Wave Optics

Lecture about the principles of

wave optics

N. Lindlein

Institute of Optics, Information and Photonics

University of Erlangen–Nürnberg,

Staudtstr. 7/B2, D–91058 Erlangen

norbert.lindlein@physik.uni-erlangen.de

First edition: October 2005

Revised version: May 2009

Introduction

The quest for the nature of light is centuries old and today there can be at least three answers to the question what light is depending on the experiment which is used to investigate the nature of light: (i) light consists of rays which propagate e.g. rectilinear in homogeneous media, (ii) light is an electromagnetic wave, (iii) light consists of small portions of energy, the so called photons. The first property will be treated in the lecture about **Geometrical Optics** and geometrical optics can be interpreted as a special case of wave optics for very small wavelengths. On the other hand the interpretation as photons is unexplainable with wave optics and first of all also contradicting to wave optics. Only the theory of quantum mechanics and quantum field theory can explain light as photons and simultaneously as an electromagnetic wave. The field of optics which treats this subject is generally called **Quantum Optics** and is also one of the lecture courses in optics.

In this lecture about **Wave Optics** the electromagnetic property of light is treated and the basic equations which describe all electromagnetic phenomena which are relevant for us are Maxwell's equations. Starting with the Maxwell equations the wave equation and the Helmholtz equation will be derived. Here, we will try to make a trade–off between theoretical exactness and a practical approach. For an exact analysis see e.g. [1]. After this, some basic properties of light waves like polarization, interference, and diffraction will be described. Especially, the propagation of coherent scalar waves is quite important in optics. Therefore, the chapter about diffraction will treat several propagation methods like the method of the angular spectrum of plane waves, which can be easily implemented in a computer, or the well–known diffraction integrals of Fresnel–Kirchhoff, Fresnel and Fraunhofer. In modern physics and engineering lasers are very important and therefore the propagation of a coherent laser beam is of special interest. A good approximation for a laser beam is a Hermite–Gaussian mode and the propagation of a fundamental Gaussian beam can be performed very easily if some approximations of paraxial optics are valid. The formula for this are treated in one of the last chapters of this lecture script. It is tried to find a tradeoff between theoretical and applied optics. Therefore, practically important subjects of wave optics like interferometry, optical image processing and filtering (Fourier optics), and holography will also be treated in this lecture.

Notes to this lecture script

The lecture **Wave Optics** (Grundkurs Optik II: Wellenoptik) is the second course in optics at the University of Erlangen–Nürnberg following the first course about **Geometrical and Technical Optics** (Grundkurs Optik I: Geometrische und Technische Optik). So, basic knowledge of geometrical optics is necessary to understand this lecture. Besides this, basic knowledge of electromagnetism is very useful. In mathematics, basic knowledge of analysis, vector calculus, and linear algebra are expected. So, in general this lecture should be attended during the advanced study period after having passed the "Vordiplom".
The lecture itself has two hours per week accompanied by an exercise course of also two hours per week. In order to get a certificate the lecture and the exercises have to be attended on a regular base and it is expected that every student performs from time to time one of the exercises at the blackboard.

Contents

List of Figures

List of Tables

Chapter 1

Maxwell's equations and the wave equation

1.1 The Maxwell equations

The well–known equations of J.C. Maxwell about electrodynamics [1] are the basis for our considerations and will be given here in international SI units. The following physical quantities are used:

- $\boldsymbol{E}$: electric (field) vector (*dt.:* Elektrische Feldstärke), unit $[E]$=1 V/m
- $\boldsymbol{D}$: electric displacement (*dt.:* Elektrische Verschiebungsdichte), unit $[D]$=1 A s/m^2
- $\boldsymbol{H}$: magnetic (field) vector (*dt.:* Magnetische Feldstärke), unit $[H]$=1 A/m
- $\boldsymbol{B}$: magnetic induction (*dt.:* Magnetische Induktion/Flußdichte), unit $[B]$=1 V s/m^2=1 T
- $\boldsymbol{j}$: electric current density (*dt.:* Elektrische Stromdichte), unit $[j]$=1 A/m^2
- ρ: (free) electric charge density (*dt.:* Elektrische Ladungsdichte), unit $[\rho]$=1 A s/m^3

All quantities can be functions of the spatial coordinates with position vector $\boldsymbol{r} = (x, y, z)$ and the time t. In the following this explicit functionality is mostly omitted in the equations if it is clear from the context.
The Maxwell equations are formulated in the differential form by using the so called Nabla operator

$$\boldsymbol{\nabla} := \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \tag{1.1.1}$$

The four Maxwell equations and the physical interpretation are:

$$\boldsymbol{\nabla} \times \boldsymbol{E}(\boldsymbol{r}, t) = -\frac{\partial \boldsymbol{B}(\boldsymbol{r}, t)}{\partial t} \tag{1.1.2}$$

The vortices of the electric field $\boldsymbol{E}$ are caused by temporal variations of the magnetic induction $\boldsymbol{B}$ (Faraday's law of induction).

$$\boldsymbol{\nabla} \times \boldsymbol{H}(\boldsymbol{r},t) = \frac{\partial \boldsymbol{D}(\boldsymbol{r},t)}{\partial t} + \boldsymbol{j}(\boldsymbol{r},t) \tag{1.1.3}$$

The vortices of the magnetic field $\boldsymbol{H}$ are either caused by an electric current with density $\boldsymbol{j}$ or by temporal variations of the electric displacement $\boldsymbol{D}$ (Ampère's law + Maxwell's extension). The quantity $\partial \boldsymbol{D}/\partial t$ is called the electric displacement current (*dt.:* Maxwell'scher Verschiebungsstrom).

$$\boldsymbol{\nabla} \cdot \boldsymbol{D}(\boldsymbol{r},t) = \rho(\boldsymbol{r},t) \tag{1.1.4}$$

The sources of the electric displacement $\boldsymbol{D}$ are the electric charges with density ρ (Gauss' law).

$$\boldsymbol{\nabla} \cdot \boldsymbol{B}(\boldsymbol{r},t) = 0 \tag{1.1.5}$$

The magnetic field (induction) is solenoidal (*dt.:* quellenfrei), i.e. there exist no "magnetic charges" (Gauss' law for magnetism).

1.1.1 The continuity equation

From equation (1.1.3) and (1.1.4) the conservation of the electric charge can be obtained by using the mathematical identity $\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \times \boldsymbol{H}) = 0$

$$\frac{\partial \rho}{\partial t} + \boldsymbol{\nabla} \cdot \boldsymbol{j} = 0 \tag{1.1.6}$$

This equation is called the **continuity equation** of electrodynamics because it is analogue to the continuity equation of hydrodynamics. By integrating over a volume V with a closed surface A and applying Gauss' theorem the following equation is obtained:

$$\int\limits_V \frac{\partial \rho}{\partial t}\, dV = -\int\limits_V \boldsymbol{\nabla} \cdot \boldsymbol{j}\, dV = -\oint\limits_A \boldsymbol{j} \cdot \boldsymbol{dA} \tag{1.1.7}$$

Note: The integral $\int_V f dV$ always indicates here and in the following a volume integral of a scalar function f over the volume V whereas the symbol $\oint_A \boldsymbol{a} \cdot \boldsymbol{dA}$ always indicates a surface integral of the vector function $\boldsymbol{a}$ over the closed surface A which borders the volume V. The vector $\boldsymbol{dA}$ always points outwards of the closed surface.
Therefore, the left side of equation (1.1.7) is the temporal variation of the total electric charge Q in the volume V and the right side of equation (1.1.7) is the net electric current I_{net} (i.e. current of positive charges flowing out of the surface plus current of negative charges flowing in the surface minus current of positive charges flowing in the surface minus current of negative charges flowing out of the surface) which flows through the closed surface A:

$$\frac{\partial Q}{\partial t} = -I_{net} \tag{1.1.8}$$

If the net current I_{net} is positive the total charge in the volume decreases during time, i.e. the volume is charged negatively.

1.1.2 Energy conservation in electrodynamics

From the equations (1.1.2) and (1.1.3) a law of energy conservation of electrodynamics can be deduced by calculating the scalar product of $\boldsymbol{E}$ with (1.1.3) minus the scalar product of $\boldsymbol{H}$ with (1.1.2):

$$\boldsymbol{E} \cdot (\boldsymbol{\nabla} \times \boldsymbol{H}) - \boldsymbol{H} \cdot (\boldsymbol{\nabla} \times \boldsymbol{E}) = \boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{E} \cdot \boldsymbol{j} + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t}$$

According to the rules of calculating with a Nabla operator the following equation is obtained:

$$\boldsymbol{\nabla} \cdot \boldsymbol{S} = \boldsymbol{\nabla} \cdot (\boldsymbol{E} \times \boldsymbol{H}) = -\left[\boldsymbol{E} \cdot (\boldsymbol{\nabla} \times \boldsymbol{H}) - \boldsymbol{H} \cdot (\boldsymbol{\nabla} \times \boldsymbol{E})\right]$$

The quantity $\boldsymbol{S}$

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{1.1.9}$$

is called the **Poynting vector** and has the physical unit of an intensity: $[S]$=1 V A m^{-2}=1 W/m^2, i.e. power per surface area. The Poynting vector is due to the property of a cross product of two vectors perpendicular to both the electric and magnetic vector and its absolute value describes the flow of energy per unit area and time unit through a surface perpendicular to the Poynting vector. It therefore describes the energy transport of the electromagnetic field. The sources of $\boldsymbol{S}$ are connected with temporal variations of the electric displacement or the magnetic induction or with explicit electric currents.

$$\boldsymbol{\nabla} \cdot \boldsymbol{S} = -\left(\boldsymbol{E} \cdot \frac{\partial \boldsymbol{D}}{\partial t} + \boldsymbol{E} \cdot \boldsymbol{j} + \boldsymbol{H} \cdot \frac{\partial \boldsymbol{B}}{\partial t}\right) \tag{1.1.10}$$

In the next section it will be shown for the special case of an isotropic dielectric material that this equation can be interpreted as an equation of energy conservation.

1.1.3 Energy conservation in the special case of isotropic dielectric materials

In section 1.3 we will see that isotropic dielectric materials are described with the following equations. The charge density and the electric currents are both zero.

$$\rho = 0, \quad \boldsymbol{j} = 0 \tag{1.1.11}$$

Additionally, there are the following linear interrelations between the electric and magnetic quantities:

$$\boldsymbol{D}(\boldsymbol{r}, t) = \epsilon_0 \epsilon(\boldsymbol{r})\, \boldsymbol{E}(\boldsymbol{r}, t)\,, \quad \boldsymbol{B}(\boldsymbol{r}, t) = \mu_0 \mu(\boldsymbol{r})\, \boldsymbol{H}(\boldsymbol{r}, t)\,, \tag{1.1.12}$$

ϵ is the dielectric function (*dt.:* Dielektrizitätszahl) of the material and μ the magnetic permeability (*dt.:* Permeabilitätszahl). Both are functions of the position $\boldsymbol{r}$. The dielectric constant of the vacuum (*dt.:* Elektrische Feldkonstante oder Influenzkonstante) $\epsilon_0 = 8.8542 \cdot 10^{-12}$ A s V^{-1} m^{-1} and the magnetic permeability of the vacuum (*dt.:* Magnetische Feldkonstante oder Induktionskonstante) $\mu_0 = 4\pi \cdot 10^{-7}$ V s A^{-1} m^{-1} are related with the light speed in vacuum (*dt.:* Vakuumlichtgeschwindigkeit) c via:

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \tag{1.1.13}$$

with $c = 2.99792458 \cdot 10^8$ m s^{-1}. In fact the light speed in vacuum is defined in the SI system as a fundamental constant of nature to exactly this value so that the basic unit of length (1 m)

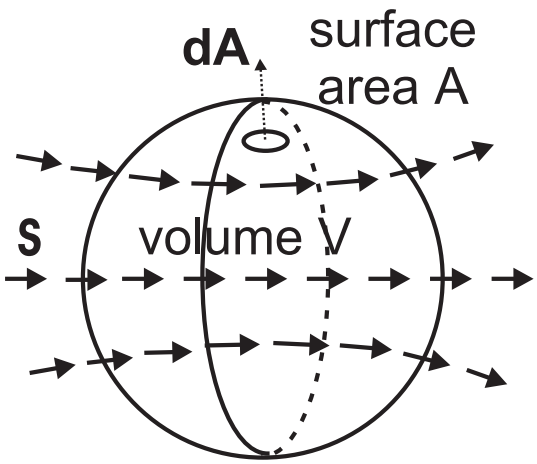


Figure 1.1: Illustration of the quantities used for applying Gauss theorem. The surface A need not to be the surface of a sphere with volume V but can be an arbitrary closed surface. The small dotted vector symbolizes the infinitesimal surface vector $\boldsymbol{dA}$, whereas the other vectors represent the vector field of the local Poynting vectors $\boldsymbol{S}$ at a fixed time.

can be connected with the basic unit of time (1 s). The magnetic permeability of the vacuum is also defined in order to connect the basic SI unit of the electric current 1 A with the mechanical basic SI units of mass (1 kg), length (1 m) and time (1 s). So, only the dielectric constant of the vacuum has to be determined by experiments, whereas c and μ_0 are defined constants in the SI system.
In dielectrics equation (1.1.10) reduces by using equations (1.1.11) and (1.1.12) to the following equation:

$$\boldsymbol{\nabla}\cdot\boldsymbol{S} = -\left(\epsilon_0\epsilon\boldsymbol{E}\cdot\frac{\partial\boldsymbol{E}}{\partial t} + \mu_0\mu\boldsymbol{H}\cdot\frac{\partial\boldsymbol{H}}{\partial t}\right) = -\frac{1}{2}\frac{\partial}{\partial t}\left(\epsilon_0\epsilon\boldsymbol{E}\cdot\boldsymbol{E} + \mu_0\mu\boldsymbol{H}\cdot\boldsymbol{H}\right) \tag{1.1.14}$$

By integrating both sides of the equation over a volume V which is bounded by a closed surface A (see fig. 1.1) Gauss' theorem can be applied:

$$\begin{aligned} P_{net} &:= \oint\limits_A \boldsymbol{S}\cdot\boldsymbol{dA} = \int\limits_V \boldsymbol{\nabla}\cdot\boldsymbol{S}\,dV = \\ &= \int\limits_V \left[-\frac{1}{2}\frac{\partial}{\partial t}\left(\epsilon_0\epsilon\boldsymbol{E}\cdot\boldsymbol{E} + \mu_0\mu\boldsymbol{H}\cdot\boldsymbol{H}\right)\right] dV = \\ &= -\frac{\partial}{\partial t}\int\limits_V \left(\frac{1}{2}\,\epsilon_0\epsilon\boldsymbol{E}\cdot\boldsymbol{E} + \frac{1}{2}\,\mu_0\mu\boldsymbol{H}\cdot\boldsymbol{H}\right) dV = -\frac{\partial}{\partial t}\int\limits_V w\,dV \end{aligned} \tag{1.1.15}$$

The integral $\int_V \boldsymbol{\nabla}\cdot\boldsymbol{S}\,dV$ symbolizes the volume integral of $\boldsymbol{\nabla}\cdot\boldsymbol{S}$ over the volume V whereas the integral $\oint_A \boldsymbol{S}\cdot\boldsymbol{dA}$ symbolizes the surface integral of the Poynting vector over the closed surface A. The vector $\boldsymbol{dA}$ in the integral always points outwards in the case of a closed surface. Therefore, P_{net} is equal to the net amount of the electromagnetic power (difference between the power flowing out of the closed surface and the power flowing in the closed surface) which flows through the closed surface. So, a positive value of P_{net} indicates that more energy flows out

of the surface than in. Since the right side of equation (1.1.15) must therefore also have the physical unit of a power (unit 1 W=1 J/s) it is clear that the quantity

$$w := \frac{1}{2}\left(\epsilon_0\epsilon \boldsymbol{E}\cdot\boldsymbol{E} + \mu_0\mu \boldsymbol{H}\cdot\boldsymbol{H}\right) = w_e + w_m \tag{1.1.16}$$

is the **energy density** of the electromagnetic field in isotropic dielectric materials having the unit 1 J/m^3. The first term

$$w_e = \frac{1}{2}\epsilon_0\epsilon \boldsymbol{E}\cdot\boldsymbol{E} \tag{1.1.17}$$

is the **electric energy density** and the second term

$$w_m = \frac{1}{2}\mu_0\mu \boldsymbol{H}\cdot\boldsymbol{H} \tag{1.1.18}$$

the **magnetic energy density**. The negative sign on the right side of equation (1.1.15) just indicates that the amount of energy in the volume decreases over the time if the net amount of power P_{net} flowing through the surface is positive because this means that in total energy flows out of the closed surface. If the net electromagnetic power flow through the surface is zero, i.e. $P_{net} = 0$, the total amount of the electromagnetic energy $\int_V w\, dV$ in the volume is constant. This again shows that it is useful to interpret w as an energy density.

1.1.4 The wave equation in homogeneous dielectrics

In this section the behavior of light in homogeneous dielectric materials will be discussed. In homogeneous materials the dielectric function ϵ and the magnetic permeability μ are both constants. A special case is the vacuum where both constants are one ($\epsilon = 1$, $\mu = 1$). The conclusion that electromagnetic waves can also exist in vacuum without any matter was one of the most important discoveries in physics in the 19th century.
In homogeneous dielectrics the Maxwell equations (1.1.2)–(1.1.5) can be simplified by using equations (1.1.11) and (1.1.12) with ϵ and μ constant:

$$\boldsymbol{\nabla}\times\boldsymbol{E}(\boldsymbol{r},t) = -\mu_0\mu\frac{\partial \boldsymbol{H}(\boldsymbol{r},t)}{\partial t} \tag{1.1.19}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H}(\boldsymbol{r},t) = \epsilon_0\epsilon\frac{\partial \boldsymbol{E}(\boldsymbol{r},t)}{\partial t} \tag{1.1.20}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{E}(\boldsymbol{r},t) = 0 \tag{1.1.21}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{H}(\boldsymbol{r},t) = 0 \tag{1.1.22}$$

These equations are completely symmetrical to a simultaneous replacement of $\boldsymbol{E}$ with $\boldsymbol{H}$ and $\epsilon_0\epsilon$ with $-\mu_0\mu$.
In the following the vector identity

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - (\boldsymbol{\nabla}\cdot\boldsymbol{\nabla})\boldsymbol{E} = \boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - \Delta\boldsymbol{E} \tag{1.1.23}$$

has to be used. Thereby, the Laplacian operator $\Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$ has to be applied to each component of $\boldsymbol{E}$!
Equation (1.1.19) together with (1.1.21) then results in:

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = -\Delta\boldsymbol{E} = -\mu_0\mu\boldsymbol{\nabla}\times\frac{\partial \boldsymbol{H}}{\partial t} = -\mu_0\mu\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{H}) \tag{1.1.24}$$

Using equation (1.1.20) the **wave equation** for the electric vector in a homogeneous dielectric is obtained:

$$-\Delta \boldsymbol{E} = -\mu_0 \mu \frac{\partial}{\partial t}\left(\epsilon_0 \epsilon \frac{\partial \boldsymbol{E}}{\partial t}\right) \quad \Rightarrow \quad \Delta \boldsymbol{E} - \epsilon_0 \mu_0 \epsilon \mu \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0 \tag{1.1.25}$$

By using (1.1.13) this is usually written as:

$$\Delta \boldsymbol{E} - \frac{n^2}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = 0 \tag{1.1.26}$$

The **refractive index** (*dt.:* Brechzahl) n of a homogeneous dielectric is defined as:

$$n = \sqrt{\epsilon \mu} \tag{1.1.27}$$

Because of the symmetry in $\boldsymbol{E}$ and $\boldsymbol{H}$ of the Maxwell equations (1.1.19)–(1.1.22) in homogeneous dielectrics the same equation also holds for the magnetic vector:

$$\Delta \boldsymbol{H} - \frac{n^2}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} = 0 \tag{1.1.28}$$

1.1.5 Plane waves in homogeneous dielectrics

By defining the so called **phase velocity** (*dt.:* Phasengeschwindigkeit)

$$v = \frac{c}{n} \tag{1.1.29}$$

a solution of equation (1.1.26) or (1.1.28) is:

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{f}(\boldsymbol{e} \cdot \boldsymbol{r} \mp vt) \tag{1.1.30}$$

$$\boldsymbol{H}(\boldsymbol{r}, t) = \boldsymbol{g}(\boldsymbol{e} \cdot \boldsymbol{r} \mp vt) \tag{1.1.31}$$

This can be seen by using

$$u := \boldsymbol{e} \cdot \boldsymbol{r} \mp vt = e_x x + e_y y + e_z z \mp vt \quad \text{with} \quad e_x^2 + e_y^2 + e_z^2 = 1 \tag{1.1.32}$$

so that it holds

$$\Delta \boldsymbol{E} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} = \left(e_x^2 + e_y^2 + e_z^2\right)\frac{\partial^2 \boldsymbol{f}(u)}{\partial u^2} = \frac{\partial^2 \boldsymbol{f}(u)}{\partial u^2}$$

$$\frac{\partial^2 \boldsymbol{E}}{\partial t^2} = v^2 \frac{\partial^2 \boldsymbol{f}(u)}{\partial u^2}$$

and the same is valid for $\boldsymbol{H}$ in equation (1.1.31). The quantity nu has the physical unit of a path and the first term $n\boldsymbol{e} \cdot \boldsymbol{r}$ is called optical path difference OPD because it is the product of the geometrical path times the refractive index n.
A solution of the type (1.1.30) or (1.1.31) is called a plane wave because of the following reason. The value of $\boldsymbol{E}$ remains constant for a constant argument $u = u_0$ defined by equation (1.1.32). The same is valid for $\boldsymbol{H}$. Now, if we consider e.g. $u = u_0 = 0$ and take the negative sign in equation (1.1.32) we obtain:

$$u = 0 \quad \Rightarrow \quad \boldsymbol{e} \cdot \boldsymbol{r} = vt \tag{1.1.33}$$

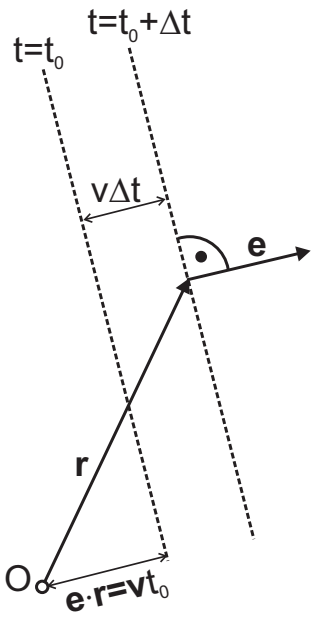


Figure 1.2: The plane surfaces of constant optical path in the case of a plane wave. O is the origin of the coordinate system and the dashed lines indicate the planes at two different times for the fixed value $u = 0$.

The geometrical path at the time $t = 0$ has then also to be zero and this means that the plane then passes through the origin of the coordinate system. For a fixed value t_0 equation (1.1.33) describes a plane surface in space (see fig. 1.2) at a distance vt_0 from the origin. At the time $t_0 + \Delta t$ it describes again a plane surface parallel to the plane at $t = 0$ but with a distance $v(t_0 + \Delta t)$ from the origin. The unit vector $\boldsymbol{e}$ is perpendicular to the planes and points in the direction of propagation if the negative sign is used in equation (1.1.32) (what we have done here and what we will do in the following) and points in the opposite direction if the positive sign is used.

1.1.6 The orthogonality condition for plane waves in homogeneous dielectrics

The Maxwell equations (1.1.19)–(1.1.22) do not allow all orientations of the electric and magnetic vector relative to the propagation direction $\boldsymbol{e}$ of a plane wave. Equations (1.1.30) and (1.1.32) deliver:

$$\frac{\partial \boldsymbol{E}}{\partial t} = \mp v \frac{\partial \boldsymbol{f}}{\partial u}$$

and

$$(\boldsymbol{\nabla} \times \boldsymbol{E})_x = \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} = \frac{\partial f_z}{\partial u} e_y - \frac{\partial f_y}{\partial u} e_z = \left(\boldsymbol{e} \times \frac{\partial \boldsymbol{f}}{\partial u} \right)_x$$

$$\Rightarrow \quad \boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{e} \times \frac{\partial \boldsymbol{f}}{\partial u}$$

An analogue expression is valid for $\boldsymbol{H}$.

Therefore, the Maxwell equations (1.1.19) and (1.1.20) deliver

$$\boldsymbol{e} \times \frac{\partial \boldsymbol{f}}{\partial u} = \pm \mu_0 \mu v \frac{\partial \boldsymbol{g}}{\partial u} \tag{1.1.34}$$

$$\boldsymbol{e} \times \frac{\partial \boldsymbol{g}}{\partial u} = \mp \epsilon_0 \epsilon v \frac{\partial \boldsymbol{f}}{\partial u} \tag{1.1.35}$$

These equations can be integrated with respect to the variable u and by setting the integration constant to zero and using equations (1.1.13), (1.1.27) and (1.1.29) the result is:

$$\boldsymbol{E} = \boldsymbol{f} = \mp \sqrt{\frac{\mu_0 \mu}{\epsilon_0 \epsilon}} \boldsymbol{e} \times \boldsymbol{H} \tag{1.1.36}$$

$$\boldsymbol{H} = \boldsymbol{g} = \pm \sqrt{\frac{\epsilon_0 \epsilon}{\mu_0 \mu}} \boldsymbol{e} \times \boldsymbol{E} \tag{1.1.37}$$

These two equations show that $\boldsymbol{E}$ is perpendicular to $\boldsymbol{e}$ and $\boldsymbol{H}$ and that $\boldsymbol{H}$ is perpendicular to $\boldsymbol{e}$ and $\boldsymbol{E}$. This can only be the case if $\boldsymbol{e}$, $\boldsymbol{E}$ and $\boldsymbol{H}$ form an orthogonal triad of vectors. Therefore, a plane wave in a homogeneous dielectric is always a transversal wave.

1.1.7 The Poynting vector of a plane wave

In this section, the physical interpretation of the Poynting vector will be illustrated for plane waves. The Poynting vector defined with equation (1.1.9) is parallel to $\boldsymbol{e}$:

$$\begin{aligned} \boldsymbol{S} &= \boldsymbol{E} \times \boldsymbol{H} = \left(\mp \sqrt{\frac{\mu_0 \mu}{\epsilon_0 \epsilon}} \boldsymbol{e} \times \boldsymbol{H} \right) \times \left(\pm \sqrt{\frac{\epsilon_0 \epsilon}{\mu_0 \mu}} \boldsymbol{e} \times \boldsymbol{E} \right) = \\ &= -\left[(\boldsymbol{e} \times \boldsymbol{H}) \cdot \boldsymbol{E} \right] \boldsymbol{e} + \left[(\boldsymbol{e} \times \boldsymbol{H}) \cdot \boldsymbol{e} \right] \boldsymbol{E} \end{aligned} \tag{1.1.38}$$

The second scalar product is zero so that only the first term remains. By using equation (1.1.36) this finally results in:

$$\boldsymbol{S} = -\left[(\boldsymbol{e} \times \boldsymbol{H}) \cdot \boldsymbol{E} \right] \boldsymbol{e} = \pm \sqrt{\frac{\epsilon_0 \epsilon}{\mu_0 \mu}} \left(\boldsymbol{E} \cdot \boldsymbol{E} \right) \boldsymbol{e} \tag{1.1.39}$$

This means that the energy transport is along the direction of propagation of the plane wave and that the absolute value of the Poynting vector, i.e. the intensity of the light wave, is proportional to $|\boldsymbol{E}|^2$.
By using the vector identity $(\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{c} = -(\boldsymbol{a} \times \boldsymbol{c}) \cdot \boldsymbol{b}$ and equation (1.1.37) we can also write:

$$\boldsymbol{S} = -\left[(\boldsymbol{e} \times \boldsymbol{H}) \cdot \boldsymbol{E} \right] \boldsymbol{e} = \left[(\boldsymbol{e} \times \boldsymbol{E}) \cdot \boldsymbol{H} \right] \boldsymbol{e} = \pm \sqrt{\frac{\mu_0 \mu}{\epsilon_0 \epsilon}} \left(\boldsymbol{H} \cdot \boldsymbol{H} \right) \boldsymbol{e} \tag{1.1.40}$$

This means that the absolute value of the Poynting vector is also proportional to $|\boldsymbol{H}|^2$ and that the equality holds:

$$\sqrt{\frac{\mu_0 \mu}{\epsilon_0 \epsilon}} |\boldsymbol{H}|^2 = \sqrt{\frac{\epsilon_0 \epsilon}{\mu_0 \mu}} |\boldsymbol{E}|^2 \quad \Rightarrow \quad \mu_0 \mu |\boldsymbol{H}|^2 = \epsilon_0 \epsilon |\boldsymbol{E}|^2 \tag{1.1.41}$$

Comparing this with the equations (1.1.16), (1.1.17) and (1.1.18) for the energy density of the electromagnetic field we have for a plane wave in a homogeneous dielectric

$$w_e = w_m = \frac{1}{2} w \quad \Rightarrow \quad w = \epsilon_0 \epsilon |\boldsymbol{E}|^2 = \mu_0 \mu |\boldsymbol{H}|^2 \tag{1.1.42}$$

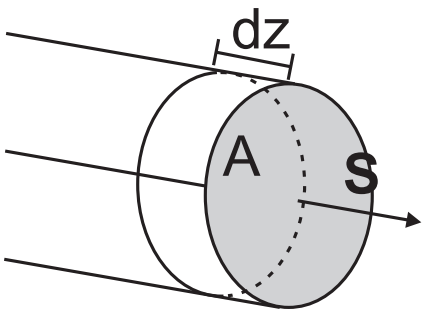


Figure 1.3: Illustration of the Poynting vector $\boldsymbol{S}$ as transporting the energy of the electromagnetic field. A is the area of the circular surface.

By using again equations (1.1.13), (1.1.27) and (1.1.29) the equations (1.1.39) and (1.1.40) can be transformed to

$$\boldsymbol{S} = \pm v\mu_0\mu \left|\boldsymbol{H}\right|^2 \boldsymbol{e} = \pm v\epsilon_0\epsilon \left|\boldsymbol{E}\right|^2 \boldsymbol{e} = \pm vw\boldsymbol{e} \tag{1.1.43}$$

This means that the absolute value of the Poynting vector is in a homogeneous dielectric the product of the energy density (energy per volume) of the electromagnetic field and the phase velocity of light. This confirms the interpretation of the Poynting vector as being the vector of the electromagnetic wave transporting the energy of the electromagnetic field with the phase velocity of light. Fig. 1.3 illustrates this. In the infinitesimal time interval dt the light covers the also infinitesimal distance $dz = v\,dt$. We assume that the distance dz is so small that the local energy density w of the electromagnetic field is constant in the volume $dV = A\,dz$, whereby A is the area of a small surface perpendicular to the Poynting vector. Therefore, all the energy $dW = w\,dV$ that is contained in the infinitesimal volume dV passes the surface area A in the time dt and we have for the intensity I (electromagnetic power per area):

$$I = \frac{dW}{A\,dt} = \frac{w\,dV}{A\,dt} = \frac{wAv\,dt}{A\,dt} = wv = |\boldsymbol{S}| \tag{1.1.44}$$

This is exactly the absolute value of the Poynting vector $\boldsymbol{S}$.
The light intensity on a surface which is not perpendicular to the direction of the Poynting vector is calculated by the equation

$$I = \boldsymbol{S} \cdot \boldsymbol{N} \tag{1.1.45}$$

whereby $\boldsymbol{N}$ is a local unit vector perpendicular to the surface.
However, we have to be a little bit careful with the interpretation that the speed of the energy transport is really the phase velocity of light. Normally, the refractive index n is larger than one and therefore the phase velocity v is smaller than the vacuum speed of light c. But, there are also cases where n is smaller than one and therefore $v > c$ (e.g. in the case of X–rays). Of course, this does not mean that the energy is transported faster than the vacuum speed of light. In fact, a plane wave is an idealization of a real wave because a plane wave would be spatially and temporarily infinitely extended. So, it would be quite impossible to measure the speed of energy transport of a plane wave. To do this, a wave packet is necessary with a finite temporal extension and the group velocity of this wave packet has to be taken.

1.1.8 A time–harmonic plane wave

Up to now a plane wave was defined with equations (1.1.30) and (1.1.31) to be $\boldsymbol{E}(\boldsymbol{r},t)=\boldsymbol{f}(u)$ and $\boldsymbol{H}(\boldsymbol{r},t)=\boldsymbol{g}(u)$. The argument u is defined by equation (1.1.32) to be $u=\boldsymbol{e}\cdot\boldsymbol{r}-vt$. This means that all points with position vector $\boldsymbol{r}$ at a fixed point t in time lie on a plane surface for a constant value u. Additionally, we saw that $\boldsymbol{e}$, $\boldsymbol{E}$ and $\boldsymbol{H}$ have to form an orthogonal triad of vectors (see equations (1.1.36) and (1.1.37)). But the concrete form of the functions $\boldsymbol{f}$ and $\boldsymbol{g}$ can be quite arbitrary to fulfill these conditions. A wave which is very important in optics because of its simple form is a time–harmonic wave. Additionally, it should be **linearly polarized**, i.e. the direction of the electric and magnetic vector are each constant. A linearly polarized time–harmonic plane wave is represented by the equations:

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E_0}\cos\left(\frac{2\pi n}{\lambda}u\right)=\boldsymbol{E_0}\cos\left(\frac{2\pi n}{\lambda}\left[\boldsymbol{e}\cdot\boldsymbol{r}-vt\right]\right) \tag{1.1.46}$$

$$\boldsymbol{H}(\boldsymbol{r},t) = \boldsymbol{H_0}\cos\left(\frac{2\pi n}{\lambda}u\right)=\boldsymbol{H_0}\cos\left(\frac{2\pi n}{\lambda}\left[\boldsymbol{e}\cdot\boldsymbol{r}-vt\right]\right) \tag{1.1.47}$$

Here, we have introduced the value λ which has the physical unit of a length so that the argument of the cosine function has no physical unit. Its meaning will be clear soon.
The characteristic property of a time–harmonic wave is that it has for a fixed point $\boldsymbol{r}$ periodically the same value after a certain time interval. The smallest time interval for which this is the case is called **period** T:

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r},t+T) &= \boldsymbol{E_0}\cos\left(\frac{2\pi n}{\lambda}\left[\boldsymbol{e}\cdot\boldsymbol{r}-v\left(t+T\right)\right]\right)= \\ &= \boldsymbol{E_0}\cos\left(\frac{2\pi n}{\lambda}\left[\boldsymbol{e}\cdot\boldsymbol{r}-vt\right]\right)=\boldsymbol{E}(\boldsymbol{r},t) \\ \Rightarrow \qquad & \frac{2\pi n}{\lambda}vT=2\pi \quad\Rightarrow\quad vT=\frac{\lambda}{n} \quad\text{or}\quad cT=\lambda\end{aligned} \tag{1.1.48}$$

Therefore, λ/n is the distance which the light covers in the material in the period T and is called the **wavelength** of the harmonic wave in the material. The quantity λ itself is the wavelength in vacuum. The reciprocal of T is called the **frequency** ν of the wave and the term $2\pi\nu=2\pi/T$ is called the **angular frequency** ω of the wave. Therefore the two following equations are valid:

$$cT=\lambda \quad\Rightarrow\quad c=\lambda\nu \tag{1.1.49}$$

$$\frac{2\pi}{\lambda}c=\frac{2\pi}{T}=2\pi\nu=\omega \tag{1.1.50}$$

Additionally, we introduce the **wave vector** $\boldsymbol{k}$ which is defined by:

$$\boldsymbol{k}=\frac{2\pi n}{\lambda}\boldsymbol{e} \tag{1.1.51}$$

Then the equations (1.1.46) and (1.1.47) for $\boldsymbol{E}$ and $\boldsymbol{H}$ can be written as:

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E_0}\cos\left(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t\right) \tag{1.1.52}$$

$$\boldsymbol{H}(\boldsymbol{r},t) = \boldsymbol{H_0}\cos\left(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t\right) \tag{1.1.53}$$

Because of the orthogonality condition $\boldsymbol{k}$ (or $\boldsymbol{e}$ which is parallel to $\boldsymbol{k}$), $\boldsymbol{E_0}$ and $\boldsymbol{H_0}$ have to form an orthogonal triad. This can be explicitly seen in this case by using Maxwell's first equation (1.1.19) in a homogeneous dielectric and the mathematical rules for the Nabla operator:

$$\begin{aligned}
&\boldsymbol{\nabla} \times \boldsymbol{E} = \boldsymbol{\nabla} \times [\boldsymbol{E_0} \cos(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)] = [\boldsymbol{\nabla} \cos(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)] \times \boldsymbol{E_0} = \\
&= -\boldsymbol{k} \times \boldsymbol{E_0} \sin(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)\,; \\
&-\mu_0 \mu \frac{\partial \boldsymbol{H}}{\partial t} = -\omega \mu_0 \mu \boldsymbol{H_0} \sin(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t) \\
&\Rightarrow \quad \omega \mu_0 \mu \boldsymbol{H_0} = \boldsymbol{k} \times \boldsymbol{E_0} \\
&\Rightarrow \quad \boldsymbol{H_0} = \frac{1}{\omega \mu_0 \mu} \boldsymbol{k} \times \boldsymbol{E_0} = \frac{\lambda}{2\pi c \mu_0 \mu} \boldsymbol{k} \times \boldsymbol{E_0} = \sqrt{\frac{\epsilon_0 \epsilon}{\mu_0 \mu}} \boldsymbol{e} \times \boldsymbol{E_0}
\end{aligned} \tag{1.1.54}$$

In the last step equations (1.1.50), (1.1.13) and (1.1.27) are used and the geometrical interpretation of the result is that $\boldsymbol{H_0}$ is perpendicular to both $\boldsymbol{e}$ and $\boldsymbol{E_0}$. The third Maxwell equation (1.1.21) delivers:

$$\begin{aligned}
\boldsymbol{\nabla} \cdot \boldsymbol{E} &= \boldsymbol{\nabla} \cdot [\boldsymbol{E_0} \cos(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)] = \boldsymbol{E_0} \cdot [\boldsymbol{\nabla} \cos(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t)] = \\
&= -\boldsymbol{E_0} \cdot \boldsymbol{k} \sin(\boldsymbol{k} \cdot \boldsymbol{r} - \omega t) = 0 \\
&\Rightarrow \quad \boldsymbol{E_0} \cdot \boldsymbol{k} = 0
\end{aligned} \tag{1.1.55}$$

This means that also $\boldsymbol{k}$ (or $\boldsymbol{e}$) and $\boldsymbol{E_0}$ are perpendicular to each other. The two other Maxwell equations (1.1.20) and (1.1.22) are automatically fulfilled because of the symmetry in $\boldsymbol{E}$ and $\boldsymbol{H}$.

1.2 The complex representation of time–harmonic waves

In section 1.1.8 a linearly polarized time–harmonic plane wave is expressed with real cosine functions for the electric and the magnetic vector. Because $\boldsymbol{E}$ and $\boldsymbol{H}$ are observable physical quantities they have of course to be expressed by real functions. The fact that usual detectors are not fast enough to detect the electric and magnetic vector of light waves directly does not matter here. Nevertheless, the calculation with a complex exponential function is more convenient than the calculation with real cosine or sine functions. Now, the Maxwell equations (1.1.2)–(1.1.5) are linear. Therefore, if the functions $\boldsymbol{E_1}$, $\boldsymbol{D_1}$, $\boldsymbol{H_1}$, $\boldsymbol{B_1}$, $\boldsymbol{j_1}$ and ρ_1 on the one hand and $\boldsymbol{E_2}$, $\boldsymbol{D_2}$, $\boldsymbol{H_2}$, $\boldsymbol{B_2}$, $\boldsymbol{j_2}$ and ρ_2 on the other are both solutions of the Maxwell equations, also a linear combination of these functions is a solution:

$$\begin{aligned}
&\boldsymbol{\nabla} \times \boldsymbol{E_1} = -\frac{\partial \boldsymbol{B_1}}{\partial t} \\
&\boldsymbol{\nabla} \times \boldsymbol{H_1} = \frac{\partial \boldsymbol{D_1}}{\partial t} + \boldsymbol{j_1} \\
&\boldsymbol{\nabla} \cdot \boldsymbol{D_1} = \rho_1 \\
&\boldsymbol{\nabla} \cdot \boldsymbol{B_1} = 0
\end{aligned}$$

$$\begin{aligned}
&\boldsymbol{\nabla} \times \boldsymbol{E_2} = -\frac{\partial \boldsymbol{B_2}}{\partial t} \\
&\boldsymbol{\nabla} \times \boldsymbol{H_2} = \frac{\partial \boldsymbol{D_2}}{\partial t} + \boldsymbol{j_2} \\
&\boldsymbol{\nabla} \cdot \boldsymbol{D_2} = \rho_2 \\
&\boldsymbol{\nabla} \cdot \boldsymbol{B_2} = 0
\end{aligned}$$

$$\Rightarrow$$

$$\boldsymbol{\nabla} \times [\alpha \boldsymbol{E_1} + \beta \boldsymbol{E_2}] = -\frac{\partial [\alpha \boldsymbol{B_1} + \beta \boldsymbol{B_2}]}{\partial t} \tag{1.2.1}$$

$$\boldsymbol{\nabla} \times [\alpha \boldsymbol{H_1} + \beta \boldsymbol{H_2}] = \frac{\partial [\alpha \boldsymbol{D_1} + \beta \boldsymbol{D_2}]}{\partial t} + \alpha \boldsymbol{j_1} + \beta \boldsymbol{j_2} \tag{1.2.2}$$

$$\boldsymbol{\nabla} \cdot [\alpha \boldsymbol{D_1} + \beta \boldsymbol{D_2}] = \alpha \rho_1 + \beta \rho_2 \tag{1.2.3}$$

$$\boldsymbol{\nabla} \cdot [\alpha \boldsymbol{B_1} + \beta \boldsymbol{B_2}] = 0 \tag{1.2.4}$$

α and β are arbitrary real or complex constants.
The Euler equation delivers:

$$e^{ix} = \cos x + i \sin x \tag{1.2.5}$$

or

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} = \frac{e^{ix} + \left(e^{ix}\right)^*}{2} \tag{1.2.6}$$

Due to the linearity of the Maxwell equations it is obvious that if a function containing a cosine function is a solution of the Maxwell equations the replacement of the cosine function by a complex exponential function will then also be a solution. Therefore, it is quite normal that waves are expressed by using a complex function although only the real part of this function represents the real physical quantity. The addition, subtraction, integration and differentiation of such a complex function is also a linear operation, so that we can build at the end the real part and have the real solution:

$$\begin{array}{l} z_1(x) = a_1(x) + i b_1(x) \\ z_2(x) = a_2(x) + i b_2(x) \end{array} \quad \Rightarrow \quad \begin{array}{l} \mathrm{Re}\{z_1 + z_2\} = \mathrm{Re}\{z_1\} + \mathrm{Re}\{z_2\} \\ \mathrm{Re}\{z_1 - z_2\} = \mathrm{Re}\{z_1\} - \mathrm{Re}\{z_2\} \\ \mathrm{Re}\left\{\frac{dz_1}{dx}\right\} = \frac{d}{dx}\mathrm{Re}\{z_1\} \\ \mathrm{Re}\left\{\int z_1 dx\right\} = \int \mathrm{Re}\{z_1\}\, dx \\ \mathrm{Re}\{f z_1\} = f \mathrm{Re}\{z_1\} \end{array} \tag{1.2.7}$$

Here, f is an arbitrary real function or constant. Only if two complex functions have to be multiplied or divided or the absolute value has to be built we have to be careful:

$$\begin{array}{l} \mathrm{Re}\{z_1 z_2\} = a_1 a_2 - b_1 b_2 \neq a_1 a_2 = \mathrm{Re}\{z_1\}\,\mathrm{Re}\{z_2\} \\ \mathrm{Re}\left\{\frac{z_1}{z_2}\right\} = \frac{a_1 a_2 + b_1 b_2}{a_2^2 + b_2^2} \neq \frac{a_1}{a_2} = \frac{\mathrm{Re}\{z_1\}}{\mathrm{Re}\{z_2\}} \\ \mathrm{Re}\{z_1 z_1\} = a_1^2 - b_1^2 \neq a_1^2 + b_1^2 = |z_1|^2 \end{array} \tag{1.2.8}$$

So, if the Poynting vector or products of the electric or magnetic vectors have to be calculated it is not allowed to just take the complex functions. Nevertheless, there are some useful applications of the complex notation. As we mentioned before the frequency of a light wave is so high that no usual detector can directly measure the vibrations. For a typical wavelength of visible light of 500 nm the frequency of a wave in vacuum is according to equation (1.1.49):

$$\nu = \frac{c}{\lambda} = \frac{2.998 \cdot 10^8 \mathrm{m\ s^{-1}}}{5.0 \cdot 10^{-7} \mathrm{m}} = 5.996 \cdot 10^{14} \mathrm{s^{-1}}$$

$$\Rightarrow \quad T = \frac{1}{\nu} = 1.668 \cdot 10^{-15} \mathrm{s} = 1.668\ \mathrm{fs}$$

So, the period is just a little bit more than a femtosecond. This means, that in most cases only the time average of the light intensity over many periods is measured.
A general time–harmonic wave with the angular frequency ω has the representation:

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{r},t) &= \begin{pmatrix} A_x(\boldsymbol{r})\cos(\Phi_x(\boldsymbol{r})-\omega t) \\ A_y(\boldsymbol{r})\cos(\Phi_y(\boldsymbol{r})-\omega t) \\ A_z(\boldsymbol{r})\cos(\Phi_z(\boldsymbol{r})-\omega t) \end{pmatrix} = \mathrm{Re}\left\{ \begin{pmatrix} A_x(\boldsymbol{r})\,e^{i\Phi_x(\boldsymbol{r})-i\omega t} \\ A_y(\boldsymbol{r})\,e^{i\Phi_y(\boldsymbol{r})-i\omega t} \\ A_z(\boldsymbol{r})\,e^{i\Phi_z(\boldsymbol{r})-i\omega t} \end{pmatrix} \right\} = \\ &= \mathrm{Re}\left\{ e^{-i\omega t} \begin{pmatrix} A_x(\boldsymbol{r})\,e^{i\Phi_x(\boldsymbol{r})} \\ A_y(\boldsymbol{r})\,e^{i\Phi_y(\boldsymbol{r})} \\ A_z(\boldsymbol{r})\,e^{i\Phi_z(\boldsymbol{r})} \end{pmatrix} \right\} =: \mathrm{Re}\left\{ e^{-i\omega t}\hat{\boldsymbol{E}}(\boldsymbol{r}) \right\} \end{aligned} \tag{1.2.9}$$

A_x, A_y, A_z, Φ_x, Φ_y and Φ_z are all real functions which depend only on the position $\boldsymbol{r}$. Additionally, A_x, A_y and A_z which are called the components of the **amplitude** are slowly varying functions of the position. On the other hand, the exponential terms with the components of the **phase** Φ_x, Φ_y and Φ_z are rapidly varying functions of the position. The components of the complex vector $\hat{\boldsymbol{E}}(\boldsymbol{r})$ are often called the **complex amplitudes** of the electric vector of the wave.
Equation (1.1.39) gives the relation between the Poynting vector and the electric vector of a plane wave in a homogeneous dielectric. Without proof, we can assume that this relation is also valid for a general time–harmonic wave as long as it deviates not too far from a plane wave (of course it is for example not valid in the focus of a high numerical aperture lens as is later shown):

$$\boldsymbol{S} = \sqrt{\frac{\epsilon_0\epsilon}{\mu_0\mu}}\,(\boldsymbol{E}\cdot\boldsymbol{E})\,\boldsymbol{e}$$

Now, the time average $\bar{S}$ of the absolute value of the Poynting vector, i.e. the intensity which is really measured by a common detector, will be calculated for the general time–harmonic wave. Therefore, we have to integrate the absolute value S of the Poynting vector over one period T and divide it by T:

$$\bar{S}(\boldsymbol{r}) := \frac{1}{T}\int_0^T |\boldsymbol{S}(\boldsymbol{r},t)|\,dt = \sqrt{\frac{\epsilon_0\epsilon}{\mu_0\mu}}\frac{1}{T}\int_0^T \boldsymbol{E}(\boldsymbol{r},t)\cdot\boldsymbol{E}(\boldsymbol{r},t)\,dt \tag{1.2.10}$$

Using equation (1.2.9) for a general time–harmonic wave and equation (1.1.13) we obtain:

$$\begin{aligned}\bar{S} &= \frac{\epsilon_0 c}{T}\sqrt{\frac{\epsilon}{\mu}}\int_0^T \left[A_x^2\cos^2(\Phi_x-\omega t)+A_y^2\cos^2(\Phi_y-\omega t)+A_z^2\cos^2(\Phi_z-\omega t)\right]dt = \\ &= \sqrt{\frac{\epsilon}{\mu}}\frac{\epsilon_0 c}{2}\left[A_x^2+A_y^2+A_z^2\right] \end{aligned} \tag{1.2.11}$$

But, if we calculate directly the square of the absolute value of the time–independent vector $\hat{\boldsymbol{E}}$ we also obtain:

$$\left|\hat{\boldsymbol{E}}\right|^2 = \hat{\boldsymbol{E}}\cdot\hat{\boldsymbol{E}}^* = A_x^2+A_y^2+A_z^2 \tag{1.2.12}$$

By combining equations (1.2.11) and (1.2.12) we finally obtain:

$$\bar{S}(\boldsymbol{r}) = \sqrt{\frac{\epsilon}{\mu}}\frac{\epsilon_0 c}{2}\left|\hat{\boldsymbol{E}}(\boldsymbol{r})\right|^2 = \sqrt{\frac{\epsilon}{\mu}}\frac{\epsilon_0 c}{2}\hat{\boldsymbol{E}}(\boldsymbol{r})\cdot\left(\hat{\boldsymbol{E}}(\boldsymbol{r})\right)^* \tag{1.2.13}$$

Therefore, the complex representation of time–harmonic waves allows a fast calculation of the time average of the Poynting vector, i.e. of the intensity of the light wave.

1.2.1 Time–averaged Poynting vector for general time–harmonic waves with complex representation

In the derivation of above we have used equation (1.1.39) which was strictly derived only for a plane wave. For a real general time–harmonic wave (for example also in the focus of a high numerical aperture lens) we can nevertheless derive an equation for the time–averaged Poynting vector using the electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{H}$. Again, we assume that we have both fields in the complex representation for a time–harmonic wave. Then, we have:

$$\boldsymbol{E}(\boldsymbol{r},t) = \mathrm{Re}\left\{e^{-i\omega t}\hat{\boldsymbol{E}}(\boldsymbol{r})\right\} \tag{1.2.14}$$

$$\boldsymbol{H}(\boldsymbol{r},t) = \mathrm{Re}\left\{e^{-i\omega t}\hat{\boldsymbol{H}}(\boldsymbol{r})\right\} \tag{1.2.15}$$

Here, $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{H}}$ are the complex electric and magnetic vectors of a time–harmonic wave which depend only on the position $\boldsymbol{r}$ in space.
Then, the time–averaged real Poynting vector $\bar{\boldsymbol{S}}$ can be written:

$$\begin{aligned}
\bar{\boldsymbol{S}}(\boldsymbol{r}) &= \frac{1}{T}\int_0^T \boldsymbol{S}(\boldsymbol{r},t)dt = \frac{1}{T}\int_0^T \mathrm{Re}\left\{e^{-i\omega t}\hat{\boldsymbol{E}}(\boldsymbol{r})\right\}\times\mathrm{Re}\left\{e^{-i\omega t}\hat{\boldsymbol{H}}(\boldsymbol{r})\right\}dt = \\
&= \frac{1}{T}\int_0^T \left[\mathrm{Re}\left\{\hat{\boldsymbol{E}}\right\}\cos(\omega t) + \mathrm{Im}\left\{\hat{\boldsymbol{E}}\right\}\sin(\omega t)\right]\times\left[\mathrm{Re}\left\{\hat{\boldsymbol{H}}\right\}\cos(\omega t) + \mathrm{Im}\left\{\hat{\boldsymbol{H}}\right\}\sin(\omega t)\right]dt = \\
&= \frac{1}{2}\left[\mathrm{Re}\left\{\hat{\boldsymbol{E}}\right\}\times\mathrm{Re}\left\{\hat{\boldsymbol{H}}\right\} + \mathrm{Im}\left\{\hat{\boldsymbol{E}}\right\}\times\mathrm{Im}\left\{\hat{\boldsymbol{H}}\right\}\right] = \frac{1}{2}\mathrm{Re}\left\{\hat{\boldsymbol{E}}\times\hat{\boldsymbol{H}}^*\right\}
\end{aligned} \tag{1.2.16}$$

This definition of the time–averaged Poynting vector is valid for each time–harmonic wave. This means that it is for example also valid in the focus of a lens with a high numerical aperture.

1.3 Material equations

In the last two sections we concentrated often on the electromagnetic field in an isotropic and homogeneous dielectric material where the Maxwell equations are simplified to (1.1.19)–(1.1.22). In other materials the general Maxwell equations (1.1.2)–(1.1.5) have to be used and more complex interrelations between the electric displacement and the electric vector on the one hand and the magnetic induction and the magnetic vector on the other have to be found. Since the atomic distances are small compared to the wavelength of light a macroscopic description with smooth functions is possible. To calculate the influence of the material, first of all the interrelations between $\boldsymbol{D}$ and $\boldsymbol{E}$ on the one hand and $\boldsymbol{B}$ and $\boldsymbol{H}$ on the other are considered

in vacuum. These equations in vacuum are obtained from (1.1.12) for the case $\mu = \epsilon = 1$. Then, additional terms are added to the equations in vacuum. The **electric polarization** (*dt.:* Polarisierung) $\boldsymbol{P}$ and the **magnetization** (*dt.:* Magnetisierung) $\boldsymbol{M}$ are introduced by:

$$\boldsymbol{D}(\boldsymbol{r},t) = \epsilon_0 \boldsymbol{E}(\boldsymbol{r},t) + \boldsymbol{P}(\boldsymbol{r},t) \tag{1.3.1}$$

$$\boldsymbol{B}(\boldsymbol{r},t) = \mu_0 \boldsymbol{H}(\boldsymbol{r},t) + \boldsymbol{M}(\boldsymbol{r},t) \tag{1.3.2}$$

The atomic theory goes far beyond our scope. But in a macroscopic theory the effect of the atoms (i.e. mainly the electrons of the atoms) on the electric polarization is that it is a function of the electric vector. In the same way the magnetization is a function of the magnetic vector. The most general equations are:

$$\begin{aligned} P_i(\boldsymbol{r},t) &= P_0(\boldsymbol{r},t) + \epsilon_0 \sum_{j=1}^{3} \eta_{ij}^{(1)}(\boldsymbol{r},t)\, E_j(\boldsymbol{r},t) + \\ &+ \sum_{j=1}^{3}\sum_{k=1}^{3} \eta_{ijk}^{(2)}(\boldsymbol{r},t)\, E_j(\boldsymbol{r},t)\, E_k(\boldsymbol{r},t) + \\ &+ \sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} \eta_{ijkl}^{(3)}(\boldsymbol{r},t)\, E_j(\boldsymbol{r},t)\, E_k(\boldsymbol{r},t)\, E_l(\boldsymbol{r},t) + \ldots \end{aligned} \tag{1.3.3}$$

$$\begin{aligned} M_i(\boldsymbol{r},t) &= M_0(\boldsymbol{r},t) + \mu_0 \sum_{j=1}^{3} \chi_{ij}^{(1)}(\boldsymbol{r},t)\, H_j(\boldsymbol{r},t) + \\ &+ \sum_{j=1}^{3}\sum_{k=1}^{3} \chi_{ijk}^{(2)}(\boldsymbol{r},t)\, H_j(\boldsymbol{r},t)\, H_k(\boldsymbol{r},t) + \\ &+ \sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} \chi_{ijkl}^{(3)}(\boldsymbol{r},t)\, H_j(\boldsymbol{r},t)\, H_k(\boldsymbol{r},t)\, H_l(\boldsymbol{r},t) + \ldots \end{aligned} \tag{1.3.4}$$

There, the lower indices running from 1 to 3 indicate the components of the respective electromagnetic vectors, e.g. $E_1 := E_x$, $E_2 := E_y$ and $E_3 := E_z$. The tensor functions $\eta_{ij}^{(1)}$, $\eta_{ijk}^{(2)}$ and so on describe the influence of the electric vector on the electric polarization and the same is valid for the tensor functions $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$ and so on in the magnetic case. The tensor functions are defined here in such a way that $\eta_{ij}^{(1)}$ and $\chi_{ij}^{(1)}$ have no physical unit and are pure numbers. Nevertheless, the tensor functions of higher degree have here different physical units. The tensors $\eta_{ij}^{(1)}$, $\eta_{ijk}^{(2)}$ and so on are called the tensors of the **dielectric susceptibility** (*dt.:* elektrische Suszeptibilität). The tensors $\chi_{ij}^{(1)}$, $\chi_{ijk}^{(2)}$ and so on are called the tensors of the **magnetic susceptibility** (*dt.:* magnetische Suszeptibilität). In equation (1.3.3) also a bias P_0 for the polarization is assumed and the same is made for the magnetization. In our general material equations the different terms can depend on the position $\boldsymbol{r}$ as well as on the time t. But in most cases the material functions will not depend explicitly on the time t.

Additionally, there have to be equations for the current density $\boldsymbol{j}$ and the charge density ρ. In optics the most important materials are either **dielectrics** (*dt.:* Dielektrika) or metals (e.g. for mirrors). In both cases we can assume $\rho = 0$. For the current density we can in most cases take the equation:

$$\boldsymbol{j} = \sigma \boldsymbol{E} \tag{1.3.5}$$

The **conductivity** (*dt.:* Leitfähigkeit) σ indicates how good an electric current is conducted in a material and has the physical unit $[\sigma] = 1\ \mathrm{A\ V^{-1}\ m^{-1}}$. For ideal dielectric materials σ is zero so that we obtain $\boldsymbol{j} = 0$. In this case the material does not absorb light. For metals σ is of course not zero and for an ideal conductor it would become infinity, so that all light would be absorbed or reflected at once. There are also anisotropic absorbing materials like special crystals where σ is not a scalar but a tensor [1]. But this is out of our scope.

1.3.1 Discussion of the general material equations

1.3.1.1 Polarization

The term $\sum_{j=1}^{3} \eta_{ij}^{(1)} E_j$ in equation (1.3.3) is responsible for **linear** responses of the electric polarization on the electric vector and is the most important effect. The following terms and the bias term are responsible for so called **nonlinear** effects and are subject of the **nonlinear optics** [2] (e.g. second harmonic generation or self focusing effects). In the following the bias and all tensors with upper index (2) and more of the dielectric susceptibility $\eta_{ijk}^{(2)}$, $\eta_{ijkl}^{(3)}$, ... will be set to zero because only the **linear optics** will be treated in this lecture. In "normal" materials like different glass types the linear case is the normal case. Only if the absolute value of the electric vector of the electromagnetic field is in the range of the atomic electric field nonlinear effects occur in these materials.
An estimation of the electric fields in atoms and in a light wave is helpful. In a typical atom the electric field on an outer electron can be estimated by applying Coulomb's law and assuming an effective charge of the nucleus of one elementary charge and a distance of the electron of $r = 10^{-10}\mathrm{m}$:

$$E = |\boldsymbol{E}| = \frac{e}{4\pi\epsilon_0 r^2} \tag{1.3.6}$$

$$r = 10^{-10}\mathrm{m}, e = 1.6022 \cdot 10^{-19}\mathrm{A\ s}, \epsilon_0 = 8.8542 \cdot 10^{-12}\mathrm{A\ s\ V^{-1}m^{-1}}$$

$$\Rightarrow \quad E \approx 1.4 \cdot 10^{11}\mathrm{V/m}$$

The electric field oscillates very rapidly in a light wave. Therefore, to estimate the electric field in a light wave (here in vacuum) the amplitude $\left|\hat{E}\right|$ of the modulus of the time–independent complex–valued electric field is calculated. This can be done by using equation (1.2.13) for the relation between the time average $\bar{S}$ of the modulus of the Poynting vector and the modulus of the time–independent complex–valued electric field. Here, the values are calculated in vacuum ($\epsilon = \mu = 1$):

$$\bar{S} = \frac{c\epsilon_0}{2}\left|\hat{E}\right|^2 \quad \Rightarrow \quad \left|\hat{E}\right| = \sqrt{\frac{2\bar{S}}{c\epsilon_0}} \tag{1.3.7}$$

The result for the electric field of the light on a sunny day is:

$$\bar{S} \approx 1\ \mathrm{kW/m^2} \quad \Rightarrow \left|\hat{E}\right| \approx 868\ \mathrm{V/m}$$

In the focused spot of a medium power continuous wave (cw) laser beam we have e.g.:

$$\bar{S} = 1\ \mathrm{W}/\mu\mathrm{m}^2 = 10^{12}\mathrm{W/m^2} \quad \Rightarrow \left|\hat{E}\right| \approx 2.74 \cdot 10^7\mathrm{V/m}$$

This shows that in normal materials and with "normal" light intensities the electric field of a light wave is quite small compared to the electric field of the atoms. Therefore, the electrons are

only moved a little bit and this results normally in a linear response of the dielectric function to the exciting electric field of the light wave. Of course, there are also so called nonlinear materials which show for smaller electric fields nonlinear effects. In addition, ultra–short pulsed lasers, e.g. so called femtosecond lasers, can achieve much higher intensities in their focus so that electric fields which are comparable to or higher than the electric field in atoms result. Then the response is of course nonlinear.

1.3.1.2 Magnetization

In practice, there are nearly no materials relevant to optics which show nonlinear magnetic effects, so that $\chi^{(2)}_{ijk}$ and all higher order tensors are zero. In fact, most optically interesting materials are non–magnetic at all, so that the remaining tensor of the magnetic susceptibility $\chi^{(1)}_{ij}$ is also zero. In some materials the magnetic susceptibility $\chi^{(1)}_{ij}$ is not zero but it can be written as a scalar constant χ times a 3x3 unit matrix. χ is a negative constant for diamagnetic materials or a positive constant for paramagnetic materials. The **magnetic permeability** μ of the material, which is a pure real number without a physical unit, is then defined as:

$$\mu := 1 + \chi \tag{1.3.8}$$

Then we have due to the equations (1.3.2) and (1.3.4):

$$\boldsymbol{B} = \mu\mu_0\boldsymbol{H} \tag{1.3.9}$$

This equation, which is also used in (1.1.12) will be used in the rest of this script and in many cases μ will be really a constant that does not depend on the position $\boldsymbol{r}$.

1.3.2 Specialization to the equations of linear and non–magnetic materials

For linear materials only the tensor of lowest degree of the dielectric susceptibility $\eta^{(1)}_{ij}$ is different from zero. Then, it can be expressed as a matrix

$$\begin{pmatrix} \eta^{(1)}_{11} & \eta^{(1)}_{12} & \eta^{(1)}_{13} \\ \eta^{(1)}_{21} & \eta^{(1)}_{22} & \eta^{(1)}_{23} \\ \eta^{(1)}_{31} & \eta^{(1)}_{32} & \eta^{(1)}_{33} \end{pmatrix} \tag{1.3.10}$$

The **dielectric tensor** is defined as

$$\begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix} := \begin{pmatrix} 1+\eta^{(1)}_{11} & \eta^{(1)}_{12} & \eta^{(1)}_{13} \\ \eta^{(1)}_{21} & 1+\eta^{(1)}_{22} & \eta^{(1)}_{23} \\ \eta^{(1)}_{31} & \eta^{(1)}_{32} & 1+\eta^{(1)}_{33} \end{pmatrix} \tag{1.3.11}$$

Using equations (1.3.1) and (1.3.3) the relation between the dielectric displacement and the electric vector is:

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \epsilon_0 \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \epsilon_{13} \\ \epsilon_{21} & \epsilon_{22} & \epsilon_{23} \\ \epsilon_{31} & \epsilon_{32} & \epsilon_{33} \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \tag{1.3.12}$$

In anisotropic materials like non–cubic crystals or originally isotropic materials that are subject to mechanical stresses the dielectric tensor has this general matrix form and the effects which

occur are e.g. birefringence [1][3]. It can be shown that the dielectric tensor is symmetric, i.e. $\epsilon_{ij} = \epsilon_{ji}$. But the treatment of anisotropic materials is out of the scope of this lecture so that we will have in the following only isotropic materials. Then the dielectric tensor reduces to a scalar ϵ times a unit matrix whereby ϵ is in general a function of the position $\boldsymbol{r}$ (and of the wavelength of the light).

1.3.3 Material equations for linear and isotropic materials

If the material is isotropic the dielectric tensor and all other material quantities are scalars times a unit matrix. Due to equations (1.3.12) and (1.3.9) we have in this case the well–known relations between the electric displacement and the electric vector on the one hand and between the magnetic induction and the magnetic vector on the other, which we also used in equation (1.1.12):

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{r},t) &= \epsilon_0 \epsilon(\boldsymbol{r})\, \boldsymbol{E}(\boldsymbol{r},t) \\ \boldsymbol{B}(\boldsymbol{r},t) &= \mu_0 \mu(\boldsymbol{r})\, \boldsymbol{H}(\boldsymbol{r},t) \end{aligned} \tag{1.3.13}$$

$\epsilon(\boldsymbol{r})$, $\mu(\boldsymbol{r})$ means that the material functions will in general depend on the position. An explicit dependence on the time is mostly not the case so that it is omitted here.
Additionally, we assume that the charge density is zero and equation (1.3.5) is valid:

$$\rho = 0 \tag{1.3.14}$$

$$\boldsymbol{j}(\boldsymbol{r},t) = \sigma(\boldsymbol{r})\, \boldsymbol{E}(\boldsymbol{r},t) \tag{1.3.15}$$

If ϵ, μ and σ are constant, i.e. independent of the position, the material is called **homogeneous**. Due to the dispersion theory which will not be treated here, the material functions will in general depend on the frequency of the stimulating electric or magnetic fields. Therefore, the Fourier components of the electric and magnetic field have to be calculated and treated separately. The electric vector and the electric displacement are written as Fourier integrals, i.e. as superposition of time–harmonic waves with the angular frequency ω:

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{r},t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \tilde{\boldsymbol{E}}(\boldsymbol{r},\omega)\, e^{-i\omega t} d\omega \\ \boldsymbol{D}(\boldsymbol{r},t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \tilde{\boldsymbol{D}}(\boldsymbol{r},\omega)\, e^{-i\omega t} d\omega \end{aligned} \tag{1.3.16}$$

The magnetic vector and the magnetic induction are treated in the same way so that we can omit this. If the function $\boldsymbol{E}$ is given $\tilde{\boldsymbol{E}}$ is calculated by:

$$\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \boldsymbol{E}(\boldsymbol{r},t)\, e^{i\omega t} dt \tag{1.3.17}$$

Since $\boldsymbol{E}$ is a real function the complex Fourier components have to fulfill the condition:

$$\tilde{\boldsymbol{E}}(\boldsymbol{r},-\omega) = \tilde{\boldsymbol{E}}^*(\boldsymbol{r},\omega) \tag{1.3.18}$$

The same is valid for the electric displacement, the magnetic vector and the magnetic induction.

Totally, the material equations can be written for isotropic and linear materials with the Fourier components of the four electromagnetic vector quantities:

$$\begin{aligned} \tilde{\boldsymbol{D}}(\boldsymbol{r},\omega) &= \epsilon_0\epsilon(\boldsymbol{r},\omega)\,\tilde{\boldsymbol{E}}(\boldsymbol{r},\omega) \\ \tilde{\boldsymbol{B}}(\boldsymbol{r},\omega) &= \mu_0\mu(\boldsymbol{r},\omega)\,\tilde{\boldsymbol{H}}(\boldsymbol{r},\omega)\ , \end{aligned} \tag{1.3.19}$$

In the following the tilde on the different quantities will be mostly omitted to simplify the notation. This is equivalent to just deal with time–harmonic waves of a certain angular frequency ω, where the quantities ϵ and μ are functions of ω.

1.4 The wave equations

The Maxwell equations (1.1.2)–(1.1.5) contain the five vector fields $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, $\boldsymbol{B}$ and $\boldsymbol{j}$ and the scalar field ρ. These quantities are related to each other by the material equations. Here, only the case of isotropic, linear and uncharged ($\rho = 0$) materials will be treated. Additionally, all material parameters like ϵ, μ and σ will be independent of the time t, but functions of the position $\boldsymbol{r}$ (and the frequency or wavelength of the light). In the following the explicit dependence of the functions on $\boldsymbol{r}$ and t will be omitted but there are the following functionalities: $\boldsymbol{E}(\boldsymbol{r},t)$, $\boldsymbol{H}(\boldsymbol{r},t)$, $\mu(\boldsymbol{r})$, $\epsilon(\boldsymbol{r})$, $\sigma(\boldsymbol{r})$.
Using equations (1.3.13) and (1.3.15) for this case results in the following specialized Maxwell equations:

$$\boldsymbol{\nabla}\times\boldsymbol{E} = -\mu_0\mu\frac{\partial\boldsymbol{H}}{\partial t} \tag{1.4.1}$$

$$\boldsymbol{\nabla}\times\boldsymbol{H} = \epsilon_0\epsilon\frac{\partial\boldsymbol{E}}{\partial t} + \sigma\boldsymbol{E} \tag{1.4.2}$$

$$\boldsymbol{\nabla}\cdot[\epsilon\boldsymbol{E}] = 0 \tag{1.4.3}$$

$$\boldsymbol{\nabla}\cdot[\mu\boldsymbol{H}] = 0 \tag{1.4.4}$$

These equations contain for given material functions ϵ, μ and σ now only the electric and the magnetic vector. To eliminate the magnetic vector the cross product of the Nabla operator with equation (1.4.1) is taken:

$$\boldsymbol{\nabla}\times(\boldsymbol{\nabla}\times\boldsymbol{E}) = -\mu_0\boldsymbol{\nabla}\times\left[\mu\frac{\partial\boldsymbol{H}}{\partial t}\right] = -\mu_0\mu\frac{\partial}{\partial t}(\boldsymbol{\nabla}\times\boldsymbol{H}) - \mu_0(\boldsymbol{\nabla}\mu)\times\frac{\partial\boldsymbol{H}}{\partial t} \tag{1.4.5}$$

By using equations (1.4.1), (1.4.2) and the Nabla operator calculus for a double cross product this results in:

$$\boldsymbol{\nabla}(\boldsymbol{\nabla}\cdot\boldsymbol{E}) - \Delta\boldsymbol{E} = -\epsilon_0\mu_0\epsilon\mu\frac{\partial^2\boldsymbol{E}}{\partial t^2} - \mu_0\mu\sigma\frac{\partial\boldsymbol{E}}{\partial t} + (\boldsymbol{\nabla}(\ln\mu))\times(\boldsymbol{\nabla}\times\boldsymbol{E}) \tag{1.4.6}$$

Here, $\Delta = \boldsymbol{\nabla}\cdot\boldsymbol{\nabla}$ is the Laplacian operator which has to be applied on each component of $\boldsymbol{E}$. Equation (1.4.3) can be used to eliminate the term $\boldsymbol{\nabla}\cdot\boldsymbol{E}$ from equation (1.4.6).

$$\begin{aligned} &\boldsymbol{\nabla}\cdot[\epsilon\boldsymbol{E}] = \boldsymbol{E}\cdot\boldsymbol{\nabla}\epsilon + \epsilon\boldsymbol{\nabla}\cdot\boldsymbol{E} = 0 \\ \Rightarrow\quad &\boldsymbol{\nabla}\cdot\boldsymbol{E} = -\frac{1}{\epsilon}\boldsymbol{E}\cdot\boldsymbol{\nabla}\epsilon = -\boldsymbol{E}\cdot\boldsymbol{\nabla}(\ln\epsilon) \end{aligned}$$

So, equation (1.4.6) gives the final so called **wave equation** for the electric vector $\boldsymbol{E}$:

$$\Delta \boldsymbol{E} + \boldsymbol{\nabla}\left[\boldsymbol{E}\cdot\boldsymbol{\nabla}\left(\ln\epsilon\right)\right] - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \mu_0\mu\sigma\frac{\partial \boldsymbol{E}}{\partial t} + \left(\boldsymbol{\nabla}\left(\ln\mu\right)\right)\times\left(\boldsymbol{\nabla}\times\boldsymbol{E}\right) = 0 \tag{1.4.7}$$

Additionally, equation (1.1.13) was used to replace $\epsilon_0\mu_0$ by $1/c^2$.
An analogue equation for the magnetic vector can be found using equations (1.4.1), (1.4.2) and (1.4.4):

$$\begin{aligned}
&\boldsymbol{\nabla}\times\left(\boldsymbol{\nabla}\times\boldsymbol{H}\right) = \boldsymbol{\nabla}\underbrace{\left(\boldsymbol{\nabla}\cdot\boldsymbol{H}\right)}_{=-\boldsymbol{H}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)} - \Delta\boldsymbol{H} = -\boldsymbol{\nabla}\left[\boldsymbol{H}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)\right] - \Delta\boldsymbol{H} = \\
&= \epsilon_0\boldsymbol{\nabla}\times\left[\epsilon\frac{\partial \boldsymbol{E}}{\partial t}\right] + \boldsymbol{\nabla}\times\left[\sigma\boldsymbol{E}\right] = \\
&= \epsilon_0\epsilon\boldsymbol{\nabla}\times\frac{\partial \boldsymbol{E}}{\partial t} + \epsilon_0\left(\boldsymbol{\nabla}\epsilon\right)\times\frac{\partial \boldsymbol{E}}{\partial t} + \sigma\boldsymbol{\nabla}\times\boldsymbol{E} + \left(\boldsymbol{\nabla}\sigma\right)\times\boldsymbol{E} = \\
&= -\epsilon_0\mu_0\epsilon\mu\frac{\partial^2 \boldsymbol{H}}{\partial t^2} + \epsilon_0\left(\boldsymbol{\nabla}\epsilon\right)\times\frac{\partial \boldsymbol{E}}{\partial t} - \mu_0\mu\sigma\frac{\partial \boldsymbol{H}}{\partial t} + \left(\boldsymbol{\nabla}\sigma\right)\times\boldsymbol{E}
\end{aligned} \tag{1.4.8}$$

Using again equation (1.4.2) this equation can be resolved with respect to $\partial\boldsymbol{E}/\partial t$:

$$\boldsymbol{\nabla}\times\boldsymbol{H} = \epsilon_0\epsilon\frac{\partial \boldsymbol{E}}{\partial t} + \sigma\boldsymbol{E} \quad \Rightarrow \quad \frac{\partial \boldsymbol{E}}{\partial t} = \frac{1}{\epsilon_0\epsilon}\left[\boldsymbol{\nabla}\times\boldsymbol{H} - \sigma\boldsymbol{E}\right]$$

Then, $\partial\boldsymbol{E}/\partial t$ can be eliminated in equation (1.4.8) resulting in:

$$\begin{aligned}
&\Delta\boldsymbol{H} + \boldsymbol{\nabla}\left[\boldsymbol{H}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)\right] - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} - \mu_0\mu\sigma\frac{\partial \boldsymbol{H}}{\partial t} + \left(\boldsymbol{\nabla}\left(\ln\epsilon\right)\right)\times\left(\boldsymbol{\nabla}\times\boldsymbol{H}\right) + \\
&+ \left[\boldsymbol{\nabla}\sigma - \sigma\boldsymbol{\nabla}\left(\ln\epsilon\right)\right]\times\boldsymbol{E} = 0
\end{aligned} \tag{1.4.9}$$

Unfortunately, it is not possible to completely eliminate the electric vector from this wave equation of the magnetic field.
Equations (1.4.7) and (1.4.9) are nearly symmetrical for a replacement of $\boldsymbol{E}$ with $\boldsymbol{H}$ and ϵ with μ. Only the terms containing the conductivity σ are not symmetrical. Nevertheless, there are two important special cases which provide symmetries of the wave equations of the electric and magnetic vector.

1.4.1 Wave equations for pure dielectrics

If the material is a pure dielectric the conductivity σ which is responsible for absorption is zero. Then equations (1.4.7) and (1.4.9) reduce to:

$$\Delta\boldsymbol{E} + \boldsymbol{\nabla}\left[\boldsymbol{E}\cdot\boldsymbol{\nabla}\left(\ln\epsilon\right)\right] - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \left(\boldsymbol{\nabla}\left(\ln\mu\right)\right)\times\left(\boldsymbol{\nabla}\times\boldsymbol{E}\right) = 0 \tag{1.4.10}$$

$$\Delta\boldsymbol{H} + \boldsymbol{\nabla}\left[\boldsymbol{H}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)\right] - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} + \left(\boldsymbol{\nabla}\left(\ln\epsilon\right)\right)\times\left(\boldsymbol{\nabla}\times\boldsymbol{H}\right) = 0 \tag{1.4.11}$$

Here, the equations are really symmetrical for a replacement of $\boldsymbol{E}$ with $\boldsymbol{H}$ and ϵ with μ. In practice, there are of course no materials which are completely transparent to light. But, in the visible or infrared region most glasses can be assumed to be dielectrics with a very good approximation.

1.4.2 Wave equations for homogeneous materials

The second interesting special case is for homogeneous materials. Then, ϵ, μ and σ are constants which do not depend on $\boldsymbol{r}$ and their gradients are zero. In this case equations (1.4.7) and (1.4.9) reduce to:

$$\Delta \boldsymbol{E} - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} - \mu_0\mu\sigma\frac{\partial \boldsymbol{E}}{\partial t} = 0 \tag{1.4.12}$$

$$\Delta \boldsymbol{H} - \frac{\epsilon\mu}{c^2}\frac{\partial^2 \boldsymbol{H}}{\partial t^2} - \mu_0\mu\sigma\frac{\partial \boldsymbol{H}}{\partial t} = 0 \tag{1.4.13}$$

These equations are symmetrical to a replacement of $\boldsymbol{E}$ with $\boldsymbol{H}$. In practice, homogeneous materials are the most important case because all conventional lenses (with the exception of **gr**aded **in**dex lenses, shortly called GRIN lenses) are made of homogeneous glasses or at least of glasses with a very small inhomogeneity.

