from the right, one will "see" the object in the same way as if the object was really there. One can look at it from various positions and one will "see" the object from different points of view. This demonstrates that the hologram stores 3D information indeed. The fourth term gives rise to the conjugated object wave. One can prove that this corresponds to a real image in front of the hologram (see figure 6.40).

The real image being formed has a peculiar property however: it is a so-called "pseudoscopic" image. Take for example the case where the object is a face mask (as used during carnival). The nose of the mask is pointing towards the hologram plate. In the real pseudoscopic image the nose will also point towards the hologram plate (which is the opposite from the more common case of imaging by means of a lens, which produces a normal "orthoscopic" image). In other words the pseudoscopic image features depth reversal.

Another way of interpreting this, is that the terms u_1 and u_2 are a zeroth order diffraction of the reference beam through the hologram (with a quasi-periodic transmission function). The terms u_3 and u_4 are the first and minus first diffraction order. The wave that is most interesting is therefore the first order diffraction.

One can also consider the angular spectrum of the four terms

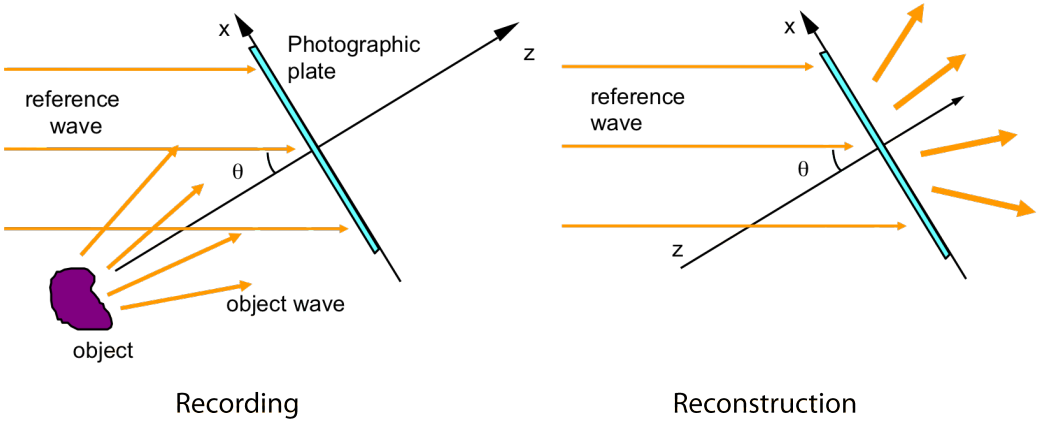


Figure 6.40: Recording and reconstruction of a transmission hologram

$$\begin{array}{rcl} U_1(f_x,f_y) & = & F(t_0 r e^{j2\pi f_x^r x}) = t_0 r \delta(f + f_x^r, f_y) \\ U_2(f_x,f_y) & = & F(\beta Tr\,|o(x,y)|^2\, e^{j2\pi f_x^r x}) = \beta Tr R_o(f_x,f_y) * \delta(f + f_x^r, f_y) \\ U_3(f_x,f_y) & = & F(\beta Tr^2 o(x,y)) = \beta Tr^2 O(f_x,f_y) \\ U_4(f_x,f_y) & = & F(\beta Tr^2 o^*(x,y) e^{j4\pi f_x^r x}) = \beta Tr^2 O^*(-f_x,-f_y) * \delta(f + 2f_x^r, f_y) \end{array} \tag{6.132}$$

In these equations $O(f_x, f_y)$ and $R_o(f_x, f_y)$ represent the Fourier transform of the object wave $o(x, y)$ and the autocorrelation of $O(f_x, f_y)$ respectively. The term $U_3(f_x, f_y)$ represents, disregarding a constant factor, the Fourier transform of the object wave and is located in the center of the spatial frequency plane. The term $U_1(f_x, f_y)$ corresponds to the spatial frequency of the carrier and therefore results in a δ-function in $(-f_x^r, 0)$. The second term $U_2(f_x, f_y)$ is centered around U_1 and contains the autocorrelation of $O(f_x, f_y)$ as a factor, which has double the spectral width of $o(x, y)$. Analogously, U_4 is centered around $(-2f_x^r, 0)$. From this one can derive that when one wants to avoid that the diffracted waves overlap, the spatial frequency of the reference wave has to be sufficiently large:

$$f_x^r \geq 3 f_{x,\max} \tag{6.133}$$

in which $f_{x,\max}$ is the highest spatial frequency occurring in the object wave. From this we can conclude that the angle θ between the reference beam and the surface normal of the photosensitive plate has to be made sufficiently large to avoid any overlap (see figure 6.41).

Questions: consider a small part of the hologram: which information is contained in this small part? Half of the hologram is covered. What information is lost during the reconstruction?

6.8 Appendix - reciprocal lattice as a Fourier transform

6.8.1 Real lattice

In a material with a periodic refractive index $n(\mathbf{r})$, $n(\mathbf{r})$ can be written as

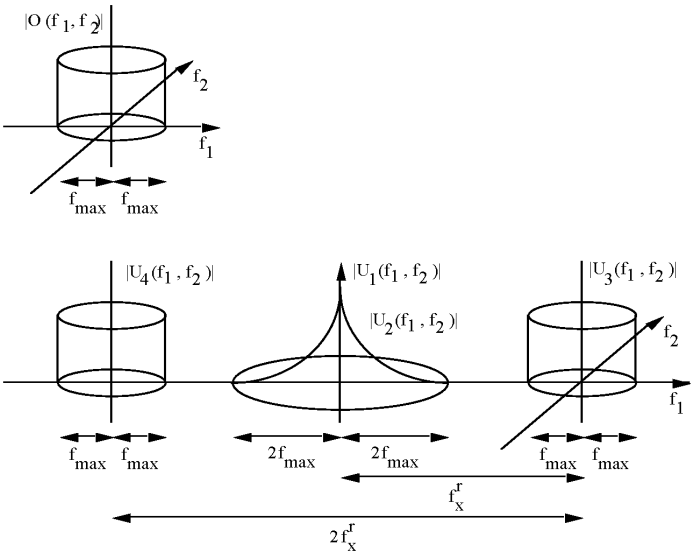


Figure 6.41: Representation in the spatial frequency domain

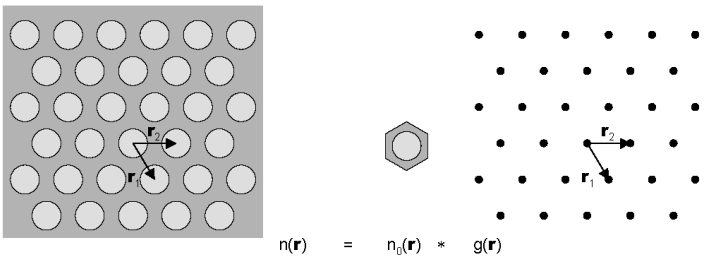


Figure 6.42: The periodic refractive index can be written as the convolution of the unit cell function and the lattice function

$$n\left(\vec{r}\right) = n_0\left(\vec{r}\right) * g\left(\vec{r}\right) = \int n_0\left(\vec{r}' - \vec{r}\right) g\left(\vec{r}\right) d\vec{r}' \tag{6.134}$$

or a convolution of the refractive index $n_0(\mathbf{r})$ of the unit cell and the lattice function $g(\mathbf{r})$, consisting of a number of delta-functions. For a two dimensional lattice, the lattice function becomes

$$g\left(\vec{r}\right) = \sum_{s,t} \delta\left(\vec{r} - s\vec{r}_1 - t\vec{r}_2\right) \tag{6.135}$$

in which $\vec{r}_1$ and $\vec{r}_2$ are the translation invariance vectors.

6.8.2 Reciprocal lattice

The reciprocal lattice can be considered as a spatial Fourier transform of the real lattice

$$N\left(\vec{k}\right) = \mathbf{F}\left(n\left(\vec{r}\right)\right) = \mathbf{F}\left(n_0\left(\vec{r}\right)\right).\mathbf{F}\left(g\left(\vec{r}\right)\right) \tag{6.136}$$

or as a product of the Fourier transform of the unit cell and the Fourier transform of the lattice function (convolution property of the Fourier transform). By applying the translation property we find

$$\begin{aligned} G\left(\vec{k}\right) &= \mathbf{F}\left(g\left(\vec{r}\right)\right) \\ &= \sum_{s,t} \mathbf{F}\left(\delta\left(\vec{r} - s.\vec{r}_1 - t.\vec{r}_2\right)\right) \\ &= \sum_{s,t} e^{js\vec{r}_1\cdot\vec{k}} e^{jt\vec{r}_2\cdot\vec{k}} \mathbf{F}\left(\delta\left(\vec{r}\right)\right) \\ &= \sum_{s} e^{js\vec{r}_1\cdot\vec{k}} \sum_{t} e^{jt\vec{r}_2\cdot\vec{k}} \\ &= G_1\left(\vec{k}\right).G_2\left(\vec{k}\right) \end{aligned} \tag{6.137}$$

As is known from the theory of distributions that

$$G_1\left(\vec{k}\right) = \sum_{s} e^{js\vec{r}_1\cdot\vec{k}} = \frac{1}{2\pi}\sum_{p} \delta\left(\bar{r}_1\cdot\vec{k} - 2\pi p\right) \tag{6.138}$$

we see that this function is non-zero when

$$\vec{k} = p.\vec{K}_1 + \vec{K}_{1\perp} \tag{6.139}$$

with

$$\begin{aligned} \vec{r}_1\cdot\vec{K}_1 &= 2\pi \\ \vec{r}_1\cdot\vec{K}_{1\perp} &= 0 \end{aligned} \tag{6.140}$$

Similarly we find that the function $G_2\left(\vec{k}\right)$ is non-zero when

$$\vec{k} = q\cdot\vec{K}_2 + \vec{K}_{2\perp} \tag{6.141}$$

with

$$\begin{aligned} \vec{r}_2\cdot\vec{K}_2 &= 2\pi \\ \vec{r}_2\cdot\vec{K}_{2\perp} &= 0 \end{aligned} \tag{6.142}$$

When we equalize equation 6.139 and equation 6.141 we find the k-vectors for which $G\left(\vec{k}\right)$ is non-zero.

$$p.\vec{K}_1 + \vec{K}_{1\perp} = q.\vec{K}_2 + \vec{K}_{2\perp} \tag{6.143}$$

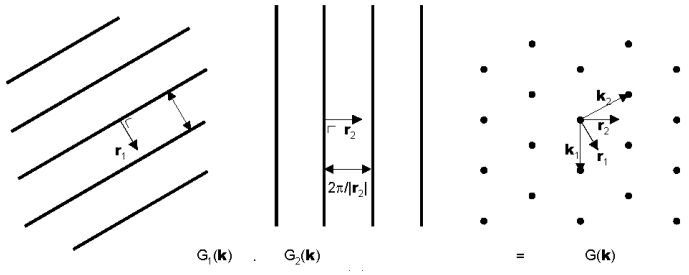


Figure 6.43: The 2D periodic lattice function is the product of two 1D lattice functions. These functions are zero except for some equidistant lines perpendicular to $\vec{r}_1$ and $\vec{r}_2$ respectively

and this for all values of p and q

As $\vec{K}_1 \neq \vec{0}$ and $\vec{K}_2 \neq \vec{0}$ (because of $\vec{r}_1 \cdot \vec{K}_1 = 2\pi$ and $\vec{r}_2 \cdot \vec{K}_2 = 2\pi$) this is only possible if

$$\begin{aligned} \vec{K}_{1\perp} &= q.\vec{K}_2 \\ \vec{K}_{2\perp} &= p.\vec{K}_1 \end{aligned} \tag{6.144}$$

or

$$\begin{aligned} \vec{r}_1 \cdot \vec{K}_2 &= 0 \\ \vec{r}_2 \cdot \vec{K}_1 &= 0 \end{aligned} \tag{6.145}$$

The reciprocal lattice function becomes non-zero when

$$\vec{k} = p.\vec{K}_1 + q.\vec{K}_2 \tag{6.146}$$

with

$$\vec{r}_i \cdot \vec{K}_j = 2\pi\delta_{ij} \tag{6.147}$$

and p,q integer.

From this discussion it is clear that when $\vec{r}_1$ and $\vec{r}_2$ are perpendicular (rectangular lattice), $\vec{K}_1$ and $\vec{K}_2$ will also be perpendicular. $\vec{K}_1$ has the same direction as $\vec{r}_2$ and $\vec{K}_2$ has the same direction as $\vec{r}_1$. In non-rectangular lattices the situation is different however. For the example in this appendix of a triangular lattice, this is no longer the case.

The above discussion can be easily extended to three dimensions.

Chapter 7

Photonic components

Contents

7.1 Introduction

In this chapter different basic photonic components will be described. The next chapter then will describe more complex photonic systems built from these components. One way to classify optical components is based on their implementation. We then distinguish the following types of components:

- Macro-optic components : bulk optical components with dimensions in the order of at least several millimeters up to centimeters (lenses, prisms, diffraction gratings ...).
- Micro-optic components : micro-version of the macro-optic components, dimensions in the sub-millimeter range.
- Fiber-based optical components : the light stays in the optical fiber (attenuators, splitters and combiners, fiber bragg gratings, ...).
- Integrated waveguide devices : light is guided by integrated waveguides, devices are integrated on a common substrate.

In this chapter however we will classify optical components according to their function. In a first section we will focus on devices affecting the polarization state of light (polarizers, polarization conversion and isolators). In a second section different optical modulators (temporal switches), both macro-optic and integrated devices, will be described. A third section will deal with beam

scanning devices and spatial switches, while in a last section we will deal with wavelength dependent devices. For each type of component different applications will be identified.

Important metrics for optical components are:

- Operational wavelength range (which may be determined by the intrinsic properties of the component or by the properties of anti-reflective coatings applied to the surface)
- Transmission, Reflection, Loss
- Wavefront distortion (which should be much smaller than the wavelength for interferometric components)
- Transmitted beam deviation
- Aperture diameter
- Operational temperature range
- Threshold damage
- Price

Depending on the application, different specifications for these metrics may be required. They will be discussed in the following sections.

7.2 Polarization controlling devices

7.2.1 Polarizers

A linear *polarizer* is anything, which when placed in an incident unpolarized beam, produces a beam of light whose electric vector is vibrating primarily in one plane, with only a small component vibrating in the plane perpendicular to it. If a polarizer is placed in a plane-polarized beam and is rotated around an axis parallel to the beam direction, the transmittance T will vary between a maximum T_1 and a minimum value T_2 according to the law[1]:

$$T = (T_1 - T_2)\, cos^2\theta + T_2 \tag{7.1}$$

With θ the angle between the principal axis of the polarizer and the angle of polarization of the incident beam. The polarizing properties of the polarizer are in general defined in terms of *its degree of polarization* P

$$P = \frac{T_1 - T_2}{T_1 + T_2} \tag{7.2}$$

[1]What will be the transmission for an unpolarized beam going through a single polarizer ? Going through two polarizers which are crossed or parallel ?

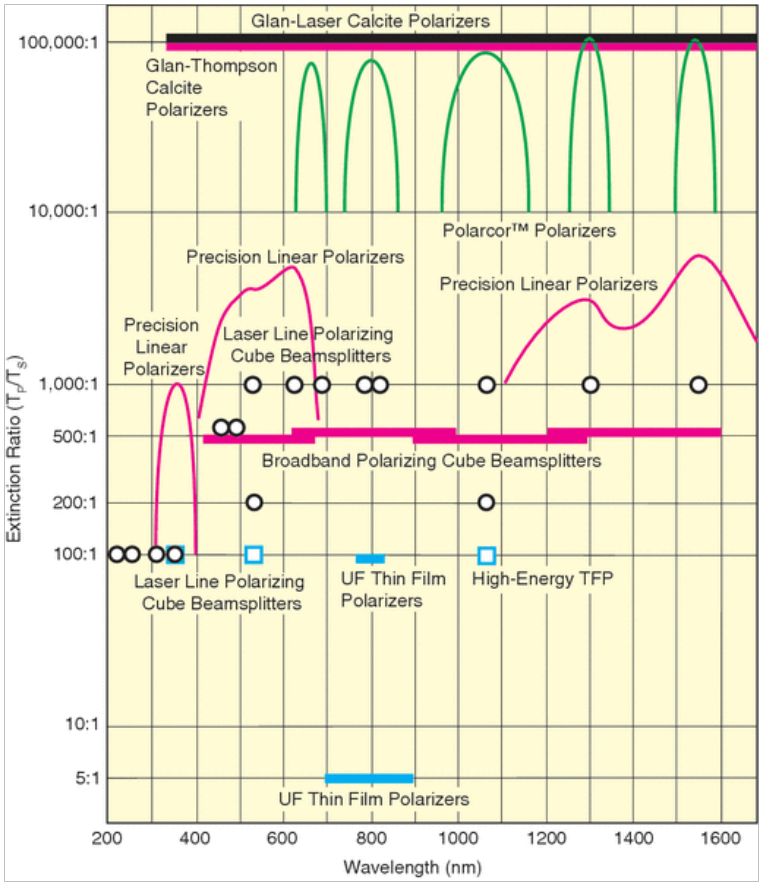


Figure 7.1: Different types of polarizers with operational wavelength range and extinction ratio (Figure taken from `http://www.newport.com/Polarization-Optics-Selection-Guide/141146/1033/catalog.aspx`)

or its *extinction ratio* ρ_p

$$\rho_p = \frac{T_2}{T_1} \tag{7.3}$$

Note that a polarizer cannot only be used to select a polarization state but also to analyze the polarization state of a linearly polarized beam. We then call the component an *analyzer*.

Many different mechanisms exist to realize a linear polarizer. Figure 7.1 gives an overview of several types of polarizers commercially available. The most important ones will be discussed below.

Anisotropic prisms

As can be seen from Figure 7.1, *calcite polarizers* exhibit the best performance, both in terms of operational wavelength range as in terms of extinction ratio. Calcite is a crystal showing strong birefringence. Birefringent materials have a different refractive index for light having its electrical field orthogonal to the optical axis (ordinary waves) compared to light having its electrical field

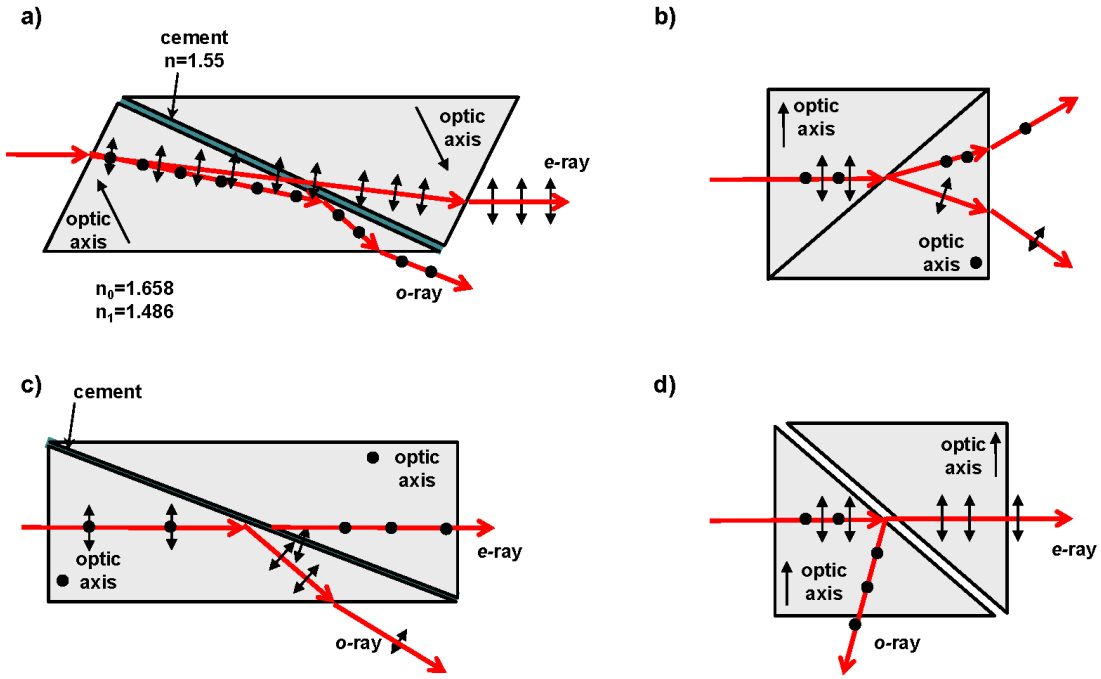


Figure 7.2: a) Nicol prism. b) Wollaston prism. c) Glan-Thompson prism. d) Glan-Taylor or Glan-Laser prism

parallel to the optical axis (extraordinary waves). For calcite the ordinary and extraordinary refractive index are respectively given by $n_0 = 1.658$ and $n_e = 1.486$. This anisotropy can be used to split linear polarization states.

Nicol Prism (Figure 7.2a) this was the first type of polarizing prism, invented in 1828 by William Nicol. It consists of a piece of calcite cut along a 68 degree angle, split diagonally and glued back together (with the refractive index of the glue chosen such that it lies between the ordinary and extraordinary index of the calcite). The optical axis of the calcite lies in the plane of incidence. Starting from unpolarized light, two waves will originate, an ordinary wave and an extraordinary wave, having a different propagation direction. By appropriately choosing the angle of incidence, we can achieve total internal reflection of the ordinary wave at the calcite-glue interface, while the extraordinary wave is transmitted. This way both polarization states are spatially separated. Although they were once widely used, Nicol prisms are now mostly replaced by Glan-Thompson and Glan-Taylor type prisms. The most important drawback of the Nicol prism is the large deviation between the input beam and the output beam.

Wollaston prism (Figure 7.2b). This prism is made up from two glued prisms with orthogonal optical axes as shown in 7.2b. Light is perpendicularly incident to the prism in such a way that the optical axis of the first prism lies in the plane of incidence. Both ordinary and extraordinary waves are generated. Due to the different orientation of the optical axis of the second prism, the ordinary wave becomes the extraordinary wave and vice versa. The two

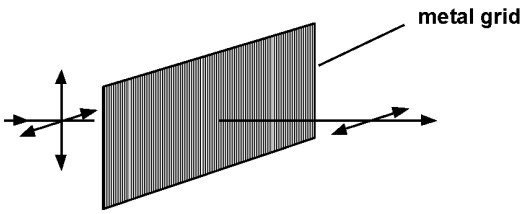


Figure 7.3: Wire grid polarizer

waves are transmitted under different angles. Like this, both polarization states are spatially separated.

Glan-Thompson prism (Figure 7.2c). This prism consists of two right-angled calcite prisms cemented together by their long faces. The optical axes are parallel to each other and perpendicular to the plane of incidence. Compared to the Glan-Taylor prism it has a much wider angle of acceptance but a much lower damage threshold due to the cement (approximately 100 times lower)

Glan-Taylor and Glan-Laser prism (Figure 7.2d). These prism consists of two right-angled calcite prisms with an air gap in between. The optical axes are parallel to each other and parallel to the plane of incidence. Because of the air gap the damage threshold is much higher than for the Glan-Thompson prism, but the angle of acceptance is reduced. In the Glan-Laser prism, the prism angle is steeper, which decreases reflection. These prisms are typically optimized for operation with very high optical powers (as those coming from lasers) and often the material is selected for high damage threshold.

Dichroic materials

A *dichroic material* shows an anisotropy in the imaginary part of the dielectric constant. One linear polarization state is preferentially absorbed. After a sufficiently long propagation through a dichroic material, only the polarization state that experienced the least absorption remains. These type of polarizers are quite lossy. An example of the use of a dichroic material is a Polaroid film, consisting of oriented dichroic molecules in a plastic sheet.

Dichroic polarizers can be large, light weight and have any shape. They are low cost but show a lower extinction ratio than crystals and may also have a relatively large insertion loss.

The polarizers denoted by *Precision Linear Polarizers* in Figure 7.1 are polymer films exhibiting anisotropic loss held between two glass plates. The polarizers indicated by *Polarcor* form a modern variation on classical organic dichroic materials. They consist of elongated silver particles in a glass surface and they also show preferential absorption on one direction (electrical field parallel to long axis of particles). They have much higher transmission and extinction ratio. They can also be looked at as a variation of the wire-grid polarizers discussed below.

A *wire grid polarizer* consists of a very fine grid (period much smaller than the wavelength of the incident light) of metallic wires or metallic tracks on a substrate as shown in figure 7.3. The linear polarization state with the electric field parallel to the wires is strongly attenuated due to the boundary condition at the metallic surface: the total electric field of this polarization state (tangential to the metallic wire surface) must vanish. When the pitch of the wires is sufficiently small, this means that the field in between the wires nearly vanishes as well and the wave doesn't propagate through the wire grid. The orthogonal polarization state is attenuated but transmitted by the grid.

These devices were historically mainly used in the far infrared and for radio waves and operate over a wide wavelength range. With the improved fabrication technology coming available, allowing for finer grids, they are now also used for shorter wavelength ranges.

Thin film polarizers

Thin film polarizers consist of a stack of dielectric layers. Interference effects cause them to act as beam-splitting polarizers. The substrate can be a glass plate (inserted in the beam at a particular angle) or a prism, with a second prism cemented to the film (then we call them beam splitting cubes). These should not be confused with the birefringent prisms discussed above.

Polarizers based on Brewster angle

The reflection and transmission of a plane wave at the interface between two dielectric materials is described by the fresnel equations 7.4 [**?**]. Reflection and transmission coefficients do not only depend on the incidence angle but also on the polarization state (transverse electric polarization is defined as the polarization state with the electric field orthogonal to the plane of incidence, transverse magnetic polarization is defined as the polarization state with the electric field in the plane of incidence).

$$\begin{cases} r_{TE} &= \dfrac{n_1\cos(\theta_1) - n_2\cos(\theta_2)}{n_1\cos(\theta_1) + n_2\cos(\theta_2)} \\ t_{TE} &= \dfrac{2n_1\cos(\theta_1)}{n_1\cos(\theta_1) + n_2\cos(\theta_2)} \\ r_{TM} &= \dfrac{n_2\cos(\theta_1) - n_1\cos(\theta_2)}{n_2\cos(\theta_1) + n_1\cos(\theta_2)} \\ t_{TM} &= \dfrac{2n_1\cos(\theta_1)}{n_2\cos(\theta_1) + n_1\cos(\theta_2)} \end{cases} \tag{7.4}$$

From equations 7.4 and Snell's law,

$$n_1\sin(\theta_1) = n_2\sin(\theta_2) \tag{7.5}$$

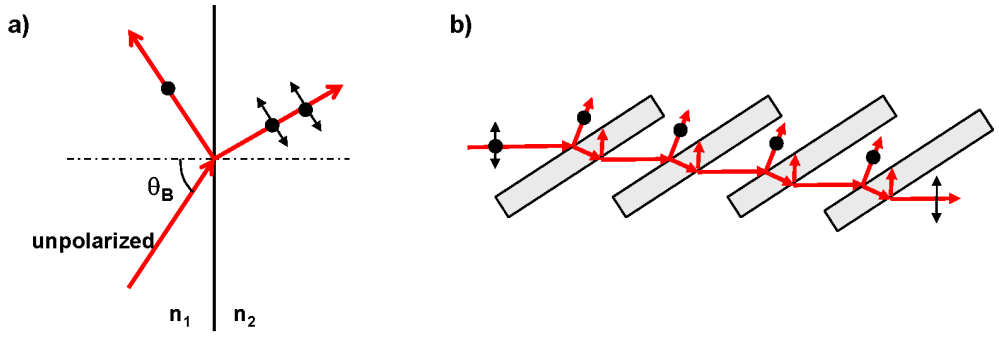


Figure 7.4: a) Light incident at the Brewster angle on an interface between two dielectrics. b) Concatenation of dielectric interfaces acting as polarizer

we can easily calculate that for a particular angle, the Brewster angle, the reflection coefficient of the transverse magnetic polarization state is zero and all power is transmitted for this polarization state. This angle is given by

$$\tan(\theta_B) = \frac{n_2}{n_1} \tag{7.6}$$

This means that an unpolarized beam incident to an interface of two dielectric materials at the Brewster angle results in a purely transverse electrical beam in reflection and a transmitted beam which has a relatively stronger contribution from the transverse magnetic polarization as shown in figure 7.4a. Concatenating these interfaces as shown in figure 7.4b results in a separation of the two polarization states. Thick plates are used to loose the coherence of the light beam and avoid interference effects.