1.5 The Helmholtz equations

Assume a time–harmonic wave with angular frequency ω represented in its complex notation (see equation (1.2.9)):

$$\boldsymbol{E}(\boldsymbol{r},t) = \mathrm{Re}\left(\hat{\boldsymbol{E}}(\boldsymbol{r})\, e^{-i\omega t}\right) \tag{1.5.1}$$

$$\boldsymbol{H}(\boldsymbol{r},t) = \mathrm{Re}\left(\hat{\boldsymbol{H}}(\boldsymbol{r})\, e^{-i\omega t}\right) \tag{1.5.2}$$

As long as linear operations like differentiation are made it is allowed to execute this operation on the complex function and finally build the real part. The partial derivatives with respect to t can be calculated directly whereby again the functionalities are omitted in our notation:

$$\begin{aligned}
\frac{\partial \boldsymbol{E}}{\partial t} &= \mathrm{Re}\left(-i\omega\hat{\boldsymbol{E}}e^{-i\omega t}\right)\\
\frac{\partial \boldsymbol{H}}{\partial t} &= \mathrm{Re}\left(-i\omega\hat{\boldsymbol{H}}e^{-i\omega t}\right)\\
\frac{\partial^2 \boldsymbol{E}}{\partial t^2} &= -\omega^2\mathrm{Re}\left(\hat{\boldsymbol{E}}e^{-i\omega t}\right)\\
\frac{\partial^2 \boldsymbol{H}}{\partial t^2} &= -\omega^2\mathrm{Re}\left(\hat{\boldsymbol{H}}e^{-i\omega t}\right)
\end{aligned}$$

These equations can be used in the wave equations (1.4.7) and (1.4.9) for linear and isotropic materials. The intermediate result is:

$$\begin{aligned}
&\mathrm{Re}\left(\left\{\Delta\hat{\boldsymbol{E}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{E}}\cdot\boldsymbol{\nabla}(\ln\epsilon)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{E}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{E}} + \right.\right.\\
&\left.\left. + (\boldsymbol{\nabla}(\ln\mu))\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{E}}\right)\right\}e^{-i\omega t}\right) = 0\\
&\mathrm{Re}\left(\left\{\Delta\hat{\boldsymbol{H}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{H}}\cdot\boldsymbol{\nabla}(\ln\mu)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{H}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{H}} + \right.\right.\\
&\left.\left. + (\boldsymbol{\nabla}(\ln\epsilon))\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{H}}\right) + [\boldsymbol{\nabla}\sigma - \sigma\boldsymbol{\nabla}(\ln\epsilon)]\times\hat{\boldsymbol{E}}\right\}e^{-i\omega t}\right) = 0
\end{aligned}$$

Since the real part of the product of a complex function $\hat{\boldsymbol{f}}(\boldsymbol{r})$ with the also complex function $\exp(-i\omega t)$ is

$$\mathrm{Re}\left(\hat{\boldsymbol{f}}e^{-i\omega t}\right) = \mathrm{Re}\left(\hat{\boldsymbol{f}}\right)\cos(\omega t) + \mathrm{Im}\left(\hat{\boldsymbol{f}}\right)\sin(\omega t)$$

and this function has to be zero at each time t, i.e. also at $\omega t = 0$ and $\omega t = \pi/2$, the complex function $\hat{\boldsymbol{f}}$ has to be zero itself. So, the final result for both differential equations of $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{H}}$ is:

$$\begin{aligned}
&\Delta\hat{\boldsymbol{E}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{E}}\cdot\boldsymbol{\nabla}\left(\ln\epsilon\right)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{E}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{E}} + \\
&+ \left(\boldsymbol{\nabla}\left(\ln\mu\right)\right)\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{E}}\right) = 0 \qquad (1.5.3)\\
&\Delta\hat{\boldsymbol{H}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{H}}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{H}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{H}} + \\
&+ \left(\boldsymbol{\nabla}\left(\ln\epsilon\right)\right)\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{H}}\right) + \left[\boldsymbol{\nabla}\sigma - \sigma\boldsymbol{\nabla}\left(\ln\epsilon\right)\right]\times\hat{\boldsymbol{E}} = 0 \qquad (1.5.4)
\end{aligned}$$

These two time–independent equations are called the Helmholtz equations for the electric and the magnetic vector. Since only the position–dependent, but time–independent complex electric and magnetic vectors $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{H}}$ are used, only time–averaged stationary quantities can be calculated using the Helmholtz equations. Again, two special cases are of interest.

1.5.1 Helmholtz equations for pure dielectrics

For pure dielectric materials, which do not absorb any radiation in the regarded wavelength range, the conductivity is zero ($\sigma = 0$). In this case, equations (1.5.3) and (1.5.4) can be simplified and result in:

$$\begin{aligned}
\Delta\hat{\boldsymbol{E}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{E}}\cdot\boldsymbol{\nabla}\left(\ln\epsilon\right)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{E}} + \left(\boldsymbol{\nabla}\left(\ln\mu\right)\right)\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{E}}\right) &= 0 \qquad (1.5.5)\\
\Delta\hat{\boldsymbol{H}} + \boldsymbol{\nabla}\left[\hat{\boldsymbol{H}}\cdot\boldsymbol{\nabla}\left(\ln\mu\right)\right] + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{H}} + \left(\boldsymbol{\nabla}\left(\ln\epsilon\right)\right)\times\left(\boldsymbol{\nabla}\times\hat{\boldsymbol{H}}\right) &= 0 \qquad (1.5.6)
\end{aligned}$$

Again, both equations are symmetric to a replacement of $\hat{\boldsymbol{E}}$ with $\hat{\boldsymbol{H}}$ and ϵ with μ. Therefore, it is sufficient to solve one of these equations.

1.5.2 Helmholtz equations for homogeneous materials

In practice, we often have (at least approximately) homogeneous materials like glasses, air or vacuum. For homogeneous materials the gradients of ϵ, μ and σ are zero. In this case, equations (1.5.3) and (1.5.4) obtain a quite simple form:

$$\Delta\hat{\boldsymbol{E}} + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{E}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{E}} = 0 \qquad (1.5.7)$$

$$\Delta\hat{\boldsymbol{H}} + \omega^2\frac{\epsilon\mu}{c^2}\hat{\boldsymbol{H}} + i\omega\mu_0\mu\sigma\hat{\boldsymbol{H}} = 0 \qquad (1.5.8)$$

Both equations are completely symmetric in $\hat{\boldsymbol{E}}$ and $\hat{\boldsymbol{H}}$. The angular frequency ω is defined as $2\pi\nu$, and the frequency ν and the wavelength in vacuum λ are connected by equation (1.1.49): $\nu\lambda = c$. Therefore, equations (1.5.7) and (1.5.8) can be written as:

$$\left[\Delta + \hat{k}^2\right]\hat{\boldsymbol{E}} = 0 \qquad (1.5.9)$$

$$\left[\Delta + \hat{k}^2\right]\hat{\boldsymbol{H}} = 0 \qquad (1.5.10)$$

with

$$\hat{k}^2 = \omega^2 \frac{\epsilon\mu}{c^2} + i\omega\mu_0\mu\sigma = \left(\frac{2\pi}{\lambda}\right)^2 \left[\epsilon\mu + i\frac{\lambda}{2\pi c\epsilon_0}\mu\sigma\right] = \left(\frac{2\pi\hat{n}}{\lambda}\right)^2 \tag{1.5.11}$$

Here, the generally complex–valued refractive index $\hat{n}$ is defined as:

$$\hat{n}^2 = \epsilon\mu + i\frac{\lambda}{2\pi c\epsilon_0}\mu\sigma =: \mu\left(\epsilon + i\epsilon_I\right) = \mu\hat{\epsilon} \tag{1.5.12}$$

This means that $\hat{n}$ is a complex number if the conductivity σ is different from zero. In this case, the dielectric function can also be defined as a complex function $\hat{\epsilon}$ with the real part ϵ and the imaginary part ϵ_I:

$$\hat{\epsilon} := \epsilon + i\epsilon_I \quad \text{with} \quad \epsilon_I := \frac{\lambda\sigma}{2\pi c\epsilon_0} \tag{1.5.13}$$

The real part n and the imaginary part n_I of $\hat{n}$ can be calculated:

$$\hat{n} = n + in_I \quad \Rightarrow \quad \hat{n}^2 = n^2 - n_I^2 + 2inn_I \tag{1.5.14}$$

$$\Rightarrow \quad n^2 - n_I^2 = \epsilon\mu \quad \text{and} \quad 2nn_I = \frac{\lambda}{2\pi c\epsilon_0}\mu\sigma = \mu\epsilon_I \tag{1.5.15}$$

$$\Rightarrow \quad n^2 n_I^2 = \frac{\mu^2\epsilon_I^2}{4}$$

$$\Rightarrow \quad n^4 - \epsilon\mu n^2 - \frac{\mu^2\epsilon_I^2}{4} = 0$$

$$\Rightarrow \quad n = \sqrt{\frac{\epsilon\mu + \sqrt{\epsilon^2\mu^2 + \epsilon_I^2\mu^2}}{2}} = \sqrt{\mu\frac{\epsilon + \sqrt{\epsilon^2 + \epsilon_I^2}}{2}}; \tag{1.5.16}$$

$$n_I = \frac{\mu\epsilon_I}{2n} = \frac{\mu\epsilon_I}{\sqrt{2\mu\left(\epsilon + \sqrt{\epsilon^2 + \epsilon_I^2}\right)}} \tag{1.5.17}$$

Only the positive solution of the two solutions of the quadratic equation with the variable n^2 is taken since n should be a real number and additionally only the positive square root of n^2 is taken since n should be a positive real number (so called negative refractive index materials, which are nowadays very popular in basic research, are excluded in our considerations).
For a pure dielectric ($\sigma = 0$) the imaginary parts of $\hat{\epsilon}$ and $\hat{n}$ (see equations (1.5.13) and (1.5.17)) vanish and the refractive index is a real number like it was defined in equation (1.1.27):

$$\sigma = 0 \quad \Rightarrow \quad \hat{n} = n = \sqrt{\epsilon\mu} \tag{1.5.18}$$

On the other side, it has to be mentioned that there exist many natural materials like many metals with a negative value of ϵ. This can be seen from equation (1.5.15): $n^2 - n_I^2 = \epsilon\mu$. If n_I is larger than n the real part of the dielectric function ϵ will be negative (since we assume here that μ is positive). Of course, this implies that a negative value of ϵ is connected with large absorption. As example let us consider aluminium at a wavelength of $\lambda = 517$ nm. There, we have $n = 0.826$ and $n_I = 6.283$ [4]. With $\mu = 1$ there is $\epsilon = n^2 - n_I^2 = -38.8$ and $\epsilon_I = 2nn_I = 10.4$.

Only, if μ would also be negative there would be the possibility of a negative ϵ without absorption. These materials which are called left–handed materials or negative index materials [5] are an actual subject of research because they would have many interesting applications like a "perfect lens" [6]. But, up to now all experimental realizations (which are mostly not for visible light but for microwaves) show very high absorption and/or are only valid in the near field.

1.5.3 A simple solution of the Helmholtz equation in a homogeneous material

A simple solution of equation (1.5.9) is e.g. a linearly polarized plane wave propagating in the z–direction:

$$\hat{\boldsymbol{E}} = \hat{\boldsymbol{E}}_0 e^{i\hat{k}z} = \hat{\boldsymbol{E}}_0 e^{i2\pi\hat{n}z/\lambda} \tag{1.5.19}$$

Here, $\hat{\boldsymbol{E}}_0$ is a constant vector and its modulus represents the amplitude of the electric vector at $z = 0$. If $\hat{n}$ is complex the effective position–dependent amplitude decreases exponentially:

$$\hat{n} = n + in_I \quad \Rightarrow \quad \hat{\boldsymbol{E}} = \hat{\boldsymbol{E}}_0 e^{-2\pi n_I z/\lambda}\, e^{i2\pi n z/\lambda} \tag{1.5.20}$$

So, the extinction of a wave can be formally included in the notation of a wave using complex exponential terms by just assuming an also complex refractive index. The real part of this complex refractive index is responsible for the "normal" refractive properties and the imaginary part is responsible for absorption. For metals n_I can be larger than one so that the wave can enter the metal for only a fraction of a wavelength before the electric (and magnetic) vector vanishes.
Instead of using the imaginary part n_I of the refractive index the so called **absorption coefficient** α is often used. It is defined by

$$\alpha := \frac{4\pi}{\lambda} n_I \quad \Rightarrow \quad \hat{\boldsymbol{E}} = \hat{\boldsymbol{E}}_0 e^{-\alpha z/2}\, e^{i2\pi n z/\lambda} \tag{1.5.21}$$

After having propagated a distance $z = 1/\alpha$ the electric energy density and the intensity of the wave (modulus of the Poynting vector), which are both proportional to $|\hat{\boldsymbol{E}}|^2$, decrease to $1/e$ of their starting values.
Whereas α is according to our definition always positive in lossy materials, there are also active gain media, e.g. in lasers, which have a negative coefficient α. Then α is not an absorption but an amplification or gain coefficient.

1.5.4 Inhomogeneous plane waves

The solution of the Helmholtz equation (1.5.9) in a homogeneous but lossy material defined by equation (1.5.20) is the simplest form of a so called **inhomogeneous plane wave** [1],[7]. The general inhomogeneous plane wave is obtained from equation (1.5.9) by the approach

$$\hat{\boldsymbol{E}} = \hat{\boldsymbol{E}}_0 e^{i\hat{\boldsymbol{k}}\cdot\boldsymbol{r}} = \hat{\boldsymbol{E}}_0 e^{-\boldsymbol{g}\cdot\boldsymbol{r}} e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \quad \Rightarrow \quad -\hat{\boldsymbol{k}}\cdot\hat{\boldsymbol{k}} + \hat{k}^2 = 0 \tag{1.5.22}$$

where $\hat{\boldsymbol{k}} = \boldsymbol{k} + i\boldsymbol{g}$ is a constant but complex wave vector with the real part $\boldsymbol{k}$ and the imaginary part $\boldsymbol{g}$. In the general case, also $\hat{\boldsymbol{E}}_0$ is a constant but complex electric vector so that all polarization states can be represented (see section 2.3 on page 33).

By using equations (1.5.11), (1.5.14) and (1.5.21) the complex quantity $\hat{k}$ is defined as

$$\hat{k} = \frac{2\pi}{\lambda}\hat{n} = \frac{2\pi}{\lambda}(n + in_I) = \frac{2\pi n}{\lambda} + i\frac{\alpha}{2} \quad \Rightarrow \quad \hat{k}^2 = \frac{4\pi^2 n^2}{\lambda^2} - \frac{\alpha^2}{4} + i\frac{2\pi n\alpha}{\lambda} \tag{1.5.23}$$

This means that the two vectors $\boldsymbol{k}$ and $\boldsymbol{g}$ have to fulfill the conditions:

$$\hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{k}} = (\boldsymbol{k} + i\boldsymbol{g}) \cdot (\boldsymbol{k} + i\boldsymbol{g}) = |\boldsymbol{k}|^2 - |\boldsymbol{g}|^2 + 2i\boldsymbol{g} \cdot \boldsymbol{k} = \hat{k}^2 = \frac{4\pi^2 n^2}{\lambda^2} - \frac{\alpha^2}{4} + i\frac{2\pi n\alpha}{\lambda} \tag{1.5.24}$$

A separation of the real and imaginary part gives:

$$|\boldsymbol{k}|^2 - |\boldsymbol{g}|^2 = \frac{4\pi^2 n^2}{\lambda^2} - \frac{\alpha^2}{4} \tag{1.5.25}$$

$$\boldsymbol{g} \cdot \boldsymbol{k} = \frac{\pi n\alpha}{\lambda} \tag{1.5.26}$$

So, the projection of the vector $\boldsymbol{g}$ onto the vector $\boldsymbol{k}$ has to fulfill the second equation. An important and interesting case is that of a lossless material, i.e. $\alpha = 0$. Then $\boldsymbol{g}$ and $\boldsymbol{k}$ have to be perpendicular to each other. This means that the planes of constant phase, which are perpendicular to $\boldsymbol{k}$, and the planes of constant amplitude, which are perpendicular to $\boldsymbol{g}$, are also perpendicular to each other.

Inhomogeneous plane waves do not exist in the whole space because the amplitude decreases exponentially along the direction of $\boldsymbol{g}$, but on the other side, it increases exponentially for the direction antiparallel to $\boldsymbol{g}$ and would tend to infinity. Therefore, only the half space with the exponentially decreasing part can exist in the real world whereas in the other direction there has to be a limit. An example of an inhomogeneous plane wave is an **evanescent wave** in the case of total internal reflection at an interface between two dielectric materials with different refractive indices. There, a plane wave propagating in the material with higher refractive index with an angle of incidence at the interface of more than the critical angle of total internal reflection is reflected. But, besides the reflected wave there exists an evanescent wave in the material with the lower refractive index. Its vector $\boldsymbol{k}$ is parallel to the interface between the two materials while its amplitude decreases exponentially with increasing distance from the interface. The evanescent wave transports no energy into the material with the lower refractive index but the total energy is in the reflected wave.

Nevertheless, if there is in a short distance a second interface to a medium with a higher refractive index so that there the refracted wave can again propagate, there will be an energy transport to this propagating wave over the evanescent wave. This is equivalent to the tunnelling effect in quantum mechanics. Of course, this effect is only important if the distance between both interfaces is of the order of a wavelength or smaller.

In the following sections some basic properties of light waves will be described. For more information see text books of optics like e.g. [1], [8], [9], [10], [11], [12], [13], [14].

Chapter 2

Polarization

In section 1.1 it is shown that e.g. a so called linearly polarized plane wave is a solution of Maxwell's equations. There, the electric vector has a well–defined direction which remains constant during the propagation of the wave. There are other solutions of Maxwell's equations where the direction of the electric vector does not remain constant during the propagation, but nevertheless, it has at a certain point and at a certain time a well–defined direction. All these solutions are called **polarized light**.
Contrary to this, light which is emitted by an electric bulb is unpolarized. This means that there are many light waves which have stochastically distributed phase relations to each other, i.e. incoherent light, and where the polarization varies in time. So, these light waves are added incoherently and there is no preferred direction of the electric vector. In practice, light is often partially polarized, i.e. some of the light is unpolarized and the other is polarized. Natural sun light on the earth is e.g. partially polarized because of the influence of the atmosphere onto the originally unpolarized light of the sun.
Here, only the case of a fully polarized plane wave in a homogeneous dielectric material will be investigated. In section 1.1.6 it is shown that the electric vector $\boldsymbol{E}$ and the magnetic vector $\boldsymbol{H}$ of a plane wave in a homogeneous and isotropic linear material are always perpendicular to each other and both are perpendicular to the direction of propagation $\boldsymbol{e}$ of the plane wave. Therefore, for a given direction of propagation $\boldsymbol{e}$ it is sufficient to consider only the electric vector. The magnetic vector is then automatically defined by equation (1.1.37):

$$\boldsymbol{H} = \pm\sqrt{\frac{\epsilon_0\epsilon}{\mu_0\mu}}\boldsymbol{e} \times \boldsymbol{E}$$

The electric vector $\boldsymbol{E}$ has to fulfill the wave equation (1.4.12) with $\sigma = 0$:

$$\Delta\boldsymbol{E} - \frac{n^2}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} = 0$$

Without loss of generality, the direction of propagation will be parallel to the z–axis, i.e. $\boldsymbol{e} = (0, 0, 1)$. Because of the orthogonality relation $\boldsymbol{E}$ can then only have a x– and a y–component. A quite general plane wave solution of the wave equation is in this case:

$$\boldsymbol{E}(z,t) = \begin{pmatrix} E_x(z,t) \\ E_y(z,t) \\ E_z(z,t) \end{pmatrix} = \begin{pmatrix} A_x\cos(kz - \omega t + \delta_x) \\ A_y\cos(kz - \omega t + \delta_y) \\ 0 \end{pmatrix} \tag{2.0.1}$$

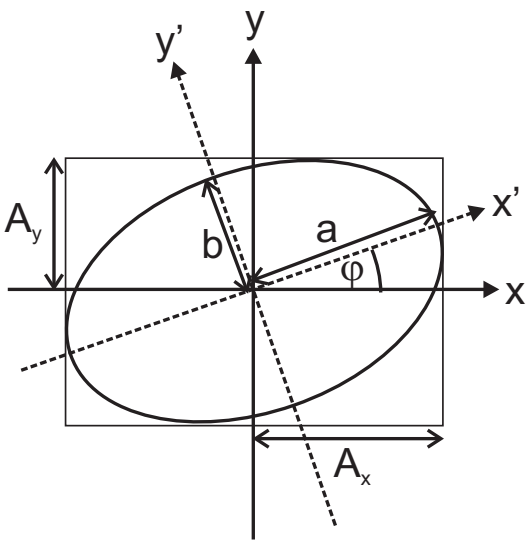


Figure 2.1: The polarization ellipse on which the apex of the electric vector moves if the time t or the coordinate z changes.

Here, $k = 2\pi n/\lambda = \omega n/c$ (see equation (1.1.50)) is the modulus of the wave vector $\boldsymbol{k} = 2\pi n\boldsymbol{e}/\lambda$. It holds $A_x \geq 0$ and $A_y \geq 0$. By applying a trigonometric theorem, introducing the parameter $\alpha := kz - \omega t + \delta_x$ and the phase difference $\delta := \delta_y - \delta_x$ this equation can be written as:

$$\begin{pmatrix} E_x(\alpha) \\ E_y(\alpha) \end{pmatrix} = \begin{pmatrix} A_x \cos\alpha \\ A_y \cos(\alpha + \delta) \end{pmatrix} = \begin{pmatrix} A_x \cos\alpha \\ A_y \cos\alpha \cos\delta - A_y \sin\alpha \sin\delta \end{pmatrix} \tag{2.0.2}$$

This equation is the parametric representation of an ellipse which is formed by the apex of the two–dimensional vector (E_x, E_y) in the x–y–plane for different values of the parameter α. Unfortunately, in general the principal axes of this ellipse will be rotated with respect to the x–axis and y–axis. Therefore, a transformation has to be done to calculate the principal axes of this ellipse with lengths $2a$ and $2b$. To do this, the following quantity is calculated where the argument α of E_x and E_y is omitted in the notation:

$$\begin{aligned}
&\left(\frac{E_x}{A_x}\right)^2 + \left(\frac{E_y}{A_y}\right)^2 - 2\frac{E_x E_y}{A_x A_y}\cos\delta = \\
&= \cos^2\alpha + (\cos\alpha\cos\delta - \sin\alpha\sin\delta)^2 - \\
&\quad -2\cos\alpha\cos\delta\,(\cos\alpha\cos\delta - \sin\alpha\sin\delta) = \\
&= \cos^2\alpha + (\cos\alpha\cos\delta - \sin\alpha\sin\delta)\,(-\cos\alpha\cos\delta - \sin\alpha\sin\delta) = \\
&= \cos^2\alpha\left(1 - \cos^2\delta\right) + \sin^2\alpha\sin^2\delta = \sin^2\delta \\
&\Rightarrow \quad \left(\frac{E_x}{A_x\sin\delta}\right)^2 + \left(\frac{E_y}{A_y\sin\delta}\right)^2 - 2\frac{E_x E_y}{A_x A_y \sin^2\delta}\cos\delta = 1
\end{aligned} \tag{2.0.3}$$

This is the implicit representation of an ellipse which is rotated with respect to the x– and y–axis (see fig. 2.1). On the other side an ellipse with the half axes a and b parallel to the coordinate axes x' and y' and the coordinates of the electric vector E'_x and E'_y is written as:

$$\left(\frac{E'_x}{a}\right)^2 + \left(\frac{E'_y}{b}\right)^2 = 1 \tag{2.0.4}$$

This equation is transformed into a coordinate system with axes x and y whereby the system (x', y') is rotated by an angle φ relative to the system (x, y) by applying the well–known rotation matrix to the coordinates:

$$\begin{pmatrix} E'_x \\ E'_y \end{pmatrix} = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} E_x\cos\varphi + E_y\sin\varphi \\ -E_x\sin\varphi + E_y\cos\varphi \end{pmatrix}$$

So, the equation of the rotated ellipse in the coordinate system (x, y) indexed by terms in E_x^2, E_y^2 and E_xE_y is:

$$E_x^2\left(\frac{\cos^2\varphi}{a^2} + \frac{\sin^2\varphi}{b^2}\right) + E_y^2\left(\frac{\sin^2\varphi}{a^2} + \frac{\cos^2\varphi}{b^2}\right) + E_xE_y\sin(2\varphi)\left(\frac{1}{a^2} - \frac{1}{b^2}\right) = 1$$

In the last step the trigonometric equality $\sin(2\varphi) = 2\sin\varphi\cos\varphi$ is used.
By comparing this equation with equation (2.0.3) the coefficients of the terms in E_x^2, E_y^2 and E_xE_y have to be equal to calculate the rotation angle φ and the principal axes. This results in three equations:

$$\begin{aligned} \frac{\cos^2\varphi}{a^2} + \frac{\sin^2\varphi}{b^2} &= \frac{1}{A_x^2\sin^2\delta} \\ \frac{\sin^2\varphi}{a^2} + \frac{\cos^2\varphi}{b^2} &= \frac{1}{A_y^2\sin^2\delta} \\ \sin(2\varphi)\left(\frac{1}{a^2} - \frac{1}{b^2}\right) &= -2\frac{\cos\delta}{A_xA_y\sin^2\delta} \end{aligned}$$

By (i) adding the second equation to the first equation and (ii) subtracting the second equation from the first equation and applying the trigonometric equation $\cos(2\varphi) = \cos^2\varphi - \sin^2\varphi$ two new equations are obtained and notated together with the old third equation:

$$\begin{aligned} \frac{1}{a^2} + \frac{1}{b^2} &= \frac{1}{\sin^2\delta}\left(\frac{1}{A_x^2} + \frac{1}{A_y^2}\right) \\ \cos(2\varphi)\left(\frac{1}{a^2} - \frac{1}{b^2}\right) &= \frac{1}{\sin^2\delta}\left(\frac{1}{A_x^2} - \frac{1}{A_y^2}\right) \\ \sin(2\varphi)\left(\frac{1}{a^2} - \frac{1}{b^2}\right) &= -2\frac{\cos\delta}{A_xA_y\sin^2\delta} \end{aligned}$$

Division of the third equation by the second equation gives the tangent of twice the rotation angle φ of the ellipse:

$$\tan(2\varphi) = \frac{-2\cos\delta}{A_xA_y\left(\frac{1}{A_x^2} - \frac{1}{A_y^2}\right)} = \frac{-2A_xA_y\cos\delta}{A_y^2 - A_x^2} \tag{2.0.5}$$

So, the rotation angle φ can be calculated from the known variables A_x, A_y and δ. There is a null of the denominator for the case $A_x = A_y$. Then the argument of the tangent is $2\varphi = \pm\pi/2$, i.e. $\varphi = \pm\pi/4$. Only, if additionally the numerator is also zero, i.e. $\cos\delta = 0$ and therefore $\delta = (2n+1)\pi/2$, the angle φ is not defined. In this case, the ellipse is degenerated into a circle as we will see later and then the rotation angle is of course not defined. It should also be

mentioned that the rotation angle φ is only defined between $-\pi/4$ and $\pi/4$ and therefore 2φ between $-\pi/2$ and $\pi/2$ so that the arc tangent function is well defined. This is sufficient because at first a rotation of an ellipse by π (=180 degree) does not change anything. And second, the principal axes of the ellipse can be chosen in such a way that either a is the large axis or b. So, the rotation angle φ has to be defined only in an interval of the length $\pi/2$, i.e. $[-\pi/4; \pi/4]$. After having calculated φ also the half axes a and b of the polarization ellipse can be calculated using the first two of the above equations:

$$
\begin{aligned}
&(i) \quad \frac{1}{a^2} + \frac{1}{b^2} = \frac{1}{\sin^2\delta}\left(\frac{1}{A_x^2} + \frac{1}{A_y^2}\right) \\
&(ii) \quad \frac{1}{a^2} - \frac{1}{b^2} = \frac{1}{\cos(2\varphi)\sin^2\delta}\left(\frac{1}{A_x^2} - \frac{1}{A_y^2}\right) \\
&(i)+(ii) \;\Rightarrow\; \frac{1}{a^2} = \frac{1}{2\sin^2\delta}\left[\frac{1}{A_x^2} + \frac{1}{A_y^2} + \frac{1}{\cos(2\varphi)}\left(\frac{1}{A_x^2} - \frac{1}{A_y^2}\right)\right] \qquad (2.0.6) \\
&(i)-(ii) \;\Rightarrow\; \frac{1}{b^2} = \frac{1}{2\sin^2\delta}\left[\frac{1}{A_x^2} + \frac{1}{A_y^2} - \frac{1}{\cos(2\varphi)}\left(\frac{1}{A_x^2} - \frac{1}{A_y^2}\right)\right] \qquad (2.0.7)
\end{aligned}
$$

Using a trigonometric relation and equation (2.0.5) it holds

$$
\begin{aligned}
\frac{1}{\cos(2\varphi)} &= \sqrt{1+\tan^2(2\varphi)} = \sqrt{1 + 4\frac{A_x^2 A_y^2 \cos^2\delta}{\left(A_y^2 - A_x^2\right)^2}} = \\
&= \frac{\sqrt{\left(A_y^2 - A_x^2\right)^2 + 4A_x^2 A_y^2 \cos^2\delta}}{\left|A_y^2 - A_x^2\right|}
\end{aligned}
$$

Remember, that 2φ is only defined between $-\pi/2$ and $\pi/2$ so that $\cos(2\varphi) \geq 0$. Defining the value s by

$$
s = \begin{cases} +1 \text{ if } A_y^2 - A_x^2 \geq 0 \\ -1 \text{ if } A_y^2 - A_x^2 < 0 \end{cases} \qquad (2.0.8)
$$

the reciprocal values of the squares of the half axes can be explicitly written as:

$$
\frac{1}{a^2} = \frac{1}{2A_x^2 A_y^2 \sin^2\delta}\left[A_y^2 + A_x^2 + s\sqrt{\left(A_y^2 - A_x^2\right)^2 + 4A_x^2 A_y^2 \cos^2\delta}\right] \qquad (2.0.9)
$$

$$
\frac{1}{b^2} = \frac{1}{2A_x^2 A_y^2 \sin^2\delta}\left[A_y^2 + A_x^2 - s\sqrt{\left(A_y^2 - A_x^2\right)^2 + 4A_x^2 A_y^2 \cos^2\delta}\right] \qquad (2.0.10)
$$

Now it is easy to calculate the ratio of the squares of the half axes, whereby the equality $\cos^2\delta - 1 = -\sin^2\delta$ is used:

$$
\begin{aligned}
\frac{b^2}{a^2} &= \frac{A_y^2 + A_x^2 + s\sqrt{\left(A_y^2 + A_x^2\right)^2 - 4A_x^2 A_y^2 \sin^2\delta}}{A_y^2 + A_x^2 - s\sqrt{\left(A_y^2 + A_x^2\right)^2 - 4A_x^2 A_y^2 \sin^2\delta}} = \\
&= \frac{1 + s\sqrt{1 - 4\sin^2\delta \frac{A_x^2 A_y^2}{\left(A_y^2 + A_x^2\right)^2}}}{1 - s\sqrt{1 - 4\sin^2\delta \frac{A_x^2 A_y^2}{\left(A_y^2 + A_x^2\right)^2}}} \qquad (2.0.11)
\end{aligned}
$$

So, for given parameters A_x, A_y and δ the half axes a and b of the ellipse can be calculated by the equations (2.0.9) and (2.0.10). The rotation angle can be calculated by (2.0.5) and the ratio of the half axes by (2.0.11). There are several interesting special cases of polarization states and these will be discussed in the following.

2.1 Different states of polarization

2.1.1 Linear polarization

An important and quite simple polarization state is the case of **linear polarization**. In this case the polarization ellipse degenerates to a line and the apex of the electric vector just oscillates on a line. This is the case if either the numerator or the denominator of equation (2.0.11) is zero so that a or b is zero. This means:

$$1 - \sqrt{1 - 4\sin^2\delta \frac{A_x^2 A_y^2}{\left(A_y^2 + A_x^2\right)^2}} = 0 \;\Rightarrow\; \sin\delta = 0 \text{ or } A_x = 0 \text{ or } A_y = 0 \tag{2.1.1}$$

The two cases $A_x = 0$ or $A_y = 0$ are obvious because in this case there is only an x– or a y–component of the electric vector. If both components are different from zero there is nevertheless linear polarization if the phase difference δ between the two components of the electric vector is $\delta = 0$ or $\delta = \pi$.

2.1.2 Circular polarization

If the apex of the electric vector moves on a circle the polarization state is called **circular polarization**. This means that both half axes have to be equal: $a = b$. Using equation (2.0.11) this requires:

$$\frac{b^2}{a^2} = 1 \;\Rightarrow\; 2\sin\delta \frac{A_x A_y}{A_y^2 + A_x^2} = \pm 1$$

Since $|\sin\delta| \leq 1$ this requires (A_x and A_y are both positive):

$$\frac{A_x A_y}{A_y^2 + A_x^2} \geq \frac{1}{2} \;\Rightarrow\; 2A_x A_y \geq A_x^2 + A_y^2 \;\Rightarrow\; 0 \geq (A_x - A_y)^2$$

This condition can only be fulfilled for the case $A_x = A_y$ and then there is the additional condition $\sin\delta = \pm 1$. Finally, the conditions for circular polarization are:

$$A_x = A_y \;\wedge\; \delta = \frac{\pm\pi}{2} \tag{2.1.2}$$

Using the original equation (2.0.1) this means for the electric vector:

$$\begin{pmatrix} E_x(z,t) \\ E_y(z,t) \end{pmatrix} = \begin{pmatrix} A_x \cos\left(kz - \omega t + \delta_x\right) \\ A_x \cos\left(kz - \omega t + \delta_x \pm \frac{\pi}{2}\right) \end{pmatrix} = \begin{pmatrix} A_x \cos\left(kz - \omega t + \delta_x\right) \\ \mp A_x \sin\left(kz - \omega t + \delta_x\right) \end{pmatrix} \tag{2.1.3}$$

So, the two different signs of the phase difference δ correspond to different directions of rotation of the apex of the electric vector. These two cases are called **right–handed circular polarization**

($\delta = -\pi/2$) and **left–handed circular polarization** ($\delta = +\pi/2$). The definition of right–handed and left–handed is not always identical in text books so that we use this definition for **right–handed circular polarization**:

$$\begin{pmatrix} E_x(z,t) \\ E_y(z,t) \end{pmatrix} = \begin{pmatrix} A_x \cos(kz - \omega t + \delta_x) \\ A_x \cos\left(kz - \omega t + \delta_x - \frac{\pi}{2}\right) \end{pmatrix} = \begin{pmatrix} A_x \cos(kz - \omega t + \delta_x) \\ A_x \sin(kz - \omega t + \delta_x) \end{pmatrix} \tag{2.1.4}$$

If we take a fixed point in space, for example z=0, the electric vector rotates for right–handed circular polarization in time clockwise, if we look from the observation plane to the incident beam, i.e. anti-parallel to the z-direction or propagation direction of the beam. If we look in the direction of propagation and take a fixed point in time the electric vector forms a helix in space. For right–handed circular polarization the fingers of the right hand show the chirality of this helix if the thumb points in the direction of propagation. Therefore, it is named right–handed circular polarization. One has to be very careful with these definitions, because it is quite important whether the electric vector is taken for a fixed point in time or in space and also the orientation of the different axes x, y and z have to form together a right-handed system.

2.1.3 Elliptic polarization

The general polarization state is of course the so called **elliptic polarization**. In this case the apex of the electric vector moves on an ellipse if the time t or the position z is changed. This state is the case if neither $\delta = 0$ or $\delta = \pi$ nor $\delta = \pm\pi/2$. Also if $\delta = \pm\pi/2$ the light is elliptically polarized if $A_x \neq A_y$. There, we have again to distinguish between right–handed and left–handed elliptic polarization.

2.2 The Poincaré sphere

A method to visualize the different states of polarization is the so called Poincaré sphere which was introduced by H. Poincaré in 1892. By using equation (2.0.2) as definition of the electric field, the so called **Stokes parameters** of a plane monochromatic wave can be defined [1]:

$$\begin{aligned} s_0 &:= A_x^2 + A_y^2 \\ s_1 &:= A_x^2 - A_y^2 \\ s_2 &:= 2A_xA_y \cos\delta \\ s_3 &:= 2A_xA_y \sin\delta \end{aligned} \tag{2.2.1}$$

It is obvious that the four quantities are connected by the relation:

$$s_1^2 + s_2^2 + s_3^3 = A_x^4 + A_y^4 - 2A_x^2A_y^2 + 4A_x^2A_y^2 = A_x^4 + A_y^4 + 2A_x^2A_y^2 = s_0^2 \tag{2.2.2}$$

So, only three of the parameters are independent and the parameter s_0 is proportional to the intensity of the wave. If s_1, s_2 and s_3 are now used as the cartesian coordinates of a point in space, all allowed combinations will be situated according to equation (2.2.2) on a sphere with radius s_0. The radius s_0 is proportional to the intensity of the wave and the sphere is called the **Poincaré sphere** (see fig. 2.2).
The different polarization states correspond to different positions on the Poincaré sphere. For linearly polarized light it is e.g. either $A_x = 0$ or $A_y = 0$ or $\delta = 0$ or $\delta = \pi$. In all four cases the parameter s_3 will be zero. This means that points lying in the equatorial plane of the Poincaré

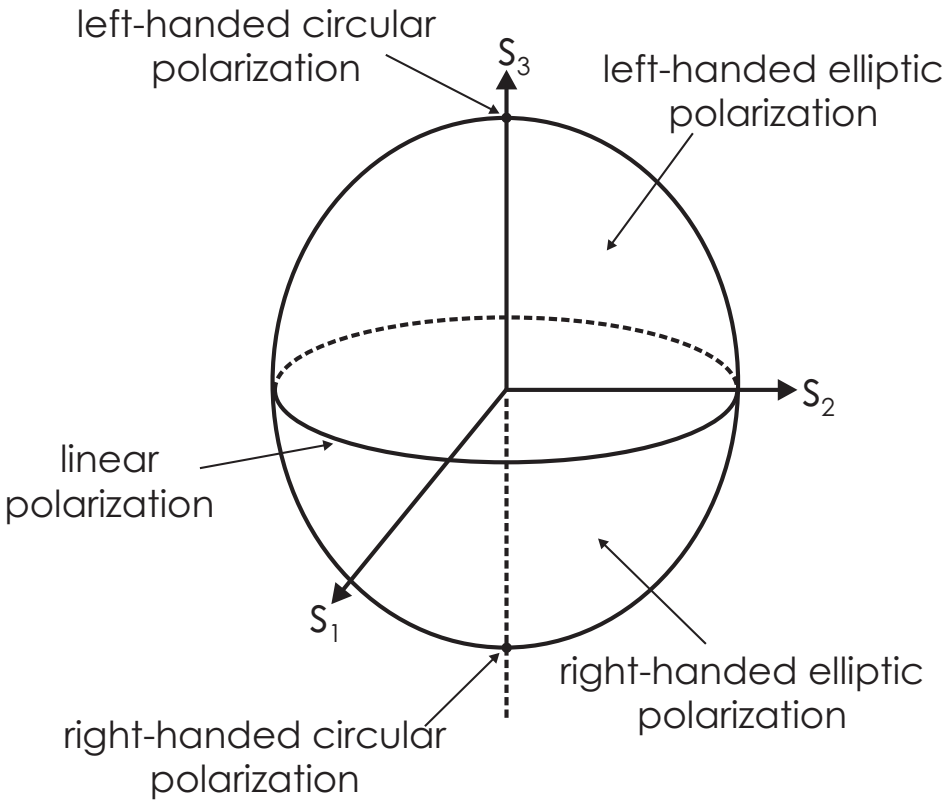


Figure 2.2: The Poincaré sphere and the visualization of the different states of polarization.

sphere represent linear polarization. Another interesting case is circular polarization. In this case the conditions are according to equation (2.1.2): $A_x = A_y$ and $\delta = \pm\pi/2$. Therefore, s_1 and s_2 are both zero and circular polarization corresponds to the poles of the Poincaré sphere. At the north pole ($s_1 = s_2 = 0$ and $s_3 = s_0$) the light is left–handed circularly polarized ($\delta = \pi/2$). At the south pole ($s_1 = s_2 = 0$ and $s_3 = -s_0$) the light is right–handed circularly polarized ($\delta = -\pi/2$). All other states of polarization (elliptic polarization) correspond to points somewhere else on the Poincaré sphere. In the upper hemisphere of the Poincaré sphere ($s_3 > 0$) the light is always left–handed polarized and in the lower hemisphere ($s_3 < 0$) the light is right–handed polarized.
Again, it has to be mentioned that the definition of left–handed and right–handed or the sign of the definition of the phase difference δ may be different in different text books. Therefore, in some text books the hemispheres for left–handed and right–handed polarization on the Poincaré sphere may be exchanged compared to our definition.

2.2.1 The helicity

A quite interesting quantity which gives a measure of the "degree of circular polarization" is the so called **helicity** σ. Using our sign definitions it is defined as:

$$\sigma = \frac{2A_xA_y \sin\delta}{A_x^2 + A_y^2} = \frac{s_3}{s_0} \tag{2.2.3}$$

The modulus of the helicity can only vary between zero for linear polarization and one for circular polarization. General elliptic polarization means a value of $0 < |\sigma| < 1$. On the Poincaré sphere $|\sigma|$ describes the "degree of latitude". It is zero in the equatorial plane and increases towards one at the poles.
Without being able to give at this point a concrete derivation it should be mentioned that the helicity is the ratio of the actual component of the angular momentum (*dt.:* Drehimpuls) of a beam along its direction of propagation (say z–direction) to the maximum possible z–component of the angular momentum for a given light power. Of course, the maximum z–component of the angular momentum is obtained for circular polarization with $|\sigma| = 1$, whereas a linearly polarized beam has no z–component of the angular momentum ($|\sigma| = 0$).
Concerning the sign definition, a value of $\delta = +\pi/2$, which means left–handed circular polarization, gives a positive helicity [15].

2.3 Complex representation of a polarized wave

In equation (2.0.1) the electric vector is expressed as a real quantity. As we have seen in other sections it is in many cases useful to take a complex notation, where only the real part has a physical meaning:

$$\begin{aligned} \boldsymbol{E}(z,t) &= \begin{pmatrix} E_x(z,t) \\ E_y(z,t) \\ E_z(z,t) \end{pmatrix} = \begin{pmatrix} A_x\cos(kz - \omega t + \delta_x) \\ A_y\cos(kz - \omega t + \delta_y) \\ 0 \end{pmatrix} = \\ &= \mathrm{Re}\left\{ \begin{pmatrix} A_x e^{i\delta_x} \\ A_y e^{i\delta_y} \\ 0 \end{pmatrix} e^{ikz} e^{-i\omega t} \right\} = \mathrm{Re}\left\{ \hat{\boldsymbol{A}} e^{ikz} e^{-i\omega t} \right\} \end{aligned} \tag{2.3.1}$$

Polarization state	Jones vector
Linear polarization in x–direction	$E_0 \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
Linear polarization in y–direction	$E_0 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
Linear polarization with 45^o rotation	$\frac{E_0}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
Right–handed circular polarization	$\frac{E_0}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix}$
Left–handed circular polarization	$\frac{E_0}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}$

Table 2.1: Jones vectors of some important polarization states. The absolute value is in all cases equal to $|E_0|$, where E_0 itself can be in the general case a complex valued quantity in order to allow the representation of a phase–offset.

In this case the vector $\hat{\boldsymbol{A}}$ is a constant but complex vector in order to can represent all possible states of polarization. With a real vector only a linear polarization state could be represented. The complex notation has the advantage that it is quite easy to calculate the time average of the intensity. According to equation (1.2.13) the time average of the intensity I is proportional to the square of the absolute value of the time–independent complex electric vector $\hat{\boldsymbol{E}}$:

$$\hat{\boldsymbol{E}}(z) = \hat{\boldsymbol{A}} e^{ikz} \;\Rightarrow\; I \propto \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{E}}^* = A_x^2 + A_y^2 \tag{2.3.2}$$

Therefore, it is clear that a detector which is only sensitive to the intensity of a light wave cannot distinguish between different polarization states.

2.4 Simple polarizing optical elements and the Jones calculus

There are of course optical elements which influence the polarization state like polarizers, quarter–wave plates, half–wave plates and many other. Here, only the basic idea of their effects can be discussed. For more information about the treatment of polarizing optical elements see e.g. [16],[17],[18]. In this section only fully polarized light is treated. This can be produced from natural unpolarized light with the help of a **polarizer**. In the following the word *polarizer* is always used for a polarization filter whereas other polarizing elements are simply called *polarizing elements* or *polarizing optical elements*. The case of partially polarized light will not be treated. A quite useful algorithm for the treatment of fully polarized light is the so called Jones calculus which was invented by R.C. Jones [19]. If we have again a plane wave propagating in the z–direction the polarization state can be described by a two–dimensional vector $\boldsymbol{J}$ in x– and y–direction containing the x– and y–components of the vector $\hat{\boldsymbol{A}}$ of equation (2.3.1):

$$\boldsymbol{J} = \begin{pmatrix} J_x \\ J_y \end{pmatrix} = \begin{pmatrix} A_x e^{i\delta_x} \\ A_y e^{i\delta_y} \end{pmatrix} \tag{2.4.1}$$

This vector is called **Jones vector** and the Jones vectors of some important polarization states are listed in table 2.1.
The square of the absolute value of the Jones vector is proportional to the intensity of the plane wave:

$$I \propto \boldsymbol{J} \cdot \boldsymbol{J}^* = |J_x|^2 + |J_y|^2 = A_x^2 + A_y^2 \tag{2.4.2}$$

Now each polarizing optical element can be represented by a 2x2 matrix, the **Jones matrix P**. The resulting Jones vector $\boldsymbol{J'}$ of the light that has passed such a polarizing optical element is calculated by multiplying the Jones vector $\boldsymbol{J}$ of the incident light with the Jones matrix $\mathbf{P}$:

$$\boldsymbol{J'} = \mathbf{P}\boldsymbol{J} \tag{2.4.3}$$

Several polarizing optical elements can be passed by multiplying simply the Jones matrices of these elements. Of course, the order of the matrices has to be correct. The matrix of the first polarizing optical element is rightmost and the matrix of the last element is leftmost.

2.4.1 Polarizer

A polarizer is a device that produces linearly polarized light from an arbitrary polarization state. If the polarizer lets pass only light polarized in x–direction its Jones matrix is

$$\mathbf{P_x} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{2.4.4}$$

Similarly Jones matrices of polarizers in other directions (y–direction, 45 degree or -45 degree) can be represented by:

$$\mathbf{P_y} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix};\ \mathbf{P_{45^o}} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix};\ \mathbf{P_{-45^o}} = \frac{1}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}; \tag{2.4.5}$$

The matrix of a polarizer with axis of transmission rotated by an angle φ can of course be calculated from the matrix of the polarizer in x–direction by a coordinate transformation. Mathematically, this means a multiplication of the matrix of the polarizer from left and right with the respective rotation matrix and its inverse matrix:

$$\mathbf{P}_\varphi = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \tag{2.4.6}$$

As example using polarizers consider a plane wave linearly polarized in x–direction which passes first a polarizer in x–direction and then in the direction 45 degree. The resulting vector is:

$$\boldsymbol{J}_{final} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} E_0 \\ 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} E_0 \\ 0 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} E_0 \\ E_0 \end{pmatrix}$$

Of course, the first polarizer in x–direction has no effect and the second polarizer selects the component of the electric vector in 45 degree direction. The resulting intensity of the wave is then proportional to $|\boldsymbol{J}_{final}|^2 = |E_0|^2/2$, i.e. half of the intensity is absorbed. An also well–known effect is the combination of two crossed polarizers (e.g. in x– and y–direction). Their matrix is of course:

$$\mathbf{P_{crossed}} = \mathbf{P_y}\mathbf{P_x} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

This means that no light passes this combination. Only if other polarizing elements are between the two crossed polarizers light can pass. If e.g. a polarizer with direction 45 degree is inserted between the two crossed polarizers light can pass this combination if it has originally a component in x–direction. The resulting light will of course only have a y–component:

$$\mathbf{P_{crossed+45^o}} = \mathbf{P_y P_{45^o} P_x} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$$

Though this is a well–known example it is nevertheless a typical example that polarizing optical elements can produce quite astonishing results by inserting additional elements.

2.4.2 Quarter–wave plate

Another elementary polarizing optical element is a quarter–wave plate ($\lambda/4$–plate) which consists of a birefringent material. If the axes of the material are correctly oriented the refractive index for light polarized in x–direction is e.g. different from the refractive index for polarized light in y–direction. The resulting effect is a phase difference between the two components of the Jones vector of $\pi/2$. The Jones matrix for a higher phase velocity in y–direction is e.g.

$$\mathbf{P}_{\lambda/4} = e^{i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \tag{2.4.7}$$

This means that linearly polarized light with components in only x– or y–direction remains linearly polarized and the intensity is unchanged (in practice some light is of course absorbed). But for light which is linearly polarized and has equal components in x– and y–direction (i.e. linearly polarized with a direction of 45 degree) the resulting polarization state is circularly polarized light:

$$\boldsymbol{J}_{final} = e^{i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} E_0 \\ E_0 \end{pmatrix} = \frac{1}{\sqrt{2}} e^{i\pi/4} \begin{pmatrix} E_0 \\ iE_0 \end{pmatrix}$$

Again, the intensity of the light is unchanged, only the polarization state has changed. Linearly polarized incident light with other directions of polarization will result in elliptically polarized light.
It has to be mentioned that the phase factor $\exp(i\pi/4)$ in front of the Jones matrix of the quarter wave plate is to some degree arbitrary since it depends on the point of reference for the phase.

2.4.3 Half–wave plate

A third interesting case is now that the circularly polarized light passes an identical quarter–wave plate a second time. Then the Jones vector is:

$$\boldsymbol{J}_{final} = e^{i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} \frac{1}{\sqrt{2}} e^{i\pi/4} \begin{pmatrix} E_0 \\ iE_0 \end{pmatrix} = \frac{1}{\sqrt{2}} e^{i\pi/2} \begin{pmatrix} E_0 \\ -E_0 \end{pmatrix}$$

The result is again linearly polarized light but with a rotation of the direction of polarization of 90 degree. The effect of two identical quarter–wave plates is of course the effect of a half–wave plate ($\lambda/2$–plate). So, a half–wave plate rotates the direction of polarization of linearly polarized light by 90 degree if the incident light is polarized in the direction of 45 degree. If the incident

light is polarized in x– or y–direction nothing happens. The matrix of a half–wave plate with the principal axes along x– and y–direction is:

$$\mathbf{P}_{\lambda/\mathbf{2}} = e^{i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} e^{i\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix} = e^{i\pi/2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{2.4.8}$$

For linearly polarized light with an arbitrary angle of rotation φ ($|\varphi| \leq \pi/4$) relative to one of the axes of the half–wave plate, the light behind the plate is again linearly polarized but rotated by 2φ.

Chapter 3

Interference

Interference is the property of all types of waves to form characteristic stationary variations of the intensity by the superposition of two or more waves. Of course, in the case of light some conditions have to be fulfilled because with natural light of the sun or from a bulb it is quite difficult to get interference effects. On the other hand it is no problem to obtain interference effects with the help of a laser. In fact, the condition is that the light has to be coherent or at least partially coherent[20]. There are complete books about interference effects and the application of these optical effects in the field of **interferometry** [21],[22],[23],[24]. So, in this chapter only the basic ideas can be treated.

3.1 Interference of two plane waves

First, the interference of two monochromatic plane waves in a homogeneous and isotropic material will be treated. The two plane waves with the angular frequency ω are propagating in the direction of their wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$. The corresponding unit vectors in the direction of propagation are $\boldsymbol{e}_1 = \boldsymbol{k}_1/|\boldsymbol{k}_1|$ and $\boldsymbol{e}_2 = \boldsymbol{k}_2/|\boldsymbol{k}_2|$. Their polarization state shall be arbitrary. In section 2 we investigated the different polarization states and used the fact that for one plane wave the coordinate system can be chosen so convenient that the direction of propagation is in the z–direction and the electric vector can only have x– and y–components because of the orthogonality condition. For two plane waves which are not propagating parallel a more generalized description has to be found. Therefore, the unit vector $\boldsymbol{e}_\perp$ which is perpendicular to the plane formed by the two propagation vectors $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ is defined as

$$\boldsymbol{e}_\perp := \frac{\boldsymbol{e}_1 \times \boldsymbol{e}_2}{|\boldsymbol{e}_1 \times \boldsymbol{e}_2|} \tag{3.1.1}$$

Only for the case that $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are parallel or antiparallel, i.e. $\boldsymbol{e}_2 = \pm\boldsymbol{e}_1$, $\boldsymbol{e}_\perp$ is not defined and we define in this case $\boldsymbol{e}_1 := (0, 0, 1)$ and $\boldsymbol{e}_\perp = (0, 1, 0)$. But in the following, it is sufficient to assume that $\boldsymbol{e}_\perp$ is well defined via equation (3.1.1) or otherwise. So, for each wave we can define a unit vector $\boldsymbol{e}_{\parallel,1}$ or $\boldsymbol{e}_{\parallel,2}$ which is lying in the plane defined by the directions of propagation of the two waves but perpendicular to the respective propagation vector:

$$\boldsymbol{e}_{\parallel,1} := \boldsymbol{e}_\perp \times \boldsymbol{e}_1 \tag{3.1.2}$$

$$\boldsymbol{e}_{\parallel,2} := \boldsymbol{e}_\perp \times \boldsymbol{e}_2 \tag{3.1.3}$$

Now, each plane wave will only have components along $\boldsymbol{e}_\perp$ and the respective vector $\boldsymbol{e}_{\parallel,1}$ or $\boldsymbol{e}_{\parallel,2}$. The components along $\boldsymbol{e}_\perp$ are called **TE–components**, i.e. transversal electric. The components along $\boldsymbol{e}_{\parallel,1}$ or $\boldsymbol{e}_{\parallel,2}$ are called **TM–components**, i.e. transversal magnetic, because in this case the corresponding component of the magnetic vector is perpendicular to the plane of propagation.
Therefore, using a generalization of equation (2.3.1) the electric vectors of both plane waves can be represented as:

$$\begin{aligned}
\boldsymbol{E}_1(\boldsymbol{r},t) &= \mathrm{Re}\left\{\left(A_{\parallel,1}e^{i\delta_{\parallel,1}}\boldsymbol{e}_{\parallel,1} + A_{\perp,1}e^{i\delta_{\perp,1}}\boldsymbol{e}_\perp\right)e^{i\boldsymbol{k}_1\cdot\boldsymbol{r}}e^{-i\omega t}\right\} = \\
&= \mathrm{Re}\left\{\left(\hat{\boldsymbol{A}}_{\parallel,1} + \hat{\boldsymbol{A}}_{\perp,1}\right)e^{i\boldsymbol{k}_1\cdot\boldsymbol{r}}e^{-i\omega t}\right\} \qquad (3.1.4)\\
\boldsymbol{E}_2(\boldsymbol{r},t) &= \mathrm{Re}\left\{\left(A_{\parallel,2}e^{i\delta_{\parallel,2}}\boldsymbol{e}_{\parallel,2} + A_{\perp,2}e^{i\delta_{\perp,2}}\boldsymbol{e}_\perp\right)e^{i\boldsymbol{k}_2\cdot\boldsymbol{r}}e^{-i\omega t}\right\} = \\
&= \mathrm{Re}\left\{\left(\hat{\boldsymbol{A}}_{\parallel,2} + \hat{\boldsymbol{A}}_{\perp,2}\right)e^{i\boldsymbol{k}_2\cdot\boldsymbol{r}}e^{-i\omega t}\right\} \qquad (3.1.5)
\end{aligned}$$

The quantities with a hat are complex, the others are real. Using this representation the orthogonality condition for electromagnetic waves in isotropic media is automatically fulfilled. The magnetic vector is not explicitly notated here because it is automatically defined by equation (1.1.37). Moreover, the interaction of an electromagnetic wave with matter is normally due to the electric field. Therefore, the electric vector is used in our calculation.
The interference of these two plane waves just means that the electric vectors have to be added. Since this is a linear operation and we are at the end only interested in the time average of the intensity it is sufficient to add the time–independent complex electric vectors $\hat{\boldsymbol{E}}_1$ and $\hat{\boldsymbol{E}}_2$. The resulting electric vector $\hat{\boldsymbol{E}}_{1+2}$ is:

$$\hat{\boldsymbol{E}}_{1+2} = \hat{\boldsymbol{E}}_1 + \hat{\boldsymbol{E}}_2 = \left(\hat{\boldsymbol{A}}_{\parallel,1} + \hat{\boldsymbol{A}}_{\perp,1}\right)e^{i\boldsymbol{k}_1\cdot\boldsymbol{r}} + \left(\hat{\boldsymbol{A}}_{\parallel,2} + \hat{\boldsymbol{A}}_{\perp,2}\right)e^{i\boldsymbol{k}_2\cdot\boldsymbol{r}} \qquad (3.1.6)$$

The intensity of a plane wave measured on a surface perpendicular to the direction of propagation is according to equation (1.2.13) proportional to $|\hat{\boldsymbol{E}}|^2$. The proportionality factor is $\sqrt{\epsilon/\mu}\epsilon_0 c/2$. The intensity of the plane wave on a plane surface that is not perpendicular to the direction of the energy flow is decreased by the cosine of the angle of incidence. In the following the plane on which we define the intensity of our interference pattern is perpendicular to the effective direction of the energy flow, i.e. perpendicular to $\boldsymbol{k}_1 + \boldsymbol{k}_2$ if the two waves are not antiparallel, or perpendicular to $\boldsymbol{k}_1$ if $\boldsymbol{k}_1 = -\boldsymbol{k}_2$. The cosine factors are then identical for both waves and the intensity I_{1+2} and the square of the modulus of the resulting electric vector $|\hat{\boldsymbol{E}}_{1+2}|^2$ are really proportional to each other with a constant of proportionality a. So, for the interference pattern holds:

$$\begin{aligned}
\left|\hat{\boldsymbol{E}}_{1+2}\right|^2 &= \hat{\boldsymbol{E}}_{1+2}\cdot\hat{\boldsymbol{E}}^*_{1+2} = \\
&= \left[\left(\hat{\boldsymbol{A}}_{\parallel,1} + \hat{\boldsymbol{A}}_{\perp,1}\right)e^{i\boldsymbol{k}_1\cdot\boldsymbol{r}} + \left(\hat{\boldsymbol{A}}_{\parallel,2} + \hat{\boldsymbol{A}}_{\perp,2}\right)e^{i\boldsymbol{k}_2\cdot\boldsymbol{r}}\right]\cdot \\
&\quad \cdot\left[\left(\hat{\boldsymbol{A}}^*_{\parallel,1} + \hat{\boldsymbol{A}}^*_{\perp,1}\right)e^{-i\boldsymbol{k}_1\cdot\boldsymbol{r}} + \left(\hat{\boldsymbol{A}}^*_{\parallel,2} + \hat{\boldsymbol{A}}^*_{\perp,2}\right)e^{-i\boldsymbol{k}_2\cdot\boldsymbol{r}}\right] = \\
&= \hat{\boldsymbol{A}}_{\parallel,1}\cdot\hat{\boldsymbol{A}}^*_{\parallel,1} + \hat{\boldsymbol{A}}_{\perp,1}\cdot\hat{\boldsymbol{A}}^*_{\perp,1} + \hat{\boldsymbol{A}}_{\parallel,2}\cdot\hat{\boldsymbol{A}}^*_{\parallel,2} + \hat{\boldsymbol{A}}_{\perp,2}\cdot\hat{\boldsymbol{A}}^*_{\perp,2} + \\
&\quad + \left(\hat{\boldsymbol{A}}_{\parallel,1}\cdot\hat{\boldsymbol{A}}^*_{\parallel,2} + \hat{\boldsymbol{A}}_{\perp,1}\cdot\hat{\boldsymbol{A}}^*_{\perp,2}\right)e^{i\,(\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}} + \\
&\quad + \left(\hat{\boldsymbol{A}}^*_{\parallel,1}\cdot\hat{\boldsymbol{A}}_{\parallel,2} + \hat{\boldsymbol{A}}^*_{\perp,1}\cdot\hat{\boldsymbol{A}}_{\perp,2}\right)e^{-i\,(\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}} \qquad (3.1.7)
\end{aligned}$$

All other terms vanish because of the orthogonality of the respective vectors. To evaluate this equation further, the scalar product of $\boldsymbol{e}_{\parallel,1}$ and $\boldsymbol{e}_{\parallel,2}$ has to be calculated. It is:

$$\begin{aligned} \boldsymbol{e}_{\parallel,1}\cdot\boldsymbol{e}_{\parallel,2} &= (\boldsymbol{e}_\perp\times\boldsymbol{e}_1)\cdot(\boldsymbol{e}_\perp\times\boldsymbol{e}_2) = (\boldsymbol{e}_1\times(\boldsymbol{e}_\perp\times\boldsymbol{e}_2))\cdot\boldsymbol{e}_\perp = \\ &= ((\boldsymbol{e}_1\cdot\boldsymbol{e}_2)\,\boldsymbol{e}_\perp - (\boldsymbol{e}_1\cdot\boldsymbol{e}_\perp)\,\boldsymbol{e}_2)\cdot\boldsymbol{e}_\perp = \boldsymbol{e}_1\cdot\boldsymbol{e}_2 \end{aligned} \tag{3.1.8}$$

This relation is of course obvious and we can use it to evaluate the interference pattern. To abbreviate the notation the phase differences between the two waves are defined as $\delta_\parallel := \delta_{\parallel,1}-\delta_{\parallel,2}$ and $\delta_\perp := \delta_{\perp,1}-\delta_{\perp,2}$:

$$\begin{aligned} \left|\hat{\boldsymbol{E}}_{1+2}\right|^2 &= A_{\parallel,1}^2 + A_{\perp,1}^2 + A_{\parallel,2}^2 + A_{\perp,2}^2 + \\ &\quad + \left(A_{\parallel,1}A_{\parallel,2}e^{i\delta_\parallel}\,(\boldsymbol{e}_1\cdot\boldsymbol{e}_2) + A_{\perp,1}A_{\perp,2}e^{i\delta_\perp}\right)e^{i\,(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}} + \\ &\quad + \left(A_{\parallel,1}A_{\parallel,2}e^{-i\delta_\parallel}\,(\boldsymbol{e}_1\cdot\boldsymbol{e}_2) + A_{\perp,1}A_{\perp,2}e^{-i\delta_\perp}\right)e^{-i\,(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}} = \\ &= A_{\parallel,1}^2 + A_{\parallel,2}^2 + 2A_{\parallel,1}A_{\parallel,2}\,(\boldsymbol{e}_1\cdot\boldsymbol{e}_2)\cos\left((\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}+\delta_\parallel\right) + \\ &\quad + A_{\perp,1}^2 + A_{\perp,2}^2 + 2A_{\perp,1}A_{\perp,2}\cos\left((\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}+\delta_\perp\right) \end{aligned} \tag{3.1.9}$$

The two terms which depend on parameters of both waves and which are functions of the position are called **interference terms**. These interference terms distinguish the superposition of coherent waves and incoherent waves. In the case of incoherent waves the interference terms vanish and the resulting intensity is just the sum of the single intensities of both waves. For so called partially coherent light there would be an additional factor in front of the interference terms which depends on the degree of coherence of the two waves and which can vary from zero (i.e. incoherent light) to one (i.e. fully coherent light). But, in this chapter we will concentrate on fully coherent light so that equation (3.1.9) is valid.
It can be seen that the interference pattern resolves into terms which depend only on TM–components and those which depend only on TE–components. Since both are perpendicular to each other the intensities of both waves can also be divided into the sum of a "TM–intensity" and a "TE–intensity":

$$\begin{aligned} I_1 &= a\left(A_{\parallel,1}^2 + A_{\perp,1}^2\right) = I_{\parallel,1} + I_{\perp,1} \\ I_2 &= a\left(A_{\parallel,2}^2 + A_{\perp,2}^2\right) = I_{\parallel,2} + I_{\perp,2} \end{aligned} \tag{3.1.10}$$

Here, the constant of proportionality a is used which is explained above. Then, the intensity of the interference pattern is:

$$\begin{aligned} I_{1+2} &= I_{\parallel,1} + I_{\parallel,2} + 2\sqrt{I_{\parallel,1}I_{\parallel,2}}\,(\boldsymbol{e}_1\cdot\boldsymbol{e}_2)\cos\left((\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}+\delta_\parallel\right) + \\ &\quad + I_{\perp,1} + I_{\perp,2} + 2\sqrt{I_{\perp,1}I_{\perp,2}}\cos\left((\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r}+\delta_\perp\right) \end{aligned} \tag{3.1.11}$$

3.1.1 The grating period and the fringe period

Equation (3.1.11) shows that the surfaces of constant intensity are plane surfaces with

$$(\boldsymbol{k}_1-\boldsymbol{k}_2)\cdot\boldsymbol{r} = \text{constant} \tag{3.1.12}$$

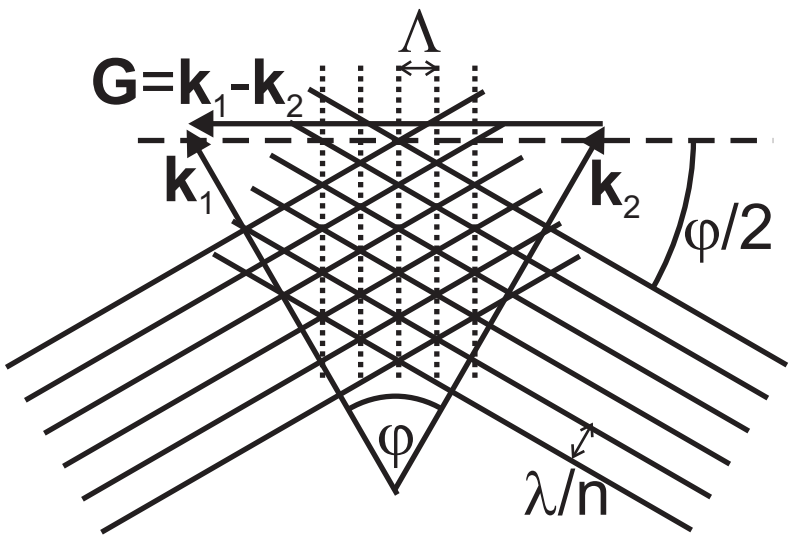


Figure 3.1: Interference of two plane waves. The solid lines indicate the planes of constant phase of the two plane waves at a fixed time having the distance λ/n respectively whereas the dashed lines indicate the interference planes with constant intensity having a distance Λ. The planes itself are perpendicular to the drawing plane.

The planes are perpendicular to the so called grating vector $\boldsymbol{G}$ (see fig. 3.1) with

$$\boldsymbol{G} = \boldsymbol{k}_1 - \boldsymbol{k}_2 \tag{3.1.13}$$

Since the cosine function is periodic the distance between two neighboring planes of equal intensity is called the grating period Λ of the interference pattern. It can be calculated by taking a point $\boldsymbol{r}_1$ on a first plane and a point $\boldsymbol{r}_2$ on a neighboring second plane with the same intensity value as on the first plane, so that the vector $\Delta\boldsymbol{r} := \boldsymbol{r}_2 - \boldsymbol{r}_1$ is parallel to $\boldsymbol{G}$ and simultaneously perpendicular to the planes. Its modulus is the grating period:

$$\begin{aligned}
&\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}_2\right) = \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}_1\right) \\
&\Rightarrow\ (\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}_2 = (\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r}_1 + 2\pi \\
&\Rightarrow\ \boldsymbol{G}\cdot\Delta\boldsymbol{r} = |\boldsymbol{G}|\,|\Delta\boldsymbol{r}| = 2\pi \\
&\Rightarrow\ \Lambda = |\Delta\boldsymbol{r}| = \frac{2\pi}{|\boldsymbol{G}|} = \frac{2\pi}{\frac{2\pi n}{\lambda}\,|\boldsymbol{e}_1 - \boldsymbol{e}_2|} = \frac{\lambda}{n\,|\boldsymbol{e}_1 - \boldsymbol{e}_2|}
\end{aligned} \tag{3.1.14}$$

Here, n is the refractive index of the material in which the waves propagate and $\lambda = 2\pi c/\omega$ would be the wavelength in vacuum. So, the grating period is infinity if both waves are propagating parallel, i.e. $\boldsymbol{e}_1 = \boldsymbol{e}_2$, and the smallest grating period can be obtained if both waves are propagating antiparallel, i.e. $\boldsymbol{e}_1 = -\boldsymbol{e}_2$. Then the grating period Λ_{min} is

$$\Lambda_{min} = \frac{\lambda}{2n}$$

In the general case, the grating period can be expressed by using the angle φ between the two wave vectors (see fig. 3.1):

$$|\boldsymbol{e}_1 - \boldsymbol{e}_2| = \sqrt{(\boldsymbol{e}_1 - \boldsymbol{e}_2)\cdot(\boldsymbol{e}_1 - \boldsymbol{e}_2)} = \sqrt{2 - 2\boldsymbol{e}_1\cdot\boldsymbol{e}_2}\ \Rightarrow\ \Lambda = \frac{\lambda}{n\sqrt{2\,(1-\cos\varphi)}} = \frac{\lambda}{2n\sin\left(\frac{\varphi}{2}\right)} \tag{3.1.15}$$

Here, the trigonometric identity $\sqrt{2(1-\cos\varphi)} = 2\sin(\varphi/2)$ has been used. The result for Λ can also be graphically derived from fig. 3.1.
What is really observed are the lines of intersection of the planes of constant intensity with a detector plane. The resulting lines are called the **interference fringes**. In the case of the interference of two plane waves the interference fringes are straight, parallel and equidistant lines with the distance p, called the period of the fringes. Only in the case that the grating vector $\boldsymbol{G}$ is parallel to the detector plane the fringe period p is equal to the grating period Λ. In the general case only the component of the grating vector parallel to the detector plane has to be used to calculate the fringe period. This relation can be easily seen by taking the plane with $z = 0$ as detector plane. Then the fringes in the x–y–plane are described in analogy to equation (3.1.12) by:

$$(\boldsymbol{k}_{x,1} - \boldsymbol{k}_{x,2})\, x + (\boldsymbol{k}_{y,1} - \boldsymbol{k}_{y,2})\, y = \boldsymbol{G}_{\parallel} \cdot \boldsymbol{r} = \text{constant} \tag{3.1.16}$$

In analogy to equation (3.1.14) the result is

$$p = \frac{2\pi}{\left|\boldsymbol{G}_{\parallel}\right|} = \frac{2\pi}{\left|\boldsymbol{G} - (\boldsymbol{G}\cdot\boldsymbol{N})\,\boldsymbol{N}\right|} \tag{3.1.17}$$

whereby $\boldsymbol{N}$ is a unit vector perpendicular to the detector plane. By defining the two angles of incidence β_1 and β_2 of the plane waves onto the detector plane as

$$\cos\beta_1 := \boldsymbol{e}_1 \cdot \boldsymbol{N}; \quad \cos\beta_2 := \boldsymbol{e}_2 \cdot \boldsymbol{N};$$

the fringe period p can be written as:

$$\begin{aligned} p &= \frac{2\pi}{\sqrt{\left(\boldsymbol{G} - (\boldsymbol{G}\cdot\boldsymbol{N})\,\boldsymbol{N}\right)^2}} = \frac{2\pi}{\sqrt{|\boldsymbol{G}|^2 - (\boldsymbol{G}\cdot\boldsymbol{N})^2}} = \\ &= \frac{2\pi}{\sqrt{|\boldsymbol{k}_1 - \boldsymbol{k}_2|^2 - ((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{N})^2}} = \frac{\lambda}{n\sqrt{2\,(1-\cos\varphi) - (\cos\beta_1 - \cos\beta_2)^2}} \\ \text{or} & \\ p &= \frac{2\pi}{\sqrt{|\boldsymbol{G}|^2 - (\boldsymbol{G}\cdot\boldsymbol{N})^2}} = \frac{\Lambda}{\sqrt{1-\cos^2\alpha}} = \frac{\Lambda}{\sin\alpha} = \frac{\lambda}{n\sqrt{2\,(1-\cos\varphi)}\sin\alpha} \\ &= \frac{\lambda}{2n\sin\left(\frac{\varphi}{2}\right)\sin\alpha} \end{aligned} \tag{3.1.18}$$

In the last equation the angle α between the grating vector $\boldsymbol{G}$ and the surface normal $\boldsymbol{N}$, i.e. the angle α between the grating planes and the detector plane, is defined as $\cos\alpha = (\boldsymbol{G}\cdot\boldsymbol{N})/|\boldsymbol{G}|$. From equation (3.1.18) it can be seen that the fringe period is infinity if the grating planes are parallel to the detector plane, and that the fringe period is minimal if the grating planes are perpendicular to the detector plane, i.e. if $\boldsymbol{G}$ is parallel to the detector plane. Then, the fringe period p is equal to the grating period Λ.

3.2 Interference of plane waves with different polarization

In the calculation of the grating period and the fringe period we assumed that the interference terms in equation (3.1.11) are different from zero so that interference occurs. But, this is not always the case as we will discuss now.

A quite interesting quantity in interferometry is the **visibility** V of the interference fringes. It is defined as:

$$V := \frac{I_{max} - I_{min}}{I_{max} + I_{min}} \tag{3.2.1}$$

whereby I_{max} is the maximum intensity and I_{min} the minimum intensity at a point in the interference pattern when the phase (i.e. the argument of the cosine function) of the interference terms is varied in a range of 2π. For the interference of plane waves the maximum and minimum intensity can also be taken at different points because the intensity of the two single waves is then independent of the position. The visibility can vary between 0 for $I_{min} = I_{max}$, i.e. no interference occurs, and 1 for $I_{min} = 0$.

3.2.1 Linearly polarized plane waves

For the case of linearly polarized plane waves the phase constants of each wave $\delta_{\parallel,1}$ and $\delta_{\perp,1}$ on the one hand and $\delta_{\parallel,2}$ and $\delta_{\perp,2}$ on the other are equal or differ only by π. There are in fact two effective different cases

$$\left.\begin{array}{r} \delta_{\parallel,1} = \delta_{\perp,1} \ \wedge \ \delta_{\parallel,2} = \delta_{\perp,2} \\ \delta_{\parallel,1} = \delta_{\perp,1} + \pi \ \wedge \ \delta_{\parallel,2} = \delta_{\perp,2} + \pi \end{array}\right\} \Rightarrow \quad \begin{array}{l} \delta_\perp = \delta_{\perp,1} - \delta_{\perp,2} = \\ = \delta_{\parallel,1} - \delta_{\parallel,2} = \delta_\parallel = \\ =: \delta \text{ and } s := +1 \end{array} \tag{3.2.2}$$

$$\left.\begin{array}{r} \delta_{\parallel,1} = \delta_{\perp,1} + \pi \ \wedge \ \delta_{\parallel,2} = \delta_{\perp,2} \\ \delta_{\parallel,1} = \delta_{\perp,1} \ \wedge \ \delta_{\parallel,2} = \delta_{\perp,2} + \pi \end{array}\right\} \Rightarrow \quad \begin{array}{l} \delta_\perp = \delta_{\perp,1} - \delta_{\perp,2} = \\ \delta_{\parallel,1} - \delta_{\parallel,2} \mp \pi = \delta_\parallel \mp \pi = \\ =: \delta \mp \pi \text{ and } s := -1 \end{array} \tag{3.2.3}$$

The parameter s characterizes the different cases and is either +1 or -1. Then, the intensity of the interference pattern (see equation (3.1.11)) can be expressed as:

$$\begin{aligned} I_{1+2} \ &= \ I_{\parallel,1} + I_{\parallel,2} + 2\sqrt{I_{\parallel,1} I_{\parallel,2}} \cos\varphi \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) + \\ &\quad + I_{\perp,1} + I_{\perp,2} + 2s\sqrt{I_{\perp,1} I_{\perp,2}} \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) \end{aligned} \tag{3.2.4}$$

Here, the angle φ between the directions of propagation of the two waves is used. So, there are several interesting special cases of equation (3.2.4):

- Both waves have only TE–components, i.e. $I_{\parallel,1} = I_{\parallel,2} = 0$, $I_{\perp,1} \neq 0$ and $I_{\perp,2} \neq 0$. Then, we obtain:

  $$I_{TE,TE} = I_{\perp,1} + I_{\perp,2} + 2s\sqrt{I_{\perp,1} I_{\perp,2}} \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) \tag{3.2.5}$$

  This is the well–known interference equation like it is also used for scalar waves where an arbitrary component of the electric vector, but for a small angle φ between the two wave vectors of the waves, is regarded. In this case the visibility $V_{TE,TE}$ (defined with equation (3.2.1)) of the interference pattern is:

  $$V_{TE,TE} = \frac{2\sqrt{I_{\perp,1} I_{\perp,2}}}{I_{\perp,1} + I_{\perp,2}} \tag{3.2.6}$$

  If the intensities of both waves are equal the visibility is 1.

- One wave has only a TE–component, i.e. $I_{\parallel,1} = 0$ and $I_{\perp,1} \neq 0$, and the other wave has only a TM–component, i.e. $I_{\parallel,2} \neq 0$ and $I_{\perp,2} = 0$. Then the interference terms vanish and the intensity is constant:

$$I_{TE,TM} = I_{\perp,1} + I_{\parallel,2} = \text{constant} \tag{3.2.7}$$

  This just means that orthogonally polarized waves cannot interfere.

- Both waves have only TM–components, i.e. $I_{\perp,1} = I_{\perp,2} = 0$, $I_{\parallel,1} \neq 0$ and $I_{\parallel,2} \neq 0$. Then, we obtain:

$$I_{TM,TM} = I_{\parallel,1} + I_{\parallel,2} + 2\sqrt{I_{\parallel,1} I_{\parallel,2}} \cos\varphi \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) \tag{3.2.8}$$

  The visibility $V_{TM,TM}$ is in this case:

$$V_{TM,TM} = \frac{2\sqrt{I_{\parallel,1} I_{\parallel,2}} \cos\varphi}{I_{\parallel,1} + I_{\parallel,2}} \tag{3.2.9}$$

  So, the visibility is always smaller than 1 and the interference term vanishes if both waves are propagating perpendicular to each other, i.e. $\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = \cos\varphi = 0$. For the case $I_{\parallel,1} = I_{\parallel,2}$, where the visibility is 1 in the TE–polarized case, the visibility in the TM–polarized case is $V_{TM,TM} = \cos\varphi$. Only for small angles between the directions of propagation of the two plane waves the visibility is high. Of course, for small angles there is in fact no real difference between TE– and TM–polarized light, and for $\varphi = 0$ or $\varphi = \pi$ the difference between TE and TM vanishes at all.

- TE– and TM–components of both waves are present. Then the value of the constant s is important. For $s = +1$ the interference terms have identical signs and add. But for $s = -1$ the interference terms have different signs and cancel out each other if

$$\sqrt{I_{\parallel,1} I_{\parallel,2}} \cos\varphi = \sqrt{I_{\perp,1} I_{\perp,2}} \tag{3.2.10}$$

  The meaning of this is that for $s = -1$ the electric vectors of both waves are oscillating in different quadrants. If the above equation is fulfilled they are again perpendicular to each other and cannot interfere. This is explained in fig. 3.2.

3.2.2 Circularly polarized plane waves

Circularly polarized plane waves exist according to equation (2.1.2) only if:

$$I_{\parallel,1} = I_{\perp,1} = \frac{1}{2} I_1; \quad \delta_{\parallel,1} - \delta_{\perp,1} = \pm\frac{\pi}{2}$$
$$I_{\parallel,2} = I_{\perp,2} = \frac{1}{2} I_2; \quad \delta_{\parallel,2} - \delta_{\perp,2} = \pm\frac{\pi}{2}$$

Then, the intensity of the interference pattern is according to equation (3.1.11):

$$\begin{aligned} I_{1+2} &= \frac{1}{2}\left[I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\varphi\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta_\parallel\right)\right] + \\ &\quad + \frac{1}{2}\left[I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta_\perp\right)\right] = \\ &= I_1 + I_2 + \\ &\quad + \sqrt{I_1 I_2}\left[\cos\varphi\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta_\parallel\right) + \cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta_\perp\right)\right] \end{aligned} \tag{3.2.11}$$

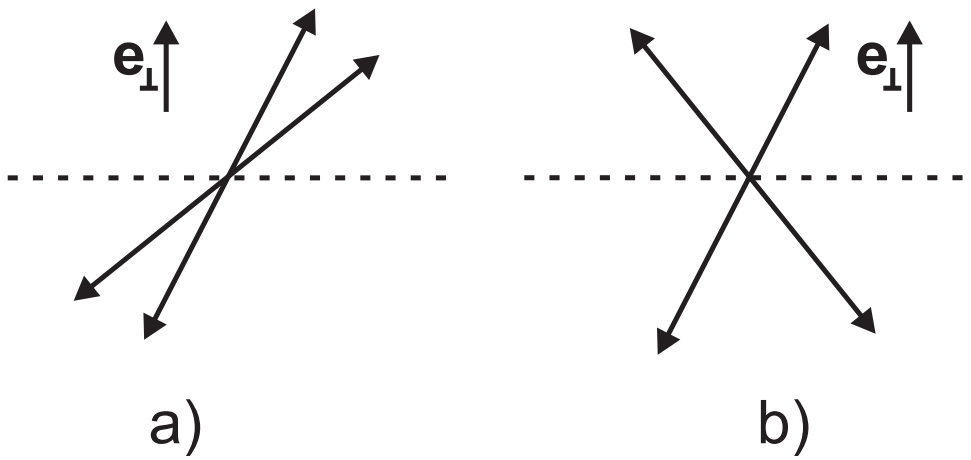


Figure 3.2: The different possibilities for the interference of linearly polarized waves. a) $s = +1$, i.e. the two electric vectors are oscillating in the same quadrant; b) $s = -1$, i.e. the two electric vectors are oscillating in different quadrants. The dashed line indicates the plane in which the wave vectors of both waves are situated.

Now, we have to differ between several cases:

- Both waves have the same chirality, i.e. either $\delta_{\parallel,1} - \delta_{\perp,1} = \pi/2$ and $\delta_{\parallel,2} - \delta_{\perp,2} = \pi/2$ or $\delta_{\parallel,1} - \delta_{\perp,1} = -\pi/2$ and $\delta_{\parallel,2} - \delta_{\perp,2} = -\pi/2$. Then, the phase differences are

$$\delta_\parallel = \delta_{\parallel,1} - \delta_{\parallel,2}; \quad \delta_\perp = \delta_{\perp,1} - \delta_{\perp,2} = \delta_\parallel =: \delta$$

  and equation (3.2.11) reduces to:

$$I_{\uparrow\uparrow} = I_1 + I_2 + \sqrt{I_1 I_2}\,[\cos\varphi + 1]\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) \tag{3.2.12}$$

  The visibility (see equation (3.2.1)) is:

$$V_{\uparrow\uparrow} = \frac{\sqrt{I_1 I_2}\,[\cos\varphi + 1]}{I_1 + I_2} \tag{3.2.13}$$

  For small angles φ between the directions of propagation of the two waves the intensity is identical to the interference pattern of two linearly TE–polarized waves and the visibility can reach 1 for equal intensities in both waves. If both waves are propagating perpendicular to each other the interference term has only half the size and the visibility is for equal intensities only 1/2. If the angle φ is larger than $\pi/2$ and approaches π the interference term and the visibility vanish. This means that waves which are propagating antiparallel and have the same chirality cannot interfere.

- Both waves have different chirality, i.e. either $\delta_{\parallel,1} - \delta_{\perp,1} = \pi/2$ and $\delta_{\parallel,2} - \delta_{\perp,2} = -\pi/2$ or $\delta_{\parallel,1} - \delta_{\perp,1} = -\pi/2$ and $\delta_{\parallel,2} - \delta_{\perp,2} = \pi/2$. Then, the phase differences are

$$\delta_\parallel = \delta_{\parallel,1} - \delta_{\parallel,2} =: \delta; \quad \delta_\perp = \delta_{\perp,1} - \delta_{\perp,2} = \delta_\parallel \pm \pi = \delta \pm \pi$$

  and equation (3.2.11) reduces to:

$$I_{\uparrow\downarrow} = I_1 + I_2 + \sqrt{I_1 I_2}\,[\cos\varphi - 1]\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) \tag{3.2.14}$$

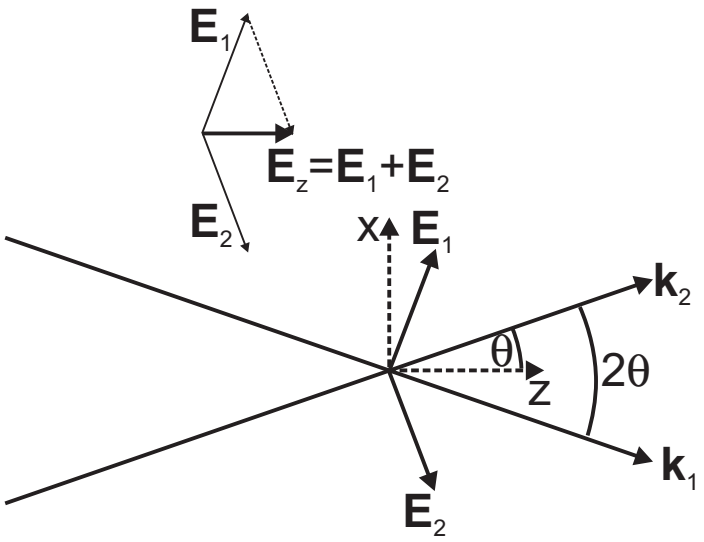


Figure 3.3: The interference of two linearly polarized waves with TM polarization and π phase difference results in a longitudinal component $\boldsymbol{E}_z$ of the electric field.

The visibility (see equation (3.2.1)) is:

$$V_{\uparrow\downarrow} = \frac{\sqrt{I_1 I_2}\,[1 - \cos\varphi]}{I_1 + I_2} \tag{3.2.15}$$

So, for different chirality the behavior is reverse to that of equal chirality. Waves which are propagating parallel ($\varphi = 0$) and have different chirality cannot interfere, whereas waves which are propagating antiparallel ($\varphi = \pi$) and have different chirality interfere very well.

3.2.3 The application of two beam interference for an electron accelerator

A quite interesting modern application of the interference of two waves is the laser–driven electron accelerator [25],[26]. This accelerator can in principle also be used for any other charged particle whereby the efficiency increases if the velocity of the particle is nearly identical to the speed of light. So, heavy particles need a lot of kinetic energy (at best a multiple of their rest energy (*dt.:* Ruheenergie) m_0c^2) if they shall be accelerated by laser light with a good efficiency. Here, only the basic principle can be discussed.
Fig. 3.3 shows two interfering waves which both have linear TM polarization, equal amplitudes and a phase difference of π on points along the z–axis. First of all we assume that the waves are plane. The z–axis bisects the angle 2θ between the wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ of both waves so that θ is the angle between the z–axis and one of the wave vectors. Then, the electric vectors $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are oriented as indicated and the components perpendicular to the z–axis, i.e. parallel to the x–axis, cancel each other. The magnetic vectors of both waves which are perpendicular to the drawing plane, i.e. parallel to the y–axis, also cancel each other on points along the z–axis because they are antiparallel. However, due to the configuration of the two interfering waves there is a resulting longitudinal component $\boldsymbol{E}_z$ of the electric vector for points on the z–axis. From a mathematical point of view we have for the electric vectors using the coordinate system

of fig. 3.3:

$$\begin{aligned}
\boldsymbol{E}_1(x,z,t) &= \begin{pmatrix} E_0\cos\theta\cos\left(\frac{2\pi}{\lambda}\left(-x\sin\theta+z\cos\theta\right)-\omega t\right) \\ 0 \\ E_0\sin\theta\cos\left(\frac{2\pi}{\lambda}\left(-x\sin\theta+z\cos\theta\right)-\omega t\right) \end{pmatrix} \\
\boldsymbol{E}_2(x,z,t) &= \begin{pmatrix} -E_0\cos\theta\cos\left(\frac{2\pi}{\lambda}\left(x\sin\theta+z\cos\theta\right)-\omega t\right) \\ 0 \\ E_0\sin\theta\cos\left(\frac{2\pi}{\lambda}\left(x\sin\theta+z\cos\theta\right)-\omega t\right) \end{pmatrix} \\
\Rightarrow\ \boldsymbol{E}_z(x{=}0,z,t) &= \boldsymbol{E}_1(x{=}0,z,t)+\boldsymbol{E}_2(x{=}0,z,t) = \\
&= \begin{pmatrix} 0 \\ 0 \\ 2E_0\sin\theta\cos\left(\frac{2\pi}{\lambda}z\cos\theta-\omega t\right) =: E_z(0,z,t) \end{pmatrix}
\end{aligned} \tag{3.2.16}$$

Hereby, E_0 is the maximum amplitude of the electric vector of one of the two interfering waves ($E_0 = \max|\boldsymbol{E}_1| = \max|\boldsymbol{E}_2|$), λ is the wavelength of the waves and ω is their angular frequency. So, we see that along the z–axis with $x = 0$ there exists only a component of the electric field parallel to the z–axis and by using the relations $\lambda\nu = c$ and $\omega = 2\pi\nu = 2\pi c/\lambda$ (ν is the frequency) for our waves which are situated in vacuum, we obtain:

$$E_z(x{=}0,z,t) = 2E_0\sin\theta\cos\left(2\pi\frac{c}{\lambda}\left(z\frac{\cos\theta}{c}-t\right)\right) \tag{3.2.17}$$

Now, a relativistic electron, i.e. the speed v of the electron should be nearly the speed of light c, travels along the z–axis from left to right. It should pass the regarded point at that time where $E_z = -2E_0\sin\theta$. Then, the electron will be maximally accelerated along the z–axis due to its negative electric charge. But, the interference maximum with $E_z = -2E_0\sin\theta$ seems to propagate along the z–axis with the phase velocity $c/\cos\theta$ which is for $\theta > 0$ faster than the speed of light. Of course, this is in no contradiction to special relativity because no information nor energy is transported with this speed. So, for a small angle θ the relativistic electron with $v \approx c$ (but nevertheless $v < c$) travels a certain distance nearly in phase with the accelerating electric field before it comes out of phase because of $c/\cos\theta > c > v$. The distance after which the electron is out of phase can also be calculated very easily. The electron travels in the laboratory framework in the time interval t a distance $z = vt$, i.e. $t = z/v$. The velocity v can be assumed to be constant during the acceleration process because it should be nearly the speed of light and therefore it does not change considerably although the electron may gain a lot of kinetic energy. By introducing this into equation (3.2.17) the argument Φ of the cosine function as a function of the position z on the z–axis is:

$$\Phi(z) = 2\pi\frac{c}{\lambda}\left(z\frac{\cos\theta}{c}-t\right) = 2\pi\frac{c}{\lambda}z\left(\frac{\cos\theta}{c}-\frac{1}{v}\right) \tag{3.2.18}$$

If the phase Φ of the electric field which affects the electron changes by $\pm\pi$ the electric field which first accelerated the electron will now slow it down. So, the distance Δz on the z–axis between being accelerated and decelerated by the electric field is:

$$\Delta\Phi = 2\pi\frac{c}{\lambda}\Delta z\left(\frac{\cos\theta}{c}-\frac{1}{v}\right) = \pm\pi \ \Rightarrow\ \Delta z = \pm\frac{\lambda}{2\left(\cos\theta - c/v\right)} \tag{3.2.19}$$

For $v \approx c$ the longest distance for being in phase would be achieved by $\theta = 0$. But then, the accelerating electric field itself would be zero because of the factor $\sin\theta$ in equation (3.2.17). So, in practice a trade–off has to be found between a big value $\sin\theta$ and an also big value $\cos\theta$. Additionally, the velocity of the electron (or other charged particle) should be nearly the speed of light.

However, we see that for the interference of two infinitely extended plane waves the electron would be accelerated and slowed down periodically. But, we can replace the plane waves by focused laser beams, i.e. Gaussian beams (see section 6), so that the region with a high electric field has a quite limited length smaller than Δz. The beam waist of each laser beam should be at the crossing point of the two Gaussian beams on the z–axis to achieve a high amplitude of the electric vector. Then, the electric field amplitude E_0 is not constant along the z–axis but decreases outside of the beam waist like a Gaussian curve. So, it is possible to achieve a net acceleration of the electron if it is harmonized with the phase of the electric field of the laser beams when it crosses the beam waist. If the electron is not harmonized with the phase of the laser beams it can also be slowed down. The concrete calculation in the case of Gaussian beams is of course a little bit more complex than with plane waves because the wave vector and therefore the direction of the electric vector changes locally in the case of a Gaussian beam. By using ultra–short and focused laser pulses the resulting electric field in the beam waist which accelerates the electron can be some GV/m up to about 1 TV/m. Of course, the acceleration distance is only as long as the beam waist, i.e. several μm for a strongly focused laser beam. So, some keV or even MeV of kinetic energy can be gained by the electron. But, by repeating many acceleration devices in series (and of course harmonized in phase) the effective acceleration distance can be increased so that the laser–driven electron accelerator may become an alternative to conventional particle accelerators in future.

3.3 Interference of arbitrary scalar waves

The interference phenomena are of course not restricted to plane waves but can occur for arbitrary waves. There, the polarization can change locally, but we will neglect the polarization in this section and concentrate on so called **scalar waves**.

3.3.1 Some notes to scalar waves

In the case of scalar waves only one cartesian component of the electric (or magnetic) vector is regarded and the complete polarization state is neglected. Nevertheless, for two linearly polarized interfering waves which are both TE–polarized the result of the scalar calculation is identical to the exact result. Also, for linearly TM–polarized waves the result is approximately correct as long as the angle between both waves is not too high. This can be seen quite well by looking at equation (3.2.8) which contains the cosine of the angle between both waves as additional factor in the interference equation. But, for angles of less than 25^o the cosine factor just varies between 1.0 and 0.9 so that the intensity pattern and its visibility just vary slightly. Nevertheless, since the orthogonality condition is neglected for scalar waves the result of the scalar wave equation is not automatically a solution of the Maxwell equations.

A scalar wave which is often used in optics is a spherical wave with its center of curvature at

the point $\boldsymbol{r}_0$ with the complex amplitude

$$u(\boldsymbol{r}) = a\frac{e^{ik|\boldsymbol{r}-\boldsymbol{r}_0|}}{|\boldsymbol{r}-\boldsymbol{r}_0|} \tag{3.3.1}$$

and the modulus of the wave vector $k = 2\pi n/\lambda$ which is also called **wave number** (*dt.:* Wellenzahl). a is a constant. It should be mentioned that a spherical wave is a solution of the scalar Helmholtz equation of homogeneous materials (see section 1.5), where only one component of the electric or magnetic vector is regarded. But, a spherical wave is not a solution of the Maxwell equations itself because this would violate the orthogonality conditions of an electromagnetic wave. But, a dipole radiation is in the far field and in the plane perpendicular to its dipole axis a good approximation for a spherical wave.
Here, the scalar complex amplitude u was introduced which can stand, apart from a constant of proportionality, for one component of the time–independent complex electric or magnetic vector $\hat{\boldsymbol{E}}$ or $\hat{\boldsymbol{H}}$ of a monochromatic wave. Again, the intensity of this scalar wave is at least proportional to the square of the modulus of u and in the following we just define the scalar complex amplitude u in such a way that it is:

$$I := uu^* \tag{3.3.2}$$

A general scalar wave can be described by

$$u(\boldsymbol{r}) = A(\boldsymbol{r})\, e^{i\Phi(\boldsymbol{r})} \tag{3.3.3}$$

In this case, A is a real function which changes only slowly with the position $\boldsymbol{r}$, whereby Φ is also a real function but the complex exponential factor $\exp(i\Phi)$ varies rapidly with the position $\boldsymbol{r}$.

3.3.2 The interference equation for scalar waves

By using two general scalar waves (see equation (3.3.3))

$$u_1(\boldsymbol{r}) = A_1(\boldsymbol{r})\, e^{i\Phi_1(\boldsymbol{r})}; \quad u_2(\boldsymbol{r}) = A_2(\boldsymbol{r})\, e^{i\Phi_2(\boldsymbol{r})}$$

instead of using plane waves in equation (3.1.11) or equation (3.2.5) the interference equation of scalar waves is obtained:

$$I_{1+2} = (u_1 + u_2)(u_1^* + u_2^*) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Phi \tag{3.3.4}$$

with $I_1 = A_1^2$, $I_2 = A_2^2$ and $\Phi = \Phi_1 - \Phi_2$. In some cases, it is more convenient to write this equation as:

$$I_{1+2} = I_0\,[1 + V\cos\Phi] \tag{3.3.5}$$

Here, $I_0 = I_1 + I_2$ is defined as the resulting intensity for incoherent light, where only the intensities of the single waves have to be added. The visibility V is defined in equation (3.2.1) and is here:

$$V = \frac{I_{max} - I_{min}}{I_{max} + I_{min}} = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} = \frac{2\sqrt{I_1 I_2}}{I_0} \tag{3.3.6}$$

For general scalar waves the fringe period in the detector plane with coordinates (x, y) is not constant but will vary. But at a point (x, y) in the neighborhood of a fixed point (x_0, y_0) the

phase function Φ can be written as a Taylor expansion neglecting all terms of second and higher order:

$$\begin{aligned}\Phi(x,y) &\approx \Phi(x_0,y_0) + \begin{pmatrix} \frac{\partial\Phi(x_0,y_0)}{\partial x} \\ \frac{\partial\Phi(x_0,y_0)}{\partial y} \end{pmatrix} \cdot \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} = \\ &= \Phi(x_0,y_0) + \boldsymbol{\nabla}_\perp \Phi(x_0,y_0) \cdot \boldsymbol{\Delta r} \end{aligned} \tag{3.3.7}$$

Here, the two–dimensional Nabla operator $\boldsymbol{\nabla}_\perp$ is introduced. The local fringe period p is defined as the distance between two fringes, i.e. the distance where the phase function increases or decreases by 2π taken along a path parallel to the local phase gradient. Therefore, we have for the vector $\boldsymbol{\Delta r}$ pointing from one fringe to the next neighbored fringe at the position (x_o, y_0):

$$\begin{aligned}&\boldsymbol{\Delta r} = p\frac{\boldsymbol{\nabla}_\perp\Phi}{|\boldsymbol{\nabla}_\perp\Phi|} \;\Rightarrow\; \boldsymbol{\nabla}_\perp\Phi\cdot\boldsymbol{\Delta r} = p|\boldsymbol{\nabla}_\perp\Phi| = 2\pi \\ &\Rightarrow\; p = \frac{2\pi}{|\boldsymbol{\nabla}_\perp\Phi|} = \frac{2\pi}{\sqrt{\left(\frac{\partial\Phi}{\partial x}\right)^2 + \left(\frac{\partial\Phi}{\partial y}\right)^2}} \end{aligned} \tag{3.3.8}$$

All quantities have to be calculated at the point (x_0, y_0). By comparing this equation with equation (3.1.17) it is clear that the component $\boldsymbol{G}_\parallel$ of the grating vector in the x–y–plane is defined by:

$$\boldsymbol{G}_\parallel = \begin{pmatrix} \frac{\partial\Phi}{\partial x} \\ \frac{\partial\Phi}{\partial y} \\ 0 \end{pmatrix} \tag{3.3.9}$$

The grating vector $\boldsymbol{G}$ itself is defined as:

$$\boldsymbol{G} = \boldsymbol{\nabla}\Phi = \begin{pmatrix} \frac{\partial\Phi}{\partial x} \\ \frac{\partial\Phi}{\partial y} \\ \frac{\partial\Phi}{\partial z} \end{pmatrix} \tag{3.3.10}$$

with Φ defined as a function of (x, y, z).
The local fringe frequency ν is defined as the reciprocal of the local fringe period

$$\nu = \frac{1}{p} = \frac{\sqrt{\left(\frac{\partial\Phi}{\partial x}\right)^2 + \left(\frac{\partial\Phi}{\partial y}\right)^2}}{2\pi} \tag{3.3.11}$$

and describes the number of fringes per length unit. If the fringe frequency is too high the interference pattern cannot be resolved in practice since common detector arrays like a CCD camera have only a limited number of pixels per length unit and integrate the light intensity over the area of one pixel.

3.3.3 Interference of scalar spherical and plane waves

The interference of two plane waves is investigated in detail in the last section for general polarization states. Simple examples for the interference of two scalar waves are the interference

of a spherical wave and a plane wave or the interference of two spherical waves. In principle, the general statements to the effects of different polarization states hold also for spherical waves as long as the numerical aperture is not too high. Moreover, the local fringe period or grating vector for a general polarization state is calculated correctly with the scalar approach. It can only be that for a general polarization state the visibility of the interference pattern may be so small that the interference is not visible in practice (for example for orthogonally polarized waves). So, the investigation of scalar waves is not really a restriction as long as we keep in mind that for generally polarized light there may be some changes in the visibility of the interference pattern.
Two spherical waves with the wavelength λ and the corresponding wave number $k = 2\pi n/\lambda$ having their center of curvature at the points $\boldsymbol{r}_1 = (x_1, y_1, z_1)$ and $\boldsymbol{r}_2 = (x_2, y_2, z_2)$, respectively, have the complex amplitude functions:

$$\begin{aligned} u_1(\boldsymbol{r}) &= a_1 \frac{e^{ik\,|\boldsymbol{r}-\boldsymbol{r}_1|}}{|\boldsymbol{r}-\boldsymbol{r}_1|} \\ u_2(\boldsymbol{r}) &= a_2 \frac{e^{ik\,|\boldsymbol{r}-\boldsymbol{r}_2|}}{|\boldsymbol{r}-\boldsymbol{r}_2|} \end{aligned}$$

In the following only the interference pattern in the x–y–plane at $z = 0$ in an area centered around the origin of the coordinate system at $x = y = z = 0$ is evaluated. Additionally, the distances $|\boldsymbol{r}_i|$ $(i = 1, 2)$ of the centers of curvature of both spherical waves to the origin of the coordinate system shall be large compared to the maximum distance $|\boldsymbol{r}| = \sqrt{x^2 + y^2}$ of the origin of the coordinate system to a point lying in the evaluated aperture of the interference pattern. Then the amplitude of the spherical waves can be assumed to be constant because of:

$$\begin{aligned} &|\boldsymbol{r}_i| \gg |\boldsymbol{r}| \;\Rightarrow \\ &|\boldsymbol{r}-\boldsymbol{r}_i| = \sqrt{(\boldsymbol{r}-\boldsymbol{r}_i)\cdot(\boldsymbol{r}-\boldsymbol{r}_i)} = \sqrt{|\boldsymbol{r}_i|^2 + |\boldsymbol{r}|^2 - 2\,|\boldsymbol{r}_i|\,|\boldsymbol{r}|\cos\alpha} \approx |\boldsymbol{r}_i| \end{aligned}$$

Here, α is the angle between the two vectors $\boldsymbol{r}$ and $\boldsymbol{r}_i$. Therefore, the two spherical waves are written as

$$\begin{aligned} u_1(\boldsymbol{r}) &= A_1 e^{ik\,|\boldsymbol{r}-\boldsymbol{r}_1|} \\ u_2(\boldsymbol{r}) &= A_2 e^{ik\,|\boldsymbol{r}-\boldsymbol{r}_2|} \end{aligned}$$

with constant amplitudes $A_1 = a_1/|\boldsymbol{r}_1|$ and $A_2 = a_2/|\boldsymbol{r}_2|$. The arguments of the complex exponential functions can of course not be replaced by the constant terms $|\boldsymbol{r}_i|$ because the complex exponential functions are very fast oscillating functions. Then, the intensity I_{1+2} of the interference pattern is according to equation (3.3.4):

$$\begin{aligned} &I_{1+2}(x,y) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Phi(x,y) \quad \text{with} \quad I_1 = A_1^2;\; I_2 = A_2^2 \quad \text{and} \\ &\Phi(x,y) = k\,(|\boldsymbol{r}-\boldsymbol{r}_1| - |\boldsymbol{r}-\boldsymbol{r}_2|) = \\ &= k\left(|\boldsymbol{r}_1|\sqrt{1 + \frac{x^2+y^2-2xx_1-2yy_1}{|\boldsymbol{r}_1|^2}} - \right. \\ &\left. - |\boldsymbol{r}_2|\sqrt{1 + \frac{x^2+y^2-2xx_2-2yy_2}{|\boldsymbol{r}_2|^2}}\right) \end{aligned} \tag{3.3.12}$$

The square roots can be developed into a Taylor series according to $\sqrt{1+x} \approx 1 + x/2 - x^2/8$ for $x \ll 1$. Since $|\boldsymbol{r}_i| \gg |\boldsymbol{r}|$ the most important terms are:

$$\begin{aligned} &|\boldsymbol{r}_i| \sqrt{1 + \frac{x^2 + y^2 - 2xx_i - 2yy_i}{|\boldsymbol{r}_i|^2}} \approx \\ &\approx |\boldsymbol{r}_i| + \frac{x^2 + y^2}{2\,|\boldsymbol{r}_i|} - \frac{xx_i + yy_i}{|\boldsymbol{r}_i|} - \frac{\left(x^2 + y^2 - 2xx_i - 2yy_i\right)^2}{8\,|\boldsymbol{r}_i|^3} \end{aligned} \tag{3.3.13}$$

The last term can be neglected for

$$k\frac{\left(x^2 + y^2 - 2xx_i - 2yy_i\right)^2}{8\,|\boldsymbol{r}_i|^3} \ll 1 \;\Rightarrow\; \frac{\left(x^2 + y^2 - 2xx_i - 2yy_i\right)^2}{8\,|\boldsymbol{r}_i|^3} \ll \frac{\lambda}{2\pi n} \tag{3.3.14}$$

By using spherical coordinates $r_i, \vartheta_i, \varphi_i$ for $\boldsymbol{r}_i$, whereby r_i is the distance from the origin, ϑ_i is the polar angle and φ_i is the azimuthal angle,

$$x_i = r_i \cos\varphi_i \sin\vartheta_i \tag{3.3.15}$$

$$y_i = r_i \sin\varphi_i \sin\vartheta_i \tag{3.3.16}$$

$$z_i = r_i \cos\vartheta_i \tag{3.3.17}$$

the condition for neglecting the last term is:

$$\begin{aligned} &\frac{\left(x^2 + y^2\right)^2}{4r_i^3} - \frac{\left(x^2 + y^2\right)(x\cos\varphi_i + y\sin\varphi_i)\sin\vartheta_i}{r_i^2} + \\ &+ \frac{(x\cos\varphi_i + y\sin\varphi_i)^2 \sin^2\vartheta_i}{r_i} \ll \frac{\lambda}{\pi n} \end{aligned} \tag{3.3.18}$$

If this condition is fulfilled for both spherical waves the phase function of the interference pattern can be written as:

$$\begin{aligned} \Phi(x,y) \approx \delta &+ \frac{\pi n}{\lambda}\left(\frac{1}{r_1} - \frac{1}{r_2}\right)\left(x^2 + y^2\right) - \\ &- \frac{2\pi n}{\lambda}\left(\cos\varphi_1 \sin\vartheta_1 - \cos\varphi_2 \sin\vartheta_2\right) x - \frac{2\pi n}{\lambda}\left(\sin\varphi_1 \sin\vartheta_1 - \sin\varphi_2 \sin\vartheta_2\right) y \end{aligned} \tag{3.3.19}$$

The phase constant δ is defined as $\delta = 2\pi n(r_1 - r_2)/\lambda$.
The term which depends on $x^2 + y^2$ is called **defocus** and is proportional to the difference of the curvatures of both spherical waves. The linear terms in x and y are called **tilts** and are only present if the centers of curvature of the two spherical waves and the origin are not lying on a common line. In the interferometric testing of spherical surfaces or in the measurement of the wave aberrations of lenses there appear often defocus and tilts due to an axial ($\Rightarrow$ defocus) or lateral ($\Rightarrow$ tilts) misalignment of the test object. Then the coefficients of these terms are determined by a least squares fit of the function

$$\Phi_{misalign} = a + bx + cy + d\left(x^2 + y^2\right) \tag{3.3.20}$$

to the measured phase function $\Phi_{measured}$. Afterwards, the phase function $\Phi_{reduced}$ which is freed from misalignment aberrations is calculated by

$$\Phi_{reduced} = \Phi_{measured} - \Phi_{misalign} \tag{3.3.21}$$

using the fitted coefficients a, b, c and d. The phase function $\Phi_{reduced}$ contains then only the desired wave aberrations or the desired surface deviations from the ideal surface plus systematic errors of the experimental setup.
A special case of equation (3.3.19) is that one of the waves is a plane wave. Without loss of generality the second wave shall be plane. This means that the parameter r_2 is infinity. Then equation (3.3.19) reduces to:

$$\begin{aligned}\Phi(x,y) \approx \delta + \frac{\pi n}{\lambda r_1}\left(x^2+y^2\right) - \\ -\frac{2\pi n}{\lambda}\left(\cos\varphi_1 \sin\vartheta_1 - e_{x,2}\right)x - \frac{2\pi n}{\lambda}\left(\sin\varphi_1 \sin\vartheta_1 - e_{y,2}\right)y\end{aligned} \tag{3.3.22}$$

whereby $e_{x,2} := \cos\varphi_2 \sin\vartheta_2$ and $e_{y,2} := \sin\varphi_2 \sin\vartheta_2$ are the x– and y–components of the unit vector $\boldsymbol{e}_2 = \boldsymbol{k}_2/|\boldsymbol{k}_2|$ parallel to the wave vector $\boldsymbol{k}_2$ of the plane wave. Of course, the phase constant δ which is only defined modulus 2π is then not infinity but has a certain value which depends on r_1 and the phase offset of the plane wave at $(x = 0, y = 0, z = 0)$. In the case of the interference of a spherical and a plane wave the defocus term is directly proportional to the curvature of the spherical wave. The tilt terms depend again on both waves but for the case that either the spherical wave has its center of curvature at $x_1 = y_1 = 0$ ($\Rightarrow \sin\vartheta_1 = 0$) or the plane wave is perpendicular to the x–y–plane ($\Rightarrow e_{x,2} = e_{y,2} = 0$) the tilt terms depend only on the parameters of one wave.

3.3.4 Two examples of interference patterns

Assume that we have two interfering monochromatic waves with a wavelength λ=0.5 μm. Additionally, we know that the first of the waves is a plane wave which propagates parallel to the optical axis which is defined to be parallel to the surface normal of the detector and to intersect the detector in its center. So, the parameters of the first wave in equation (3.3.19) are $r_1 \to \infty$, $\vartheta_1 = 0$ and $\varphi_1 = 0$ and the phase function Φ of the interference pattern depends only on the parameters of the second wave:

$$\Phi(x,y) \approx \delta - \frac{\pi}{\lambda r_2}\left(x^2+y^2\right) + \frac{2\pi}{\lambda}\cos\varphi_2 \sin\vartheta_2\, x + \frac{2\pi}{\lambda}\sin\varphi_2 \sin\vartheta_2\, y \tag{3.3.23}$$

The refractive index n has been set to 1 because the measurements should be made in air.
Let us now assume that we detect the interference pattern which is displayed in fig. 3.4. Such an interference pattern is often called an **interferogram**. In this case it has straight, parallel and equidistant fringes parallel to the y–axis. Therefore, the second wave has also to be a plane wave ($r_2 \to \infty$) which can be described by the two angles φ_2 and ϑ_2. The interference pattern changes only along the x–axis and there the period p is 0.2 mm. Therefore, the phase function Φ has to be of the form $\Phi(x,y) = ax$ with $a = 2\pi/p = 10\pi/\text{mm}$. Comparing this with equation

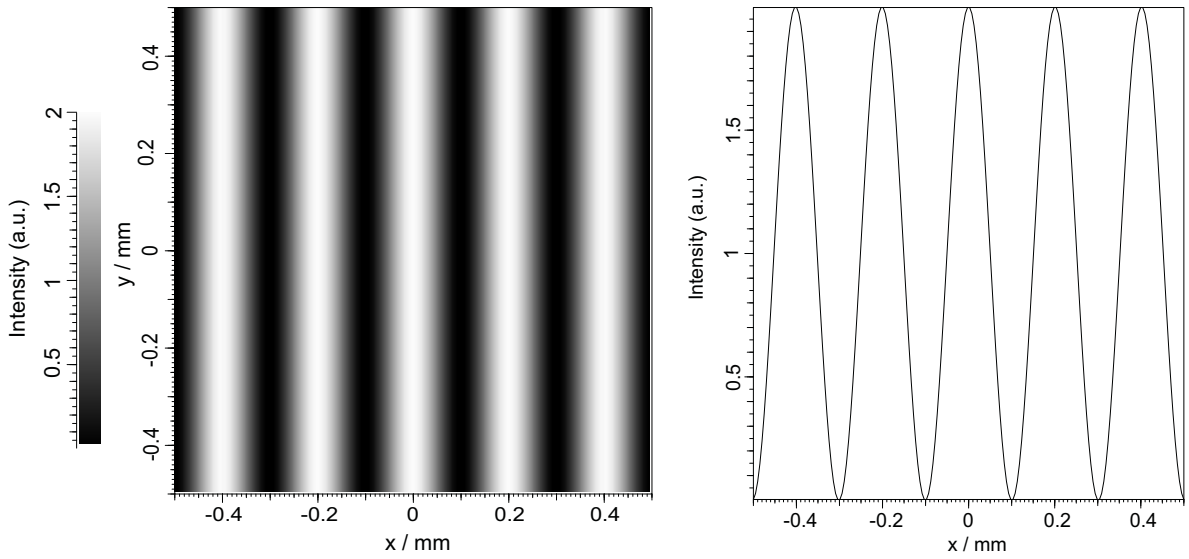


Figure 3.4: Example of an interference pattern (interferogram) with straight, parallel and equidistant fringes. Left: (simulated) camera picture as it can also be seen directly with the eye, right: section through the intensity function.

(3.3.23) results in:

$$\delta + \frac{2\pi}{\lambda}\cos\varphi_2 \sin\vartheta_2\, x + \frac{2\pi}{\lambda}\sin\varphi_2 \sin\vartheta_2\, y = ax$$
$$\Rightarrow \quad \delta = 0 \ \wedge \ \varphi_2 = 0 \ \wedge \ \sin\vartheta_2 = \frac{a\lambda}{2\pi} = \frac{\lambda}{p} = 0.0025 \tag{3.3.24}$$

By looking at the intensity pattern it can also be seen that the visibility has the maximum value of $V = (I_{max} - I_{min})/(I_{max} + I_{min}) = (2-0)/(2+0) = 1$. Therefore, the second plane wave has the same intensity as the first wave.

In a second (simulated) measurement the interferogram of fig. 3.5 is obtained. It can be estimated that the local fringe frequency increases linearly with the distance from the center and that we have therefore a defocus term with a quadratic phase function. No linear phase function is present and therefore the tilt angle ϑ_2 of the second wave has to be zero. A more detailed evaluation of the intensity pattern using the interference equation (3.3.4) confirms the estimation that the phase function is $\Phi(x,y) = b(x^2+y^2)$ with $b = 20\pi/\text{mm}^2$. Additionally, it can be seen that the visibility is $V = (1.28 - 0.72)/(1.28 + 0.72) = 0.28$.

Using equation (3.3.23) to calculate the radius of curvature r_2 of the second wave results in:

$$\Phi(x,y) = -\frac{\pi}{\lambda r_2}\left(x^2+y^2\right) = b\left(x^2+y^2\right) \ \Rightarrow \ |r_2| = \frac{\pi}{b\lambda} = 100 \text{ mm} \tag{3.3.25}$$

The sign of r_2 cannot be detected in this case where only one interferogram without a carrier frequency is known. This is quite clear because the cosine function is an even function and so $\cos\Phi = \cos(-\Phi)$.

By using equation (3.3.6) and the approach $I_2 = \alpha I_1$ the coefficient α can be calculated using

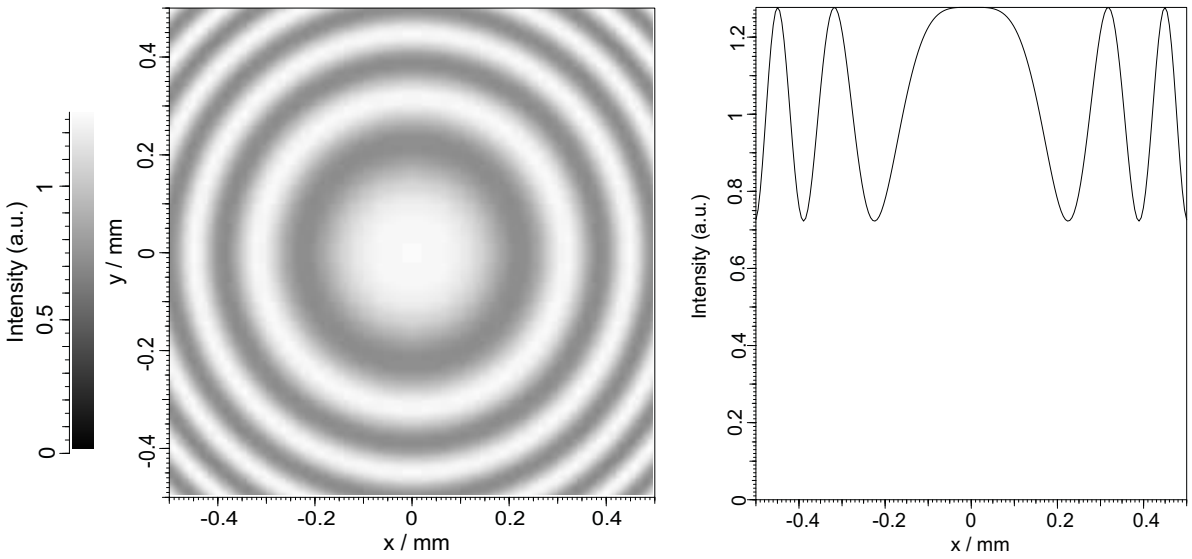


Figure 3.5: Example of an interference pattern showing a defocus term with a low contrast. Left: (simulated) camera picture as it can also be seen directly with the eye, right: section through the intensity function.

the visibility V:

$$V = \frac{2\sqrt{\alpha}}{1+\alpha} \;\Rightarrow\; \alpha = \frac{2}{V^2} - 1 \pm \sqrt{\left(\frac{2}{V^2} - 1\right)^2 - 1} = \frac{2}{V^2} - 1 \pm \frac{2}{V^2}\sqrt{1 - V^2} = \begin{cases} 49 \\ 0.02 \end{cases} \quad (3.3.26)$$

This means that the intensity of the second wave is either about 50 times higher than that of the first wave or about 50 times smaller. Which one of these two values is valid can be determined by measuring first the intensity I_1 of the first wave alone and then the intensity of the interference pattern and therefore also $I_1 + I_2$. The example also shows that the visibility decreases quite slowly if the intensity difference factor α increases. This is the reason why e.g. a scattered spherical wave of a dust particle which has quite a small intensity compared to the intensity of the illuminating coherent wave produces in many cases quite high–contrast fringes which disturb the measurement.
These two examples are of course quite simple and can be evaluated manually. However, already in the second example it can be seen that it is not so easy to decide whether it is really a pure defocus term or mixed with some other terms. Therefore, in practice an automated evaluation of the interference pattern has to be made [24]. One step to this is the phase shifting technique which will be discussed shortly on page 61.

3.4 Some basics of interferometry

The basic principle of an interferometer is that an incident wave is divided into two waves which can then interfere with each other. In most interferometers, like e.g. a **Michelson** or a **Mach–Zehnder interferometer**, there exist a so called reference arm and an object arm. The object arm contains often an object to be tested which changes the object wave. Together with the unchanged wave of the reference arm the interference pattern is formed and carries

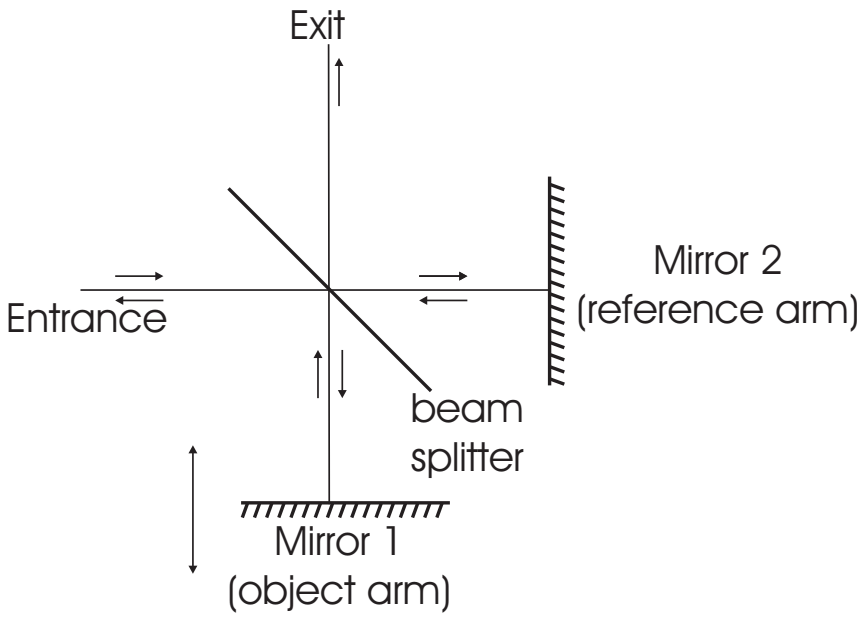


Figure 3.6: Basic principle of a Michelson interferometer.

information about the test object. Nevertheless, there are also interferometers, like e.g. **shearing interferometers**, which have not an object and a reference arm but two copies of an object wave which interfere.

The applications of interferometers are the measurement of surface deviations or aberrations of optical elements and the wavefront characterization. Another application is the high–precision length measurement. In the following the principles of the most important two–beam interferometers with monochromatic light are described. Other types of interferometers which cannot be treated here are interferometers with two or more wavelengths and so called white–light interferometers with a broad spectrum of wavelengths. Additionally, there are multiple beam interferometers where three or more light beams interfere, e.g. a Fabry–Perot interferometer which will be shortly discussed in section 3.5. For more information about interferometry we refer to [21],[22],[23],[24],[27],[28],[29],[30],[31],[32].

3.4.1 Michelson interferometer

One of the simplest interferometers is the Michelson interferometer (see fig. 3.6). A plane wave is divided by a beam splitter into two plane waves. One of these plane waves hits the reference mirror and one the object mirror. Both waves are then reflected back and each wave is again divided into two plane waves. Therefore, the Michelson interferometer has two exits whereby one is identical to the entrance so that only the other can really be used. If the beam splitter is exactly oriented by 45 degree relative to the incoming plane wave and both mirrors are exactly perpendicular to the plane waves the wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ of the two plane waves at the exit are parallel, i.e. $\boldsymbol{k}_1 = \boldsymbol{k}_2$. Then, according to equation (3.2.5) which is also valid for scalar waves the intensity in the whole space behind the exit, e.g. in the detector plane which shall be perpendicular to the z–axis, depends only on the phase difference δ which is constant over the

whole exit pupil.

$$I_{1+2} = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left((\boldsymbol{k}_1 - \boldsymbol{k}_2)\cdot\boldsymbol{r} + \delta\right) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta \tag{3.4.1}$$

In this case one speaks of **fluffed out fringes** because no interference fringes are present. Depending on the optical path difference δ between object and reference arm the intensity can have a maximum or a minimum. If the object mirror is axially shifted the intensity on the detector changes periodically and one period corresponds to an axial shift Δz of half a wavelength in the material with refractive index n, in which the light propagates (normally air), because of the double pass arrangement.

$$\Delta z = \frac{\lambda}{2n} \tag{3.4.2}$$

If the beam splitter or one of the mirrors are tilted there are interference fringes on the detector. An axial shift of the object mirror then causes a lateral movement of the fringes. Again, the fringes move by one period if the axial shift is $\lambda/(2n)$.
A typical application of a Michelson interferometer is the length measurement whereby the relative shift of the object mirror can be measured with an accuracy of a small fraction of the wavelength. In high–precision measurements like in interferometers for the measurement of gravitational waves an accuracy of less than 1/1000 wavelength is necessary. Additionally, the optical path can be folded several times by using additional optical elements so that the effect of an axial shift of the object mirror on the relative phase shift between object and reference wave is also multiplied. It has to be pointed out that an accuracy of less than 1/1000 wavelength using visible light is in the range of the diameter of an atom. Of course, such small shifts can only be interpreted as a shift of a large ensemble of atoms like in the case of a mirror because the light averages over many atoms.
There are several variations of the Michelson interferometer. One is e.g. a Twyman–Green interferometer where a lens is placed in the object arm and instead of using a plane object mirror a spherical mirror is used. In the case of the ideal adjustment of the interferometer the focus of the lens and the center of curvature of the spherical mirror have to coincide. If the quality of either the lens or the spherical mirror (and of all other components of the interferometer) is known, the resulting interference pattern can be used to determine the errors of the other component.

3.4.2 Mach–Zehnder interferometer

Another very important interferometer is a Mach–Zehnder interferometer (see fig. 3.7). There the light of an incoming plane wave is again divided by a beam splitter into two waves. Then the transmitted plane wave is reflected at the upper mirror and passes the second beam splitter or is reflected at it. The plane wave which is reflected at the first beam splitter is reflected at the lower mirror and can pass an optional transmissive object to be tested. At the second beam splitter this wave can be transmitted or reflected. So, the Mach–Zehnder interferometer has two exits which can both be used. The exit with equal number of reflections and passages of object and reference wave at beam splitters is called symmetric exit (exit 1 in fig. 3.7). Similar like in the case of the Michelson interferometer there are fluffed out fringes if all mirrors and beam splitters are oriented by exactly 45 degree relative to the incoming plane wave. If one of the mirrors or beam splitters is tilted there are fringes on the detector.

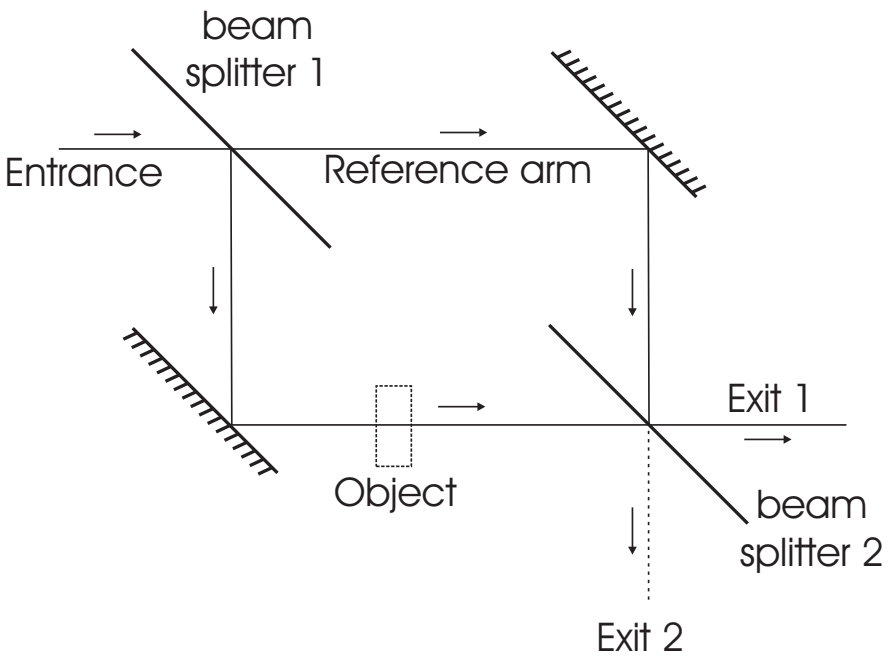


Figure 3.7: Basic principle of a Mach–Zehnder interferometer.

A Mach–Zehnder interferometer can be used to detect inhomogeneities of an optional object in the object arm. A special type is e.g. an interferometer where the object is a tube filled with gas. By changing the pressure or the temperature in the tube the refractive index n is changed. Then, also the optical path difference between object and reference arm changes and the fringes move or in the case of fluffed out fringes the overall intensity changes. So, the dependence of the refractive index of the gas on the pressure and temperature can be measured. In another type of Mach–Zehnder interferometer the object is a combination of a well–known lens and a lens to be tested which form together a telescope. Using phase shifting interferometry (see later) by shifting one of the mirrors the errors of the lens to be tested can be determined. Of course, in practice there are always so called adjustment errors which have to be eliminated from the measurement results and there are also systematic errors of the other components in the setup. Additionally, a very important fact is that the object to be tested has to be imaged onto the detector using auxiliary optics. Then, it is possible to say that the measured errors correspond to errors of the object at a certain point. This is of course also valid for the Twyman–Green interferometer and all other interferometers which are used for the measurement of an optical object with a limited depth. In an interferometer for the measurement of refractive index changes where the object is very long it is of course not possible to image the complete object sharply onto the detector.

3.4.3 Shearing interferometer

An interesting interferometer which needs no external reference arm is a shearing interferometer. There, by some means a copy of the object wave is generated which is either laterally or radially sheared [23],[33]. Here, only the case of lateral shearing will be discussed. The coordinate system is chosen in such a way that the shearing is along the x–axis by a distance Δx. Then, the phases of the two copies are $\Phi_1(x, y) = \Phi_o(x + \Delta x, y)$ and $\Phi_2(x, y) = \Phi_o(x, y)$, whereby Φ_o is the phase

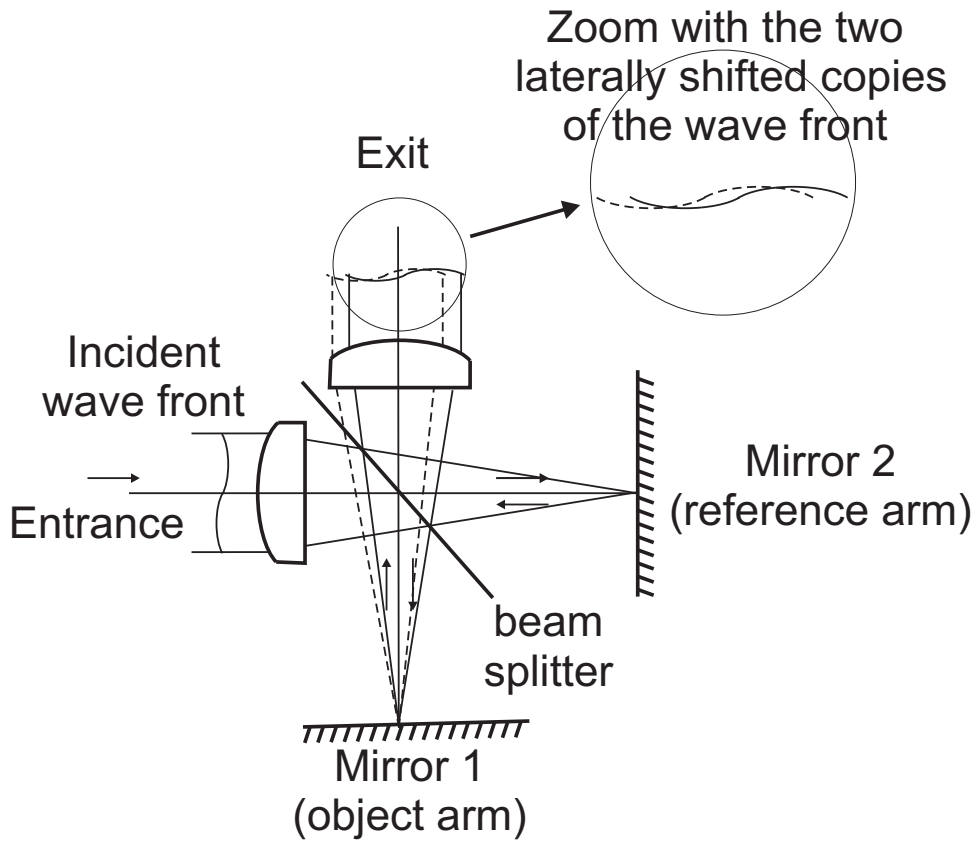


Figure 3.8: Basic principle of a shearing interferometer based on a Michelson type interferometer.

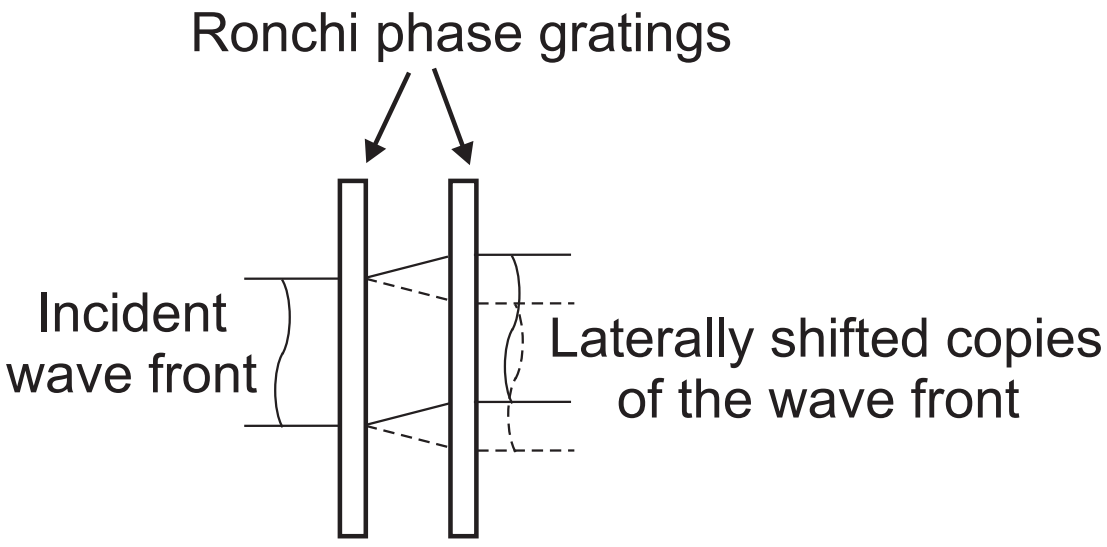


Figure 3.9: Basic principle of a shearing interferometer based on two Ronchi phase gratings.

of the object wave itself. The phase difference Φ which appears in the interference term of the interference equation (3.3.4) is then:

$$\Phi(x,y) = \Phi_1(x,y) - \Phi_2(x,y) = \Phi_o(x+\Delta x, y) - \Phi_o(x,y) \approx \Delta x \frac{\partial \Phi_o(x,y)}{\partial x} \tag{3.4.3}$$

The approximation of taking the first partial derivative at the point (x, y) is valid for small shearing distances Δx. The shearing interferometer can be similarly evaluated as other interferometers with phase shifting techniques and phase unwrapping whereby a continuous function for $\Phi \approx \Delta x \partial \Phi_o / \partial x$ results. To obtain the phase Φ_o of the wave front itself a kind of integration has to be made [23], [34]. To obtain an unambiguous wave front both partial derivatives of Φ_o in x– and y–direction have to be determined before the integration can be performed.
A lateral shear between object and reference wave can be generated by different methods:
One possibility [34] is to use a kind of Michelson imterferometer (see fig. 3.8), but with a lens in front of the beam splitter which focusses the plane wave onto the object and the reference mirror. By tilting one of these mirrors and using a second collimating lens behind the beam splitter a lateral shift between object and reference wave results. Since, in this case the object and the reference wave are identical copies of the incident wave, two laterally shifted (and inverted) copies of the incident wave result. The lateral shear distance is proportional to the tilt angle of the mirror. However, a disadvantage of this type of shearing interferometer is that the beam splitter is passed by a spherical wave. If the beam splitter is a common beam splitter cube the waves will therefore have some spherical aberration after having passed the Michelson type interferometer. Since the two copies are laterally shifted the spherical aberration will not cancel.
Another possibility [35] is the use of two identical Ronchi phase gratings (see fig. 3.9) with adapted etching depth, so that the diffraction efficiency in the first diffraction orders is about 40.5% each and no even diffraction orders exist. The first grating generates in the plus first and the minus first diffraction order two copies of the incident wave. The second grating directs the light back onto axis and two laterally shifted copies of the incident wave front result. Of course, there are also other diffraction orders, e.g. the light which is diffracted at both gratings in the same first diffraction order or light diffracted in higher orders. But, most of these waves will be off–axis and can be filtered out easily by using a telescope behind the shearing unit. Only those waves which are diffracted at the first grating into the order m and at the second grating into the order $-m$ are again on–axis, for example the waves with $m = \pm 3$ and a diffraction efficiency of 4.5% at each grating, which are the next efficient waves for such a Ronchi phase grating. So, these waves can disturb the interference pattern of the first order waves by multiple beam interference. A method to overcome this problem is an additional telescope between both gratings which allows also a filtering behind the first grating (only both first orders are allowed to pass) and which images the first grating onto a plane with a certain distance to the second grating. By changing the distance between both gratings (or between the image of the first grating and the second grating if there is an additional filtering telescope behind the first grating) the lateral shear distance can be changed continuously.

3.4.4 Fringe evaluation in interferometers

The fringes of an interferogram have to be evaluated in order to extract the information about the object. Typically the following procedures are necessary:

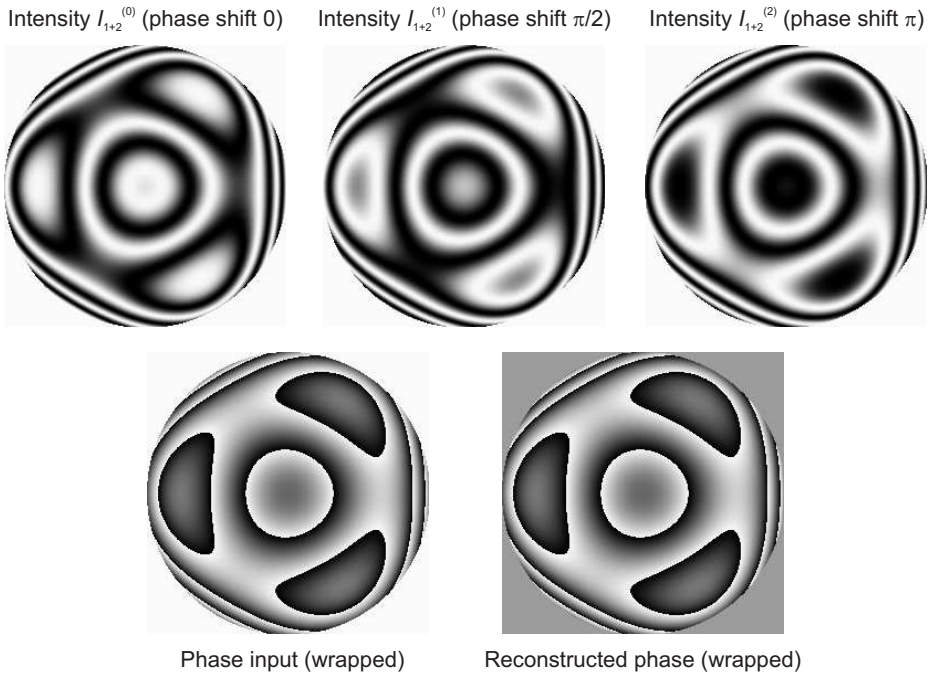


Figure 3.10: Reconstruction of phase data with phase shifting interferometry using ideal intensity data. Interferograms (top row from left to right) with phase shifts of 0, 0.5π and π. Wrapped input phase (bottom row left) and reconstructed phase (bottom row right).

- The so called raw phase or wrapped phase (phase modulo 2π) has to be calculated from the intensity distribution of one or more interferograms.
- The wrapped phase has to be unwrapped to obtain a continuous phase.
- Misalignment errors have to be subtracted (see page 52 for the case of testing a spherical object wave).
- In the case of shearing interferometry an integration of the obtained phase function has to be made since the measured phase function is only the partial derivative of the wanted phase.

The first two steps will be considered shortly in the following. However, we can only give a simple introduction since there are complete books and conferences about fringe evaluation.

3.4.4.1 Phase shifting interferometry

A typical technique to extract the object phase from the measured intensity values is the **phase shifting interferometry** [29],[36]. There, the reference mirror (or the object) is axially shifted by a well–known small distance and at least three different intensity distributions with different but well–known reference phases have to be observed, whereby there are also other possibilities to shift the phase [37]. By shifting the reference phase by $\delta\phi$ which is a well–defined integer multiple m of $\pi/2$ the intensity of the interference pattern $I_{1+2}^{(m)}$ changes according to equation

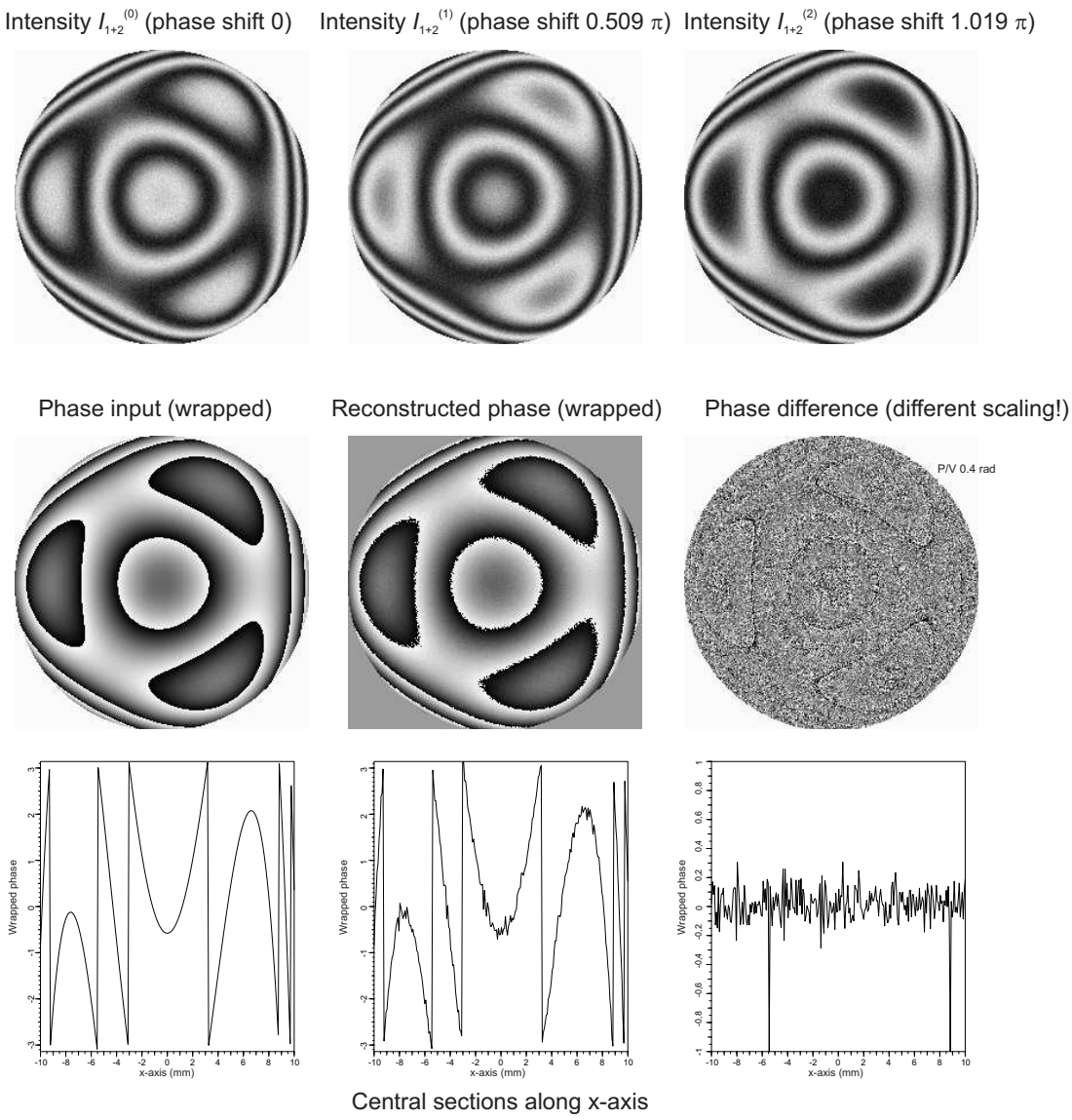


Figure 3.11: Reconstruction of phase data with phase shifting interferometry using noisy intensity data with a small linear phase shift error of 2% of the ideal value. Top row: interferograms (from left to right) with phase shifts of 0, 0.509π and 1.019π and a Gaussian noise with rms–value of 5% of the total intensity peak–to–valley value. Central row: Wrapped input phase (left), reconstructed phase (center) and difference between reconstructed and input phase (right). Bottom row: central sections of wrapped input phase (left), reconstructed phase (center) and difference between reconstructed and input phase (right). Please, pay attention to the different scale for the difference phase!

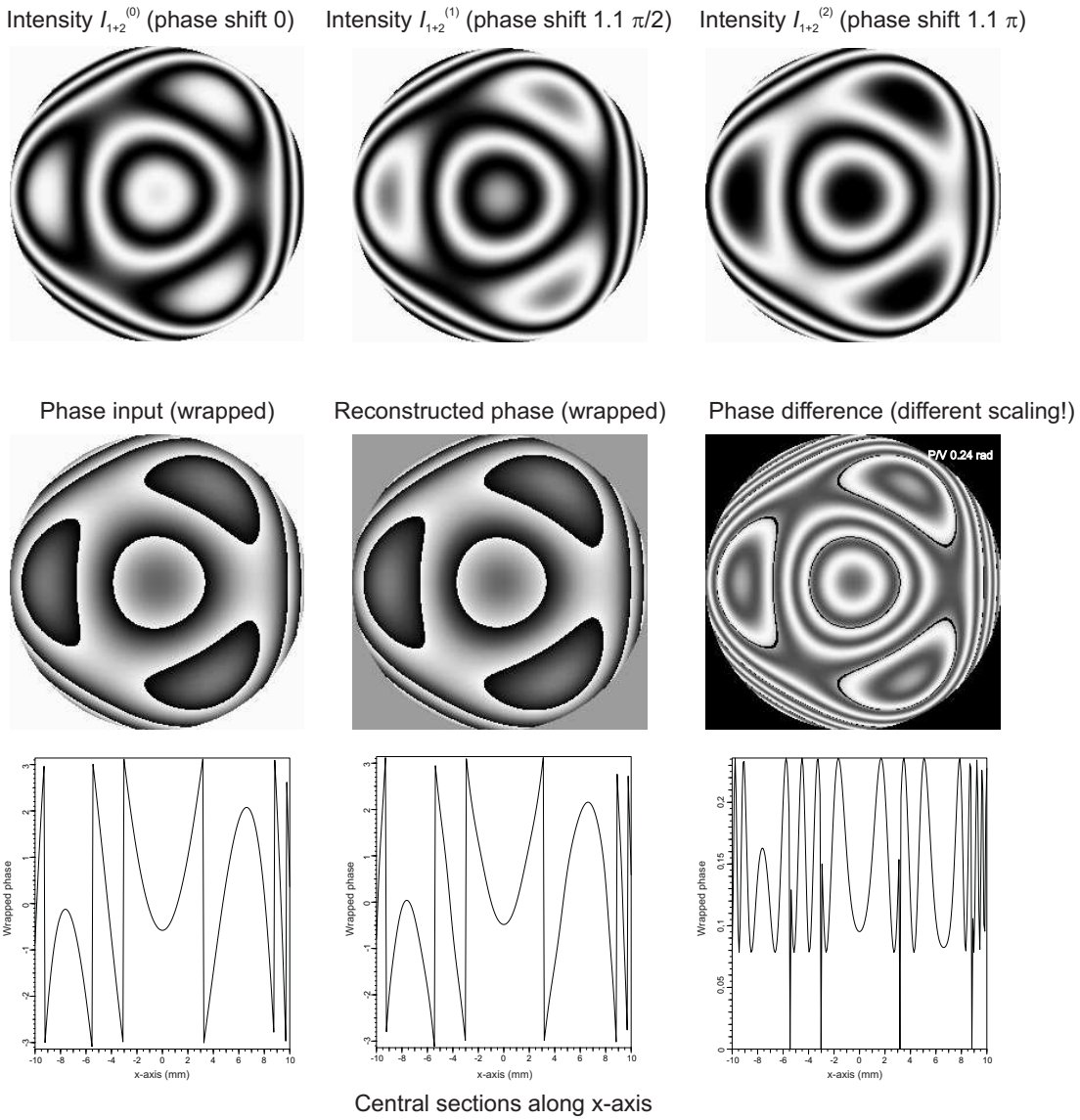


Figure 3.12: Reconstruction of phase data with phase shifting interferometry in the case of a strong linear phase shift error of 10% of the ideal value. Top row: interferograms (from left to right) with phase shifts of 0, $1.1\pi/2$ and 1.1π. Central row: Wrapped input phase (left), reconstructed phase (center) and difference between reconstructed and input phase (right). Bottom row: central sections of wrapped input phase (left), reconstructed phase (center) and difference between reconstructed and input phase (right). Please, pay attention to the different scale for the difference phase!

(3.3.4) as:

$$I_{1+2}^{(m)} = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\Phi + \delta\phi\right) = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\Phi + m\frac{\pi}{2}\right) \tag{3.4.4}$$

Three measurements with different reference phases are in principle enough because equation (3.4.4) contains the three unknowns I_1, I_2 and the desired phase Φ. It is quite simple to combine the three measured intensity distributions $I_{1+2}^{(0)}$, $I_{1+2}^{(1)}$ and $I_{1+2}^{(2)}$ with different values m in order to calculate Φ (fig. 3.10 shows examples of phase–shifted interferograms):

$$\begin{aligned}
I_{1+2}^{(0)} &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Phi \\
I_{1+2}^{(1)} &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\Phi + \frac{\pi}{2}\right) = I_1 + I_2 - 2\sqrt{I_1 I_2}\sin\Phi \\
I_{1+2}^{(2)} &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\Phi + \pi\right) = I_1 + I_2 - 2\sqrt{I_1 I_2}\cos\Phi
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad & I_{1+2}^{(0)} + I_{1+2}^{(2)} = 2\left(I_1 + I_2\right) \\
& I_{1+2}^{(0)} - I_{1+2}^{(2)} = 4\sqrt{I_1 I_2}\cos\Phi \\
& I_{1+2}^{(0)} + I_{1+2}^{(2)} - 2I_{1+2}^{(1)} = 4\sqrt{I_1 I_2}\sin\Phi \\
\Rightarrow \quad & \tan\Phi = \frac{I_{1+2}^{(0)} + I_{1+2}^{(2)} - 2I_{1+2}^{(1)}}{I_{1+2}^{(0)} - I_{1+2}^{(2)}}
\end{aligned} \tag{3.4.5}$$

However, this simple phase shifting algorithm with only three measurements is quite sensitive to phase shifting errors, i.e. if the phase shifts are not integer multiples of $\pi/2$. If more measurements are made the correction of phase shifting errors is possible [24],[30],[36]. Then, mostly a kind of least squares fit is used to calculate the three unknown variables (or mostly only the phase function Φ) of equation (3.4.4) from the larger number of measured intensity values.
Figures 3.10, 3.11 and 3.12 show the simulation of the reconstruction of phase data using the simple phase shifting algorithm of equation (3.4.5) with three intensity distributions. The simulated phase is a superposition of spherical aberration and trifoil error. In fig. 3.10 ideal data are taken as input, whereas in fig. 3.11 the phase shift has a slight linear error of 2% of the ideal value and a Gaussian noise with an rms–value (rms=root mean squares) of 5% of the total intensity P/V (P/V=peak to valley) has been added to the intensity data. Finally, figure 3.12 shows data with a strong linear phase shift error of 10% of the ideal value. For the ideal input data there is no difference between the input phase (bottom left) and the reconstructed phase (bottom right) if the correct phase offset is taken for the input data (which is that of the last interferogram with phase shift π). For the noisy data with a small linear phase shift error the reconstructed phase shows mainly some stochastic errors (see fig. 3.11), where the difference phase is displayed with a different scale to emphasize the errors. The evaluation of the data with the strong linear phase shift error shows on the first look nearly no difference between ideal and reconstructed data (see fig. 3.12 central and bottom row, left and central column). However, by building the difference between ideal and reconstructed phase (see fig. 3.12 central and bottom row, right column) there is a small phase error visible with the doubled frequency of the input phase. This is of course well–known in the literature [30] and measures to overcome this problem are the four–phase or multiple–phase algorithms for phase shifting interferometry [30, 36, 38].

3.4.4.2 Evaluation of carrier frequency interferograms

The disadvantage of phase shifting interferometry is that at least three interferograms with shifted reference phase are necessary. Therefore, these interferograms are normally made one after the other by shifting the reference mirror or some other means. Of course, the measurement of a rapidly varying phase distribution (e.g. an object which vibrates) is not possible with this method. As mentioned before, there is also the possibility to shift the phase temporally in parallel [37] by some special means. But, this also needs additional hardware and mostly a quite complex evaluation method.
In principle most of the information about the phase is contained in only one interferogram. But, the computer based automatic evaluation of only one fringe–contour map is not easy. Another more principle drawback is that the sign of the phase cannot be determined from one interferogram since the cosine function is an even function ($\cos(-\Phi) = \cos(\Phi)$). So, it is not clear whether the phase decreases or increases by looking at only one fringe–contour map.
However, there is another fringe evaluation technique developed by Takeda et al. [39], the so called Takeda algorithm, that needs only one single interferogram and which can also extract the correct sign of the phases. The precondition of using this technique is that there is besides the wanted phase Φ a spatial carrier frequency ν_0 in the interferogram so that there are no closed fringes (see for example fig. 3.14 top row). This can be achieved in practice by tilting for example the reference mirror in a Michelson or Mach–Zehnder type interferometer. Then, assuming a carrier frequency in x–direction the intensity distribution in the interferogram is of the form (see equation (3.3.5)):

$$\begin{aligned} I(x,y) &= \underbrace{I_0(x,y)}_{=:a(x,y)} + \underbrace{I_0(x,y)V(x,y)}_{=:b(x,y)} \cos\left[\Phi(x,y) + 2\pi\nu_0 x\right] = \\ &= a(x,y) + \underbrace{\frac{b(x,y)e^{i\Phi(x,y)}}{2}}_{=:c(x,y)} e^{2\pi i\nu_0 x} + \underbrace{\frac{b(x,y)e^{-i\Phi(x,y)}}{2}}_{=:c^*(x,y)} e^{-2\pi i\nu_0 x} \end{aligned} \tag{3.4.6}$$

Then, a Fourier transformation is applied to the intensity data field (in the computer a Fast Fourier transformation = FFT is used). Capital letters A and C are used to denote the Fourier transforms of a and c, where the Fourier transform G of a function g is defined by:

$$\mathrm{FT}\left\{g(x,y)\right\} = G(\nu_x,\nu_y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)e^{-2\pi i(\nu_x x + \nu_y y)}\mathrm{d}x\,\mathrm{d}y \tag{3.4.7}$$

Additionally, the Fourier transform shift theorem is used, i.e. multiplication by a phase factor of the form $\exp(2\pi i\nu_0 x)$ just means a lateral shift of the Fourier transformed function with variables ν_x and ν_y (spatial frequencies in x– and y–direction) by ν_0 in x–direction:

$$\mathrm{FT}\left\{I\right\}(\nu_x,\nu_y) = A(\nu_x,\nu_y) + C(\nu_x-\nu_0,\nu_y) + C^*(\nu_x+\nu_0,\nu_y) \tag{3.4.8}$$

Here, it has to be noted that C^* shall denote the Fourier transform of c^* and not, as usual, the complex conjugate of C! Otherwise, the signs of our arguments would not be correct.
If the spatial carrier frequency ν_0 is larger than each other spatial frequency component in the spectrum, especially $\nu_0 > |\partial\Phi/\partial x|/(2\pi)$, the three regions, where A, C and C^* are remarkably different from zero, are spatially separated (see for example left part of fig. 3.13). So, by

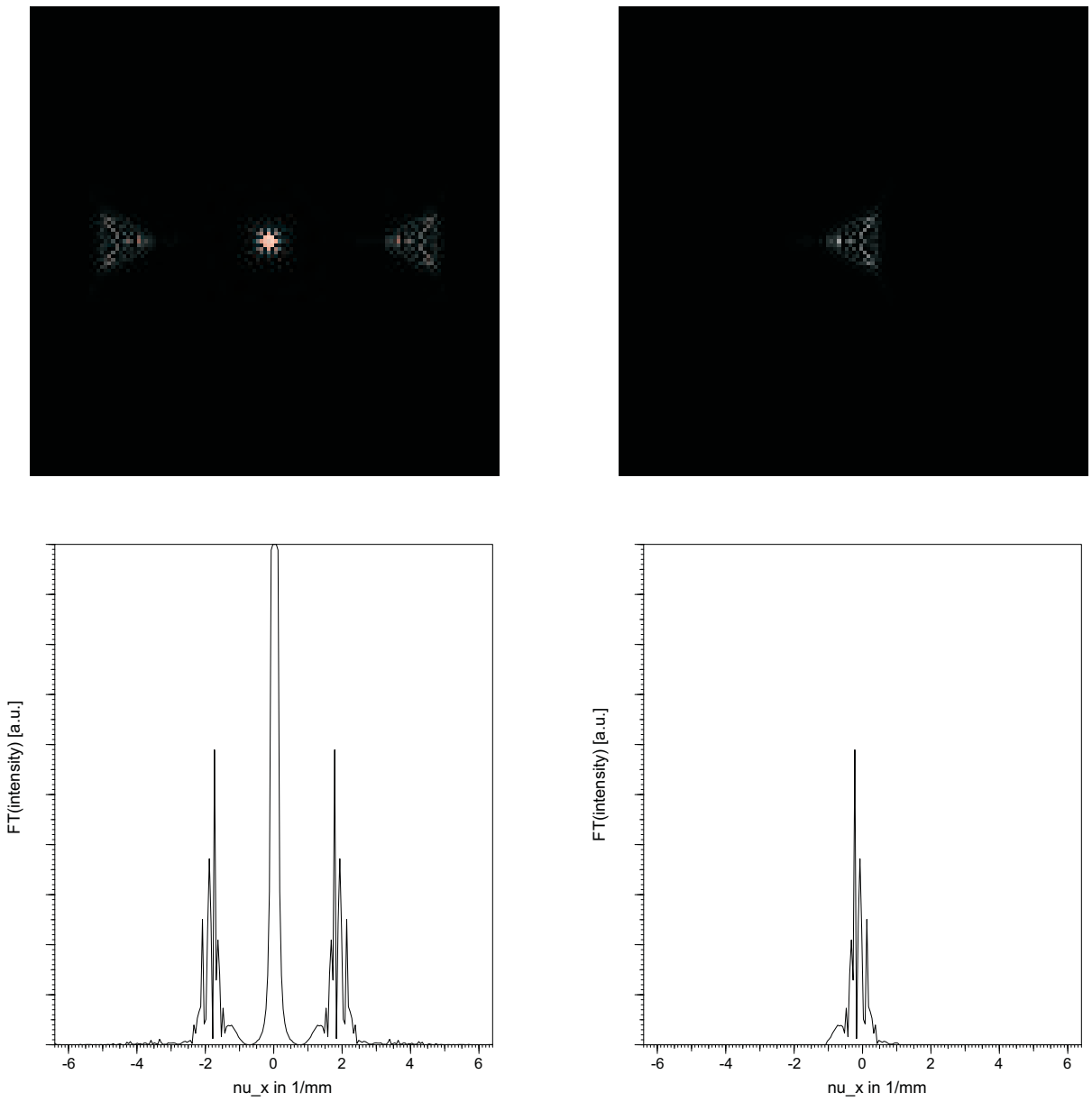


Figure 3.13: Squared modulus of the Fourier transform of the fringe pattern (Takeda algorithm). Left: before filtering and shifting (central peak clipped), right: after filtering and shifting. Top row: zoomed part, bottom row: section along the x–axis. Carrier frequency 2 lines/mm, diameter of interferogram 20 mm.