These devices are used in the infrared and UV region where dichroic materials and calcite prisms do not work.

Specular reflection

Scattering from a particle smaller than the wavelength of light creates polarized light. Light scattered in a direction normal to the incident ray is linearly polarized. The vertically polarized component of the incident light cannot be scattered in the vertical direction because the **E**-field would become parallel to the direction of propagation. Therefore light scattered in the vertical direction will be highly horizontally polarized, and likewise, light scattered in the horizontal direction is highly vertically polarized.

A chamber filled with either N_2 or CO_2 molecules makes a polarizer. Although the amount of scattered light is small, the degree of polarization is good. This phenomenon also explains why sun light is partly polarized, through scattering at water molecules (see Figure 7.5). This effect is used by some animals as a means of orientation.

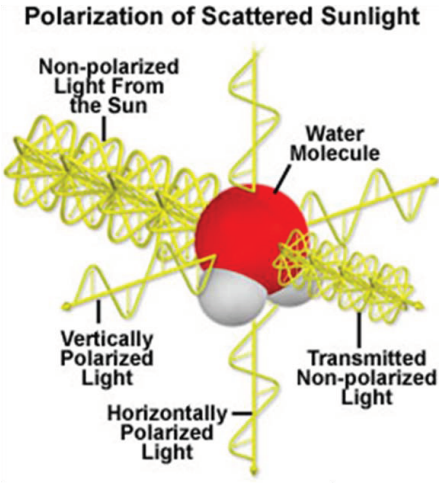


Figure 7.5: Polarization of sunlight through scattering at water molecules

7.2.2 Polarization conversion

Also polarization conversion is an important optical function. Again, the use birefringent materials is the preferred approach for realizing the desired polarization conversion action. As discussed in chapter 2, the Jones matrix of a plate of uniaxially anisotropic material with the optical axis in the plane of the plate can be written as

$$\begin{bmatrix} e^{-j\phi_o} & 0 \\ 0 & e^{-j\phi_e} \end{bmatrix} = e^{-j\phi_o} \begin{bmatrix} 1 & 0 \\ 0 & e^{-j\Delta\phi} \end{bmatrix} \tag{7.7}$$

with $\phi_i = k_0 n_i d$ and d the thickness of the plate. n_o and n_e are the ordinary and extraordinary refractive indices of the material. The total retardation of the plate is then given by

$$\Delta\phi = k_0 \left(n_e - n_0\right) d = \frac{2\pi \left(n_e - n_0\right) d}{\lambda} \tag{7.8}$$

Depending on the thickness of the plate and the angle of the optical axis with respect to the polarization of the incident beam different conversion actions can be realized.

Wave plates

- Quarter wave plate

  As already discussed in a previous chapter, a quarter wave plate can in general convert a linear polarization state (angle $\frac{\pi}{2} - \theta$ with respect to the optical axis) to an elliptical polarization state with its main axis parallel to the optical axis and ellipticity $tan(\theta)$ as can be seen from its Jones matrix description 7.9 ($\Delta\phi = \pm\frac{\pi}{2}$)

$$e^{-j\phi_o} \begin{bmatrix} 1 & 0 \\ 0 & \pm j \end{bmatrix} \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} = e^{-j\phi_o} \begin{bmatrix} \cos(\theta) \\ \pm j \sin(\theta) \end{bmatrix} \tag{7.9}$$

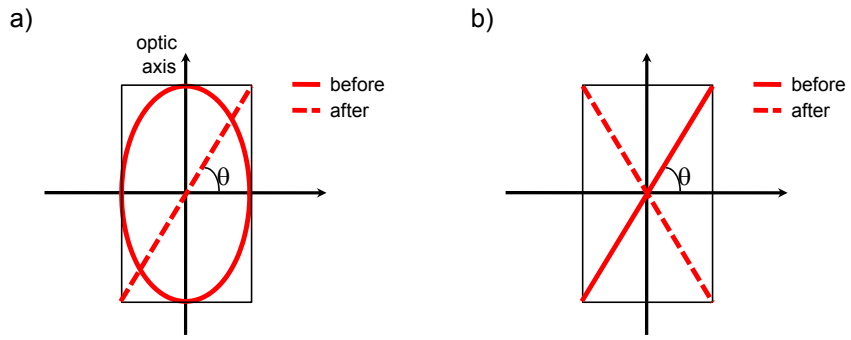


Figure 7.6: Polarization conversion in a) a quarter wave plate and b) a half wave plate

This is illustrated in figure 7.6.

For the special case of $\theta = \pm\frac{\pi}{4}$ the resulting polarization state is circular.

- Halve wave plate

  Besides a quarter wave plate, also a halve wave plate finds its application in polarization conversion. Considering the Jones matrix 7.7 for which $\Delta\phi = (n_e - n_o)\frac{2\pi}{\lambda}d = \pi$, it is easy to see that a right handed circularly polarized beam is converted into a left handed circularly polarized beam and vice versa as shown in equation 7.10.

$$e^{-j\phi_o}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ j \end{bmatrix} = e^{-j\phi_o}\begin{bmatrix} 1 \\ -j \end{bmatrix} \tag{7.10}$$

  A linear polarization state with polarization angle θ is transformed in a linear polarization state with polarization angle $-\theta$ as shown in equation 7.11. This is illustrated in Figure 7.6b.

$$e^{-j\phi_o}\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} = e^{-j\phi_o}\begin{bmatrix} \cos(\theta) \\ -\sin(\theta) \end{bmatrix} \tag{7.11}$$

  By rotating a halve wave plate any linear polarization state can be generated, when a linear polarization state is incident on the halve wave plate. By cascading a rotatable quarter wave plate and a halve wave plate any elliptical polarization state can be transformed into any linear polarization state.

- Practical wave plates

  The thickness of a wave plate is determined by equation (7.8). From this we can see that, to obtain mechanically stable plates with sufficient thickness d, the birefringence of the material should not be too large. In practice, quartz is often used (with $n_e = 1.555, n_o = 1.546$). Even then, a single order phase plate would be to thin to handle and therefore mostly so-called *multiple order phase plate* are used, for which the retardation is given by $\Delta\phi = 2m\pi + \Delta\phi_{desired}$. These plates show good performance, at a given wavelength, but is very wavelength dependent. In addition, the thinner the wave plate, the less sensitive to temperature variations, angular tilt and material defects.

  Multiple order wave plates can be designed to perform at two different wavelengths. However, if wide wavelength operation is required, zero-order wave plates are required. These

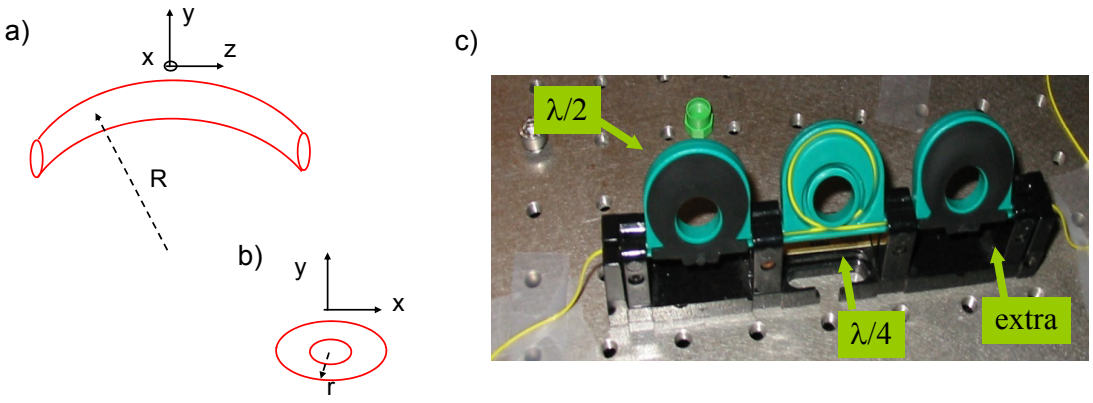


Figure 7.7: a) Fiber bend with radius R. b) Cross section of bend fiber with r the radius of the fiber core. c) Practical implementation of fiber polarization rotator.

can be fabricated as so-called *compound zero order retardation plates* or as *true zero order retardation plates*. The *compound zero order retardation plates* are formed from two multiple order plates assembled together with orthogonal optical axis, and with thicknesses chosen such that they yield the desired retardation. A variation are the *achromatic compound zero order retardation plates* which are built from two materials with opposing material dispersion($d\Delta n/d\lambda$), so that they can operate over a very broad wavelength range. *True zero order retardation plates* are formed by a birefringent polymer cemented between two glass plates.

Fiber loop polarization controller

Also fiber-optic polarization controllers can be realized based on birefringence. When bending an optical fiber birefringence is induced. In the radial direction (y-direction) the fiber is compressed, leading to a lowering of the refractive index. In the lateral direction (x-direction) the fiber is stretched and the refractive index raises. This is shown in figure 7.7a and b. Together this leads to a difference in effective index (propagation constant) for the wave polarized in the radial direction and the wave polarized in the lateral direction given by $\Delta n = 0.0439 n^3 \left(r/R\right)^3$ with r the radius of the fiber core and R the radius of the fiber bend.

A fiber based polarization controller can be realized as shown in Figure 7.7c. Several rotatable fiber coils are placed in series. The radius R and number of turns N of the first ring are chosen such that halve wave plate conditions are obtained ($\Delta\phi = 2\pi RN(\frac{2\pi}{\lambda}\Delta n) = \pi$) and the number of turns of the second ring is halved to obtain quarter wave plate conditions. The optical axes of both plates can be rotated with respect to the input polarization, by orienting the rings. A third ring is added to allow a larger flexibility.

Babinet-Soleil compensator

A Babinet-Soleil compensator, as shown in figure 7.8, consists of two cascaded uniaxially anisotropic plates. The first plate consists out of two prisms that can be shifted along each other. This allows modifying the thickness of the first plate. The optical axes of the two

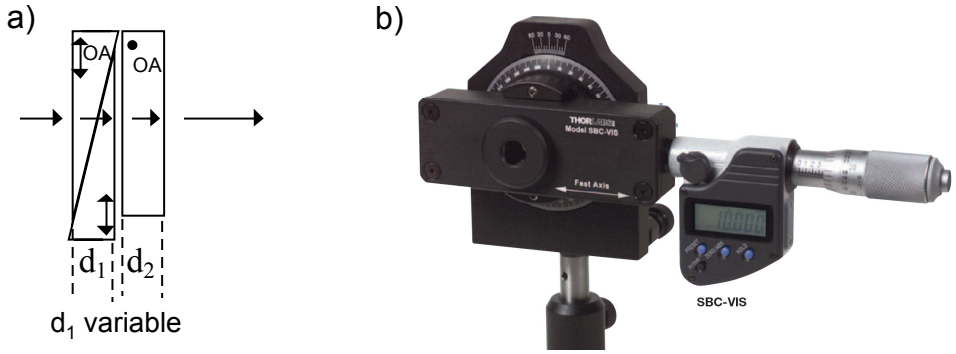


Figure 7.8: Babinet-Soleil compensator with a) schematic representation and b) practical implementation

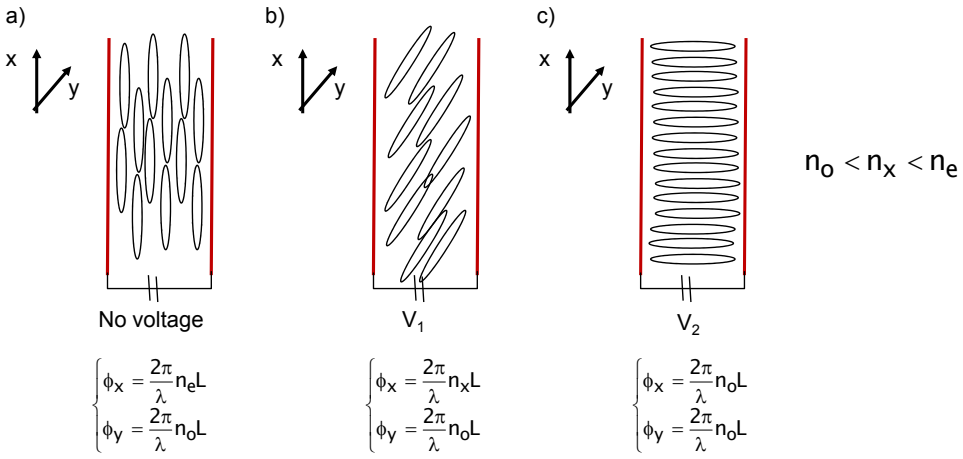


Figure 7.9: Liquid crystal variable retardation plate (liquid crystal molecules NOT drawn to scale!)

prisms are parallel and are perpendicular to the optical axis of the second plate. The optical axis of the first plate lies in the plane of incidence and light is incident perpendicular to the plates. Ordinary and extraordinary waves are generated in the first plate. The ordinary wave becomes the extraordinary wave in the second plate and vice versa due to the perpendicular optical axes. Because the thickness of the first plate is variable, any phase difference can be induced between both polarization states, given by

$$\Delta\phi_{tot} = \Delta\phi_1 + \Delta\phi_2 = -\frac{2\pi}{\lambda} d_1(n_e - n_o) - \frac{2\pi}{\lambda} d_2(n_o - n_e) = \frac{2\pi}{\lambda}(d_2 - d_1)(n_e - n_o) \qquad (7.12)$$

Additional rotation of the complete compensator allows to generate any possible Jones matrix, thereby allowing to convert any polarization state into any other.

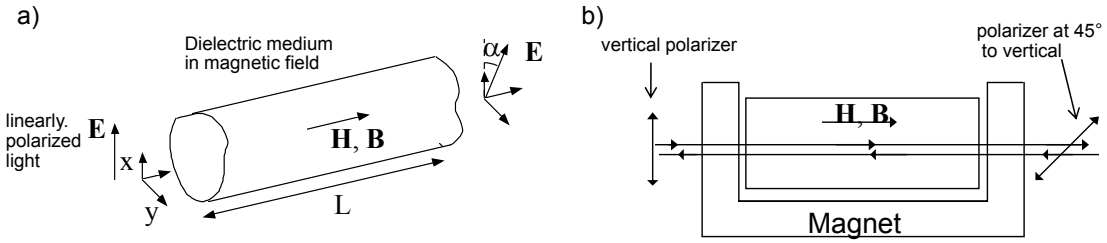


Figure 7.10: a) Faraday effect. b) Optical isolator

Liquid Crystal Polarizing Optics

It is well known that liquid crystals show strong uniaxial anisotropy (see also section 7.3.8). In addition, the optical axis can be controlled by applying an electrical field. The working principle is illustrated in Figure 7.9. If no field is applied the orientation of the liquid crystal molecules is determined by a passive alignment layer on the surfaces of the holder. In this case the alignment is chosen in such a way that all molecules (and hence the optical axis) are oriented along the x-direction (Figure 7.9a). As a consequence, a wave incident on the plate with polarization along the x-axis will see a different refractive index then a wave with its polarization oriented along the y-direction. The total retardation in this case will be given by $\Delta\phi = \phi_x - \phi_y = \frac{2\pi}{\lambda}(n_e - n_o)d$.

When a large enough electrical field is applied, all molecules (and the optical axis) are oriented in the direction of the electrical field and both x and y polarized waves will see the ordinary refractive index. Hence, the total retardation is equal to zero in this case(Figure 7.9c).

By applying an intermediate electrical field, all retardations between both extreme cases can be realized (Figure 7.9b).

7.2.3 Isolators

An optical isolator is a non-reciprocal component allowing light propagation in one direction and inhibiting light propagation in the opposite direction. Optical isolators are mainly used to avoid light coupling back from an optical system into a laser diode (leading to extra noise in the laser output or possible coherence collapse of the laser diode) while still allowing light to be coupled from the laser diode into the optical system.

Optical isolators are based on the Faraday effect. The Faraday effect occurs when placing a dielectric material in a magnetic field as shown in Figure 7.10left. The incident optical field induces movement of the electrons in the dielectric medium. The static magnetic field induces a Lorentz force onto the charges given by $-q\mathbf{v} \times \mathbf{B}$. This effect has influence on the dielectric tensor so that a left circular wave has a different velocity than right circular waves, leading to a rotation of the incident linear polarization (this can be understood by looking at a linear polarization state as the superposition of a right-handed and left-handed circular polarization state). The rotation angle α is given by

$$\alpha = VB_zL \tag{7.13}$$

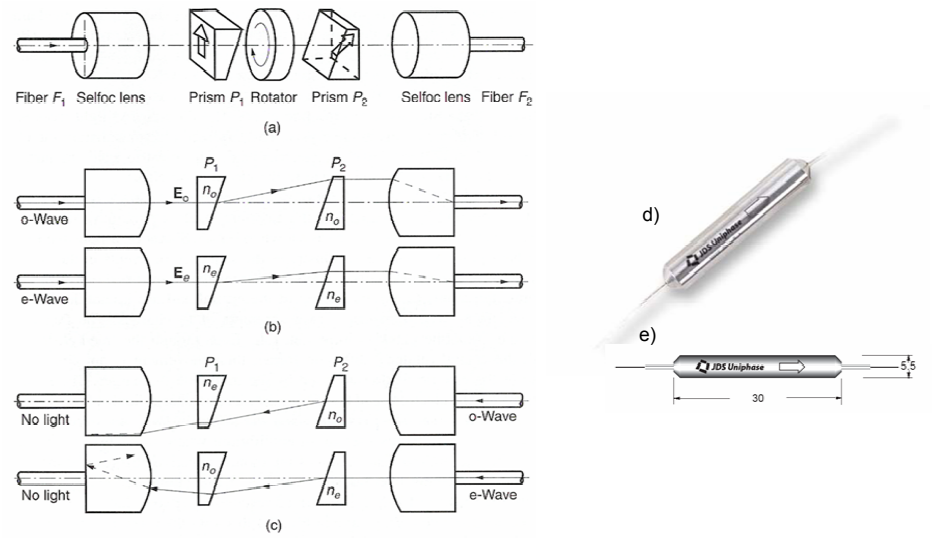


Figure 7.11: Fiber integrated polarization independent isolator. Panels a) to c) schematically show the operation principle, d) and e) show a practical implementation

with V being the Verdet constant of the material and is dependent on the direction of the applied magnetic field. The latter is the basis of the non-reciprocal behavior of the component. A popular material exhibiting a strong Faraday effect is YIG (Yttrium ion garnet or Y3Fe5O12). When a strong magnetic field is applied, all electrons spin in the same direction and the effect saturates. This allows for easy control over the effect. Note that the Faraday effect is fundamentally different from the *optical active materials* such as sugar described in chapter 2. In the case of an optical active materials light is rotated in the same direction when propagating forwards and backwards through the material. Optical active materials are reciprocal.

To realize a practical optical isolator, a material with a large Verdet is placed in between the poles of a magnet. A vertically oriented polarizer is placed at the left hand side of the device and a second polarizer, oriented by $\pi/4$ with respect to the vertical axis is placed at the right hand side, as shown in Figure 7.10right. Through the Faraday effect, the material in between the magnet behaves as a polarization rotator and the length of the device is chosen to rotate incoming linearly polarized light over an angle $\pi/4$, resulting in a high transmission from the left hand side to the right hand side. Light incident from the right hand side ($\pi/4$ polarization angle) is further rotated towards the left hand side, resulting in a polarization state perpendicular to the vertical polarizer. Hence there is very low transmission from the right hand side to the left hand side and an optical isolation effect is accomplished. Practically, isolation levels of 30dB are achieved.

One drawback of this device is its limited optical bandwidth as the Faraday effect is wavelength sensitive and the anti-reflection coatings on both ends are also wavelength sensitive.

A further drawback of the device described above is that it only works for polarized beams (because of the polarizer used at the entrance). If operation for unpolarized light is required, e.g. when used in combination with fiber optics, more complex schemes are required. Figure 7.11 illustrates a possible scheme for a fiber integrated optical isolator employing birefringent crystals at the entrance and output sides.

7.3 Modulators

7.3.1 System level description

As telecommunication systems require the optical transmission of data, a light beam needs to be modulated in time to form a data signal. The most commonly used modulation scheme is on-off keying (OOK). In this modulation scheme a *1* bit is encoded by the presence of a light pulse in the bit interval or by turning a light source (laser or LED) on. A *0* bit is encoded by the absence of a light pulse in the bit interval or by turning the light source off. To achieve this we can either directly modulate the light source by turning it on and off or use an external modulator behind the source to perform the same function. Using an external modulator is the preferred approach for high-speed transmission over long distances. Recently also more complex modulation schemes such as *PSK*(Phase Shift Keying), *DPSK* (Differential Phase Shift Keying) and *QDPSK* (Quadrature Differential Phase Shift Keying). In those cases, also the phase of the light needs to be modulated. Modulators used in telecommunication networks are modulated at speeds ranging from 10GBits/s to 100Gbit/s. Typical telecom modulators have a single input and a single output.

Also in other applications, light modulators are very important components. E.g. in projectors, light needs to be modulated for every pixel. Although the speed is typically much lower in these cases, these modulators can be considered as N to N devices. This type of modulator is also often called a *Spatial Light Modulator* (SLM). Further applications for optical modulators include pulsed lasers, printing applications, some specialized camera applications and more complex optical systems.

Many different mechanisms can be employed for modulating the amplitude or phase of an optical beam. They include refractive effects such as the electro-optic, the acousto-optic effect, the magneto-optic effect and the thermo-optic effect, which change the real part of the refractive index (and hence induce a phase change, which then can be converted in a amplitude modulation through the use of an interferometer. The electro-absorption effect changes the imaginary part of the refractive index and hence directly impacts the amplitude of the beam. Other effects include the use of liquid crystals and electro-mechanical effects. The parameters determining the figure of merit (FOM) for a modulator are its speed, size, power consumption and cost. Depending on the application a different type of modulator may present the best compromise between these parameters.

7.3.2 Index Ellipsoid

Within some materials anisotropy can be induced or modified, by applying an electric field or a strain field. For example, the linear Pockels effect in non-centrosymmetric crystals results in a change of the refractive index experienced by a wave propagating in this material when applying an electric field.

An attractive way to describe material anisotropy is to define an index ellipsoid determined by

$$\left(\frac{x}{n_x}\right)^2 + \left(\frac{y}{n_y}\right)^2 + \left(\frac{z}{n_z}\right)^2 = 1 \tag{7.14}$$

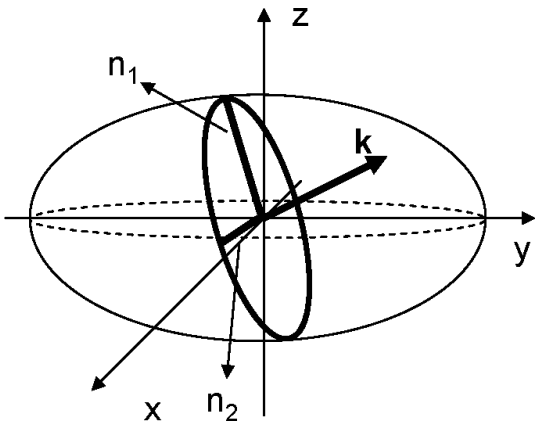


Figure 7.12: Index ellipsoid and determination of effective indices

with x, y and z being the principal axes of the material. Using a graphical representation of the index ellipsoid it is easy to derive the effective index of the ordinary and extraordinary wave propagating in the material in a certain direction (without proof). For a given propagation direction $\mathbf{k}$, find the plane through the origin that is perpendicular to this direction $\mathbf{k}$. This plane intersects the index ellipsoid along an ellipse. Half the length of the major and minor axis determine the effective index of respectively the ordinary and extraordinary wave as shown in figure 7.12.

The influence of the application of a field (electrical or strain induced) on the anisotropy of the material can then be described by a change in coefficients describing the index ellipsoid as in equation 7.15.

$$(\frac{1}{n_x^2}+\Delta_1)x^2+(\frac{1}{n_y^2}+\Delta_2)y^2+(\frac{1}{n_z^2}+\Delta_3)z^2+2\Delta_4 yz+2\Delta_5 xz+2\Delta_6 xy=1 \tag{7.15}$$

The form of the coefficients Δ_i will depend strongly on the effect being considered. E.g. in the case of the linear Pockels effect, the coefficients are linearly dependent on the applied electrical field and can be written as

$$\begin{bmatrix}\Delta_1\\ \Delta_2\\ \Delta_3\\ \Delta_4\\ \Delta_5\\ \Delta_6\end{bmatrix}=\begin{bmatrix}r_{11} & r_{12} & r_{13}\\ r_{21} & r_{22} & r_{23}\\ r_{31} & r_{32} & r_{33}\\ r_{41} & r_{42} & r_{43}\\ r_{51} & r_{52} & r_{53}\\ r_{61} & r_{62} & r_{63}\end{bmatrix}\begin{bmatrix}E_x\\ E_y\\ E_z\end{bmatrix} \tag{7.16}$$

Figure 7.13 shows the electro-optic tensor for some important materials exhibiting a large Pockels effect. Note that only a few of the components r_{ij} are non zero, which will simplify the analysis considerably (see below).

Name of Substance	Chemical Symbol	Pockels Coefficient (10^{-12} m/V)	Index of Refraction	Wavelength (μm)	Crystal Symmetry	Electrooptic Tensor
Potassium dihydrogen phosphate	KH_2PO_4 or KDP	$r_{41} = 8.6$ $r_{63} = 10.5$ (T)	$n_x = n_y = 1.51$ $n_z = 1.47$	0.63	$\bar{4}2m$	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{41} & 0 \\ 0 & 0 & r_{63} \end{bmatrix}$
Ammonium dihydrogen phosphate	$NH_4H_2PO_4$ or ADP	$r_{41} = 23.1$ $r_{63} = 8.5$ (T)	$n_x = n_y = 1.52$ $n_z = 1.48$	0.63	$\bar{4}2m$	
D-KDP	KD_2PO_4	$r_{41} = 8.8$ $r_{63} = 26.4$ (T)	$n_x = n_y = 1.51$ $n_z = 1.47$	0.63	$\bar{4}2m$	
Quartz	SiO_2	$r_{11} = 0.29$ (S) $r_{41} = 0.2$ (T)	$n_x = n_y = 1.546$ $n_z = 1.555$	0.63	32	$\begin{bmatrix} r_{11} & 0 & 0 \\ -r_{11} & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & -r_{41} & 0 \\ 0 & -r_{11} & 0 \end{bmatrix}$
Cinnabar	HgS	$r_{11} = 3.1$ $r_{41} = 1.4$	$n_x = n_y = 2.885$ $n_z = 3.232$	0.63		
Lithium niobate	$LiNbO_3$	$r_{13} = \begin{cases} 10 \text{ (T)} \\ 8.6 \text{ (S)} \end{cases}$ $r_{33} = \begin{cases} 32.2 \text{ (T)} \\ 30.8 \text{ (S)} \end{cases}$ $r_{22} = \begin{cases} 6.7 \text{ (T)} \\ 3.4 \text{ (S)} \end{cases}$ $r_{51} = \begin{cases} 32 \text{ (T)} \\ 28 \text{ (S)} \end{cases}$	$n_x = n_y = 2.286$ $n_z = 2.200$	0.63	$3m$	$\begin{bmatrix} 0 & -r_{22} & r_{13} \\ 0 & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{51} & 0 \\ r_{51} & 0 & 0 \\ -r_{22} & 0 & 0 \end{bmatrix}$
Lithium tantalate	$LiTaO_3$	$r_{13} = 7.0$ (S) $r_{33} = 27$ (S) $r_{22} = 1.0$ (S) $r_{51} = 20$ (S)	$n_x = n_y = 2.176$ $n_z = 2.180$	0.63	$3m$	

Figure 7.13: Pockels Coefficient for some important electro-optical materials

In the case of the quadratic Kerr effect, the coefficients in the electro-optic tensor depend on the square of the electric field components and we have:

$$\begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \Delta_4 \\ \Delta_5 \\ \Delta_6 \end{bmatrix} = \mathbf{S} \begin{bmatrix} E_x E_x \\ E_y E_y \\ E_z E_z \\ E_x E_y \\ E_z E_x \\ E_y E_z \end{bmatrix} \tag{7.17}$$

Now the tensor **S** has 36 components. Luckily, in practical situations, only a few tensor elements are independent of each other. E.g. for crystals with $\bar{4}3m$ symmetry (used for realizing semiconductor integrated optical modulators)we find:

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & s_{12} & 0 & 0 & 0 \\ s_{12} & s_{11} & s_{12} & 0 & 0 & 0 \\ s_{12} & s_{12} & s_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & s_{11} - s_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 & s_{11} - s_{12} & 0 \\ 0 & 0 & 0 & 0 & 0 & s_{11} - s_{12} \end{bmatrix} \tag{7.18}$$

In the case of the elasto-optic effect, the coefficients Δ_i are proportional to the applied strain field.