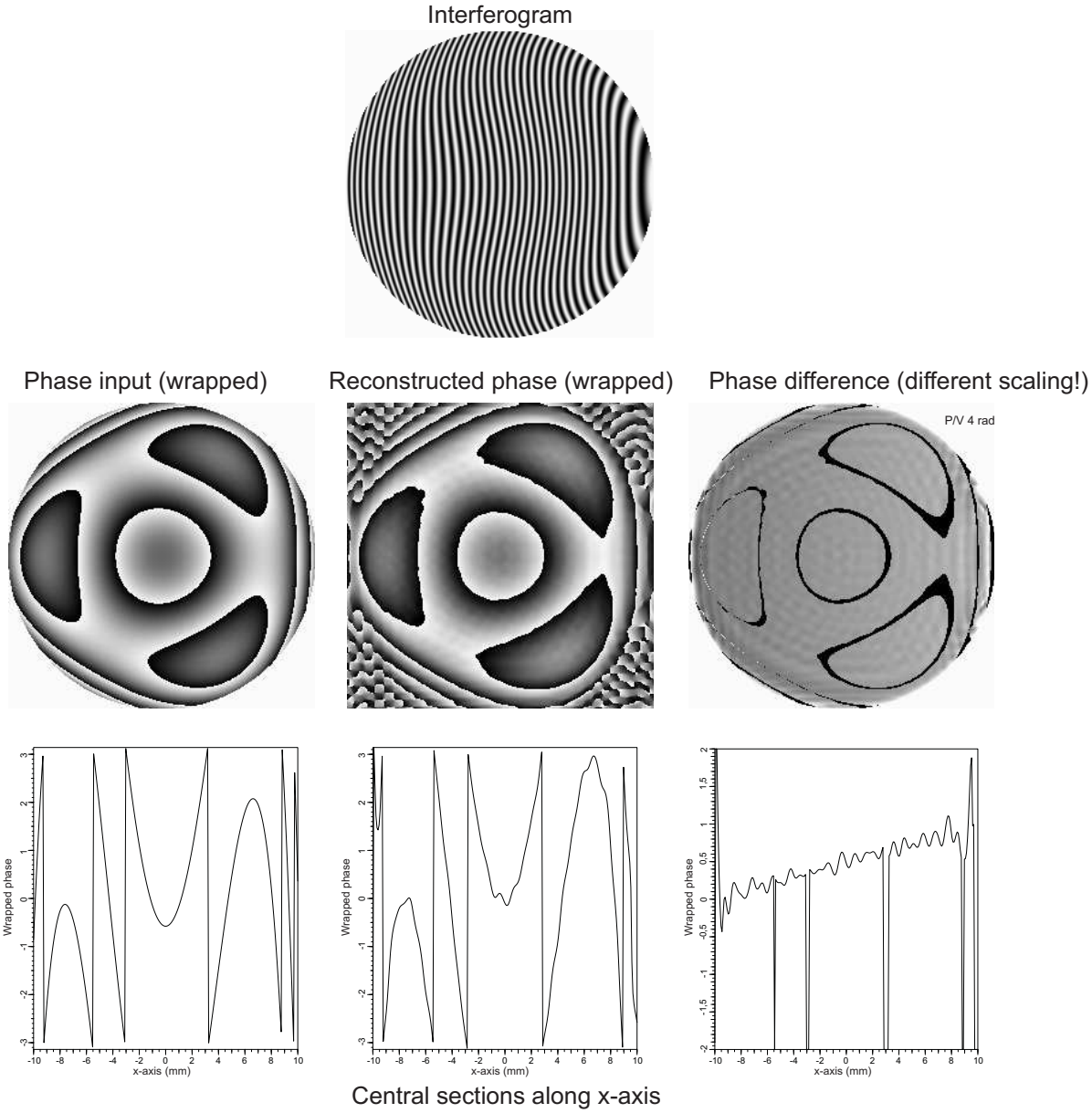


Figure 3.14: Interferogram (top row) and resulting wrapped phase of the Takeda algorithm (central row and bottom row) in comparison with the wrapped input phase. Left: input phase, center: reconstructed phase, right: difference between reconstructed and input phase (different scale!). Carrier frequency 2 lines/mm, diameter of interferogram 20 mm.

applying a mask to the field which sets all values to zero which are not in a small band around the coordinate ν_0 in x–direction, only the function $C(\nu_x - \nu_0, \nu_y)$ remains. Then, the field can be shifted towards the coordinate origin resulting in the non–shifted function $C(\nu_x, \nu_y)$ (see for example right part of fig. 3.13):

$$C(\nu_x, \nu_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} c(x, y) e^{-2\pi i(\nu_x x + \nu_y y)} \mathrm{d}x \, \mathrm{d}y \tag{3.4.9}$$

So, by using an inverse Fourier transformation the function c itself is obtained:

$$c(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} C(\nu_x, \nu_y) e^{+2\pi i(\nu_x x + \nu_y y)} \mathrm{d}\nu_x \, \mathrm{d}\nu_y \tag{3.4.10}$$

Since $c = b \exp(i\Phi)/2 = (b \cos \Phi)/2 + i(b \sin \Phi)/2$ and b is a real–valued function the phase Φ can be extracted via:

$$\tan \Phi(x, y) = \frac{\sin \Phi(x, y)}{\cos \Phi(x, y)} = \frac{\mathrm{Im}\{c(x, y)\}}{\mathrm{Re}\{c(x, y)\}} \tag{3.4.11}$$

As in the case of phase shifting interferometry the wrapped phase with values between $-\pi$ and $+\pi$ can then be calculated by taking into account the signs of real and imaginary part of c separately.

Simulations of the Takeda algorithm are displayed in figures 3.13–3.18. The input phase and the interferogram without carrier frequency are the same as in fig. 3.10, where the phase shift algorithm is simulated. The three Takeda algorithm simulations have a carrier frequency of 2 lines/mm without noise (see fig. 3.13 and 3.14), 2 lines/mm with intensity noise of 5% rms–value of the total intensity P/V (see fig. 3.15 and 3.16), and 4 lines/mm without noise (see fig. 3.17 and 3.18). The diameter of the interferograms is in all cases 20 mm, so that there are 40 fringes for the carrier frequency of 2 lines/mm and 80 fringes for the carrier frequency of 4 lines/mm. The number of pixels is in all cases 256x256. So, for the high carrier frequency of 4 lines/mm there are only about 3 pixels per fringe period.

The three interferograms of the input wave front with different carrier frequency and with or without intensity noise are displayed in the top rows of fig. 3.14, 3.16 and 3.18. Figures 3.13, 3.15 and 3.17 show the squared modulus of the Fourier transform of the intensity of the carrier frequency interferograms before and after filtering and shifting. Finally, the central and bottom rows of figures 3.14, 3.16 and 3.18 give a comparison of the wrapped phase resulting from the Takeda algorithm (central columns) with the wrapped input phase (left columns). The differences of the reconstructed and the input wrapped phase are displayed in the right columns, where different scales are used to display the phase differences more clearly. It can be seen that there are reconstructed phase values outside of the valid region. These have of course to be clipped since they contain no valid information and the intensity of the reconstructed wave, which is not shown here, is (nearly) zero in these regions.

Of course, the carrier frequency method has also its disadvantages as can be seen in the figures. If Φ varies locally quite fast, the carrier frequency has also to be large. However, for a given setup the maximum carrier frequency is given by the sampling theorem, i.e. the detector has to resolve more than two pixels per period. So, the reconstruction will be worse if the carrier frequency is at its limit as can be seen by comparing figures 3.14 and 3.18. Moreover, there are numerical limitations using discrete arrays of values. For a given field diameter D in the spatial domain the minimum spatial frequency distance $\Delta\nu$ in the Fourier domain is given by

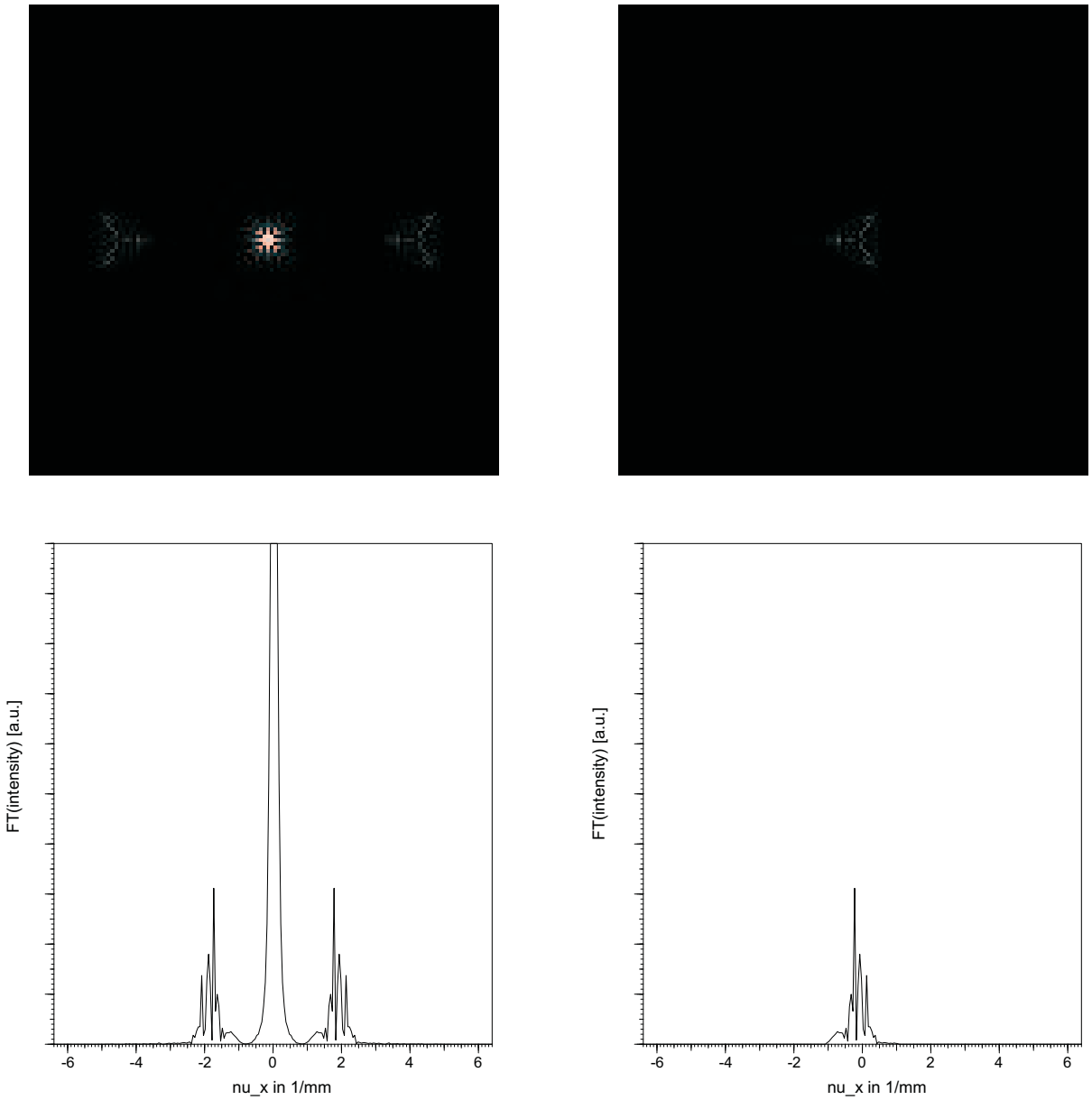


Figure 3.15: Squared modulus of the Fourier transform of the fringe pattern (Takeda algorithm). Left: before filtering and shifting (central peak clipped), right: after filtering and shifting. Top row: zoomed part, bottom row: section along the x–axis. Carrier frequency 2 lines/mm, diameter of interferogram 20 mm, noise added to the intensity data with an rms–value of 5% of the total intensity P/V.

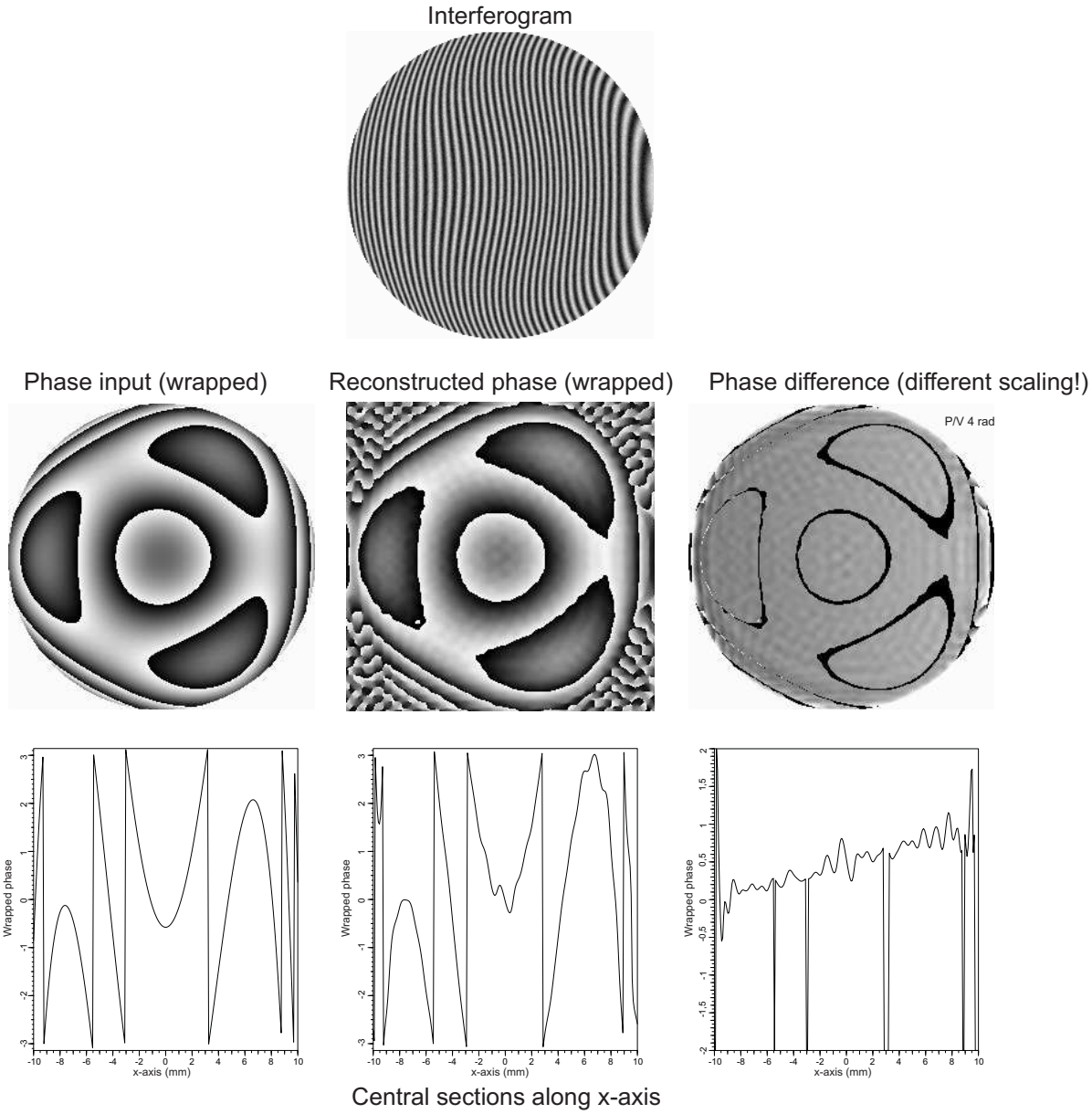


Figure 3.16: Interferogram (top row) and resulting wrapped phase of the Takeda algorithm (central row and bottom row) in comparison with the wrapped input phase. Left: input phase, center: reconstructed phase, right: difference between reconstructed and input phase (different scale!). Carrier frequency 2 lines/mm, diameter of interferogram 20 mm, noise added to the intensity data with an rms–value of 5% of the total intensity P/V.

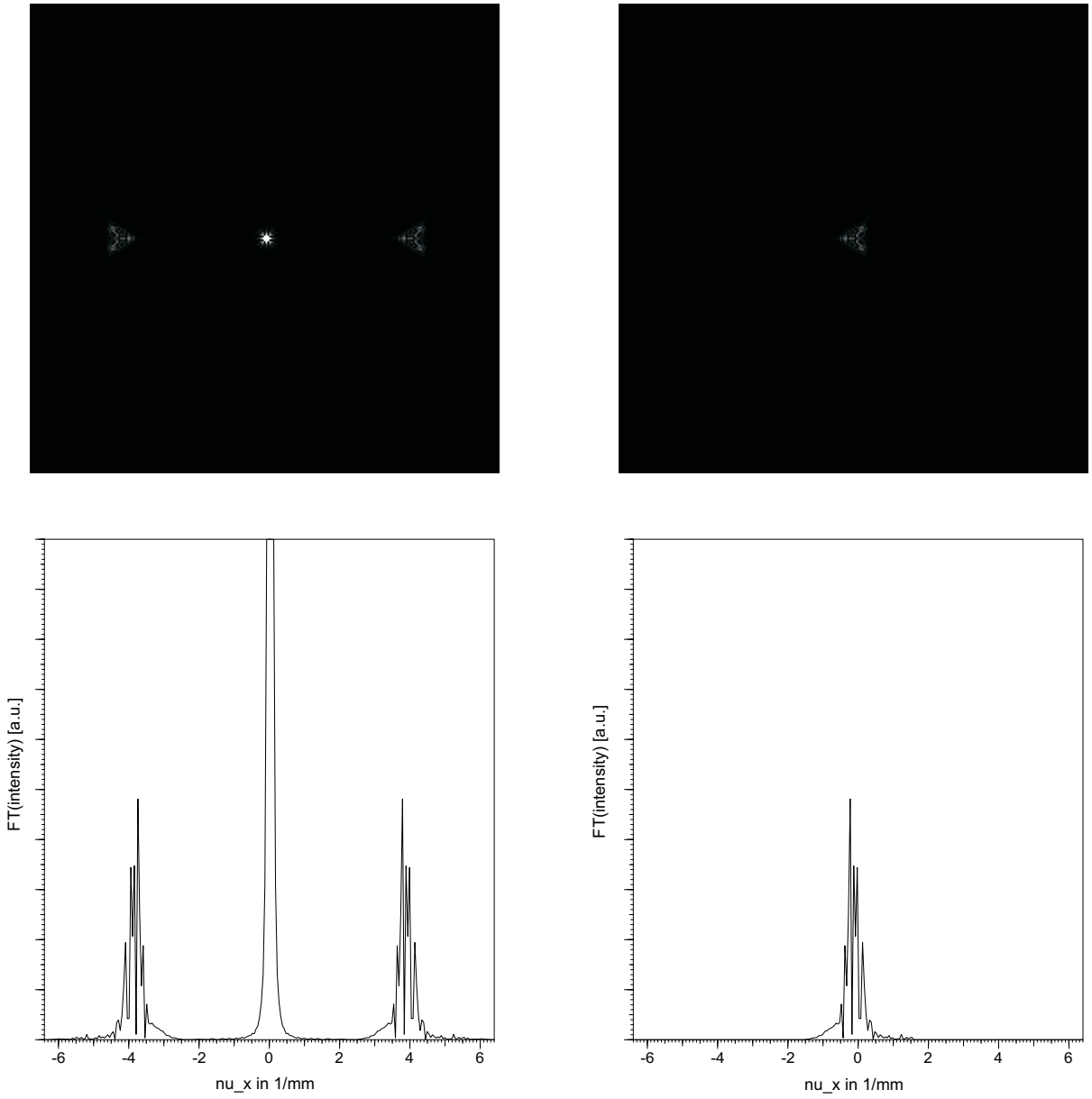


Figure 3.17: Squared modulus of the Fourier transform of the fringe pattern (Takeda algorithm). Left: before filtering and shifting (central peak clipped), right: after filtering and shifting. Top row: zoomed part, bottom row: section along the x–axis. Carrier frequency 4 lines/mm, diameter of interferogram 20 mm.

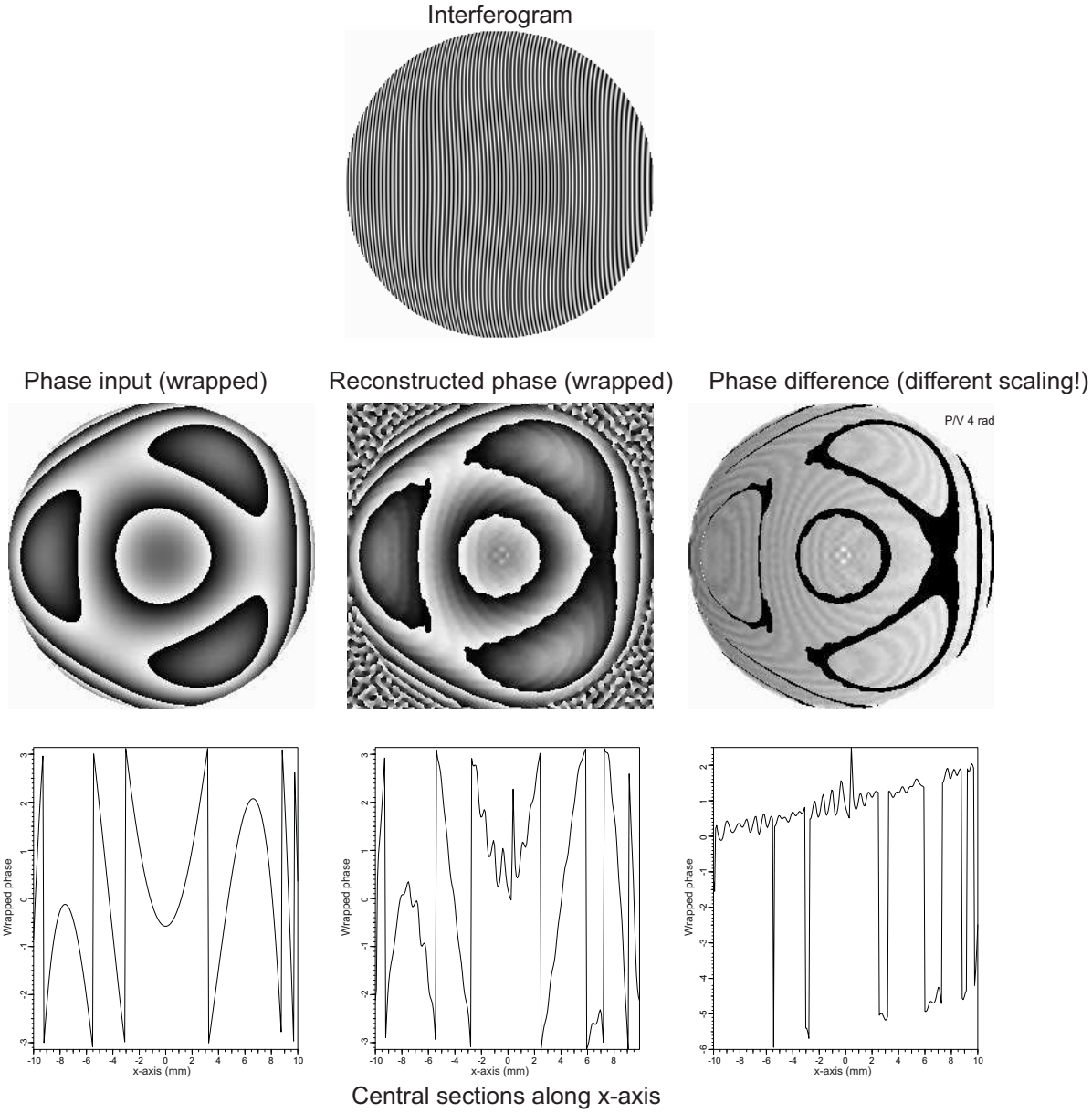


Figure 3.18: Interferogram (top row) and resulting wrapped phase of the Takeda algorithm (central row and bottom row) in comparison with the wrapped input phase. Left: input phase, center: reconstructed phase, right: difference between reconstructed and input phase (different scale!). Carrier frequency 4 lines/mm, diameter of interferogram 20 mm.

$\Delta\nu = 1/D$. So, if the carrier frequency ν_0 is not an integer multiple of $\Delta\nu$, the shifting of the Fourier transformed and filtered field towards the origin by a distance ν_0 cannot be done exactly. However, this means first of all that a small linear term remains in the phase values.

3.4.4.3 Comparison between fringe evaluation using phase shifting or carrier frequency interferograms

The main difference between the fringe evaluation using the phase shift method or a single carrier frequency interferogram is that in the first case the interferogram is evaluated at each pixel independently on other pixels. In contrary to this the carrier frequency interferogram contains the information encoded in the deformation of the fringes so that many pixels are necessary to get the information of the phase at one point. So, a variation of the sensitivity between different pixels of the detector are no problem in the case of the phase shift method, but it deteriorates the result of the Takeda algorithm since there many pixels contribute to one phase value.
On the other side, there is no additional hardware required in an interferometer to generate a carrier frequency interferogram (besides a possibility to tilt the reference mirror which is necessary in any case to adjust the interferometer). For the phase shift method a controlled piezo transducer or another possibility to shift the phase by hardware is required.
By comparing the results of the phase shift method and the Takeda algorithm, the phase shift method seems to be superior as long as the phase shift can be controlled quite good. Nevertheless, the additional hardware and the fact that at least three interferograms with shifted phase are needed for the phase shift method compared to only one interferogram for the carrier frequency interferogram method (very fast) have to be taken into account to select the most appropriate method for a certain application.

3.4.4.4 Phase unwrapping

A principal problem of two beam interferometry is that the raw phase values obtained with a phase shifting algorithm, like e.g. with equation (3.4.5), or with other fringe evaluation methods, like e.g. the Takeda algorithm, are first of all only defined modulo 2π.[1]
Therefore, so called **phase unwrapping algorithms** [24] have to be used in order to obtain a continuous phase profile of e.g. the surface deviations or wave aberrations of a lens. Of course, there exist also cases where a phase unwrapping is not possible, e.g. if diffraction effects can be seen in the interferogram.
The precondition for all phase unwrapping algorithms is that the modulus of the phase difference between two neighbored pixels is less than π for the continuous real phase (fulfillment of sampling theorem). Phase steps of more than π between neighbored pixels violate the sampling theorem and therefore such phase steps are not allowed in single wavelength interferometry in order to have a unique measurement of the object wave. So, if there are phase differences between neighbored pixels with a modulus of more than π in the wrapped raw data and the sampling theorem is fulfilled, this can only be since the phase is measured modulo 2π. From the sign of

[1] The arc tangent function itself is only unambiguously defined between $-\pi/2$ and $+\pi/2$. But, it is $\tan\Phi = \sin\Phi/\cos\Phi = \mathrm{Im}(\exp(i\Phi))/\mathrm{Re}(\exp(i\Phi))$. So, by evaluating the signs of the numerator and denominator of equation (3.4.5) separately, the phase Φ can be unambiguously calculated between $-\pi$ and $+\pi$. In many programming languages there exists for this purpose a special function, like e.g. the atan2 function of the language C.

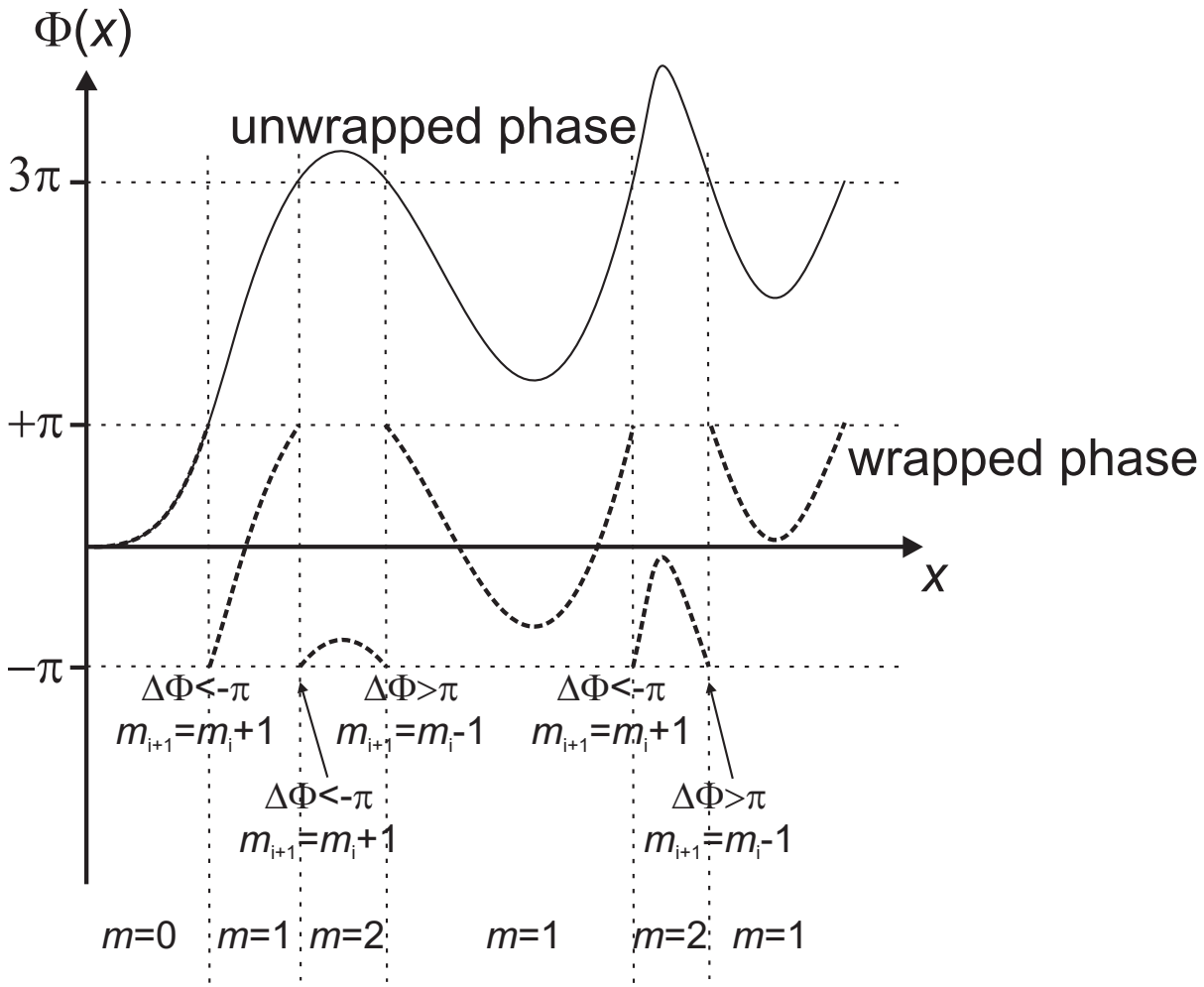


Figure 3.19: Basic principle of phase unwrapping along a line of the raw phase array. The dashed curves represent the measured wrapped raw phase and the continuous curve represents the wanted unwrapped phase.

the phase difference it can be seen whether the phase of the neighbored pixel has to be increased or decreased by 2π (or an integer multiple of 2π since the phase offset of the starting pixel has to be taken into account) in order to obtain the unwrapped phase. So, in principle the phase can be unwrapped pixel by pixel starting at a certain point. Figure 3.19 shows the basic principle of unwrapping along a line of the normally two–dimensional raw phase array. For each pixel number i we define an integer number m_i, which defines the phase offset $2\pi m_i$ which has to be added at the end. The phase of the pixel at the left rim is taken as reference for all other pixels and therefore the integer m_1 of this first pixel is by definition $m_1 = 0$. Then, for each pixel number i and its neighbor $i+1$ the phase difference $\Delta\Phi = \Phi_{i+1} - \Phi_i$ is calculated. If the modulus of $\Delta\Phi$ is larger than π we know that the number m_{i+1} of the neighbored pixel has to be increased or decreased by one:

$$\begin{array}{lll} m_{i+1} = m_i + 1 & \text{if} & \Phi_{i+1} - \Phi_i < -\pi \\ m_{i+1} = m_i - 1 & \text{if} & \Phi_{i+1} - \Phi_i > +\pi \\ m_{i+1} = m_i & \text{if} & |\Phi_{i+1} - \Phi_i| < \pi \end{array} \tag{3.4.12}$$

If all pixels i of the array have obtained in this way an integer number m_i, the final unwrapped phase is calculated via:

$$\Phi_{unwrapped,i} = \Phi_i + 2\pi m_i \tag{3.4.13}$$

However, in practice there will be pixels without a sufficient modulation in the field or other errors. By using phase shifting interferometry to determine the phase, the visibility of a pixel can be used as criterion to detect defected pixels. The visibility can be calculated from the intensities of the three interferograms with shifted reference phase and the calculated phase itself (see equation (3.4.5)) via:

$$V = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} = \frac{I_{1+2}^{(0)} - I_{1+2}^{(2)}}{\left(I_{1+2}^{(0)} + I_{1+2}^{(2)}\right)\cos\Phi} \tag{3.4.14}$$

If the visibility is below a certain threshold (which depends on the application and wanted accuracy) the pixel has to be marked as defect.
If we unwrap the phase of a two–dimensional array only line by line one defect in a line would divide the line into two regions which could not be correlated to each other. However, in a two–dimensional array there is the possibility to bypass defects as long as the defects do not build complete lines which separate the array into two discontiguous regions. Therefore, the phase unwrapping algorithms which are used in practice have to find their path in a two–dimensional plane via bypassing defect pixels. These algorithms can be quite complex and beyond our scope of the basic ideas of interferometry. In two–dimensional phase arrays it is also possible to check the unwrapping process since the continuation along an arbitrary way has to give the same phase value at a certain pixel.

3.4.5 Some ideas to the energy conservation in interferometers

Here, some principal ideas to the conservation of energy in an interferometer shall be given because the laws of energy conservation have to be fulfilled everywhere in optics.
Let us consider e.g. the Mach–Zehnder interferometer of fig. 3.7. We assume that each beam splitter has a splitting ratio of 1:1, i.e. half of the light power is transmitted and half is reflected.

So, if the intensity of the incoming plane wave is I_0 at the entrance, the intensities of the transmitted and reflected plane waves are each $I_0/2$. The two mirrors are assumed to reflect all light without losses and diffraction effects at apertures are also neglected because we handle with plane waves. So, at the second beam splitter each of the plane waves is again divided into two waves with equal intensity, i.e. each of the four waves has now the intensity $I_0/4$. At the exit 1 two of these waves interfere and we assume that the phase difference Φ between these two waves shall be zero or an integer number times 2π. Then, the resulting intensity I_{1+2} is according to equation (3.3.4):

$$I_{1+2} = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Phi = \frac{I_0}{4} + \frac{I_0}{4} + 2\sqrt{\frac{I_0^2}{16}} = I_0 \tag{3.4.15}$$

But this means that all of the incoming light power has to be at exit 1. Therefore, the intensity I'_{1+2} at exit 2 has automatically to be zero. Since the intensities of the two single waves which interfere at exit 2 are also $I_0/4$ this is only possible if the phase difference Φ' is an odd–numbered multiple of π:

$$I'_{1+2} = \frac{I_0}{4} + \frac{I_0}{4} + 2\sqrt{\frac{I_0^2}{16}}\cos\Phi' = 0 \;\Rightarrow\; \Phi' = (2m+1)\,\pi \tag{3.4.16}$$

with an integer number m. So, if we take the basic solution $\Phi' = \pi$ this requires that the phase shift between the transmitted and the reflected wave at each beam splitter is half of this value, i.e. $\pi/2$, if the function can be split between the two beam splitters. Then, the law of energy conservation is fulfilled. To explain this a little bit more fig. 3.7 is regarded.
The phase differences between the two waves interfering at exits 1 and 2 respectively due to the geometrical paths are identical for both exits. The same is valid for reflections at the mirrors because each wave is exactly reflected once at a mirror. So, there has to be a phase shift between a reflected and a transmitted wave at a beam splitter in order to fulfill the law of energy conservation. The two waves which interfere at exit 1 (symmetrical exit) are each reflected one time at a beam splitter and have transmitted one time a beam splitter. So, the assumed phase shifts of $\pi/2$ due to a reflection at a beam splitter cancel out each other because the phase difference between both interfering waves is taken. But, at the exit 2 (antisymmetric exit) the first wave is transmitted by both beam splitters and the other wave is reflected at both beam splitters. Therefore, the phase difference between the two interfering waves is in this case:

$$\Phi' = \Phi + 2\frac{\pi}{2} = \Phi + \pi \tag{3.4.17}$$

This guarantees that the sum of the intensities $I_{1+2} + I'_{1+2}$ at both exits is equal to the intensity at the entrance:

$$\begin{aligned} I_{1+2} + I'_{1+2} &= \left(\frac{I_0}{4} + \frac{I_0}{4} + 2\sqrt{\frac{I_0^2}{16}}\cos\Phi\right) + \left(\frac{I_0}{4} + \frac{I_0}{4} + 2\sqrt{\frac{I_0^2}{16}}\cos\Phi'\right) = \\ &= \frac{I_0}{2}\,(1+\cos\Phi) + \frac{I_0}{2}\,(1-\cos\Phi) = I_0 \end{aligned} \tag{3.4.18}$$

A more detailed analysis of the behavior of the beam splitters in our interferometer shows that the assumption of a phase shift between the reflected and transmitted wave at each beam splitter of $\pi/2$ is not really correct, but we have to regard the construction of such a beam splitter. Then,

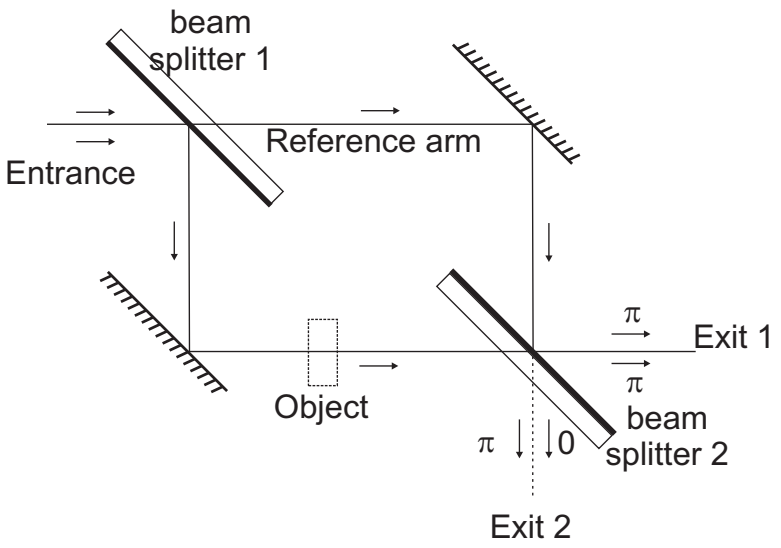


Figure 3.20: Illustration of the energy conservation in a Mach–Zehnder interferometer by showing the phase shifts of each ray at both exits which are purely due to the reflections at the beam splitters because the optical ray paths are identical for both rays of each exit, respectively. The thick line at a beam splitter indicates the silvered side where the reflection takes place. The tilt of the ray in the glass plate due to refraction is neglected in the drawing. The arrows indicate the different paths of the rays. The ray at exit 2 which is reflected at both beam splitters has e.g. a phase shift of only π because there is only at the first beam splitter a phase shift due to the reflection at an interface from an optically less dense material to a denser material. There is no phase shift at the second beam splitter where the reflection is at an interface from an optically denser material to a less dense material. The second ray at exit 2 which only transmits both beam splitters has no phase shift. Both rays at exit 1 have a phase shift of π because both are reflected once at a beam splitter with an interface from an optically less dense material to a denser material. So, totally there is a phase shift difference of π between both exits.

we see that it is important at which side of the beam splitter the wave is reflected because later we will see that there is a phase shift of π if the light is reflected at an interface from an optically less dense material to an optically denser material, whereas there is no phase shift for reflection at an interface from an optically denser material to an optically less dense material. A beam splitter can e.g. be constructed by taking a glass plate which is silvered at one side. Then, we have such a phase difference of π between the two exits which is illustrated in fig. 3.20.

But, in the end, independent of the real construction of the beam splitter, the important thing is that there is always a phase shift of π between the two exits of the interferometer which guarantees the energy conservation.

3.5 Multiple beam interference

Up to now, we discussed only two beam interferences. But, there are also important interferometers which are based on multiple beam interferences, like e.g. the Fabry–Perot interferometer for measuring wavelength differences. In this section, we will therefore discuss a little bit the principles of multiple beam interference and some applications. For more information look e.g. at [1].

3.5.1 Optical path difference at a plane–parallel glass plate

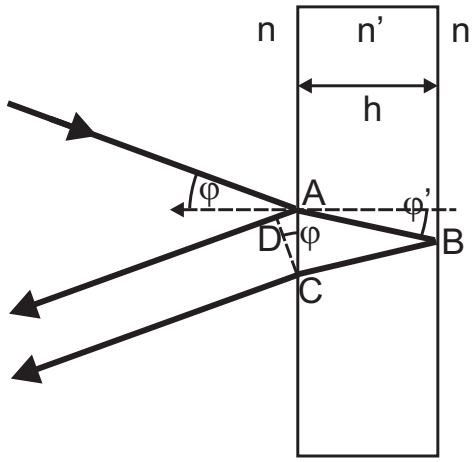


Figure 3.21: Scheme for calculating the optical path difference at a plane–parallel glass plate for an incident plane wave.

In most cases, a plane–parallel glass plate which is illuminated by a plane wave is used for the creation of controlled multiple beam interferences. As preliminary practice we have therefore to calculate the optical path difference between a ray which is directly reflected at the first surface of the glass plate and a ray which is reflected at the back side of the glass plate (see fig. 3.21) with thickness h. The ray represents of course a plane wave which is extended over the whole aperture. The vacuum wavelength of this wave is λ, the refractive index of the glass plate is n' and the refractive index of the surrounding medium (which is normally air) is n. Then, the optical path difference for a wave impinging with the angle φ is according to fig. 3.21:

$$\begin{aligned} \text{OPD} &= n'[AB] + n'[BC] - n[AD] = 2n'\frac{h}{\cos\varphi'} - n\sin\varphi \cdot 2h\tan\varphi' = \\ &= 2n'h\left(\frac{1}{\cos\varphi'} - \frac{\sin^2\varphi'}{\cos\varphi'}\right) = 2n'h\cos\varphi' \end{aligned} \tag{3.5.1}$$

Here, Snellius law $n\sin\varphi = n'\sin\varphi'$ has been used, where φ' is the angle between the refracted ray in the glass plate and the surface normal. So, the phase difference due to the propagation in the glass plate is:

$$\phi = \frac{2\pi}{\lambda}\text{OPD} = \frac{4\pi}{\lambda}n'h\cos\varphi' \tag{3.5.2}$$

Here, a possible phase shift of π due to reflection at an interface between an optically less dense material and a denser material is not taken into account, but ϕ is the phase which is purely due to the different path lengths.

3.5.2 Calculation of the intensity distribution of the multiple beam interference pattern

Now, the intensity of the multiple beam interference pattern in the far field of a plane–parallel glass plate will be calculated (see fig. 3.22). There, we have reflected beams and transmitted

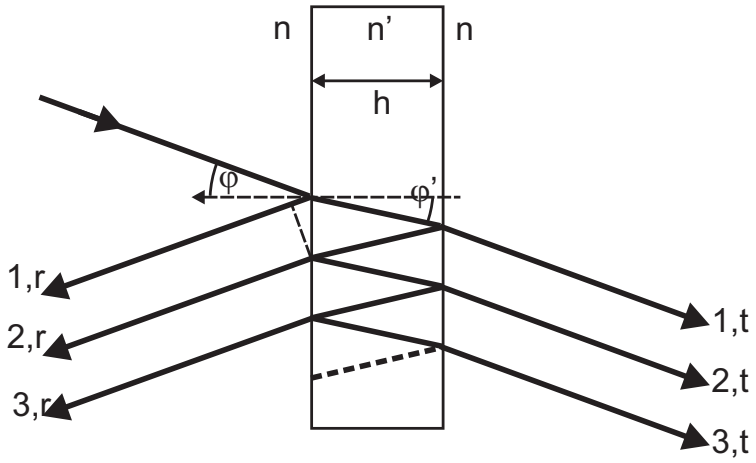


Figure 3.22: Scheme for calculating the intensities of the multiple beam interference patterns at a plane–parallel glass plate for an incident plane wave. The numbers mark the different reflected and transmitted waves due to multiple reflections.

beams. The amplitudes of the reflected beams all interfere in the far field, and the same is valid for the transmitted beams. The amplitude of the incident plane wave with wavelength λ and angle of incidence φ is A_0. Then, there is a directly reflected wave which is indicated in fig. 3.22 with the number 1,r. The amplitude coefficient (complex value) for reflection at the interface from the material with refractive index n to that with refractive index n' shall be r. So, the amplitude of this directly reflected wave is:

$$A_{1,r} = A_0 r \tag{3.5.3}$$

The wave marked with number 2,r first transmits the interface from the material with refractive index n to that with refractive index n'. The amplitude transmittance coefficient for this is t. Then, the wave is reflected at the back side. There, we have an amplitude reflection coefficient r'. After reflection, the wave number 2,r transmits the interface from the material with refractive index n' to the material with refractive index n. The amplitude transmittance coefficient for this is t'. Due to the phase difference ϕ (equation (3.5.2)) because of the optical path difference between wave 1,r and 2,r the amplitude of wave 2,r is in total:

$$A_{2,r} = A_0 t t' r' e^{i\phi} \tag{3.5.4}$$

It is easy to see from fig. 3.22 that the amplitudes of the other reflected waves can be calculated by a similar scheme:

$$\begin{aligned} A_{3,r} &= A_0 t t' r'^3 e^{2i\phi} \\ A_{4,r} &= A_0 t t' r'^5 e^{3i\phi} \\ &\dots \end{aligned} \tag{3.5.5}$$

Since we did not treat the Fresnel formulae for reflection and transmission at an interface up to now, we will not explicitly express t and t' or r and r'. The concrete values may also depend on dielectric layers on the surface to increase for example the reflectivity. But, for us it is interesting

that the transmissivity T (i.e. the ratio between the transmitted light power and the incident light power) of each surface is

$$T = tt' \tag{3.5.6}$$

Additionally, there is the relation

$$r' = -r \tag{3.5.7}$$

due to the phase difference of π between reflection at an interface from an optically less dense material to a denser material and reflection at an interface from an optically denser material to a less dense material. Here, we do not explicitly define which material is optically denser (normally the glass plate will be in air so that $n' > n$) because for us it is only interesting that there is this relative phase difference of π between the reflection coefficients r and r'. Moreover, there is the reflectivity R (i.e. the ratio between the reflected light power and the incident light power) of each surface

$$R = r^2 = r'^2 \tag{3.5.8}$$

and the relation

$$R + T = 1 \ , \tag{3.5.9}$$

since no light is absorbed in our glass plate.
The amplitudes of all reflected waves with numbers 1,r; 2,r; 3,r and so on have to be added to achieve the amplitude of the interference pattern in the far field. So, we have for the total reflected amplitude A_r:

$$\begin{aligned} A_r &= A_{1,r} + A_{2,r} + A_{3,r} + A_{4,r} + \ldots = A_0 r + A_0 tt' r' e^{i\phi} + A_0 tt' r'^3 e^{2i\phi} + A_0 tt' r'^5 e^{3i\phi} + \ldots = \\ &= A_0 r \left[1 - T e^{i\phi} \left(1 + R e^{i\phi} + R^2 e^{2i\phi} + \ldots\right)\right] \end{aligned} \tag{3.5.10}$$

The term in the parenthesis is an infinite geometric series $1 + x + x^2 + \ldots$ with $x = R\exp(i\phi)$ and $|x| < 1$. So, it converges to the value $1/(1-x)$ and the result using the relations above is:

$$A_r = A_0 r \left(1 - \frac{T e^{i\phi}}{1 - R e^{i\phi}}\right) = A_0 r \frac{1 - e^{i\phi}}{1 - R e^{i\phi}} \tag{3.5.11}$$

The total intensity of the reflected waves is:

$$I_r = A_r A_r^* = I_0 R \frac{2\,(1 - \cos\phi)}{1 + R^2 - 2R\cos\phi} = I_0 \frac{4R\sin^2\left(\frac{\phi}{2}\right)}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} \tag{3.5.12}$$

where $I_0 = A_0 A_0^*$ is the intensity of the incident light and the trigonometric relation $1 - \cos\phi = 2\sin^2(\phi/2)$ is used.
Similar, the amplitudes of the transmitted beams can be calculated (see fig. 3.22). There, the phase differences due to optical path differences between the waves number 1,t; 2,t and so on are again expressed by multiples of ϕ. So, we have:

$$\begin{aligned} A_{1,t} &= A_0 tt' \\ A_{2,t} &= A_0 tt' r'^2 e^{i\phi} \\ A_{3,t} &= A_0 tt' r'^4 e^{2i\phi} \\ &\ldots \end{aligned} \tag{3.5.13}$$

Then, the total transmitted amplitude A_t is again the sum of all transmitted amplitudes:

$$\begin{aligned} A_t &= A_{1,t} + A_{2,t} + A_{3,t} + \ldots = A_0 tt' + A_0 tt' r'^2 e^{i\phi} + A_0 tt' r'^4 e^{2i\phi} + \ldots = \\ &= A_0 tt' \left(1 + Re^{i\phi} + R^2 e^{2i\phi} + \ldots\right) = A_0 \frac{T}{1 - Re^{i\phi}} \end{aligned} \tag{3.5.14}$$

The total intensity of the transmitted light is therefore:

$$I_t = A_t A_t^* = I_0 \frac{T^2}{1 + R^2 - 2R\cos\phi} = I_0 \frac{(1-R)^2}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} \tag{3.5.15}$$

It can be seen that the sum of the intensities of the reflected light (3.5.12) and the transmitted light (3.5.15) is equal to the intensity of the incident light:

$$I_r + I_t = I_0 \frac{4R\sin^2\left(\frac{\phi}{2}\right)}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} + I_0 \frac{(1-R)^2}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} = I_0 \tag{3.5.16}$$

This just means that no light is absorbed in the (ideal) glass plate.

3.5.3 Discussion of the intensity distribution

Now, equations (3.5.12) and (3.5.15) have to be discussed in dependence on the phase difference ϕ defined by equation (3.5.2). For a given wavelength λ, ϕ changes for different angles of incidence φ (and therefore also different angles φ'). So, there are the interesting cases where

$$\frac{\phi}{2} = \frac{2\pi}{\lambda} n' h \cos\varphi' = m\pi \quad \Rightarrow \quad \cos\varphi' = m\frac{\lambda}{2n'h} \tag{3.5.17}$$

or

$$\frac{\phi}{2} = \frac{2\pi}{\lambda} n' h \cos\varphi' = (2m'+1)\frac{\pi}{2} \quad \Rightarrow \quad \cos\varphi' = (2m'+1)\frac{\lambda}{4n'h} \tag{3.5.18}$$

with the integers $m = 1, 2, \ldots$ and $m' = 0, 1, 2, \ldots$.
If we regard the reflected light the ratio of the reflected light and the incident light will have in the first case a minimum and in the second case a maximum:

$$\frac{\phi}{2} = m\pi \quad \Rightarrow \quad \frac{I_r}{I_0} = 0 \tag{3.5.19}$$

$$\frac{\phi}{2} = (2m'+1)\frac{\pi}{2} \quad \Rightarrow \quad \frac{I_r}{I_0} = \frac{4R}{(1+R)^2} \tag{3.5.20}$$

So, for small values of the reflectivity R the maxima will also be quite small. But, if the reflectivity tends to one, the maxima will have the intensity $I_r/I_0 \to 1$.
The transmitted light shows an opposed behavior so that we have maxima in the first case and minima in the second case:

$$\frac{\phi}{2} = m\pi \quad \Rightarrow \quad \frac{I_t}{I_0} = 1 \tag{3.5.21}$$

$$\frac{\phi}{2} = (2m'+1)\frac{\pi}{2} \quad \Rightarrow \quad \frac{I_t}{I_0} = \frac{(1-R)^2}{(1+R)^2} \tag{3.5.22}$$

R	F
0.04	0.174
0.2	1.25
0.5	8
0.8	80
0.95	1520

Table 3.1: Parameter F as function of the reflectivity R.

So, for small values of the reflectivity R the minima will be less pronounced. But, if the reflectivity tends to one, the minima will tend to zero $I_t/I_0 \to 0$.
It is useful to write equations (3.5.12) and (3.5.15) by introducing the parameter

$$F = \frac{4R}{(1-R)^2} \quad . \tag{3.5.23}$$

Then, we have for the ratio of the intensities between the reflected and incident light on the one hand and between the transmitted and incident light on the other, the following equations:

$$\frac{I_r}{I_0} = \frac{F\sin^2\left(\frac{\phi}{2}\right)}{1+F\sin^2\left(\frac{\phi}{2}\right)} \tag{3.5.24}$$

$$\frac{I_t}{I_0} = \frac{1}{1+F\sin^2\left(\frac{\phi}{2}\right)} \tag{3.5.25}$$

The ratio of the intensities of the reflected light and the incident light as function of the phase difference ϕ can be seen in fig. 3.23, where the different curves correspond to different values of the reflectivity R. Of course different values R mean also different values F (see table 3.1). It can be clearly seen that there are minima at $\phi = 2m\pi$ (integer value m) which become more pronounced and narrower with increasing value R or F. If the reflectivity R tends to one and therefore F to infinity there is only in the immediate neighborhood of $\phi = 2m\pi$ a value different from one and the minimum with the value zero is very narrow.
The behavior of the transmitted light is opposite to that of the reflected light as can be seen in fig. 3.24. There are maxima at $\phi = 2m\pi$ which become more pronounced and narrower with increasing value R or F. If the reflectivity R tends to one and therefore F to infinity there is only in the immediate neighborhood of $\phi = 2m\pi$ a value different from zero and the maximum with the value one is very narrow.
The half–intensity width, or shortly half–width or FWHM (full width at half maximum), ϵ of the minima or maxima for the reflected or transmitted light is the interval of the phase difference ϕ in which the relative intensity I_r/I_0 or I_t/I_0 is smaller or larger as $1/2$, respectively. Since these minima or maxima are at $\phi_m = 2m\pi$ we have to calculate:

$$\left.\begin{array}{l} \dfrac{I_r\left(\phi = 2m\pi \pm \epsilon/2\right)}{I_0} = \dfrac{F\sin^2\left(m\pi \pm \frac{\epsilon}{4}\right)}{1+F\sin^2\left(m\pi \pm \frac{\epsilon}{4}\right)} = \dfrac{F\sin^2\left(\frac{\epsilon}{4}\right)}{1+F\sin^2\left(\frac{\epsilon}{4}\right)} = \dfrac{1}{2} \\ \dfrac{I_t\left(\phi = 2m\pi \pm \epsilon/2\right)}{I_0} = \dfrac{1}{1+F\sin^2\left(m\pi \pm \frac{\epsilon}{4}\right)} = \dfrac{1}{1+F\sin^2\left(\frac{\epsilon}{4}\right)} = \dfrac{1}{2} \end{array}\right\} \quad \Rightarrow \quad \epsilon = \frac{4}{\sqrt{F}} \tag{3.5.26}$$

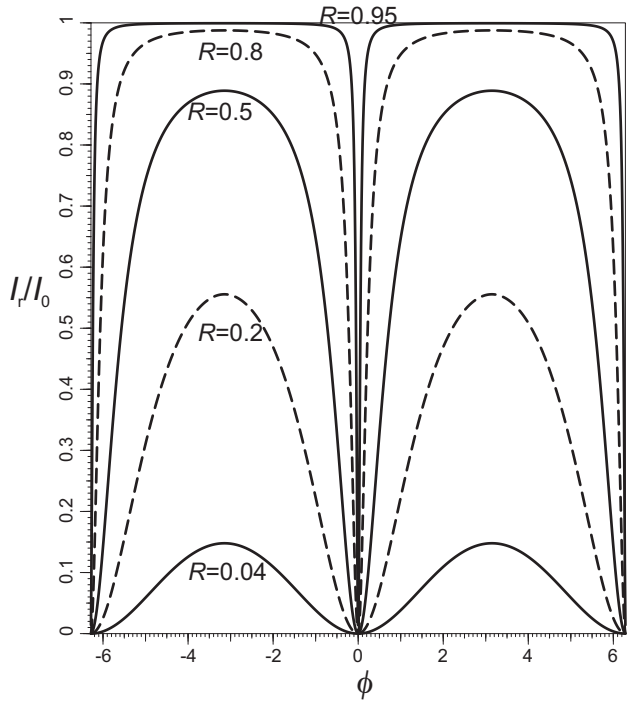


Figure 3.23: Relative intensity I_r/I_0 of the reflected light at a plane–parallel glass plate due to multiple beam interference as function of the phase difference ϕ for different values of the reflectivity R.

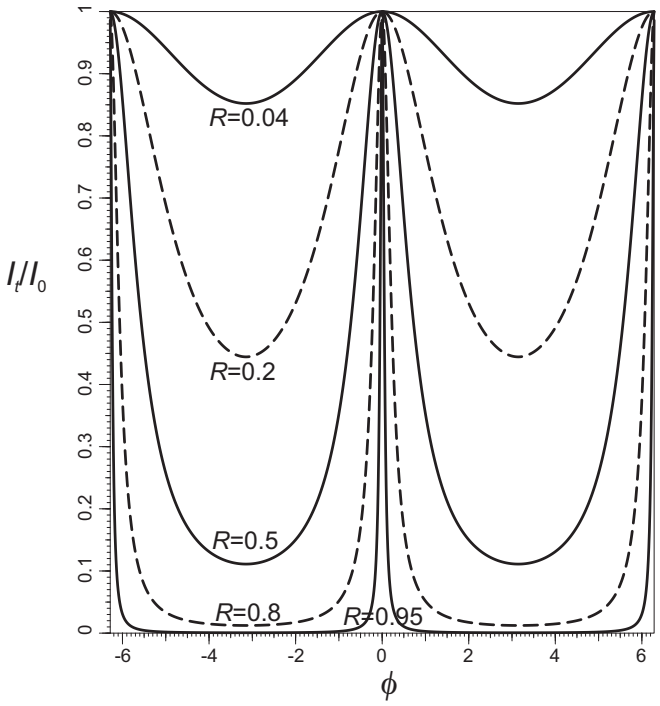


Figure 3.24: Relative intensity I_t/I_0 of the transmitted light at a plane–parallel glass plate due to multiple beam interference as function of the phase difference ϕ for different values of the reflectivity R.

Here, the approximation $\sin\varphi \approx \varphi$ is used because ϵ has a quite small value if the reflectivity R tends to one so that we have pronounced minima and maxima.
Finally, we define the so called **finesse** $\mathcal{F}$ of the multiple beam interference fringes as the ratio between the separation of two fringes, which is 2π, and the half–width ϵ. So, by using equation (3.5.23) we have for the finesse:

$$\mathcal{F} = \frac{2\pi}{\epsilon} = \frac{\pi\sqrt{F}}{2} = \frac{\pi\sqrt{R}}{(1-R)} \tag{3.5.27}$$

3.5.4 Spectral resolution of the multiple beam interference pattern

For quasi–monochromatic light with a very small band width $\Delta\lambda$ around a center wavelength λ_0 the resulting intensity will be the sum of the intensities of all spectral components. Now, the phase difference ϕ depends according to equation (3.5.2) on the wavelength λ. Then, the change $\Delta\phi$ of the phase difference due to changing the wavelength from λ_0 to $\lambda_0 \pm \Delta\lambda$ is:

$$\frac{\partial\phi}{\partial\lambda} = -\frac{4\pi}{\lambda^2} n' h \cos\varphi' \quad \Rightarrow \quad \Delta\phi = \frac{4\pi}{\lambda_0^2} n' h \cos\varphi' \Delta\lambda \tag{3.5.28}$$

As long as this phase difference change $\Delta\phi$ is much smaller than the half–width ϵ of the maximum/minimum for monochromatic light of the wavelength λ_0, there will be only one slightly broadened maximum/minimum and the different spectral components of the quasi–monochromatic light cannot be resolved.
On the other side, if there are two monochromatic spectral components with the wavelengths λ_0 and $\lambda_0 + \Delta\lambda$ these spectral lines can be resolved if the superposition of the intensities of both components (assuming equal intensity in both spectral lines) shows a clearly visible dip. It can be easily shown by taking equation (3.5.25) that there is a dip of about $0.8 I_{max}$ (which is similar to the Rayleigh criterion in a grating or prism spectrograph) if the separation $\Delta\phi$ of the spectral lines using equation (3.5.28) is equal to the half–width ϵ of a multiple beam interference peak (equation (3.5.26)). Figure 3.25 shows a simulation of the superposition of two fringes (reflectivity $R = 0.9$) with different separations $\Delta\phi$. For $\Delta\phi = 0.5\epsilon$ there is no dip at all visible. For $\Delta\phi = \epsilon$ there is a dip with a height of about 0.8 times the maximum height of the superposed intensities, i.e. the Rayleigh criterion is fulfilled. For larger separations the two peaks become more and more resolved. Of course, the Rayleigh criterion with the $0.8 I_{max}$ dip is to some degree arbitrary and in practice it may be the case that also fringes with a little bit less distinct dip or only fringes with a more distinct dip can be resolved.
So, in total a condition for the smallest wavelength interval $\Delta\lambda$ between two spectral components, which can just be resolved using multiple beam interference, is obtained:

$$\Delta\phi = \frac{4\pi}{\lambda_0^2} n' h \cos\varphi' \Delta\lambda \geq \epsilon = \frac{4}{\sqrt{F}} \quad \Rightarrow \quad \Delta\lambda \geq \frac{\lambda_0^2}{\pi n' h \cos\varphi' \sqrt{F}} = \frac{\lambda_0^2}{2 n' h \cos\varphi' \mathcal{F}} \tag{3.5.29}$$

The **spectral resolution** itself, which is defined as the ratio of λ_0 to $\Delta\lambda$, is then

$$\frac{\lambda_0}{\Delta\lambda} \leq \frac{2 n' h \cos\varphi' \mathcal{F}}{\lambda_0} \tag{3.5.30}$$

The maximum of the spectral resolution is obtained near normal incidence, i.e. $\cos\varphi' \approx 1$, provided that the finesse $\mathcal{F}$ is independent of the angle of incidence. Of course, this is not

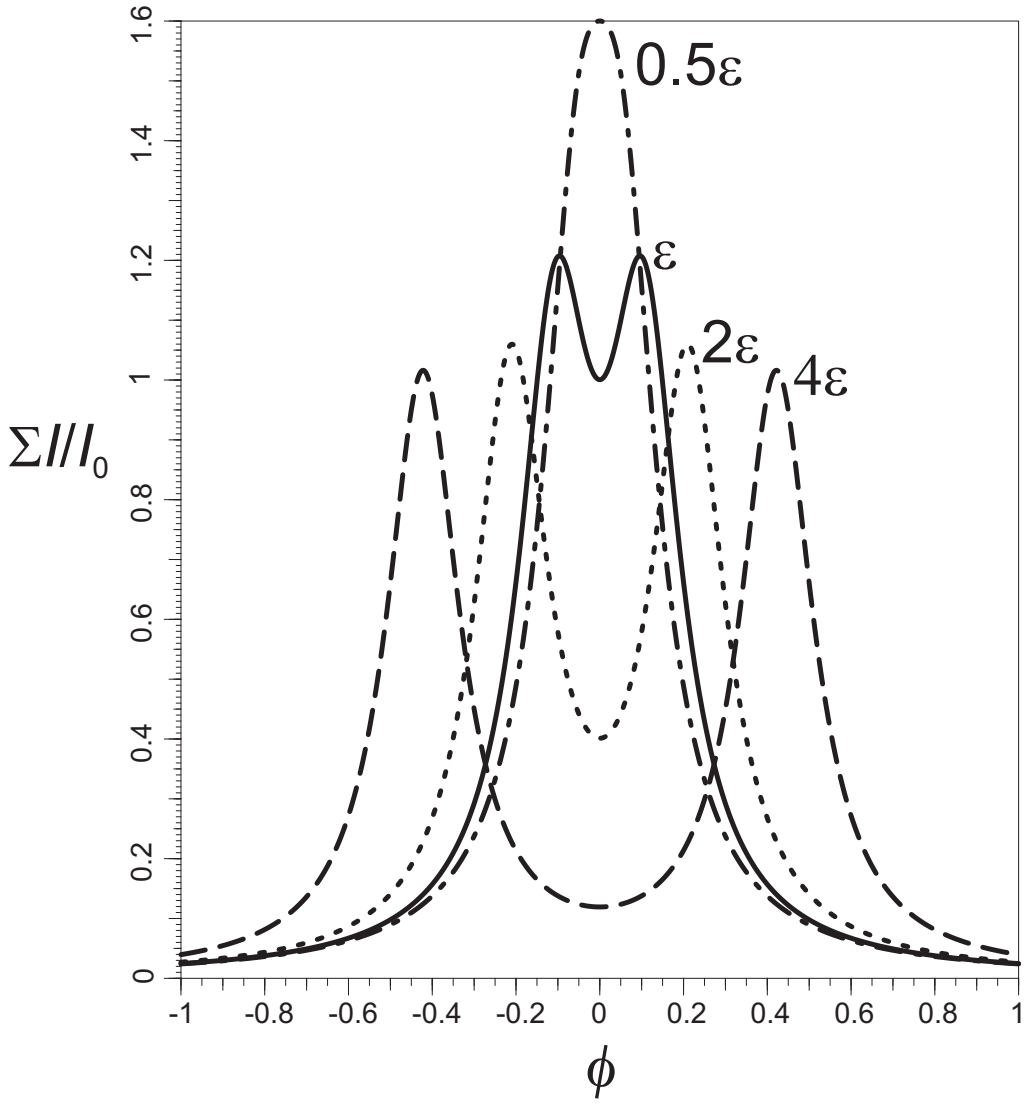


Figure 3.25: Superposition of two multiple beam interference fringes with different angular separations $\Delta\phi$. If the angular separation $\Delta\phi$ is equal to the half–width ϵ a dip with about $0.8I_{max}$ is obtained and the two fringes can be resolved.

the case if we have a pure glass plate. As we will see later the reflectivity R of a glass plate approaches the value one if the angle of incidence (in air) is near to 90 degree. However, for normal incidence the reflectivity of a glass plate with refractive index $n = 1.5$ and air outside is only $R = (n-1)^2/(n+1)^2 = 0.04$. On the other side, a high reflectivity is necessary to achieve a high finesse and a good spectral resolution. But, it is not practical to use the glass plate at a very grazing angle of incidence (i.e. φ near to 90 degree). So, the glass plate is in practice coated with either a thin metal layer or dielectric layers. Then, our derivation of the intensity distribution of the multiple beam interference has to be changed by taking into account losses (for metal layers) or other phase changes than 0 or π for reflection at the air/glass interface. But, the result for the intensity in transmission (equation (3.5.15)) is still valid if we replace equation (3.5.2) by

$$\phi = \frac{4\pi}{\lambda} n' h \cos\varphi' + \delta \tag{3.5.31}$$

with the phase change δ ($|\delta| \leq \pi$) which depends on the type of coating, the angle of incidence and the polarization. Additionally, losses are taken into account by replacing equation (3.5.9) by

$$R + T + A = 1 \tag{3.5.32}$$

with the absorption coefficient A ($A < 1$).
Then, we have:

$$\frac{I_t}{I_0} = \frac{T^2}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} = \frac{(1-R-A)^2}{(1-R)^2 + 4R\sin^2\left(\frac{\phi}{2}\right)} = \left(1 - \frac{A}{(1-R)}\right)^2 \frac{1}{1 + F\sin^2\left(\frac{\phi}{2}\right)} \tag{3.5.33}$$

So, the absorption just decreases the height of the maximum and the phase change δ just alters the effective thickness of the plate by a fraction of the wavelength.
Totally, the spectral resolution remains the same for glass plates with and without coatings and equation (3.5.30) can also be used in these cases. The only difference is that the finesse will in practice have a finite value due to defects of the glass plate (e.g. deviations from planarity), also in the case that the reflectivity R tends to one. It has also to be taken into account that a high reflectivity R in the case of a metal layer causes a certain amount of absorption A so that the peak–transmission $(I_t/I_0)_{max}$ decreases according to equation (3.5.33) with increasing reflectivity. Of course, it is possible to take dielectric layers instead of a metal layer, so that the absorption is very low. But, dielectric layers are very sensitive to wavelength changes so that they can only be used for a quite small wavelength range. So, in practice a thin metal layer is used and a compromise between a high finesse and a high peak–transmission has to be found. But, a finesse of about 10–50 can be easily obtained.
As example, we take a glass plate of optical thickness $n'h = 10$ mm and reflectivity $R = 0.9$. Then, we have for the finesse (3.5.27) $\mathcal{F} = 29.8$. At normal incidence ($\varphi = \varphi' = 0$) and for a wavelength $\lambda_0 = 500$ nm the spectral resolution is

$$\frac{\lambda_0}{\Delta\lambda} \leq \frac{2n'h\cos\varphi'\mathcal{F}}{\lambda_0} = 1.2 \cdot 10^6$$

So, two spectral lines with a wavelength distance of only $\Delta\lambda = (500/\left(1.2 \cdot 10^6\right))$ nm$= 4.2 \cdot 10^{-4}$ nm$= 0.42$ pm can be resolved, if there are no other restricting conditions.

3.5.5 Fabry–Perot interferometer

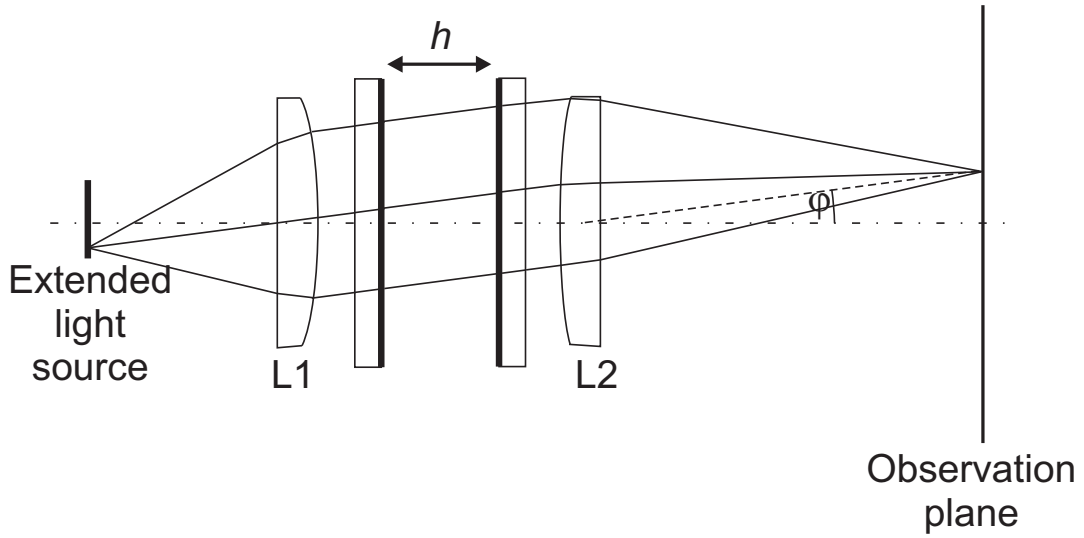


Figure 3.26: Scheme of a Fabry–Perot interferometer for measuring the spectral lines of an extended light source.

A Fabry–Perot interferometer (see fig. 3.26) can be used to measure the spectral composition of light sources with narrow adjacent spectral lines. The light source is assumed to be extended. The lens L1 collimates the light and then it enters a so called **Fabry–Perot etalon**. The Fabry–Perot etalon consists normally of two plane glass plates which have each a coated inner surface (e.g. coated with a thin silver layer). These coated surfaces are exactly parallel, whereas the non coated surfaces of the glass plates have a small tilt angle relative to the other surfaces in order to suppress reflections from these surfaces which would disturb the multiple interference pattern of the coated surfaces. So, the coated surfaces of the Fabry–Perot etalon form a plane–parallel plate with air inside. Since the surfaces are coated they have a quite high reflectivity R and their distance h is the thickness of the 'air' plate. The considerations of the last sections about multiple beam interferences can be easily transformed to this 'air' plate using $n' = 1$ and $\varphi' = \varphi$, where φ is the angle of the collimated beam relative to the optical axis. For a point of the extended light source with a distance Δx from the axis and the focal length f_1 of the lens L1, we have $\varphi = \Delta x / f_1$ for $|\Delta x| \ll |f_1|$. Behind the Fabry–Perot etalon the lens L2 focusses the collimated light beam to the observation plane. Now, for a given wavelength λ there are only multiple beam interference maxima for certain angles φ_m which fulfill the condition $\phi = 2m\pi$ with ϕ defined in equation (3.5.31) and an integer m. Therefore, the Fabry–Perot interferograms consist of concentric rings with radii $r_m = \varphi_m f_2'$ if there is only one spectral line. For another spectral line there is a similar system of concentric rings with other radii.

The thickness h of the Fabry–Perot etalon can be changed in many cases slightly so that the thickness can be adjusted in such a way that for a given wavelength there is a maximum on–axis. This is also a reason why the Fabry–Perot etalon is taken in the Fabry–Perot interferometer instead of just taking a plane–parallel glass plate with coated surfaces. But in principle, a Fabry–Perot interferometer can also be built by using a coated plane–parallel glass plate.

The interferogram in the image plane of the Fabry–Perot interferometer consists of concentric

rings. Now, we have a focussing lens with focal length $f_2' \gg r_{max}$, where r_{max} is the maximum considered radius in the image plane. Then, the field angle $\varphi = \varphi'$ for $r \leq r_{max}$ can be approximated by:

$$\varphi' \approx \frac{r}{f_2'} \ll 1 \quad \Rightarrow \quad \cos\varphi' \approx 1 - \frac{r^2}{2{f_2'}^2} \tag{3.5.34}$$

So, equation (3.5.31) delivers for the phase ϕ

$$\frac{\phi(r)}{2} = \frac{2\pi}{\lambda} n'h\cos\varphi' + \frac{\delta}{2} = \frac{2\pi}{\lambda} n'h + \frac{\delta}{2} - \frac{\pi}{\lambda} n'h \frac{r^2}{{f_2'}^2} \tag{3.5.35}$$

and the intensity distribution of the transmitted light is described by equation (3.5.33):

$$\frac{I_t(r)}{I_0} = \left(1 - \frac{A}{(1-R)}\right)^2 \frac{1}{1 + F\sin^2\left(\frac{2\pi}{\lambda} n'h + \frac{\delta}{2} - \frac{\pi}{\lambda} n'h \frac{r^2}{{f_2'}^2}\right)} \tag{3.5.36}$$

By assuming that the thickness of the Fabry–Perot etalon is adjusted in such a way that the term $2\pi n'h/\lambda + \delta/2$ is an integer multiple of π there is a maximum at $r = 0$ and the other maxima are at:

$$\frac{\pi}{\lambda} n'h \frac{r_m^2}{{f_2'}^2} = m\pi \quad \Rightarrow \quad r_m = f_2' \sqrt{\frac{m\lambda}{n'h}} \tag{3.5.37}$$

3.5.5.1 Simulation examples of Fabry–Perot interferograms

The following simulation example with focal length of lens 2 $f_2' = 100$ mm, wavelength $\lambda = 500$ nm, reflectivity $R = 0.8$ (absorption is neglected, i.e. $A = 0$) and optical thickness of the Fabry–Perot etalon $n'h = 10$ mm (assuming $\delta = 0$) delivers for the radius of the first off–axis maximum $r_1 = 0.707$ mm. This can also be seen in the numerical simulation of figure 3.27. Additionally, two simulations are shown assuming that the light source emits two very narrow spectral lines with equal intensity. Then, the intensity distributions of the two spectral lines can just be added to produce the resulting interferogram. Fig. 3.28 shows the simulation for two spectral lines with $\lambda_1 = 500$ nm and $\lambda_2 = 500.001$ nm, i.e. a wavelength distance of 1 pm. It can be seen that these two spectral lines can just be resolved. By using equations (3.5.27) and (3.5.30) we see that the minimum wavelength range which can be resolved is:

$$\mathcal{F} = \frac{\pi\sqrt{R}}{1-R} = 14.05 \quad \Rightarrow \quad \Delta\lambda \geq \frac{\lambda_0^2}{2n'h\mathcal{F}} = 0.89 \text{ pm}$$

Here, $\cos\varphi' = 1$ and $\lambda_0 = \lambda_1 = 500$ nm has been used. So, the minimum resolvable wavelength range is just a little bit smaller than 1 pm which we used in the simulation.
Another simulation with the two wavelengths $\lambda_1 = 500$ nm and $\lambda_2 = 500.0025$ nm is shown in fig. 3.29. Since here, the two wavelengths have a distance of 2.5 pm, the rings of both wavelengths can be very well resolved.
So, Fabry–Perot interferometers can be used to investigate the hyper fine structure of chemical elements which have in practice wavelength distances of about 1 pm to 10 pm. Of course, if there are other spectral lines in the spectrum a prism or grating spectrograph should be used in front of the Fabry–Perot interferometer which selects only a small wavelength range around the wanted wavelength. Then, there are no disturbing rings of other spectral lines. The reason for

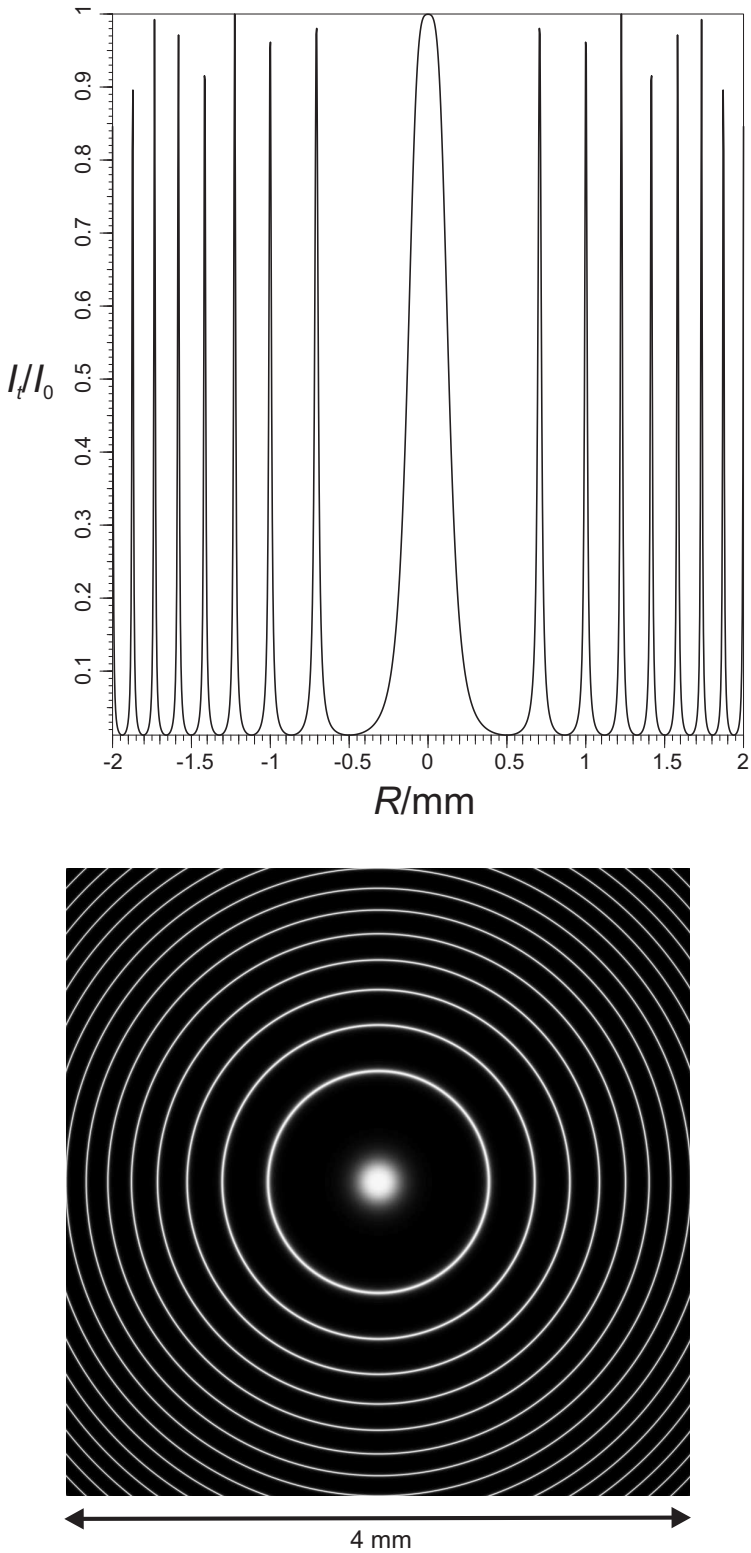
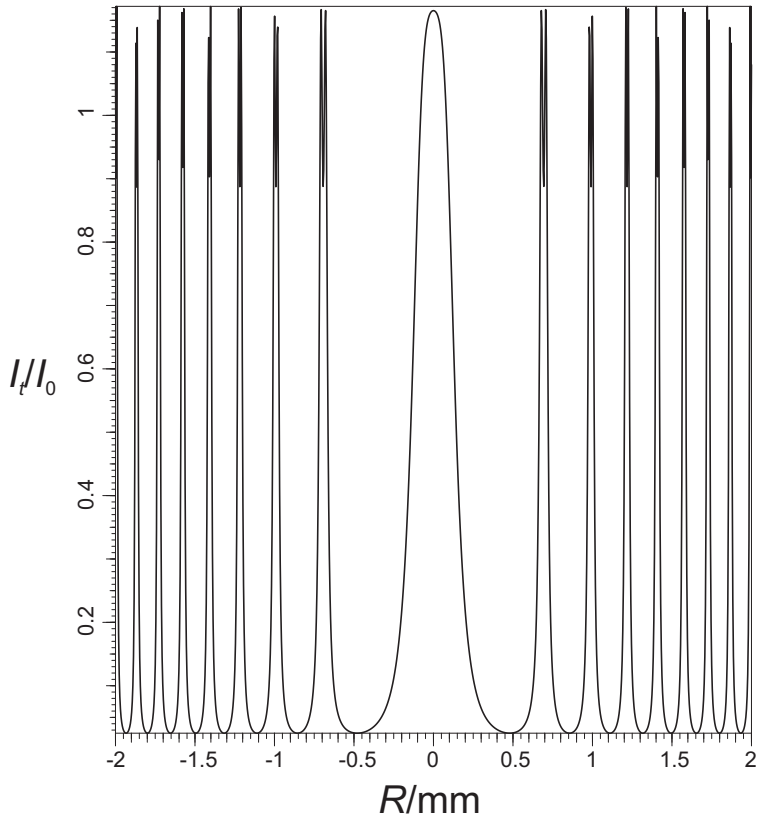


Figure 3.27: Simulation of a Fabry–Perot interferogram for $\lambda = 500$ nm. Top: central section of the intensity distribution, bottom: intensity distribution as seen by a CCD camera.



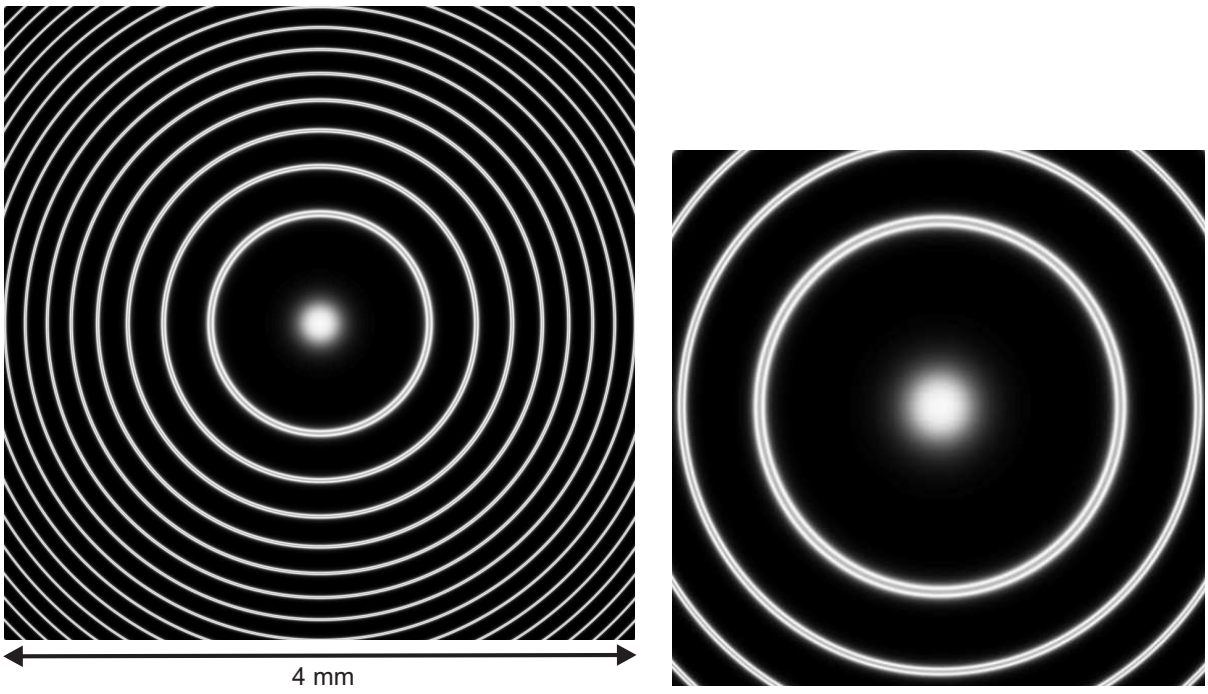


Figure 3.28: Simulation of a Fabry–Perot interferogram for two spectral lines with $\lambda_1 = 500$ nm and $\lambda_2 = 500.001$ nm. Top: central section of the intensity distribution, bottom: intensity distribution as seen by a CCD camera and zoom of the central rings.

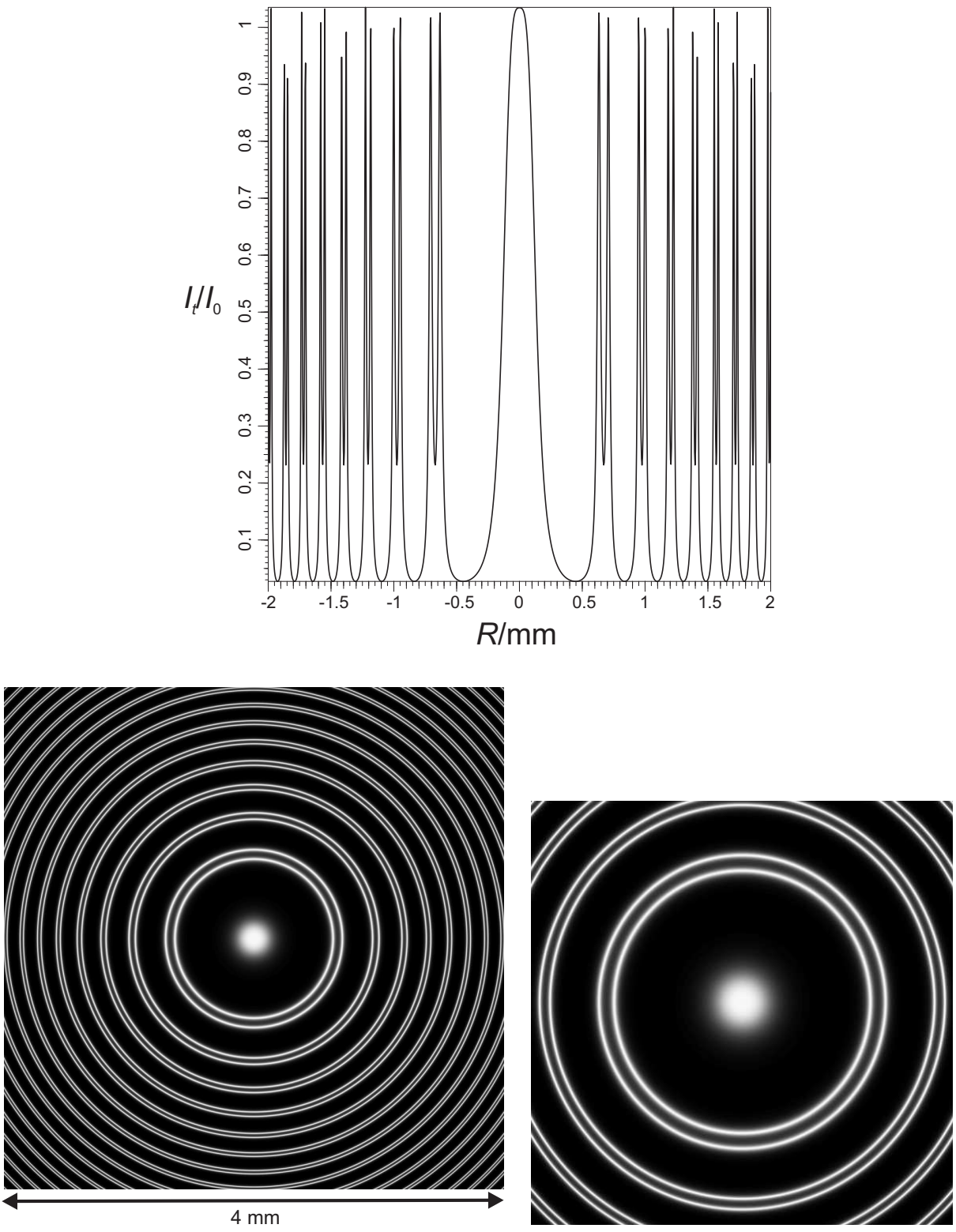


Figure 3.29: Simulation of a Fabry–Perot interferogram for two spectral lines with $\lambda_1 = 500$ nm and $\lambda_2 = 500.0025$ nm. Top: central section of the intensity distribution, bottom: intensity distribution as seen by a CCD camera and zoom of the central rings.

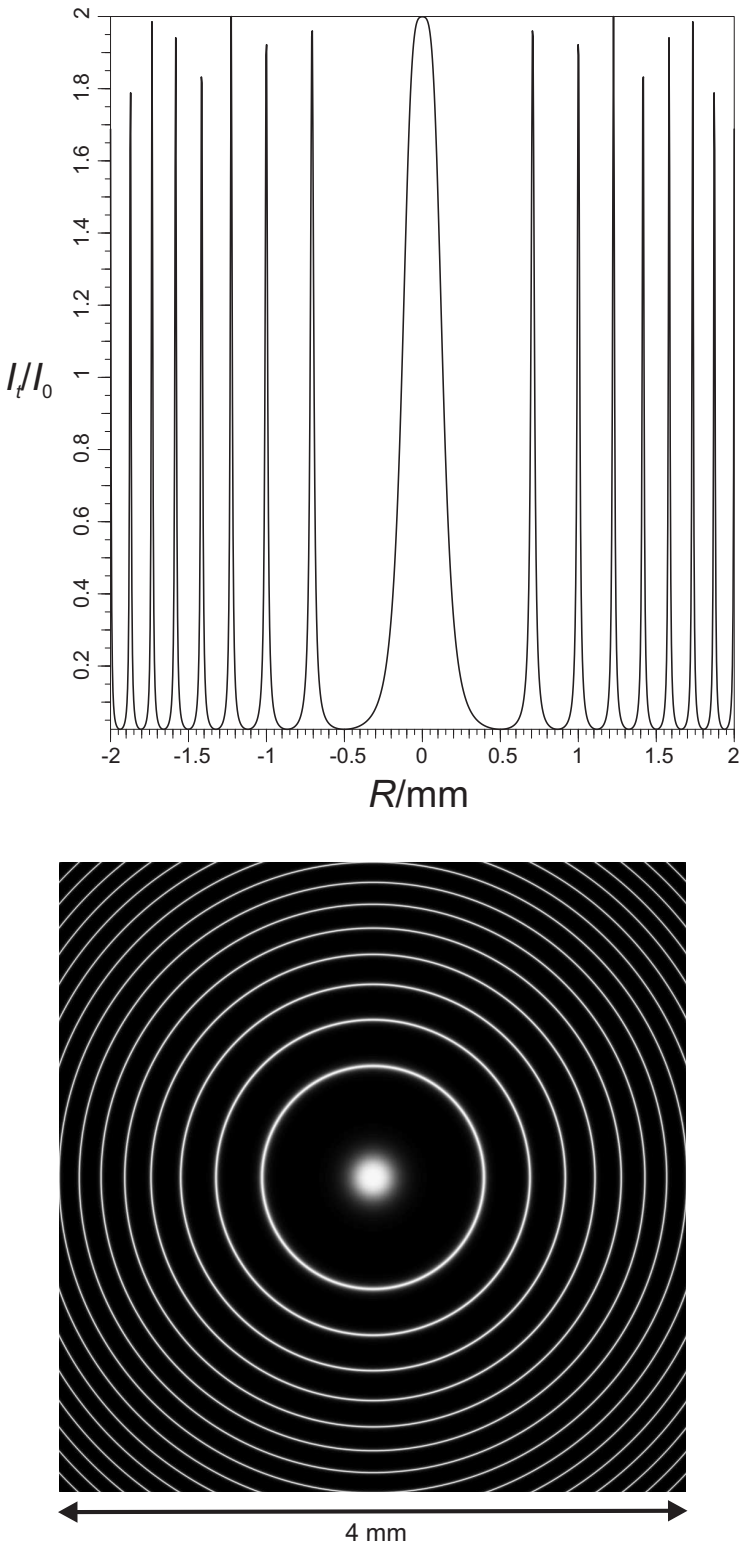


Figure 3.30: Simulation of a Fabry–Perot interferogram for two spectral lines with $\lambda_1 = 500$ nm and $\lambda_2 = 500.0125$ nm. Top: central section of the intensity distribution, bottom: intensity distribution as seen by a CCD camera.

using an additional spectrograph is the small spectral range of a Fabry–Perot interferometer, which can be uniquely resolved, and which is only:

$$(\Delta\lambda)_{max} = \mathcal{F}\Delta\lambda = \frac{\lambda_0^2}{2n'h\cos\varphi'} = 12.5 \text{ pm} \tag{3.5.38}$$

Here, $\lambda_0 = 500$ nm, $n'h = 10$ mm and $\cos\varphi' = 1$, i.e. nearly normal incidence on the Fabry–Perot etalon. It is easy to see that by changing the wavelength from λ_0 to $\lambda_0 \pm (\Delta\lambda)_{max}$ (with $(\Delta\lambda)_{max} \ll \lambda_0$) the variable $\phi/2$ of equation (3.5.35) changes by $\pm\pi$ and, therefore, the intensity distribution described by equation (3.5.36) is not changed:

$$\frac{\phi(\lambda_0)}{2} = \frac{2\pi}{\lambda_0} n'h\cos\varphi' + \frac{\delta}{2} \quad \Rightarrow$$

$$\begin{aligned}
\frac{\phi\left(\lambda_0 \pm (\Delta\lambda)_{max}\right)}{2} &= \frac{2\pi}{\lambda_0\left(1 \pm (\Delta\lambda)_{max}/\lambda_0\right)} n'h\cos\varphi' + \frac{\delta}{2} \approx \\
&\approx \frac{2\pi}{\lambda_0} n'h\cos\varphi' + \frac{\delta}{2} \mp \frac{(\Delta\lambda)_{max}}{\lambda_0}\frac{2\pi}{\lambda_0} n'h\cos\varphi' = \\
&= \frac{2\pi}{\lambda_0} n'h\cos\varphi' + \frac{\delta}{2} \mp \pi
\end{aligned}$$

So, $(\Delta\lambda)_{max}$ is really the spectral range of the Fabry–Perot etalon which can be uniquely determined. This can also be seen in fig. 3.30, where two spectral lines with a distance of 12.5 pm were used. There is only one set of rings like in the case of one spectral line. The only difference is, that the intensity is now doubled because of the two spectral lines having equal intensity. But, this can only be seen in the simulation and not in practice, where the absolute intensity is not known.

Chapter 4

Diffraction

Up to now we investigated mostly the propagation of plane waves and other waves which are not affected by any limiting apertures. A plane wave has e.g. an infinite spatial extension and therefore it does not exist in the real world. Nevertheless, if the diameter of the limiting aperture is very large compared to the wavelength of the light, a plane wave can be a quite good approximation if the propagation distance is not very large. But, also in this case there are disturbances at the rim of the wave which are called **diffraction** effects. In this section the diffraction theory will mostly be treated for scalar waves and only at the end of this section the influence of polarization effects to the electric energy density in the focal region of a lens will be treated [40],[41]. In contrast to most text books of optics like e.g. [1],[42] we will not start historically with the **Huygens–Fresnel principle** or with the **integral theorem of Helmholtz and Kirchhoff**, but we will start with the **angular spectrum of plane waves**. Only **Kirchhoff's boundary conditions** will be used, i.e. a wave which is incident on an absorbing screen with a hole will be undisturbed in the area of the hole and completely absorbed in the other parts of the screen. Starting from the angular spectrum of plane waves the **Fresnel–Kirchhoff diffraction integral** will be derived and it will be shown that both formulations are nearly equivalent [43],[44],[45]. The approximations of **Fresnel diffraction** and **Fraunhofer diffraction** will be discussed afterwards. A quite interesting application of Fraunhofer diffraction is e.g. the calculation of the intensity distribution in the focal region of a lens [1],[46]. Afterwards, some ideas to the numerical implementation of scalar diffraction formula are given [45]. At the end of this section, we will reflect a little bit about the combination of polarization and diffraction by using the superposition of plane waves taking into account their polarization states. This is used to calculate the influence of polarization effects to the electric energy density in the focal region of a lens.
There are many modern applications of diffraction and interference effects in optical holography [47],[48],[49],[50] and computer–generated diffractive optics [51],[52],[53],[54],[55],[56],[57],[58]. Some of these subjects will be treated in chapter 7 about holography. For more information we refer to the literature.

4.1 The angular spectrum of plane waves

The knowledge of the angular spectrum of plane waves allows the exact propagation of a complex amplitude function u from one plane (which is chosen perpendicular to the z–axis) to another

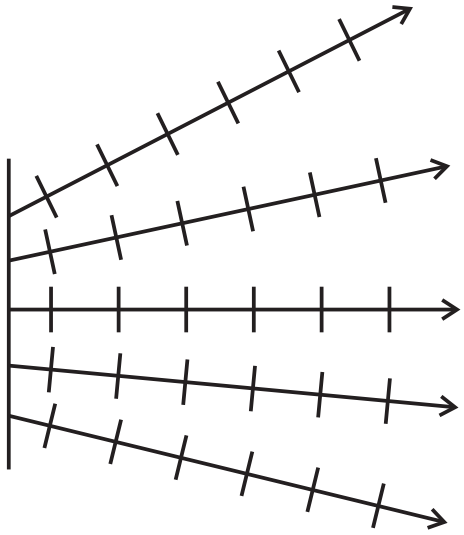


Figure 4.1: Arbitrary scalar wave as superposition of plane waves.

parallel plane in the distance z_0 in a homogeneous and isotropic material with the refractive index n. The only approximation is here that u is assumed to be a scalar function. But, since a plane wave can be defined easily taking into account polarization (see section 2) an extension of this formalism is possible, but it cannot be treated here.
According to equation (1.1.52) a scalar plane wave

$$u(\boldsymbol{r}) = u_0 e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \tag{4.1.1}$$

fulfills the Helmholtz equation (1.5.9) which is written for scalar waves as

$$\left(\boldsymbol{\nabla}^2 + k^2\right) u\left(x, y, z\right) = 0 \tag{4.1.2}$$

The condition for the modulus $|\boldsymbol{k}| = k$ of the wave vector is:

$$|\boldsymbol{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2} = \frac{2\pi n}{\lambda} \tag{4.1.3}$$

According to the linearity of the Helmholtz equation a sum of plane waves with different directions of propagation is also a solution of the Helmholtz equation and in the limit a continuous spectrum of plane waves is a solution (see fig. 4.1). There, the integration has to be done over two angles or more exactly over two components of the wave vector. The third component is then automatically defined by equation (4.1.3) as long as only plane waves propagating in the positive z–direction are taken into account, what will be the case here. Since the complex amplitude is here always regarded in a plane perpendicular to the z–axis, the two components of the wave vector used for the integration will be the x– and y–components. To obtain a symmetrical formulation the vector $\boldsymbol{\nu}$ of the **spatial frequencies** (*dt.:* Ortsfrequenzen) is introduced by

$$\boldsymbol{\nu} = \frac{1}{2\pi}\boldsymbol{k} = \begin{pmatrix} \nu_x \\ \nu_y \\ \nu_z \end{pmatrix} \quad \text{with} \quad |\boldsymbol{\nu}| = \sqrt{\nu_x^2 + \nu_y^2 + \nu_z^2} = \frac{n}{\lambda} \tag{4.1.4}$$

The complex amplitude $u(\boldsymbol{r})$ of a wave can then be written as a superposition of plane waves:

$$u(\boldsymbol{r}) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}\left(\nu_x, \nu_y\right) e^{2\pi i \boldsymbol{\nu}\cdot\boldsymbol{r}} d\nu_x d\nu_y \tag{4.1.5}$$

The integration takes into account arbitrary spatial frequencies ν_x and ν_y, so that later the mathematical formalism of the Fourier transformation [59], [60] can be used. To fulfill equation (4.1.4) the z–component of the vector of the spatial frequencies is defined by the x– and y–components since we allow only waves propagating in the positive z–direction:

$$\nu_z = \sqrt{\frac{n^2}{\lambda^2} - \nu_x^2 - \nu_y^2} \tag{4.1.6}$$

Nevertheless, the square root delivers only for a positive argument a real solution. Therefore, two cases have to be distinguished

$$\nu_x^2 + \nu_y^2 \leq \frac{n^2}{\lambda^2} \quad \Rightarrow \quad e^{2\pi i \nu_z z} = e^{2\pi i z\sqrt{\frac{n^2}{\lambda^2} - \nu_x^2 - \nu_y^2}} \tag{4.1.7}$$

$$\nu_x^2 + \nu_y^2 > \frac{n^2}{\lambda^2} \quad \Rightarrow \quad e^{2\pi i \nu_z z} = e^{-2\pi z\sqrt{\nu_x^2 + \nu_y^2 - \frac{n^2}{\lambda^2}}} \tag{4.1.8}$$

whereby in both cases the result of the square root is a real number. The second case corresponds to an exponentially decreasing amplitude and so the waves with such high spatial frequencies propagate only along very small distances z of the range of some few wavelengths and are called **evanescent waves**. The other mathematically possible solution with exponentially increasing amplitude is useless for free space propagation and therefore excluded here.
If the complex amplitude u_0 of a wave is known in a plane and the coordinate system is chosen such that this plane is perpendicular to the z–axis at $z = 0$, it holds according to equation (4.1.5):

$$u_0(x, y, 0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0\left(\nu_x, \nu_y, 0\right) e^{2\pi i(\nu_x x + \nu_y y)} d\nu_x d\nu_y \tag{4.1.9}$$

$\tilde{u}_0(\nu_x, \nu_y, 0)$ is the Fourier transform of u_0 in the plane at $z = 0$ and can be calculated using the Fourier relation

$$\tilde{u}_0\left(\nu_x, \nu_y, 0\right) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u_0(x, y, 0) e^{-2\pi i(\nu_x x + \nu_y y)} dx dy \tag{4.1.10}$$

Since now $\tilde{u}_0$ is known the complex amplitude u in another parallel plane at $z = z_0$ can be calculated with equation (4.1.5) and (4.1.6):

$$u(x, y, z_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0\left(\nu_x, \nu_y, 0\right) e^{2\pi i\left(\nu_x x + \nu_y y\right)} e^{2\pi i \nu_z z_0} d\nu_x d\nu_y \tag{4.1.11}$$

$$\Rightarrow\ u(x,y,z_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0\left(\nu_x,\nu_y,0\right) e^{2\pi i \frac{nz_0}{\lambda}\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}} \cdot e^{2\pi i\left(\nu_x x+\nu_y y\right)} d\nu_x d\nu_y \tag{4.1.12}$$

So, this is an inverse Fourier transformation whereby the function

$$\tilde{u}\left(\nu_x,\nu_y,z_0\right) = \tilde{u}_0\left(\nu_x,\nu_y,0\right) e^{2\pi i \frac{nz_0}{\lambda}\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}} \tag{4.1.13}$$

has to be Fourier transformed. In total, by using a Fourier transformation (see equation (4.1.10)), multiplying $\tilde{u}_0$ with the propagation factor $\exp(2\pi i \nu_z z_0)$ and applying an inverse Fourier transformation (see equation (4.1.12) for both operations) the complex amplitude in a plane parallel to the original plane in the distance z_0 can be calculated. Hereby, it has to be taken into account that according to equations (4.1.7) and (4.1.8) the propagation factor $\exp(2\pi i \nu_z z_0)$ can either be a pure phase factor (for $\nu_x^2+\nu_y^2 \leq n^2/\lambda^2$) or an exponentially decreasing real term.
The propagation factor is also known as the **transfer function of free space** H (*dt.:* Übertragungsfunktion der Freiraumausbreitung):

$$H\left(\nu_x,\nu_y,z_0\right) = e^{2\pi i \frac{nz_0}{\lambda}\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}} \tag{4.1.14}$$

Then, equation (4.1.12) can be written as:

$$u(x,y,z_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0(\nu_x,\nu_y,0)\, H(\nu_x,\nu_y,z_0)\, e^{2\pi i\left(\nu_x x+\nu_y y\right)} d\nu_x d\nu_y \tag{4.1.15}$$

4.2 Rayleigh–Sommerfeld diffraction formula and angular spectrum of plane waves

In this section, the equivalence of the angular spectrum of plane waves and the Rayleigh–Sommerfeld diffraction formula will be shown.
According to the convolution theorem of Fourier mathematics equation (4.1.15) can be written as a convolution of two functions, whereby these two functions are the inverse Fourier transforms of $\tilde{u}_0$ and H. The inverse Fourier transform of $\tilde{u}_0$ is due to equation (4.1.9) the complex amplitude distribution u_0 at $z=0$. The other term is not so obvious. But in [43] and [44] it is shown that the following relation is valid:

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{2\pi i \frac{nz_0}{\lambda}\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}} e^{2\pi i\left(\nu_x x+\nu_y y\right)} d\nu_x d\nu_y =$$

$$= -\frac{1}{2\pi}\frac{\partial}{\partial z_0}\left(\frac{e^{2\pi i \frac{nr}{\lambda}}}{r}\right) = -\frac{1}{2\pi}\frac{\partial}{\partial z_0}\left(\frac{e^{ikr}}{r}\right) = -\frac{1}{2\pi}\left(ik-\frac{1}{r}\right)\frac{z_0}{r}\frac{e^{ikr}}{r} \tag{4.2.1}$$

whereby $r := \sqrt{x^2 + y^2 + z_0^2}$ and $k := 2\pi n/\lambda$.
So, in total equation (4.1.12) can be written as:

$$\begin{aligned} u(x,y,z_0) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0(\nu_x,\nu_y,0)\, e^{ikz_0\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}}\cdot \\ &\quad \cdot e^{2\pi i\left(\nu_x x+\nu_y y\right)} d\nu_x d\nu_y = \\ &= -\frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u_0\left(x',y',0\right)\left(ik-\frac{1}{l}\right)\frac{z_0}{l}\frac{e^{ikl}}{l}dx'dy' \end{aligned} \tag{4.2.2}$$

with

$$l := \sqrt{(x-x')^2+(y-y')^2+z_0^2} \tag{4.2.3}$$

The right–hand side of equation (4.2.2) is known as the **general Rayleigh–Sommerfeld diffraction formula**. So, the complex amplitude $u(x,y,z_0)$ can be either expressed as a superposition of plane waves or as a convolution of the original complex amplitude $u_0(x,y,0)$ with a **spherical Huygens wavelet** h of the form

$$h(x,y,z_0) = -\frac{1}{2\pi}\left(ik-\frac{1}{r}\right)\frac{z_0}{r}\frac{e^{ikr}}{r} \tag{4.2.4}$$

which is the inverse Fourier transform of the transfer function of free space H. h is also called the **impulse response** (*dt.:* Impulsantwort) since it results if the stimulating complex amplitude $u_0(x,y,0)$ has the form of a δ–function. Equation (4.2.2) is a mathematical expression of the **Huygens–Fresnel principle**. The term z_0/l is the cosine obliquity factor. By using equations (4.2.2) and (4.2.4) the complex amplitude can be written as

$$u(x,y,z_0) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u_0\left(x',y',0\right) h\left(x-x',y-y',z_0\right) dx'dy' \tag{4.2.5}$$

i.e. as a convolution of the original complex amplitude u_0 in the first plane and the impulse response h.
In most cases, the term r is much larger than the wavelength in the medium λ/n. Then the relation $k \gg 1/r$ is valid and the impulse response of equation (4.2.4) can be written as:

$$h(x,y,z_0) = -\frac{1}{2\pi}\left(ik-\frac{1}{r}\right)\frac{z_0}{r}\frac{e^{ikr}}{r} \approx -i\frac{n}{\lambda}\frac{z_0}{r}\frac{e^{ikr}}{r} \tag{4.2.6}$$

Then equation (4.2.2) results in the more familiar but less general Rayleigh–Sommerfeld diffraction formula

$$u(x,y,z_0) \approx -i\frac{n}{\lambda}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u_0\left(x',y',0\right)\frac{z_0}{l}\frac{e^{ikl}}{l}dx'dy' \tag{4.2.7}$$

In the case that the complex amplitude u_0 is different from zero inside an aperture A and zero outside of the aperture A, equation (4.2.7) is also known as the **Fresnel–Kirchhoff diffraction**

integral [1]. Then, the effective integration is not carried out from minus infinity to plus infinity, but, on the area of the aperture A.
It has to be mentioned that equation (4.2.2) can also be written in a similar form which does not explicitly assume that the original complex amplitude which is now named $u_0(\boldsymbol{r'})$ has to be known in a plane and that the complex amplitude $u(\boldsymbol{r})$ which has to be calculated is also defined in a plane. The generalization is:

$$u(\boldsymbol{r}) = -\frac{1}{2\pi}\iint\limits_A u_0(\boldsymbol{r'})\left(ik - \frac{1}{l}\right)\frac{\boldsymbol{N}\cdot(\boldsymbol{r}-\boldsymbol{r'})}{l}\frac{e^{ikl}}{l}dS \tag{4.2.8}$$

Here, $\boldsymbol{r'} = (x', y', z')$ defines an arbitrary point on a curved surface (the aperture A), where the original complex amplitude u_0 is defined, and $\boldsymbol{r} = (x, y, z)$ is an arbitrary point on the second curved surface, on which the complex amplitude u has to be calculated. Additionally, $\boldsymbol{N}$ is a unit vector perpendicular to the aperture A at the point $\boldsymbol{r'}$. The integration is done over the aperture A and the integration element dS just indicates a two–dimensional integration. Additionally, the distance l is then defined by:

$$l := \sqrt{|\boldsymbol{r}-\boldsymbol{r'}|^2} \tag{4.2.9}$$

Nevertheless, in the following we will always assume that both complex amplitudes u_0 and u are defined in parallel planes. This allows e.g., as we will see in the next section, that approximate integrals like the Fresnel diffraction integral can be numerically calculated using the efficient Fast Fourier transformation.

4.3 The Fresnel and the Fraunhofer diffraction integral

In the following, we always assume that the plane aperture A, on which the integration of the diffraction integral is carried out, is limited and has a maximum diameter of D. There is no other restriction on the form of the aperture which can be circular, rectangular or irregularly formed. The parameter D is therefore the diameter of a circle which contains the aperture and which is centered around the z–axis. Again, we have the complex amplitude $u_0(x', y', 0)$ in a first plane at $z = 0$, where u_0 is zero outside of the aperture A. Additionally, the distance z_0 of the second plane to the first parallel plane is much larger than the diameter D of the aperture A, i.e. $D \ll z_0$.
Then, the distance l (see equation (4.2.3)) of a point $P' = (x', y', 0)$ in the first plane and a point $P = (x, y, z_0)$ in the second plane can be written as (see figure 4.2):

$$\begin{aligned} l &= \sqrt{(x-x')^2 + (y-y')^2 + z_0^2} = \\ &= \sqrt{x^2 + y^2 + z_0^2 + x'^2 + y'^2 - 2xx' - 2yy'} = \\ &= \sqrt{x^2 + y^2 + z_0^2}\sqrt{1 + \frac{x'^2 + y'^2 - 2xx' - 2yy'}{x^2 + y^2 + z_0^2}} \end{aligned} \tag{4.3.1}$$

We define the term l_0 as

$$l_0 := \sqrt{x^2 + y^2 + z_0^2} \gg D \Rightarrow \frac{D}{l_0} \ll 1 \tag{4.3.2}$$

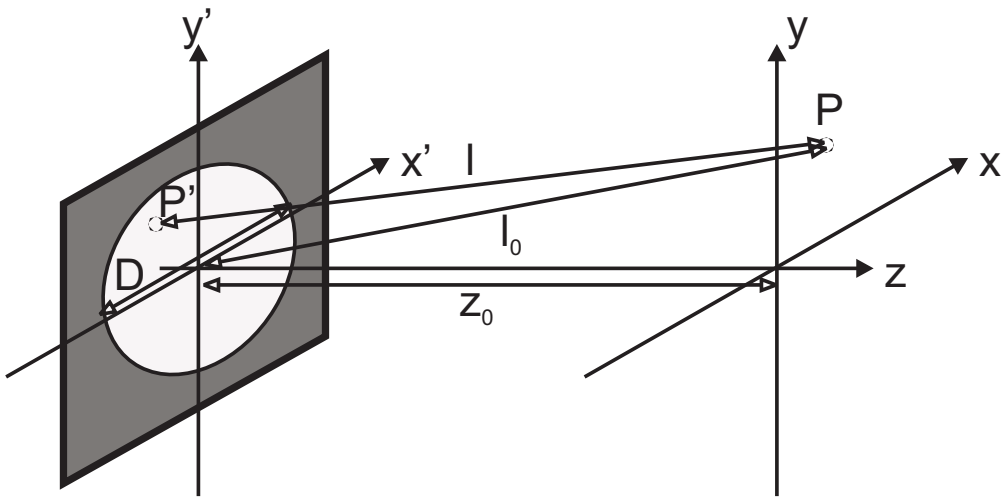


Figure 4.2: Coordinate systems used in the calculation of the different diffraction integrals.

Because of $x' \leq D/2$, $y' \leq D/2$ and equation (4.3.2) all terms of the order x'^3/l_0^3, y'^3/l_0^3 and higher can be neglected. Then l can be approximated by the first terms of its Taylor expansion:

$$\begin{aligned} l &= l_0\sqrt{1+\frac{x'^2+y'^2-2xx'-2yy'}{l_0^2}} \approx \\ &\approx l_0\left[1+\frac{x'^2+y'^2}{2l_0^2}-\frac{xx'+yy'}{l_0^2}-\frac{\left(x'^2+y'^2-2xx'-2yy'\right)^2}{8l_0^4}\right] = \\ &= l_0+\frac{x'^2+y'^2}{2l_0}-\frac{xx'+yy'}{l_0}-\frac{\left(xx'+yy'\right)^2}{2l_0^3}+ \\ &\quad +\frac{\left(x'^2+y'^2\right)\left(xx'+yy'\right)}{2l_0^3}-\frac{\left(x'^2+y'^2\right)^2}{8l_0^3} \end{aligned} \tag{4.3.3}$$

The last two terms are of the order x'^3/l_0^3, y'^3/l_0^3 or higher. Therefore, they can be neglected and the result is:

$$l \approx l_0+\frac{x'^2+y'^2}{2l_0}-\frac{xx'+yy'}{l_0}-\frac{\left(xx'+yy'\right)^2}{2l_0^3} \tag{4.3.4}$$

l in the denominator of the Huygens wavelet of equation (4.2.7) can then be replaced with a good approximation by only the first term of the Taylor expansion, i.e. l_0, and the cosine obliquity factor z_0/l can be replaced by the term z_0/l_0. This means that the Fresnel–Kirchhoff diffraction integral (4.2.7) can be approximated by

$$u(x,y,z_0) \approx -i\frac{n}{\lambda l_0}\frac{z_0}{l_0}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} u_0\left(x',y',0\right)e^{ikl}dx'dy' \tag{4.3.5}$$

whereby l is defined via equation (4.3.4). This equation is valid as long as the last two terms of equation (4.3.3) can be neglected in the exponential factor $\exp(ikl)$. Therefore, two conditions

have to be valid:

$$k\frac{\left(x'^2+y'^2\right)\left(xx'+yy'\right)}{2l_0^3} \ll \pi \Rightarrow n\frac{\left(x'^2+y'^2\right)\left(xx'+yy'\right)}{\lambda l_0^3} \ll 1 \tag{4.3.6}$$

$$k\frac{\left(x'^2+y'^2\right)^2}{8l_0^3} \ll \pi \Rightarrow n\frac{\left(x'^2+y'^2\right)^2}{4\lambda l_0^3} \ll 1 \tag{4.3.7}$$

There are two especially interesting approximations of the Fresnel–Kirchhoff diffraction integral. The first is the Fresnel diffraction integral where only points P in the neighborhood of the axis, i.e. in the paraxial regime, are considered, and the second is the Fraunhofer diffraction integral where only points in the far field or in the focal plane of a lens are considered.

4.3.1 The Fresnel diffraction integral

In the paraxial regime, i.e. $x^2+y^2 \le D^2$, we can write

$$x^2+y^2 \le D^2 \ll z_0^2 \Rightarrow l_0 = z_0\sqrt{1+\frac{x^2+y^2}{z_0^2}} \approx z_0 + \frac{x^2+y^2}{2z_0} \tag{4.3.8}$$

The second term is only interesting in the rapidly oscillating exponential factor $\exp(ikl_0)$. In the other cases, we can write $l_0 \approx z_0$ and $z_0/l_0 \approx 1$. Additionally, we have to approximate $1/l_0$ in some terms in the exponential factor:

$$\frac{1}{l_0} = \frac{1}{z_0}\left(1+\frac{x^2+y^2}{z_0^2}\right)^{-1/2} \approx \frac{1}{z_0} - \frac{x^2+y^2}{2z_0^3} \tag{4.3.9}$$

Since the term $1/l_0$ appears only in terms which are themselves of the order x'^2, y'^2 or xx', yy', the second term $(x^2+y^2)/(2z_0^3)$ can be neglected because it would lead to terms of higher order. In the paraxial regime the last term in equation (4.3.4) can also be neglected. Finally, we obtain from equation (4.3.5):

$$\begin{aligned} u(x,y,z_0) &= -i\frac{n}{\lambda z_0}e^{i\frac{2\pi n z_0}{\lambda}}\, e^{i\pi n\frac{x^2+y^2}{\lambda z_0}} \cdot \\ &\quad \cdot \iint\limits_A u_0\left(x',y',0\right) e^{i\pi n\frac{x'^2+y'^2}{\lambda z_0}}\, e^{-2\pi i n\frac{xx'+yy'}{\lambda z_0}}\, dx'dy' \end{aligned} \tag{4.3.10}$$

This is the **Fresnel diffraction integral**, whereby the integration is made over the aperture A. The condition for the validity of the Fresnel diffraction integral is according to equations (4.3.6) and (4.3.7):

$$\begin{aligned} & n\frac{\left(x'^2+y'^2\right)\left(xx'+yy'\right)}{\lambda l_0^3} \ll 1 \text{ and } n\frac{\left(x'^2+y'^2\right)^2}{4\lambda l_0^3} \ll 1 \\ & \Rightarrow Q_{Fresnel} := \frac{n(D/2)^4}{\lambda z_0^3} = \frac{nD^4}{16\lambda z_0^3} = \frac{\lambda}{nz_0}F^2 \ll 1 \end{aligned} \tag{4.3.11}$$

Here, the quantity $F = n(D/2)^2/(\lambda z_0)$ is introduced which is known as the **Fresnel number** (see equation (4.3.53) on page 113).
As an example, we take n=1, λ=0.5 μm, D=10 mm and z_0=1 m. Then, $F = 50$ and the term $Q_{Fresnel}$ has the value $Q_{Fresnel}$=0.00125. Therefore, the condition for the validity of the Fresnel diffraction integral is very well fulfilled. If the distance z_0 is only 0.1 m, we have $F = 500$ and the term $Q_{Fresnel}$ is 1.25. Then, the Fresnel approximation is at the limit of its validity. This shows that the Fresnel diffraction integral is a good approximation in a distance between the near and the far field. In the near field (which ranges from $z_0 = 0$ up to a distance z_0 of several times D) the Fresnel–Kirchhoff diffraction integral or the angular spectrum of plane waves has to be used. In the far field, there is another more simple approximation, the Fraunhofer diffraction formula, which will be discussed in the next section. But, before doing this, the Fresnel diffraction integral of equation (4.3.10) will be discussed a little bit more.
Equation (4.3.10) shows that the integral itself is formally the Fourier transformation of the function

$$f\left(x',y'\right) = \begin{cases} u_0\left(x',y',0\right)\exp\left(i\pi n\frac{x'^2+y'^2}{\lambda z_0}\right) & \text{if} \quad (x',y') \in A \\ 0 & \text{if} \quad (x',y') \notin A \end{cases} \tag{4.3.12}$$

This gives the quite efficient possibility of calculating the integral numerically by using a fast Fourier transformation (FFT) [60]. But equation (4.3.10) can also be written in a different form:

$$u(x,y,z_0) = -i\frac{n}{\lambda z_0}e^{i\frac{2\pi n z_0}{\lambda}} \iint\limits_A u_0\left(x',y',0\right) e^{i\pi n\frac{(x-x')^2+(y-y')^2}{\lambda z_0}} dx'dy' \tag{4.3.13}$$

So, this form shows the Fresnel diffraction integral as a convolution of the functions u_0 and $\exp(i\pi n(x'^2+y'^2)/(\lambda z_0))$. According to the convolution theorem the Fresnel diffraction integral can then formally be written as:

$$\begin{aligned} u(x,y,z_0) &= -i\frac{n}{\lambda z_0}e^{i\frac{2\pi n z_0}{\lambda}} \cdot \\ &\quad \cdot \mathrm{FT}^{-1}\left\{\mathrm{FT}\left\{u_0(x',y',0)\right\}\cdot \mathrm{FT}\left\{e^{i\pi n\frac{x'^2+y'^2}{\lambda z_0}}\right\}\right\} \end{aligned} \tag{4.3.14}$$

So, to solve this equation we have first to find the Fourier transforms of the functions u_0 and $\exp(i\pi n(x'^2+y'^2)/(\lambda z_0))$.
The first Fourier pair is trivial because according to equation (4.1.10) we have just formally defined $\tilde{u}_0$ as the Fourier transform of the complex amplitude u_0 in the starting plane at $z = 0$ and a concrete evaluation can only be done if u_0 is known analytically or numerically. So, we have formally:

$$\begin{aligned} \tilde{u}_0\left(\nu_x,\nu_y,0\right) &= \int\limits_{-\infty}^{+\infty}\int\limits_{-\infty}^{+\infty} u_0(x',y',0)e^{-2\pi i(\nu_x x'+\nu_y y')}dx'dy' = \\ &= \mathrm{FT}\left\{u_0(x',y',0)\right\} \end{aligned} \tag{4.3.15}$$

The second Fourier pair is:

$$\begin{aligned}
&\mathrm{FT}\left\{ e^{i\pi n\frac{x'^2+y'^2}{\lambda z_0}} \right\} = \\
&= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{i\pi n\frac{x'^2+y'^2}{\lambda z_0}} e^{-2\pi i(\nu_x x' + \nu_y y')} dx'dy' = \\
&= i\frac{\lambda z_0}{n} e^{-i\pi\frac{\lambda z_0}{n}\left(\nu_x^2+\nu_y^2\right)}
\end{aligned} \tag{4.3.16}$$

In total, equation (4.3.14) results in:

$$\begin{aligned}
u(x,y,z_0) \;=\; & e^{i\frac{2\pi n z_0}{\lambda}} \cdot \\
& \cdot \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \tilde{u}_0(\nu_x,\nu_y,0)\, e^{-i\pi\frac{\lambda z_0}{n}\left(\nu_x^2+\nu_y^2\right)} e^{2\pi i(\nu_x x+\nu_y y)} d\nu_x d\nu_y
\end{aligned} \tag{4.3.17}$$

This is the Fresnel diffraction integral expressed in the Fourier domain. The same equation can also be obtained from the angular spectrum of plane waves (equation (4.1.12)) by expanding the square root of the transfer function of free space in a Taylor series:

$$\begin{aligned}
&\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)} \approx 1-\frac{\lambda^2}{2n^2}\left(\nu_x^2+\nu_y^2\right) \\
&\Rightarrow\; e^{2\pi i\frac{nz_0}{\lambda}\sqrt{1-\frac{\lambda^2}{n^2}\left(\nu_x^2+\nu_y^2\right)}} \approx e^{2\pi i\frac{nz_0}{\lambda}} e^{-i\pi\frac{\lambda z_0}{n}\left(\nu_x^2+\nu_y^2\right)}
\end{aligned} \tag{4.3.18}$$

This approximation is valid as long as the higher order terms do not contribute to a considerable variation of the exponential factor. The condition for this is:

$$\begin{aligned}
&2\pi\frac{nz_0}{\lambda}\frac{\lambda^4}{8n^4}\left(\nu_x^2+\nu_y^2\right)^2 = \pi\frac{\lambda^3 z_0}{4n^3}\left(\nu_x^2+\nu_y^2\right)^2 \ll \pi \\
&\Rightarrow\; Q_{Fresnel,Fourier} := \frac{\lambda^3 z_0}{4n^3}\left(\nu_x^2+\nu_y^2\right)^2 \ll 1
\end{aligned} \tag{4.3.19}$$

To give an estimation of this term a spherical wave with the half aperture angle φ is regarded. Then, the maximum spatial frequency of the spherical wave will be $n\sin\varphi/\lambda$. The function $\tilde{u}_0$ will be considerably different from zero only for spatial frequencies with $\nu_x^2+\nu_y^2 < n^2\sin^2\varphi/\lambda^2$. Therefore, the condition (4.3.19) can be transformed to:

$$Q_{Fresnel,Fourier} = \frac{nz_0}{4\lambda}\sin^4\varphi \ll 1 \tag{4.3.20}$$

It is obvious that the error term $Q_{Fresnel,Fourier}$ increases with increasing distance z_0 whereas the error term $Q_{Fresnel}$ of equation (4.3.11) decreases with increasing z_0.

So, the formulation of the Fresnel diffraction integral in the Fourier domain (4.3.17) is better used for the near field whereas the Fresnel diffraction integral of equation (4.3.10) is better used for the medium or far field. So, both formulations are contrary. The numerical evaluation of both diffraction integrals can be done using FFT's whereas only one FFT is necessary in the formulation of equation (4.3.10) and two FFT's (one for calculating $\tilde{u}_0$ and one for the integral itself) in the formulation of equation (4.3.17).

4.3.2 The Fraunhofer diffraction formula

An approximation of the Fresnel–Kirchhoff diffraction integral for the far field can be obtained from equations (4.3.4) and (4.3.5). First, we define the direction cosines α and β as

$$\alpha := \frac{x}{l_0}\,;\ \beta := \frac{y}{l_0} \tag{4.3.21}$$

so that equation (4.3.4) can be written as:

$$l \approx l_0 + \frac{x'^2 + y'^2}{2l_0} - \left(\alpha x' + \beta y'\right) - \frac{\left(\alpha x' + \beta y'\right)^2}{2l_0} \tag{4.3.22}$$

With increasing l_0 the second and fourth term decrease more and more and only the first and the third term remain. The condition that only these two terms have to be considered is that the contribution of the other terms does not remarkably vary the exponential factor in equation (4.3.5). This is fulfilled if:

$$\pi n \frac{x'^2 + y'^2}{\lambda l_0} \ll \pi \Rightarrow Q_{Fraunhofer} := \frac{nD^2}{4\lambda z_0} = F \ll 1 \tag{4.3.23}$$

In the last step it is used that all points in the aperture A fulfill the condition $x'^2 + y'^2 \leq (D/2)^2$, whereby D is again the maximum diameter of the aperture. We also used that according to equation (4.3.2) $l_0 \geq z_0$. So, the condition is that the Fresnel number F is much smaller than one.
Then, equation (4.3.5) can be written as:

$$u(\alpha, \beta, z_0) = -i \frac{n}{\lambda l_0} \frac{z_0}{l_0} e^{ikl_0} \iint\limits_A u_0\left(x', y', 0\right) e^{-2\pi i \frac{n}{\lambda}\left(\alpha x' + \beta y'\right)} dx' dy' \tag{4.3.24}$$

This is the well–known **Fraunhofer diffraction integral**. It means that the complex amplitude in the far field is the Fourier transform of the complex amplitude at $z = 0$.
The importance of equation (4.3.24) would be quite marginal if it is only valid for the far field. This can be seen by the following example. We assume $n = 1$, λ=0.5 μm, D=10 mm. Then equation (4.3.23) would require:

$$z_0 \gg \frac{nD^2}{4\lambda} = 50 \text{ m} \tag{4.3.25}$$

But, there is another quite important case: The complex amplitude in the focal plane of a lens.

4.3.3 The complex amplitude in the focal plane of a lens

We assume a complex amplitude u_0 in the starting plane which is defined as different from zero in the aperture A and zero outside of the aperture. The influence of an ideal thin lens which is positioned in the starting plane would be that u_0 has to be multiplied by the transmission function $t_{lens,ideal}$ of the lens which is an exponential phase factor of the form:

$$t_{lens,ideal}\left(x',y'\right) = e^{-ik\left(f'\sqrt{1+\frac{x'^2+y'^2}{f'^2}}-f'\right)} =: e^{-ikl_{lens}} \tag{4.3.26}$$

Here, f' is the focal length of the lens and a positive value f' corresponds to a positive lens, whereas in our case we will only have a positive lens. Of course, an ideal lens does not exist in reality and a more adapted transmission function is

$$t_{lens}\left(x',y'\right) = e^{-ikl_{lens}+iW\left(x',y'\right)} \tag{4.3.27}$$

where W are the wave aberrations of the lens. However, in reality the wave aberrations of a lens will depend on the incident wave front and a wave–optical simulation of a lens including aberrations is not so easy. But here, we assume that W is known for a given complex amplitude u_0. For an ideal lens we just have to put W equal to zero.
So, the new complex amplitude u_0' behind the lens is defined by

$$u_0'\left(x',y',0\right) = u_0\left(x',y',0\right)t_{lens}\left(x',y'\right) \tag{4.3.28}$$

and the complex amplitude in a parallel plane at the distance z_0 is according to the Fresnel–Kirchhoff diffraction integral (see equation (4.2.7)):

$$u(x,y,z_0) = -i\frac{n}{\lambda}\iint_A u_0'\left(x',y',0\right)\frac{z_0}{l}\frac{e^{ikl}}{l}dx'dy' \tag{4.3.29}$$

with l defined by equation (4.2.3)

$$l = \sqrt{\left(x-x'\right)^2+\left(y-y'\right)^2+z_0^2}$$

Now, we are only interested in points in the neighborhood of the Gaussian focus of the lens at $(0,0,f')$ and since the radius $D/2$ of the aperture shall be several times smaller than the focal length f' of the lens we have the following conditions and approximations:

$$\begin{aligned}
&f' = z_0\left(1+\epsilon\right) \quad \text{with} \quad |\epsilon| \ll 1\\
&f' \gg \frac{D}{2} \gg \frac{\lambda}{n}\\
&x'^2+y'^2 \le \frac{D^2}{4} \ll z_0^2\\
&\frac{x}{z_0} \ll \frac{D/2}{z_0} \ll 1 \quad \text{and} \quad \frac{y}{z_0} \ll \frac{D/2}{z_0} \ll 1
\end{aligned}$$

Then, similar to the case of the Fresnel diffraction integral the cosine obliquity factor z_0/l is one and the distance l in the denominator of equation (4.3.29) can be replaced by z_0. Only l in

the exponential phase factor of equation (4.3.29) has to be considered carefully since the phase factor will rapidly oscillate if l varies by more than one wavelength λ/n. So, using equations (4.3.27), (4.3.29) and the approximations the intermediate result is:

$$u(x,y,z_0) = -i\frac{n}{\lambda z_0}\iint\limits_A u_0\left(x',y',0\right)e^{iW\left(x',y'\right)}e^{ik\left[l-l_{lens}\right]}dx'dy' \tag{4.3.30}$$

The term $l - l_{lens}$ has to be evaluated:

$$\begin{aligned} l - l_{lens} &= z_0\sqrt{1+\frac{x^2+y^2+x'^2+y'^2-2\left(xx'+yy'\right)}{z_0^2}} - \\ &\quad -f'\sqrt{1+\frac{x'^2+y'^2}{f'^2}} + f' \approx \\ &\approx z_0 + \frac{x^2+y^2}{2z_0} + \frac{x'^2+y'^2}{2z_0} - \frac{xx'+yy'}{z_0} - \frac{x'^2+y'^2}{2f'} = \\ &= z_0 + \frac{x^2+y^2}{2z_0} + \frac{\Delta z}{f'z_0}\,\frac{x'^2+y'^2}{2} - \frac{xx'+yy'}{z_0} \end{aligned} \tag{4.3.31}$$

with

$$\Delta z := f' - z_0$$

As in the case of the Fresnel diffraction integral the terms of higher than second order in x, y, x' or y' have been neglected because of our conditions. Nevertheless, it should be pointed out that the restrictions on the sine of the half aperture angle φ of the lens ($\sin\varphi \approx D/(2f')$) are not so severe as in the case of the Fresnel diffraction integral since only points in the neighborhood of the Gaussian focus, i.e. x, y and Δz are small, are interesting. To make this clear the higher order terms have to be estimated whereby only one section along x and x' is considered. But this is no restriction if the aperture is circular or if the section along x, x' has a larger diameter than along y, y'. The maximum value of x' is $D/2$ and the maximum interesting value of x is just some wavelengths if we are near the focal plane of the lens. Additionally, we can replace the term $D/(2z_0)$ with a good approximation by $\sin\varphi$ because of $f' \approx z_0$. Because of the same reason we also have $1/z_0^3 - 1/f'^3 \approx 3\Delta z/f'^4$[1]. The fourth order terms of $k(l - l_{lens})$ are therefore of the following form:

$$\begin{aligned} &k\left(\frac{\left(x^2+x'^2-2xx'\right)^2}{8z_0^3} - \frac{x'^4}{8f'^3}\right) = \\ &= 2\pi n\left(\frac{x^4}{8\lambda z_0^3} + \left(\frac{1}{8\lambda z_0^3} - \frac{1}{8\lambda f'^3}\right)x'^4 + \frac{3x^2x'^2}{4\lambda z_0^3} - \frac{x^3x'}{2\lambda z_0^3} - \frac{xx'^3}{2\lambda z_0^3}\right) \leq \\ &\leq \pi n\left(\frac{x_{max}^4}{4\lambda f'^3} + \frac{3\Delta z}{4\lambda}\sin^4\varphi + \frac{3x_{max}^2\sin^2\varphi}{2\lambda f'} - \frac{x_{max}^3\sin\varphi}{\lambda f'^2} - \frac{x_{max}\sin^3\varphi}{\lambda}\right) \end{aligned} \tag{4.3.32}$$

Again, these terms have to be much smaller than π to be negligible. A numerical example can illustrate this: A lens with a focal length of f'=100 mm and $D/2$=30 mm, i.e. $\sin\varphi$=0.29, is

[1] It is $f'^3 = z_0^3(1+\epsilon)^3 \approx z_0^3 + 3\epsilon z_0^3$. Therefore: $1/z_0^3 - 1/f'^3 \approx (1-1/(1+3\epsilon))/z_0^3 \approx (1-1+3\epsilon)/z_0^3 \approx 3\Delta z/f'^4$

illuminated with light of the wavelength λ=0.5μm. The refractive index for the light propagation is n=1. The propagation distance behind the lens is z_0=99.9 mm and the maximum value of x, which is interesting for us, is $x_{max} = 10\mu$m. Later (see equation (4.3.51)) we will see that the radius of the diffraction limited Airy disc of a lens with sinφ=0.29 and $\lambda = 0.5\mu$m is $\rho_0 = 0.61\lambda/NA = 0.61\lambda/(n\sin\varphi) \approx 1\mu$m. Therefore, an area with radius $x_{max} = 10\mu$m contains all interesting structures of the intensity distribution of the focus. Using these values, the higher order terms are:

$$\begin{aligned}
\pi n \frac{x_{max}^4}{4\lambda f'^3} &= 5 \cdot 10^{-12}\pi \\
\pi n \frac{3\Delta z}{4\lambda} \sin^4\varphi &= 1.1\,\pi \\
\pi n \frac{3x_{max}^2 \sin^2\varphi}{2\lambda f'} &= 2.5 \cdot 10^{-4}\pi \\
\pi n \frac{x_{max}^3 \sin\varphi}{\lambda f'^2} &= 6 \cdot 10^{-8}\pi \\
\pi n \frac{x_{max} \sin^3\varphi}{\lambda} &= 0.49\,\pi
\end{aligned}$$

We see that the "defocus" term of higher order is $1.1\,\pi$ and cannot be neglected. But, if we go directly in the focal plane, i.e. $z_0 = f'$, this term will completely vanish. The second term which cannot be neglected is the last term ($0.49\,\pi$) which is proportional to x_{max}/λ and the third power of sinφ. But, in the direct neighborhood of the airy disc which has in the diffraction limited case a radius of about 1μm this term will be a factor 10 smaller. This means that for sinφ of 0.3 (and the given other parameters) the higher order terms can only be neglected in the direct neighborhood of the focus and the complex amplitude calculated outside may have some errors. But, by decreasing the numerical aperture of the lens the accuracy of the calculation increases.

Moreover, if we would have performed the calculations not starting from a plane but on a sphere with radius f' around the ideal focal point, the so called focal sphere, most of the "critical" higher order terms will vanish also for quite high numerical apertures. But, for the demonstration of the principle we selected here the approximation of a thin flat lens and starting in the plane of the lens.

So, finally we have the following result for the complex amplitude in the neighborhood of the Gaussian focus by neglecting higher order terms:

$$\begin{aligned}
u(x,y,z_0) = -i\frac{n}{\lambda z_0}\, e^{ikz_0} e^{ik\frac{x^2+y^2}{2z_0}} \cdot \\
\cdot \iint\limits_A u_0(x',y',0)\, e^{iW(x',y')} e^{ik\frac{\Delta z}{z_0 f'}\frac{x'^2+y'^2}{2}} e^{-ik\frac{xx'+yy'}{z_0}} dx'dy'
\end{aligned} \tag{4.3.33}$$

This integral is similar to the **Debye integral**. In [1] (pp. 435–449) a little different but in fact nearly identical form of this integral is evaluated for an ideal lens, i.e. $W = 0$, using

the Lommel functions[2]. For us it is especially interesting that equation (4.3.33) expresses the complex amplitude in the neighborhood of the Gaussian focus as a Fourier transformation of the pupil function G:

$$G\left(x',y'\right)=\begin{cases} u_0(x',y',0)\exp(iW(x',y'))\exp\left(ik\frac{\Delta z}{f'z_0}\frac{x'^2+y'^2}{2}\right) & \text{if } (x',y')\in A \\ 0 & \text{if } (x',y')\notin A \end{cases} \tag{4.3.34}$$

Here, u_0 is the complex amplitude of the incident wave, the term $\exp(iW)$ describes the influence of the wave aberrations of the lens and the third term is a defocus term. In the focal plane itself the defocus term vanishes since then $z_0 = f' \Rightarrow \Delta z = 0$. So, in fact we see that in the focal plane of an ideal lens (i.e. $W = 0$ and $\Delta z = 0$) we have again a kind of Fraunhofer diffraction and the complex amplitude u is calculated by a Fourier transformation of u_0.
The intensity distribution I in the focal plane for an incident plane wave which is on–axis, i.e. $u_0(x', y', 0) = a$=constant with $I_0 = a^2$, is according to equation (4.3.33):

$$I(x,y,z_0=f')=I_0\frac{n^2}{\lambda^2 f'^2}\left|\iint\limits_A e^{iW\left(x',y'\right)}e^{-2\pi in\frac{xx'+yy'}{\lambda f'}}dx'dy'\right|^2 \tag{4.3.35}$$

whereby I_0 is the intensity of the incident plane wave. The intensity distribution in the focal plane is also called the **point spread function (PSF)** of the lens. It is often usual to normalize the PSF by dividing it by the intensity I_F of an non–aberrated lens of the same type at the Gaussian focus. Using equation (4.3.35) I_F is obtained by setting $W = 0$ and $(x, y) = (0, 0)$:

$$I_F(0,0,f')=I_0\frac{n^2}{\lambda^2 f'^2}\left|\iint\limits_A dx'dy'\right|^2=I_0\frac{n^2}{\lambda^2 f'^2}S^2 \tag{4.3.36}$$

Here, $S = \iint_A dx'dy'$ is the surface area of the aperture A. Then, the normalized point spread function PSF is:

$$PSF(x,y)=\frac{I(x,y,f')}{I_F(0,0,f')}=\frac{1}{S^2}\left|\iint\limits_A e^{iW\left(x',y'\right)}e^{-2\pi in\frac{xx'+yy'}{\lambda f'}}dx'dy'\right|^2 \tag{4.3.37}$$

The dimensionless number $\sigma = PSF(0, 0)$ of the aberrated lens at the Gaussian focus is called the **Strehl ratio** of the lens and is defined by:

$$\sigma=PSF(0,0)=\frac{1}{S^2}\left|\iint\limits_A e^{iW\left(x',y'\right)}dx'dy'\right|^2 \tag{4.3.38}$$

In this section, we calculated the PSF and the Strehl ratio only for a thin lens with the aperture in the plane of the lens. For general optical systems the same concept and the same equations are used, but then the aperture A is the **exit pupil** of the optical system and f' is the distance of the Gaussian focus from the exit pupil. Moreover, for an optical system, which fulfills the sine condition, the PSF can be calculated with the given equations also for quite high numerical apertures. The only restriction is that in the case of very high numerical apertures polarization effects and the so called apodization factor have to be taken into account (see section 4.5).

[2]E. Lommel invented these functions when he was a professor of physics at the University of Erlangen in 1868–1886.

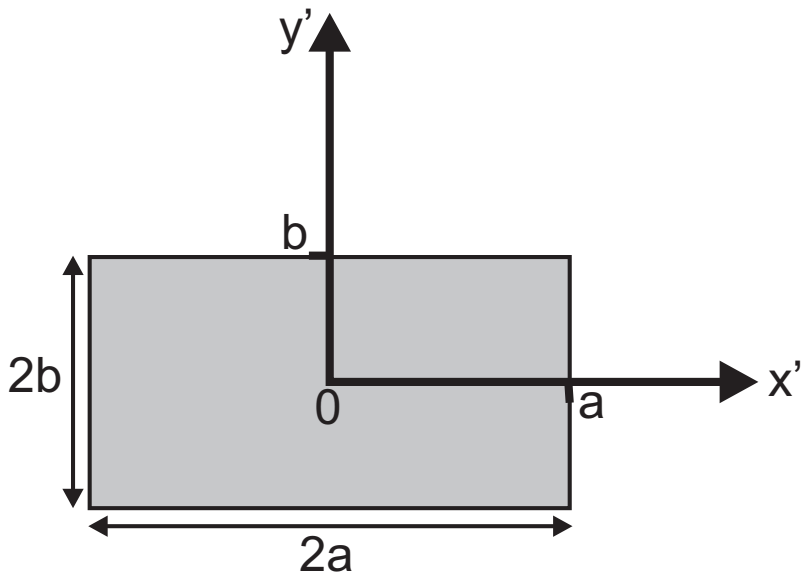


Figure 4.3: Parameters of a rectangular transparent aperture in an opaque screen.

4.3.4 Two examples for Fraunhofer diffraction

Two simple examples which can be solved analytically are the Fraunhofer diffraction at a rectangular aperture or at a circular aperture.

4.3.4.1 Fraunhofer diffraction at a rectangular aperture

The intensity distribution in the focal plane of an ideal lens (focal length f') with a rectangular aperture of the diameter $2a$ in x–direction and $2b$ in y–direction has to be calculated (see fig. 4.3). The wavelength of the light is λ and the refractive index of the material in which the wave propagates is n. The lens itself is illuminated with a uniform plane on–axis wave with the intensity I_0. Then, according to equation (4.3.35) the intensity in the focal plane of the ideal lens (i.e. $W = 0$) is:

$$\begin{aligned} I(x, y, z_0 = f') &= I_0 \frac{n^2}{\lambda^2 f'^2} \left| \iint\limits_A e^{-2\pi i n \frac{xx' + yy'}{\lambda f'}} dx' dy' \right|^2 = \\ &= I_0 \frac{n^2}{\lambda^2 f'^2} \left| \int_{-a}^{a} e^{-2\pi i n \frac{xx'}{\lambda f'}} dx' \right|^2 \left| \int_{-b}^{b} e^{-2\pi i n \frac{yy'}{\lambda f'}} dy' \right|^2 \end{aligned} \tag{4.3.39}$$

The first integral is:

$$\int_{-a}^{a} e^{-2\pi i n \frac{xx'}{\lambda f'}} dx' = \left[\frac{-\lambda f'}{2\pi i n x} e^{-2\pi i n \frac{xx'}{\lambda f'}} \right]_{x'=-a}^{x'=a} = \frac{2\lambda f'}{2\pi n x} \sin\left(2\pi n \frac{xa}{\lambda f'} \right) \tag{4.3.40}$$

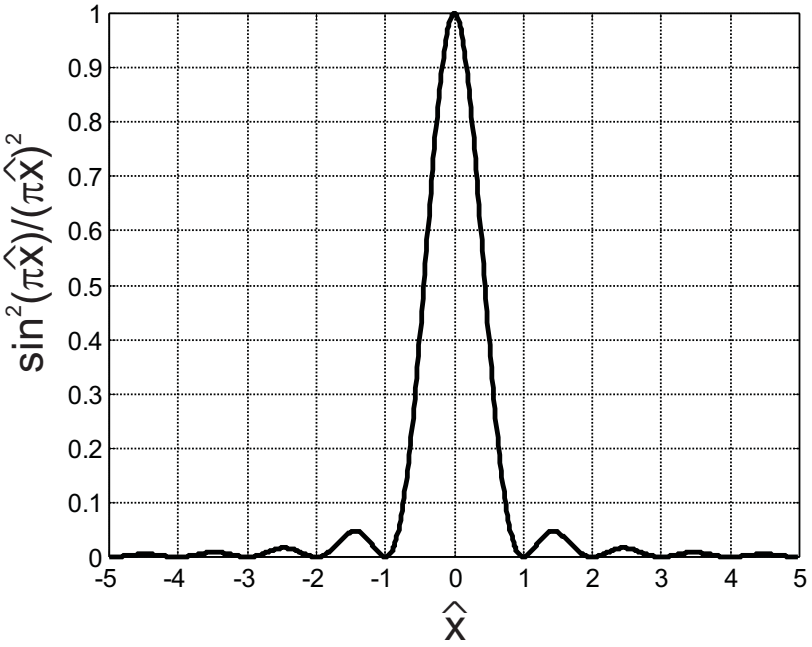


Figure 4.4: Normalized intensity distribution along the x–axis in the focal plane of a lens for a rectangular aperture. Shown is the function $(\sin(\pi\hat{x})/(\pi\hat{x}))^2$. The function $(\sin(\pi\hat{y})/(\pi\hat{y}))^2$ along the y–axis is identical.

The second integral is analogous and so the intensity in the focal plane is:

$$I(x,y,z_0=f') = I_0\left(\frac{n\,2a\,2b}{\lambda f'}\right)^2\left(\frac{\sin\left(\pi\frac{nx\,2a}{\lambda f'}\right)}{\pi\frac{nx\,2a}{\lambda f'}}\right)^2\left(\frac{\sin\left(\pi\frac{ny\,2b}{\lambda f'}\right)}{\pi\frac{ny\,2b}{\lambda f'}}\right)^2 \tag{4.3.41}$$

By introducing the variables $\hat{x} = nx\,2a/(\lambda f')$ and $\hat{y} = ny\,2b/(\lambda f')$, which are pure numbers without a physical unit, the normalized intensity distribution along one of the axes (along x– or y–axis) can be easily calculated and is shown in fig. 4.4. The minima of the intensity distribution along the x–axis are at:

$$\hat{x} = m \text{ with } m = 1,\,2,\,3\,\ldots \;\Rightarrow\; x = m\frac{\lambda f'}{2na} \approx m\frac{\lambda}{2NA_x} \tag{4.3.42}$$

Hereby, the numerical aperture $NA_x := n\sin\varphi \approx na/f'$ with the half aperture angle $\varphi \ll 1$ of the lens in x–direction has been used.

4.3.4.2 Fraunhofer diffraction at a circular aperture

The intensity distribution in the focal plane of an ideal lens (focal length f') with a circular aperture of the radius a can also be calculated using equation (4.3.35) (see fig. 4.5). The wavelength of the light is again λ and the refractive index of the material in which the wave propagates is n. The lens itself is again illuminated with a uniform plane on–axis wave with the

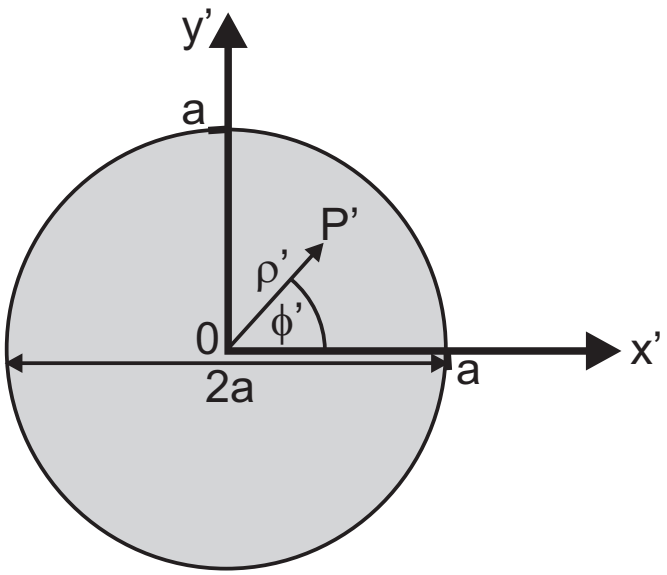


Figure 4.5: Parameters of a circular transparent aperture in an opaque screen.

intensity I_0. Because of the circular symmetry it is useful to introduce polar coordinates:

$$\begin{array}{ll} x' = \rho' \cos\phi'; & x = \rho\cos\phi \\ y' = \rho' \sin\phi'; & y = \rho\sin\phi \end{array} \quad \Rightarrow \; xx' + yy' = \rho\rho' \cos\left(\phi' - \phi\right) \tag{4.3.43}$$

Then, according to equation (4.3.35) the intensity in the focal plane of the ideal lens (i.e. $W = 0$) is written in polar coordinates:

$$\begin{aligned} I(\rho, \phi, z_0 = f') &= I_0 \frac{n^2}{\lambda^2 f'^2} \left| \iint\limits_A e^{-2\pi i n \frac{xx' + yy'}{\lambda f'}} dx'dy' \right|^2 = \\ &= I_0 \frac{n^2}{\lambda^2 f'^2} \left| \int\limits_0^a \int\limits_0^{2\pi} e^{-2\pi i n \frac{\rho\rho' \cos(\phi' - \phi)}{\lambda f'}} \rho' d\rho' d\phi' \right|^2 = \\ &= I_0 \frac{n^2}{\lambda^2 f'^2} \left| \int\limits_0^a \int\limits_0^{2\pi} e^{-2\pi i n \frac{\rho\rho' \cos\phi'}{\lambda f'}} \rho' d\rho' d\phi' \right|^2 \end{aligned} \tag{4.3.44}$$

To solve the double integral the well–known Bessel functions [59] of the first kind $J_m(x)$ are introduced by the integral representation:

$$J_m(x) = \frac{i^{-m}}{2\pi} \int\limits_0^{2\pi} e^{ix\cos\alpha} e^{im\alpha} d\alpha \tag{4.3.45}$$

For $m = 0$ we obtain:

$$J_0(x) = \frac{1}{2\pi} \int\limits_0^{2\pi} e^{ix\cos\alpha} d\alpha \tag{4.3.46}$$

Therefore, the intensity is

$$I(\rho,\phi,z_0=f') = I_0\left(\frac{2\pi n}{\lambda f'}\right)^2 \left|\int_0^a J_0\!\left(-2\pi n\frac{\rho\rho'}{\lambda f'}\right)\rho' d\rho'\right|^2 \tag{4.3.47}$$

There is another integral relation which connects the two Bessel functions J_0 and J_1:

$$xJ_1(x) = \int_0^x x' J_0(x')dx' \tag{4.3.48}$$

By substituting $x' = -2\pi n\rho\rho'/(\lambda f')$ and $dx' = -2\pi n\rho/(\lambda f')\,d\rho'$ we obtain:

$$\begin{aligned} I(\rho,\phi,z_0=f') &= I_0\left(\frac{2\pi n}{\lambda f'}\right)^2 \left|\frac{\lambda^2 f'^2}{4\pi^2 n^2\rho^2}\int_0^{-2\pi n\frac{\rho a}{\lambda f'}} x' J_0\big(x'\big)\,dx'\right|^2 = \\ &= I_0\left(\frac{2\pi n}{\lambda f'}\right)^2 \left[-\frac{\lambda f' a}{2\pi n\rho}J_1\!\left(-2\pi n\frac{\rho a}{\lambda f'}\right)\right]^2 = \\ &= I_0\left(\frac{2\pi n a^2}{\lambda f'}\right)^2 \left[\frac{J_1\!\left(2\pi n\frac{\rho a}{\lambda f'}\right)}{2\pi n\frac{\rho a}{\lambda f'}}\right]^2 \end{aligned} \tag{4.3.49}$$

Here, the symmetry of the Bessel function $J_1(-x) = J_1(x)$ has been used. By defining the variable $\hat{\rho} := 2n\rho a/(\lambda f')$, which is a pure number without a physical unit, the intensity in the focal plane of the ideal lens can be written as:

$$I(\hat{\rho},z_0=f') = I_0\left(\frac{n\pi a^2}{\lambda f'}\right)^2 \left[2\frac{J_1(\pi\hat{\rho})}{\pi\hat{\rho}}\right]^2 \tag{4.3.50}$$

The function $\left(2J_1(\pi\hat{\rho})/(\pi\hat{\rho})\right)^2$ is shown in fig. 4.6. The first minimum is at the value $\hat{\rho}_0 = 1.22$, i.e. at the radius ρ_0 with

$$\rho_0 = 1.22\frac{\lambda f'}{2na} = 0.61\frac{\lambda f'}{na} = 0.61\frac{\lambda}{NA} \tag{4.3.51}$$

Here, again the numerical aperture of the lens $NA \approx na/f'$ is used. The area inside the first minimum of the diffraction limited focus is called the **Airy disc**. It is also interesting to compare the maximum intensity $I(\hat{\rho}=0,z_0=f')$ in the central peak of the focus with the intensity I_0 of the incident plane wave. The ratio is:

$$\frac{I(\hat{\rho}=0,z_0=f')}{I_0} = \left(\frac{n\pi a^2}{\lambda f'}\right)^2 = \pi^2 F^2 \tag{4.3.52}$$

For a lens with a focal lens of f'=10 mm, an aperture radius of a=1 mm and a wavelength of λ=0.5 μm and n=1 we obtain e.g. $I(\hat{\rho}=0,z_0=f')/I_0 = (200\pi)^2 \approx 4\cdot 10^5$. The quantity $na^2/(\lambda f')$ is also known as the **Fresnel number** F of a lens which is the number of Fresnel zones of the lens in the paraxial case. This can be easily seen by calculating the optical path

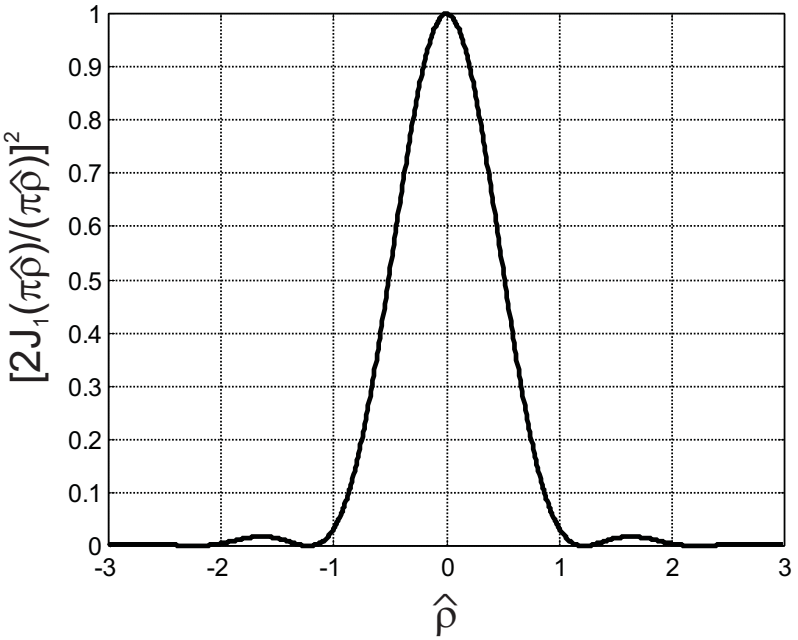


Figure 4.6: Normalized intensity distribution in the focal plane of a lens with circular aperture. Shown is the function $(2J_1(\pi\hat{\rho})/(\pi\hat{\rho}))^2$.

difference OPD between a ray from the center of the lens to the focus and a ray from the rim of the lens to the focus. It is:

$$OPD = n\left(\sqrt{a^2 + f'^2} - f'\right) = n\left(f'\sqrt{1 + \frac{a^2}{f'^2}} - f'\right) \approx \frac{na^2}{2f'} = F\frac{\lambda}{2}$$

$$\Rightarrow\ F = \frac{na^2}{\lambda f'} \tag{4.3.53}$$

This means that the intensity in the central peak of the focus of an ideal lens with circular aperture is proportional to the square of the Fresnel number of this lens.

4.4 Numerical implementation of some diffraction methods

Many of the proposed diffraction integrals can be solved by performing one or two Fourier transformations (see table 4.1). For a numerical implementation a discrete Fourier transformation is necessary and to increase the speed of the calculation it makes of course sense to take a **Fast Fourier transformation (FFT)** [60]. Of course, it has to be noticed that the sampling theorem is fulfilled and that the size of the field is large enough. In practice, large field sizes of e.g. more than 2048x2048 samples need a lot of computer memory and computing time. To use an FFT the field of the complex amplitude u_0 is in each spatial direction x and y uniformly sampled at N_xxN_y points, whereby N_x and N_y are powers of two. The diameters of the field in the spatial domain are called D_x in x– and D_y in y–direction. In most cases the field will be quadratic and sampled with equal number of points, i.e. $N_x = N_y$ and $D_x = D_y$. Nevertheless, there are

Diffraction method	Spectrum of plane waves	Fresnel (Fourier domain)	Fresnel (convolution)	Fraunhofer, Debye integral
Equation	(4.1.12)	(4.3.17)	(4.3.10)	(4.3.24), (4.3.33)
Conjugated variables	$(x,y) \leftrightarrow (\nu_x,\nu_y)$	$(x,y) \leftrightarrow (\nu_x,\nu_y)$	$(x',y') \leftrightarrow \left(\frac{nx}{\lambda z_0},\frac{ny}{\lambda z_0}\right)$	$(x',y') \leftrightarrow \left(\frac{n\alpha}{\lambda},\frac{n\beta}{\lambda}\right)$, $(x',y') \leftrightarrow \left(\frac{nx}{\lambda z_0},\frac{ny}{\lambda z_0}\right)$
Number of FFTs	2	2	1	1

Table 4.1: Conjugated variables and number of FFTs for calculating the different diffraction integrals.