7.3.3 Pockels effect in macro-optic devices

In this part we will describe the amplitude and phase modulation of a collimated beam traversing a crystal exhibiting a linear Pockels effect. As can be seen from Figure 7.13 in many such crystals the change of anisotropy when applying an electric field is described by an r_{ij} tensor of the form

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ r_{41} & 0 & 0 \\ 0 & r_{52} & 0 \\ 0 & 0 & r_{63} \end{bmatrix} \tag{7.19}$$

Semiconductor materials (GaAs, InP, ZnSe) with a Zinc Blende crystal and materials like potassium dihydrogen phosphate (KDP or KH_2PO_4) exhibit this behavior (with $r_{41} = r_{52}$).

The index ellipsoid when applying an electrical field is then described by

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} + 2r_{41}E_x yz + 2r_{52}E_y xz + 2r_{63}E_z xy = 1 \tag{7.20}$$

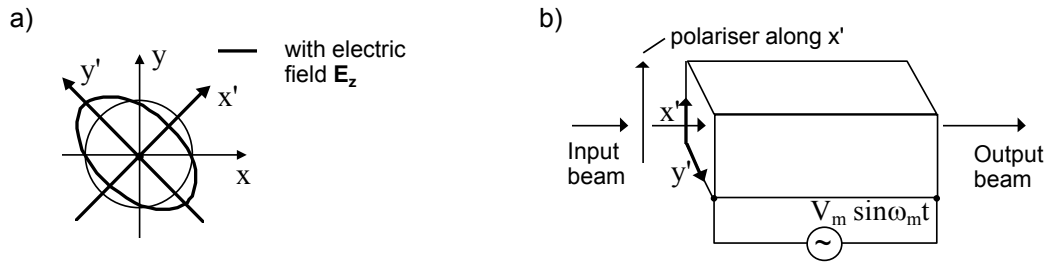


Figure 7.14: a) Change of index ellipsoid when applying an electric field. b) Longitudinal phase modulator

longitudinal configuration

If we assume that the applied static electric field lies along the optical axis and the optical axis is along the propagation direction of the beam, i.e. the z-direction (referred to as the *longitudinal configuration*), we can rewrite equation 7.20 as

$$\frac{x^2}{n_o^2} + \frac{y^2}{n_o^2} + \frac{z^2}{n_e^2} + 2r_{63}E_z xy = 1 \tag{7.21}$$

Rotating the x and y-axis over $\frac{\pi}{4}$, equation 7.21 can be rewritten as

$$\frac{x'^2}{n_x'^2} + \frac{y'^2}{n_y'^2} + \frac{z^2}{n_e^2} = 1 \tag{7.22}$$

with

$$\begin{cases} \frac{1}{n_x'^2} = \frac{1}{n_0^2} + r_{63}E_z \\ \frac{1}{n_y'^2} = \frac{1}{n_0^2} - r_{63}E_z \end{cases} \tag{7.23}$$

The index ellipsoid in the xy plane is shown in figure 7.14a.

If the refractive index change due to the electric field is small, we can rewrite equation 7.23 as

$$\begin{cases} n_x' = n_o - \frac{n_o^3 r_{63} E_z}{2} \\ n_y' = n_o + \frac{n_o^3 r_{63} E_z}{2} \end{cases} \tag{7.24}$$

Consider the phase modulator configuration as shown in figure 7.14b with an input polarizer parallel to the x' axis.

When the optical input beam is written as $E = A\cos(\omega t)$, the output electric field can be written as

$$E = A\cos(\omega t - \frac{2\pi}{\lambda}(n_0 - \frac{n_0^3 r_{63} E_m \sin(\omega_m t)}{2})L) \tag{7.25}$$

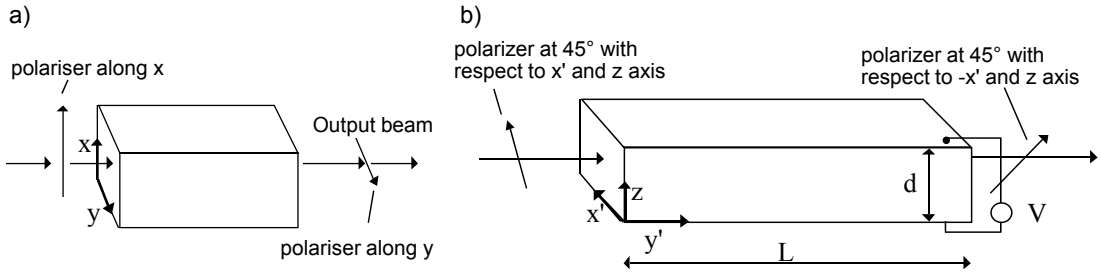


Figure 7.15: Amplitude modulator in the a) longitudinal configuration and b) transversal configuration.

with $E_m L = V_m$. From these equations the phase modulation operation is clear.

We can now consider the amplitude modulator configuration as shown in figure 7.15a. In this case the input polarizer is along the x-axis and there is an additional polarizer along the y-axis at the output.

The optical field at the input can be decomposed by its field components along x' and y', each propagating with their own phase velocities.

$$\begin{cases} E_{x',in} = \frac{A}{\sqrt{2}} \\ E_{y',in} = -\frac{A}{\sqrt{2}} \end{cases} \tag{7.26}$$

The output field strength along the y-direction is then given by

$$E_{y,out} = \frac{A}{2}\left[\exp(-j\phi_{x'}) - \exp(-j\phi_{y'})\right] \tag{7.27}$$

or

$$< |E_{y,out}|^2 >= A^2 \sin^2\left(\frac{\phi_{x'} - \phi_{y'}}{2}\right) \tag{7.28}$$

The phase difference $\phi_{x'} - \phi_{y'}$ can be written as

$$\phi_{x'} - \phi_{y'} = \frac{2\pi}{\lambda}(n_0 - \frac{n_o^3 r_{63} E_m}{2})L - \frac{2\pi}{\lambda}(n_0 + \frac{n_o^3 r_{63} E_m}{2})L = -\frac{2\pi}{\lambda} n_o^3 r_{63} E_m L \tag{7.29}$$

The ratio of output power to input power can be written as

$$\frac{I_{out}}{I_{in}} = \sin^2(\frac{\pi}{\lambda} n_o^3 r_{63} E_m L) = \sin^2(\frac{\pi}{2}\frac{V}{V_\pi}) \tag{7.30}$$

with $V_\pi = \frac{\lambda}{2n_0^3 r_{63}}$

The longitudinal configuration has two disadvantages: the electrodes must be transparent and the voltage needed for modulation is independent of the length of the device. Because the Pockels effect is a small effect, large voltages (on the order of 10kV) are needed.

Transversal configuration

To circumvent the problems with the longitudinal configuration, we can consider the transversal configuration as shown in figure 7.15b.

The optical axis still lies along the z-axis but now the light is propagating along the y'-axis. The ratio of output power to input power can be written as

$$\frac{I_{out}}{I_{in}} = \sin^2(\frac{\phi_z - \phi_{x'}}{2}) \tag{7.31}$$

with

$$\phi_z - \phi_{x'} = -\frac{2\pi}{\lambda}(n_e - n_o + \frac{n_o^3 r_{63} E_m}{2})L \tag{7.32}$$

Using $V = E_m d$ we can write

$$\frac{I_{out}}{I_{in}} = \sin^2(\frac{\pi}{\lambda}L(n_o - n_e) - \frac{\pi}{2}\frac{V}{V_\pi}) \tag{7.33}$$

with $V_\pi = \lambda d / L n_o^3 r_{63}$. This voltage is a function of the length of the device and can be reduced by making the device longer.

7.3.4 Pockels effect in integrated waveguide devices

For realizing integrated waveguide modulators employing the Pockels effect either compound semiconductor materials (GaAs, InP) or ferro-electric materials are being used. GaAs and InP are direct semiconductors and are used for fabricating efficient lasers emitting in the telecom range. It is possible to integrate such a laser together with a modulator on the same platform. Ferro-electric materials allow for very low waveguide losses and do not show parasitic effects. Therefore, they show very good performance. They cannot easily be integrated with other active devices such as lasers however.

In both cases, the configuration of the integrated modulators is rather similar to the *transversal configuration* for bulk modulators as described above. An electrode is applied on top of the waveguide and the field is applied perpendicular to the propagation direction. To translate the phase modulation in an amplitude modulation integrated interferometer structures such as an Mach-Zehnder interferometer are used. Because the optical field can be confined in a small region (the waveguide), much lower voltages are needed for reaching a desired field strength.

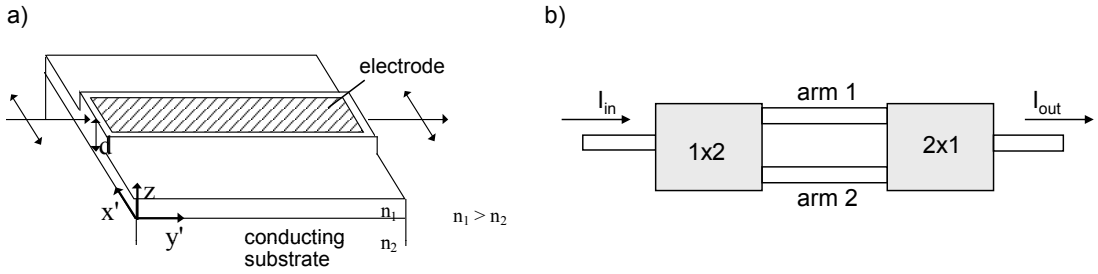


Figure 7.16: a) Integrated phase modulator in transverse configuration. b) 1x1 Mach-Zehnder interferometer

InP and GaAs based integrated modulators

InP and GaAs show an electro-optic tensor r_{ij} as described in equation 7.19 with $r_{41} = r_{52} = r_{63}$. The phase change per unit length and per Volt for a wave with its polarization in the plane of the waveguide (x'-axis) is then given by (see Figure 7.16a):

$$\frac{\Delta\phi_{x'}}{VL} = \frac{\pi}{\lambda d} n_o^3 r_{63} \tag{7.34}$$

To translate this phase modulation into an amplitude modulation, an integrated interferometer is required. We will describe the electro-optic Mach-Zehnder interferometer and the electro-optic directional coupler.

- *Electro-optic Mach-Zehnder interferometer*

  The structure of a Mach-Zehnder interferometer is depicted in figure 7.16b. It consists of an input waveguide, a 1x2 3dB power splitter, two waveguide arms, a 2x1 power combiner and an output waveguide. The power splitter and combiner (which are identical components but used in the opposite sense) can be a symmetrical y-junction, a symmetrical multimode interference coupler (MMI) or a directional coupler. The transmission characteristic of this interferometer can be written as

$$\frac{I_{out}}{I_{in}} = \left| \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} e^{-j\phi_1} & 0 \\ 0 & e^{-j\phi_2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \right|^2 = \cos^2\left(\frac{\phi_1 - \phi_2}{2}\right) \tag{7.35}$$

  With ϕ_1 and ϕ_2 the phase change in respectively the upper and lower arm of the interferometer. From this equation one can see that a Mach-Zehnder interferometer transforms a phase difference (phase modulation of one or both of the arms of the interferometer) to an intensity variation (amplitude modulation) in the output waveguide. We can now use the linear Pockels effect to modulate the refractive index of one arm of the Mach-Zehnder interferometer (as shown in figure 7.16) to modulate the phase difference between both arms. Assuming that the phase difference is zero when no electric field is applied, the transmission characteristic of the device becomes

$$\frac{I_{out}}{I_{in}} = \cos^2\left(\frac{\phi_1 - \phi_2}{2}\right) = \cos^2\left(\frac{\pi}{2\lambda d} n_0^3 r_{63} L_{arm} V\right) \tag{7.36}$$

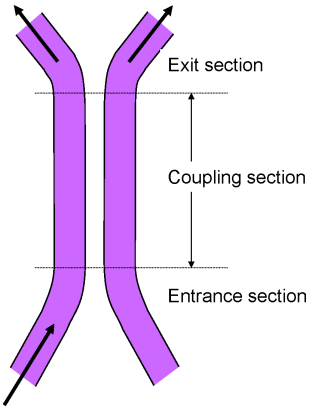


Figure 7.17: Directional coupler

- *Electro-optic directional coupler*

  Alternatively a directional coupler with an electrode on one of both arms can be used as an electro-optic modulator. As discussed in chapter 4, when two optical waveguides are brought close together, periodic power exchange between these waveguides will occur. This device is called a directional coupler and is shown in figure 7.17.

  The transmission from port 1 to port 3 can be written as

$$\frac{I_3}{I_1} = \frac{\kappa^2}{\kappa^2 + \frac{\Delta\beta^2}{4}} \sin^2(\sqrt{\kappa^2 + \frac{\Delta\beta^2}{4}} L) \tag{7.37}$$

  in which κ is the coupling strength of the device and $\Delta\beta$ is the difference in propagation constants of both unperturbed waveguides.

  By applying the electro-optic effect we can modify the difference in propagation constants $\Delta\beta$ of the unperturbed waveguides by changing the effective index of the waveguide modes. By carefully choosing this refractive index change (and the corresponding modulation voltage) light can be switched from port 3 to port 4. Hence this device shows both temporal and spatial switching behavior.

  Suppose the waveguides are physically identical ($\Delta\beta = 0$ with no electric field applied). Let the length of the directional coupler be $L = \frac{\pi}{2\kappa}$, such that all the power is coupled from the one waveguide to the other (cross state). To switch to the bar state, a phase change of $\Delta\beta L = \sqrt{3}\pi$ is needed.

Besides the linear optical effect, several other effects occur when applying an electrical field over a compound semiconductor material. These include the quadratic Kerr effect and carrier effects such as band filling, band gap shrinkage and the plasma effect. These effects induce non-linearities in the modulation transfer curve and are sometimes unwanted because they limit the transmission range of the signal.

subsubsectionFerro-electric switch

In nature, most materials are neutrally charged, meaning that their positive and negative parts are placed in such a way that there is no charge within the actual molecule. However, there

are some materials like ferro-electric materials in which the positively and negatively charged parts are slightly off-centered, thereby inducing a macroscopic electrical dipole moment. Optical modulators can be made with ferro-electric materials because the electro-optic effects are large and fast (GHz range). Typical examples of these materials are $LiNbO_3$ (lithium niobate) and $LiTaO_3$ (lithiumtantalate). Ferroelectric materials typically have a more complex r_{ij} matrix than equation 7.19. The r_{ij} matrix for $LiNbO_3$ for example is of the form

$$\begin{bmatrix} 0 & -r_{22} & r_{13} \\ 0 & r_{22} & r_{13} \\ 0 & 0 & r_{33} \\ 0 & r_{51} & 0 \\ r_{51} & 0 & 0 \\ -r_{22} & 0 & 0 \end{bmatrix} \tag{7.38}$$

$LiNbO_3$ is uniaxially anisotropic ($n_o = n_x = n_y$ and $n_e = n_z$).

Since the r_{33} component is the largest, an applied electric field along the z direction will be most efficient for electrooptic control. Assuming $E_x = E_y = 0$, the index ellipsoid is given by

$$x^2(\frac{1}{n_o^2} + r_{13}E_z) + y^2(\frac{1}{n_o^2} + r_{13}E_z) + z^2(\frac{1}{n_e^2} + r_{33}E_z) = 1 \tag{7.39}$$

This equation can be rewritten as

$$\frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2} = 1 \tag{7.40}$$

with

$$\begin{cases} n_x = n_y \approx n_o - \frac{1}{2}n_o^3 r_{13}E_z \\ n_z \approx n_e - \frac{1}{2}n_e^3 r_{33}E_z \end{cases} \tag{7.41}$$

Depending on the orientation of the crystal and the positioning of the electrodes different types of phase modulators can be realized. We will discuss the X-cut $LiNbO_3$ and the Z-cut $LiNbO_3$ phase modulator.

- *X-cut $LiNbO_3$ crystal*

  For an X-cut phase modulator two electrodes are placed symmetrically on both sides of the waveguide such that the bias field is along the z-direction as shown in figure 7.18left. Light is propagating along the waveguide in the y-direction.

  For TE polarization, the phase change due to an applied electric field is given by

$$\Delta\phi_{TE,X} = -\frac{2\pi}{\lambda}\frac{1}{2}n_e^3 r_{33}E_z L \tag{7.42}$$

  while TM polarization experiences a phase change

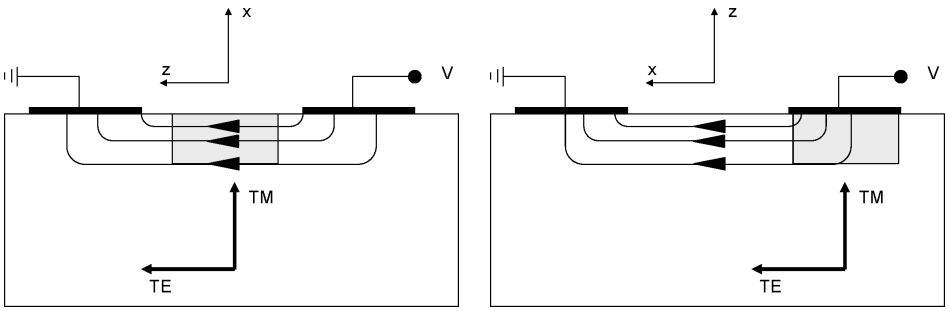


Figure 7.18: Cross section of an X-cut phase modulator(left) and Z-cut phase modulator (right)

$$\Delta\phi_{TM,X} = -\frac{2\pi}{\lambda}\frac{1}{2}n_o^3 r_{13} E_z L \tag{7.43}$$

Therefore, TE polarization should be used for most efficient phase modulation since $r_{33} > r_{13}$.

- *Z-cut $LiNbO_3$ crystal*

  For a Z-cut phase modulator the electrodes are placed such that the waveguide is below one of the two electrodes where the field is perpendicular to the Z-cut surface and the field is still along the z-direction as shown in Figure 7.18right.

  TE polarization acquires a phase change

$$\Delta\phi_{TE,Z} = -\frac{2\pi}{\lambda}\frac{1}{2}n_o^3 r_{13} E_z L \tag{7.44}$$

  while for TM polarization

$$\Delta\phi_{TM,Z} = -\frac{2\pi}{\lambda}\frac{1}{2}n_e^3 r_{33} E_z L \tag{7.45}$$

In this case TM polarization is preferred for most efficient phase modulation.

Again, an integrated interferometer is used to translate the induced phase change into an amplitude change.

7.3.5 Thermo-optical modulators

The thermo-optical effect denotes the change of refractive index with temperature. In a limited temperature range the refractive index varies linearly with temperature

$$n(T) = n(T_0) + \left.\frac{dn}{dT}\right|_0 (T - T_0) \tag{7.46}$$

The thermo-optical effect can be applied in an integrated waveguide device, by integrating a resistor on top of an arm of a Mach-Zehnder interferometer. Dissipating electrical power in this resistor

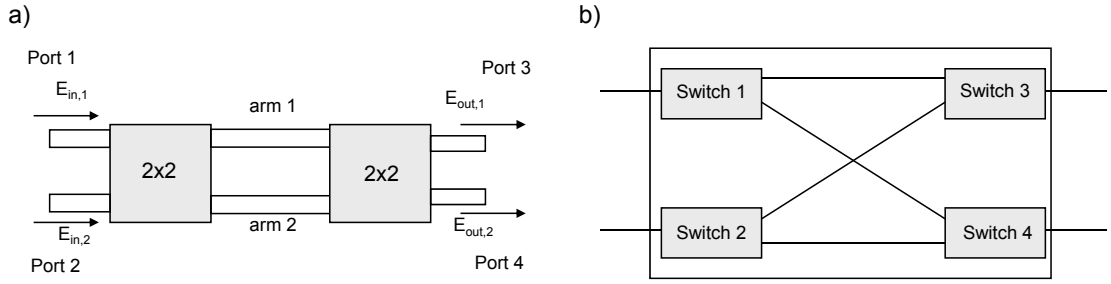


Figure 7.19: a) Mach-Zehnder interferometer as a spatial optical switch. b) Dilated switch configuration

heats an arm of the interferometer, thereby inducing a phase difference between both arms. Both arms must be sufficiently separated to allow a substantial temperature difference between both arms. The thermo-optical effect can be used in materials which do not exhibit an electro-optical effect, such as amorphous materials (and in particular silica based waveguide devices). Although this is a simple principle for amplitude modulation, the power consumption is rather high and it is slow (due to the large specific heat of the materials applied). Therefore, this type of device is rather used as a (slow) spatial switch than as a temporal switch (see also next section). Using a Mach-Zehnder interferometer in which 2x2 power combiners and splitters are used (i.e. a directional coupler or a 2x2 multimode interference coupler), as shown in figure 7.19a (compare to Figure 7.16b), light can be switched from one output waveguide to the other output waveguide by introducing a phase difference in both arms

The transmission characteristics are given by

$$\begin{bmatrix} E_{out,1} \\ E_{out,2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -j\frac{1}{\sqrt{2}} \\ -j\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} e^{-j\phi_1} & 0 \\ 0 & e^{-j\phi_2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -j\frac{1}{\sqrt{2}} \\ -j\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} E_{in,1} \\ E_{in,2} \end{bmatrix} \tag{7.47}$$

If we assume $E_{in,2} = 0$, then equation 7.47 becomes

$$\begin{aligned} \left|\frac{E_{out,1}}{E_{in,1}}\right|^2 &= \sin^2(\tfrac{\phi_2-\phi_1}{2}) \\ \left|\frac{E_{out,2}}{E_{in,1}}\right|^2 &= \cos^2(\tfrac{\phi_2-\phi_1}{2}) \end{aligned} \tag{7.48}$$

¿From this equations it is clear that if there is no phase difference between both arms (or a phase difference $(2k\pi)$, all light is switched from port 1 to port 4. If the phase difference $\Delta\phi = \phi_2 - \phi_1$ equals $(2k+1)\pi$, all power exits at port 3. A 1x2 spatial beam splitter can be achieved by replacing the 2x2 power splitter by a 1x2 3dB power splitter.

To further decrease the crosstalk level switches can be combined to form a dilated switch as shown in figure 7.19b. Here a 2x2 switch is fabricated by cross-connecting four 1x2 switches. The crosstalk level is reduced by a factor of 2 on a dB scale.

Integrated optical switches can easily be combined into larger switching fabrics. A simple switching fabric used as a 4x4 crossbar is shown in figure 7.20. This switch uses 16 2x2 switches and the interconnection between inputs and outputs is achieved by appropriately setting the states

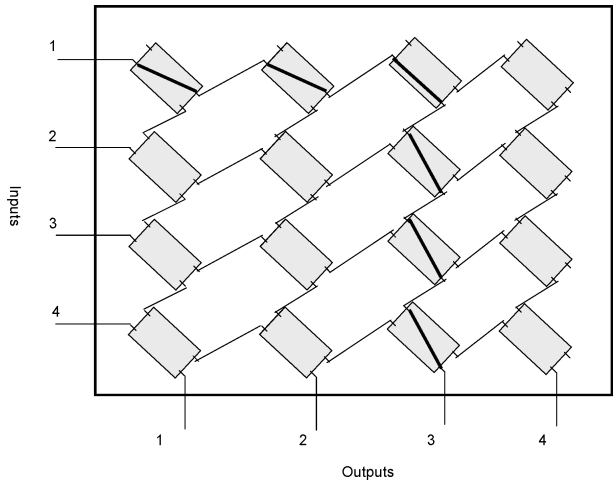


Figure 7.20: Crossbar switch

of these 2x2 switches. The settings of the 2x2 switch required to connect input 1 to output 3 are shown.

Architectures different from the simple version shown in figure 7.20 can be used to minimize the number of optical switches needed to realize a crossbar (important switching architectures were named after Spanke and Benes). Two types of switch matrices are being discerned: blocking and nonblocking. A switch is said to be nonblocking if an unused input port can be connected to any unused output port. Thus a nonblocking switch is capable of realizing every interconnection pattern between the inputs and the outputs. If some interconnection pattern cannot be realized, the switch is said to be blocking.

7.3.6 Electro-absorption modulators

Electro-absorption modulators are based on a change of the absorption spectrum of a semiconductor near its band edge by applying an electric field. Electroabsorption effects include the interband photon-assisted tunneling or Franz-Keldysh effect and exciton absorption effects in quantum well structures (quantum confined Stark effect).

The Franz-Keldysh effect can be understood from figure 7.21a in which the band structure of a direct bandgap semiconductor is depicted with and without applying an electric field. When no electric field is applied, the absorption of light with energy $h\nu$ by the material is ideally zero when $h\nu < h\nu_0$. When $h\nu > h\nu_0$ the material absorption coefficient raises like $\sqrt{h\nu - h\nu_0}$. When an electric field is applied, the valence and conduction band are tilted and light with a bandgap energy just below the bandgap of the semiconductor can tunnel through the remaining potential barrier $h\nu_0 - h\nu$ and get absorbed as well. The change in absorption spectrum due to the application of an electric field is shown in figure 7.21b.

The quantum confined Stark effect (QCSE) in quantum well devices can be understood from figure 7.21c. In figure 7.21c left, the wave function of electrons and holes inside a quantum well is shown when no electric field is present. When an electrical field is present, the wave functions shift oppositely and the effective bandgap of the quantum well shrinks due to the tilting of the bands.

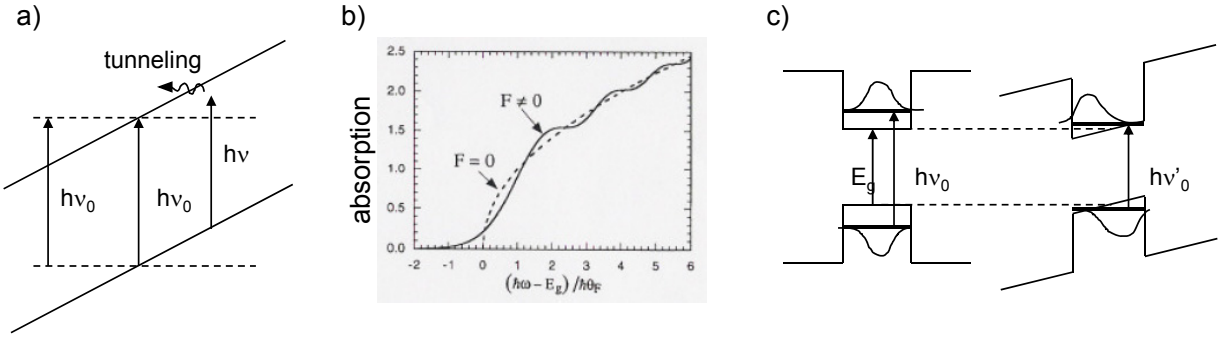


Figure 7.21: a) Franz-Keldysh effect. b) Absorption spectrum before applying field ($F = 0$) and after applying field ($F \neq 0$) taking into account FK-effect. c) Quantum confined Stark effect

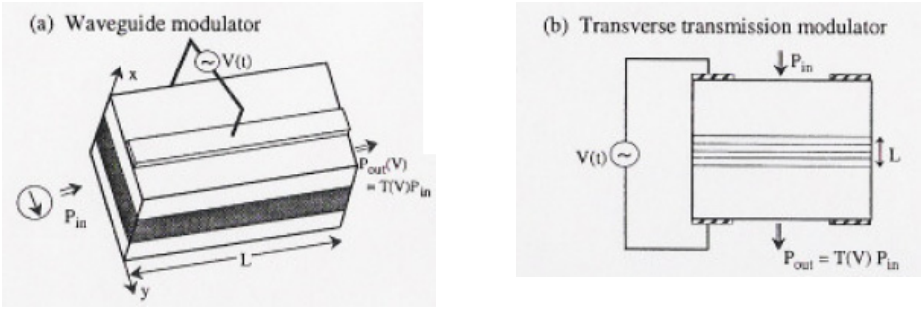


Figure 7.22: Electroabsorption modulators

In addition, this will lead to a reduction of overlap between the electron and hole wave functions. This again results in a change of absorption spectrum.

Electro-absorption modulators can be designed using a waveguide configuration or a transverse transmission configuration as shown in figure 7.22a and 7.22b.

The transmission coefficient is proportional to

$$T(V) = e^{-\alpha'(V)L} \tag{7.49}$$

For the waveguide structure, $\alpha'(V)$ is the absorption coefficient multiplied by the mode confinement factor and is L the length of the device. For the transverse transmission modulator, $\alpha'(V)$ is the average absorption coefficient of the multi-quantum-well region (MQW) and L is the total thickness of the MQW stack.

The on/off ratio is defined as

$$R_{on/off}(dB) = 10\log(\frac{T(V_{on}=0)}{T(V_{off}=V)}) = 4.343[\alpha'(V) - \alpha'(0)]L \tag{7.50}$$

By increasing the length, the on/off ratio can be made as large as desired. However, increasing the length will also increase the transmission loss of the device in the on-state, which is undesired.

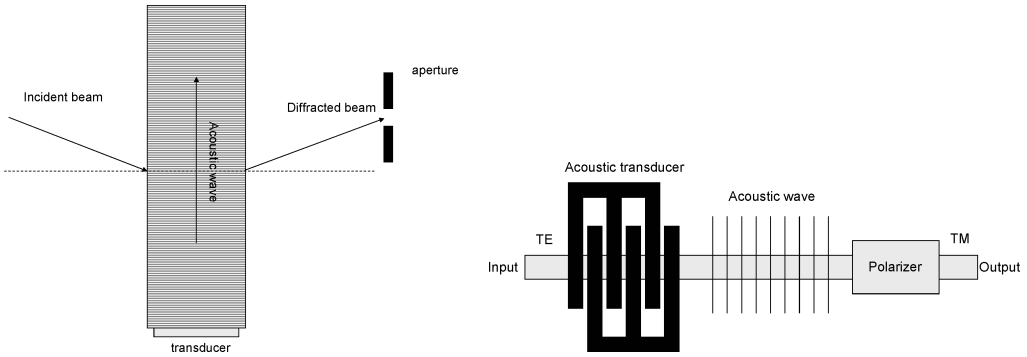


Figure 7.23: a) Macro acousto-optic modulator as temporal switch. b) Acousto-optic integrated modulator

7.3.7 Acousto-optical modulators

Macro-optic devices

As discussed in the chapter on periodic structures, an acoustic wave can induce a spatially periodic variation of the refractive index by the mechanical stress induced in a semiconductor material. A beam that is incident onto such a structure is diffracted by the grating. By varying the frequency of the acoustic wave, the angle of diffraction of an incident beam is modulated. If an aperture is placed after the acousto-optical device, the light passing through the aperture can be switched on and off as shown in figure 7.23a.

This can be used in laser welding/cutting applications, when the beam needs to be turned off when moving the sample to be welded or cut between two different positions.

Integrated waveguide devices

The use of acoustic waves to generate a bragg grating in a material can also be used in an integrated waveguide device. A simple device exploiting this effect is shown in figure 7.23b.

An acoustic transducer generates an acoustic wave along a waveguide, thereby inducing a bragg grating (period Λ). Assuming that the input field is totally TE polarized (having an effective index n_{TE} differing from the effective index n_{TM} of the TM mode), this TE mode can couple to the TM mode if the period Λ satisfies the equation

$$\frac{2\pi}{\lambda}\left|n_{TE} - n_{TM}\right| = \frac{2\pi}{\Lambda} \tag{7.51}$$

If the length of the bragg grating is well chosen, all light can couple from the TE mode to the TM mode. By placing a polarizer at the end of the device and by switching the acoustic wave on and off, light intensity can be modulated. Due to the wavelength dependence of equation 7.51, this device has a relatively small bandwidth.

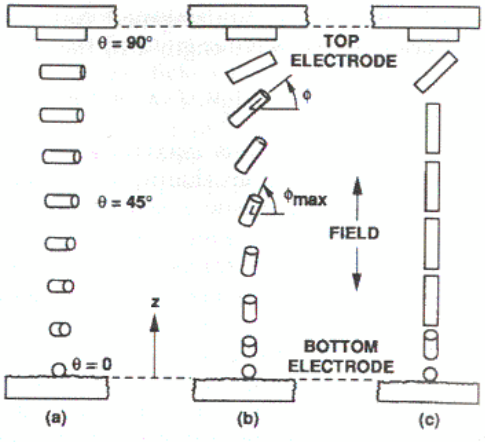


Figure 7.24: 90 degree TN cell for various degrees of electrical field strength

Applications of acousto-optic modulators

The switching speed of this type of modulator is below 1GHz, but components are cheaper than the ones based on the electro-optic effect. Furthermore they require a low drive power, provide a high extinction ratio, need relatively simple drive electronics and are insensitive to temperature variations. These components are used in telecom applications, for laser machining and inside a laser cavity for Q-switching and mode-locking.

7.3.8 Liquid crystal based modulators

Liquid crystals possess physical properties that are intermediate between conventional fluids and solids. They are fluid like, yet the arrangement of molecules within them exhibits structural order. Different types of liquid crystals exist, but only two types have reached a mature application stage: thermotropic liquid crystals and polymer dispersed liquid crystals. Thermotropic liquid crystals are liquid crystals in which the transition to the liquid crystalline state is induced thermally. Polymer-dispersed liquid crystals consist of micro droplets of liquid crystal dispersed in a polymer matrix. Both types of liquid crystals are applied in micro displays in which the temporal switching characteristics are applied. Liquid crystals show typically an uniaxially anisotropic behavior: the dielectric constant (both at low frequencies and at optical frequencies) along the axis of the molecule is different from the dielectric constant in the plane perpendicular to the axis of the molecule. The anisotropy at low frequencies allows to modify the orientation of the liquid crystal molecules by applying an electric field: liquid crystals molecules tend to align along the applied electric field (or perpendicular to the electric field). The anisotropy at optical frequencies can be used to induce a polarization rotation of the incident light or allows to scatter the light in a polymer-dispersed liquid crystal.

- *Twisted nematic liquid crystal cell*

  Nematic liquid crystals are thermotropic liquid crystals in which the molecules align approximately along a preferential direction (called the director).

  A 90 degree twisted nematic liquid crystal cell is shown in figure 7.24b.

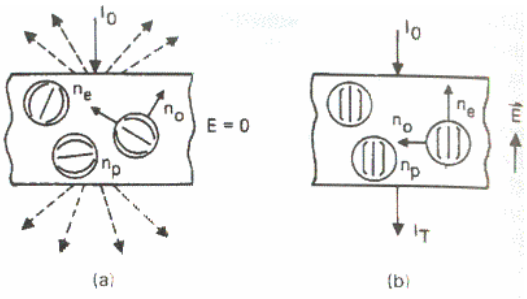


Figure 7.25: Polymer-dispersed liquid crystal cell

Consider figure 7.24a, in which no electric field is applied to the liquid crystal cell. At the plates confining the liquid crystal, the orientation of the molecule can be fixed by appropriate surface treatment of these plates. In between the plates, the molecules are twisted uniformly along the liquid crystal cell. When linearly polarized light traverses through a 90 degree twisted nematic cell, the polarization direction follows the twist of the liquid crystal directors if the Mauguin's condition is satisfied:

$$d\Delta n >> \lambda \tag{7.52}$$

with d being the thickness of the liquid crystal cell and Δn being the degree of anisotropy of the liquid crystal

$$\Delta n = \left| n_{\perp} - n_{\parallel} \right| \tag{7.53}$$

One can say that this condition implies that the twist of the molecules must vary sufficiently slow for the polarization to be able to follow this twist.

When a polarizer is put at the entrance of the liquid crystal cell along the orientation of the molecule and placing an analyzer at the exit end of the liquid crystal cell orthogonal to the polarizer high transmission is achieved.

If a sufficiently strong electric field is applied perpendicular to the liquid crystal cell (using optically transparent electrodes) the molecules are rotated and align along the vertical direction (figure 7.24c). Because there is no anisotropy in the plane perpendicular to the director, there is no polarization rotation in this configuration. Because polarizer and analyzer are placed orthogonally, no light is transmitted through the cell.

- *Polymer-dispersed liquid crystal cell*

  A simple diagram illustrating the electro-optical effect of a polymer-dispersed liquid crystal (PDLC) shutter is illustrated in figure 7.25 for the off and on states respectively.

  Suppose the polymer matrix is optically isotropic and has a refractive index n_p. The LC directors within the droplets are determined by the polymer-liquid crystal interaction at each droplet boundary. They have no preferred orientation but vary nearly randomly from droplet to droplet in the absence of an external field. The index mismatch between the droplets and the polymer matrix results in light scattering.

  When the applied electric field is sufficiently strong, the directors in the liquid crystal are oriented along the electric field. If the refractive index in the plane perpendicular to the axis

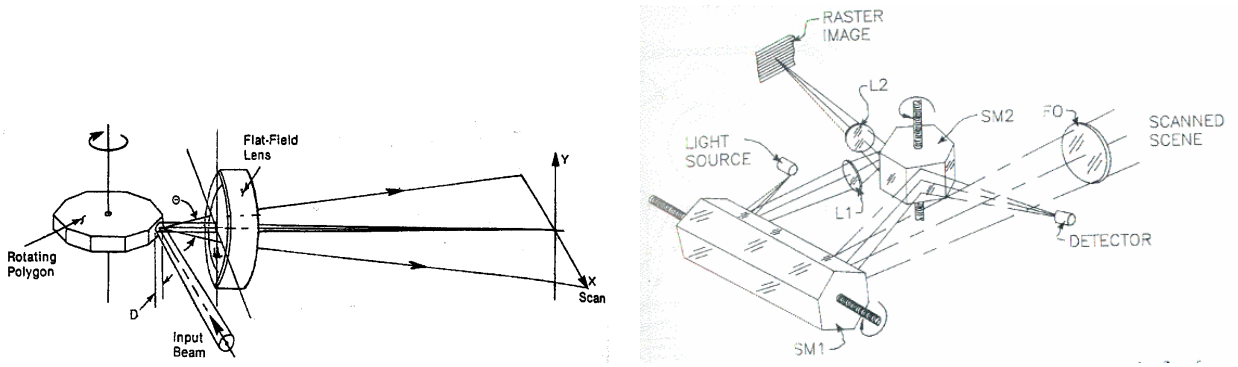


Figure 7.26: a) Input/output sensing using polygon mirrors. b) Remote sensing using polygon mirrors.

of the molecule n_o equals the refractive index of the polymer host, no light is scattered and the cell becomes transparent

When applying an electric field to a liquid crystal, the molecules will react by aligning along this electric field. The time scale for reorientation is in the millisecond range however, limiting its application to (slow) display applications or (slow) reconfiguration of photonic integrated circuits for telecom.

7.4 Beam scanning and spatial switches

7.4.1 System level description

Spatial switching involves switching of a light beam to 1 of N distinct directions (for example the switching of a light beam to 1 out of N detectors). Besides this discrete switching application, a light beam may need to be scanned along a certain path (for example in a bar code scanner or a printer). This is referred to as beam scanning. The relevant parameters determining the performance of such systems are the crosstalk level, the transmission loss and the maximum power level that can be handled.

7.4.2 Macro-optic devices

From its operation principle it is obvious that an *acousto-optical* modulator can be applied for spatial beam switching (using a discrete set of acoustic frequencies) and for beam scanning (continuously varying acoustic frequency). It can provide random beam positioning with extremely short access time, or generate linear scans at very high rates.

Beam scanning can also be achieved using *polygon mirrors*. Such scanners can be used for input/output imaging like bar-code reading (input) or printing (output). A simple system generating a linear displacement of a beam along a straight scan line is depicted in figure 7.26a. Rotating the polygon changes the angle θ of the light beam. The flat-field lens converts this change in angle into a linear displacement along a straight scan line.

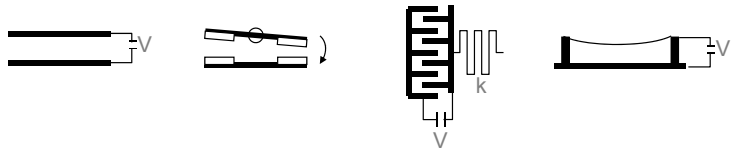


Figure 7.27: Electrostatic actuation of mems structures

Polygon mirrors can also be used in remote sensing applications (for example in medics or space applications). A simple system illustrating the use of polygon mirrors is shown in figure 7.26b. The scene is imaged using the focussing optics FO and the mirrors SM1 and SM2 onto a detector. A raster image of the scene is formed by imaging the light source using lenses L1 and L2. The display image and the scanned scene are synchronized using the same pair of mirrors. The light source is modulated by the output of the detector. Mirror SM1 moves the beam in vertical direction at a slow rate while mirror SM2 generates the high-speed horizontal scan.

7.4.3 Fiber-optic devices

Besides bulk devices, also fiber-optic switching systems exist. An example of such a (mechanical) system uses a fiber-optic directional coupler. Bending or stretching the fiber in the coupling region changes the coupling ratio of the coupler and can be used to switch light from an input port to different output ports. These type of switches have low insertion losses, low polarization sensitivity, low crosstalk and are relatively inexpensive. Switching speeds are on the order of a few milliseconds and the number of ports is fairly small.

Another family of devices consists of optical fibers connected via micro-optics (lenses, beamsplitters) that are moved electromechanically, or fibers themselves that are physically moved. Especially when single mode fibers are used (9 micrometer core size), the alignment criteria are very stringent and makes the mechanical design non-trivial. For multimode fibers (125 micrometer core size) the alignment problem is reduced.

7.4.4 Micro-electro-mechanical systems

Micro-electro mechanical systems (MEMS) are miniature mechanical devices typically fabricated using silicon substrates. A variety of structural elements can be realised such as beams, membranes, tips, hinged plates and integrated micro-lenses. These structures are deflected from one position to another using a variety of electronic actuation techniques, such as electromagnetic, electrostatic or piezo-electric methods. Of these methods, electrostatic deflection is particularly power efficient. Electrostatic actuation can be employed in parallel plate structures (acting as a capacitor), through torsional actuation, using comb actuators or using full membranes (see Figure 7.27).

MEMS are fabricated employing standard microelectronic processes such as deposition, lithography and etching. In a final step, the moving structures are released using a selective etch process. E.g. using hydrofluoric (HF) acid, SiO_2 is removed but supporting structures and moving structures made out of (poly-)silicon are unaffected. This is illustrated in Figure 7.28.

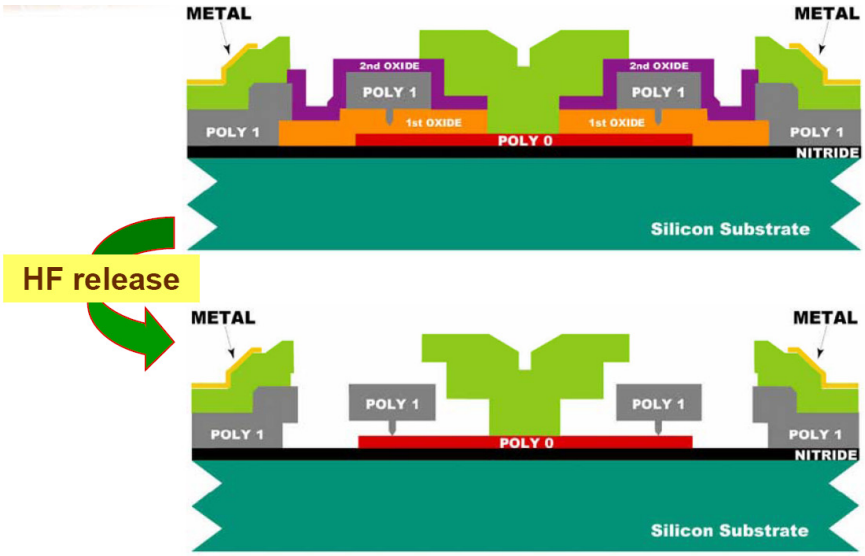


Figure 7.28: Basic principle of MEMS-fabrication using standard microelectronic processes and HF release.

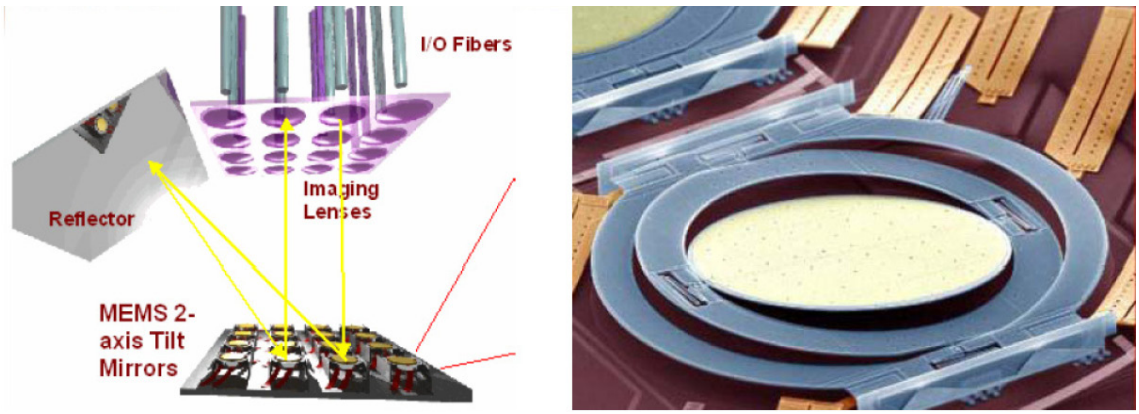


Figure 7.29: Analog MEMS-mirror supported by a double set of hinges and its application in an optical crossconnect.

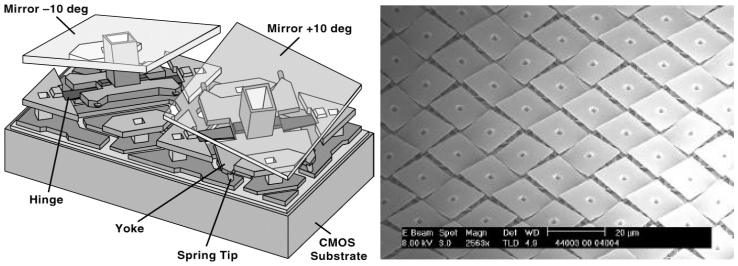


Figure 7.30: Schematic and image of fabricated MEMS mirrors.

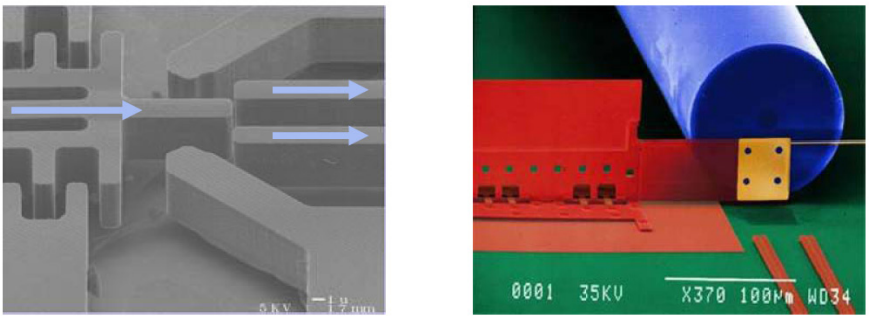


Figure 7.31: a) 1x2 MEMS-based optical waveguide switch. b) MEMS-based on-off switch.

In the context of optical switches, the acronym MEMS usually refers to miniature movable mirrors fabricated in silicon, with dimensions ranging from a few hundred micrometers to a few millimeters. A single silicon wafer can yield a large number of mirrors, which means that these mirrors can be manufactured and packaged as arrays. An example of an analog beam steering mirror is shown in figure 7.29. The mirror is connected through flexures to an inner frame, which is in turn connected to an outer frame through another set of flexures. The flexures allow the mirror to be rotated freely on two distinct axes. By combining a 2D-array of such mirrors with an optical fiber array, an array of microlenses and a mirror, a full optical crossconnect can be built, as also shown in Figure 7.29.

The control of these mirrors is not a trivial matter, with fairly sophisticated servo control mechanisms to deflect the mirrors to the correct position and hold them there. Therefore, in cases where full analog control is not needed, digital micro-mirrors are used, which can only switch between an on-state and an off-state. The most well known example of these, the *DMD* or digital micro-mirror device (Texas instruments) is shown in figure 7.30. This type of mirror is applied in projector systems for image formation (DMD: digital micromirror device). Micro-mirrors provide higher brightness and a higher contrast compared to liquid crystal devices.

Figure 7.31 shows two further examples of optical MEMS switches. In the left picture, a comb actuator is used to switch an input waveguide to one of two output waveguides. In the right picture, a shutter can be moved up and down and block light transmission from the input fiber (on top) to the output fiber (bottom, not shown in picture for clarity).

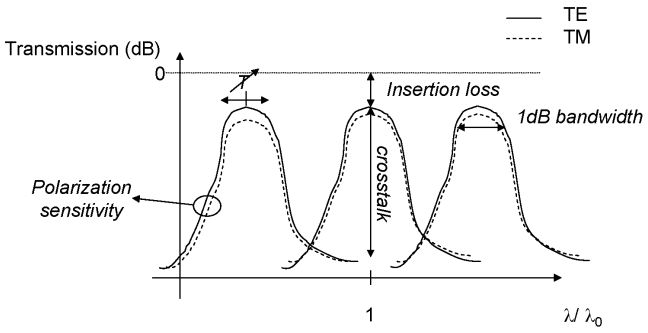


Figure 7.32: Filter characteristics

7.5 Wavelength dependent devices

7.5.1 System level description

Wavelength selection is a core operation in optical transmission systems based on wavelength division multiplexing (see the subsection on applications of wavelength selective devices). Key characteristics for use in telecom systems are:

- *Insertion loss* Good optical filters should have low insertion losses. The insertion loss is the input-to-output loss of the filter.
- *Polarization insensitivity* Although sometimes hard to achieve, good optical filters should be polarization insensitive.
- *Temperature insensitivity* In order to get sufficient temperature insensitivity (the central wavelength of the filter shifts due to temperature changes) sometimes a temperature controller is needed.
- *Passband flatness* In order to get good transmission characteristics (certainly after cascading different optical filters) the passband of the optical filter should be very flat. This is measured by the 1dB bandwidth of the optical filter.
- *Crosstalk* Crosstalk levels, related to the suppression of wavelengths outside the passband should be sufficiently low

Wavelength dependent devices are also used for other applications like mode selection inside a gas laser cavity and for AR and HR coatings.

7.5.2 Macro-optic devices

Thin film dielectric filters

Stacks of thin dielectric films are probably the most commonly used optical filters in optics. Bulk filters with application dependent tailored transmission characteristics can be designed and fabri-

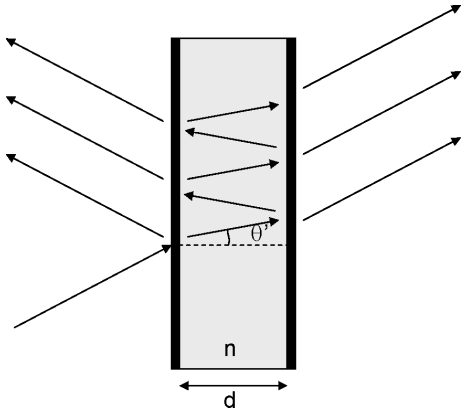


Figure 7.33: Fabry-Perot etalon

cated. A common application is the use of thin dielectric layers onto sunglasses, where UV light needs to be blocked as much as possible and the transmission in the visible range needs to be decreased (but with a flat transmission spectrum). They are also used as AR-coating on regular glasses and optical elements, as reflective coatings on windows, and much more ... In combination with micro-optic elements they can also be used to realise filters for telecom systems with very high performance.

In integrated optics, stacks of thin dielectric films are used for high reflectivity (HR) and anti-reflectivity (AR) coatings on laser diodes and semiconductor optical amplifiers.

Dedicated software tools exist for the design of thin film dielectric filters, using matrix procedures as outlined in chapter 3. When optimizing for a given function, it is important to take into account the wavelength dependent characteristics of the film.

Fabry-Perot etalon

As discussed before, a Fabry-Perot etalon consists of a cavity formed between two high-reflective mirrors as shown in figure 7.33. The cavity can be made out of any transparent material (mostly glass), or can be filled with air or gas (in this case the mirrors are separated by spacers). The transmission spectrum of the filter is comb-like and is given by (see chapter 3):

$$T(\lambda) = \frac{1}{1 + \frac{4R}{(1-R)^2}\sin^2\phi} \tag{7.54}$$

with $\phi = \frac{2\pi}{\lambda} nd\cos(\theta')$ (assuming lossless mirrors).

This type of device can be used inside a gas laser cavity to select a single laser wavelength (necessary due to the inhomogeneous broadening of the emission due to large spread in velocity of the gas molecules, referred to as Doppler broadening).

A Fabry-Perot etalon can also be used to measure the refractive index of gases. The mirrors are separated by spacers and gas flows in between the mirrors forming the medium of the cavity. As

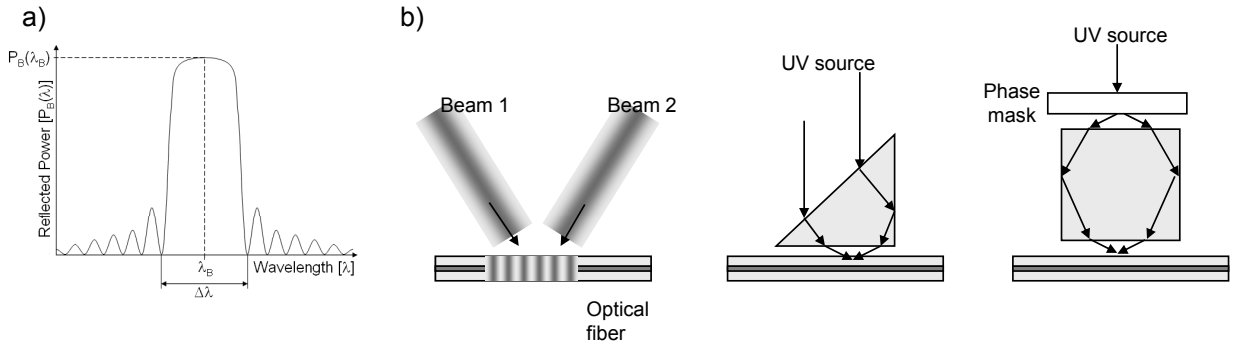


Figure 7.34: a) Reflection characteristic of fiber bragg grating. b) Fabrication of fiber bragg gratings

the transmission wavelengths directly depend on the refractive index of the cavity, this refractive index can be measured.

Gratings

As already discussed before, gratings can be used in spectrometer designs as wavelength selective elements (Czerny-Turner monochromator). These will be discussed in more detail in the next chapter.