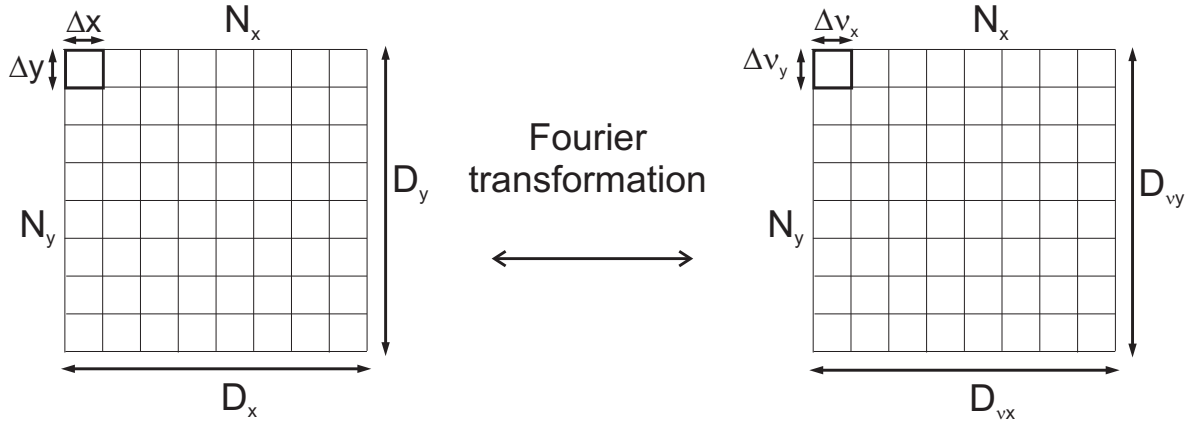


Figure 4.7: Discrete fields for solving diffraction integrals by using an FFT.

cases (e.g. systems with cylindrical or toric optical elements) where it is useful to have different sampling rate and number of sampling points along x and y. Then, the sampling interval Δx and Δy in x– and y–direction between two neighbored sampling points in the spatial domain are (see fig. 4.7):

$$\begin{aligned} \Delta x &= \frac{D_x}{N_x} \\ \Delta y &= \frac{D_y}{N_y} \end{aligned} \tag{4.4.1}$$

In some cases, it is also useful to perform only a one–dimensional simulation because it is much faster. In this case, N_y is set to one and only a one–dimensional FFT is taken.
We call the conjugated variables in the Fourier domain ν_x and ν_y, whereby these can be really the spatial frequencies like defined in equation (4.1.4) or other variables like in the Fresnel diffraction integral which just have the physical dimension of a spatial frequency. Table 4.1 shows the conjugated variables for the different diffraction integrals. The diameters of the fields in the Fourier domain are called $D_{\nu x}$ and $D_{\nu y}$ along the ν_x– and ν_y–direction. The associated sampling intervals between two neighbored sampling points in the Fourier domain are:

$$\begin{aligned} \Delta\nu_x &= \frac{D_{\nu x}}{N_x} \\ \Delta\nu_y &= \frac{D_{\nu y}}{N_y} \end{aligned} \tag{4.4.2}$$

For a discrete Fourier transformation like the FFT the product of the respective diameters in the spatial domain and in the Fourier domain is equal to the number of sampling points. Therefore,

the two relations are valid:

$$\begin{aligned} D_x D_{\nu x} &= N_x \\ D_y D_{\nu y} &= N_y \end{aligned} \tag{4.4.3}$$

By using equations (4.4.1) and (4.4.2) it is clear that for the product of the sampling intervals in the spatial and in the Fourier domain the following equations apply:

$$\begin{aligned} \Delta x \Delta \nu_x = \tfrac{1}{N_x} \quad &\Rightarrow \quad \Delta \nu_x = \tfrac{1}{D_x}; \quad \Delta x = \tfrac{1}{D_{\nu x}} \\ \Delta y \Delta \nu_y = \tfrac{1}{N_y} \quad &\Rightarrow \quad \Delta \nu_y = \tfrac{1}{D_y}; \quad \Delta y = \tfrac{1}{D_{\nu y}} \end{aligned} \tag{4.4.4}$$

The variables are in our case always symmetrical around the origin of the coordinate system. Therefore, the minimum and maximum values of the respective variables are:

$$\begin{aligned} x_{min} &= -\tfrac{D_x}{2}; & x_{max} &= \tfrac{D_x}{2} \\ y_{min} &= -\tfrac{D_y}{2}; & y_{max} &= \tfrac{D_y}{2} \\ \nu_{x,min} &= -\tfrac{N_x}{2}\Delta\nu_x = -\tfrac{N_x}{2D_x}; & \nu_{x,max} &= +\tfrac{N_x}{2}\Delta\nu_x = +\tfrac{N_x}{2D_x} \\ \nu_{y,min} &= -\tfrac{N_y}{2}\Delta\nu_y = -\tfrac{N_y}{2D_y}; & \nu_{y,max} &= +\tfrac{N_y}{2}\Delta\nu_y = +\tfrac{N_y}{2D_y} \end{aligned} \tag{4.4.5}$$

Since a discrete Fourier transformation has periodic boundary conditions, the function values at the left boundary will be equal to those at the right boundary, i.e. $u(x_{min}, y) = u(x_{max}, y)$, $u(x, y_{min}) = u(x, y_{max})$, and so on. This fact is important because it generates aliasing effects if the field size and the sampling are not correct. Additionally, because of the periodic boundary conditions only the values of the left boundary will be stored in the data field. Therefore, the coordinates of the stored sampling points range from $x_{min} = -D_x/2$ to $x_{max} - \Delta x = D_x(N_x - 2)/(2N_x)$, etc. This especially means, that the data point number $N_x/2$ is at $x = 0$ if we count the data points starting with number zero (otherwise starting with 1 the data point number $N_x/2 + 1$ is at $x = 0$). The same is valid for the other spatial and spatial frequency coordinates.
In the following some special aspects of the different diffraction integrals which can be calculated using one or two FFTs are presented.

4.4.1 Numerical implementation of the angular spectrum of plane waves or the Fresnel diffraction in the Fourier domain

To solve the diffraction integrals (4.1.12) and (4.3.17) two FFTs are necessary. The first is to transform the complex amplitude u_0 into the Fourier domain with the spatial frequencies ν_x and ν_y. In order to can represent all propagating waves the maximum spatial frequencies have to fulfill the conditions

$$\begin{aligned} \nu_{x,max} \geq \tfrac{n}{\lambda} \quad &\Rightarrow \quad N_x \geq \tfrac{2nD_x}{\lambda} \\ \nu_{y,max} \geq \tfrac{n}{\lambda} \quad &\Rightarrow \quad N_y \geq \tfrac{2nD_y}{\lambda} \end{aligned} \tag{4.4.6}$$

Here, equations (4.1.4) and (4.4.5) have been used. However, it is in general not necessary that all spatial frequencies belonging to propagating waves can be represented. Especially, in the case of the Fresnel diffraction integral (4.3.17) formulated in the Fourier domain high spatial frequencies will in most cases not be allowed because the equation is, depending on the

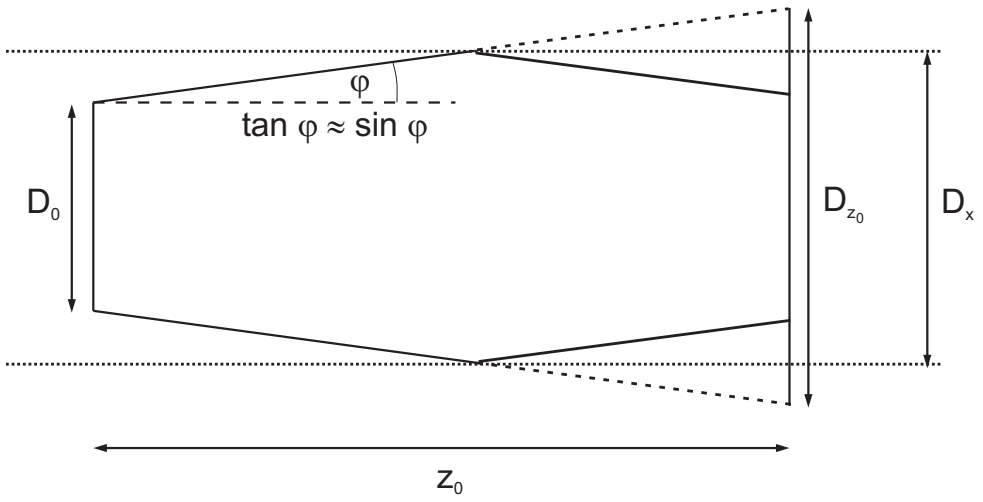


Figure 4.8: Illustration of aliasing effects. The parts of the wave which leave the field at a boundary enter the field at the opposite boundary and interfere with the other parts of the wave.

propagation distance z_0, only valid for small spatial frequencies (see equation (4.3.19)). Let us assume that the angular spectrum of plane waves has a maximum tilt angle φ for which the Fourier transform $\tilde{u}_0$ of u_0 has a function value which is noticeable different from zero. Then, it is sufficient that instead of equation (4.4.6) the following conditions are used:

$$\begin{aligned} \nu_{x,max} \geq \tfrac{n}{\lambda}\sin\varphi \quad &\Rightarrow \quad N_x \geq \tfrac{2nD_x}{\lambda}\sin\varphi \\ \nu_{y,max} \geq \tfrac{n}{\lambda}\sin\varphi \quad &\Rightarrow \quad N_y \geq \tfrac{2nD_y}{\lambda}\sin\varphi \end{aligned} \tag{4.4.7}$$

If these conditions are not fulfilled aliasing effects occur and the numerical result is wrong. It is clear that only for micro–optical elements the condition (4.4.6), i.e. $\sin\varphi$ is allowed to be 1, can be fulfilled. For a wavelength of λ=0.5 μm, refractive index n=1 and a field diameter in x–direction of D_x=1 mm we obtain e.g. $N_x \geq 2nD_x/\lambda = 4000$. For a two–dimensional FFT a field size of 4096x4096 samples is at the upper limit of modern PCs. Using double precision (i.e. 64 bit=8 Bytes per number) and keeping in mind that the field is complex (i.e. two real numbers per sampling point) we need e.g. 256 MB computer memory for 4096x4096 sampling points.
Also in the case of the second (inverse) FFT which calculates u using equations (4.1.12) or (4.3.17) aliasing effects can occur. Graphically, this means that parts of a diverging propagating wave leave the border of the field. Because of the periodic boundary conditions these parts of the waves will then enter the field at the opposite boundary. Fig. 4.8 illustrates this effect for a diverging spherical wave which has a diameter D_0 at the starting plane and would have a diameter $D_{z_0} > D_x$ in the distance z_0. If the half aperture angle of the spherical wave is φ, whereby we assume that φ is so small that $\sin\varphi \approx \tan\varphi$, the diameter D_{z_0} will be:

$$D_{z_0} \approx D_0 + 2z_0\sin\varphi \tag{4.4.8}$$

If $D_{z_0} > D_x$ aliasing effects will appear. This effect can be useful to simulate e.g. numerically the self–imaging Talbot effect for infinitely extended periodic structures [61]. However, in most cases aliasing effects are disturbing and have to be avoided. For converging waves aliasing effects will not appear as long as the propagation distance is not so large that the wave passes the focus

and becomes diverging. In practice, the limitations on the number of samples will limit the application of the propagation of a wave using the angular spectrum of plane waves. It is also very interesting to note that the field diameters D_x and D_y normally do not change between the two planes at $z = 0$ and $z = z_0$ in the case of the propagation using the angular spectrum of plane waves or the Fresnel integral in the Fourier domain formulation. Only, if the field size in the Fourier domain is manipulated the field size in the spatial domain will change. This is e.g. possible by taking only each second sampling point and embedding the new field with so many zeros that the total number of sampling points remains unchanged. Then the effective diameters in the Fourier domain $D_{\nu x}$ and $D_{\nu y}$ are doubled and therefore, according to equation (4.4.3) the diameters D_x and D_y in the spatial domain will be halved. But, of course this manipulation reduces the sampling density in the Fourier domain and is therefore a kind of high–pass filtering operation suppressing the long–periodic spatial structures.

4.4.2 Numerical implementation of the Fresnel (convolution formulation) and the Fraunhofer diffraction

Using the method of the angular spectrum of plane waves guarantees that the field sizes D_x and D_y do not change during the propagation as long as no manipulations are made in the Fourier domain. The reason for this is that two FFTs are used, one "normal" and one inverse FFT. If the Fresnel diffraction integral of equation (4.3.10) or the Fraunhofer diffraction integrals (4.3.24) or (4.3.33) are used, only one FFT is made. Therefore, the field size changes because the conjugated variables are now according to table 4.1 (x', y') and $(nx/(\lambda z_0), ny/(\lambda z_0))$ (or $(n\alpha/\lambda, n\beta/\lambda)$ where $\alpha = x/z_0$ and $\beta = y/z_0$). This means, that according to equation (4.4.3) the following relations are valid if we introduce the field diameters $D_{x,0} = D_x$ and $D_{y,0} = D_y$ in the first plane and the field diameters D_{x,z_0} and D_{y,z_0} in the second plane, whereby $D_{\nu x} = nD_{x,z_0}/(\lambda z_0)$ and $D_{\nu y} = nD_{y,z_0}/(\lambda z_0)$:

$$\begin{aligned} D_x D_{\nu x} = N_x \quad &\Rightarrow \quad D_{x,z_0} = N_x \tfrac{\lambda z_0}{n D_{x,0}} \\ D_y D_{\nu y} = N_y \quad &\Rightarrow \quad D_{y,z_0} = N_y \tfrac{\lambda z_0}{n D_{y,0}} \end{aligned} \tag{4.4.9}$$

Of course, the sampling densities Δx_{z_0} and Δy_{z_0} in the second plane are then:

$$\begin{aligned} \Delta x_{z_0} &= \tfrac{\lambda z_0}{n D_{x,0}} \\ \Delta y_{z_0} &= \tfrac{\lambda z_0}{n D_{y,0}} \end{aligned} \tag{4.4.10}$$

Let us e.g. calculate the intensity distribution in the focal plane of a lens with focal length f' ($z_0 = f'$) using equation (4.3.33). The aperture of the lens shall be quadratic with diameters $2a = 2b$. Then, it is enough to just consider one dimension, e.g. the x–direction. If the field size $D_{x,0}$ would now only be $2a$ the sampling interval would be

$$\Delta x_{z_0} = \frac{\lambda f'}{n 2a} \approx \frac{\lambda}{2NA} \tag{4.4.11}$$

whereby, the numerical aperture of the lens $NA := n \sin\varphi \approx na/f'$ has been used. But, by comparing this result with equation (4.3.42) it is clear that then the sampling density is so low that the secondary maxima are not observed because the sampling is only in the minima. Therefore, it is necessary that the aperture of the lens is embedded into a field of zeros so that the effective field diameter $D_{x,0}$ is at least doubled $D_{x,0} \geq 4a$. So, by embedding the aperture

of the lens by more and more zeros the effective sampling density in the focal plane is increased. Increasing the field size by a factor m and filling the new area with zeros reduces the sampling interval by a factor m to $\Delta x_{z_0} = \lambda/(2mNA)$. In other words we can say: The total field size D_{x,z_0} in the focal plane is proportional to the number N_x/m of samples with which the lens aperture is sampled whereas the sampling density is proportional to the factor m of the zero embedding.
Of course, there is also another reason for embedding the lens aperture with zeros: the aliasing effects. A quadratic lens aperture with $D_{x,0} = 2a$ would due to the periodic boundary conditions mean that the aperture is repeated periodically without spaces in between and fills the whole space. Therefore, no diffraction at all would occur and the focus would be a delta peak as in geometrical optics.

4.5 Polarization effects in the focus of a lens

In the previous cases of diffraction we considered only the scalar case. But, in this paragraph we want to discuss the influence of polarization effects to the intensity distribution in the focal region of a lens. However, we will only discuss a simple numerical simulation method which is in fact identical to the semi–analytical vectorial Debye integral formulation discussed e.g. in [41] or [62]. Richards and Wolf [41] were one of the first who calculated that the light distribution in the focus of a lens with rotationally invariant aperture which is illuminated by a linearly polarized plane wave is rotationally variant, i.e. approximately elliptical, for a high numerical aperture of the lens.

4.5.1 Some elementary qualitative explanations

To illustrate the influence of polarization effects to the focus look at fig. 4.9. Assume a plane wave with linear polarization (electric vector in y–direction) which is focused by a lens. Then, there is one plane (x–z–plane) where the electric vectors are perpendicular to the plane of refraction of the rays (fig. 4.9a)). There, the electric vectors in the focus add like scalars and a quite large transversal component is obtained. But, there is also the y–z–plane where the electric vectors change their direction after being refracted by the lens. Then, they add in the focus like real vectors and a smaller transversal component than in the x–z–plane is obtained (fig. 4.9b)) leading to a larger diameter of the focus in y–direction (for incident linear polarization in y–direction). Especially, for very steep rays in the y–z–plane corresponding to high numerical aperture rays the electric vectors nearly cancel each other in the focus. Due to the broken symmetry of this problem the intensity distribution in the focus will also be rotationally variant. Of course, this effect is only visible for very high numerical apertures because otherwise the vector character of the electric field is not so obvious. It is especially good visible for annular apertures and a high numerical aperture lens.
But, there is a rotationally symmetric polarization pattern, called **radially polarized light** [40], where the direction of the electric vector varies locally, so that it always points away from the optical axis in radial direction. Fig. 4.10 shows in a) the electric vectors in the aperture of the lens at a certain time for linearly polarized light, i.e. they all point in the same direction. In b) the electric vectors for radially polarized light are drawn at a certain time. They all point radially away from the optical axis. Of course, in the case of radially polarized light there has to be the intensity zero on the optical axis due to the symmetry of the problem. So, the intensity

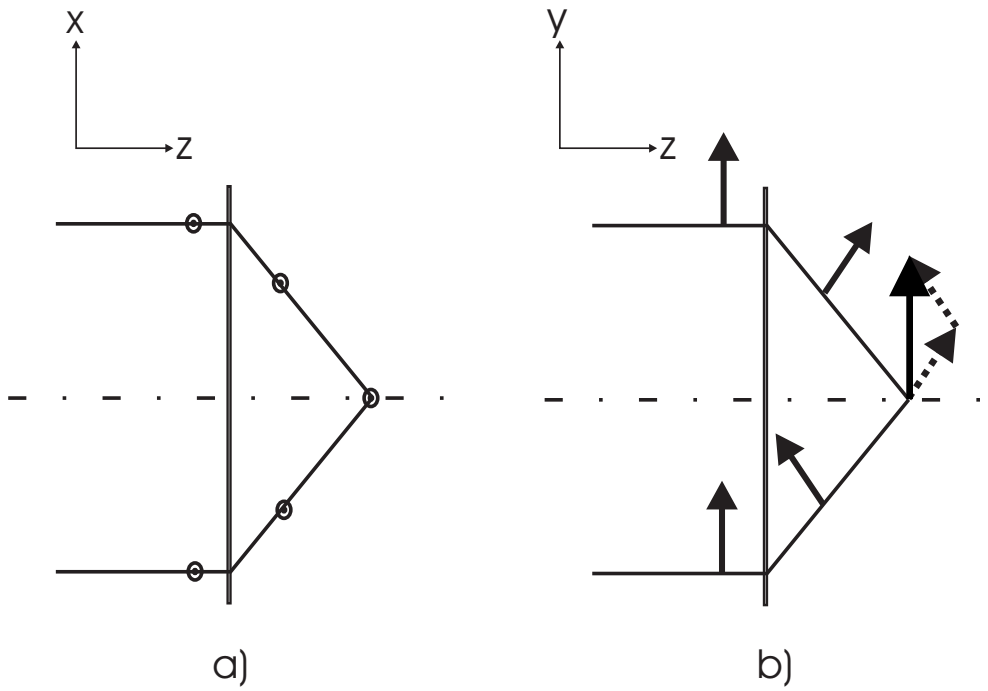


Figure 4.9: Addition of the electric vectors in the focus of a lens for linearly polarized light. a) The electric vectors are perpendicular to this x–z–plane, i.e. they add arithmetically like scalars; b) the electric vectors are lying in the y–z–plane and add vectorially to a transversal component.

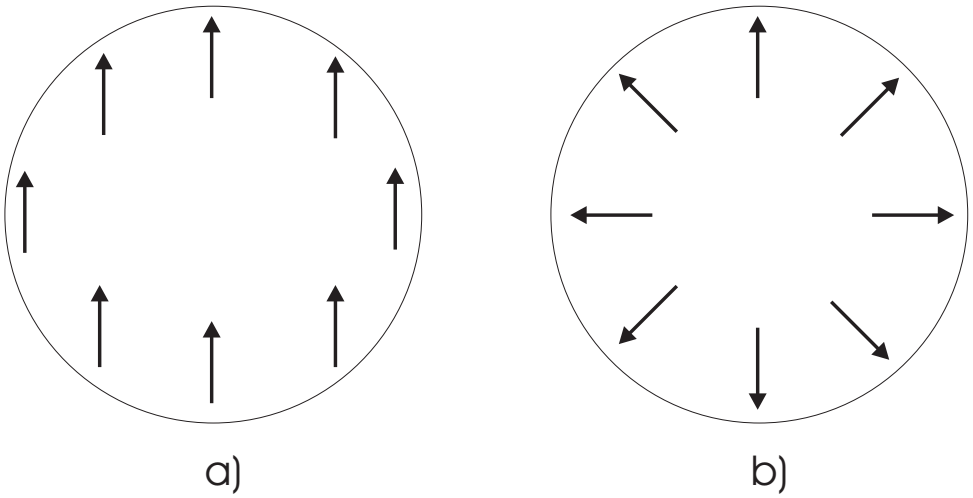


Figure 4.10: Local polarization vectors in the aperture in front of the lens for a) linearly polarized light and b) a radially polarized doughnut mode.

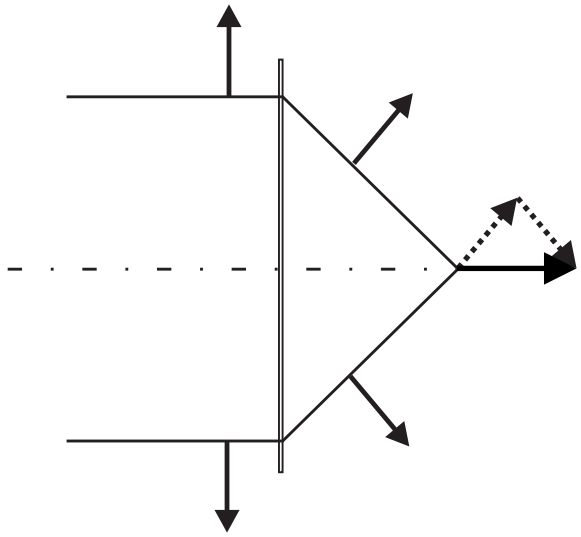


Figure 4.11: Orientation of the electric vectors in front of and behind a lens for the radially polarized doughnut mode. The electric vectors add to a longitudinal component parallel to the optical axis.

distribution in the aperture of the lens cannot be homogeneous in the case of radially polarized light and the amplitude distribution is often a so called doughnut mode. This radially polarized doughnut mode is in fact the superposition of a Hermite–Gaussian TEM_{10} and a TEM_{01} mode (see fig. 6.3 at page 182) with relative phase difference zero and perpendicular polarization. Then, the time–independent generally complex–valued electric vector $\hat{\boldsymbol{E}}_{rad}$ before the lens has the value

$$\hat{\boldsymbol{E}}_{rad}\left(x,y,z=0\right) = E_0 \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} e^{-\left(x^2+y^2\right)/w_0^2} \tag{4.5.1}$$

Hereby, w_0 is the beam waist of the Gaussian function and E_0 is a constant. The maximum of the amplitude of the electric vector is at $\sqrt{x^2+y^2} = w_0/\sqrt{2}$ as can be easily calculated.
But now, for the radially polarized doughnut mode, in all planes containing the optical axis the electric vectors are oriented like in fig. 4.11. Then, the electric vectors add in the focus to a longitudinal component and this is the case for all planes containing the optical axis. Therefore, the focus is completely axially symmetric.

4.5.2 Numerical calculation method

To calculate the electric energy density in the focal region of a lens the vectorial Debye integral of [41] or the method of [63] can be used. Both say that the electric vector in the focal region can be written as a superposition of plane waves which propagate along the direction of the rays which run from the exit pupil of the lens to the geometrical focus. In this model diffraction effects at the rim of the aperture are neglected how it is also done in the scalar formulation of the Debye integral. But, this is a quite good approximation as long as the diameter of the aperture of the lens is large compared to the wavelength ($2r_{aperture} \gg \lambda$) and as long as the numerical aperture of the lens is sufficiently high (what is here always the case because polarization effects are only interesting for high numerical apertures).

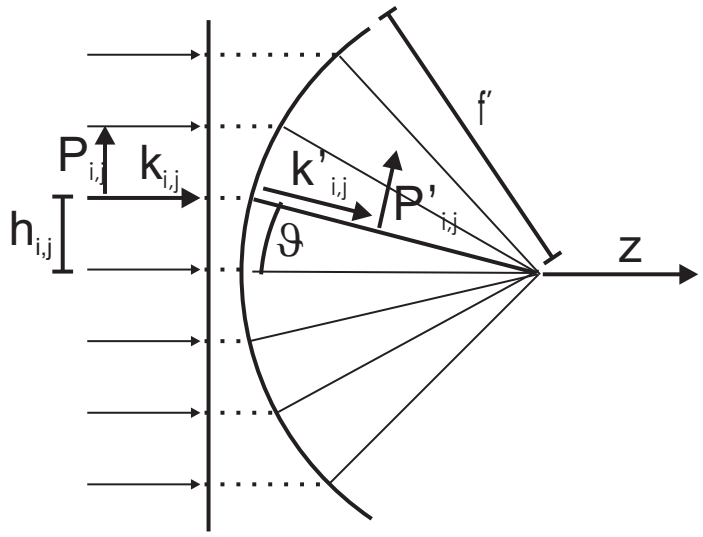


Figure 4.12: Principal scheme of the distribution of rays used to calculate the electric energy density in the focus. The lens has to fulfill the sine condition so that the ray height $h = \sqrt{x^2 + y^2}$ is $h = f' \sin\vartheta$.

For the numerical calculation we make a uniform sampling of rays in the planar entrance pupil of the lens, e.g. NxN rays in an orthogonal pattern, whereby the rays outside of the aperture (e.g. circular or annular aperture) have zero amplitude. The optical axis shall be the z–axis. Each ray number (i, j) at the coordinate (x_i, y_j, z_0) in the entrance pupil has a wave vector $\boldsymbol{k}_{i,j} = (0, 0, k) = (0, 0, 2\pi/\lambda)$ and a generally complex polarization vector $\boldsymbol{P}_{i,j} = (P_{x,i,j}, P_{y,i,j}, 0)$ which is in fact proportional to the electric vector $\hat{\boldsymbol{E}}$ at this point if the sampling is uniform. Of course, $\boldsymbol{P}_{i,j}$ is orthogonal to $\boldsymbol{k}_{i,j}$ because of the orthogonality condition of plane waves (see the paragraph on page 7). Examples for different polarization states are (see also the paragraph about the Jones calculus on page 34):

- A linearly polarized homogeneous plane wave (polarization in y–direction) with the polarization vector $\boldsymbol{P}_{i,j} = (0, P_0, 0)$ with a constant real value P_0.
- A circularly polarized homogeneous plane wave with the polarization vector $\boldsymbol{P}_{i,j} = (P_0/\sqrt{2}, iP_0/\sqrt{2}, 0)$ whereby the factor $1/\sqrt{2}$ is introduced to have $|\boldsymbol{P}_{i,j}|^2 = P_0^2$.
- The radially polarized doughnut mode with the polarization vectors according to equation (4.5.1).

The lens itself shall fulfill the sine condition so that we have for the refracted ray behind the lens

$$h = \sqrt{x^2 + y^2} = f' \sin\vartheta \quad \Rightarrow \quad \sin\vartheta = \frac{\sqrt{x^2 + y^2}}{f'} \tag{4.5.2}$$

whereby, f' is the focal length of the lens and ϑ the polar angle (see fig. 4.12). By defining the azimuthal angle φ we have the following relations for the rays in front of the lens and behind the lens:

$$\begin{aligned} x &= h\cos\varphi \\ y &= h\sin\varphi \end{aligned} \quad \Rightarrow \quad \varphi = \arctan\frac{y}{x} \tag{4.5.3}$$

and

$$\boldsymbol{k}' = k \begin{pmatrix} -\cos\varphi \sin\vartheta \\ -\sin\varphi \sin\vartheta \\ \cos\vartheta \end{pmatrix} \tag{4.5.4}$$

The polarization vector $\boldsymbol{P}$ of each ray has to be separated into a component lying in the plane of refraction and a component perpendicular to it. The unit vector $\boldsymbol{e}_{\parallel}$ along the component in the plane of refraction in front of the lens is:

$$\boldsymbol{e}_{\parallel} = \begin{pmatrix} \cos\varphi \\ \sin\varphi \\ 0 \end{pmatrix} \tag{4.5.5}$$

Behind the lens the new unit vector $\boldsymbol{e}'_{\parallel}$ along the component in the plane of refraction is:

$$\boldsymbol{e}'_{\parallel} = \begin{pmatrix} \cos\varphi \cos\vartheta \\ \sin\varphi \cos\vartheta \\ \sin\vartheta \end{pmatrix} \tag{4.5.6}$$

So, the polarization vector $\boldsymbol{P}'$ behind the lens can be calculated by keeping in mind that the component perpendicular to the plane of refraction remains unchanged and that the component in the plane of refraction keeps its amplitude, but, is now parallel to $\boldsymbol{e}'_{\parallel}$. In total this means for $\boldsymbol{P}'$:

$$\begin{aligned} \boldsymbol{P}' &= g(\vartheta) \left[\boldsymbol{P} - \left(\boldsymbol{P} \cdot \boldsymbol{e}_{\parallel}\right) \boldsymbol{e}_{\parallel} + \left(\boldsymbol{P} \cdot \boldsymbol{e}_{\parallel}\right) \boldsymbol{e}'_{\parallel} \right] = \\ &= g(\vartheta) \left[\boldsymbol{P} - \left(P_x \cos\varphi + P_y \sin\varphi\right) \begin{pmatrix} \cos\varphi\,(1-\cos\vartheta) \\ \sin\varphi\,(1-\cos\vartheta) \\ -\sin\vartheta \end{pmatrix} \right] \end{aligned} \tag{4.5.7}$$

Here, the so called apodization factor $g(\vartheta)$ is necessary in order to conserve the energy of the tilted plane wave [64]. For calculating the apodization factor, which is $g(\vartheta) = 1/\sqrt{\cos\vartheta}$ for an aplanatic lens, some further considerations are necessary.

4.5.2.1 Energy conservation in the case of discrete sampling

In general, the numerically discrete sampled amplitude of a wave (i.e. the modulus of the electric vector) can be represented either by a variation of the modulus $|\boldsymbol{P}|$ of the discrete sampled polarization vector or by a variation of the density of rays/plane waves, i.e. by a variation of the inverse surface element $1/df$ associated with each ray/plane wave:

$$E := \frac{|\boldsymbol{P}|}{df} \tag{4.5.8}$$

For conserving energy the light power contained in the entrance pupil in a ring with radius h and infinitesimal thickness dh has also to be contained in the exit pupil in an annular shaped segment (assuming that there is no absorption).

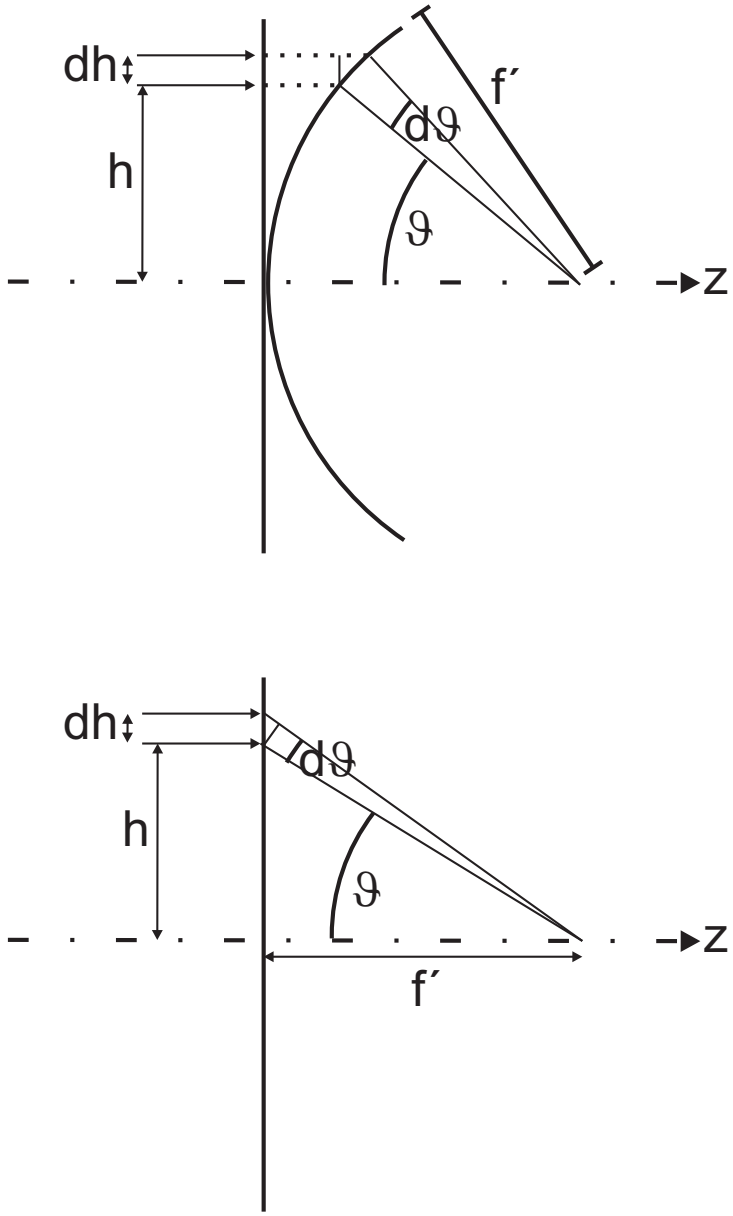


Figure 4.13: *Scheme for deriving the energy conservation conditions in the case of focusing with a high numerical aperture optical system. Top: aplanatic lens fulfilling the sine condition, bottom: ideal DOE.*

Aplanatic lens For the case of an aplanatic lens fulfilling the sine condition the exit pupil is a sphere with radius of curvature f' (see fig. 4.13 top). So, the ring in the plane entrance pupil is transformed into an annular shaped segment of a sphere in the exit pupil with polar angle ϑ and infinitesimal angular thickness $d\vartheta$. The product surface area of the ring and intensity has to be constant in both cases. Therefore, the surface areas of the plane ring dF in the entrance pupil and of the sphere segment dF' on the focal sphere (focal sphere = sphere around focus with radius of curvature identical to focal length f') and the associated intensities I (in entrance pupil) and I' (on focal sphere) are related by:

$$dF = 2\pi h\, dh \tag{4.5.9}$$

$$dF' = 2\pi f'^2 \sin\vartheta\, d\vartheta \tag{4.5.10}$$

$$I\, dF = I'\, dF' \tag{4.5.11}$$

The intensity I is proportional to the square of the amplitude E and the amplitude is given by the surface density of the modulus of the polarization vector $\boldsymbol{P}$. This means:

$$I\, dF = I'\, dF' \;\Rightarrow\; \frac{|\boldsymbol{P}|^2}{dF^2} dF = \frac{|\boldsymbol{P'}|^2}{dF'^2} dF' \;\Rightarrow \tag{4.5.12}$$

$$|\boldsymbol{P'}| = g|\boldsymbol{P}| = \sqrt{\frac{dF'}{dF}}|\boldsymbol{P}| \;\Rightarrow\; g = \sqrt{\frac{dF'}{dF}} \tag{4.5.13}$$

The factor g which is called apodization factor depends on the actual focusing optical element. Additionally, the apodization factor using our definition is not identical to the apodization factor given in most papers of the literature which is directly referred to the electric field. In our case the discrete sampled polarization vectors are considered and the electric field is represented by their surface density. So, the apodization factor describes the change of the modulus of the polarization vector on the focal sphere taking into account energy conservation. In contrary, in the literature the apodization factor describes the change of the electric field strength between plane entrance pupil and focal sphere. Therefore, the apodization factor in literature is the reciprocal value of our apodization factor:

$$I\, dF = I'\, dF' \;\Rightarrow\; E^2 dF = E'^2 dF' \;\Rightarrow\; g_{literature} = \frac{E'}{E} = \sqrt{\frac{dF}{dF'}} = \frac{1}{g} \tag{4.5.14}$$

The apodization factor g can be easily calculated for an aplanatic lens, which means a lens fulfilling the sine condition:

$$h = f' \sin\vartheta \tag{4.5.15}$$

This means:

$$dh = f' \cos\vartheta\, d\vartheta \quad \Rightarrow$$

$$g = \sqrt{\frac{dF'}{dF}} = \sqrt{\frac{f'^2 \sin\vartheta\, d\vartheta}{f'^2 \sin\vartheta \cos\vartheta\, d\vartheta}} \quad \Rightarrow$$

$$g(\vartheta) = \frac{1}{\sqrt{\cos\vartheta}} \tag{4.5.16}$$

Finally, this results in:

$$|\boldsymbol{P'}| = g|\boldsymbol{P}| = |\boldsymbol{P}|/\sqrt{\cos\vartheta} \tag{4.5.17}$$

The surface areas df and df' are here defined as the actual surface area elements represented by each ray/plane wave, whereas dF and dF' are defined as the surface areas of the rings in the plane entrance pupil and on the spherical exit pupil. However, the ratio between both is identical, i.e. $df/df' = dF/dF'$. Therefore, it is allowed to exchange both quantities in our calculations.
It has to be emphasized that the apodization factor does not depend on the polarization, but it is an effect which should also be taken into account for scalar calculations since it only relies on the concept of energy conservation. However, since scalar calculations of the point spread function are only valid with good accuracy for values of $\sin\vartheta_{max} \leq 0.5$ (with aperture angle ϑ_{max}) it can easily be estimated that for example for an aplanatic lens the error by assuming g to be constant and equal to one is only $1/\sqrt{\cos\vartheta_{max}} = 1.075$ at the rim of the aperture. Since small amplitude variations influence the intensity distribution of the PSF only quite slightly, this can be neglected in the area of validity of scalar calculations.
The apodization factor g of equation (4.5.16) is valid for an aplanatic lens. For other optical elements/systems g will of course be different. To exercise the calculation of g in other cases, we will consider in the next paragraph an idealized plane lens.

Idealized plane DOE The concept of calculating the apodization factor g shall be demonstrated as further example for an idealized diffractive optical lens (DOE). For the idealized DOE it is assumed that it is a plane lens without any aberrations having 100% diffraction efficiency independent of the polarization. Of course, this is an idealization since real DOEs with high numerical apertures, i.e. small local periods at the rim, have neither 100% or at least constant diffraction efficiency, nor is the diffraction efficiency independent of the polarization. Additionally, there is a small phase shift for small local periods of only some few wavelengths, i.e. aberrations, which also depends on the polarization. Nevertheless, we just assume in this paragraph that such an idealized plane lens/DOE exists.
Figure 4.13 (bottom) shows that for a plane lens with focal length f' a ray with height h in front of the lens results in a ray with angle ϑ behind the lens with the following equation:

$$h = \sqrt{x^2 + y^2} = f' \tan\vartheta \tag{4.5.18}$$

For a plane lens this equation just replaces the sine condition (4.5.15).
The light power in a ring with radius h and thickness dh in front of the lens has then to be transformed into a ring segment on the focal sphere with radius f'. To do this the apodization factor defined by equation (4.5.13) has to be calculated using equations (4.5.9) and (4.5.10):

$$g = \sqrt{\frac{dF'}{dF}} = \sqrt{\frac{f'^2 \sin\vartheta\, d\vartheta}{f'^2 \sin\vartheta \cos^{-3}\vartheta\, d\vartheta}} = \sqrt{\cos^3\vartheta} \tag{4.5.19}$$

For a high numerical aperture DOE this factor g has to be entered into equation (4.5.7) to obtain the correct scaling of the polarization vectors on the focal sphere. By additionally replacing the sine condition (4.5.15) by equation (4.5.18) all other equations of section 4.5.2 can be used to calculate the electric energy density in the focus of a high numerical aperture DOE.

4.5.2.2 Electric field in the focus

For an ideal lens the plane waves along the rays have to be all in phase at the focus. So, we just can set the focus to the coordinate $\boldsymbol{r}' = (x' = 0, y' = 0, z' = 0)$. To calculate the electric vector

$\hat{\boldsymbol{E}}_{\text{focus}}$ at a point $\boldsymbol{r}'$ near the focus there is according to [41] the equation:

$$\hat{\boldsymbol{E}}_{\text{focus}}\left(\boldsymbol{r}'\right) = \frac{1}{i\lambda f'} \int_{\text{focal sphere}} \hat{\boldsymbol{E}}' \mathrm{d}f' \tag{4.5.20}$$

A discretized form of this equation can be obtained by using equation (4.5.8) and taking into account that here $\hat{\boldsymbol{E}}'$ is the vector on the focal sphere:

$$\hat{\boldsymbol{E}}_{\text{focus}}\left(\boldsymbol{r}'\right) = \frac{1}{i\lambda f'} \sum_{i,j} \boldsymbol{P}'_{i,j} e^{i\boldsymbol{k}'_{i,j} \cdot \boldsymbol{r}'} \tag{4.5.21}$$

Of course, the last equality sign is only valid in the meaning that the continuous integral is replaced by a discrete sum. So, it is in fact only an approximation.
Small wave aberrations $W(x, y)$ of the lens can also be taken into account by just adding them to the phase term of each plane wave, i.e. $\boldsymbol{k}'_{i,j} \cdot \boldsymbol{r}' \rightarrow \boldsymbol{k}'_{i,j} \cdot \boldsymbol{r}' + W_{i,j}$ with $W_{i,j} := W(x_i, y_j)$. Then, we have:

$$\hat{\boldsymbol{E}}_{\text{focus}}\left(\boldsymbol{r}'\right) = \frac{1}{i\lambda f'} \sum_{i,j} \boldsymbol{P}'_{i,j} e^{i\boldsymbol{k}'_{i,j} \cdot \boldsymbol{r}' + iW_{i,j}} \tag{4.5.22}$$

It can be seen that in the focal plane $z' = 0$ and for linearly polarized light with a small numerical aperture this sum reduces to a discretized formulation of the scalar diffraction integral of equation (4.3.33) for calculating the light distribution near the focus of a lens.
By just changing the polarization vectors of the field distribution in front of the lens the effects of a quite arbitrary state of polarization to the focus can be investigated. The squares of the x–, y– or z–components of the electric field $\hat{\boldsymbol{E}}_{\text{focus}}$ can also be calculated separately. Of course, the time–averaged electric energy density w_e, which is the physical quantity which is normally detected by a light detector, can then be calculated by:

$$w_e\left(\boldsymbol{r}'\right) = \frac{1}{2}\epsilon_0 \left|\hat{\boldsymbol{E}}_{\text{focus}}\right|^2 = \frac{1}{2}\epsilon_0 \hat{\boldsymbol{E}}_{\text{focus}} \cdot \hat{\boldsymbol{E}}^*_{\text{focus}} \tag{4.5.23}$$

Here, only the case of focusing in vacuum (or air) is considered, i.e. $n = 1$ and $\epsilon = 1$. The proportionality factor $\epsilon_0/2$ can be omitted if we are only interested in relative energy distributions and not in absolute values.

4.5.3 Some simulation results

Figures 4.14, 4.15 and 4.16 show the results of some simulations for the electric energy density in the focal plane (x–y–plane) of an aplanatic lens. In all cases the numerical aperture of the lens is assumed to be 1.0 and the wavelength of the illuminating light is λ=632.8 nm. The typical number of plane waves which are used to sample the aperture of the lens is NxN=100x100 (or 200x200), whereby in the case of a circular aperture (normal case) and especially in the case of an annular aperture (see later) the effective number of plane waves is smaller (factor $\pi/4$ smaller for a circular aperture) because the sampling is made in an orthogonal uniformly sampled x–y–pattern and plane waves outside of the aperture just have zero amplitude. The number of sampling points in the focal plane is 200x200 and so a simulation on a modern 1 GHz Pentium PC takes less than two minutes.
In fig. 4.14 the lens is illuminated by a linearly polarized plane wave (polarization in y–direction). The squares of the different components of the electric vector in the focal plane are displayed.

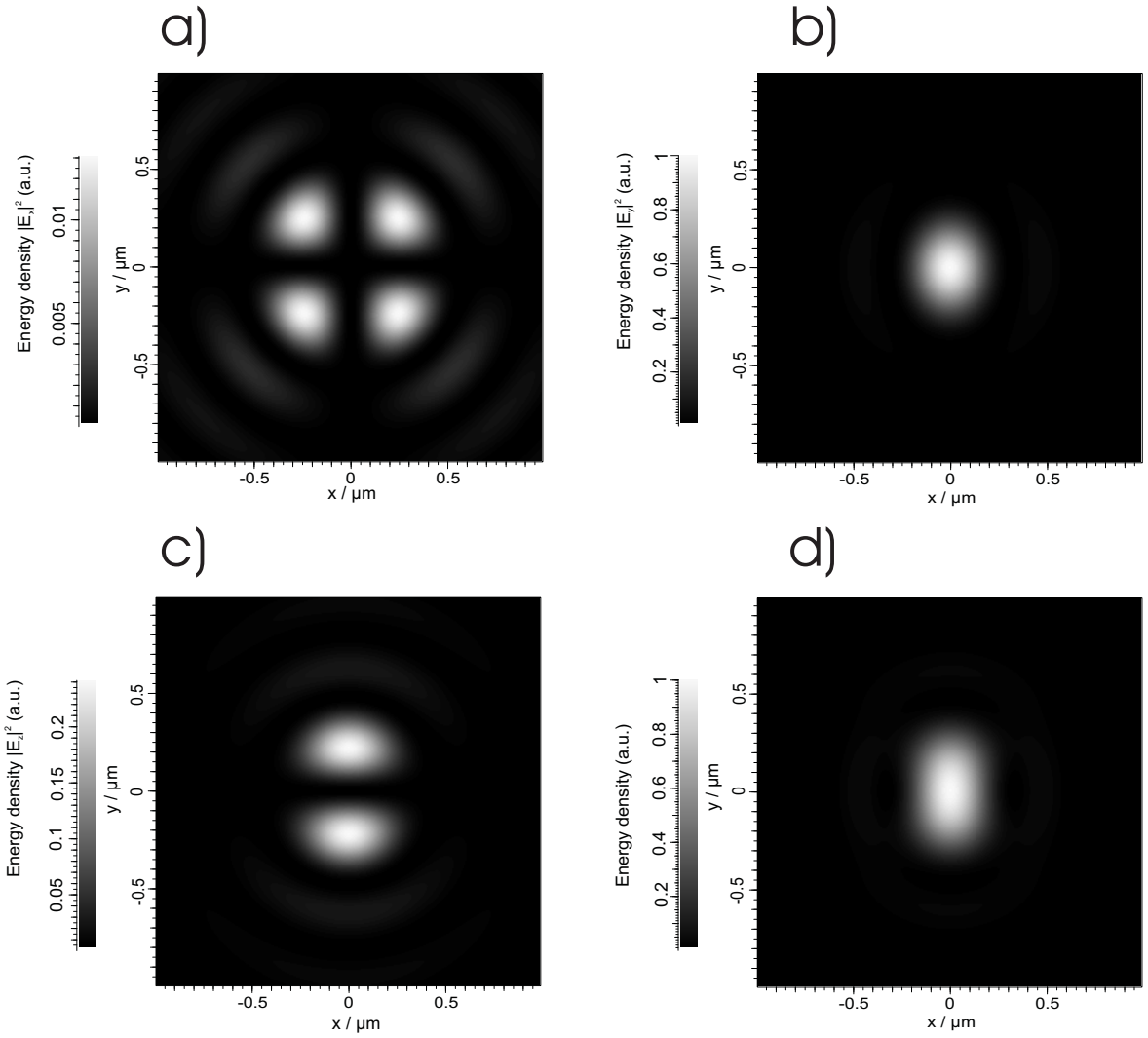


Figure 4.14: Squares of the components of the electric vector in the focal plane of an aplanatic lens with NA=1.0 which is illuminated by a plane linearly polarized (in y–direction) wave with λ=632.8 nm. a) x–component, b) y–component, c) z–component and d) sum of all components, i.e. total electric energy density.

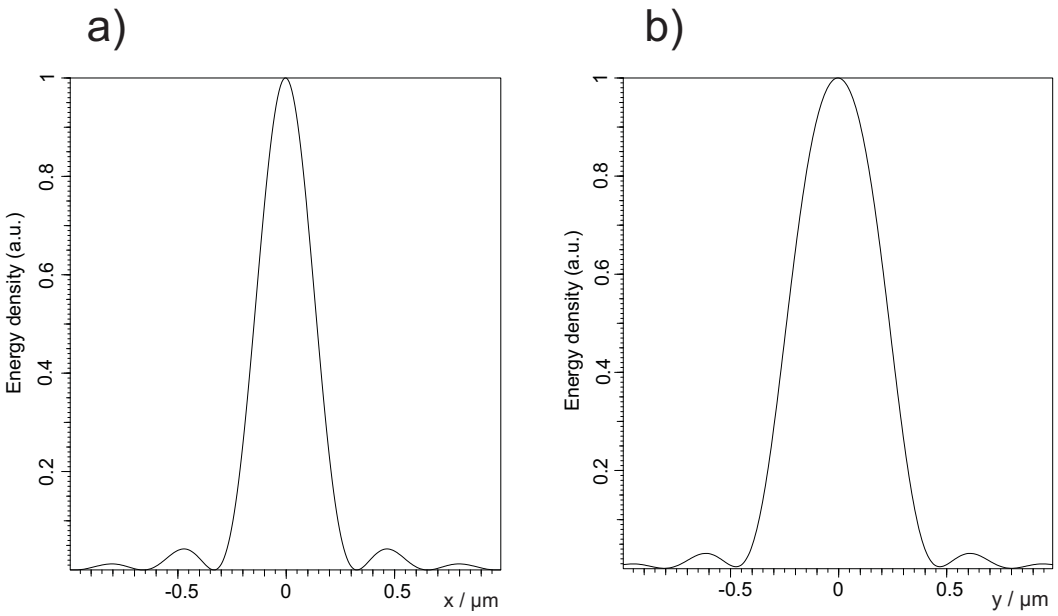


Figure 4.15: Sections along a) the x–axis and b) the y–axis of the total electric energy density in the focal plane of an aplanatic lens with NA=1.0 which is illuminated by a plane linearly polarized (in y–direction) wave with λ=632.8 nm (see also fig. 4.14 d)).

The x–component in a) is quite small and vanishes in the center of the geometrical focus (pay attention to the scale). The biggest component is the y–component in b), i.e. along the direction in which the light is polarized in front of the lens. But, there is also a z–component (see c)) for points besides the geometrical focus. This component is mainly responsible for the asymmetric shape of the total electric energy density which is displayed in d). Fig. 4.15 shows sections along the x– and y–axis of the total electric energy density. It can be seen, that the diameter of the central maximum along the y–axis is increased, whereas the diameter along the x–axis is even a little bit smaller than the value of the scalar calculation $d_{focus} = 1.22\lambda/\text{NA}=0.77\ \mu$m. All quantities are normalized in such a way that the total energy density has a maximum value of 1.

In fig. 4.16 the lens is illuminated by a radially polarized doughnut mode, whereby the beam waist w_0 is at 95% of the lens aperture radius. Then, there are radial i.e. transversal components of the electric field in a) which are of course rotationally symmetric due to the symmetry of the field. But, the strongest component is the longitudinal z–component in b) which is also rotationally symmetric and which has a central maximum with a diameter smaller than the value of the scalar calculation and even slightly smaller than the diameter of the small axis of the spot for linear polarization. The total electric energy density in c) is also rotationally symmetric, but, the diameter of the central maximum is increased due to the transversal components of the electric field. However, the surface area S which is covered in the focus by a total electric energy density of more than half the maximum value is $S = 0.29\lambda^2$ in the linearly polarized case and only $S = 0.22\lambda^2$ for the radially polarized doughnut mode. So, if the light spot is used to write into a nonlinear material which is only sensitive to a total electric energy density beyond a certain threshold a tighter spot can be obtained using the radially polarized doughnut mode.

Fig. 4.17 shows for the same parameters as in fig. 4.15 and 4.16 the total electric energy density

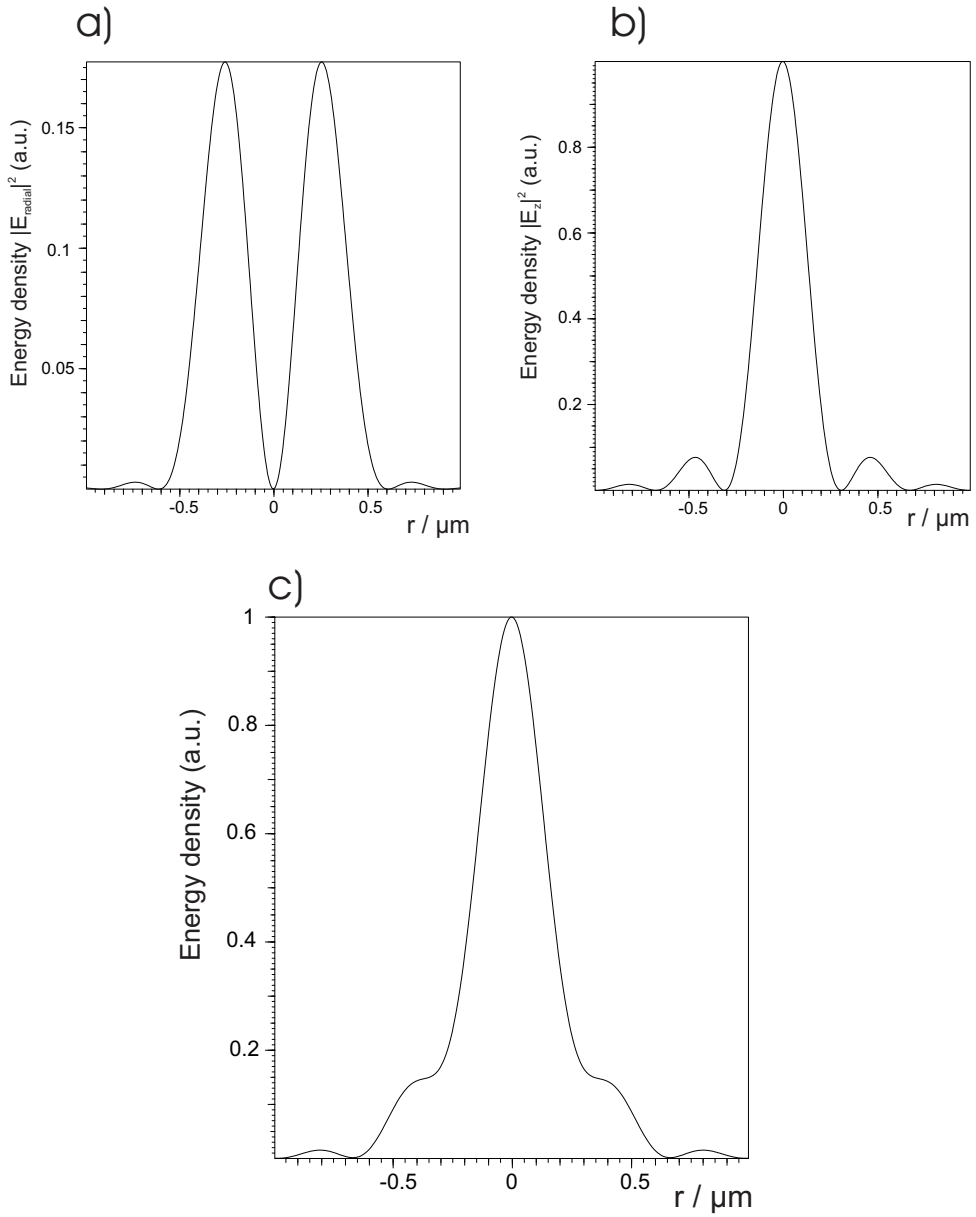


Figure 4.16: Squares of the components of the electric vector in the focal plane of an aplanatic lens with NA=1.0 which is illuminated by a radially polarized doughnut mode with λ=632.8 nm. The beam waist w_0 of the Gaussian function is $w_0 = 0.95 r_{aperture}$, whereby $r_{aperture}$ is the illuminated aperture radius of the lens. a) radial components, i.e. $|E_x|^2 + |E_y|^2$, b) z–component, c) sum of all components, i.e. total electric energy density.

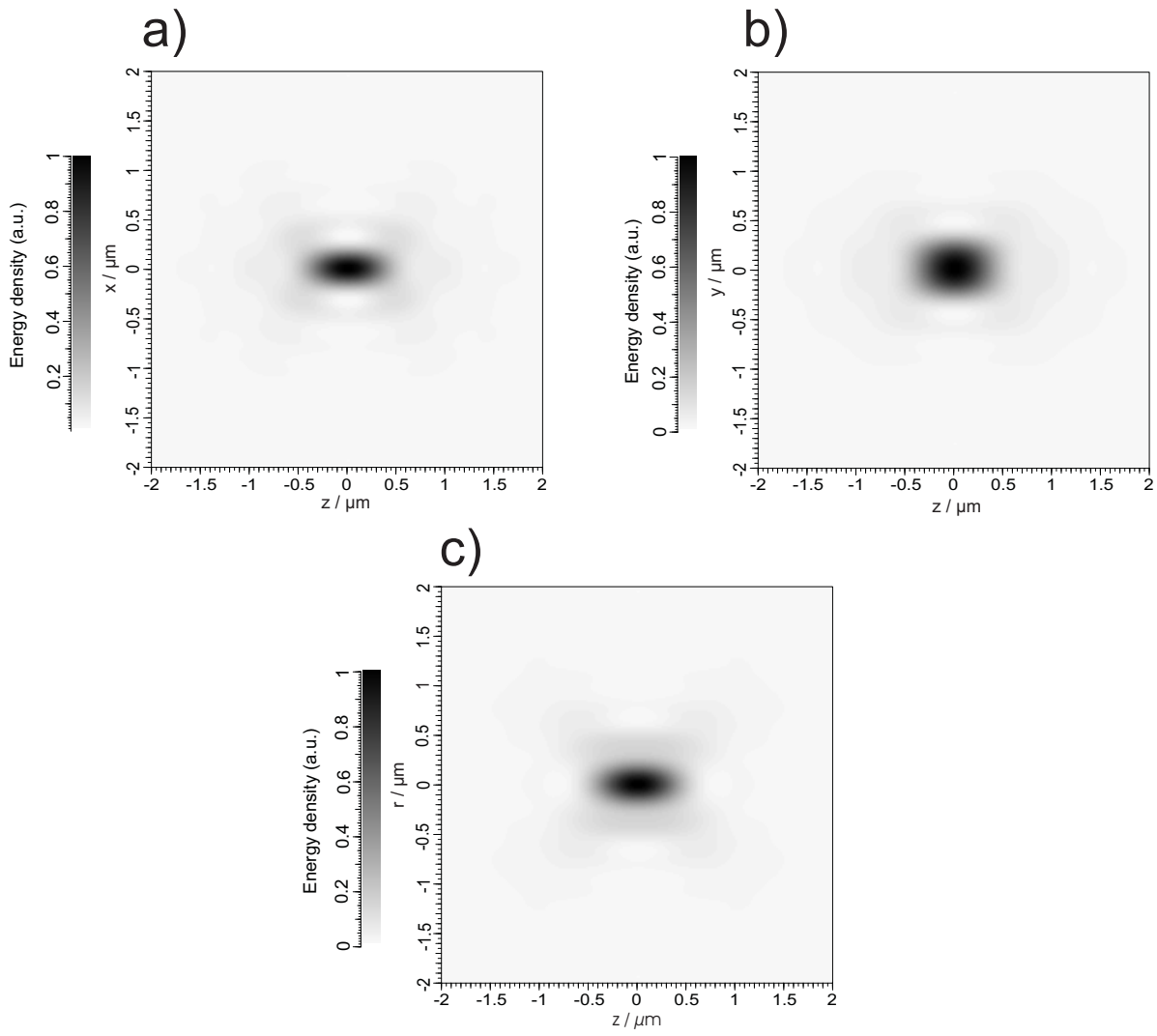


Figure 4.17: Simulation of the total electric energy density in the focal region by using a circular aperture (NA=1.0, λ=632.8 nm). a) linearly polarized homogeneous light (x–z–plane), b) linearly polarized homogeneous light (y–z–plane), c) radially polarized doughnut mode with $w_0 = 0.95 r_{aperture}$ (x–z–plane or y–z–plane because of rotational symmetry).

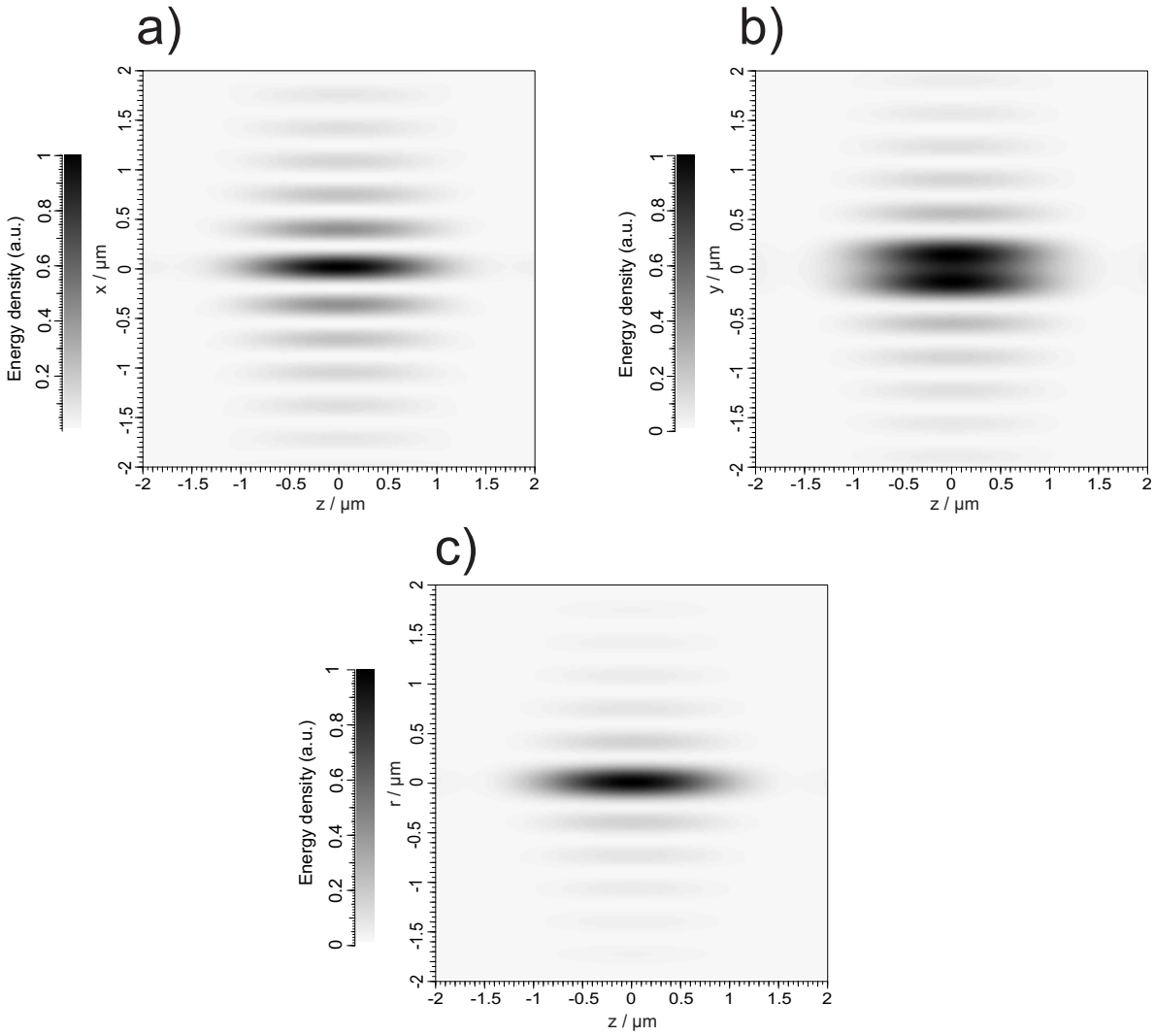


Figure 4.18: Simulation of the total electric energy density in the focal region by using an annular aperture with an inner radius of 90% of the full aperture (NA=1.0, λ=632.8 nm). a) linearly polarized homogeneous light (x–z–plane), b) linearly polarized homogeneous light (y–z–plane), c) radially polarized doughnut mode with $w_0 = 0.95 r_{aperture}$ (x–z–plane or y–z–plane because of rotational symmetry).

near the focus in planes containing the optical axis (z–axis) and one of the lateral coordinates. In a) and b) the x–z–plane and the y–z–plane are shown for the case of an illumination of the lens with linearly polarized homogeneous light (polarization in y–direction). c) shows the same for the case of the radially polarized doughnut mode whereby here the focal region is completely rotationally symmetric so that only one section has to be shown. It can again be seen that in the case of linear polarization the lateral diameter of the focus is broader in the y–z–plane than in the x–z–plane. The focus of the radially polarized doughnut mode has a lateral diameter of nearly the small axis of the focus of the linearly polarized case.

From the scalar theory it is well–known that the diameter of the central maximum of the focus is decreased if an annular aperture is used instead of a circular one. It is also well–known from the scalar theory that in this case the secondary maxima increase compared to a circular aperture. So, by using an annular aperture a smaller focus should be reached and this should be especially the case for the radially polarized doughnut mode. The reason is that in this case the transversal components of the electric field in the focus decrease so that the total electric energy density is more and more dominated by the longitudinal z–component which has a small diameter. So, by using a radially polarized doughnut mode, an annular aperture (e.g. $r_{annulus} = 0.9 r_{aperture}$) and a high numerical aperture lens (NA>0.9) a very tight rotationally symmetric spot with a small surface area can be obtained. This effect can be used to achieve a higher resolution in optical data storage or in direct writing optical lithography where a mask is written spot by spot by a so called laser pattern generator. Fig. 4.18 shows the simulation results of the electric energy density in the focal region for an annular aperture with $r_{annulus} = 0.9 r_{aperture}$. As in the former cases the numerical aperture itself is NA=1.0 and the wavelength is λ=632.8 nm. a) and b) show the x–z–plane and y–z–plane, respectively, for the case of linear polarization and c) shows the same for the doughnut mode ($w_0 = 0.95 r_{aperture}$) where the result is rotationally symmetric so that only one section is shown. It can be seen that in the case of an annular aperture the lateral diameter of the focus is decreased, but, on the other side the depth of focus along the optical axis is increased compared to the full aperture case. Besides the central maximum there are also as expected some secondary maxima with increased height. Nevertheless, a comparison of the cases of linear polarization and radial polarization shows that the lateral diameter of the focus is decreased for radial polarization, especially in the y–z–plane. By calculating again the surface area S which is covered in the focus by a total electric energy density of more than half the maximum value the result is $S = 0.29\lambda^2$ for the linearly polarized light and only $S = 0.12\lambda^2$ for the radially polarized doughnut mode. Moreover, the height of the secondary maxima is not so high for radial polarization than for linear polarization. In some applications these secondary maxima can be disturbing but in applications where a certain threshold value of the electric energy density has to be reached in order to obtain an effect these secondary maxima have no influence. So, the annular aperture, especially combined with the radially polarized doughnut mode, allows a high lateral resolution.

Experiments verify the simulation results for linearly polarized light [65] and for the radially polarized doughnut mode [40],[62].

Chapter 5

Fourier optics

It has been shown that the complex amplitude in the near field and in the far field are connected by each other with the Fraunhofer diffraction integral (4.3.24), which is mainly consisting of a Fourier transformation. It has also been shown that the complex amplitude in the focal plane of a lens is besides a scaling factor quite identical to the complex amplitude in the far field. Moreover, Fourier transformations are often used to calculate diffraction integrals like the Fresnel diffraction integral or the angular spectrum of plane waves. So, Fourier transformations are quite important in optics. In this chapter, some ideas to the so called Fourier optics are given, which deals for example with the transformation of a complex amplitude by a lens, with optical imaging in the case of coherent and incoherent light, or with optical filtering. Mostly, only the paraxial case will be treated, so that the Fresnel approximation can be used.

5.1 Transformation of the complex amplitude by a lens

A very interesting case is to calculate how the complex amplitude u_0, which is defined in a plane with a distance d_1 in front of a thin lens with focal length f', is transformed by this lens to a plane with distance d_2 behind the lens.
The complex amplitude $u^-(x, y, d_1)$ immediately in front of the lens which is situated in air is calculated by using the Fresnel diffraction integral (4.3.10):

$$u^-(x,y,d_1) = \frac{-i}{\lambda d_1} e^{i\frac{2\pi d_1}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_1}} \iint_{-\infty}^{+\infty} u_0(x',y')\, e^{i\pi\frac{x'^2+y'^2}{\lambda d_1}} e^{-2\pi i\frac{xx'+yy'}{\lambda d_1}} dx'dy' \quad (5.1.1)$$

The integration is done over the whole plane from minus infinity to plus infinity because a possible spatial limitation of the complex amplitude is assumed to be contained in the function u_0. As usual, λ is the wavelength of the monochromatic light.
In the paraxial approximation of Fourier optics, a thin lens changes only the phase of the complex amplitude. Therefore, the complex transmission function t of the lens which connects the complex amplitude u^- immediately in front of the lens with the complex amplitude u^+ immediately behind the lens is defined by:

$$t(x,y) = A(x,y)e^{-i\pi\frac{x^2+y^2}{\lambda f'}} \quad (5.1.2)$$

A is the pupil function of the lens which is one inside of the aperture and zero outside if no aberrations are present. Otherwise, it is defined as $\exp(iW(x,y))$ inside of the aperture and zero outside, where $W(x,y)$ are the wave aberrations of the lens (see also equation (4.3.27)). Later, the aperture will be set to infinity and aberrations will be neglected to simplify the calculations. The parabolic phase term is obtained from equation (4.3.26) in the paraxial approximation, i.e. by developing the square root in a Taylor series and taking only the first two leading terms. So, the complex amplitude $u^+(x,y,d_1)$ immediately behind the lens is:

$$u^+(x,y,d_1) = t(x,y)u^-(x,y,d_1) = A(x,y)e^{-i\pi\frac{x^2+y^2}{\lambda f'}}u^-(x,y,d_1) \tag{5.1.3}$$

To obtain the complex amplitude $u(x,y,d_1+d_2)$ in the distance d_2 behind the lens the Fresnel diffraction integral has to be used a second time:

$$\begin{aligned}
u(x,y,d_1+d_2) &= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} A(x',y') e^{i\pi\left(\frac{1}{\lambda d_1}-\frac{1}{\lambda f'}\right)(x'^2+y'^2)} \cdot \\
&\quad \cdot \left[\iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} e^{-2\pi i\frac{x'x''+y'y''}{\lambda d_1}} dx''dy''\right] \cdot \\
&\quad \cdot e^{i\pi\frac{x'^2+y'^2}{\lambda d_2}} e^{-2\pi i\frac{xx'+yy'}{\lambda d_2}} dx'dy' = \\
&= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty}\iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\
&\quad \cdot A(x',y') e^{i\pi\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}\right)\frac{x'^2+y'^2}{\lambda}} \cdot \\
&\quad \cdot e^{-\frac{2\pi i}{\lambda}\left[x'\left(\frac{x''}{d_1}+\frac{x}{d_2}\right)+y'\left(\frac{y''}{d_1}+\frac{y}{d_2}\right)\right]} dx'dy'dx''dy''
\end{aligned} \tag{5.1.4}$$

A general evaluation of this multiply dimensional integral is not so easy. Therefore, some interesting special cases will be treated.

5.1.1 Conjugated planes

The first interesting case is that the plane in the distance d_1 in front of the lens, where u_0 is defined, and the plane in the distance d_2 behind the lens are conjugated planes. This means that the lens images the original plane to the destination plane and the imaging equation $1/d_1 + 1/d_2 - 1/f' = 0$ is valid. Then, equation (5.1.4) results in:

$$\begin{aligned}
u(x,y,d_1+d_2) &= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty}\iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\
&\quad \cdot A(x',y') e^{-\frac{2\pi i}{\lambda}\left[x'\left(\frac{x''}{d_1}+\frac{x}{d_2}\right)+y'\left(\frac{y''}{d_1}+\frac{y}{d_2}\right)\right]} dx'dy'dx''dy''
\end{aligned} \tag{5.1.5}$$

For the special case that the diffraction at the finite lens aperture can be neglected A is constant one and the integration over $dx'dy'$ can be made first:

$$\begin{aligned} u(x,y,d_1+d_2) &= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\ &\quad \cdot\delta\left(\frac{x''}{\lambda d_1}+\frac{x}{\lambda d_2}\right)\delta\left(\frac{y''}{\lambda d_1}+\frac{y}{\lambda d_2}\right)dx''dy'' = \\ &= \frac{-d_1}{d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\ &\quad \cdot\delta\left(x''+\frac{d_1}{d_2}x\right)\delta\left(y''+\frac{d_1}{d_2}y\right)dx''dy'' = \\ &= \frac{-d_1}{d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} u_0\left(-\frac{d_1}{d_2}x,-\frac{d_1}{d_2}y\right) e^{i\pi\frac{d_1}{\lambda d_2^2}(x^2+y^2)} \end{aligned} \tag{5.1.6}$$

Here, the mathematical formulas for delta distributions

$$\int_{-\infty}^{+\infty} e^{-2\pi i x x'} dx' = \delta(x)$$

and

$$\delta(ax) = \frac{\delta(x)}{|a|}$$

have been used.
By defining the scaling factor $\beta = -d_2/d_1$, the result is:

$$u(x,y,d_1+d_2) = \frac{1}{\beta} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} u_0\left(\frac{x}{\beta},\frac{y}{\beta}\right) e^{i\pi\left(1-\frac{1}{\beta}\right)\frac{x^2+y^2}{\lambda d_2}} \tag{5.1.7}$$

It should again be emphasized that this equation is only valid if the diffraction at the lens aperture can be neglected. This is exactly the case when the lens aperture is infinite (what is never the case in practice) or when the complex amplitude u^- in the plane of the lens is approaching zero at the rim of the lens aperture. Then, the complex amplitude in the image plane is a scaled version of the complex amplitude u_0. Especially, the intensity $I(x,y,d_1+d_2)$ in the image plane and the intensity I_0 in the original plane are connected by:

$$I(x,y,d_1+d_2) = \frac{1}{\beta^2} I_0\left(\frac{x}{\beta},\frac{y}{\beta}\right) \tag{5.1.8}$$

The factor $1/\beta^2$ is responsible for the conservation of energy.

If the diffraction at the lens aperture cannot be neglected equation (5.1.5) can nevertheless formally be integrated over $dx'dy'$:

$$\begin{aligned} u(x,y,d_1+d_2) &= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\ &\quad \cdot \widetilde{A}\left(\frac{x''}{\lambda d_1}+\frac{x}{\lambda d_2}, \frac{y''}{\lambda d_1}+\frac{y}{\lambda d_2}\right) dx''dy'' = \\ &= \frac{-1}{\lambda^2 d_1 d_2} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\ &\quad \cdot \widetilde{A}\left(\frac{x/\beta - x''}{-\lambda d_1}, \frac{y/\beta - y''}{-\lambda d_1}\right) dx''dy'' \end{aligned} \tag{5.1.9}$$

Here, $\widetilde{A}$ is the Fourier transform of the lens pupil function A. Of course, this equation is identical to equation (5.1.7) if the diffraction at the lens aperture can be neglected. Then, $A(x,y)$ is constant one and $\widetilde{A}(\nu_x,\nu_y) = \delta(\nu_x,\nu_y)$.

5.1.2 Non–conjugated planes

The general case is

$$\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'} \neq 0$$

Also, in this case the integral (5.1.4) can be further processed by evaluating first the integral over $dx'dy'$. Therefore, the integral

$$F(\nu_x,\nu_y) = \iint_{-\infty}^{+\infty} A(x',y') e^{i\pi\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}\right)\frac{x'^2+y'^2}{\lambda}} e^{-2\pi i\left[x'\nu_x+y'\nu_y\right]} dx'dy' \tag{5.1.10}$$

has to be calculated. Here, the variables ν_x and ν_y were defined which are the conjugated variables to x' and y':

$$\nu_x = \frac{x''}{\lambda d_1}+\frac{x}{\lambda d_2} \tag{5.1.11}$$

$$\nu_y = \frac{y''}{\lambda d_1}+\frac{y}{\lambda d_2} \tag{5.1.12}$$

This is a Fourier transformation of the function $A(x',y')\exp(i\pi(1/d_1+1/d_2-1/f')(x'^2+y'^2)/\lambda)$. By using the convolution theorem of Fourier mathematics we obtain:

$$F(\nu_x,\nu_y) = \widetilde{A}(\nu_x,\nu_y) \otimes \mathcal{F}\left\{ e^{i\pi\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}\right)\frac{x'^2+y'^2}{\lambda}} \right\} \tag{5.1.13}$$

Here, $\mathcal{F}$ is used to designate symbolically a Fourier transformation which is exactly defined by:

$$G(\nu_x,\nu_y) := \quad \mathcal{F}\left\{ e^{i\pi\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}\right)\frac{x'^2+y'^2}{\lambda}} \right\} =$$

$$= \int\limits_{-\infty}^{+\infty}\!\!\int e^{i\pi\left(\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}\right)\frac{x'^2+y'^2}{\lambda}} e^{-2\pi i\left[x'\nu_x+y'\nu_y\right]} dx'dy' \quad (5.1.14)$$

This two–dimensional integral can be separated into the product of two one–dimensional integrals. For the further evaluation the parameter

$$a = \mathrm{sign}(a)\,|a| := (1/d_1 + 1/d_2 - 1/f')/\lambda$$

is introduced. Then, the one–dimensional integral over dx' is in fact a Gaussian integral and can be evaluated as follows:

$$\int\limits_{-\infty}^{+\infty} e^{i\pi a x'^2} e^{-2\pi i x'\nu_x} dx' =$$

$$= e^{-i\pi\,\mathrm{sign}(a)\frac{\nu_x^2}{|a|}} \int\limits_{-\infty}^{+\infty} e^{i\pi\,\mathrm{sign}(a)\left(\sqrt{|a|}x' - \mathrm{sign}(a)\nu_x/\sqrt{|a|}\right)^2} dx' =$$

$$= e^{-i\pi\frac{\nu_x^2}{a}} \frac{1}{\sqrt{\pi|a|}} \int\limits_{-\infty}^{+\infty} e^{i\,\mathrm{sign}(a)x''^2} dx'' = e^{-i\pi\frac{\nu_x^2}{a}} \frac{1}{\sqrt{\pi|a|}} \left(\sqrt{\frac{\pi}{2}} + i\,\mathrm{sign}(a)\sqrt{\frac{\pi}{2}}\right) =$$

$$= \frac{1+i\,\mathrm{sign}(a)}{\sqrt{2|a|}} e^{-i\pi\frac{\nu_x^2}{a}} \quad (5.1.15)$$

Here, the substitution $x'' = \sqrt{\pi|a|}x' - \mathrm{sign}(a)\sqrt{\pi/|a|}\nu_x$, the Euler equation $\exp(i\,\mathrm{sign}(a)x''^2) = \cos(x''^2) + i\,\mathrm{sign}(a)\sin(x''^2)$ and the relation

$$\int\limits_{-\infty}^{+\infty} \cos(x^2)dx = \int\limits_{-\infty}^{+\infty} \sin(x^2)dx = \sqrt{\frac{\pi}{2}} \quad (5.1.16)$$

were used. Totally, the function G is:

$$G(\nu_x,\nu_y) = \frac{i}{a}\, e^{-i\pi\frac{\nu_x^2+\nu_y^2}{a}} = \frac{i\lambda}{\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}}\, e^{-i\pi\lambda\frac{\nu_x^2+\nu_y^2}{\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}}} \quad (5.1.17)$$

In the following, only the case of an infinite lens aperture is treated. Then, the pupil function A has the constant value one and $\widetilde{A}$ is a delta–function. Therefore, F and G are identical, i.e.