7.5.3 Fiber Bragg Gratings

A commonly used fiber-optic wavelength selective filter is a *fiber bragg grating*. When a germanium-doped optical fiber is exposed to intense UV light a permanent refractive index change is induced. This effect can be used to define a grating inside an optical fiber (along the propagation axis). From chapter 6 (periodic structures) we know that the peak reflection wavelength of such a grating is defined by the Bragg Wavelength (see also Figure 7.34)a:

$$\lambda_b = 2n\Lambda \tag{7.55}$$

with Λ the period of the grating and n the average refractive index. We also know that the width of the transmission is mainly determined by the index contrast:

$$\frac{\Delta\lambda}{\lambda_b} = \frac{2\Delta n}{\pi n_0} \tag{7.56}$$

Since the induced index contrast is typically very low, very sharp filters can be realised. The peak reflection on the other hand is mainly determined by the grating length and is approximately given by:

$$R_{max} = tanh^2 \frac{2\Delta nL}{\lambda_b} \tag{7.57}$$

To suppress the first sidelobes in the reflection spectrum, the grating can be *apodized*: the refractive index contrast is graded to reach zero at both ends of the grating. Both Gaussion and raised cosine apodization are commonly used.

The grating can also be *chirped*. In that case the period is linearly varied along the grating. This broadens the spectrum but also induces dispersion in the optical signal: signals with different wavelength experience different delay. This effect can be used to compensate for fiber induced dispersion.

Fabrication

Different setups are used to illuminate an optical fiber. The holographic setup (with or without a prism) and the phasemask setup are shown in figure 7.34b. Interference between the different beams induces a periodically modulated light intensity. The prism setup is more stable than the pure holographic setup because both beams are perturbed in the same way by a possible vibration. The induced grating period is determined by the angle of incidence and the illumination wavelength.

The phase mask consists of a surface relief grating that diffracts an incident beam into the -1,0,1 diffraction order. The phase mask is sensitive to the illumination wavelength and a different mask is required for a different Bragg wavelength. However, it allows for more complex grating patterns than the simple holographic setups.

Application

Fiber bragg gratings are used both in telecom systems (as wavelength selective devices) and in sensing applications. In telecom systems they are mainly used as *notch* filters. In combination with optical circulators, they can also be used in optical multiplexers and add-drop filters. To realize tunable optical filters, the Bragg wavelength can be changed by applying strain using a piezoelectric transducer.

The most common sensing application is the measurement of stress induced in a mechanical structure (i.e. a bridge or a pressure vessel) because mechanical stretching of the optical fiber results in an enlargement of the bragg period, hence resulting in a shift of the resonant wavelength. By injecting broadband light into the fiber and monitoring the transmission (or reflection) spectrum a measure of the strain in the mechanical structure can be obtained. Also temperature variations can be measured due to the thermal expansion of the fiber when the temperature is increased.

7.5.4 Integrated waveguide devices

Many complex integrated components are designed to be used in WDM (wavelength division multiplexing) networks. In WDM networks many wavelengths are used in one fiber to transport the data. This way the bandwidth of the existing network can be increased. Another advantage is that the routing can be completely optical, i.e. without the optical-electrical-optical conversion. In a WDM network different complex components are needed: sources (tunable lasers or multiwavelength lasers), multiplexers and demultiplexers, add-drop filters, detectors, cross connects. Most

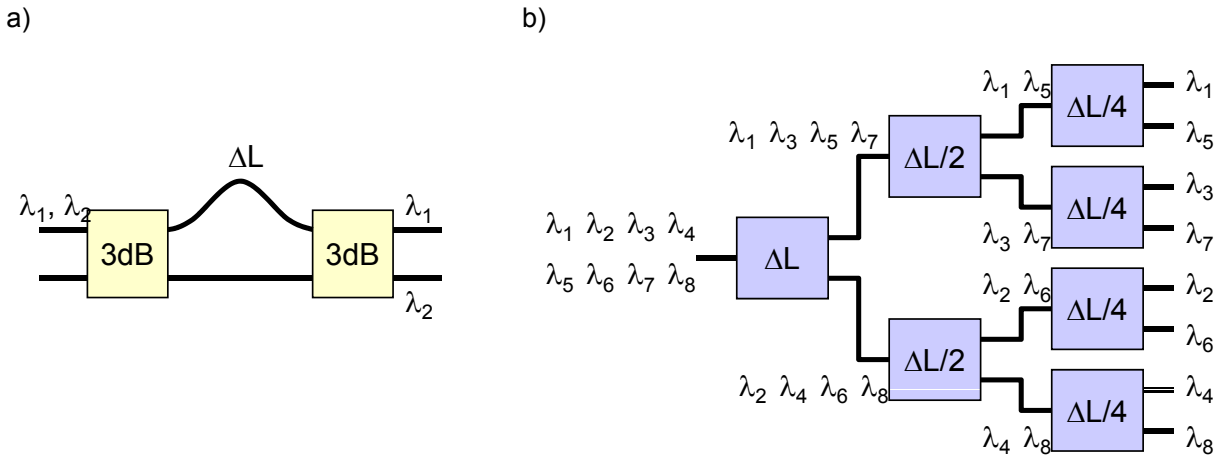


Figure 7.35: a) Basic Mach Zehnder interferometer with unequal path lengths. b) MZI based demultiplexer

of these require wavelength selective operation. Therefore, there is a need for integrated optical filters. The most commonly used integrated filters are discussed below. Next some applications are given.

Mach-Zehnder based integrated wavelength demultiplexer

A Mach-Zehnder interferometer (MZI) with a pathlength difference ΔL in one of both arms (Figure 7.35a) has a wavelength dependent response given by equation 7.48 with the phase difference

$$\phi_2 - \phi_1 = \beta \Delta L \text{ , with } \beta = 2\pi n_{eff}/\lambda. \tag{7.58}$$

For the crossport, we have maximal transmission if this phase difference is equal to $2m\pi$ or $\Delta L = m\lambda/n_{eff}$. Hence, the MZI has a sinusoidal response with a separation

$$\Delta\lambda = \lambda^2/n_g \Delta L \tag{7.59}$$

between two successive peaks. This separation is also called the free spectral range (FSR) of the device. Note that we introduced the group index $n_g = n_{eff} - \lambda \frac{dn_{eff}}{d\lambda}$ to take into account the wavelength dependence (dispersion) of the effective refractive index.

By cascading a series of MZI's with successively decreasing path length difference (and hence increasing FSR), we can realise a multichannel wavelength demultiplexer as shown in Figure 7.35b. This demultiplexer can achieve high resolution. Fabrication tolerances are very strict however.

Arrayed waveguide grating router

An arrayed waveguide grating router (AWG) can be considered as a generalised MZI: the input and output 2x2 couplers are replaced by NxN star couplers and the two arm waveguides are re-

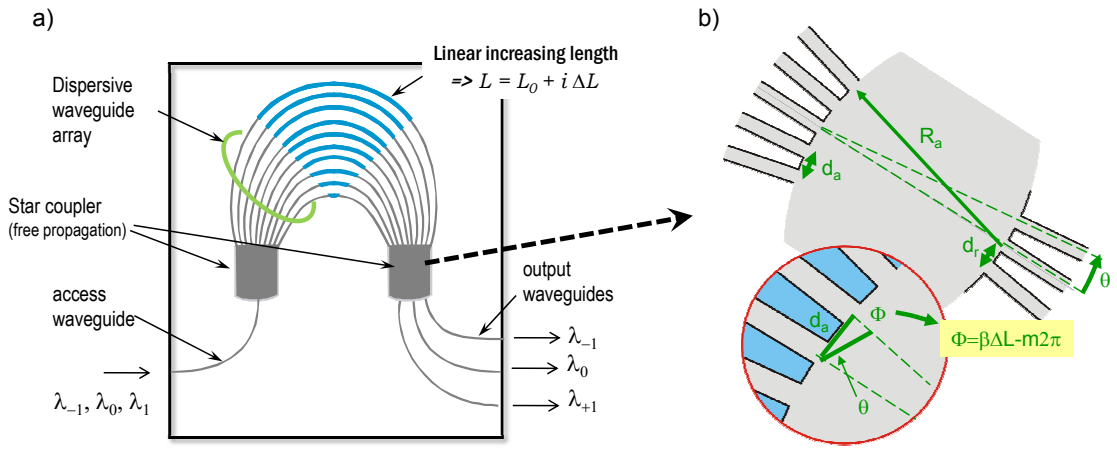


Figure 7.36: a) Global layout of AWG. b) Output starcoupler of AWG.

placed by an array of waveguides with linearly increasing path length (i.e. the path length difference between two successive arms is constant and given by ΔL). Figure 7.36a shows a schematic representation of the AWG (also sometimes called Phased Array Demultiplexer or phasar). The operation principle is remarkably simple: light entering the first starcoupler through the entrance waveguide is diffraced and coupled into the central waveguide array. This array is designed in such a way that (for the central wavelength of the component) the optical length difference between two neighbouring arms equals a multiple of the central wavelength. The consequence is that the field distribution at the entrance aperture is reproduced at the exit aperture and for this wavelength the light is focused in the middle of the image plane. An output waveguide placed at this location will then pick up the light and guide it further. When the wavelength of the incident light slightly differs from the central wavelength, the wave front at the exit aperture will be tilted over an angle θ due to the linear phase shift in the waveguide array. The focus will be shifted out of the center. If we now place exit waveguides at the right position in the image plane, wavelength channels can be demultiplexed into the various exit waveguides. When we change the propagation direction the AWG can be used as a multiplexer. The biggest advantages of the AWG are its low losses and simple fabrication procedure. At this moment the AWG is the most used integrated multiplexer. Below we will derive the most important properties of its operation.

Comparable to the MZI, also the response of the AWG is periodic. We will have maximal transmission from the central input waveguide to the central output waveguide for a wavelength fulfilling the condition $m\lambda_0/n_{eff} = \Delta L$. However, also the wavelength λ_N with $(m-1)\lambda_N/n_{eff} = \Delta L$ will exhibit maximal transmission. Hence the period or FSR is given by $FSR = \lambda_0/N$, leading to a similar expression as that given in equation 7.59.

The second important parameter is the dispersion $D = ds/d\nu$ in the focal plane. The dispersion determines how fast the focal point shifts as function of a changing signal wavelenth (or frequency). Referring to Figure 7.36b the dispersion is given by:

$$D = \frac{ds}{d\nu} = R_a \frac{d\theta}{d\nu} \tag{7.60}$$

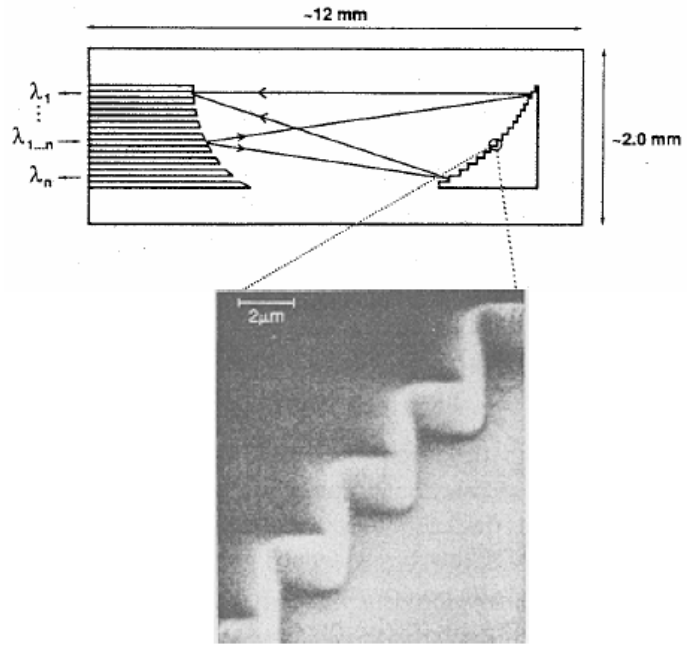


Figure 7.37: Grating demultiplexer

$d\theta$ may be determined from the phase difference between the signals coming from two successive waveguides:

$$d\theta = \frac{\lambda}{2\pi} d\Phi = \frac{\lambda}{2\pi n_s} d\left(\beta \Delta L - m2\pi\right) = \frac{\lambda}{2\pi n_s} \frac{2\pi}{c} n_{eff} \Delta L d\nu \tag{7.61}$$

with n_s the effective index of the slab area and n_{eff} the effective index of the waveguide. This then finally leads to:

$$D = \frac{ds}{d\nu} = R_a \frac{1}{\nu_c} \frac{\Delta L}{d_a} \tag{7.62}$$

where we assumed $n_{eff} \approx n_s$. From the dispersion we can calculate the channel spacing as:

$$\Delta\nu = \frac{d_r}{D} \tag{7.63}$$

with d_r the separation between two successive waveguides in the focal plane (see Figure 7.36b).

Planar concave grating demultiplexer

The PCG or planar concave grating demultiplexer uses a focussing planar integrated diffraction grating as a dispersive element. The operation principle is identical to that of free space gratings. Compared to AWG's PCG's require more complicated fabrication processes and therefore they are

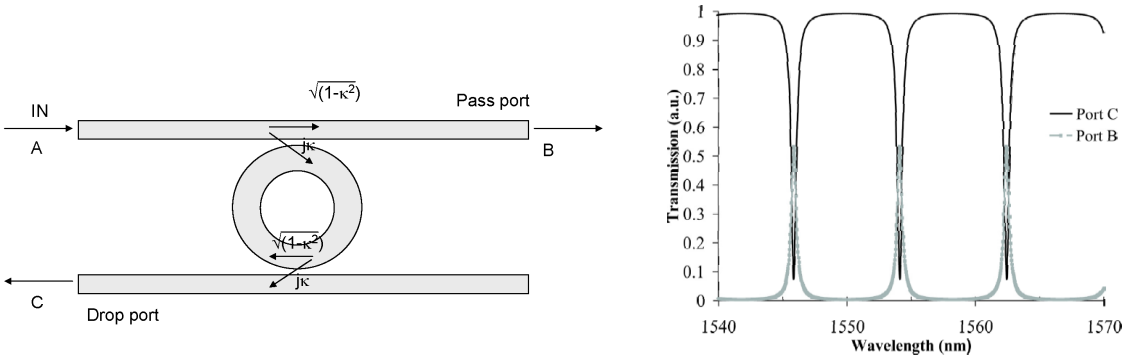


Figure 7.38: a) Basic ring resonator layout. b) Transmission characteristic (lossless operation)

currently less popular. Typically also the loss is somewhat higher and the resolution is lower but demultiplexers with a high number of channels can easily be realised.

Ring resonator

As shown in Figure 7.38, the integrated ring resonator consists of a ring shaped resonator (with radius R) coupled to two bus waveguides, which serve as input and output channels. The coupling between the ring and the bus waveguide is described by the coupling constant κ and is typically small. The operation of the integrated ring resonator is similar to that of a Fabry-Perot resonator, whereby the coupling coefficient κ of the ring resonator is equivalent to the single mirror transmission t of the Fabry-Perot resonator. The transfer T_{AC} from input to output port is sharply peaked with resonances determined by ($m\lambda/n_{eff} = 2\pi R$). The width of the resonances is determined by the finesse, which in turn is determined by the coupling coefficient κ:

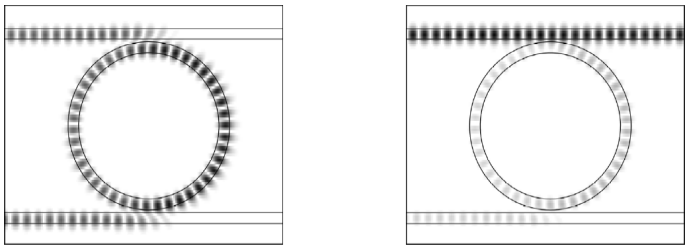
$$\mathfrak{F} = \frac{\pi\sqrt{1-\kappa^2}}{\kappa^2} \tag{7.64}$$

This assumes a lossless resonator. If the resonator exhibits loss (waveguide scattering loss, bending loss ...), κ has to be replaced in this expression by $f\kappa$ with $20\log(f)$ the roundtrip loss in dB (loss for a single roundtrip).

The transmission to port B (the pass port) is the opposite of that to port C (the drop port) and shows a notched transfer. Figure 7.39a and Figure 7.39b show field profiles of a resonator in and out of resonance.

Ring resonators are used both in fiber based versions and in integrated waveguide versions. The latter have become very popular recently with the advent of the high-index contrast material systems based on silicon or silicon nitride.

By cascading several ring resonators with slightly increasing radius allong a common bus waveguide, multi-channel demultiplexers can be realised. When organized in a 2-dimensional array, full crossconnects can be realised.



(a) In resonance (b) Out of resonance

Figure 7.39: Electrical field in ring resonator at resonance (a) and out of resonance (b).

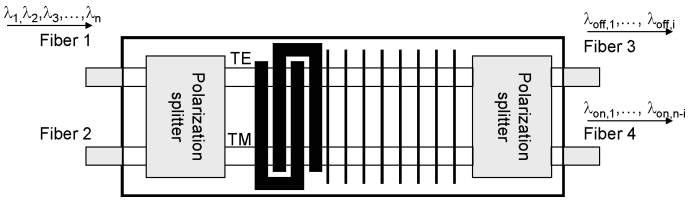


Figure 7.40: Acousto-optical cross connect filter.

7.5.5 Acousto-optical tunable filter (AOTF)

As discussed in the section on acousto-optic modulators, the optical bandwidth of these devices is small. Therefore, this type of device can be used as a wavelength selective filter or cross connect switch. An example of a 2x2 polarization independent cross connect switch based on the acousto-optic effect is shown in figure 7.40.

If light is incident from fiber 1, TE and TM polarization state are separated using a polarization splitter. If there is no acoustic wave or if the acoustic frequency doesn't match the incident wavelength no polarization conversion occurs and the light is coupled in fiber 3. If the acoustic wave frequency matches the bragg condition, TE and TM polarization are converted and light is coupled into Fiber 4.

The major advantage of the acousto-optical cross connect switch is that the individual wavelength channels can be individually addressed by superposing different acoustic frequencies. The device is rapidly tunable (by changing the acoustic frequency) over a wide wavelength range. As there are no moving parts, this is a very reliable component.

7.5.6 Tunable devices

In many applications, the possibility for tuning the wavelength of optical filters is critical. The two main mechanisms employed for tuning the transmission wavelength of an optical filter are changing its refractive index or mechanically adjusting the filter. That can easily be seen from the fact that in many of the structures we discussed above, the transmission wavelength is determined by a relation of the form:

$$\frac{2\pi}{\lambda_m} nL = \phi_m \tag{7.65}$$

Hence, either changing the refractive index n or the length L will shift the resonance wavelength λ_m. From this condition it is straightforward to calculate the relative wavelength shift as a consequence of such a change. We find:

$$\frac{\Delta\lambda_m}{\lambda_m} = \frac{\Delta n}{n} \tag{7.66}$$

for a change in the refractive index, which could be induced by any of the mechanisms described earlier in this chapter (thermo-optic effect, electro-optic effect, carrier injection, photoeleastic effect). Almost all known filters fullfill this equation (MZI, ring resonator, FP, bragg reflector ...). Since the practically attainable relative change in refractive index dn/n is typically in the order of 0.1% to 1%[2], the attainable wavelength tuning range is also rather limited however. One exception is formed by grating coupled filters (the acousto optical tunable filter is an example of these).

For a change in the length of the device we find:

$$\frac{\Delta\lambda_m}{\lambda_m} = \frac{\Delta L}{L} \tag{7.67}$$

Hence, for a given displacement ΔL, the effect is stronger if the cavity length is smaller. This effect has been succesfully employed to make widely tunable lasers by combining a VCSEL (vertical cavity surface emitting laser) which has a very short cavity with a MEMS based external mirror. Other examples of mechanically tuning the resonance wavelength of optical filters include the stretching of fiber bragg gratings, rotating thin film filters, changing the length of FP resonators or rotating an external diffraction grating.

Widely tunable devices

Even when one has to rely on small effects such as the electro-optic or thermo-optic effect, it is possible to realize widely tunable devices by cascading two optical comb filters (like a ring resonator or Fabry-Perot type filter) with slightly different free spectral ranges. This effect is called the Vernier effect and is illustrated in figure 7.41. When the comb spectrum of one of the filters is slightly shifted (by a small effect as the thermo-optic or electro-optic effect) the new resonant wavelength (where the two combs overlap) can shift much more due to the small difference in free spectral range.

7.5.7 Applications

Telecommunications

As already mentioned before, wavelength selection technologies are core operations in optical transmission systems. This is due to the fact that most optical transmission systems use wavelength division multiplexing (WDM) to send different data signals over the same optical fiber. The typical WDM scheme is shown in figure 7.42, showing the different components.

[2]One of the few exceptions is the change between the ordinary and extraordinary refractive index in liquid crystals.

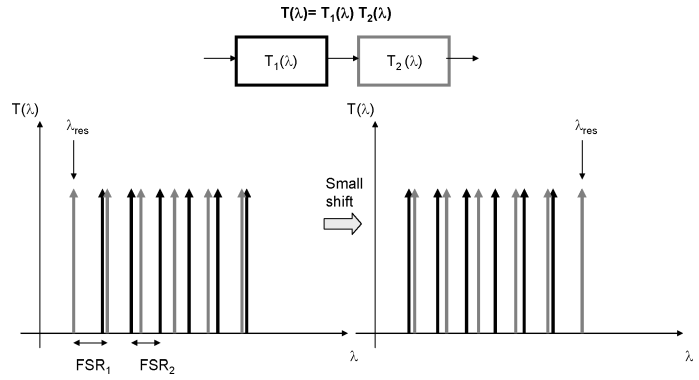


Figure 7.41: Vernier effect

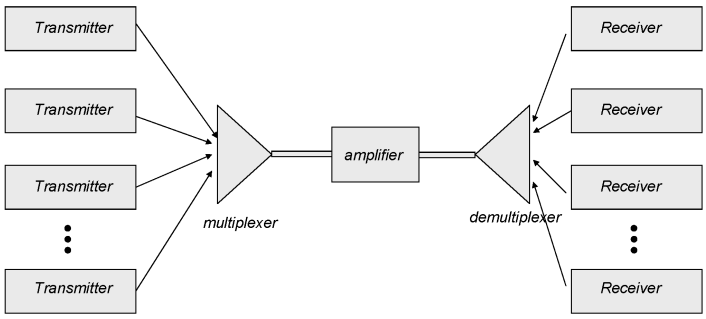


Figure 7.42: WDM transmission system

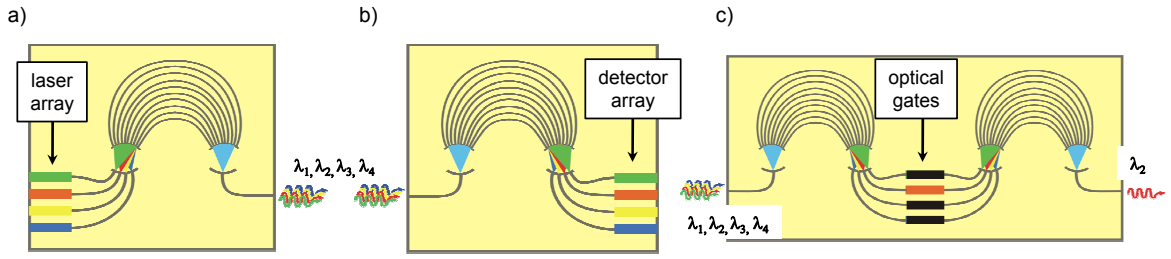


Figure 7.43: a) Multi-wavelength transmitter. b) Multiwavelength receiver. c) Multiwavelength wavelength selector.

Different transmitters modulate a laser beam at different wavelengths. These different wavelength signals are combined into one optical fiber using an *optical multiplexer*. Although losses of optical fibers can be very low, for long haul transmission intermediate amplifiers are needed to compensate the losses of the optical fibers. The gain spectra of these amplifiers are however not perfectly flat. Therefore this needs to be compensated for by using *a gain flattening filter* (for example by using a thin dielectric film filter). At the receiver the optical signal is again separated into its different wavelength components using a *demultiplexer*. These two applications (multiplexing/demultiplexing and gain flattening) are the main applications of wavelength selective devices in transmission systems. They are also used to feed the 850nm pump signal to an EDFA (and leaving the signal wavelengths unchanged).

Many of the required devices can be implemented as a photonic integrated circuit (PIC). By combining wavelength selective components with active elements such as detectors, lasers and optical gates very complex PICs can be realised. Figure 7.43a and b respectively show a multi-wavelength transmitter and receiver. Figure 7.43c shows a wavelength channel selector. The first AWG is used for demultiplexing the multi-wavelength signal. Then optical gates are used to block or pass the light. Finally, the signals are multiplexed again in the second multiplexer. The optical gates could either be semiconductor optical amplifiers (SOA), which show a high suppression ratio when unpumped, or thermo-optic MZI switches, which can be used in combination with glass based waveguide platforms.

Another typical module is a receiver/transmitter chip which contains both a 1.3 μm and a 1.55 μm photodiode. A $1.3\mu m/1.55\mu m$ wavelength demultiplexer routes the optical signal. This chip can be used in a bidirectional data link over fiber in the $1.55\mu m$ and $1.3\mu m$ telecommunication window, for fiber-to-the-home applications but also for high-speed ethernet applications.

An add-drop filter allows to drop or to add one or more of the wavelength channels. Figure 7.44 shows such a device.

Sensing

Spectrometers are used to measure all sorts of spectra. One application is the use as a gas sensing apparatus. Because gases have very distinct absorption peaks at certain wavelengths, we can illuminate a gas flow by a broadband optical beam (of which the spectrum is known) and look at the spectrum of the beam after it passed the gas flow. Depending on the position of the absorption

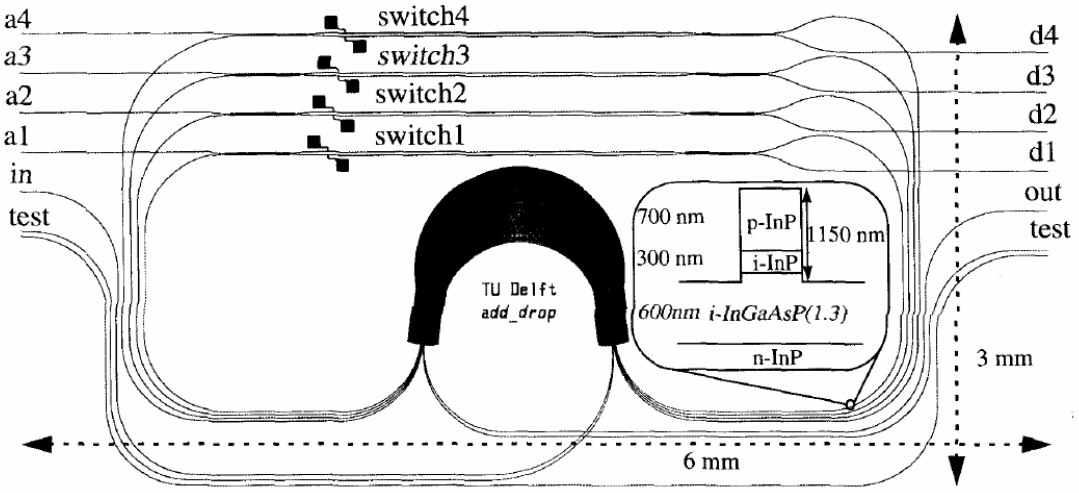


Figure 7.44: Add-drop filter.

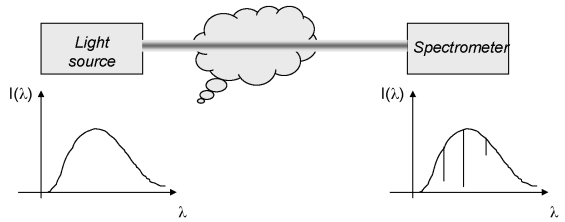


Figure 7.45: Gas sensing using spectrometry.

peaks we can identify different gases in the gas flow as shown in figure 7.45. Alternatively one can use a tunable laser source in combination with a broadband detector.

Chapter 8

Photonic Measurement Systems

Contents

8.1 Microscopy

Microscopes are optical instruments, which use visible light and a system of lenses to magnify images of small samples. The simplest magnifier is essentially a positive lens e.g. used to read small print, in which case it is often called a reading glass, or to assist the eye in examining small detail in a real object. It is often a simple convex lens but may be a doublet or a triplet, thereby providing for higher quality images. The magnification of small objects accomplished by the simple magnifier is increased further by using a compound microscope. In its simplest form, this instrument consists of two positive lenses, an objective lens of small focal length that faces the object and a magnifier functioning as an eyepiece. The eyepiece looks at the real image formed by the objective.

8.1.1 Some basic concepts

Optical aberrations

Lens errors or aberrations in optical microscopy are caused by artifacts arising from the interaction of light with glass lenses. There are two primary causes of aberration: (i) geometrical or spherical aberrations are related to the spherical nature of the lens and approximations used to obtain the Gaussian lens equation; and (ii) chromatic aberrations that arise from variations in the refractive indices of the wide range of frequencies found in visible light. In general, the effects of optical

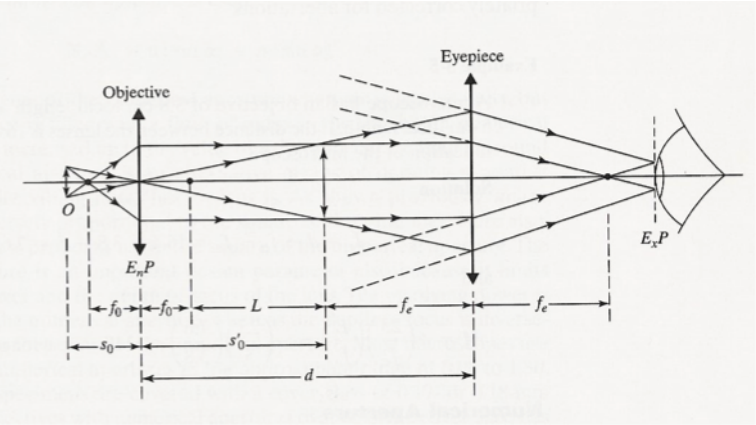


Figure 8.1: Image formation in a compound microscope)

aberrations are to induce faults in the features of an image being observed through a microscope. Todays microscope objectives do almost not suffer from these aberration effects, although careful attention must still be paid when performing quantitative high-magnification microscopy.

Spherical Aberration occurs when light waves passing through the periphery of a lens are not brought into identical focus with those passing closer to the center. Waves passing near the center of the lens are refracted only slightly, whereas waves passing near the periphery are refracted to a greater degree resulting in the production of different focal points along the optical axis. This is one of the most serious resolution artifacts because the image of the specimen is spread out rather than being in sharp focus. Spherical aberrations are very important in terms of the resolution of the lens because they affect the coincident imaging of points along the optical axis and degrade the performance of the lens, which will seriously affect specimen sharpness and clarity. These lens defects can be reduced by limiting the outer edges of the lens from exposure to light using diaphragms and also by utilizing aspherical lens surfaces within the system. The latter are however more difficult to manufacture and hence more expensive.

Chromatic Aberration is the result of the fact that white light is composed of numerous wavelengths. When white light passes through a convex lens, the component wavelengths are refracted according to their frequency. Blue light is refracted to the greatest extent followed by green and red light, a phenomenon commonly referred to as dispersion. The inability of the lens to bring all of the colors into common focus results in a slightly different image size and focal point for each predominant wavelength group. This leads to color fringes surrounding the image. By combining different types of glass (each type has a different dispersion of refractive index), the chromatic aberration can be reduced (achromatic lenses).

Resolution

The resolution of the microscope determines the smallest feature that can just be resolved, or the smallest distance that can be determined between features. Although vital for microscopy, resolution is not easily defined in general terms. The attainable resolution depends strongly on the signal-to-noise ratio of the imaging system, on the imaging mode, and on a a priori knowledge of the specimen. An often-used resolution criterion is the Rayleigh criterion, which was first de-

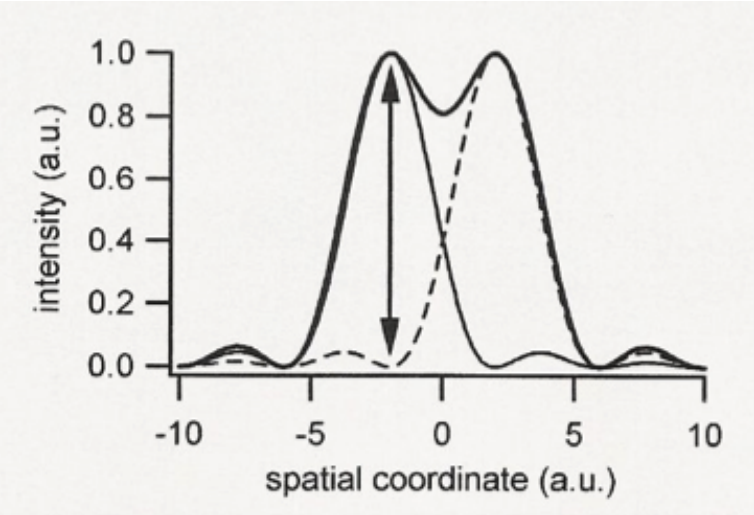


Figure 8.2: The Rayleigh criterion. Two components of equal intensity are resolved when the maximum of one coincides with the first minimum of the other. This results in an effective drop of the intensity between the two maxima of $\tilde{2}0\%$

veloped in connection with the resolving power of prism and grating spectrometers. It states that two components of equal intensity should be considered to be just resolved when the principal intensity maximum of one coincides with the first intensity minimum of the other (see Figure 8.2).

From chapter 4 (Fourier Optics) we know that the first zero off the Airy pattern (= image of a point source) lies at a distance

$$r = 1.22\frac{\lambda d_i}{D} \tag{8.1}$$

from the optical axis, with D the diameter of the lens and d_i the image location (with $d_i \simeq f$ in many optical systems). With the image side numerical aperture NA_i defined as

$$NA_i = nsin(\theta_i) = \frac{D}{2d_i} \tag{8.2}$$

with n the refractive index of the medium and θ_i the opening angle of the lens, we find for the resolution (in the image plane):

$$x_{i,min} = 0.61\frac{\lambda}{NA_i} \tag{8.3}$$

and in a similar way for the object side resolution:

$$x_{o,min} = 0.61\frac{\lambda}{NA_o} \tag{8.4}$$

Another important aspect to resolution is the axial resolving power of an objective, which is measured parallel to the optical axis and is most often referred to as depth of field. Just as in classical

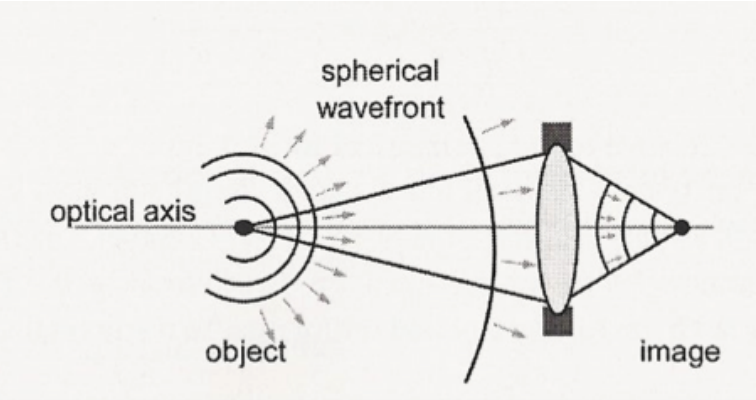


Figure 8.3: Aberration-free imaging of a point: an ideal lens converts a diverging into a converging spherical wave

photography, depth of field is determined by the distance from the nearest object plane in focus to that of the farthest plane also simultaneously in focus. In microscopy the depth of field is very limited and usually measured in terms of microns. The term depth of focus, which refers to image space, is often used interchangeably with depth of field, which refers to object space.

Imaging performance characteristics

The point spread function (PSF)

Consider the imaging of an infinitely small point by an ideal lens, i.e. a lens without optical aberrations (see Figure 8.3). The function of the lens is to transform the diverging spherical wavefront radiated by the object - the point source - to a converging spherical wavefront forming the image. Whereas the spherical waves radiated by the object extend over the full 4 radians of a sphere, the converging spherical wave is clipped by the limited size of the lens, given by its aperture. It is this clipping of the wavefront and the diffraction at the edges of the aperture that limits the quality of the image point. It causes a blurring of the image of the point source. The 3D light distribution in the image is called the point spread function (PSF) of the lens. For an ideal, aberration-free lens, the PSF is determined only by the wavelength of the light, the numerical aperture (NA) of the lens, and by diffraction. In this case, the imaging is said to be diffraction-limited. The diffraction pattern of a converging spherical wave from a circular aperture can be calculated from diffraction theory (see chapter on Fourier Optics - Figure 8.3) and forms the Airy pattern already discussed above.

8.1.2 Illumination

Proper illumination of the specimen is crucial in achieving high-quality images in microscopy. An advanced procedure for microscope illumination was first introduced in 1893 by August Köhler, as a method of providing optimum specimen illumination. All manufacturers of modern laboratory microscopes recommend this technique because it produces specimen illumination that is uniformly bright and free from glare. The collector lens and other optical components built into the base will project an enlarged and focused image of the lamp filament onto the plane of the aper-

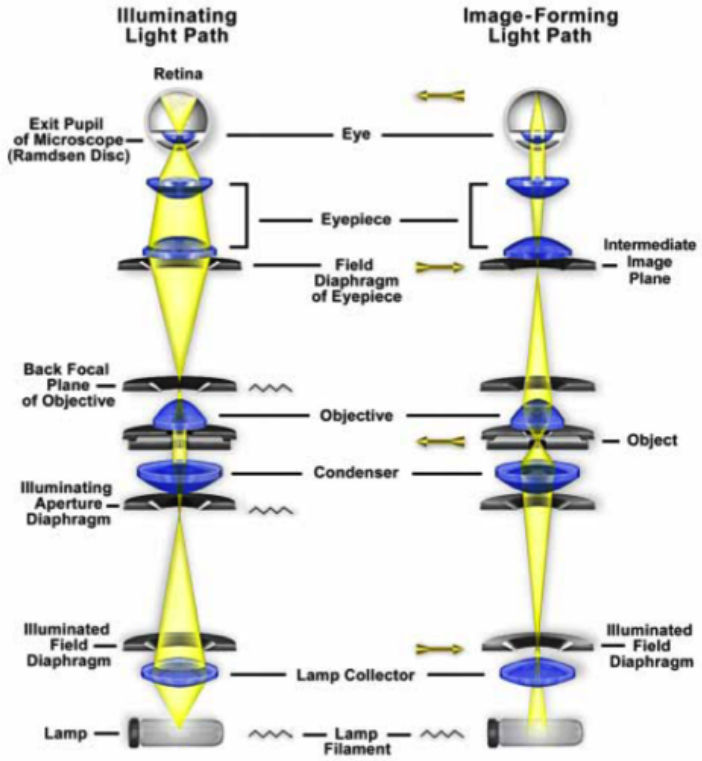


Figure 8.4: Light paths in Köhler illumination. The illuminating ray paths are illustrated on the left side and the image-forming ray paths on the right. Light emitted from the lamp passes through a collector lens and then through the field diaphragm. The aperture diaphragm in the condenser determines the size and shape of the illumination cone on the specimen plane. After passing through the specimen, light is focused at the back focal plane of the objective and then proceeds to and is magnified by the ocular before passing into the eye.

ture diaphragm of a properly positioned substage condenser. Closing or opening the condenser diaphragm controls the angle of the light rays emerging from the condenser and reaching the specimen from all azimuths. Because the light source is not focused at the level of the specimen, illumination at specimen level is essentially grainless and extended, and does not suffer deterioration from dust and imperfections on the glass surfaces of the condenser. The opening size of the condenser aperture diaphragm, along with the aperture of the objective, determines the realized numerical aperture of the microscope system. As the condenser diaphragm is opened, the working numerical aperture of the microscope increases, resulting in greater light transmittance and resolving power. Parallel light rays that pass through and illuminate the specimen are brought to focus at the rear focal plane of the objective, where the image of the variable condenser aperture diaphragm and the light source are observed in focus simultaneously.

Light pathways illustrated in Figure 8.4 are schematically drawn to represent separate paths taken by the specimen illuminating light rays and the image forming light rays. By definition, an object that is in focus at one plane is also in focus at other conjugate planes of that light path. In each light pathway (both image forming and illumination), there are four separate planes that together make up a conjugate plane set. Conjugate planes in the path of the illuminating light rays in Köhler illumination (left-hand diagram) include the lamp filament, condenser aperture diaphragm (at

the front focal plane of the condenser), the rear focal plane of the objective, and the eyepoint of the eyepiece. The eyepoint is located approximately one-half inch (one centimeter) above the top lens of the eyepiece, at the point where the observer places the front of the eye during observation. Likewise, the conjugate planes in the image-forming light path in Köhler illumination (right-hand diagram) include the field diaphragm, the focused specimen, the intermediate image plane (i.e., the plane of the fixed diaphragm of the eyepiece), and the retina of the eye or the film plane of the camera.

8.1.3 Bright Field Microscopy (transmitted or reflected)

Bright field microscopy is the simplest of all the light microscopy techniques and can either use white light illumination in transmission or reflection. Sample illumination is via transmitted white light, i.e. illuminated from below and observed from above. Limitations include low contrast of most samples and low apparent resolution due to the blur of out of focus material. The simplicity of the technique and the minimal sample preparation required are significant advantages. Reflected light microscopy is often referred to as incident light or epi-illumination microscopy. Because light is unable to pass through these specimens, it must be directed onto the surface and eventually returned to the microscope objective by either specular or diffused reflection. As mentioned above, such illumination is most often referred to as episcopic illumination, epi-illumination, or vertical illumination (essentially originating from above), in contrast to diascopic (transmitted) illumination that passes through a specimen. In reflected light microscopy, absorption and diffraction of the incident light rays by the specimen often lead to readily discernible variations in the image, from black through various shades of gray, or color if the specimen is colored. Some specimens however show so little difference in intensity and/or color that their feature details are extremely difficult to discern and distinguish in bright field reflected light microscopy. Such objects require special treatment or contrast methods that will be described in the next section.

8.1.4 Dark Field Microscopy

Dark field microscopy refers to an illumination technique whereby the background illumination is suppressed and only the scattered field is picked up. In some cases this results in much better contrast. The sample is illuminated under an angle and only diffracted rays are picked up through a suitable combination of field stops and apertures in the condenser plane and in the back focal plane. See Figure 8.5

8.1.5 Phase contrast Microscopy

In microscopy, the generation of adequate and meaningful contrast is as important as providing the needed resolution. Many specimens are transparent and differ from their surroundings only in slight differences of refractive index, reflectance or birefringence. Most objects that are black or show clear color when reasonable thick become transparent or colorless when their thickness is reduced to a few tens of a micrometer (absorption varies exponentially with thickness). Additionally the specimen is illuminated at large cone angles to maximize resolution under the microscope thus reducing the shadows and other contrast cues that aid detection of objects in macroscopic

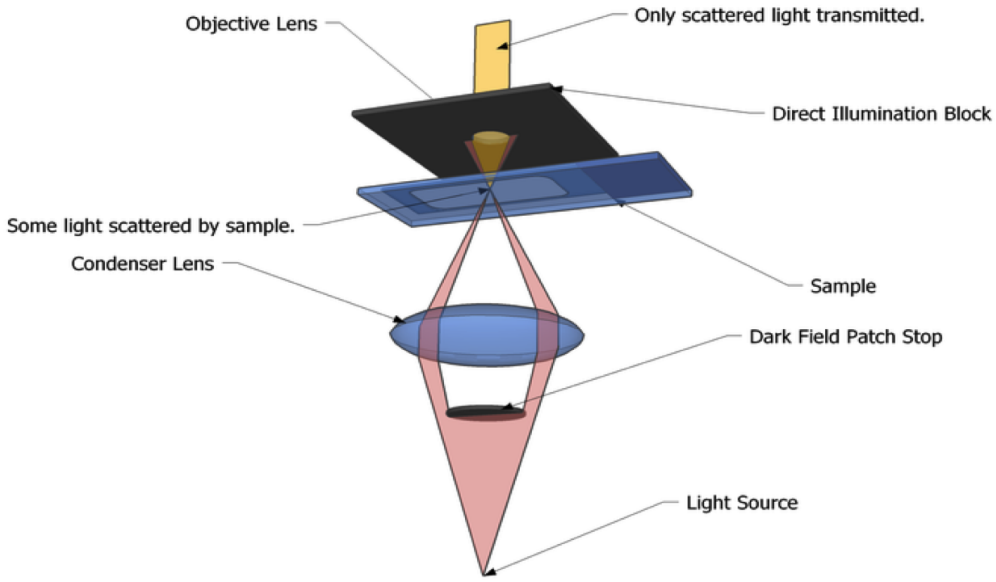


Figure 8.5: Principle of Dark Field Microscopy (picture from `http://en.wikipedia.org/wiki/Dark_field_microscopy`). A field stop in the condensor aperture plane ensures the sample is illuminated under an angle. An aperture in the objective back focal plane ensures that the light directly coming from the light source is blocked and only scattered rays are transmitted

imaging. Furthermore, contrast is reduced at high spatial frequency because of an inherent fall-off of the contrast transfer function. Many modes of contrast generation are used in microscopy partly to overcome these limitations and partly to measure, or detect, selected optical characteristics of the sample under test. Thus, in addition to simply raising contrast to make an object visible, the introduction of contrast that reflects a specific or chemical characteristic of the specimen may provide particularly important information.

Using Phase Plate

The first approach to visualize phase objects is through the introduction of a phase plate in the fourier plane of the image. The transmission of a pure phase object can be described as:

$$\tau(x,y) = e^{-j\delta(x,y)} \tag{8.5}$$

with

$$\delta(x,y) = \frac{2\pi}{\lambda}\left[n(x,y) - 1\right] d(x,y) \tag{8.6}$$

and n the refractive index of the object, d the thickness of the object. The intensity in the image plane is then proportional to $|\tau(x,y)|^2$ for which we find:

$$|\tau(x,y)|^2 = 1 \tag{8.7}$$

and hence the image is invisible to the eye. If the thickness of the samples is small (i.e. $\delta(x,y) << 1$, then

$$\tau(x,y) = 1 - j\delta(x,y) \tag{8.8}$$

In the fourier plane of the lens, the first term (basically a plane wave) will result in a non-diffracted spot in the center. The second term contains higher order spatial frequencies and will results in diffracted spots in the fourier plane. If we now introduce a phase plate in the focal plane, which shifts the central spot (the lowest spatial frequencies) by a fraction $\pi/2$ (a quarter wave plate), as shown in Figure 8.6a, the resulting transmission in the image plane will now be given by:

$$\tau(x,y) = j\left[-1 - \delta(\frac{x'}{m}, \frac{y'}{m})\right] \tag{8.9}$$

with an associated intensity profile

$$|\tau(x,y)|^2 = 1 + 2\delta\left(\frac{x'}{m}, \frac{y'}{m}\right) \tag{8.10}$$

which is visible to the eye.

Nomarski Interference Microscope

Figure 8.6b shows the schematic view of a Nomarski microscope. A birefringent plate with its optical axis rotated by 45 degrees with respect to the input polarizer splits the beam in two polarizations, each focussing at a different location. A similar setup at the image side recombines both beams in the image plate where they will interfere. Two modes of operation can be discerned:

- If the focal point of both beams is far away from each other, then the object will interfere with a uniform background, and the global phase variations will be made visible
- If both focal points are located close together (i.e. within the object), then the gradient of the phase pattern will be visualized through the interference pattern.

8.1.6 Confocal Microscopy

Confocal fluorescence microscopy is a microscopic technique that provides true three-dimensional optical resolution. In microscopy, 3D resolution is generally realized by designing the instrument so that it is primarily sensitive to a specimens response coming from an in-focus plane, or by subsequently removing the contributions from out-of-focus planes. Several techniques have been

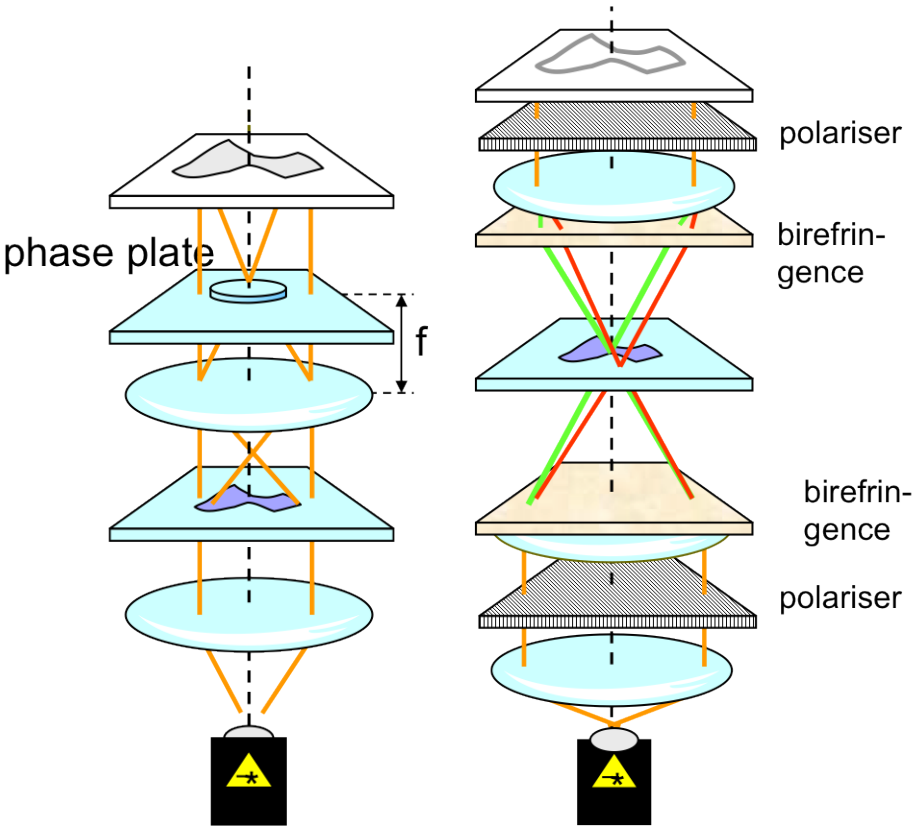


Figure 8.6: a) Schematic view of microscope with phase plate for enhancing phase contrast. b) Nomarski interference microscope

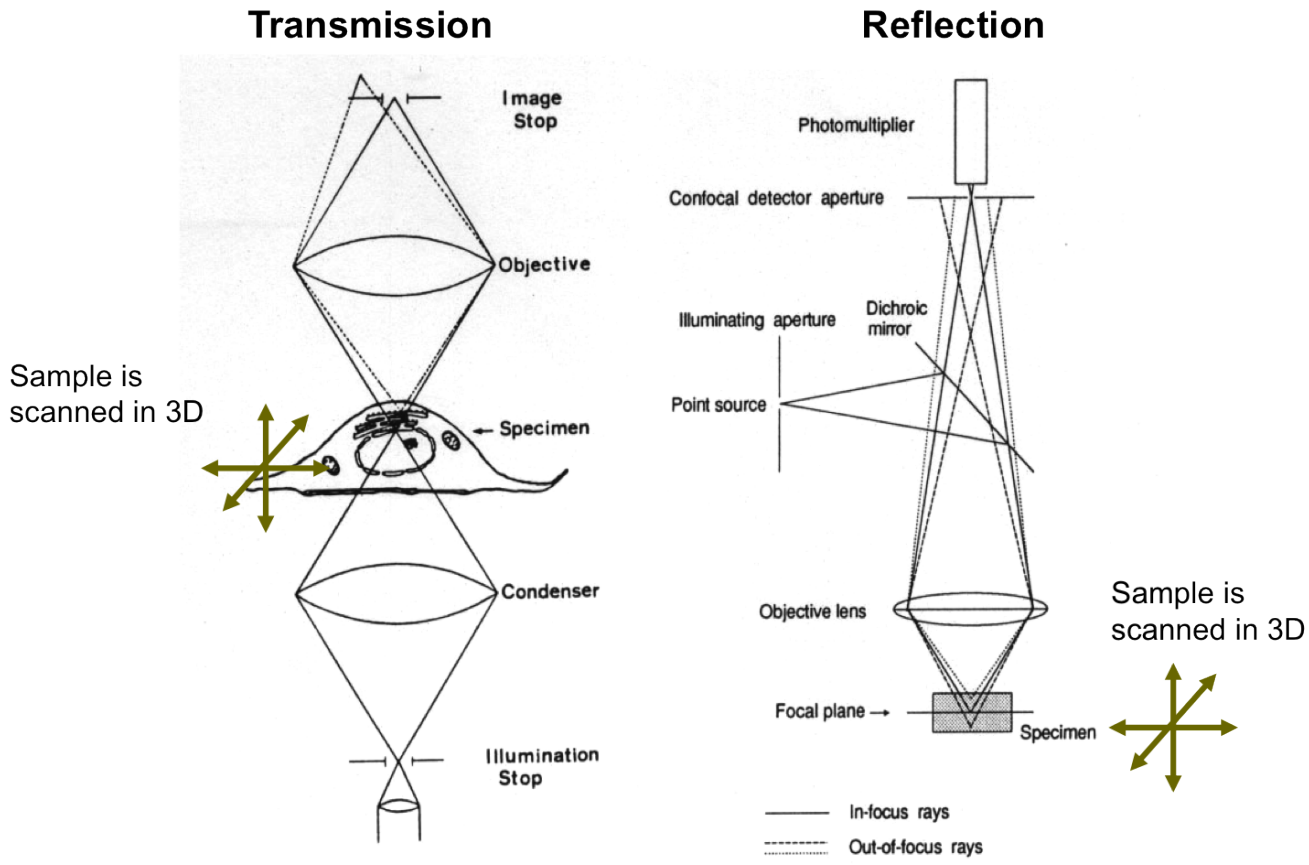


Figure 8.7: The principle of confocal fluorescence microscopy. Light coming from out-of-focus planes is largely blocked by a pinhole in front of the detector

developed to achieve this. For instance, 3D deconvolution uses both in- and out-of-focus information from a stack of images, taken at various focal planes, to reconstruct the 3D image. Another example is two- and three-photon absorption microscopy, where a non-linear interaction with the specimen is used to confine the specimens response to the focal plane only. In confocal microscopy, true 3D resolution is accomplished by actively suppressing any signal coming from out-of-focus planes. This is achieved by using a pinhole in front of the detector as shown in Figure 8.7. Light originating from an in-focus plane is imaged by the microscope objective such that it freely passes the pinhole, whereas light coming from out-of-focus planes is largely blocked by the pinhole.