$F(\nu_x, \nu_y) = G(\nu_x, \nu_y)$, and by using equation (5.1.4) the result is:

$$\begin{aligned} u(x, y, d_1 + d_2) &= \frac{-i}{\lambda(d_1 + d_2 - d_1 d_2/f')} e^{i\frac{2\pi(d_1+d_2)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_2}} \iint_{-\infty}^{+\infty} u_0(x'', y'') \cdot \\ &\cdot e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} e^{-i\pi\lambda\frac{\left(\frac{x''}{\lambda d_1}+\frac{x}{\lambda d_2}\right)^2+\left(\frac{y''}{\lambda d_1}+\frac{y}{\lambda d_2}\right)^2}{\frac{1}{d_1}+\frac{1}{d_2}-\frac{1}{f'}}} dx''dy'' \end{aligned} \tag{5.1.18}$$

This general expression will be evaluated in the following for an especially interesting case. This is, if the detector plane is in the back focal plane of the lens. Then, from geometrical optics we know that rays with the same direction in front of the lens are focussed to one point in the back focal plane. But, a ray with a certain direction corresponds to a plane wave with a certain spatial frequency. So, spatial frequencies in front of the lens are transformed to spatial coordinates in the back focal plane of the lens. This is the typical property of a Fourier transformation.

5.1.3 Detector plane in the back focal plane of the lens

If the detector plane is in the back focal plane of the lens, i.e. $d_2 = f'$, equation (5.1.18) can be further simplified:

$$\begin{aligned} u(x, y, d_1 + f') &= \frac{-i}{\lambda f'} e^{i\frac{2\pi(d_1+f')}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda f'}} \iint_{-\infty}^{+\infty} u_0(x'', y'')\, e^{i\pi\frac{x''^2+y''^2}{\lambda d_1}} \cdot \\ &\cdot e^{-i\pi\frac{d_1}{\lambda}\left(\frac{x''^2}{d_1^2}+\frac{x^2}{f'^2}+2\frac{xx''}{d_1 f'}+\frac{y''^2}{d_1^2}+\frac{y^2}{f'^2}+2\frac{yy''}{d_1 f'}\right)} dx''dy'' = \\ &= \frac{-i}{\lambda f'} e^{i\frac{2\pi(d_1+f')}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda f'}\left(1-\frac{d_1}{f'}\right)} \cdot \\ &\cdot \iint_{-\infty}^{+\infty} u_0(x'', y'')\, e^{-2\pi i\frac{xx''+yy''}{\lambda f'}} dx''dy'' \end{aligned} \tag{5.1.19}$$

This means that the complex amplitude in the back focal plane of the lens is the product of the Fourier transform of the complex amplitude u_0 in the original plane and a position dependent complex phase factor. So, the intensity in the detector plane $I(x, y, d_1 + f')$ is proportional to the square of the modulus of the Fourier transform of u_0, because the complex phase factor has the modulus one.
Another more special case is that also the distance d_1 of the original plane to the lens is equal to the focal length f' of the lens, i.e. $d_1 = f'$. Then, the position dependent complex phase factor vanishes and $u(x, y, 2f')$ is indeed proportional to the Fourier transform of u_0:

$$u(x, y, 2f') = \frac{-i}{\lambda f'} e^{i\frac{4\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} u_0(x'', y'')\, e^{-2\pi i\frac{xx''+yy''}{\lambda f'}} dx''dy'' \tag{5.1.20}$$

This is the well–known property of Fourier optics, that the Fourier transform of a complex amplitude u_0 is formed in the back focal plane of a lens if u_0 is situated in the front focal plane. Of course, this is only valid if coherent light is used.

5.2 Imaging of extended objects

The imaging of extended objects is a very important task of optics. Aberrations and/or different pupil forms (e.g. circular or annular pupil) influence the quality of optical imaging. In this section, we will only treat the imaging with either fully coherent light or with fully incoherent light. The more general case of imaging with partially coherent light will not be treated.

5.2.1 Imaging with coherent light

In section 5.1.1 we already investigated the imaging with coherent light in the paraxial case. The basic idea of imaging with coherent light is that a complex amplitude u_O in the object plane is imaged by a lens (or a complex system of several single lenses like in the case of a camera objective or microscope objective) to the image plane. There, the complex amplitude u_I is obtained.
Equation (5.1.5) is used in the paraxial case. But, u_O can formally be written as a convolution of u_O with a delta function:

$$u_O(x,y) = \iint_{-\infty}^{+\infty} u_O(x',y')\ \delta(x-x',y-y')\,dx'dy' \tag{5.2.1}$$

By introducing this into equation (5.1.5) we can write:

$$\begin{aligned}
u_I(x,y) &= a(x,y) \iint_{-\infty}^{+\infty} \iint_{-\infty}^{+\infty} u_O(x'',y'')\, e^{-i\pi \frac{x''^2+y''^2}{\lambda d_O}} \cdot \\
&\quad \cdot A(x',y') e^{-\frac{2\pi i}{\lambda}\left[x'\left(-\frac{x''}{d_O}+\frac{x}{d_I}\right)+y'\left(-\frac{y''}{d_O}+\frac{y}{d_I}\right)\right]} dx'dy'dx''dy'' = \\
&= a(x,y) \iiint_{-\infty}^{+\infty} u_O(x''',y''')\,\delta(x''-x''',y''-y''')\, e^{-i\pi \frac{x''^2+y''^2}{\lambda d_O}} \cdot \\
&\quad \cdot A(x',y') e^{-\frac{2\pi i}{\lambda d_I}\left[x'(x-\beta x'')+y'(y-\beta y'')\right]} dx'dy'dx''dy''dx'''dy''' = \\
&= \iint_{-\infty}^{+\infty} u_O(x''',y''') \left[a(x,y) \iiint_{-\infty}^{+\infty} \delta(x''-x''',y''-y''')\, e^{-i\pi \frac{x''^2+y''^2}{\lambda d_O}} \cdot \right. \\
&\quad \left. \cdot A(x',y') e^{-\frac{2\pi i}{\lambda d_I}\left[x'(x-\beta x'')+y'(y-\beta y'')\right]} dx'dy'dx''dy'' \right] dx'''dy'''
\end{aligned} \tag{5.2.2}$$

Here, d_O (corresponding to $-d_1$ in equation (5.1.5)) and d_I (corresponding to d_2) are the object and image distances, respectively. Please, keep in mind the sign conventions of geometrical optics for d_O and d_I, where e.g. $d_O < 0$ and $d_I > 0$ for the real imaging of an object in front of a positive lens (and in front of the focal point of this lens on the object side). $\beta = d_I/d_O$ is the scaling factor of the imaging. The factor a is defined as

$$a(x,y) = \frac{1}{\lambda^2 d_O d_I} e^{i\frac{2\pi(-d_O+d_I)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_I}} \tag{5.2.3}$$

and its modulus is $1/(\lambda^2|d_O d_I|)$.
Now, the integration over $dx''dy''$ can first be performed so that equation (5.2.2) delivers:

$$\begin{aligned} u_I(x,y) &= \iint_{-\infty}^{+\infty} u_O(x''',y''')\, a(x,y) e^{-i\pi\frac{x'''^2+y'''^2}{\lambda d_O}} \cdot \\ &\quad \cdot \left[\iint_{-\infty}^{+\infty} A(x',y') e^{-\frac{2\pi i}{\lambda d_I}\left[x'\left(x-\beta x'''\right)+y'\left(y-\beta y'''\right)\right]} dx'dy' \right] dx'''dy''' \end{aligned} \tag{5.2.4}$$

Finally, u_I can be written as convolution between the complex amplitude u_O in the object plane and the so called **impulse response** h_{lens} of the imaging system:

$$u_I(x,y) = \iint_{-\infty}^{+\infty} u_O(x',y')\, h_{lens}(x-\beta x', y-\beta y')\, dx'dy' \tag{5.2.5}$$

with

$$\begin{aligned} h_{lens}(x-\beta x', y-\beta y') &= \frac{1}{\lambda^2 d_O d_I} e^{i\frac{2\pi(-d_O+d_I)}{\lambda}} e^{i\pi\frac{x^2+y^2}{\lambda d_I}} e^{-i\pi\frac{x'^2+y'^2}{\lambda d_O}} \cdot \\ &\quad \cdot \iint_{-\infty}^{+\infty} A(x'',y'') e^{-\frac{2\pi i}{\lambda d_I}\left[x''(x-\beta x')+y''(y-\beta y')\right]} dx''dy'' \end{aligned} \tag{5.2.6}$$

Note, that the integration variables were renamed and now (x,y) are the coordinates in the image plane, (x',y') the coordinates in the object plane and (x'',y'') the coordinates in the plane of the (thin) lens.
In most cases, the calculation of the impulse response h_{lens} can be simplified because mostly only the intensity in the image plane is of interest. Then, the first two phase factors $\exp(2\pi i(-d_O + d_I)/\lambda)$ and $\exp(i\pi(x^2+y^2)/(\lambda d_I))$ are of no interest because they will not change the intensity in the image plane. But, the third phase factor $\exp(-i\pi(x'^2+y'^2)/(\lambda d_O))$ depends on (x',y') and so the result of the integration over $dx'dy'$ in equation (5.2.5) will depend on this phase factor. But, there is also an argument why this phase factor can be neglected in many cases [8]: The impulse response of a good imaging system will be similar to a delta function, and therefore, a point (x,y) in the image plane will only be influenced by light from points of the object

plane which are very near to $(x/\beta, y/\beta)$, which is the conjugated point. So, the phase factor $\exp(-i\pi(x'^2 + y'^2)/(\lambda d_O))$ can be replaced with a good approximation by $\exp(-i\pi(x^2 + y^2)/(\lambda d_O \beta^2))$. But now, this phase factor no longer depends on the integration variables (x', y') of the convolution (5.2.5) and if we are only interested in the intensity in the image plane it can also be neglected. For a further discussion see [8].
Totally, in many cases the interesting part of the impulse response h_{lens} is just

$$h_{lens}\big(x - \beta x', y - \beta y'\big) = \frac{e^{i\phi_0(x,y)}}{\lambda^2 d_O d_I} \iint_{-\infty}^{+\infty} A(x'', y'') e^{-\frac{2\pi i}{\lambda d_I}\left[x''\left(x - \beta x'\right) + y''\left(y - \beta y'\right)\right]} dx'' dy'' \tag{5.2.7}$$

where the phase factor $\exp(i\phi_0(x, y))$ with modulus one depends only on the coordinates (x, y) and is of no interest if we are only interested in the intensity in the image plane.
Equation (5.2.5), which describes the complex amplitude u_I in the image plane as a convolution of the complex amplitude u_O in the object plane and the impulse response h_{lens}, is a quite general equation for coherent imaging as long as the optical system is **isoplanatic** in the regarded region. Isoplanatic means that the impulse response h_{lens} does not depend on (x, y) and (x', y') themselves, but only on $(x - \beta x', y - \beta y')$. In other words, this means, that the aberrations of the optical system are identical for all regarded object points. Then, equation (5.2.5) can also be used in the non–paraxial case. Of course, the impulse response h_{lens} has then also to be calculated non–paraxially. But, equation (5.2.7) can also be used in the non–paraxial case (as long as the scalar approximation is valid) with a good approximation if the pupil function A takes into account aberrations and if the aberrations are not too large. If the aberrations are too large the before–mentioned neglecting of the dependency of one of the parabolic phase factors on (x', y') is not valid, because the image point is too large.
If the system is not isoplanatic, the more general formulation

$$u_I(x, y) = \iint_{-\infty}^{+\infty} u_O\big(x', y'\big)\, h_{lens}\big(x, y; x', y'\big)\, dx' dy' \tag{5.2.8}$$

can be used, where h_{lens} depends explicitly on both the coordinates (x, y) in the image plane and the coordinates (x', y') in the object plane. Then, h_{lens} needs to be calculated for each object point independently taking into account the different aberrations for different object points.
After having treated the imaging with coherent light, the more interesting case of imaging with incoherent light is treated in the next section.

5.2.2 Imaging with incoherent light

The disadvantage of imaging with coherent light (see last section) is that the strong interference effects between neighbored image points can deteriorate the image quality very strongly. The reason is that the complex amplitudes are added and the impulse response h_{lens}, which should be named more precisely the **amplitude impulse response**, has often quite high secondary maxima and is in general a complex–valued function.
If incoherent light is used for the imaging this means that light from different object points cannot interfere with each other and the intensities of the images of all object points have to be added or, more precisely, integrated. Of course, the light which is emitted from one object

point is always coherent. But, the light from neighbored points is incoherent to the other points. Therefore, the **intensity impulse response**, which describes the intensity distribution in the image plane which is produced by a point source in the object plane, has to be taken instead of the amplitude impulse response. The intensity impulse response is just the square of the modulus of the amplitude impulse response and is apart from a normalization factor identical to the **point spread function** PSF which we already defined in equation (4.3.37) for the case of an incident plane wave. So, the intensity impulse response is always a real–valued positive function. Therefore, destructive interference effects like in the case of coherent light are not possible with incoherent light and the imaging with incoherent light is less noisy than with coherent light.
In analogy to equation (5.2.8) we have for the imaging with incoherent light in the general case:

$$I_I(x,y) = \kappa \iint_{-\infty}^{+\infty} I_O\big(x',y'\big)\,\big|h_{lens}\big(x,y;x',y'\big)\big|^2\,dx'dy' \tag{5.2.9}$$

The factor κ is just a normalization factor for energy conservation (but with a physical unit of the square of a length) so that the integrals of I_I and I_O are identical over the image and object plane, respectively, if no absorption is present:

$$\iint_{-\infty}^{+\infty} I_I(x,y)dxdy = \iint_{-\infty}^{+\infty} I_O(x',y')dx'dy' \tag{5.2.10}$$

$$\Rightarrow\ \kappa = \frac{\iint_{-\infty}^{+\infty} I_O(x',y')dx'dy'}{\iint_{-\infty-\infty}^{+\infty+\infty} I_O(x',y')\,|h_{lens}(x,y;x',y')|^2\,dx'dy'dxdy} \tag{5.2.11}$$

In isoplanatic regions of optical systems, the intensity impulse response depends, like in the coherent case, only on the relative position $(x-\beta x', y-\beta y')$ and we can write for imaging with incoherent light in analogy to equation (5.2.5):

$$I_I(x,y) = \kappa \iint_{-\infty}^{+\infty} I_O\big(x',y'\big)\,\big|h_{lens}\big(x-\beta x', y-\beta y'\big)\big|^2\,dx'dy' \tag{5.2.12}$$

This means, that the intensity I_I in the image plane is a convolution of the intensity I_O in the object plane and the intensity impulse response $|h_{lens}|^2$. The factor κ again guarantees energy conservation.
Since $|h_{lens}|^2$ is independent of complex phase factors the assumption of coherent imaging that h_{lens} can only be described by equation (5.2.7) for nearly delta function like PSFs, i.e. small aberrations, is no longer necessary in the case of incoherent imaging. There, we can really write:

$$\big|h_{lens}\big(x-\beta x', y-\beta y'\big)\big|^2 = \frac{1}{\lambda^4 d_O^2 d_I^2}\left|\iint_{-\infty}^{+\infty} A(x'',y'')e^{-\frac{2\pi i}{\lambda d_I}\left[x''\left(x-\beta x'\right)+y''\left(y-\beta y'\right)\right]}dx''dy''\right|^2 \tag{5.2.13}$$

5.2.3 Some examples for imaging with incoherent light

5.2.3.1 Cross grating as object

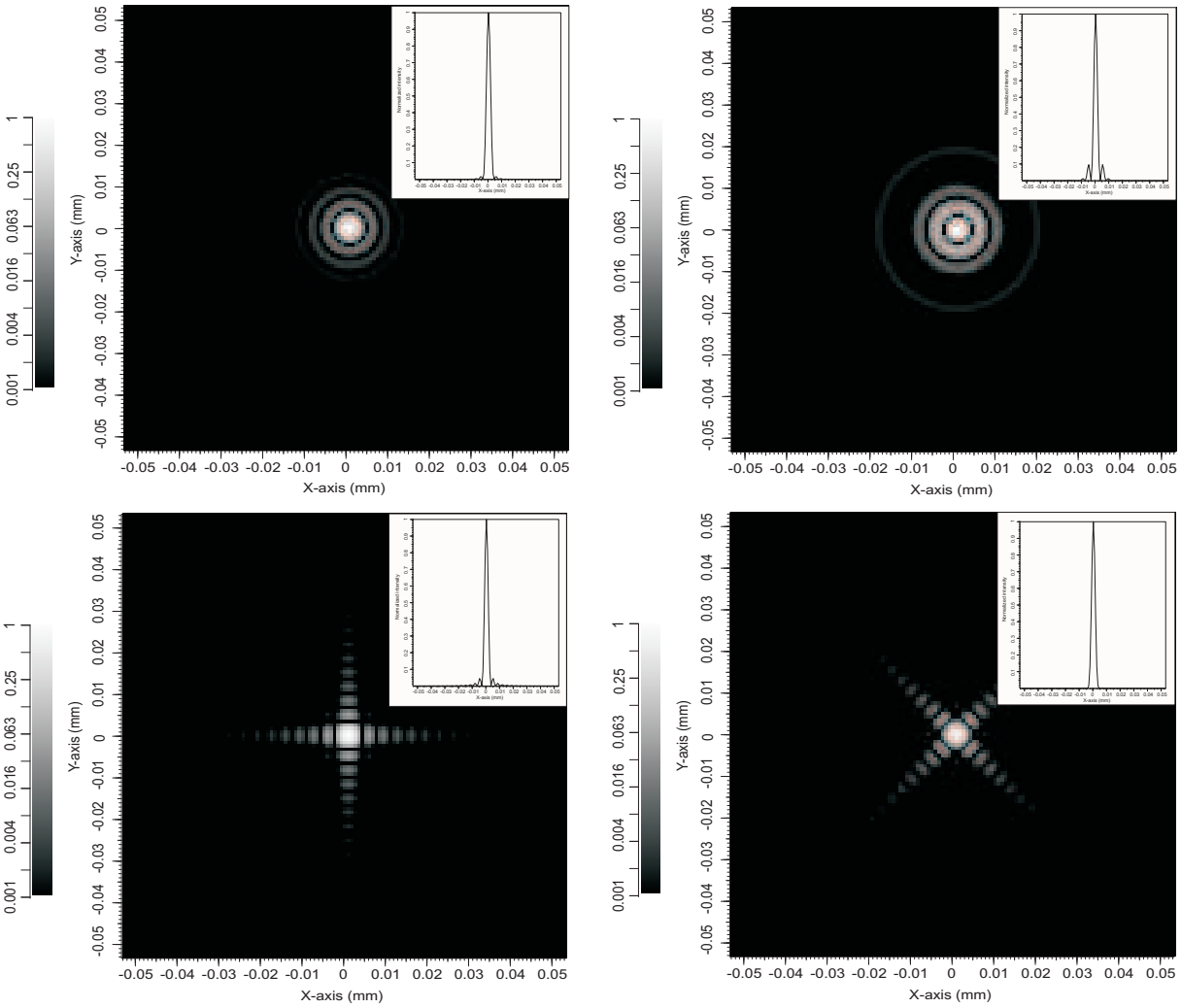


Figure 5.1: Point spread functions for different aperture shapes (wavelength $\lambda = 0.5$ μm, numerical aperture NA=0.075). Each figure shows a two–dimensional plot of the PSF with logarithmic scale and a small section along the x-axis with linear scale. Top left: circular aperture, top right: annular aperture with 50% of the diameter opaque, bottom left: quadratic aperture with axes parallel to the axes of the cross grating, bottom right: quadratic aperture rotated by 45 degree.

In the following, the influence of diffraction effects and aberrations on the imaging of a cross grating with incoherent light will be demonstrated by simulations. The optical system has a numerical aperture of NA $= 0.075$ and the wavelength is $\lambda = 0.5$ μm. In the case of a quadratic aperture the numerical aperture along the diagonal is by a factor $\sqrt{2}$ larger. The diameter of the diffraction–limited Airy disc for a circular aperture is (see equation (4.3.51)):

$$d_{Airy} = 1.22\frac{\lambda}{\mathrm{NA}} = 8.13\,\mu\mathrm{m} \tag{5.2.14}$$

The cross grating has a period of 10 μm and a duty cycle of 1:1. Therefore, the smallest

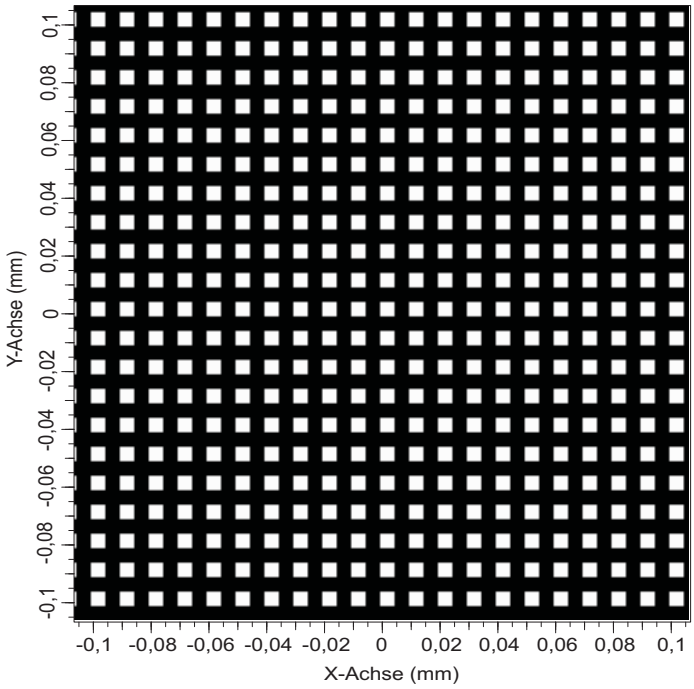


Figure 5.2: Intensity distribution of the original cross grating as object.

feature size is 5 μm. So, the smallest structures are just resolved by the circular aperture (the resolution is according to the Rayleigh condition $d_{Airy}/2$), but a significant blurring or rounding of the edges is expected. Figure 5.2 shows the original object. Figure 5.1 shows the point spread functions for different aperture forms and figure 5.3 the corresponding diffraction–limited images (i.e. without aberrations) of the cross grating. The "normal" case is a circular aperture (fig. 5.1 and 5.3 top left, respectively; see also section 4.3.4.2 for the PSF of a circular aperture). According to diffraction theory an annular aperture (here with an inner radius of 50% of the maximum radius) has a point spread function with a smaller central maximum, but with higher secondary maxima (see also figures 4.17 and 4.18 for a comparison of the PSF of a circular and an annular aperture for a high numerical aperture including polarization effects). In our case, the central maximum has a diameter of 6.7 μm, but the first secondary maxima have 10% of the height of the central maximum (see fig. 5.1 top right), whereas the first secondary maxima in the case of a circular aperture have only a height of 1.75% of the central maximum. Fig. 5.3 top right shows that for the annular aperture the image is blurred compared to the circular aperture because the higher secondary maxima deteriorate the image. A quadratic aperture with the axes parallel to the axes of the cross grating (see fig. 5.1 bottom left for the PSF and fig. 5.3 bottom left for the image of the cross grating) gives a quite similar image as in the case of a circular aperture. Here, according to section 4.3.4.1 the central maximum is smaller as in the case of a circular aperture (by a factor 1/1.22=0.82), but the secondary maxima, which are only present along the axes, are a little bit higher (4.7% of the height of the central maximum). The last case (see fig. 5.1 bottom right for the PSF and fig. 5.3 bottom right for the image of the cross grating) is a quadratic aperture, where the axes are rotated by 45 degree relative to the axes of the cross grating. In this case, the edges of the image seem to be a little bit sharper

than in all other cases.
The influence of spherical aberration (Zernike coefficient $a_S = 0.4\lambda$), astigmatism (Zernike coefficient $a_A = 0.5\lambda$) and coma (Zernike coefficient $a_C = 0.5\lambda$) is demonstrated in figures 5.4–5.6, which are all calculated for a circular aperture. In the case of spherical aberration and astigmatism the images are shown in different interesting planes so that a defocus term compensates partially the aberrations.
In the case of astigmatism it can be clearly seen that in the planes of the two focal lines, which are perpendicular to each other, either the vertical or the horizontal lines of the cross grating can be resolved (figure 5.5 top left and bottom). In the case of spherical aberration (figure 5.4) and in the case of astigmatism in the intermediate plane with a circular spot (figure 5.5 top right) a general blurring of the image can be detected. For spherical aberration the grating can hardly be resolved in the paraxial image plane (fig. 5.4 top left). In the other two shown image planes (fig. 5.4 top right and bottom) the defocus term compensates some of the spherical aberration. In the case of coma there is a certain asymmetry of the image in x– and y–direction, but not so clear as in the case of astigmatism.

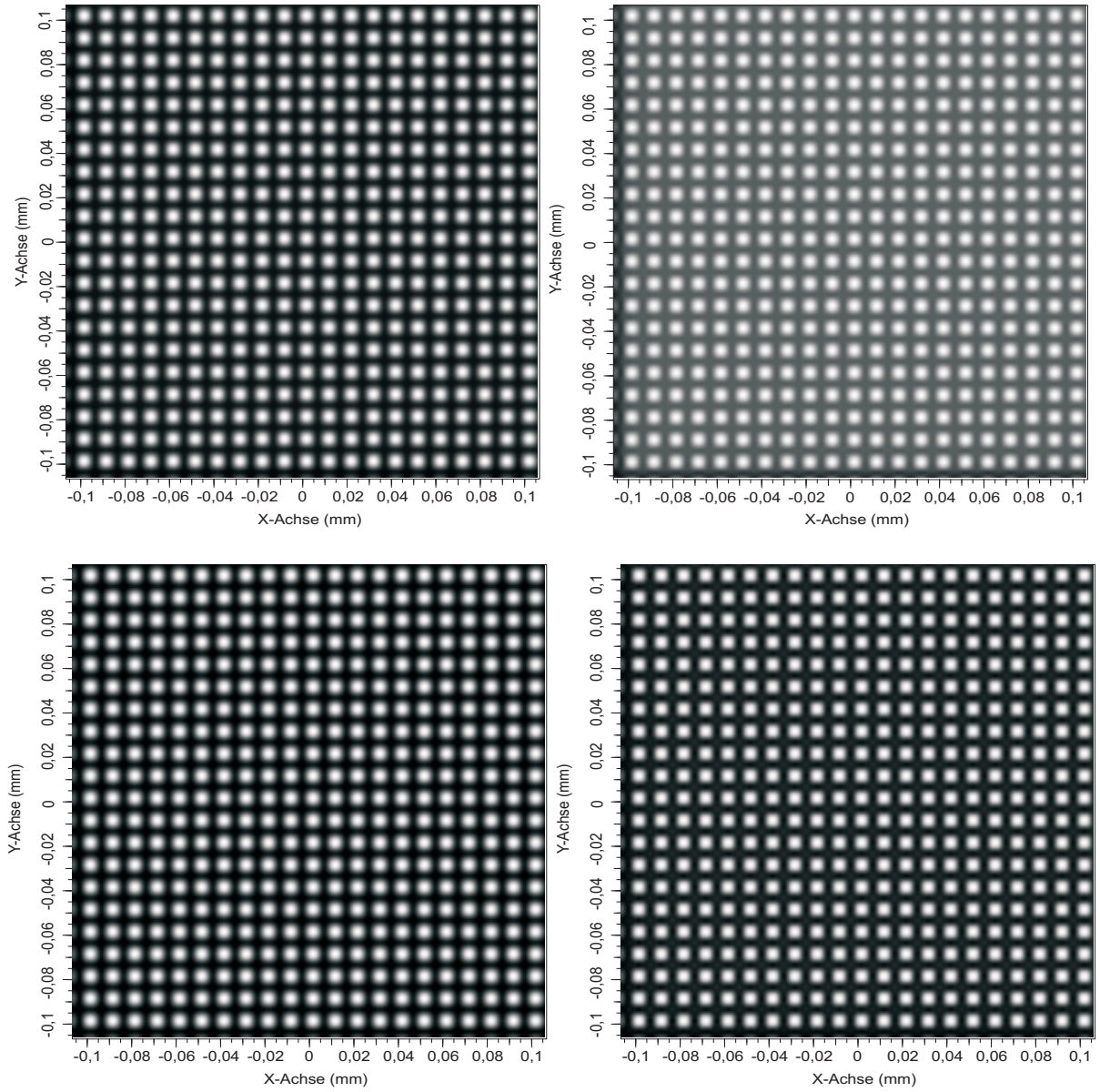


Figure 5.3: Intensity distributions of diffraction–limited images of the cross grating for different aperture forms. The corresponding point spread functions are shown in fig. 5.1. Top left: circular aperture, top right: annular aperture with 50% of the diameter opaque, bottom left: quadratic aperture with axes parallel to the axes of the cross grating, bottom right: quadratic aperture rotated by 45 degree.

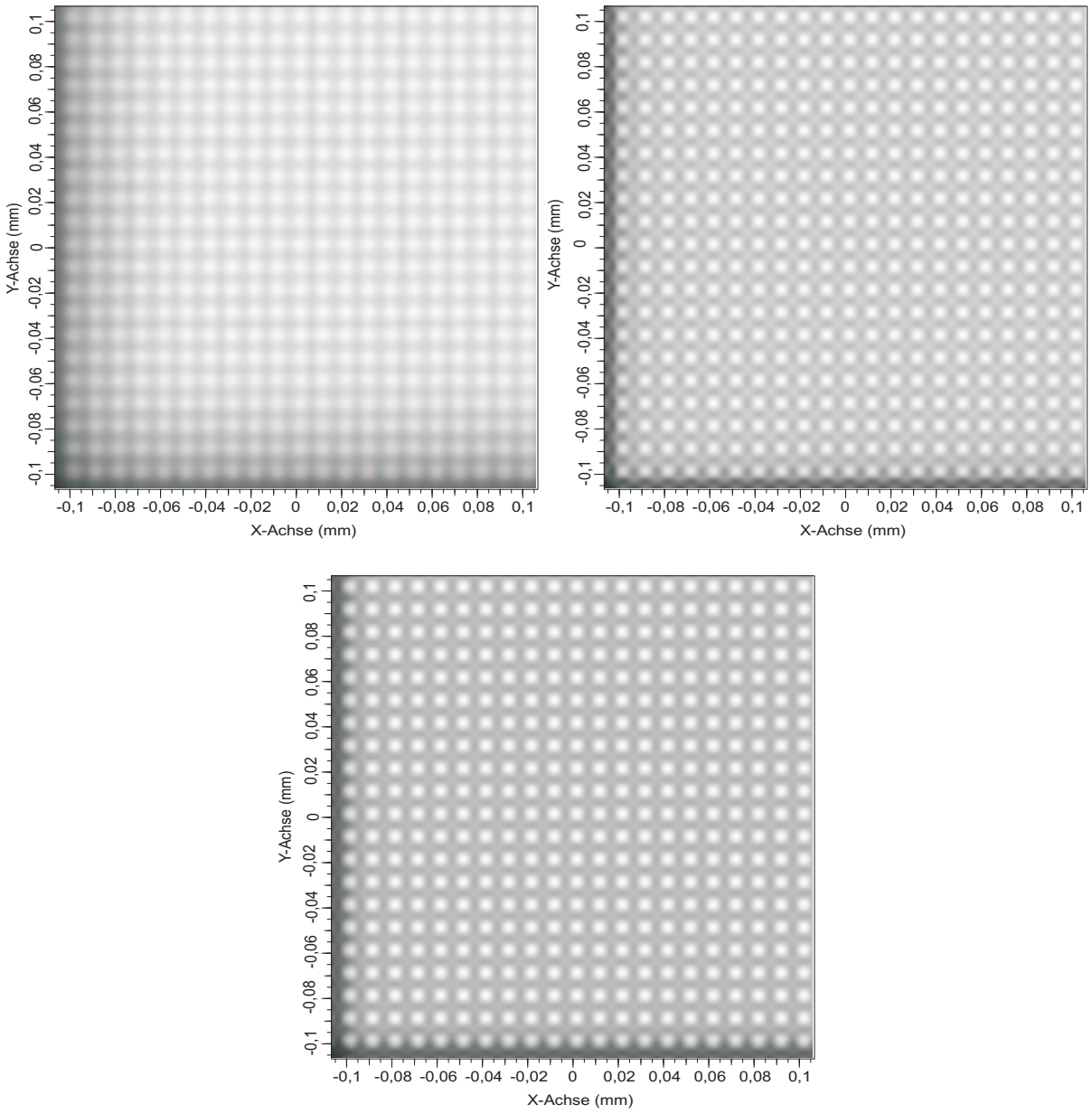


Figure 5.4: Intensity distributions of the images of the cross grating in the case of spherical aberration with Zernike coefficient $a_S = 0.4\lambda$. Left upper figure: image in the paraxial focal plane, right upper figure: image in the focal plane with the minimum wave aberrations, bottom figure: image in the focal plane with minimum ray aberrations (transverse aberrations of geometrical optics).

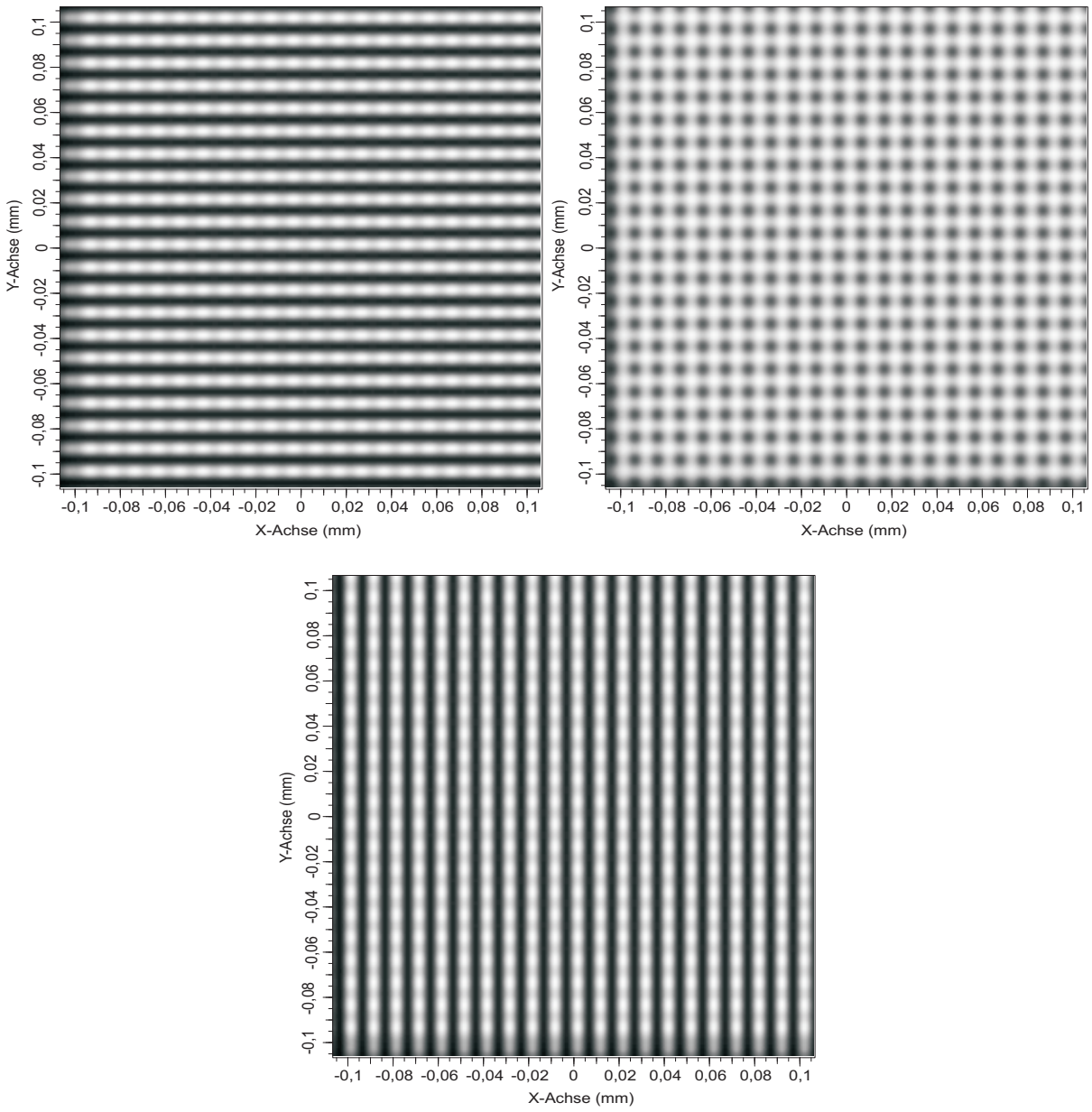


Figure 5.5: Intensity distributions of the images of the cross grating in the case of astigmatism with Zernike coefficient $a_A = 0.5\lambda$. Left upper figure: image in the plane of the first focal line, right upper figure: image in the intermediate plane with circular spot, bottom figure: image in the plane of the second focal line which is perpendicular to the first focal line.

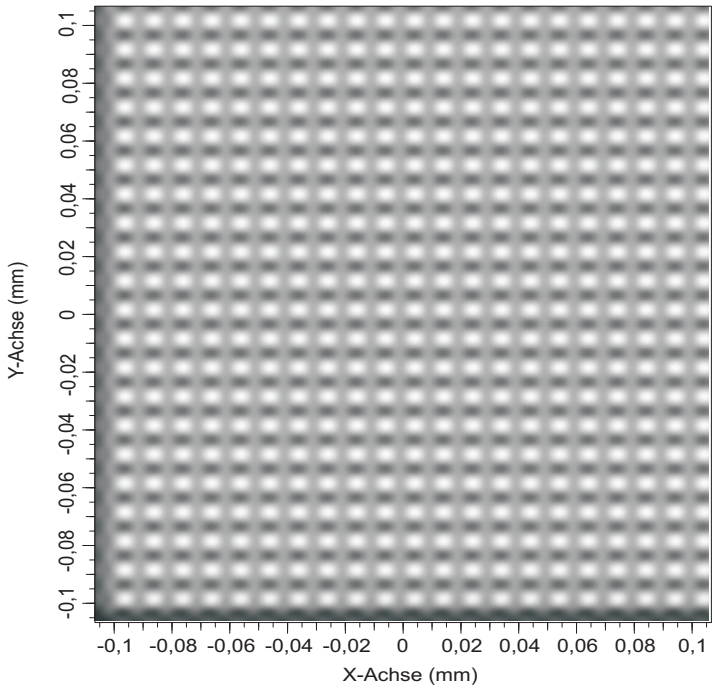


Figure 5.6: Intensity distribution of the image of the cross grating in the case of coma with Zernike coefficient $a_C = 0.5\lambda$.

5.2.3.2 "Einstein" photo as object

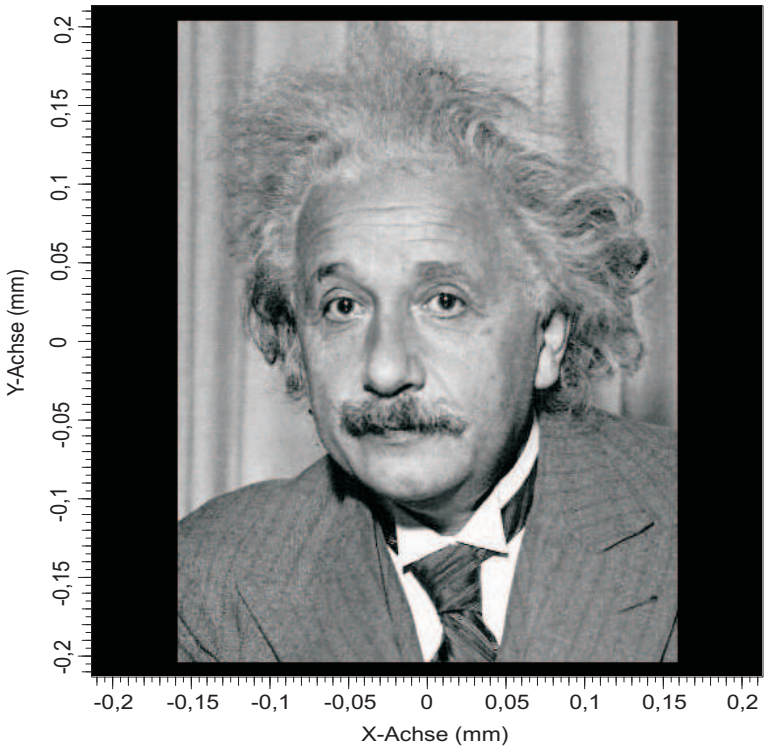


Figure 5.7: Original photo of Albert Einstein.

It is also interesting to demonstrate the influence of diffraction effects or aberrations in optical imaging (incoherent case) for the case of an object of daily life, like e.g. the image of a face. For this purpose the photo of a well–known physicist is taken (see fig. 5.7–5.12). The numerical aperture for the imaging is like in the case of the cross grating 0.075 (for a quadratic aperture by a factor $\sqrt{2}$ larger along the diagonal) and the wavelength is again $\lambda = 0.5$ μm. Therefore, the diameter of the Airy disc (circular aperture) is again about 8 μm. Additionally, it is assumed that the image of the photo has a diameter of 427 μm.

The diffraction–limited image (fig. 5.8 top left) with a circular aperture shows a kind of blurring compared to the object (fig. 5.7). However, the face can be clearly recognized. As in the case of the cross grating, the image with the annular aperture (fig. 5.8 top right) is deteriorated compared to the images of all other aperture forms. The images with quadratic apertures either parallel to the axes of the photo or rotated by 45 degree (fig. 5.8 bottom) give quite similar images like in the case of the circular aperture. An improvement of the image quality for the rotated quadratic aperture cannot be seen because a face does not only have structures with well–defined axes.

For all following cases with aberrations a circular aperture is used. In the case of spherical aberration the image in the paraxial focal plane (fig. 5.9 top) is quite blurred and low–contrast, whereas it has a higher contrast, but still blurred, in the plane of the best focus (fig. 5.9 bottom). In the case of astigmatism a kind of linear blurring along one of the axes can be seen in the image planes of the two focal lines (fig. 5.10 top left and bottom). In the intermediate image plane

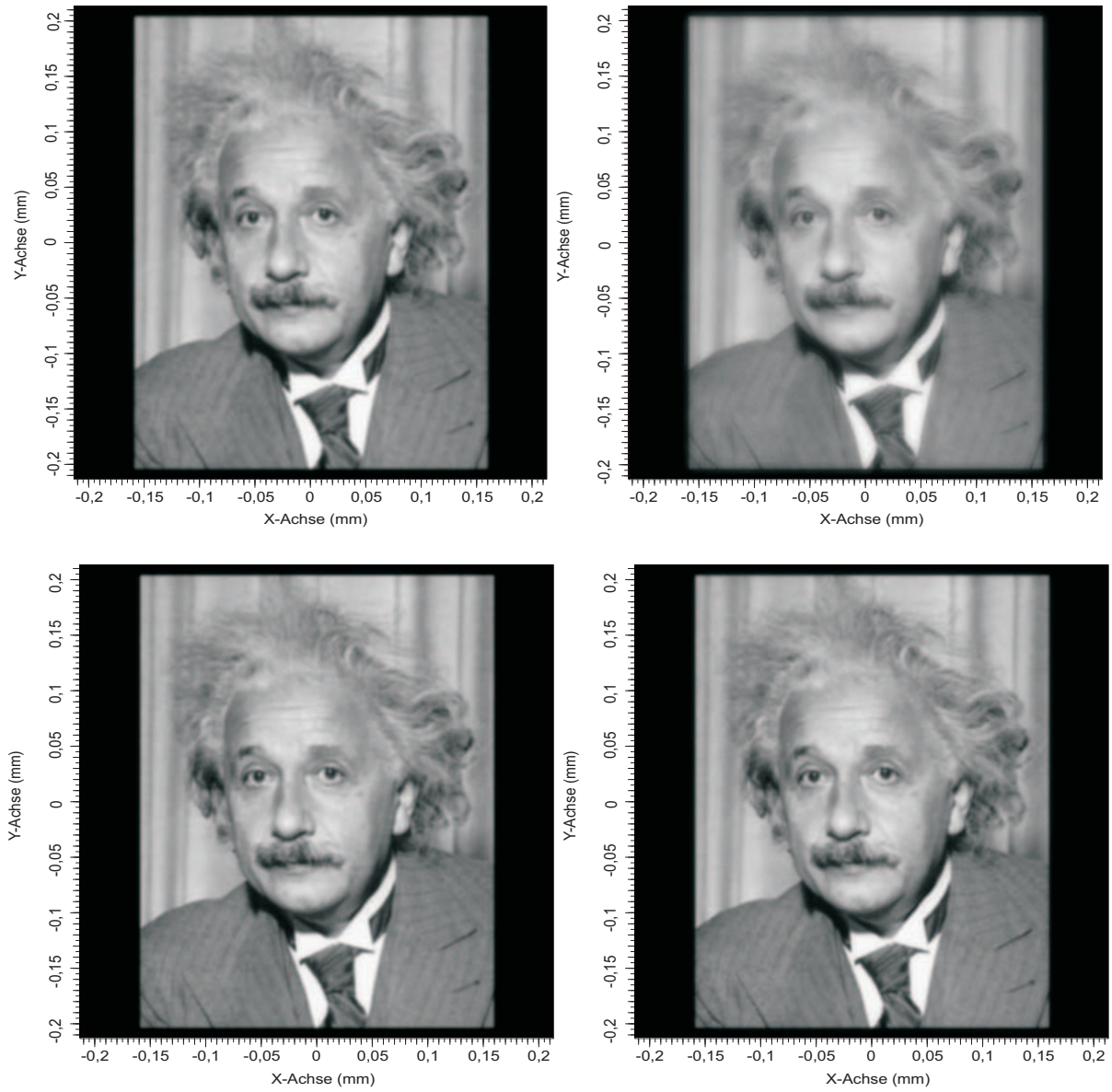


Figure 5.8: Diffraction–limited images of Einstein for different aperture forms, assuming a wavelength $\lambda = 0.5$ μm and a numerical aperture NA $= 0.075$. The corresponding point spread functions are shown in fig. 5.1. Furthermore, it is assumed that the image has a diameter of 427 μm. Top left: circular aperture, top right: annular aperture with 50% of the diameter opaque, bottom left: quadratic aperture with axes parallel to the edges of the photo, bottom right: quadratic aperture rotated by 45 degree.

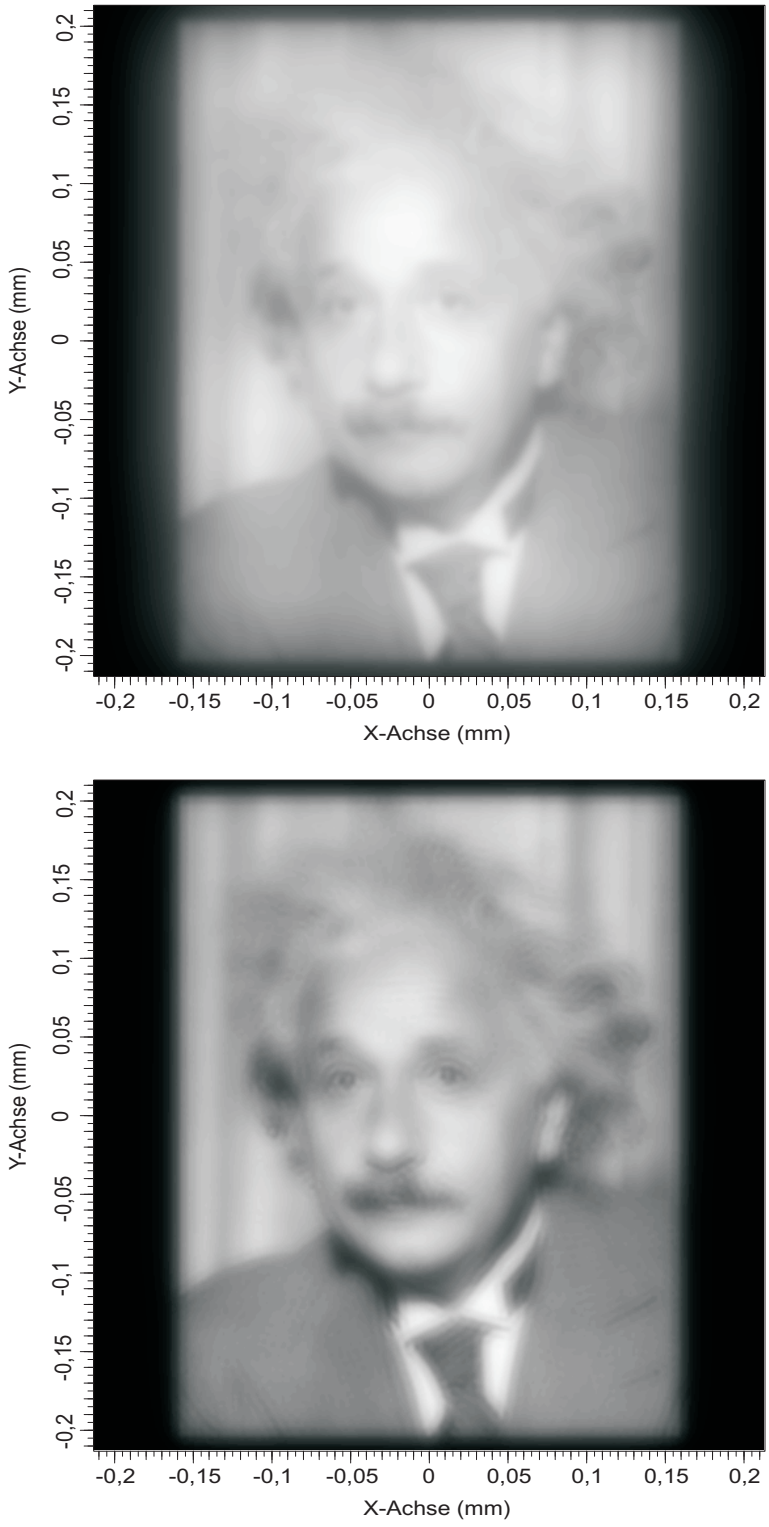


Figure 5.9: Images of Einstein in the case of spherical aberration with Zernike coefficient $a_S = 0.4\lambda$. Top: paraxial image plane, bottom: image plane with best focus (i.e. minimum wave aberrations).

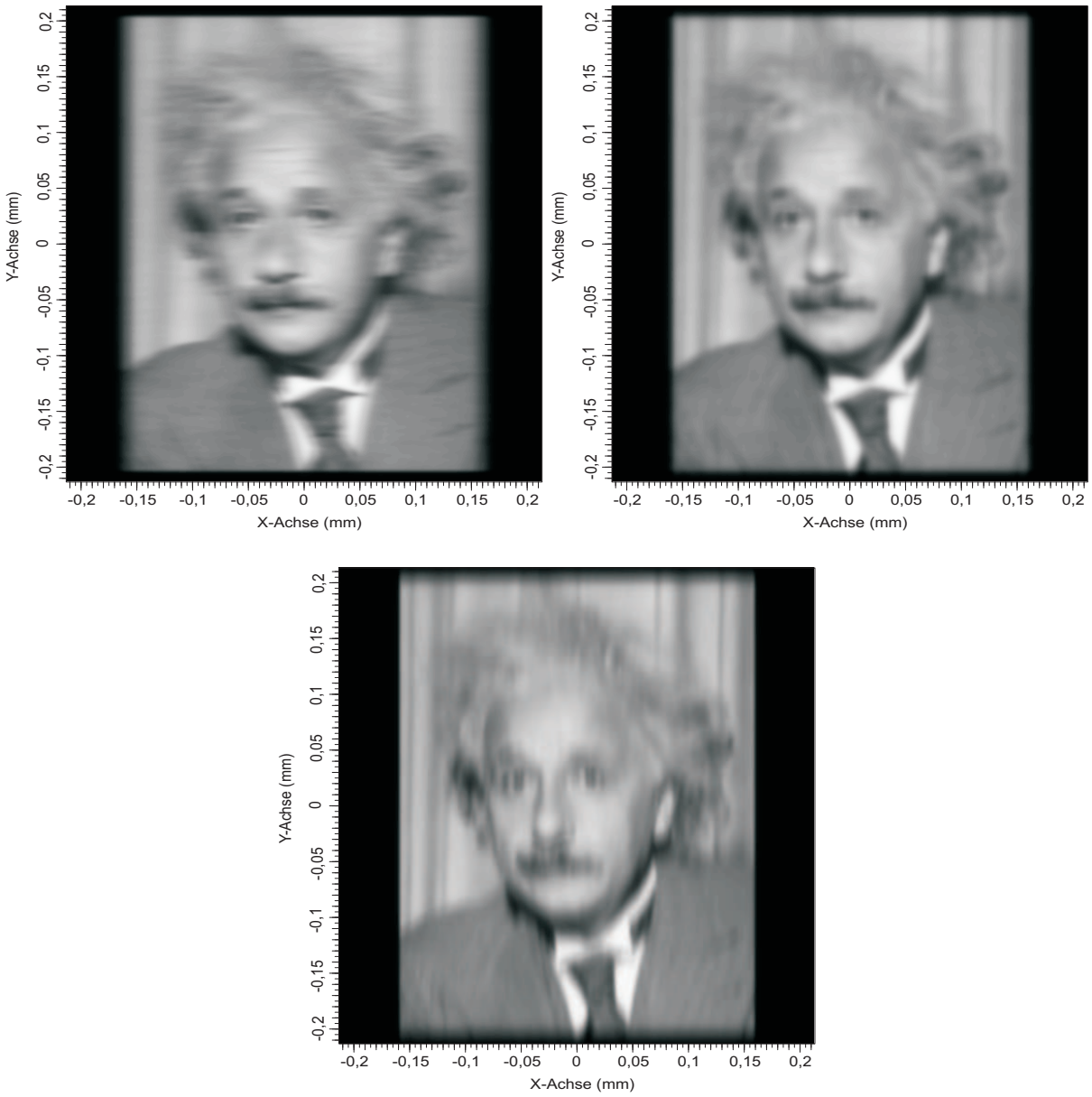


Figure 5.10: Images of Einstein in the case of astigmatism with Zernike coefficient $a_A = 0.5\lambda$. Top left: image in the plane of the first focal line, top right: image in the intermediate plane with circular spot, bottom: image in the plane of the second focal line.

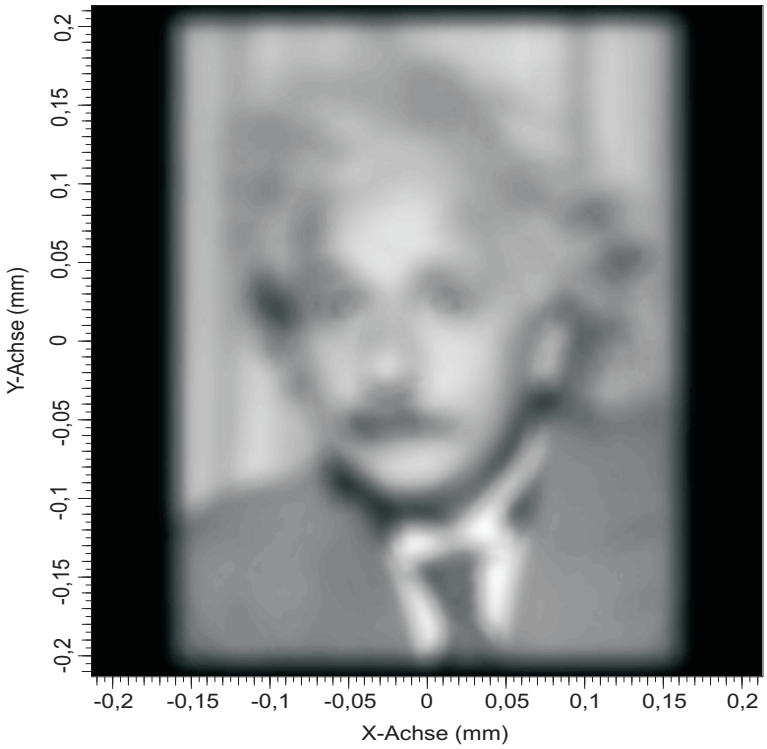


Figure 5.11: Image of Einstein in the case of defocus with Zernike coefficient $a_D = 0.5\lambda$.

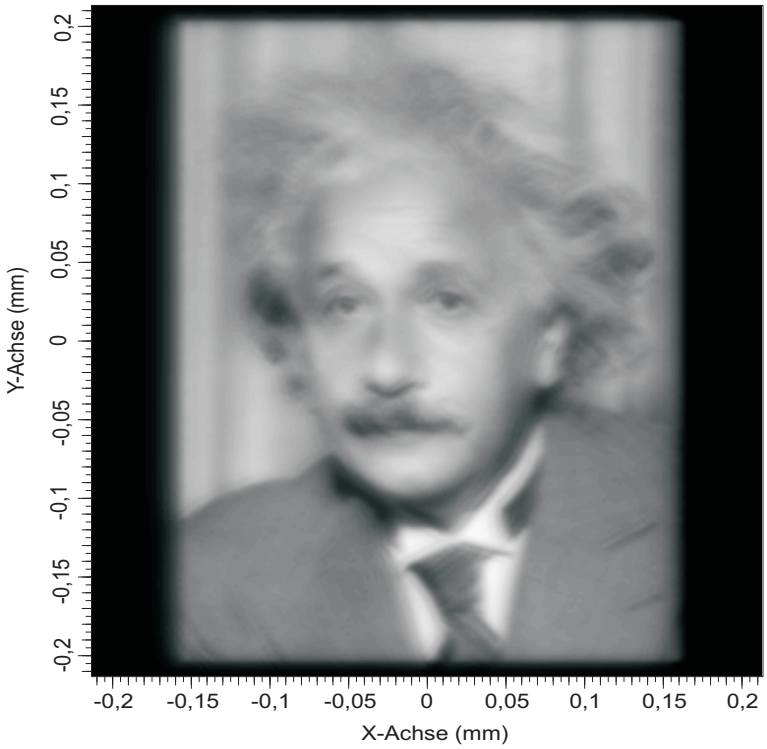


Figure 5.12: Image of Einstein in the case of coma with Zernike coefficient $a_C = 0.5\lambda$.

with circular spot only a normal blurring can be seen (fig. 5.10 top right). The same is valid in the case of pure defocus (fig. 5.11). In the case of coma (fig. 5.12) the image is a little bit less blurred, but still not quite good. In all cases, the Zernike coefficients are 0.5λ, respectively (with the exception of spherical aberration with only 0.4λ), using the definition for the normalization like in [1]. But, admittedly the same size of a Zernike coefficient does not automatically mean the same peak–to–valley or rms value of the aberrations. Therefore, also the influence on the image quality is quite different.

5.3 The optical transfer function

5.3.1 Definition of the OTF and MTF

The concept of the incoherent optical transfer function (OTF) and of the incoherent modulation transfer function (MTF) (*deutsch:* Kontrastübertragungsfunktion) can also be used to describe the incoherent imaging of extended objects.

Remark: The same concept can also be used in the coherent case. But, as we will see, the incoherent OTF is the inverse Fourier transform of the intensity impulse response or PSF. In analogy, the OTF in the coherent case is defined as the inverse Fourier transform of the amplitude impulse response which is according to equation (5.2.7) identical to the pupil function, besides a normalization factor and spatial frequencies as parameters instead of pupil coordinates. Therefore, the OTF in the coherent case is not very exciting.

In section 5.2.2 it has been shown that the intensity distribution I_I in the image plane is the convolution of the intensity distribution I_O in the object plane and the intensity impulse response or point spread function PSF of the optical imaging system. So, according to equation (5.2.12) it is:

$$I_I(x,y) = c \iint_{-\infty}^{+\infty} I_O(x',y') \cdot PSF(x-\beta x', y-\beta y') dx' dy' \tag{5.3.1}$$

Here, c is a normalization constant to ensure energy conservation. In equation (5.2.12) κ was used for this purpose, but the PSF may be normalized in another way as $|h_{lens}|^2$, so that here c is used. (x,y) and (x',y') are the coordinates in the image plane and object plane, respectively. β is again defined as the scaling factor between the coordinates in the image plane and the object plane.

Next, I_I, I_O and the PSF are written as Fourier integrals, respectively:

$$I_I(x,y) = \iint_{-\infty}^{+\infty} \widetilde{I}_I(\nu_x,\nu_y) e^{-2\pi i(\nu_x x + \nu_y y)} d\nu_x d\nu_y \tag{5.3.2}$$

$$I_O(x',y') = \iint_{-\infty}^{+\infty} \widetilde{I}_O(\nu_x,\nu_y) e^{-2\pi i(\nu_x \beta x' + \nu_y \beta y')} d\nu_x d\nu_y \tag{5.3.3}$$

$$PSF(x,y) = \iint_{-\infty}^{+\infty} \widetilde{PSF}(\nu_x,\nu_y) e^{-2\pi i(\nu_x x + \nu_y y)} d\nu_x d\nu_y \tag{5.3.4}$$

In the Fourier integral of I_O the spatial coordinates $(\beta x', \beta y')$ are used because these are the corresponding scaled coordinates in the image plane. This means that the intensity distribution

in the object plane is re–scaled in such a way as it would be imaged to the image plane if no diffraction occurs, i.e. assuming a PSF like a delta function.
Using the convolution theorem and some relations of Fourier mathematics the following conversions can be made:

$$
\begin{aligned}
&\widetilde{I}_I(\nu_x,\nu_y) = \iint\limits_{-\infty}^{+\infty} I_I(x,y)e^{2\pi i(\nu_x x+\nu_y y)}dxdy = \\
&= c\iint\limits_{-\infty}^{+\infty}\iint\limits_{-\infty}^{+\infty}\left(\iint\limits_{-\infty}^{+\infty}\widetilde{I}_O(\nu'_x,\nu'_y)e^{-2\pi i(\nu'_x\beta x'+\nu'_y\beta y')}d\nu'_x d\nu'_y\right)\cdot \\
&\cdot\left(\iint\limits_{-\infty}^{+\infty}\widetilde{PSF}(\nu''_x,\nu''_y)e^{-2\pi i\left[\nu''_x(x-\beta x')+\nu''_y(y-\beta y')\right]}d\nu''_x d\nu''_y\right)e^{2\pi i(\nu_x x+\nu_y y)}dx'dy'\,dxdy = \\
&= c\iint\limits_{-\infty}^{+\infty}\iint\limits_{-\infty}^{+\infty}\iint\limits_{-\infty}^{+\infty}\iint\limits_{-\infty}^{+\infty}\widetilde{I}_O(\nu'_x,\nu'_y)\widetilde{PSF}(\nu''_x,\nu''_y)\cdot \\
&\cdot e^{2\pi i\left[\beta(\nu''_x-\nu'_x)x'+(\nu_x-\nu''_x)x+\beta(\nu''_y-\nu'_y)y'+(\nu_y-\nu''_y)y\right]}dx'dy'\,dxdy\,d\nu'_x d\nu'_y\,d\nu''_x d\nu''_y = \\
&= c\iint\limits_{-\infty}^{+\infty}\iint\limits_{-\infty}^{+\infty}\widetilde{I}_O(\nu'_x,\nu'_y)\widetilde{PSF}(\nu''_x,\nu''_y)\,\delta\big(\beta(\nu''_x-\nu'_x),\beta(\nu''_y-\nu'_y)\big)\,\delta(\nu_x-\nu''_x,\nu_y-\nu''_y)d\nu'_x d\nu'_y\cdot \\
&\cdot d\nu''_x d\nu''_y = \\
&= \frac{c}{\beta^2}\widetilde{I}_O(\nu_x,\nu_y)\widetilde{PSF}(\nu_x,\nu_y)
\end{aligned}
$$

Summarizing we have:

$$\widetilde{I}_I(\nu_x,\nu_y) = \frac{c}{\beta^2}\widetilde{I}_O(\nu_x,\nu_y)OTF(\nu_x,\nu_y) \tag{5.3.5}$$

Here, the inverse Fourier transform of the point spread function PSF is called OTF, i.e. $OTF = \widetilde{PSF}$, and therefore, the OTF is defined as:

$$OTF(\nu_x,\nu_y) = \iint\limits_{-\infty}^{+\infty} PSF(x,y)e^{2\pi i(\nu_x x+\nu_y y)}dxdy \tag{5.3.6}$$

OTF is the shortcut for **optical transfer function** (*deutsch:* optische Übertragungsfunktion). The modulus of the optical transfer function is called the **modulation transfer function** MTF (*deutsch:* Kontrastübertragungsfunktion oder Modulationsübertragungsfunktion):

$$MTF(\nu_x,\nu_y) = |OTF(\nu_x,\nu_y)| = \left|\iint\limits_{-\infty}^{+\infty} PSF(x,y)e^{2\pi i(\nu_x x+\nu_y y)}dxdy\right| \tag{5.3.7}$$

In the following, it will be shown that the OTF is identical to the autocorrelation function of the pupil function.

Additionally, an interpretation of the concrete optical meaning of the OTF and of the MTF as giving the contrast or modulation with which a sinusoidal intensity variation with a certain spatial frequency is transferred by the optical system will be given. Then, it will also be clear why the OTF is also called **frequency response function**.

5.3.2 Interpretation of the OTF and MTF

The OTF can also be written as the autocorrelation function of the pupil function A. For this purpose, the definition of the OTF (5.3.6) and of the PSF (5.2.13) are used:

$$\begin{aligned}
OTF(\nu_x,\nu_y) &= \int_{-\infty}^{+\infty}\!\!\int PSF(x,y)e^{2\pi i(\nu_x x+\nu_y y)}dxdy = \\
&= \frac{1}{a_0}\int_{-\infty}^{+\infty}\!\!\int \left(\int_{-\infty}^{+\infty}\!\!\int A(x',y')e^{\frac{-2\pi i}{\lambda d_I}(xx'+yy')}dx'dy'\right)\cdot \\
&\quad \cdot\left(\int_{-\infty}^{+\infty}\!\!\int A^*(x'',y'')e^{\frac{2\pi i}{\lambda d_I}(xx''+yy'')}dx''dy''\right)e^{2\pi i(\nu_x x+\nu_y y)}dxdy = \\
&= \frac{1}{a_0}\int_{-\infty}^{+\infty}\!\!\int\int_{-\infty}^{+\infty}\!\!\int\int_{-\infty}^{+\infty}\!\!\int A(x',y')A^*(x'',y'')\cdot \\
&\quad \cdot e^{2\pi i\left[x\left(\nu_x+\frac{x''-x'}{\lambda d_I}\right)+y\left(\nu_y+\frac{y''-y'}{\lambda d_I}\right)\right]}dxdy\,dx'dy'\,dx''dy'' = \\
&= \frac{1}{a_0}\int_{-\infty}^{+\infty}\!\!\int\int_{-\infty}^{+\infty}\!\!\int A(x',y')A^*(x'',y')\,\delta\left(\nu_x+\frac{x''-x'}{\lambda d_I},\nu_y+\frac{y''-y'}{\lambda d_I}\right)dx'dy'\,dx''dy'' = \\
&= \frac{\lambda^2 d_I^2}{a_0}\int_{-\infty}^{+\infty}\!\!\int A(x',y')A^*(x'-\lambda d_I\nu_x,y'-\lambda d_I\nu_y)dx'dy'
\end{aligned}$$

Here, a_0 is a normalization constant for the PSF whose absolute value is not important for us. Finally, the OTF can be expressed as:

$$\begin{aligned}
OTF(\nu_x,\nu_y) &= c_{OTF}\int_{-\infty}^{+\infty}\!\!\int A(x',y')A^*(x'-\lambda d_I\nu_x,y'-\lambda d_I\nu_y)dx'dy' = \\
&= c_{OTF}\int_{-\infty}^{+\infty}\!\!\int A(x'+\lambda d_I\nu_x,y'+\lambda d_I\nu_y)A^*(x',y')dx'dy' \qquad (5.3.8)
\end{aligned}$$

Here, another constant c_{OTF} has been introduced. In most cases, the OTF is normalized in such a way, that it has the value one at the spatial frequency $\nu_x = 0, \nu_y = 0$. At this point, the OTF has also the maximum value of its modulus, since there the overlap of the pupil function with itself is at most.

If the pupil function has values different from zero only in a region with the radius r'_{max} (for a circular aperture r'_{max} is of course identical to the radius of the exit pupil), the OTF is zero for all spatial frequencies which fulfill the following conditions:

$$OTF(\nu_x, \nu_y) = 0 \quad \text{for} \quad \begin{array}{l} \lambda d_I |\nu_x| > 2r'_{max} \;\Rightarrow\; |\nu_x| > 2\dfrac{r'_{max}}{\lambda d_I} = 2\dfrac{\mathrm{NA}}{\lambda} \\ \lambda d_I |\nu_y| > 2r'_{max} \;\Rightarrow\; |\nu_y| > 2\dfrac{r'_{max}}{\lambda d_I} = 2\dfrac{\mathrm{NA}}{\lambda} \end{array} \tag{5.3.9}$$

Here, the numerical aperture $\mathrm{NA} = r'_{max}/d_I$ is used. Of course, the concept of a numerical aperture is first of all only useful for a circular aperture. But, it can also be extended to other aperture shapes. The quantity $2\mathrm{NA}/\lambda$ is called the cut–off frequency ν_{cut} of the incoherent imaging.
Since the OTF is the Fourier transform of a real–valued function (of the PSF), the following condition is valid:

$$OTF(-\nu_x, -\nu_y) = \int_{-\infty}^{+\infty}\!\!\int PSF(x,y) e^{2\pi i((-\nu_x)x + (-\nu_y)y)} dx dy = OTF^*(\nu_x, \nu_y) \tag{5.3.10}$$

For this reason, also the MTF, i.e. the modulus of the OTF, fulfills the well–known condition:

$$MTF(\nu_x, \nu_y) = |OTF(\nu_x, \nu_y)| = |OTF^*(\nu_x, \nu_y)| = |OTF(-\nu_x, -\nu_y)| = MTF(-\nu_x, -\nu_y) \tag{5.3.11}$$

This means that the MTF is symmetric to an inversion at the origin of the spatial frequency domain. Therefore, it is sufficient to display in a section of the MTF e.g. only the part with the positive spatial frequencies.
Fig. 5.13 demonstrates graphically how the OTF/MTF can be calculated in the case of imaging with an aberration–free optical system with homogeneous illumination. In this case, the pupil function is constant inside of the aperture (e.g. normalized to one) and zero outside of the aperture. Then, the autocorrelation function of the pupil function is proportional to the surface area of the overlap between the two laterally shifted copies of the aperture. Therefore, the section of the OTF/MTF along one of the axes (ν_x or ν_y) is a straight line for a quadratic aperture (fig. 5.13 left). In the case of a circular aperture, the surface area of the overlap increases faster than linear with increasing overlap (at least at the beginning). Therefore, the curve of fig. 5.13 (bottom right) results for the OTF/MTF. For the simulation of the OTF/MTF an optical system with a numerical aperture $\mathrm{NA} = 0.1$ and a wavelength $\lambda = 0.5\ \mu$m has been used. Therefore, the cut–off frequency is $\nu_{cut} = 2\mathrm{NA}/\lambda = 400/\mathrm{mm}$. For spatial frequencies, which have a smaller modulus as the cut–off frequency, the overlap of the copies of the pupil function start and the OTF/MTF has values different from zero (at least in a section along the axes ν_x or ν_y of the coordinate system).
For interpreting the meaning of the OTF or MTF an object with a sinusoidal intensity variation with the period p is taken:

$$I_O(x,y) = I_0\left(1 + K\cos\left(2\pi\frac{x}{p}\right)\right) = I_0\left(1 + \frac{K}{2}\left(e^{2\pi i x/p} + e^{-2\pi i x/p}\right)\right) \tag{5.3.12}$$

Here, I_0 is the medium intensity and K is the contrast or modulation, since we have:

$$\text{contrast} := \frac{I_{O,max} - I_{O,min}}{I_{O,max} + I_{O,min}} = \frac{I_0(1+K) - I_0(1-K)}{I_0(1+K) + I_0(1-K)} = K \tag{5.3.13}$$

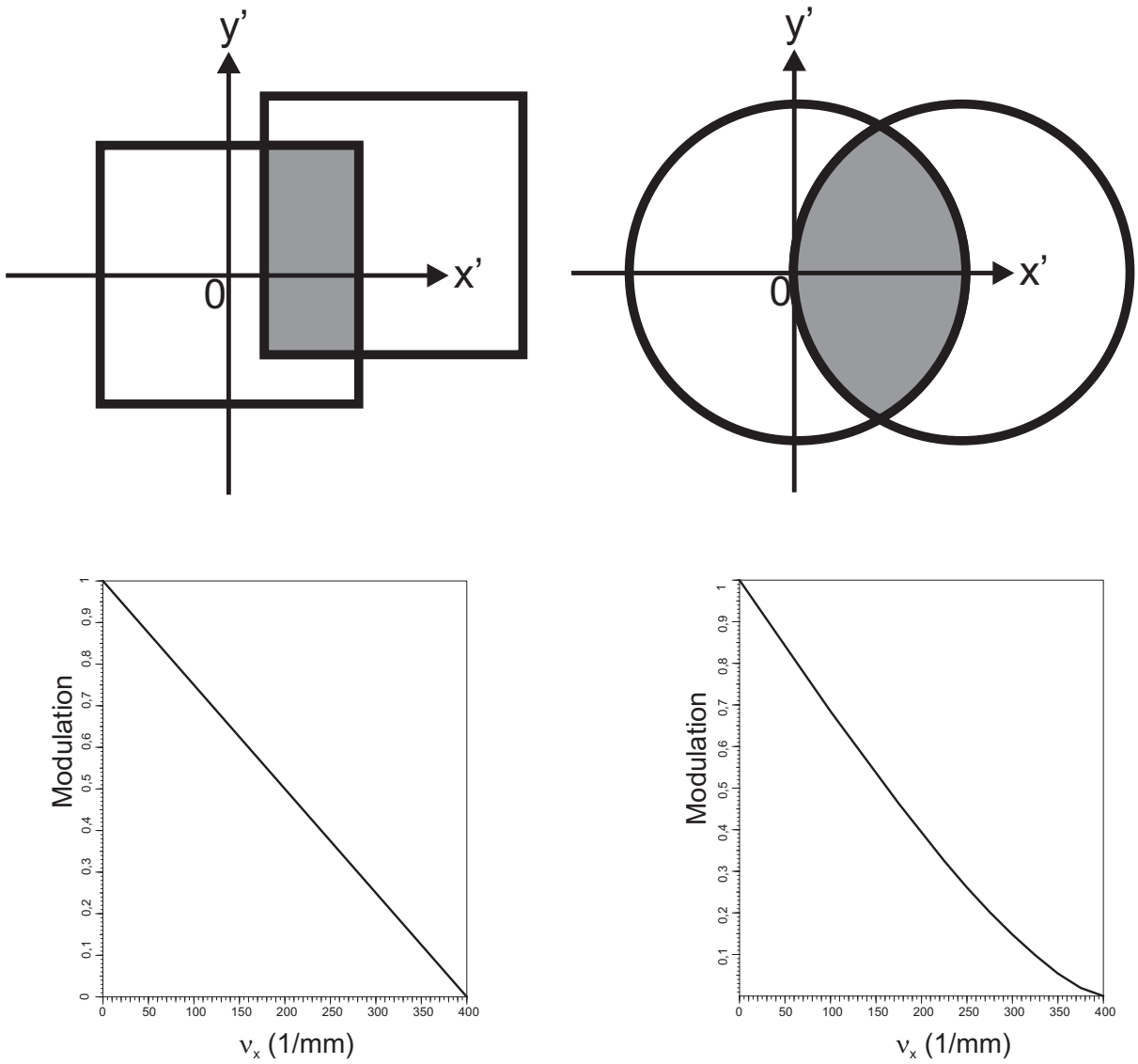


Figure 5.13: Interpretation of the OTF/MTF as autocorrelation function of the pupil function. In this scheme it is displayed for an aberration–free optical system with homogeneous illumination. Then, the surface area of the overlap between both laterally shifted copies of the pupil function is proportional to the OTF/MTF. Left: quadratic aperture, right: circular aperture. Top figures: scheme demonstrating the formation of the overlap, bottom figures: section of the MTF along the ν_x–axis.

The Fourier transform of I_O in the spatial frequency domain is:

$$\widetilde{I}_O(\nu_x,\nu_y) = I_0\delta(\nu_y)\left[\delta(\nu_x) + \frac{K}{2}\left(\delta\left(\nu_x + \frac{1}{p}\right) + \delta\left(\nu_x - \frac{1}{p}\right)\right)\right] \tag{5.3.14}$$

According to our convention the OTF is normalized in such a way that it is valid $OTF(\nu_x = 0, \nu_y = 0) = 1$ and $OTF(\nu_x = 1/p, \nu_y = 0) = a$ with $|a| < 1$. Then, the Fourier transform of the intensity distribution in the image plane I_I is according to equation (5.3.5):

$$\widetilde{I}_I(\nu_x,\nu_y) = \frac{c}{\beta^2}I_0\left[\delta(\nu_x)\delta(\nu_y) + \frac{K}{2}\left(a^*\delta\left(\nu_x + \frac{1}{p}\right) + a\delta\left(\nu_x - \frac{1}{p}\right)\right)\delta(\nu_y)\right] \tag{5.3.15}$$

Thus, the intensity distribution in the image plane I_I can be calculated using equation (5.3.2) and the relation $a := |a|\exp(i\varphi_a)$:

$$I_I(x,y) = \frac{c}{\beta^2}I_0\left[1 + \frac{K}{2}\left(a^*e^{2\pi ix/p} + ae^{-2\pi ix/p}\right)\right] = \frac{c}{\beta^2}I_0\left[1 + |a|K\cos\left(2\pi\frac{x}{p} - \varphi_a\right)\right] \tag{5.3.16}$$

Therefore, the OTF has the following meaning:

- The modulus $|a|$ of the OTF at a certain spatial frequency $1/p$, i.e. the function value of the MTF, gives the change of the modulation of a sinusoidal intensity variation of the period p during the imaging process. This explains the name modulation transfer function of the MTF.
- The phase of the OTF φ_a is responsible for a phase shift of the image compared to the object. This means for example that there is an inversion of the modulation in the case $\varphi_a = \pi$. Then, the image is dark at a point where the corresponding point of the object is bright and vice versa.

The inversion of the contrast can be seen very well in the case of the imaging of a so called Siemens star (see fig. 5.14). A Siemens star is a radially symmetric composition of spokes which are either black or white. The width of the spokes increases with increasing radius. Since the period is proportional to the radius coordinate r, there are small spatial frequencies (i.e. large periods) at the rim and high spatial frequencies towards the center ($1/p$ is proportional to $1/r$). In the case, that an optical system causes due to aberrations or defocussing an inversion of the contrast for certain spatial frequencies, a sequence of dark and bright regions can be seen along a spoke of the image of the Siemens star (see fig. 5.14, right).
As we have seen before, each optical system (also an ideal aberration–free optical system) has a cut–off frequency $\nu_{cut} = 2\mathrm{NA}/\lambda$ of the OTF or MTF. This means, that spatial frequencies which are higher than the cut–off frequency are not transferred by the optical system. Accordingly, the smallest period of a sinusoidal intensity variation, which can just be transferred (although in practice with a very small modulation), is:

$$p_{min} = \frac{1}{\nu_{cut}} = \frac{\lambda}{2\mathrm{NA}} \tag{5.3.17}$$

This is the limit of resolution of an (aberration–free) optical imaging system with the numerical aperture NA which is illuminated by incoherent light of the wavelength λ.