The confocal microscope is a point-by-point imaging device. To acquire an image, the focal plane must be raster scanned in both lateral directions. In addition, multiple sections taken at various focal planes need to be acquired to enable reconstruction of the full 3D image. In other words, to obtain a 3D image, the specimen needs to be probed point by point in three dimensions, where the 3D size of the probe is determined by the PSF of the confocal system. There are two modes of scanning: object or specimen scanning and illumination or laser scanning. In object scanning, the specimen is scanned with respect to the focal point of the microscope objective. The big advantage of this method is that the optical arrangement is stationary and the beam follows the optical axis throughout the microscope, minimizing optical aberrations. In laser scanning, the laser spot is scanned with respect to the specimen, thereby avoiding rapid movement of the specimen and

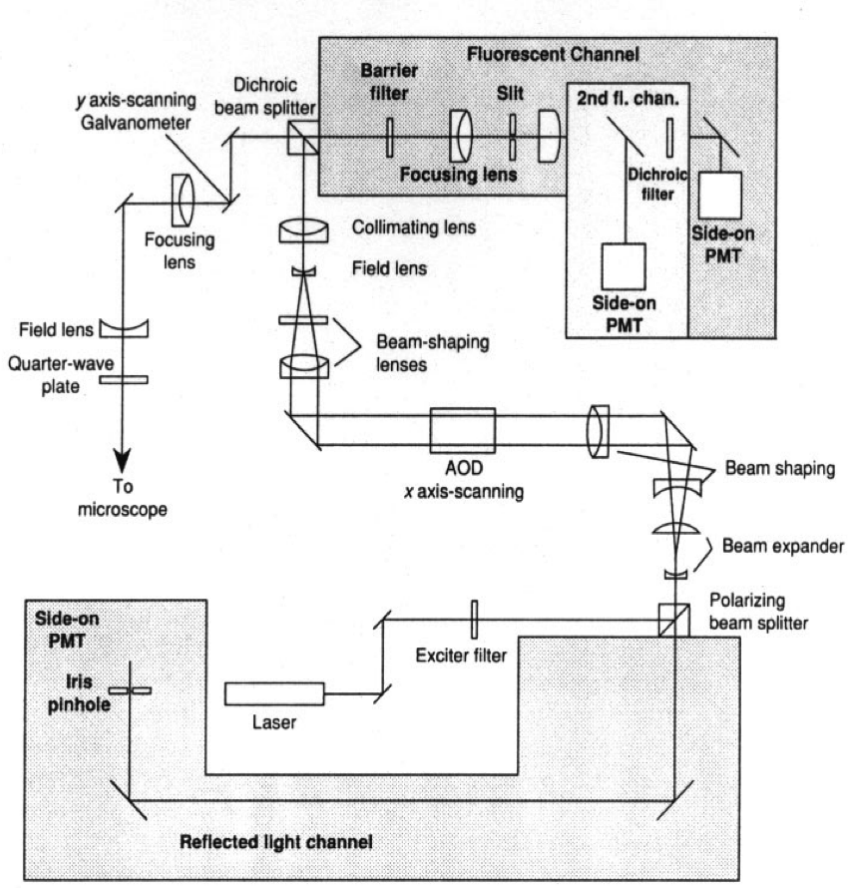


Figure 8.8: Details of practical confocal microscope. PMT = Photo-multiplier tube, AOD = Acousto-optic deflector. The specimen is located below the arrow ("to microscope").

increasing the possible scan speed. In most cases, two scan mirrors are used for fast beam steering in both lateral directions. This involves off-axis use of the microscope objective, placing high demands on its optical quality.

Confocal microscopy is often extended with spectroscopy, which can give information on the chemical structure of the specimen, or with fluorescence. In the latter case, the laser will excite fluorescent molecules. Only excited points will fluoresce, so crosstalk from other points is avoided. Figure 8.8 shows the schematic of a real-life laser scanning confocal microscope. The laser beam is scanned over the specimen through the AOD (x-axis) and the scanning galvanometer (y-axis). The dichroic beam splitter splits the incident laser beam and the excited fluorescence coming from the sample, which is then analyzed in the "fluorescent channel" [1]. The reflected laser beam on the other hand is separated from the incident beam through the combination of the quarter wave plate and the polarizing beam splitter. This beam can be used to form an image of the object under study.

[1]Note that "dichroic" can have two separate meanings. Here it refers to a color filter. In polarizing optics it refers to a material, which has different absorption for different polarizations

8.1.7 Super-resolution microscopy

As explained in section 8.1.1, traditional optical microscopes are ultimately limited by diffraction, resulting in the resolution limit defined in equation 8.1

For over a century it was believed that one would never observe anything smaller than roughly half the wavelength of light. This means for example in bio-imaging that most cell organelles and individual proteins would remain out of reach.

Abbe's principle still holds, however, in recent decades several techniques have been developed that are able to circumvent the the limit, either by making use of intrinsic properties of the dyes or fluorophores used in the imaging process (STED and STORM), or by making use of clever illumination techniques and computational methods (SIM). The importance of these techniques is illustrated by the 2014 Nobel Prize in Chemistry, which was awarded to three pioneers in the field: Eric Betzig (co-inventor of STORM), Stefan W. Hell (inventor of STED) and William E. Moerner (first optical detection and spectroscopy of a single molecule in condensed phases).

STED (STimulated Emission Depletion)

STED was the first technology developed to transcend the diffraction limit in the far-field. To achieve this, two laser pulses are focused on the sample (see figure 8.9 c). The first pulse is used to excite the fluorophores in a focal region determined by the conventional diffraction limit. The second pulse, which is slightly red-shifted with respect to the first, is able to bring the excited fluorophores back into a high vibrational energy state of the electronic ground state (see energy diagram in 8.9 a). By using a $\lambda/2$-phaseplate, the STED beam is given a zero-intensity minimum in the center of the focal region (optical vortex). By changing the intensity of this STED pulse, the region around the vortex which is not depleted can in principle be made arbitrary small. After depletion, only the emission of fluorophores in this small region will be detected.

STORM/PALM

In STORM (STochastic Optical Reconstruction Microscopy, also called PALM - Photo Activated Localization Microscopy), Abbe's diffraction limit is breached by combining two principles: "super-localization" of a sparse set of point sources and the use of so-called "switched fluorophores".

Super-localization refers to the ability to localize a single point-source with much greater precision than is prescribed by the Abbe diffraction limit. When photons hit the microscope detector, their spatial probability density will be determined by the point-spread function (PSF) of the imaging system, which typically can be approximated by a Gaussian with width Δ_{Abbe}. By fitting this known PSF to the measured intensity distribution the position of a point source can be estimated with standard error $\Delta = \Delta_{Abbe}/\sqrt{N}$, with N the number of detected photons. The resolution is thus improved by a factor $1/\sqrt{N}$. One can however only use this technique for sparse distributions of point sources, i.e. when it can be safely assumed that the PSFs of the sources will not overlap, which is generally not the case.

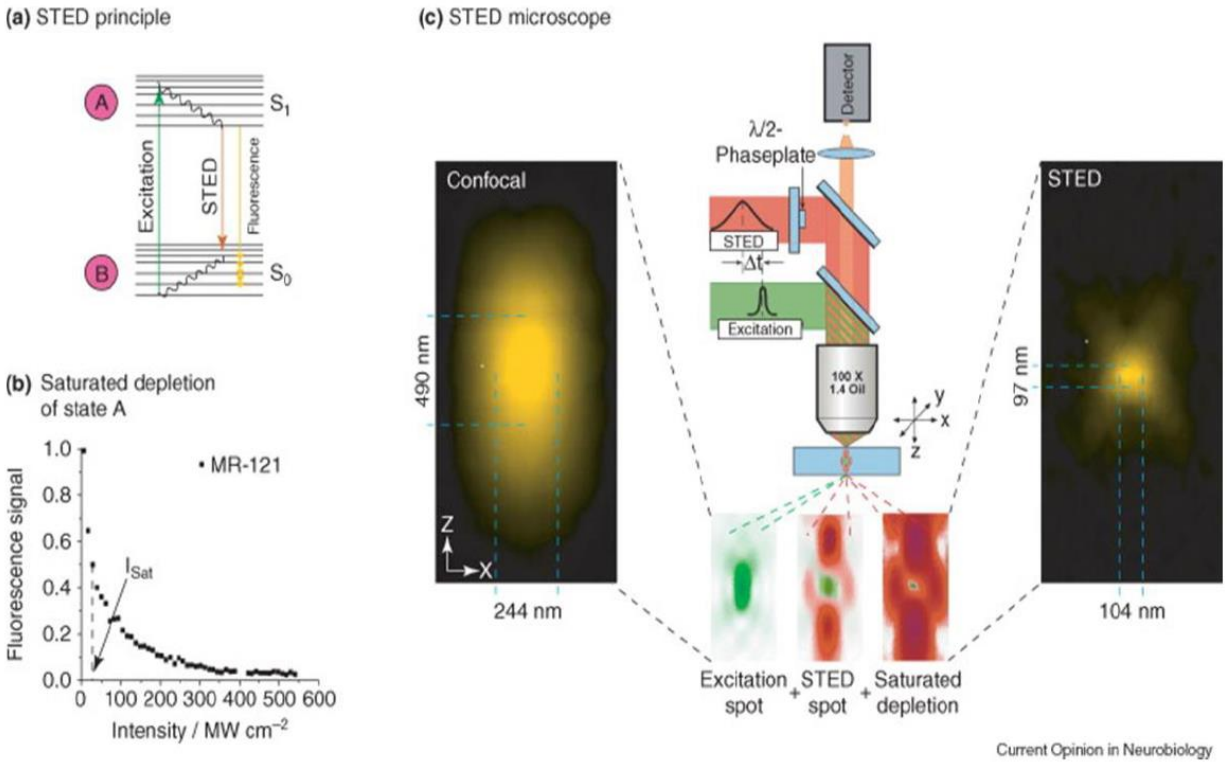


Figure 8.9: From (Hell et al., Current Opinion in Neurobiology 2004, 14:599-609). (a) Excitation from the electronic ground state S0 to S1 with green light and return to S0 with the emission of yellow light or by STED with yellow light. (b) Depletion of S1 with STED. (c) Left: yellow fluorescence density in z- and x-directions in focal plane without STED. Middle: STED microscope. Right: yellow fluorescence spot with high intensity STED.

To circumvent this problem, STORM makes use of "switched fluorphores". These fluorophores are initially optically inactive, though can be activated by irradiation with a specific activation wavelength, after which they can show strong fluorescence at a different wavelength (488 nm for mutants of the green fluorescent protein), this strong fluorescence is followed by irreversible photo-bleaching. By using low-level irradiation at the activation wavelength, a small subset op fluorophores can be activated and observed, hereby fulfilling the condition that only a sparse subset is being observed at a time, and the principle of super-localization can be used. When this first subset is bleached, a second subset can be activated and observed, and so on until all subsets have been sampled and a full super-resolution image can be constructed.

SIM (Structured Illumination)

Structured illumination(SIM) is fundamentally different from the previously described techniques. There is no fundamental resolution limit to STED and STORM, though SIM can go only twice beyond diffraction limit (this was the main motivation to exclude SIM from the 2014 Nobel Prize). However, STED and STORM suffer from some intrinsic disadvantages that are not present in SIM: extremely high labeling densities, very high irradiation intensities and long imaging times are

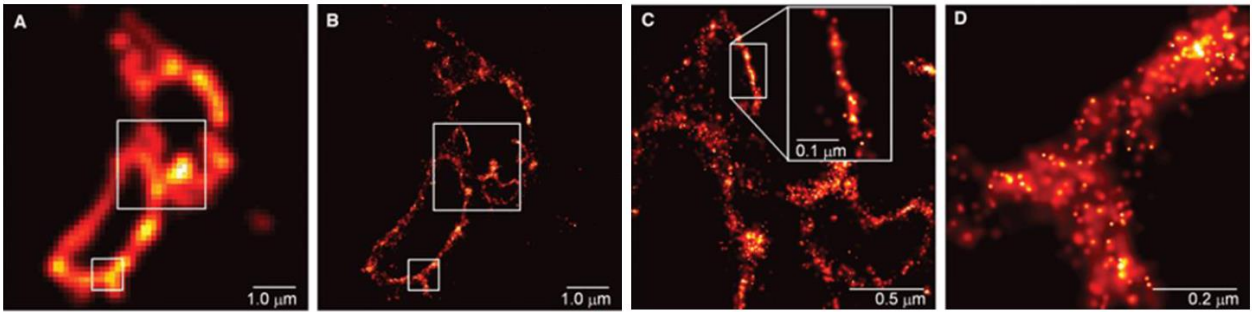


Figure 8.10: From (Betzig et al., Science, 2006, 313, 1642-1645). Distribution of GFP-labeled lysosome without (A) and with (B) PALM. Higher magnifications in (C) and (D).

necessary. These disadvantages specifically complicate imaging of living cells.

The main concept of SIM is to illuminate a sample with sinusoidally patterned light, hereby measuring the Moiré patterns between this illumination and the higher spatial frequencies in the sample. This allows collecting information from frequency space outside the otherwise observable region. This process is carried out in reciprocal space: The Fourier transform (FT) of a SIM image contains a superposition of information of different areas in the reciprocal space. By observing the sample using different phases and orientations of the illumination pattern, it is possible to computationally reconstruct the image in reciprocal space. Using an inverse FT it is then possible to construct an image with up to twice the Abbe-resolution.

8.1.8 Near-field or proximity scanning microscopy

A microscopes limit of resolution can be exceeded by narrowing the field of illumination. In confocal microscopy, the resolution is expected to be improved by up to a factor of 2,5 by illuminating the field point by point with an Airy diffraction spot and by using a small confocal exit pinhole.

The field of illumination can be reduced beyond that defined by diffraction by placing the exit aperture of a tapered light guide or a minute pinhole closely adjacent to the specimen. By scanning such an aperture relative to the specimen, one obtains a proximity-scanned image whose resolution is no longer limited by the diffraction orders captured by the objective lens. Instead, only the size of the scanning pinhole and its proximity to the surface under test limit the resolution. Because the object is in the near field of the aperture, one speaks of an SNOM or Scanning Near-field Optical Microscope.

For non-optical microscopes, e.g. in scanning tunneling, force and other proximity-scanning microscopes, resolution down to atomic dimensions can be obtained on images that reflect topological, electronic and mechanical properties of the specimen surface. In these types of proximity-scanning microscopes, a fine-tipped probe, mounted on a piezoelectric transducer that provides finely controlled x, y and z displacements of the probe, interacts with specific properties of the specimen surface. The resulting interaction signal is captured and fed back to the z-axis transducer, which generally induces the probe tip to rise and fall with the surface contour as the probe is scanned in a raster fashion along the x and y directions.

8.2 Spectrometers

A spectrometer (spectrophotometer, spectrograph or spectroscope) is an instrument used to measure properties of light over a specific portion of the electromagnetic spectrum, typically used in spectroscopic analysis to identify materials. The variable measured is most often the light's intensity but could also, for instance, be the polarization state. The independent variable is usually the wavelength of the light or a unit directly proportional to the photon energy, such as wavenumber ($\frac{1}{\lambda}$) or electron volts, which has a reciprocal relationship to wavelength. A spectrograph is a spectrometer that images a range of wavelengths simultaneously, either onto photographic film or a series of detector elements, or through several exit slits (sometimes called a polychromator). The defining characteristic of a spectrograph is that an entire section of the spectrum is recorded at once. A monochromator is a spectrometer that images a single wavelength or wavelength band at a time onto an exit slit; the spectrum is scanned by the relative motion of the entrance and/or exit optics (usually slits) with respect to the prism or grating.

8.2.1 Prism spectrometer

Prism spectrometers are the oldest spectrometers known to man. Prisms make use of the fact that the refractive index of all materials changes with wavelength and that light is refracted differently by different refractive indices. Prisms, in the sense of the word used here, are triangular, as shown in Figure 8.11. It is not required that the prism is triangular, but it is the simplest shape with the fewest surfaces that produces dispersion. Perhaps the earliest prism spectrometer was the rainbow. It had no slit, but the raindrops performed the refractive dispersion, and the eye of the spectator performed the functions of the camera lens and detector.

A prism deviates a ray that is incident on it. It is clear from the geometry of Figure 8.11 that the total deviation of the beam δ is given by

$$\delta = (\theta_1' - \theta_1) + (\theta_2 - \theta_2') \tag{8.11}$$

Since each normal is perpendicular to a side of the prism angle α, the external angle of the prism is also equal to α, and the sum of the other interior angles is equal to α. Therefore

$$\delta = (\theta_1' - \theta_1) + (\theta_2 - \theta_2') = (\theta_2 + \theta_1) - (\theta_1' + \theta_2') = \theta_2 + \theta_1 - \alpha \tag{8.12}$$

This is the general expression for the deviation. It requires the measurement of three angles. There are several techniques for obtaining a refractive index value for the prism material by finding minimum deviation, illuminating at normal incidence or others. Our interest here is in use of the prism as a tool for generating deviation of light as a function of refractive index and therefore wavelength.

As the prism is rotated (about an axis that is approximately through its center), a given spectral line (color) moves angularly until it reaches a fixed point, and then it moves back - while the prism continues to rotate in the same direction. That point is the position of minimum deviation. It can be shown that at that point

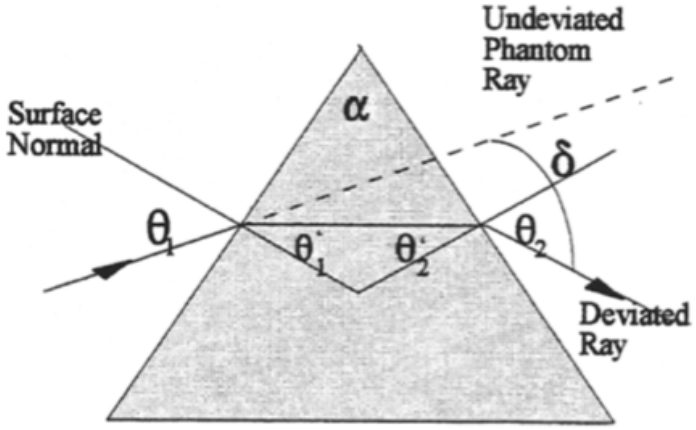


Figure 8.11: Prism angles for prism deviation of a ray

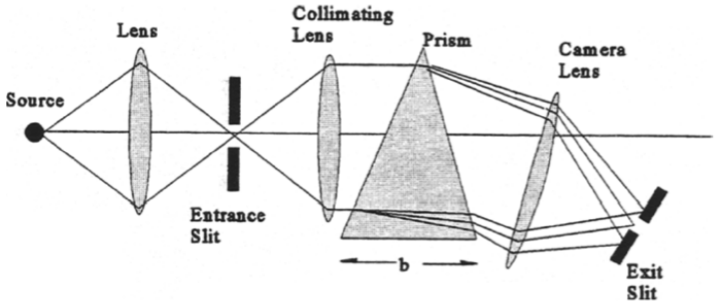


Figure 8.12: Geometry of the prism spectrometer

$$n = \frac{\sin\left(\frac{\alpha+\delta}{2}\right)}{\sin\left(\frac{\alpha}{2}\right)} \tag{8.13}$$

This is a useful way to measure the refractive index of different materials, but it is also the way most prism spectrometers are used.

A typical prism spectrometer uses a source and source optics to focus the light onto a slit. The light from the slit is collimated by a lens and passes through the prism. The arrangement is shown schematically in Figure 8.12. There is refraction at the front and rear surfaces, so that the beams of different colors (wavelengths of light) exit the prism at slightly different angles. These are then focused to slightly different positions near the exit slit, where, as the prism is rotated, they pass sequentially through the slit. There is then some sort of optical and detector system that senses the radiation. The usual operation of a laboratory or commercial spectrometer is to rotate the prism and record the sequential outputs.

This amounts to convolving the input slit across the output slit for each color. If the slits are equal in width, then the convolution is a triangle; if not, the convolution is trapezoidal.

Useful relations for the resolution and the resolving power ($R = \frac{\lambda}{\Delta\lambda} = \frac{\nu}{\Delta\nu}$) can be obtained from a consideration of the changes of the angles with respect to refractive index and therefore wavelength. The resolving power of a prism spectrometer is proportional to the effective base of the prism and the dispersion of the prism material:

$$R = b\frac{dn}{d\lambda} \tag{8.14}$$

This leads one to choose a large prism and a material that has a high dispersion in the region of interest to obtain better resolution. However, where a material has high dispersion it is close to a region of high absorption.

8.2.2 Grating spectrometer

Grating spectrometers make use of the diffraction of light from a regularly spaced, ruled surface. They disperse the light by a combination of diffraction and interference rather than the refractive index variation with wavelength, as with a prism. The spectrum is obtained by rotating the grating; this moves the grating normal relative to the incident and diffracted beams, which changes the wavelength diffracted toward the detection part. Alternatively, a much larger spectral region can be imaged at once by leaving the grating fixed and using a series of exits slits (or an array of detector elements) in the detection plane.

The Littrow spectrometer

In a Littrow-mounted system (Figure 8.13), a single focusing element is used both to collimate the light incident on the plane grating and, in the reverse direction, to focus the light onto the plate or detector placed near the slit. From the grating equation one can easily show that this happens for the angle of incidence α, for which:

$$2\sin(\alpha) = m\frac{\lambda}{\Lambda} \tag{8.15}$$

with m the diffraction order, λ the wavelength, Λ the grating period.

In a Littrow spectrometer, the spectrum is scanned by rotating the grating; this reorients the grating normal, so the angles of incidence α and diffraction β change (even though $\alpha = \beta$ for all λ). The same auxiliary optics can be used as both collimator and detector, since the diffracted rays retrace the incident rays.

The Czerny-Turner spectrometer

This design involves a classical plane grating illuminated by collimated light. The incident light is usually diverging from a source or slit, and collimated by a concave mirror (the collimator), and the diffracted light is focused by a second concave mirror (the detector); see Figure 8.14. Ideally, since the grating is planar, and used in collimated incident light, no aberrations should be introduced into the diffracted wavefronts. In practice, since spherical mirrors are often used,

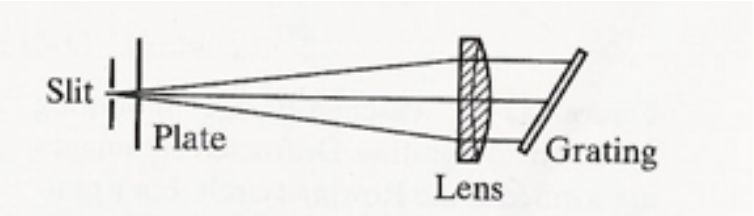


Figure 8.13: Littrow-mounted plane grating

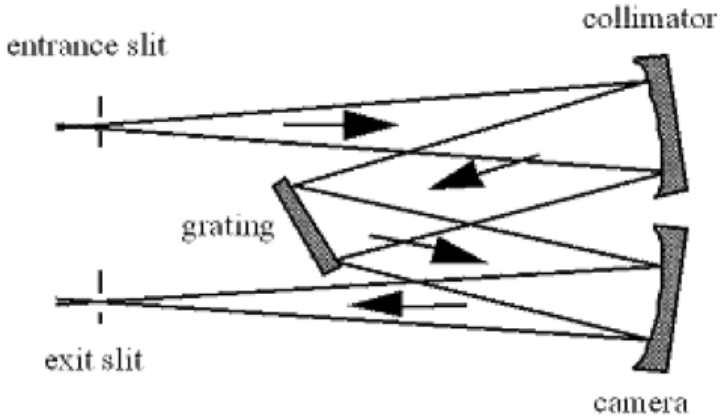


Figure 8.14: The Czerny-Turner spectrometer: the plane grating provides dispersion and the concave mirrors provide focusing.

aberrations arise from their use off-axis. Since the light incident on and diffracted by the grating is collimated, the spectrum remains in focus at the exit slit for each wavelength, since only the grating can introduce wavelength-dependent focusing properties.

Resolving power

The theoretical resolving power of a grating is given by (see chapter 6 Periodic Structures):

$$R = mN \tag{8.16}$$

with N the number of grooves and m the grating order.

from the grating equation we can then easily derive that:

$$R = N\frac{\Lambda(\sin(\alpha) + \sin(\beta)))}{\lambda} = W\frac{(\sin(\alpha) + \sin(\beta))}{\lambda} < \frac{2W}{\lambda} \tag{8.17}$$

with W the total width of the grating.

One can also derive from the grating equation that in a practical system with slit width s and focal distance f the resolving power is given by:

$$R = \frac{m\lambda f}{\Lambda s} \tag{8.18}$$

8.2.3 Fourier Transform Spectroscopy

In FTIR or Fourier Transform Spectroscopy an interferometer is used to determine the unknown spectral density of a source $S(\nu)$. Typically this spectral density is actually the product of the known spectrum of a white light source multiplied with the unknown absorption spectrum of a species (e.g. a gas) through which the light has been transmitted.

The total power in the source is given by:

$$S_0 = \int_0^\infty S(\nu) d\nu \tag{8.19}$$

we can also determine the normalised power density:

$$P(\nu) = \frac{S(\nu)}{S_0} \tag{8.20}$$

Typically the interferometer used is a Twyman-Green interferometer (see next section). If the path length difference between both beams in the interferometer is $2d$, then then the resulting phase difference δ is given by

$$\delta = 2\pi\nu \left(\frac{2d}{c}\right) \tag{8.21}$$

If we initially assume a monochromatic source $S(\nu) = S_i\delta(\nu - \nu_i)$, with frequency ν_i, then the measured intensity, as function of the displacement d is given by:

$$S(d) = S_i[1 + \cos(\delta)] \tag{8.22}$$

Often, the phase difference between both beams is expressed as function of the difference in time travelled:

$$\tau = \frac{2d}{c} \tag{8.23}$$

then eq. 8.22 can be rewritten as:

$$S(\tau) = S_i[1 + \cos(2\pi\nu_i\tau)] \tag{8.24}$$

For a non-monochromatic source $S(\nu)$, we can extend this equation to

$$S(\tau) = \int_0^\infty S(\nu)[1 + \cos(2\pi\nu\tau)] d\nu \tag{8.25}$$

This can be rewritten as the sum of a constant term and an oscillating term

$$S(\tau) = S_0 \left[1 + \gamma(\tau)\right] \tag{8.26}$$

with

$$\gamma(\tau) = \int_0^\infty P(\nu) \cos(2\pi\nu\tau) d\nu \tag{8.27}$$

The oscillating term $\gamma(\tau)$ can easily be determined from the measurement. From eq. 8.27 we can see that $\gamma(\tau)$ is the cosine transform from the unknown source spectral density $P(\nu)$. Hence the latter can be easily determined as the inverse cosine transform of the measured $\gamma(\tau)$:

$$P(\nu) = 4 \int_0^\infty \gamma(\tau) \cos(2\pi\nu\tau) d\tau \tag{8.28}$$

This is the principle of Fourier transform spectroscopy: the interferogram $\gamma(\tau)$ is captured using a suitable interferometer setup. Subsequently the source spectrum is calculated using fourier transform techniques. The technique is mostly used in the IR because they have a better sensitivity then other spectrometers.

8.3 Interferometers

8.3.1 Introduction

Optical interferometers have made feasible a variety of precision measurements using the interference phenomena produced by light waves. After a brief summary of the basic types of interferometers, this chapter will describe some of the interferometers that can be used for such applications as measurements of lengths and small changes in length; optical testing; studies of surface structure; measurements of pressure and temperature; measurements of particle velocities and vibration amplitudes.

8.3.2 Basic types of interferometers

Interferometric measurements require an optical arrangement in which two or more beams, derived from the same source, but traveling along separate paths, are made to interfere. Interferometers can be classified as two-beam interferometers or multiple-beam interferometers according to the number of interfering beams. However they can also be grouped according to the methods used to obtain these beams. The most commonly used form of beamsplitter is a partially reflecting metal or dielectric film on a transparent substrate; other devices that can be used are polarizing prisms and diffraction gratings. The best known types of two-beam interferometers are the Fizeau, the Michelson, the Mach-Zehnder and the Sagnac interferometer; the best known multiple beam interferometer is the Fabry-Perot interferometer.

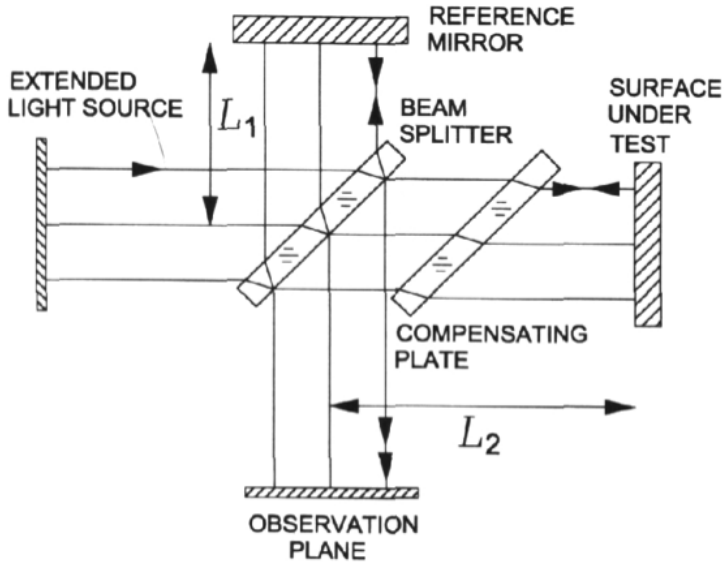


Figure 8.15: The Michelson interferometer concept

The Michelson interferometer

The Michelson interferometer is the most common configuration for optical interferometry and was invented by Albert Abraham Michelson. An interference pattern is produced by splitting a beam of light from an extended light source into two paths, bouncing the beams back and recombining them, as shown schematically in Figure 8.15.

The detector sees two virtual images of the extended light source, one on top of the other, but separated by a certain distance. The reason is that the two arms of the interferometer may have different lengths. Thus the optical path difference is given by

$$OPD = 2[L_1 - L_2 - nT] \tag{8.29}$$

where T is the effective glass thickness traveled by the light rays on one path through the beamsplitter and n is the refractive index of the glass, which is wavelength dependent. To observe interference fringes with a nonmonochromatic or white light source the optical path difference (OPD) must be zero for all wavelengths. This is possible only if the optical path is the same for the two interfering beams, at all wavelengths. We can see that the OPD can be made equal to zero by adjusting L_1 and L_2 for any desired wavelength but not for the whole spectrum, unless T is zero or n does not depend on the wavelength. Accordingly, a compensation plate of the same thickness as the beam splitter is introduced in the second beam to equalize the optical paths in glass such that the beam traverses the beam splitter material three times in both paths. If a compensating plate is used, the OPD can be zero for all wavelengths if $L_1 = L_2$. In this manner, white light fringes can be observed. When monochromatic light is used, no compensation plate is necessary.

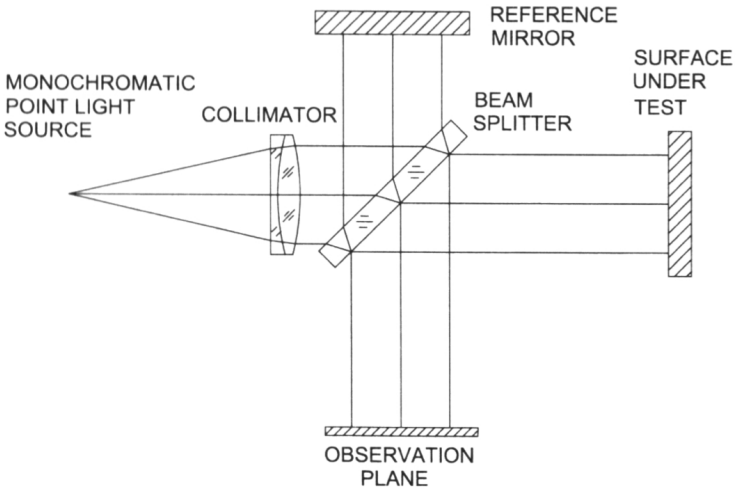


Figure 8.16: The Twyman Green interferometer concept

The Twyman Green interferometer

The Twyman Green interferometer is similar to the Michelson interferometer but uses a monochromatic and collimated optical source. Therefore it does not require the compensating plate. A schematic view is shown in Figure 8.16. Figure 8.17 shows a practical implementation of such an interferometer. The rotating ground glass directly after the laser decreases the spatial coherence of the beam. The halfwave plate in front of the polarizing beam splitter (PBS) allows to control the relative intensity of the object beam and the reference beam. The quarter wave plates turn the polarization of the incident and reflected beams with respect to each other and direct both the object and the reference beam through the PBS towards the camera (CCD) where they interfere with each other.

The Fizeau interferometer

The Fizeau interferometer uses a light source which is a monochromatic point source. The latter produces a spherical wavefront, which becomes flat after being collimated by a converging lens. This wavefront is reflected back on the partially reflecting front face of the reference plate. The transmitted beam goes to the optical element to be measured and is then reflected back to the beamsplitter.

Thus in the Fizeau interferometer, as shown in Figure 8.18, interference fringes are formed between two flat surfaces separated by an air gap and illuminated with a collimated beam. The Fizeau interferometer uses the optical path differences between optical paths L_1 (reflection from reference face) and L_2 (reflection from surface under test) in the figure. If one of the surfaces is a standard reference flat surface, the fringe pattern is a map of the errors of the test surface. Modified forms of the Fizeau interferometer are also used to test convex or concave surfaces by using a converging or diverging beam.

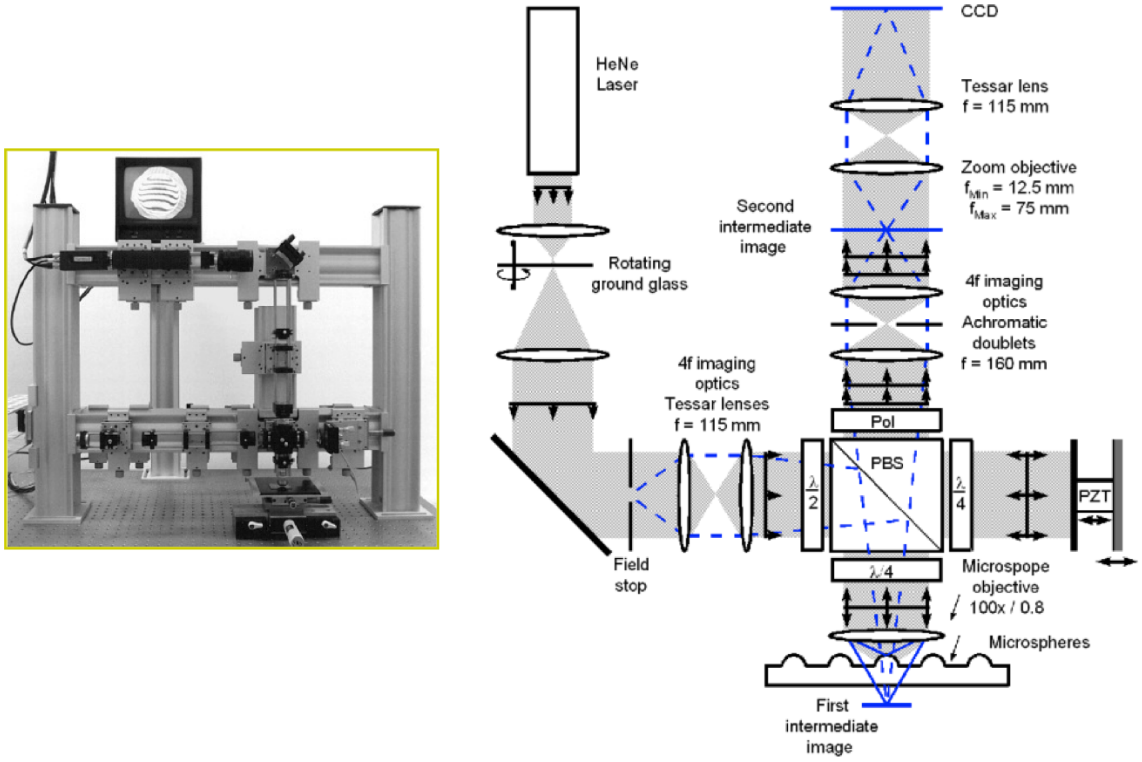


Figure 8.17: The Twyman Green interferometer in practice

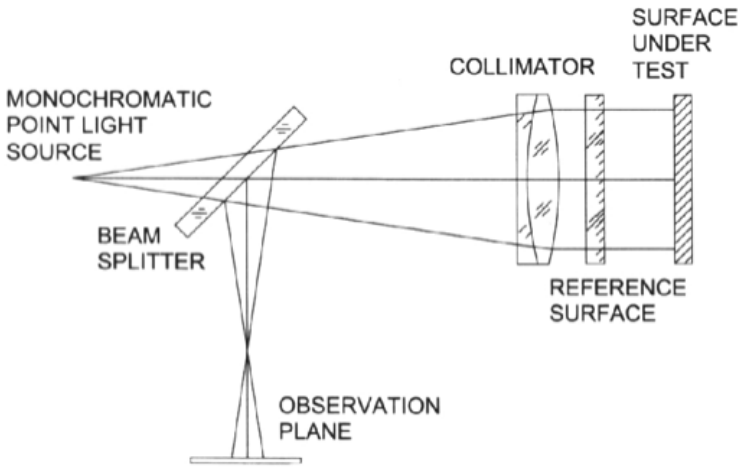


Figure 8.18: The Fizeau interferometer

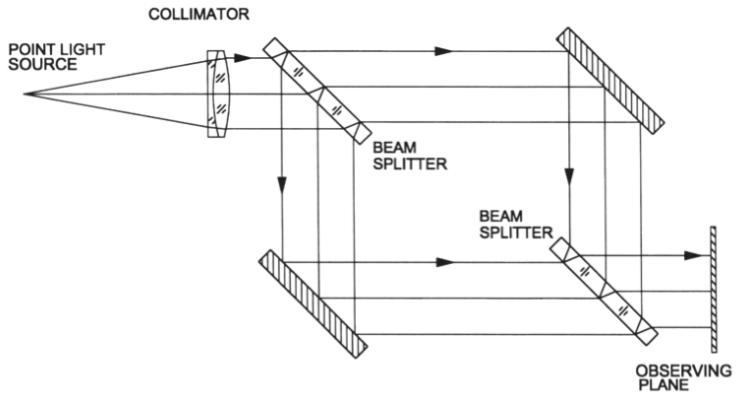


Figure 8.19: Mach-Zehnder interferometer configuration

The Mach-Zehnder interferometer

The Mach-Zehnder interferometer is a device used to determine the phase shift caused by a small sample which is placed in the path of one of two collimated beams (having plane wavefronts) from a coherent light source. The Mach-Zehnder interferometer uses two beam splitters and two mirrors to divide and recombine the beams.

The Sagnac interferometer

In Sagnac interferometers, as shown in Figure 8.20, the two beams traverse the same closed path in opposite directions. Because of this, the interferometer is extremely stable and easy to align even with a broadband light source of which the coherence is low. Usually several mirrors are used, so that the light beams follow a triangular or square trajectory. Fiber optics can also be employed to guide the light. Typically the interferometer is located on a platform that can rotate. When the platform is rotating the lines of the interference pattern are displaced as compared to the position of the interference pattern when the platform is not rotating. The amount of displacement is proportional to the angular velocity of the rotating platform. The axis of rotation does not have to be inside the enclosed area. When the platform is rotating, the point of entry/exit moves during the transit time of the light. This means that one beam has covered less distance than the other beam. This creates the shift in the interference pattern. Therefore, the interference pattern obtained at each angular velocity of the platform features a different phase-shift particular to that angular velocity.

A Sagnac interferometer measures its own angular velocity with respect to the local inertial frame, hence just as a gyroscope it can provide the reference for an inertial guidance system.

Fabry-Perot interferometer

The Fabry-Perot interferometer (also called etalon) makes use of multiple reflections between two closely spaced partially silvered surfaces (see Figure 8.21). Part of the light is transmitted each time the light reaches the second surface, resulting in multiple offset beams, which can interfere

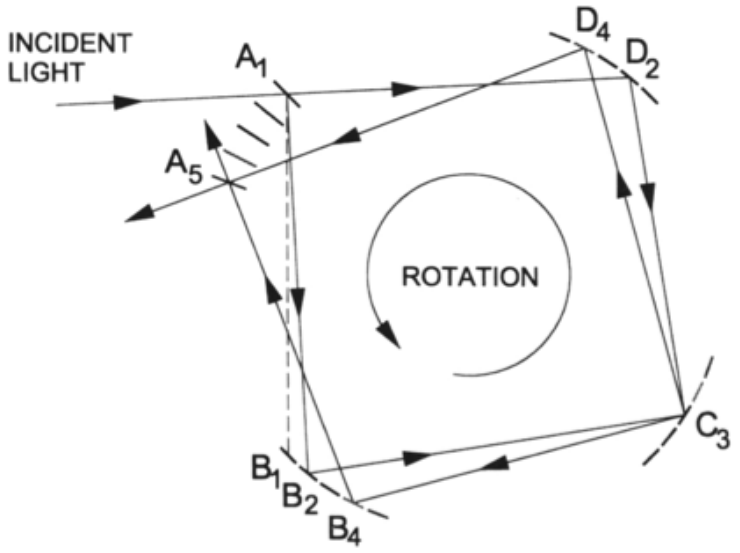


Figure 8.20: The Sagnac interferometer

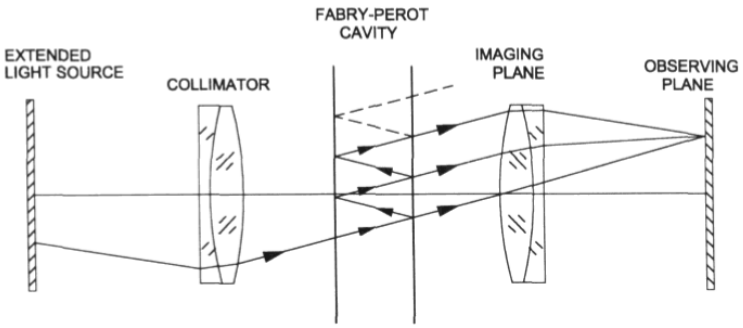


Figure 8.21: The Fabry-Perot interferometer

with each other. The large number of interfering rays produces an interferometer with extremely high resolution, somewhat like the multiple slits of a diffraction grating increase its resolution.

Evaluation of interferometric measurements

The interesting data supplied by the interferometer is the fringe pattern, which is a cosinusoidal function phase modulated by the wavefront distortions being measured. A fringe pattern or interferogram may be expressed as

$$s(x, y) = a(x, y) + b(x, y) \cos(\phi(x, y)) \tag{8.30}$$

where $a(x, y)$ is a slowly varying background illumination, $b(x, y)$ is the amplitude modulation, which is also a low frequency signal and $\phi(x, y)$ is the phase being measured. The purpose of fringe analysis is to detect the two-dimensional phase variation $\phi(x, y)$ that occurs over the inter-

ferogram due to the spatial change in the corresponding physical variable. There are a number of techniques to measure the spatial phase variation of $\phi(x, y)$. Among them we can mention phase shifting interferometry, which requires at least three phase-shifted interferograms. Phase shifting interferometry is the preferred technique whenever the turbulence and mechanical conditions of the interferometer remain constant over the time required to obtain the three phase-shifted interferograms. When the above requirements are not fulfilled, one may analyze just one interferogram and use well-known techniques like e.g. the Fourier transform technique.

The phase-shifting algorithm allows one to extract the object phase directly from a set of recorded interferograms. According to eq. 8.31 the intensity distribution $I(x, y)$ depends on 3 unknown parameters, i.e. the mean intensity, the visibility and the relative phase values. Thus we need at least 3 interferograms to be able to unambiguously calculate the phase function. In the phase-shifting or phase-stepping approach, these additional equations are generated by recording multiple interferograms (3 or more) with a set of different reference phases. It is valid only for a set of reference phase values which are spaced equidistantly over the whole period of 2π. As a conclusion one can state that to determine the object phase distribution $\phi_0(x, y)$) without ambiguity, at least three different reference phase values are necessary. However at that moment the detected phase is still wrapped to modulo 2π due to the arc tangent function involved in the phase estimation process. An unwrapping process is still necessary which is a simple matter of adding or subtracting 2π offsets at each discontinuity encountered in the phase data.

If we take four interferograms, with equally spaced reference phase values $\phi_m = (m-1)\frac{\pi}{2}$ such that the recorded data can be described as:

$$I_m(x, y) = I_0(x, y)\,[1 + V(x, y)cos(\phi_0(x, y) - \phi_m)] \tag{8.31}$$

then taking the ratio of the equations eliminates the intensity modulation term to produce a result that contains only the unknown phase 2π and the four measured intensities:

$$\frac{I_4 - I_2}{I_1 - I_3} = \frac{\sin[\phi(x, y)]}{\cos[\phi(x, y)]} = \tan[\phi(x, y)] \tag{8.32}$$

This equation can now be rearranged to produce the result for the four step PSI algorithm:

$$\phi(x, y) = \tan^{-1}\left[\frac{I_4 - I_2}{I_1 - I_3}\right] \tag{8.33}$$

This simple equation is evaluated at each measurement point to obtain a map of the measured wavefront. The latter can be easily related to the optical path difference (OPD):

$$OPD(x, y) = \lambda\phi(x, y)/2\pi \tag{8.34}$$

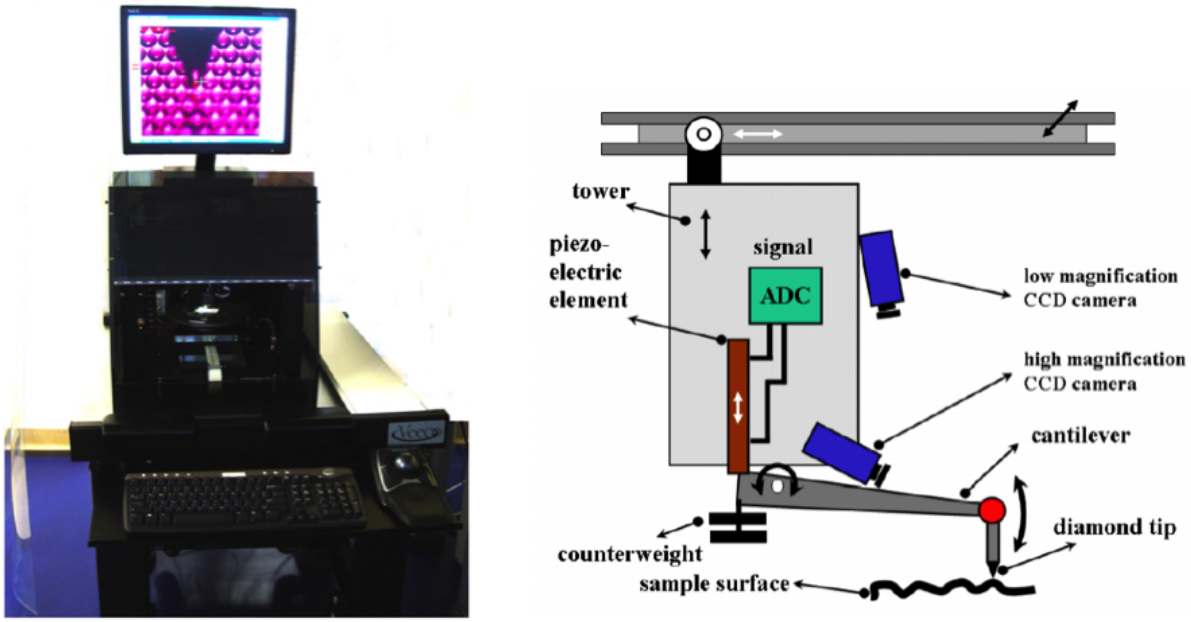


Figure 8.22: (a) Example of a commercial stylus profilometer; (b) Schematic drawing of operation principle

8.4 Profilometers

8.4.1 Stylus profilmeter

The basic principle of all instruments used for physical profile measurements is based on a sensitive detection stylus that scans the substrate surface. More in particular a stylus contact profilometer takes measurements electromechanically by moving the sample under a diamond-tipped stylus. A high-precision translation and rotation stage moves the sample under the stylus according to the user-programmed scan length, translation speed and stylus force. As the stage moves across the sample, surface variations cause the stylus to be translated vertically. The resultant vertical motion of the stylus compresses a piezoelectric element which generates a fairly linear voltage response. The analog signal is further amplified and converted by an A/D convertor into a digital format. These digitized signals can then be translated into height variation information by dedicated computer algorithms. Care has to be taken because this is a contact measurement technique and the stylus probe tip can easily damage or alter the surface of soft materials. There are also limits in lateral resolution, set by the size of the probe tip. Features that are smaller than the probe tip will be recorded as being broader than they really are, therefore a stylus profilometer is less suited to measure surface roughnesses. Also some surface geometries will pose problems, e.g. narrow troughs will prevent the probe to physically reach the bottom, or highly curved or steep slopes will be convoluted with the shape of the stylus tip. This reduced lateral resolution results in a loss of surface height information and in some cases the measurements depend on the shape of the stylus.

In todays high resolution mechanical profilometers the tip size is often reduced to a few micrometers. Consequently very high lateral resolution can be achieved if the aspect ratio of the profile

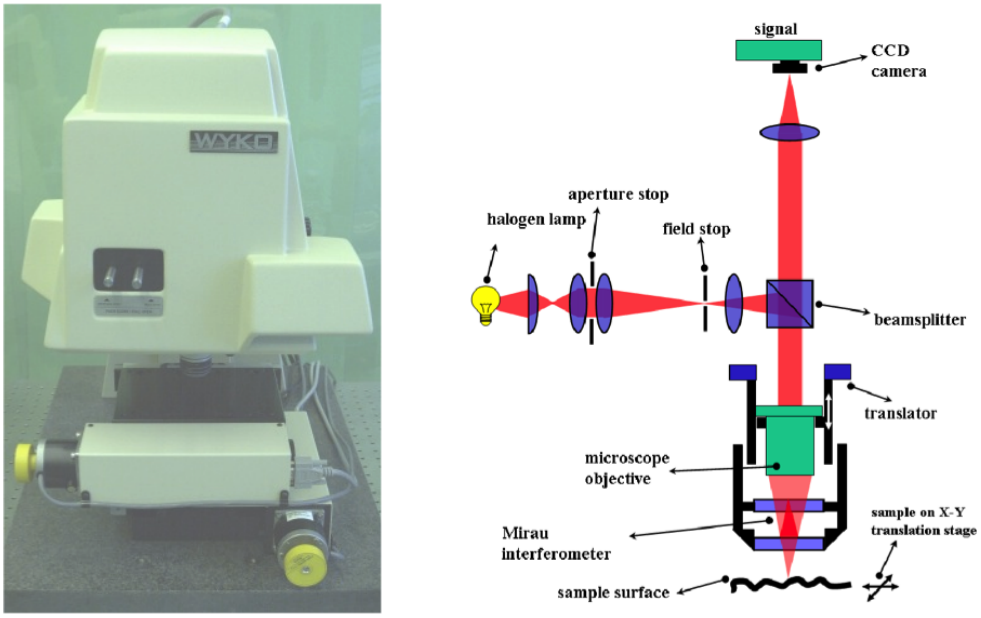


Figure 8.23: (a) Picture of a commercial white light interferometer; (b) and its working principle

is not too large. Another important drawback of profilometers is the time-consuming scanning process involved in a measurement of 3D objects. The better the lateral resolution achieved with extremely fine tip profiles, the larger the number of sampling points per area where measurements have to be performed.

8.4.2 Optical non-contact profilometer

An optical non-contact optical profiler or white light interferometer is a surface profiler system that uses vertical scanning interferometry (VSI) to measure rough surfaces and heights ranging between 160 nm and 2 mm. This non-contact profiler is based on a Mirau interference microscope where a white light beam passes through a beam splitter which reflects half of the incident beam to a reference surface and transmits the other half onto the sample (see Figure 8.23b). The light reflected from the sample and from the reference surface then recombines at the beam splitter to form interference fringes. The system measures the degree of the fringe modulation or the fringe contrast. Because white light has a short coherence length, interference fringes are present only over a very shallow depth for each focus position. Fringe contrast at a single sample point reaches a peak as the sample is translated through focus. During a measurement, the reference arm containing the interferometric objective moves vertically to scan the surface at varying heights. A linearized piezoelectric transducer (PZT) precisely controls this motion. As the system scans downward, an interference signal for each point on the surface is recorded. Finally a series of advanced computer algorithms are used to demodulate the envelope of the fringe signal and to extract the surface information.

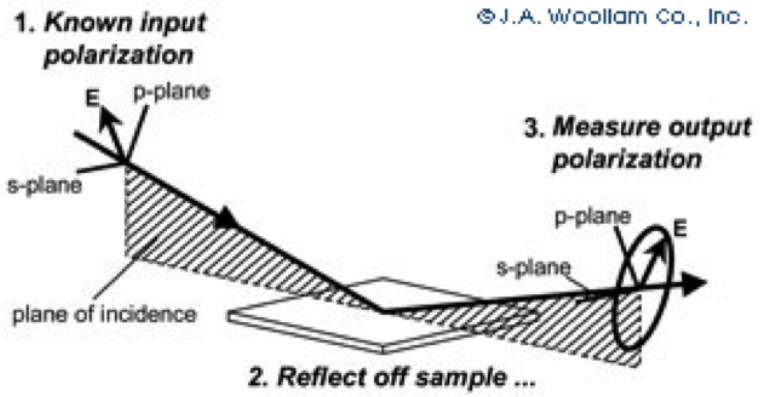


Figure 8.24: Typical ellipsometry configuration, where linearly polarized light is reflected from the sample surface and the polarization change is measured to determine the sample response

8.5 Ellipsometers

8.5.1 Introduction

Ellipsometry measures a change in polarization as light reflects or transmits from a material structure. The polarization change is represented as an amplitude ratio, Ψ, and the phase difference, Δ. The measured response depends on optical properties and thickness of individual materials. Thus, ellipsometry is primarily used to determine film thickness and optical constants. However, it is also applied to characterize composition, crystallinity, roughness, doping concentration, and other material properties associated with a change in optical response.

8.5.2 Ellipsometric measurements

Ellipsometry is primarily interested in how p- and s- components change upon reflection or transmission in relation to each other. In this manner, the reference beam is part of the experiment. A known polarization is reflected or transmitted from the sample and the output polarization is measured. The change in polarization is the ellipsometry measurement, commonly written as:

$$\rho = \tan(\Psi)e^{i\Delta} \tag{8.35}$$

An ellipsometry measurement is shown in Figure 8.24. The incident light is linear with both p- and s- components. The reflected light has undergone amplitude and phase changes for both p- and s- polarized light, and ellipsometry measures their changes.

The primary tools for collecting ellipsometry data all include the following: light source, polarization generator, sample, polarization analyzer, and detector. The polarization generator and analyzer are constructed of optical components that manipulate the polarization: polarizers, compensators, and phase modulators. Common ellipsometer configurations include rotating analyzer (RAE), rotating polarizer (RPE), rotating compensator (RCE), and phase modulation (PME). The RAE configuration is shown in Figure 8.25. A light source produces unpolarized light which is then sent through a polarizer. The polarizer allows light of a preferred electric field orientation

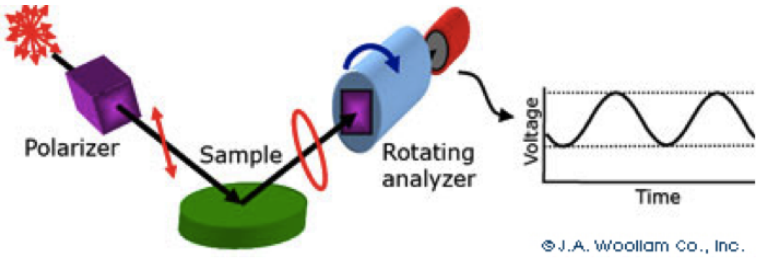


Figure 8.25: Rotating analyzer ellipsometer configuration uses a polarizer to define the incoming polarization and a rotating polarizer after the sample to analyze the outgoing light. The detector converts light to a voltage whose dependence yields the measurement of the reflected polarization.

to pass. The polarizer axis is oriented between the p- and s- planes, such that both arrive at the sample surface. The linearly polarized light reflects from the sample surface, becomes elliptically polarized, and travels through a continuously rotating polarizer (referred to as the analyzer). The amount of light allowed to pass will depend on the polarizer orientation relative to the electric field ellipse coming from the sample. The detector converts light to an electronic signal to determine the reflected polarization. This information is compared to the known input polarization to determine the polarization change caused by the sample reflection. This is the ellipsometry measurement of Ψ and Δ.