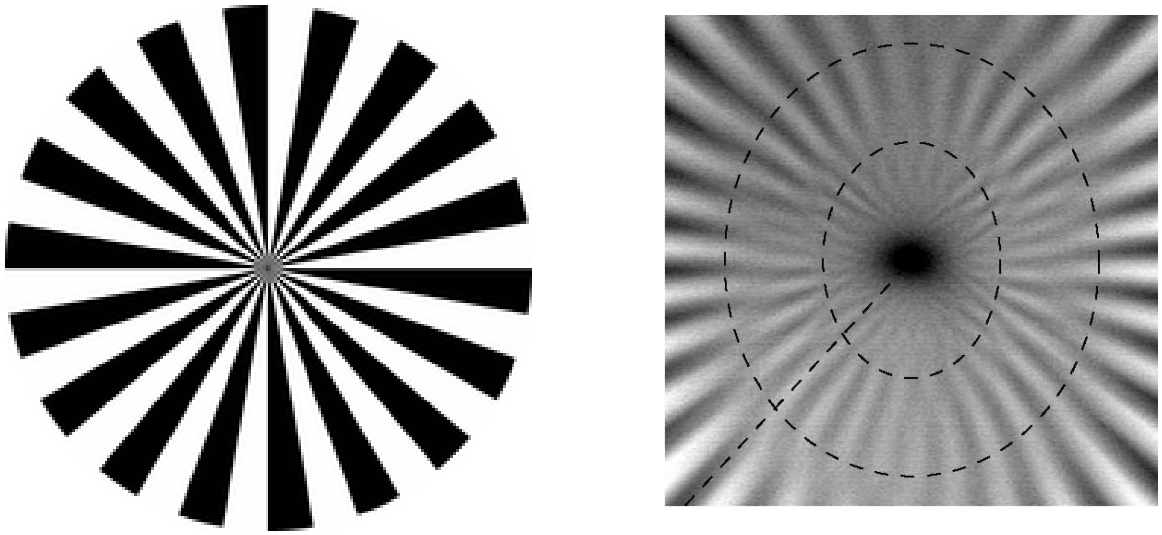


Figure 5.14: Left: schematic representation of a so called Siemens star, right: example of the imaging of a Siemens star with inversion of the contrast along the spokes (courtesy of www.informatik.hu–berlin.de/∼blaschek/studivortrag/studi.pdf).

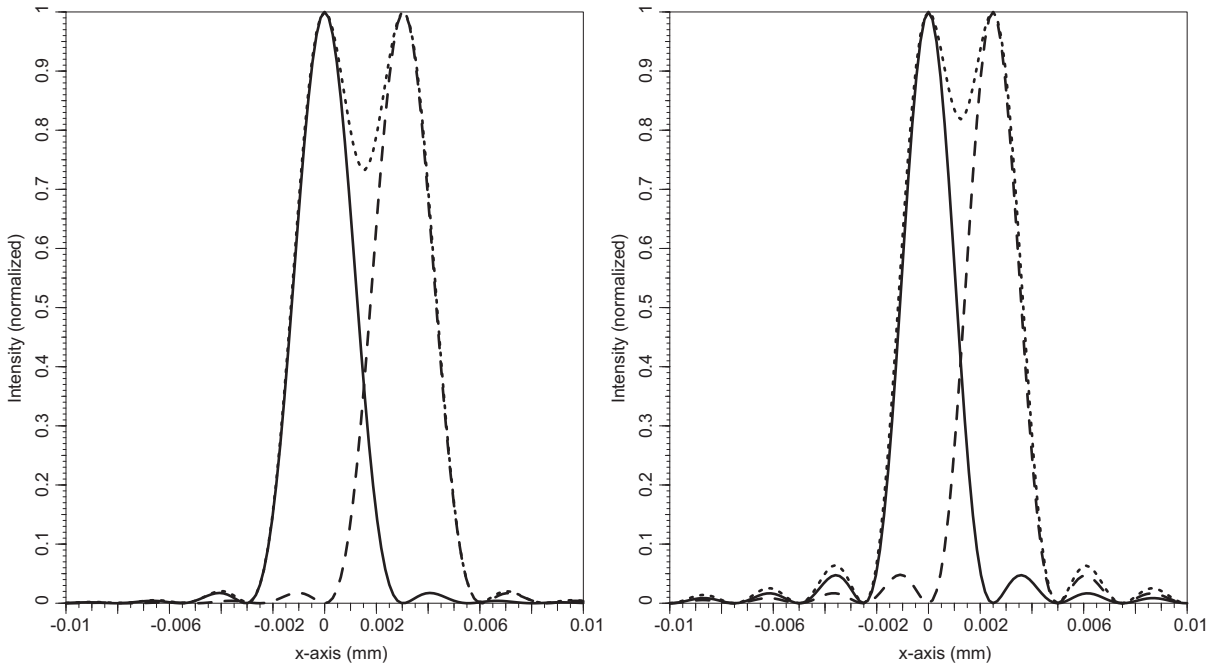


Figure 5.15: Intensity distribution of the images of two incoherent point sources (solid line and dashed line, respectively) illustrating the Rayleigh criterion of resolution. The numerical aperture of the imaging system is 0.1 and the wavelength is 0.5 μm. Left: circular exit pupil, right: quadratic exit pupil. The dotted line shows the sum of the intensities of the two image points, respectively.

Of course, this value is nearly identical to the resolution limit which is obtained by using the **Rayleigh criterion**. The Rayleigh criterion of resolution states that two incoherent point sources of equal intensity can be resolved if the first minimum of one of the image points coincides with the central maximum of the second image point.
In the case of a circular pupil this means, that the distance of the image points has to be larger or equal to $0.61\lambda/\mathrm{NA}$. Fig. 5.15 left shows a simulation for $\lambda = 0.5$ μm and NA $= 0.1$. This means that the minimum distance between the two image points has to be 3.05 μm and the dip in the sum of the intensities is 0.735 of the maximum value.
For a quadratic aperture with the numerical aperture NA in one direction the distance between the image points has to be larger or equal to $0.5\lambda/\mathrm{NA}$. Fig. 5.15 right shows a simulation for $\lambda = 0.5$ μm and NA $= 0.1$. This means that the minimum distance between the two image points has to be 2.5 μm and the dip in the sum of the intensities is 0.811 of the maximum value.

5.4 Optical filtering

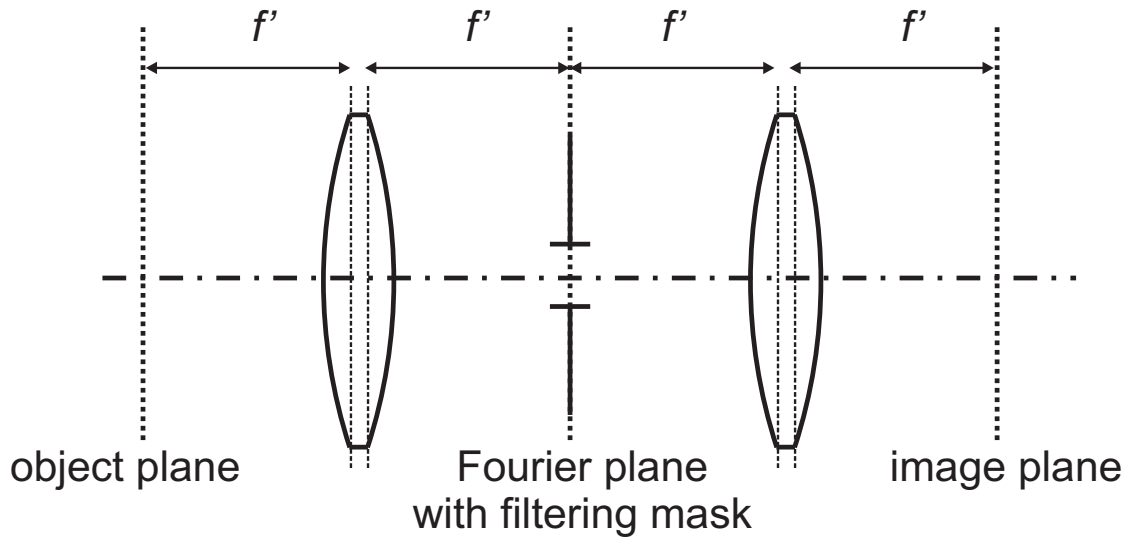


Figure 5.16: Scheme of a 4f–system which is used for optical filtering. There is an optical filtering mask in the Fourier plane which is here symbolized as a stop.

Now, after having treated the imaging with incoherent light we switch back to imaging with coherent light.
A useful application of the Fourier transform property of a lens is for example the optical filtering in image processing, where the amplitude and phase of the different spatial frequency components of an object can be tailored. As we have seen, a lens produces the Fourier transform of an object in its back focal plane if the object is situated in the front focal plane of the lens with focal length f'. For the filtering, a stop or an optical filtering mask is placed in the Fourier plane. By adding a second lens with equal focal length f' in such a way that this second lens is placed in a distance f' behind the back focal plane of the first lens, there is a filtered image of the original object in the distance f' behind this second lens (see fig. 5.16). The telescopic system with distance f' between the object and the first lens, distance $2f'$ between the lenses, and distance f' between the second lens and the image is also called a 4f–system (or 4f'–system).

In many cases, the focal lengths of the two lenses of the telescope will not have the identical values f', but will have different values f'_1 and f'_2. Then, the distance between the two lenses has to be $f'_1 + f'_2$ and the filtering plane is in a distance f'_1 behind the first lens. But, this will only introduce a different scaling factor in the image plane and so we will treat here only the case of a real 4f–system.
The complex amplitude $u(x, y, 4f')$ in the image plane, neglecting the finite aperture of the lenses, is calculated by using equation (5.1.20) two times and taking into account the transmission function t of the filtering mask in the Fourier plane of the 4f–system:

$$\begin{aligned} u(x,y,4f') &= \frac{-1}{\lambda^2 f'^2} e^{i\frac{8\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} t(x',y') \left(\iint_{-\infty}^{+\infty} u_0(x'',y'')\, e^{-2\pi i \frac{x'x''+y'y''}{\lambda f'}}\, dx''dy'' \right) \cdot \\ &\quad \cdot e^{-2\pi i \frac{xx'+yy'}{\lambda f'}}\, dx'dy' = \\ &= \frac{-1}{\lambda^2 f'^2} e^{i\frac{8\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} u_0(x'',y'') \cdot \\ &\quad \cdot \left(\iint_{-\infty}^{+\infty} t(x',y')\, e^{-2\pi i \frac{(x+x'')x'+(y+y'')y'}{\lambda f'}}\, dx'dy' \right) dx''dy'' = \\ &= \frac{-1}{\lambda^2 f'^2} e^{i\frac{8\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, \widetilde{t}\left(\frac{x+x''}{\lambda f'}, \frac{y+y''}{\lambda f'} \right) dx''dy'' \end{aligned} \tag{5.4.1}$$

Here, $\widetilde{t}$ is the Fourier transform of t.
Totally, the complex amplitude $u(x, y, 4f')$ behind the system is a convolution (or correlation) of the complex amplitude u_0 in front of the system and the Fourier transform $\widetilde{t}$ of the transmission function of the filtering mask.
In the special case, that no filtering occurs, i.e. $t = 1$ over the whole aperture, there would be a delta function for $\widetilde{t}$ and the result would be:

$$\begin{aligned} u(x,y,4f') &= \frac{-1}{\lambda^2 f'^2} e^{i\frac{8\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, \delta\left(\frac{x+x''}{\lambda f'}, \frac{y+y''}{\lambda f'} \right) dx''dy'' = \\ &= -e^{i\frac{8\pi f'}{\lambda}} \iint_{-\infty}^{+\infty} u_0(x'',y'')\, \delta(x+x'', y+y'')\, dx''dy'' = \\ &= -e^{i\frac{8\pi f'}{\lambda}}\, u_0(-x,-y) \end{aligned} \tag{5.4.2}$$

Apart from the constant phase factor, this would be the complex amplitude u_0, but, with reversed coordinates because of the scaling factor $\beta = -1$ of the telescope. Of course, this is only the case if diffraction effects at the finite apertures of the lenses can be neglected. Otherwise, it would be a convolution between u_0 and the point spread function of the lenses.

Of practical interest is the case, that there is really optical filtering. In most cases, the calculation can only be made by numerical simulations. In the following, some examples are shown.

5.4.1 Clipping of the spatial frequency spectrum

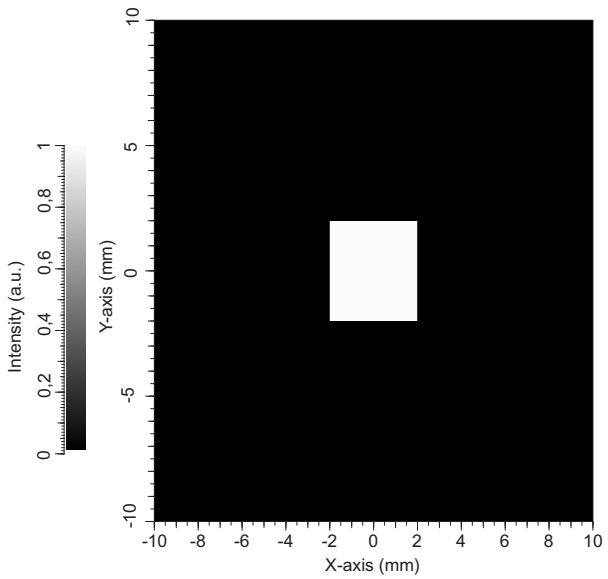


Figure 5.17: Intensity distribution of the amplitude object in the object plane of the 4f–system for optical filtering.

As first example, the effect of complete blocking of certain spatial frequency components will be illustrated. The object is a central quadratic transparent hole with 4 mm diameter in an opaque screen (see fig. 5.17). The screen has in the simulation a diameter of 20 mm and is illuminated by a plane wave with wavelength $\lambda = 0.5$ μm. The number of sampling points in the simulation is $N_x = N_y = 1024$. Therefore, the complex amplitude in the object plane is:

$$u_0(x,y) = \begin{cases} 1 & \text{for} \quad |x| \leq 2\text{mm} \wedge |y| \leq 2\text{mm} \\ 0 & \text{otherwise} \end{cases} \tag{5.4.3}$$

In the Fourier plane of the 4f–system (focal length f' of the lenses: $f' = 100$ mm) different stops are inserted. First, a pinhole with a transparent circular area of $D_{pinhole} = 100$ μm diameter is used (see fig. 5.18 left). This means, that spatial frequency components ν_x, ν_y with a modulus of larger than ν_{max} are completely blocked and the result for ν_{max} is:

$$\nu_{max} = \frac{D_{pinhole}}{2f'\lambda} = 1 \text{ mm}^{-1}$$

Fig. 5.18 shows in the right picture the case of a pinhole with $D_{pinhole} = 40$ μm. Then, the maximum spatial frequency is $\nu_{max} = 0.4$ mm^{-1}. In both cases the edges are quite blurred (especially in the second case). Additionally, a periodic intensity modulation with a period of 1 mm (fig. 5.18 left) or 2.5 mm (fig. 5.18 right) can be clearly seen in both simulations. Of course, this intensity modulation is a result of the maximum spatial frequency which is transmitted by the optical filtering system.

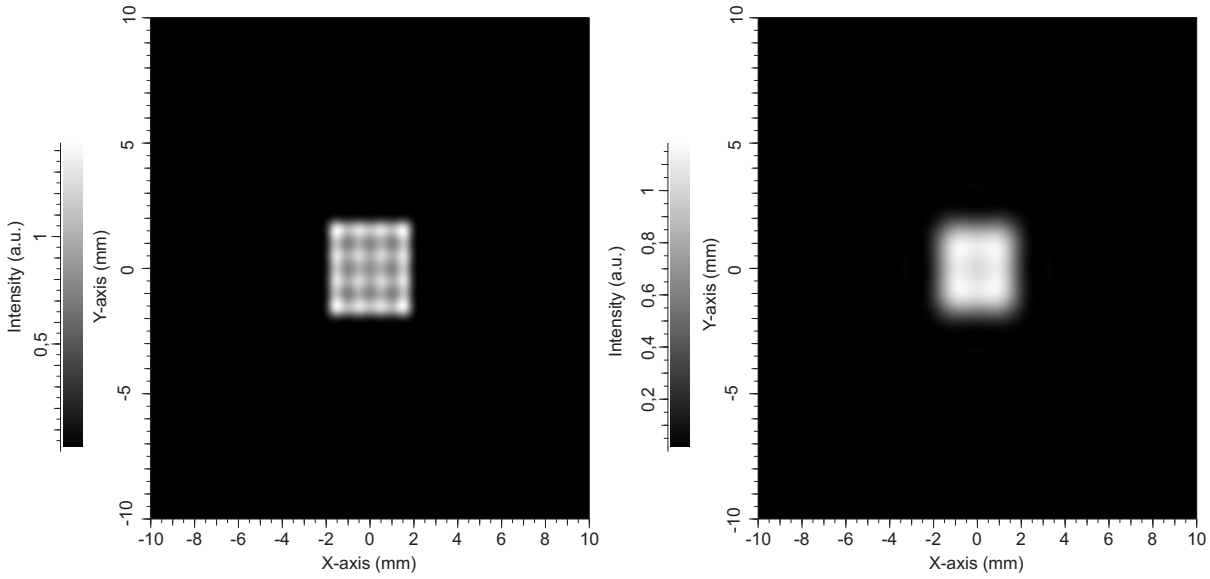


Figure 5.18: Intensity distribution of the filtered amplitude object in the image plane of the optical filtering system by using pinholes in the Fourier plane with 100 μm (left) or 40 μm (right) diameter, respectively.

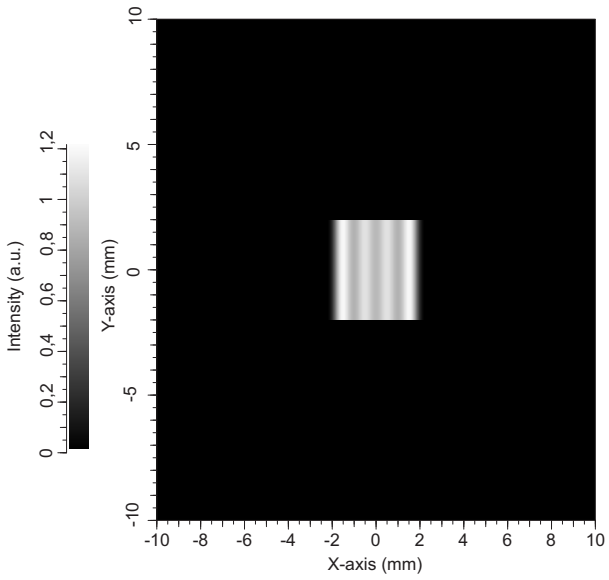


Figure 5.19: Intensity distribution of the filtered amplitude object in the image plane of the optical filtering system by using a slit aperture in the Fourier plane with 100 μm diameter in x–direction (and infinity in y–direction).

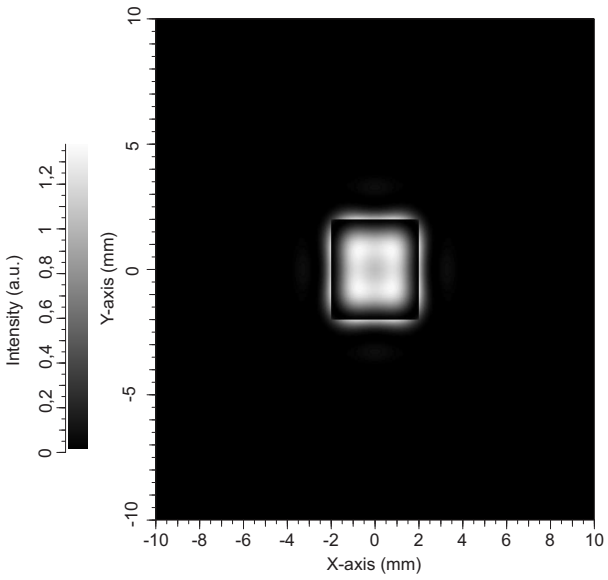


Figure 5.20: Intensity distribution of the filtered amplitude object in the image plane of the optical filtering system by using a phase mask in the Fourier plane with 40 μm diameter and a phase delay of π. The other parts of the complex amplitude in the Fourier plane are not changed.

As second example, a slit aperture with 100 μm diameter in x–direction is inserted in the Fourier plane of the filtering system. Fig. 5.19 shows that only spatial frequency components in x–direction with $|\nu_x| > 1$ mm^{-1} are blocked, whereas there is no blocking of spatial frequency components in y–direction. Therefore, the image has sharp edges in y–direction, but blurred edges in x–direction.
Up to now, only amplitude masks were inserted in the Fourier plane of the filtering system. But finally, the effect of a phase mask is demonstrated. Figure 5.20 shows the result if the filtering mask is a glass plate with a small central glass cylinder of 40 μm diameter and a height which corresponds to a phase delay of only π. In the center of the image a similar effect as in the case of the pinholes can be observed. But, the edges of the object can be clearly seen as dark structures in a bright background. The edges themselves can be clearly seen because no spatial frequency components are blocked.

5.4.2 The phase contrast method of Zernike

In the last section, a pure amplitude object has been treated which changes the amplitude of the incident wave by absorbing light. Here, a pure phase object will be considered which can for example be fabricated by etching a cavity into a glass substrate. The transmission function F of this object is

$$F(x', y') = e^{i\phi(x', y')} \tag{5.4.4}$$

with the real phase function ϕ.
This means that the object does not absorb any light, but just changes the phase or optical path length of the transmitted light by ϕ. Therefore, by imaging this object with an optical system onto a detector nearly no structure will be visible because optical detectors can only detect the intensity (i.e. amplitude), but not the phase. In practice, there may be something visible, but

with a very low contrast because each optical system blocks some spatial frequencies (especially those beyond the limit given by the numerical aperture of the system).
A method to observe especially small phase modulations of an object (without using interferometry) is the **phase contrast method of Zernike** [1] which was invented in 1935. There, a phase plate is included in the Fourier plane of the 4f–system. The phase plate is made in such a way that the central spatial frequencies near zero are changed in their phase by α. In the general case, it is not only a pure phase plate but it can also alter the amplitude of the spatial frequencies near zero by a factor $a < 1$. So, in total the transmission function t of the filtering mask has the following form:

$$t(\nu_x, \nu_y) = \begin{cases} ae^{i\alpha} & \text{for} \quad \nu_x^2 + \nu_y^2 \leq \epsilon \\ 1 & \text{in all other cases} \end{cases} \tag{5.4.5}$$

Here, ϵ is a quite small spatial frequency near zero.
The complex amplitude u_0' behind the object, which is illuminated by a plane on–axis wave with wavelength λ (i.e. u_0=const.=C), is of course

$$u_0'(x', y') = u_0 F(x', y') = CF(x', y') = C\left[\big(F(x', y') - 1\big) + 1\right] \tag{5.4.6}$$

At the end of this equation, we made a separation into light with complex amplitude $C(F-1)$, which is diffracted by the object, and non–diffracted light with complex amplitude C. Now, the non–diffracted light is a plane wave with spatial frequency ($\nu_x = 0, \nu_y = 0$) and will be focussed in the Fourier plane of the 4f–system to an on–axis point, where the transmission function of the filtering mask has the value $a\exp(i\alpha)$. On the other side, nearly all of the diffracted light will be focussed in the Fourier plane of the 4f–system to off–axis points, where the transmission function of the filtering mask is 1. That small part of the diffracted light, which is focussed in the Fourier plane to points with transmission function $a\exp(i\alpha)$, can be neglected. So, it is clear that the complex amplitude in the image plane of the 4f–system will be with a good approximation:

$$u_I(x, y) = C_I\left[(F(-x, -y) - 1) + ae^{i\alpha}\right] \tag{5.4.7}$$

The inversion of the coordinates is due to the scaling factor -1 of the 4f–system. The absolute value of the constant C_I is of no interest because we are only interested in variations of the intensity. Therefore, the intensity in the image plane is (with $|C_I|^2 = I_0$):

$$\begin{aligned} I_I(x, y) &= u_I(x, y)u_I^*(x, y) = I_0\left[e^{i\phi(-x,-y)} - 1 + ae^{i\alpha}\right]\left[e^{-i\phi(-x,-y)} - 1 + ae^{-i\alpha}\right] = \\ &= I_0\left[2 + a^2 - 2\cos\phi(-x, -y) + 2a\cos\left(\phi(-x, -y) - \alpha\right) - 2a\cos\alpha\right] = \\ &= I_0\left[a^2 + 2\left(1 - \cos\phi(-x, -y) + a\cos\left(\phi(-x, -y) - \alpha\right) - a\cos\alpha\right)\right] \end{aligned} \tag{5.4.8}$$

For phase objects with small phase variations $\phi \ll 2\pi$ we can set $\cos\phi \approx 1$ and $\cos(\phi - \alpha) = \cos\phi\,\cos\alpha + \sin\phi\,\sin\alpha \approx \cos\alpha + \phi\,\sin\alpha$. Then, equation (5.4.8) results in:

$$I_I(x, y) = I_0\left(a^2 + 2a\phi(-x, -y)\,\sin\alpha\right) = a^2 I_0\left(1 + 2\frac{\phi(-x, -y)}{a}\sin\alpha\right) \tag{5.4.9}$$

This equation shows that the intensity in the image plane increases or decreases linearly with the phase modulation ϕ and that the contrast for small phase variations ϕ can be increased by absorbing some of the non–diffracted light with $a < 1 \;\Rightarrow\; \phi/a > \phi$.

For the most common case, with a phase shift $\alpha = \pm\pi/2$ of the filtering mask and $a = 1$, the result is:

$$I_I(x,y) = a^2 I_0 \left(1 \pm 2\frac{\phi(-x,-y)}{a}\right) = I_0\left(1 \pm 2\phi(-x,-y)\right) \tag{5.4.10}$$

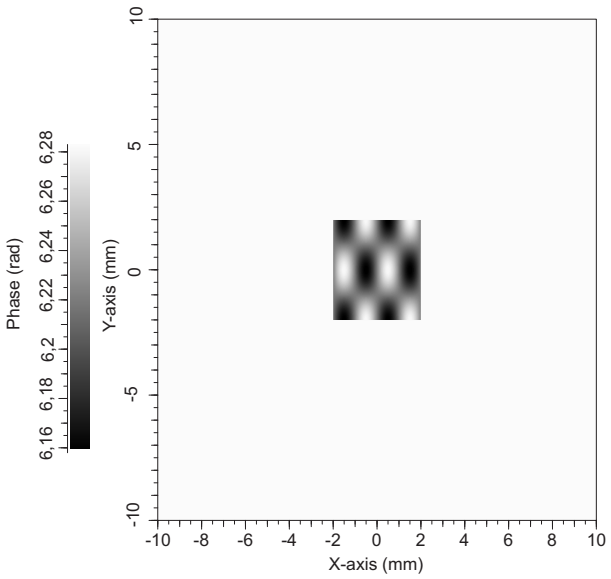


Figure 5.21: Phase distribution of the phase object in the object plane of the 4f–system for optical filtering.

Fig. 5.21 shows the phase distribution of a phase object, which has a sinusoidally modulated surface profile (made of glass with refractive index $n = 1.5$) with an amplitude of the modulation of $d = 10$ nm. This means that the amplitude of the phase modulation is $\Phi_0 = 2\pi(n-1)d/\lambda = \pi/50 = 0.0628 \ll 2\pi$ (the wavelength λ is again 0.5 μm). The peak to valley value $\Phi_{PV} = \pi/25 = 0.1257$ is of course by a factor two higher.
Figure 5.22 shows the numerically simulated intensity distribution in the image plane of the 4f–system, if the filtering mask in the Fourier plane has the following parameters.
Left figure:

$$t(x'',y'') = ae^{i\alpha} = \begin{cases} -i & \text{for} \quad \sqrt{x''^2 + y''^2} \le 2\,\mu\text{m, i.e. } a = 1,\ \alpha = -\pi/2 \\ 0 & \text{otherwise} \end{cases}$$

Right figure:

$$t(x'',y'') = ae^{i\alpha} = \begin{cases} -0.25i & \text{for} \quad \sqrt{x''^2 + y''^2} \le 2\,\mu\text{m, i.e. } a = 0.25,\ \alpha = -\pi/2 \\ 0 & \text{otherwise} \end{cases}$$

In fact this means that in the numerical simulation only the sampling point with spatial frequency zero is altered because the field in the simulation has $D = 20$ mm diameter, wavelength $\lambda = 0.5$ μm and focal lengths of the lenses $f' = 100$ mm. Therefore, the spatial frequency interval $\Delta\nu$ is $\Delta\nu = 1/D = 0.05$ mm^{-1} and this corresponds to a lateral distance of $\Delta x'' = f'\lambda\,\Delta\nu = 2.5$ μm in the Fourier plane of the 4f–system.

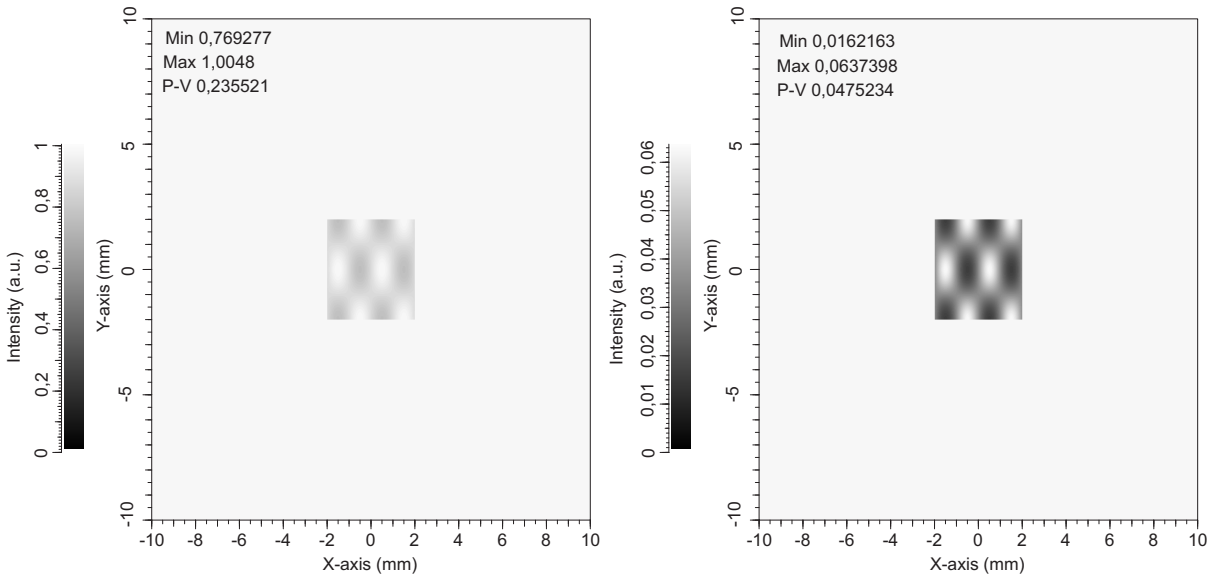


Figure 5.22: Intensity distribution in the image plane of the 4f–system using the phase contrast method of Zernike. In the left figure a pure phase plate (i.e. no absorption) with a phase delay of $\pi/2$ for spatial frequency components near zero is inserted in the Fourier plane. The right figure shows the case, that there is in addition a transmission mask for spatial frequencies near zero which reduces the amplitude to 1/4 and the intensity to 1/16 of the original value.

By comparing the simulation results with equation (5.4.9) using the peak to valley value of the phase $\phi_{PV} = 0.1257$, we see in the first case ($I_0 = 1$ unit), that the intensity should vary between 0.749 and 1. Fig. 5.22 left shows a value range between 0.769 and 1.005, which is in quite good agreement with the values given by equation (5.4.9).
In the second case (right figure 5.22) equation (5.4.9) predicts a variation of $2\phi_{PV}/a = 1.005$, i.e. larger than one. Since this equation is only valid for small variations ϕ/a, the result will not be really correct. But, by looking at the intensity range of fig. 5.22 right, it can be seen that the maximum intensity is, as expected, decreased to about 1/16=0.0625 (simulation 0.0637), since most of the light is non–diffracted light with spatial frequency zero which is partially absorbed by the filtering mask. The minimum intensity is 0.0162 corresponding to only 25% of the maximum value. However, the contrast in the right figure is clearly increased compared to the left figure and the simulation is in good agreement with equation (5.4.8) which predicts a minimum value of 0.0156 compared to 0.0162 in the simulation.

Chapter 6

Gaussian beams

Gaussian beams are a paraxial solution of the scalar Helmholtz equation and are suitable to describe the propagation of coherent laser beams [66], [13], [14]. However, the influence of apertures on the laser beam is not considered in this description because apertures would generally disturb the Gaussian beam. The transformation of laser beams at a lens is of course also only treated in a paraxial sense and aberrations of the lens are not taken into account.

First, the basic equations for the propagation of a fundamental Gaussian mode will be described. Afterwards, higher order modes will be presented. At the end, some important examples for the propagation of Gaussian beams will be treated.

6.1 Derivation of the basic equations

The typical property of a collimated laser beam is that it propagates straight on in a homogeneous material. Nevertheless, because of diffraction effects the laser beam diverges during the propagation. Depending on the diameter of the laser beam this effect will be quite small or large. However, the laser beam will behave along the direction of propagation (z–axis) nearly as a plane wave. But instead of having a constant amplitude like for a plane wave the amplitude will be a function of the transversal coordinates (x and y) and also a slowly varying function of the propagation distance along the z–axis. Mathematically, this means that the scalar complex amplitude $u(x, y, z)$ of a laser beam can be described by the product of a (generally) complex function $\Psi(x, y, z)$, which changes only slowly along the z–axis, and the complex amplitude of a plane wave propagating in the z–direction $\exp(ikz)$.

$$u(x, y, z) = \Psi(x, y, z)e^{ikz} \tag{6.1.1}$$

The constant k is again defined as $2\pi n/\lambda$. n is the refractive index of the homogeneous material in which the Gaussian beam propagates and λ is the wavelength in vacuum. To simplify the notation the wavelength $\lambda_n = \lambda/n$ in the material is used in the following.

By using the scalar Helmholtz equation (4.1.2)

$$\left(\boldsymbol{\nabla}^2 + k^2\right) u(x, y, z) = 0 \tag{6.1.2}$$

the following equation for u is obtained:

$$\begin{aligned}
\frac{\partial u}{\partial x} &= \frac{\partial \Psi}{\partial x} e^{ikz} \\
\frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 \Psi}{\partial x^2} e^{ikz} \\
\frac{\partial u}{\partial z} &= \frac{\partial \Psi}{\partial z} e^{ikz} + ik\Psi e^{ikz} \\
\frac{\partial^2 u}{\partial z^2} &= \frac{\partial^2 \Psi}{\partial z^2} e^{ikz} + 2ik\frac{\partial \Psi}{\partial z} e^{ikz} - k^2 \Psi e^{ikz}
\end{aligned}$$

$$\Rightarrow \quad \left(\boldsymbol{\nabla}^2 + k^2\right) u = \left(\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2}\right) e^{ikz} + \frac{\partial^2 \Psi}{\partial z^2} e^{ikz} + 2ik\frac{\partial \Psi}{\partial z} e^{ikz} = 0 \tag{6.1.3}$$

According to our assumption that Ψ changes only slowly along the z–direction the term $\partial^2\Psi/\partial z^2$ is assumed to be so small that it can be neglected. This is the case if the relative variation of $\partial\Psi/\partial z$ during the propagation by one wavelength is much smaller than one. In a mathematical formulation this means:

$$\left|\frac{\partial^2 \Psi}{\partial z^2}\right| \ll \left|2k\frac{\partial \Psi}{\partial z}\right| = \frac{4\pi}{\lambda_n}\left|\frac{\partial \Psi}{\partial z}\right| \quad \Rightarrow \quad \left.\frac{|\Delta(\partial\Psi/\partial z)|}{|\partial\Psi/\partial z|}\right|_{\Delta z = \lambda_n} \ll 4\pi \tag{6.1.4}$$

Using this simplification the following equation for Ψ is obtained:

$$\frac{\partial^2 \Psi}{\partial x^2} + \frac{\partial^2 \Psi}{\partial y^2} + 2ik\frac{\partial \Psi}{\partial z} = 0 \tag{6.1.5}$$

This equation is called the **paraxial Helmholtz equation** because it corresponds to the case of Fresnel diffraction (see sections 4.3.1 and 6.2). To solve it we use first of all a quite simple approach for Ψ which corresponds to a fundamental mode Gaussian beam

$$\Psi(x, y, z) = \Psi_0 e^{i\left(P(z) + \frac{k(x^2+y^2)}{2q(z)}\right)} \tag{6.1.6}$$

with the two complex functions P and q which are both functions of z. Ψ_0 is a constant which depends on the amplitude of the Gaussian beam and is determined by the boundary conditions. Using the notations $P' := dP/dz$ and $q' := dq/dz$ our approach gives:

$$\begin{aligned}
\frac{\partial \Psi}{\partial z} &= \left(iP' - \frac{ik(x^2+y^2)}{2q^2}q'\right)\Psi \\
\frac{\partial \Psi}{\partial x} &= i\frac{kx}{q}\Psi \\
\frac{\partial^2 \Psi}{\partial x^2} &= \frac{ik}{q}\Psi - \frac{k^2x^2}{q^2}\Psi
\end{aligned}$$

Inserting these equations in equation (6.1.5) results in the following conditions for P and q:

$$\frac{2ik}{q} - \frac{k^2(x^2+y^2)}{q^2} - 2kP' + \frac{k^2(x^2+y^2)}{q^2}q' = 0 \tag{6.1.7}$$

This equation has to be fulfilled for arbitrary values of x and y. Therefore, the equation finally gives two equations:

$$P' = \frac{i}{q} \quad \text{and} \quad q' = 1 \tag{6.1.8}$$

By integration we obtain:

$$q(z) = q_0 + z \tag{6.1.9}$$

$$P(z) = i \ln\left(1 + \frac{z}{q_0}\right) \tag{6.1.10}$$

The integration constant of P has been put to zero because it would just introduce a constant phase factor in Ψ.
Equation (6.1.6) has a similar form like a paraxial spherical wave, i.e. a parabolic wave, if q is interpreted as a kind of complex radius of curvature. Therefore, it is useful to split $1/q$ in a real and an imaginary part:

$$\frac{1}{q(z)} = \frac{1}{R(z)} + i\frac{\lambda_n}{\pi w^2(z)} \tag{6.1.11}$$

The real part is the curvature of the wave and R is the real **radius of curvature**. The selection of the imaginary part of $1/q$ becomes obvious by inserting equation (6.1.11) in equation (6.1.6). It shows that the real function w describes the distance $\sqrt{x^2+y^2}$ from the z–axis at which the amplitude decreases to $1/e$ of the maximum value. Therefore, w is called the **beam radius**. w and R are both real functions of z.
A further simplification can be made by choosing $q_0 = q(0)$ as an imaginary number. This means that the radius of curvature R is infinity at $z = 0$, i.e. the curvature of the wave is zero at $z = 0$.

$$\frac{1}{q_0} = i\frac{\lambda_n}{\pi w_0^2} \quad \Rightarrow \quad q_0 = -i\frac{\pi w_0^2}{\lambda_n} \tag{6.1.12}$$

The propagation constant w_0 which corresponds to the curvature $1/R_0 = 0$ is called the **beam waist**. Later it will be shown (see equation (6.3.5)) that the beam waist is the smallest beam radius of a Gaussian beam during its propagation.
In summary by using the equations (6.1.6), (6.1.9), (6.1.10), (6.1.11) and (6.1.12) the function Ψ of a fundamental mode Gaussian beam can be written as:

$$\Psi(x,y,z) = \Psi_0 \frac{1}{1 + i\frac{\lambda_n z}{\pi w_0^2}} e^{-\frac{x^2+y^2}{w^2(z)}} e^{i\frac{k(x^2+y^2)}{2R(z)}} \tag{6.1.13}$$

6.2 Fresnel diffraction and the paraxial Helmholtz equation

In section 4.3.1 the Fresnel diffraction integral is derived as a paraxial solution of the Fresnel–Kirchhoff diffraction formula. Here, it will be shown that the Fresnel diffraction integral describes the propagation of waves with complex amplitudes which fulfill the paraxial Helmholtz equation (6.1.5). So, the name "paraxial Helmholtz equation" is appropriate.
The Fresnel diffraction integral is according to equation (4.3.13), whereby the argument z_0 is substituted by z, because this equation is not only valid in a plane, and k is defined as usual as

$k = 2\pi n/\lambda$:

$$u(x,y,z) = -\frac{ik}{2\pi z} e^{ikz} \iint\limits_A u_0\big(x',y',0\big)\, e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' =$$

$$= \Psi(x,y,z) e^{ikz}$$

$$\Rightarrow\ \Psi(x,y,z) = -\frac{ik}{2\pi z} \iint\limits_A u_0\big(x',y',0\big)\, e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' \tag{6.2.1}$$

So, the function Ψ is defined in accordance with equation (6.1.1) and it has to be shown that this function Ψ is a solution of the paraxial Helmholtz equation (6.1.5). We have the following equations for the partial derivatives of Ψ:

$$\frac{\partial\Psi}{\partial x} = \frac{k^2}{2\pi z^2} \iint\limits_A u_0\left(x',y',0\right)\left(x-x'\right) e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy'$$

$$\frac{\partial^2\Psi}{\partial x^2} = i\frac{k^3}{2\pi z^3} \iint\limits_A u_0\left(x',y',0\right)\left(x-x'\right)^2 e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' +$$

$$+\frac{k^2}{2\pi z^2} \iint\limits_A u_0\left(x',y',0\right) e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy'$$

$$\frac{\partial^2\Psi}{\partial y^2} = i\frac{k^3}{2\pi z^3} \iint\limits_A u_0\left(x',y',0\right)\left(y-y'\right)^2 e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' +$$

$$+\frac{k^2}{2\pi z^2} \iint\limits_A u_0\left(x',y',0\right) e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy'$$

$$\frac{\partial\Psi}{\partial z} = \frac{ik}{2\pi z^2} \iint\limits_A u_0\big(x',y',0\big)\, e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' -$$

$$-\frac{k^2}{4\pi z^3} \iint\limits_A u_0\big(x',y',0\big)\left[\left(x-x'\right)^2+\left(y-y'\right)^2\right] e^{ik\frac{(x-x')^2+(y-y')^2}{2z}}\, dx'dy' \tag{6.2.2}$$

So, it is clear that the function Ψ of equation (6.2.1) fulfills the paraxial Helmholtz equation:

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + 2ik\frac{\partial\Psi}{\partial z} = 0$$

Therefore, the Fresnel diffraction integral and the paraxial Helmholtz equation correspond to each other. The propagation of a Gaussian beam can be either made by calculating the Fresnel

diffraction integral, if the complex amplitude u_0 is given in a plane, or the propagation rules can be directly derived from the paraxial Helmholtz equation as it has been done to obtain equations (6.1.9) and (6.1.10) [13].

6.3 Propagation of a Gaussian beam

The parameters w and R of a Gaussian beam change during the propagation of the beam along the z–axis. An explicit representation of w and R can be obtained by combining the equations (6.1.9), (6.1.11) and (6.1.12):

$$\frac{1}{R} + i\frac{\lambda_n}{\pi w^2} = \frac{1}{z - i\frac{\pi w_0^2}{\lambda_n}} = \frac{z + i\frac{\pi w_0^2}{\lambda_n}}{z^2 + \left(\frac{\pi w_0^2}{\lambda_n}\right)^2}$$

To simplify the notation the so called **Rayleigh length** is defined as

$$z_R := \frac{\pi w_0^2}{\lambda_n} \tag{6.3.1}$$

So, by separating the real and the imaginary part two equations are obtained:

$$\frac{1}{R} = \frac{z}{z^2 + z_R^2} \tag{6.3.2}$$

$$\frac{\lambda_n}{\pi w^2} = \frac{z_R}{z^2 + z_R^2} \tag{6.3.3}$$

$$\Rightarrow$$

$$R(z) = z + \frac{z_R^2}{z} = z + \frac{\pi^2 w_0^4}{\lambda_n^2 z} \tag{6.3.4}$$

$$w^2(z) = \frac{\lambda_n}{\pi}\frac{z^2 + z_R^2}{z_R} = w_0^2 + \frac{\lambda_n^2 z^2}{\pi^2 w_0^2} \tag{6.3.5}$$

The last equation shows that the beam waist w_0 is indeed the smallest value of the beam radius w and that it is obtained at $z = 0$. Simultaneously, the radius of curvature R is infinity at $z = 0$. The equation also shows that the beam radius of the Gaussian beam has the value $w{=}\sqrt{2}w_0$ at the distance $z{=}z_R$ (Rayleigh length) from the beam waist.
Another interesting limiting case is the far field, i.e. $z \to \pm\infty$. Then we have:

$$R(z) = z \tag{6.3.6}$$

$$w(z) = \frac{\lambda_n |z|}{\pi w_0} \tag{6.3.7}$$

The far field angle θ of a Gaussian beam is:

$$\theta \approx \tan\theta = \frac{w(z)}{|z|} = \frac{\lambda_n}{\pi w_0} \tag{6.3.8}$$

So, by measuring the far field angle θ and the wavelength λ_n of a laser diode its beam waist w_0 can be calculated if we assume that the fundamental Gaussian beam is a good description for the wave front of a laser diode.

By using equation (6.3.5) the function Ψ (see equation (6.1.13)) can be written in a more illustrating way:

$$\frac{1}{1+i\frac{\lambda_n z}{\pi w_0^2}} = \frac{w_0}{w_0+i\frac{\lambda_n z}{\pi w_0}} = \frac{w_0\left(w_0 - i\frac{\lambda_n z}{\pi w_0}\right)}{w_0^2+\frac{\lambda_n^2 z^2}{\pi^2 w_0^2}}$$

The term in brackets of the numerator can be expressed as:

$$w_0 - i\frac{\lambda_n z}{\pi w_0} = Ae^{i\Phi} = A\cos\Phi + iA\sin\Phi$$

with

$$A = \sqrt{w_0^2+\frac{\lambda_n^2 z^2}{\pi^2 w_0^2}} = w(z)$$

and

$$\begin{aligned} \cos\Phi &= \frac{w_0}{w(z)} \\ \sin\Phi &= -\frac{\lambda_n z}{\pi w_0 w(z)} \\ \Rightarrow \quad \tan\Phi &= -\frac{\lambda_n z}{\pi w_0^2} \end{aligned}$$

In summary we have:

$$\frac{1}{1+i\frac{\lambda_n z}{\pi w_0^2}} = \frac{w_0}{w(z)}e^{i\Phi(z)} \quad \text{with} \quad \tan\Phi(z) = -\frac{\lambda_n z}{\pi w_0^2} \tag{6.3.9}$$

The complex amplitude u of a Gaussian beam can then be expressed using the equations (6.1.1), (6.1.13) and (6.3.9):

$$u(x,y,z) = \Psi_0\frac{w_0}{w(z)}e^{-\frac{x^2+y^2}{w^2(z)}}e^{i\Phi(z)}e^{i\frac{k(x^2+y^2)}{2R(z)}}e^{ikz} \tag{6.3.10}$$

This means that a Gaussian beam has a Gaussian profile for a constant value z (see fig. 6.1). The term $w_0/w(z)$ ensures that the total power P_G of the beam is conserved during the propagation along the z–direction:

$$\begin{aligned} P_G(z) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}|u(x,y,z)|^2 dxdy = \\ &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Psi_0^2\frac{w_0^2}{w^2(z)}e^{-2\frac{x^2+y^2}{w^2(z)}}dxdy = \\ &= \Psi_0^2\frac{w_0^2}{w^2(z)}\frac{\pi w^2(z)}{2} = \Psi_0^2\frac{\pi w_0^2}{2} = \text{constant} \end{aligned} \tag{6.3.11}$$

By interpreting the beam radius w as lateral extension of the Gaussian beam it can be graphically symbolized as in fig. 6.2. At the beam waist the local curvature of the Gaussian beam is zero. In the far field, the radius of curvature R increases proportional to z like the radius of curvature of a spherical wave.

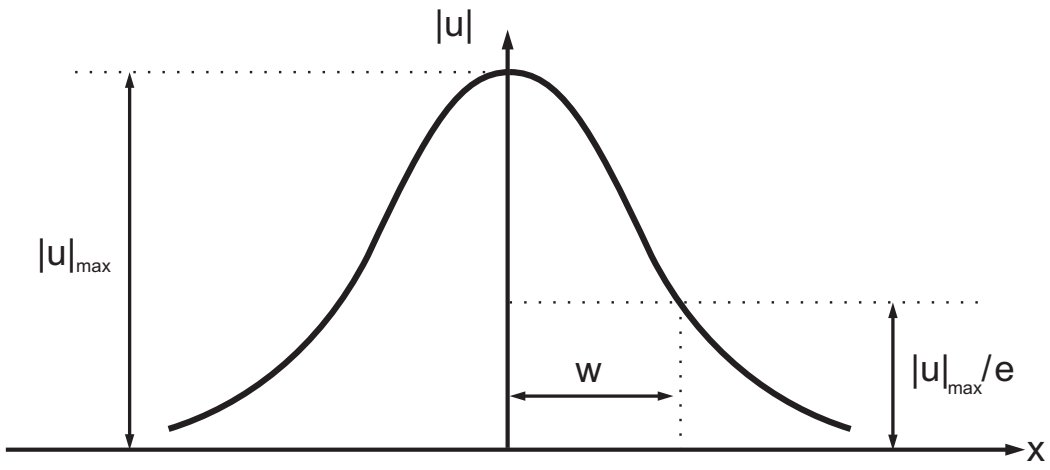


Figure 6.1: Amplitude of a Gaussian beam at a constant value z.

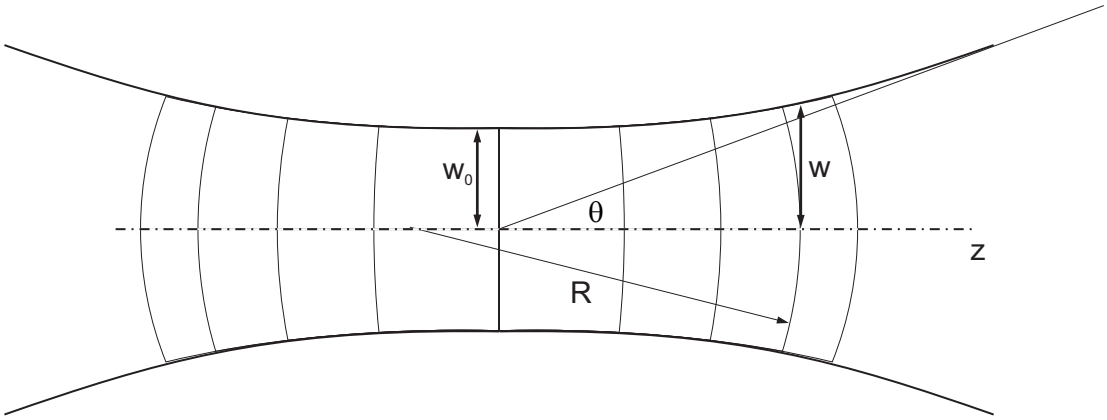


Figure 6.2: Scheme showing the propagation of a Gaussian beam along the z–axis. The Gaussian beam is laterally limited by the beam radius w and its wave front has the local radius of curvature R.

6.4 Higher order modes of Gaussian beams

In equation (6.1.6) a quite simple approach has been selected for the function Ψ which mainly describes the lateral variation of the Gaussian beam. This approach is the fundamental mode in the case of rotational symmetry. In practice, a laser can also show higher order modes where we have to distinguish between a cavity with exact circular symmetry and a cavity showing some directional preference. In the first case, the rotationally symmetric Laguerre–Gaussian modes would be the solution. Since in practice, there are often non–rotationally symmetric parts in the cavity, the more general case are the Hermite–Gaussian modes for a rectangular symmetry. Therefore, in the following the Hermite–Gaussian modes will be treated. The beam can now have two different principal curvatures along the local x– and y–direction and also higher order modes are taken into account. The different principal curvatures are e.g. useful to describe the radiation of laser diodes (e.g. edge emitter) which often have different beam radii and radii of curvature in x– and y–direction. Therefore, the following approach is taken for Ψ [66]:

$$\Psi(x,y,z) = g\left(\sqrt{2}\frac{x}{w_x(z)}\right) h\left(\sqrt{2}\frac{y}{w_y(z)}\right) e^{i\left(P(z)+\frac{kx^2}{2q_x(z)}+\frac{ky^2}{2q_y(z)}\right)} \tag{6.4.1}$$

The functions g and h have to describe the lateral variations of the amplitude of the different modes and therefore it is useful to take the normalized quantities x/w_x and y/w_y, which are pure numbers without a physical unit. Here, w_x and w_y are again the beam radii in x– and y–direction, respectively. The factor $\sqrt{2}$ in the arguments of g and h seems to be quite arbitrary in the moment. But it will be shown in the following that this leads to a well–known differential equation for g and h. This approach for Ψ has to fulfill equation (6.1.5). The functions are in the following written without arguments to simplify the notation. Additionally, the first or second derivative of a function $f(\eta)$ with respect to its argument η is written as f' or f'', respectively. However, it should be kept in mind that e.g. g is in fact a function of x and z (since w_x is a function of z) and therefore also the derivative $g'(\eta) := dg(\eta)/d\eta$ with $\eta := \sqrt{2}x/w_x(z)$ remains a function of x and z. On the other side, q_x is e.g. only a function of z so that the derivative $q_x'(z)$ is just defined as $dq_x(z)/dz$. So, using these notations, we have:

$$\begin{aligned}
\frac{\partial\Psi}{\partial z} &= \left[gh\left(iP' - \frac{ikx^2}{2q_x^2}q_x' - \frac{iky^2}{2q_y^2}q_y'\right) + \left(-gh'\sqrt{2}\frac{y}{w_y^2}w_y' - g'h\sqrt{2}\frac{x}{w_x^2}w_x'\right)\right]\cdot \\
&\quad \cdot e^{i\left(P+\frac{kx^2}{2q_x}+\frac{ky^2}{2q_y}\right)} \\
\frac{\partial\Psi}{\partial x} &= \left[gh\frac{ikx}{q_x} + \sqrt{2}\frac{g'}{w_x}h\right] e^{i\left(P+\frac{kx^2}{2q_x}+\frac{ky^2}{2q_y}\right)} \\
\frac{\partial^2\Psi}{\partial x^2} &= \left[\frac{ik}{q_x}gh - \frac{k^2x^2}{q_x^2}gh + \sqrt{2}\frac{ikx}{q_xw_x}g'h + 2\frac{g''}{w_x^2}h + \sqrt{2}\frac{ikx}{q_xw_x}g'h\right]\cdot \\
&\quad \cdot e^{i\left(P+\frac{kx^2}{2q_x}+\frac{ky^2}{2q_y}\right)}
\end{aligned}$$

$$\frac{\partial^2\Psi}{\partial y^2} = \left[\frac{ik}{q_y}gh - \frac{k^2y^2}{q_y^2}gh + \sqrt{2}\frac{iky}{q_y w_y}gh' + 2\frac{h''}{w_y^2}g + \sqrt{2}\frac{iky}{q_y w_y}gh'\right]\cdot \cdot e^{i\left(P + \frac{kx^2}{2q_x} + \frac{ky^2}{2q_y}\right)}$$

By inserting these functions in equation (6.1.5) and dividing it by gh the following equation is obtained:

$$\begin{aligned} & \frac{ik}{q_x} + \frac{ik}{q_y} - 2kP' + \frac{k^2x^2}{q_x^2}(q_x' - 1) + \frac{k^2y^2}{q_y^2}(q_y' - 1) + \\ + \;& 2\frac{g''}{gw_x^2} - 2\sqrt{2}\frac{ikx}{w_x^2}\frac{g'}{g}\left(w_x' - \frac{w_x}{q_x}\right) + \\ + \;& 2\frac{h''}{hw_y^2} - 2\sqrt{2}\frac{iky}{w_y^2}\frac{h'}{h}\left(w_y' - \frac{w_y}{q_y}\right) = 0 \end{aligned} \tag{6.4.2}$$

This equation has also to be fulfilled for $x \to \infty$ and $y \to \infty$. Then, the terms proportional to x^2 and y^2 are very large compared to the other terms and similar to the case of the fundamental mode the two conditions have to be fulfilled:

$$q_x' = 1 \quad \Rightarrow \quad q_x = q_{x,0} + z \quad \text{and} \quad q_y' = 1 \quad \Rightarrow \quad q_y = q_{y,0} + z \tag{6.4.3}$$

Additionally, q_x and q_y are analogous to the fundamental mode split up into real and imaginary part:

$$\frac{1}{q_x} = \frac{1}{R_x} + i\frac{\lambda_n}{\pi w_x^2} \quad \text{and} \quad \frac{1}{q_y} = \frac{1}{R_y} + i\frac{\lambda_n}{\pi w_y^2} \tag{6.4.4}$$

Calculating the derivative with respect to z delivers for the first equation:

$$-\frac{q_x'}{q_x^2} = -\frac{1}{q_x^2} = -\frac{R_x'}{R_x^2} - 2i\frac{\lambda_n w_x'}{\pi w_x^3} \tag{6.4.5}$$

This equation is added to the square of the first equation (6.4.4). Then, the real and the imaginary part are split resulting in two equations:

$$R_x' = 1 - \frac{\lambda_n^2 R_x^2}{\pi^2 w_x^4} \quad \text{and} \quad w_x' = \frac{w_x}{R_x} \tag{6.4.6}$$

Analogous results are obtained for R_y' and w_y'. Inserting these results into equation (6.4.2) finally delivers:

$$\begin{aligned} & \frac{ik}{q_x} + \frac{ik}{q_y} - 2kP' + \\ + \;& 2\frac{g''}{gw_x^2} - 4\sqrt{2}\frac{x}{w_x^3}\frac{g'}{g} + \\ + \;& 2\frac{h''}{hw_y^2} - 4\sqrt{2}\frac{y}{w_y^3}\frac{h'}{h} = 0 \end{aligned} \tag{6.4.7}$$

Now, the terms in the first row depend only on z, whereas the terms in the second row depend on x and z and the terms in the third row on y and z. Therefore, a separation approach has to be made:

$$\begin{aligned} \frac{ik}{q_x} + \frac{ik}{q_y} - 2kP' &= -f_x(z) - f_y(z) && (6.4.8)\\ 2\frac{g''}{gw_x^2} - 4\sqrt{2}\frac{x}{w_x^3}\frac{g'}{g} &= f_x(z) && (6.4.9)\\ 2\frac{h''}{hw_y^2} - 4\sqrt{2}\frac{y}{w_y^3}\frac{h'}{h} &= f_y(z) && (6.4.10) \end{aligned}$$

whereby f_x and f_y are functions which only depend on z. The solution of the differential equation for g (and analogous for h) shall be described shortly because it is a quite general solution scheme which is often applied in optics and physics. First, the differential equation is written by using $\eta = \sqrt{2}x/w_x$ and the abbreviation $\alpha := f_x w_x^2$ as:

$$\frac{d^2g(\eta)}{d\eta^2} - 2\eta\frac{dg(\eta)}{d\eta} - \frac{1}{2}\alpha g(\eta) = 0 \qquad (6.4.11)$$

The usual approach to solve such a differential equation is to write g as a polynomial:

$$\begin{aligned} g(\eta) &= \sum_{m=0}^{\infty} a_m\eta^m\\ \frac{dg(\eta)}{d\eta} &= \sum_{m=1}^{\infty} ma_m\eta^{m-1}\\ \frac{d^2g(\eta)}{d\eta^2} &= \sum_{m=2}^{\infty} m(m-1)a_m\eta^{m-2} && (6.4.12) \end{aligned}$$

Inserting of this approach into the differential equation (6.4.11) and arranging for equal powers of η gives:

$$\begin{aligned} &\sum_{m=2}^{\infty} m(m-1)a_m\eta^{m-2} - 2\sum_{m=1}^{\infty} ma_m\eta^m - \frac{1}{2}\alpha\sum_{m=0}^{\infty} a_m\eta^m =\\ &= \sum_{m=0}^{\infty}\left[(m+2)(m+1)a_{m+2} - (2m+\frac{1}{2}\alpha)a_m\right]\eta^m = 0 && (6.4.13) \end{aligned}$$

This equation can only be fulfilled for all possible values of η if each coefficient in front of η^m is zero, i.e.:

$$(m+2)(m+1)a_{m+2} - (2m+\frac{1}{2}\alpha)a_m = 0 \quad \Rightarrow \quad a_{m+2} = \frac{2m+\frac{1}{2}\alpha}{(m+2)(m+1)}a_m \qquad (6.4.14)$$

Now, if there would be no stop criterion for the progression of coefficients a_m this equation would tend for very large values m to

$$\lim_{m\to\infty} a_{m+2} = \frac{2}{m}a_m \qquad (6.4.15)$$

because α has a finite value. But this is the same progression of coefficients as for $\exp(\eta^2)$:

$$e^{\eta^2} = \sum_{m=0}^{\infty} \frac{\left(\eta^2\right)^m}{m!} = \sum_{m=0}^{\infty} \frac{\eta^{2m}}{m!} = \sum_{m=0,2,4,\ldots}^{\infty} \frac{1}{\left(\frac{m}{2}\right)!}\eta^m = \sum_{m=0,2,4,\ldots}^{\infty} b_m \eta^m \tag{6.4.16}$$

So, the progression of coefficients will be in this case:

$$b_m = \frac{1}{\left(\frac{m}{2}\right)!} \quad \Rightarrow \quad b_{m+2} = \frac{1}{\left(\frac{m+2}{2}\right)!} = \frac{2}{m+2} b_m \tag{6.4.17}$$

Therefore, for very large values m the progression of coefficients b_m will have the same behavior like the coefficients a_m and the amplitude $|\Psi|$ of the higher order mode Gaussian beam would tend to infinity for large values η because the compensating term (see equations (6.4.1) and (6.4.4)) has only the form $\exp(-x^2/w_x^2) = \exp(-\eta^2/2)$. But, for physical reasons $|\Psi|$ has to tend to zero for large values η. Therefore, there has to be a stop criterion for the progression of coefficients which just means that the variable α has to fulfill the following equation:

$$\alpha = f_x w_x^2 = -4j; \quad j = 0, 1, 2, \ldots \quad \Rightarrow \quad f_x = -\frac{4j}{w_x^2} \tag{6.4.18}$$

By inserting this into equation (6.4.11) the well–known differential equation for the Hermite polynomials H_j is obtained:

$$\frac{d^2 g(\eta)}{d\eta^2} - 2\eta \frac{dg(\eta)}{d\eta} + 2jg(\eta) = 0 \tag{6.4.19}$$

The progression of coefficients of the Hermite polynomials fulfill according to equations (6.4.14) and (6.4.18) the condition:

$$a_{m+2} = \frac{2m + \frac{1}{2}\alpha}{(m+2)(m+1)}\, a_m = \frac{2(m-j)}{(m+2)(m+1)}\, a_m \tag{6.4.20}$$

But, we have two coefficient progressions, one for odd numbers m and one for even numbers m. So, if j is odd only the odd coefficient progression will stop and vice versa with even j. Therefore, additionally one of the coefficients a_0 or a_1, which are the two integration constants of our second order differential equation, has to be zero. So, we have now the possibility to calculate the Hermite polynomials H_j which are the solutions of g and h. The Hermite polynomials are in most text books normalized and so by using equation (6.4.20) it is only possible to calculate the unnormalized Hermite polynomials. But this is no problem since we do not need the normalized polynomials.
If we take $a_0 \neq 0$ and $a_1 = 0$ for the even Hermite polynomials H_j and $a_0 = 0$ and $a_1 \neq 0$ for the odd Hermite polynomials we obtain up to the third order apart from the normalization constant:

$$\begin{aligned} H_0(\eta) &= 1 \\ H_1(\eta) &= \eta \\ H_2(\eta) &= -2\eta^2 + 1 \\ H_3(\eta) &= -\frac{2}{3}\eta^3 + \eta \end{aligned} \tag{6.4.21}$$

Equation (6.4.8) for P' results together with equation (6.4.18) (taking m instead of j) and the analogous equation for f_y (taking n instead of j) in:

$$\begin{aligned}\frac{dP}{dz} &= \frac{i}{2}\left(\frac{1}{q_x}+\frac{1}{q_y}\right) - m\frac{\lambda_n}{\pi w_x^2} - n\frac{\lambda_n}{\pi w_y^2} = \\ &= \frac{i}{2}\left(\frac{1}{q_x}+\frac{1}{q_y}\right) - m\mathrm{Im}\left(\frac{1}{q_x}\right) - n\mathrm{Im}\left(\frac{1}{q_y}\right)\end{aligned} \tag{6.4.22}$$

Using equation (6.4.3) and

$$\begin{aligned}\frac{1}{q_x} &= \frac{1}{q_{x,0}+z} = \frac{1}{\mathrm{Re}(q_{x,0})+i\mathrm{Im}(q_{x,0})+z} = \\ &= \frac{z+\mathrm{Re}(q_{x,0})-i\mathrm{Im}(q_{x,0})}{z^2+2z\mathrm{Re}(q_{x,0})+|q_{x,0}|^2}\end{aligned} \tag{6.4.23}$$

finally gives:

$$\begin{aligned}P(z) &= i\ln\left(\sqrt{1+\frac{z}{q_{x,0}}}\sqrt{1+\frac{z}{q_{y,0}}}\right) - m\arctan\left(\frac{z+\mathrm{Re}(q_{x,0})}{-\mathrm{Im}(q_{x,0})}\right) - \\ &\quad -n\arctan\left(\frac{z+\mathrm{Re}(q_{y,0})}{-\mathrm{Im}(q_{y,0})}\right)\end{aligned} \tag{6.4.24}$$

Note that $-\mathrm{Im}(q_{x,0})$ and $-\mathrm{Im}(q_{y,0})$ are used because these quantities are positive as will be seen in equation (6.4.26). In summary, the function Ψ of the higher order mode Gaussian beams (Hermite–Gaussian modes) in the case of a cartesian coordinate system can be written by using equations (6.4.1)–(6.4.24):

$$\begin{aligned}\Psi(x,y,z) &= H_m\left(\sqrt{2}\frac{x}{w_x(z)}\right)H_n\left(\sqrt{2}\frac{y}{w_y(z)}\right)\frac{1}{\sqrt{\left(1+\frac{z}{q_{x,0}}\right)\left(1+\frac{z}{q_{y,0}}\right)}}\cdot \\ &\quad \cdot\ e^{-i\left[m\arctan\left(\frac{z+\mathrm{Re}(q_{x,0})}{-\mathrm{Im}(q_{x,0})}\right)+n\arctan\left(\frac{z+\mathrm{Re}(q_{y,0})}{-\mathrm{Im}(q_{y,0})}\right)\right]}\cdot \\ &\quad \cdot\ e^{i\pi\left(\frac{x^2}{\lambda_n R_x(z)}+\frac{y^2}{\lambda_n R_y(z)}\right)}\cdot e^{-\left(\frac{x^2}{w_x^2(z)}+\frac{y^2}{w_y^2(z)}\right)}\end{aligned} \tag{6.4.25}$$

The functions w_x and R_x are obtained by comparing the real and imaginary parts of the two equations (6.4.4) and (6.4.23):

$$w_x^2(z) = \frac{\lambda_n}{\pi}\frac{z^2+2z\mathrm{Re}(q_{x,0})+|q_{x,0}|^2}{-\mathrm{Im}(q_{x,0})} \tag{6.4.26}$$

$$R_x(z) = \frac{z^2+2z\mathrm{Re}(q_{x,0})+|q_{x,0}|^2}{z+\mathrm{Re}(q_{x,0})} \tag{6.4.27}$$

Analogous equations are of course valid for w_y and R_y which are obtained by substituting the index x by y.

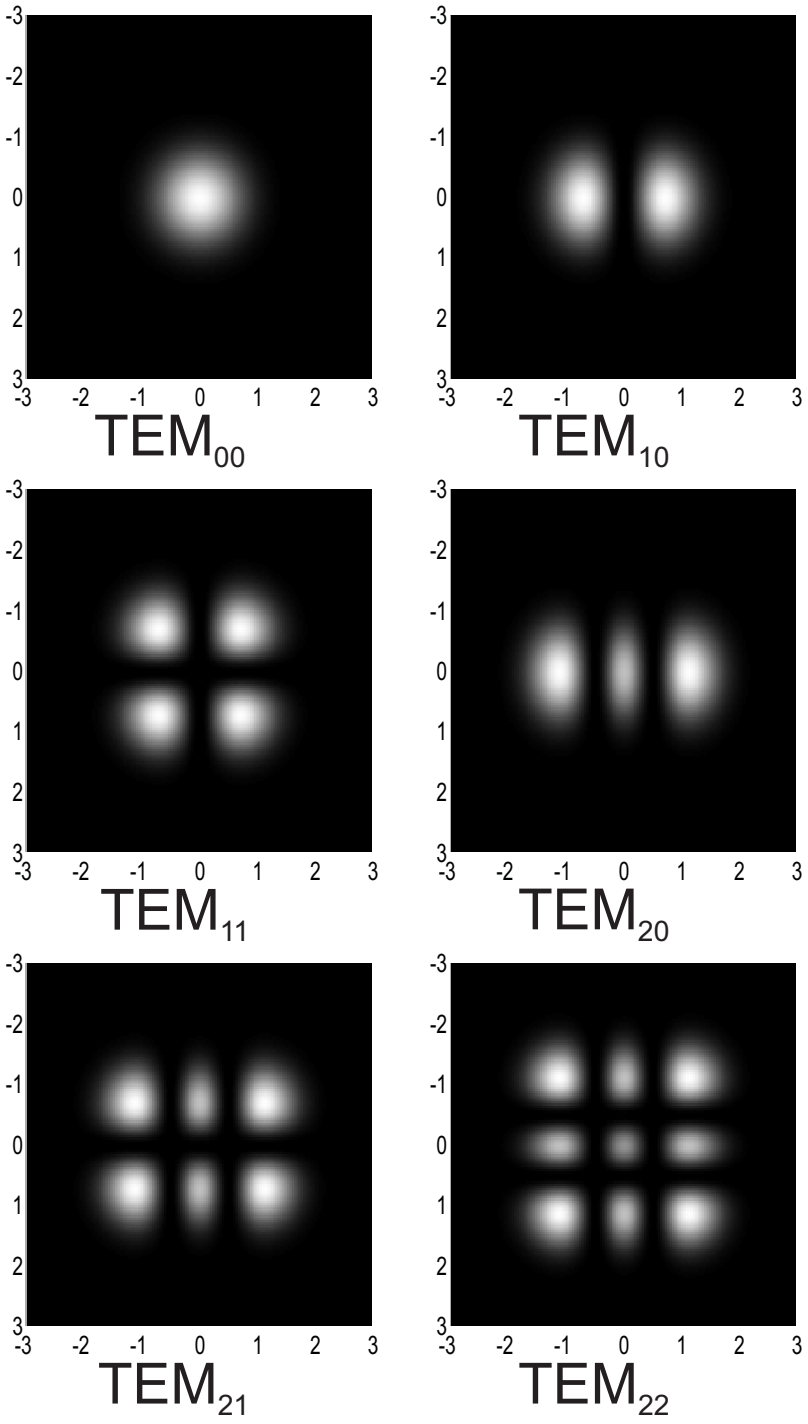


Figure 6.3: Simulation of the intensity distributions of some Hermite–Gaussian modes using the normalized coordinates x/w_x and y/w_y.

Fig. 6.3 shows the typical intensity distribution $|\Psi|^2$ of some lower order Hermite–Gaussian modes with:

$$\begin{aligned} |\Psi(x,y,z)|^2 &= \left[H_m\left(\sqrt{2}\frac{x}{w_x(z)}\right)H_n\left(\sqrt{2}\frac{y}{w_y(z)}\right)\right]^2 \cdot \\ &\cdot \left|\frac{1}{\sqrt{\left(1+\frac{z}{q_{x,0}}\right)\left(1+\frac{z}{q_{y,0}}\right)}}\right|^2 e^{-2\left(\frac{x^2}{w_x^2(z)}+\frac{y^2}{w_y^2(z)}\right)} \end{aligned} \tag{6.4.28}$$

They are named TEM_{mn}, whereby m is the index of the Hermite polynomial H_m with the argument $\sqrt{2}x/w_x$ and n the index of the Hermite polynomial H_n with the argument $\sqrt{2}y/w_y$. The number of zeros is equal to the mode number and the area covered by the modes increases with the mode number.
So, the complete behavior of the higher order Gaussian beam is well defined if the complex quantities $q_{x,0}$ and $q_{y,0}$ at the plane $z = 0$ are known. This is the case if the beam radii $w_{x,0}$ and $w_{y,0}$ and the radii of curvature $R_{x,0}$ and $R_{y,0}$ of the wave front at the plane $z = 0$ are known. Hereby, the beam waists in x– and y–direction can be in different planes. If both beam waists are in the same plane the coordinate system can be chosen such that the beam waists are in the plane $z = 0$ and $q_{x,0} = -i\pi w_{x,0}^2/\lambda_n$ with the beam waist $w_{x,0}$ in x–direction. Then a simplification similar to the case of the fundamental mode of a Gaussian beam can be made and the equations (6.4.26) and (6.4.27) reduce to equations (6.3.4) and (6.3.5). Also, equation (6.4.25) can then be simplified.

6.5 Transformation of a fundamental Gaussian beam at a lens

The transformation of a fundamental Gaussian beam at a (thin) lens is performed using a paraxial approximation. This means that it is assumed that the beam radius immediately in front of the lens is identical to the beam radius immediately behind the lens. Additionally, the radius of curvature of the Gaussian beam changes in the same way like that of a spherical wave. The sign convention is that a positive lens has a positive focal length $f > 0$ (here the focal length on the image side is meant which we earlier designated with f') and that a divergent spherical wave coming from the negative z–direction (i.e. from "left" using the optical agreement) has a positive radius of curvature $R > 0$. R_1 is the radius of curvature immediately in front of the lens and R_2 the radius of curvature immediately behind the lens. Then, a lens with focal length f transforms the radii of curvature according to the paraxial imaging equation of geometrical optics (see fig. 6.4):

$$\frac{1}{R_2} = \frac{1}{R_1} - \frac{1}{f} \tag{6.5.1}$$

Since the beam radius remains constant, the complex beam parameters q_1 immediately in front of the lens and q_2 immediately behind the lens transform also with:

$$\frac{1}{q_2} = \frac{1}{q_1} - \frac{1}{f} \tag{6.5.2}$$

In the case of a thick lens or a lens system the two principal planes of the lens system have to be taken as reference planes for q_1 and q_2 according to the laws of paraxial geometrical optics. If the

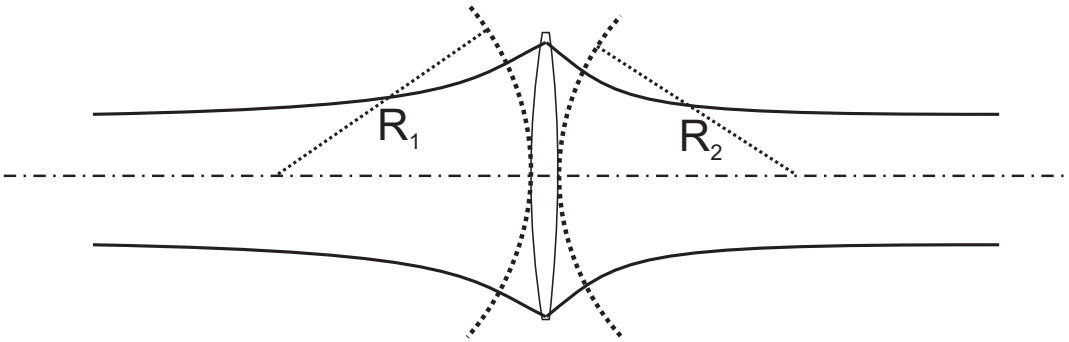


Figure 6.4: Transformation of a Gaussian beam at a thin ideal lens.

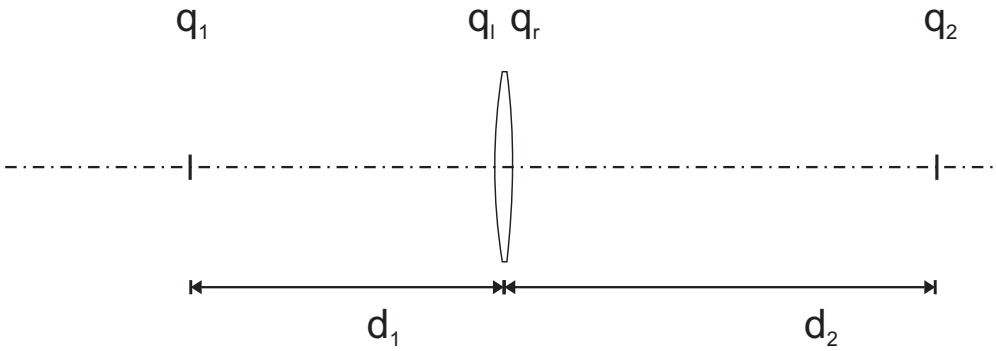


Figure 6.5: Scheme showing the complex beam parameters for the transformation of a Gaussian beam from a plane in the distance d_1 in front of a lens (with focal length f) to a plane in the distance d_2 behind the lens.

q–parameters are different in x– and y–direction as in equation (6.4.4) both sets of parameters just have to be treated separately using equation (6.5.2).

To calculate the relation between the q–parameter q_1 in the distance d_1 in front of a lens with focal length f and q_2 in the distance d_2 behind the lens equations (6.1.9) and (6.5.2) have to be combined. We call q_l as the Gaussian beam parameter immediately in front of the lens and q_r as the beam parameter immediately behind the lens, whereby the parameters are illustrated in fig. 6.5. Then we have:

$$\begin{aligned}
q_l &= q_1 + d_1 \\
\frac{1}{q_r} &= \frac{1}{q_l} - \frac{1}{f} = \frac{1}{q_1 + d_1} - \frac{1}{f} \\
\Rightarrow\ q_r &= \frac{f\,(q_1 + d_1)}{f - q_1 - d_1} \\
q_2 &= q_r + d_2 = \frac{f q_1 + f d_1 + f d_2 - d_2 q_1 - d_1 d_2}{f - q_1 - d_1} \\
\Rightarrow\ q_2 &= \frac{q_1 \left(1 - \frac{d_2}{f}\right) + \left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)}{-\frac{q_1}{f} + \left(1 - \frac{d_1}{f}\right)}
\end{aligned} \tag{6.5.3}$$

6.6 ABCD matrix law for Gaussian beams

The propagation through an optical system can be described in the paraxial geometrical optics by an ABCD matrix (see the lecture about geometrical optics or [67],[68],[66],[14]). We compare the terms of equation (6.5.3) with the paraxial ABCD matrix for the propagation from a plane with the distance d_1 in front of a lens with focal length f to a plane with distance d_2 behind the lens.

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1-\frac{d_2}{f} & d_1+d_2-\frac{d_1 d_2}{f} \\ -\frac{1}{f} & 1-\frac{d_1}{f} \end{pmatrix}, \tag{6.6.1}$$

We see that the Gaussian beam parameter transforms as

$$q_2 = \frac{Aq_1 + B}{Cq_1 + D} \tag{6.6.2}$$

It can be shown that this ABCD matrix law is valid quite generally as long as the paraxial approximation holds. In the following it will be shown for a sequence of (thin) lenses and free space propagation.

6.6.1 Free space propagation

The free space propagation in a homogeneous material with refractive index n is described by equation (6.1.9): $q_2 = q_1 + z$. On the other side the paraxial matrix for free space propagation between two planes with a distance z is:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & z \\ 0 & 1 \end{pmatrix} \quad \Rightarrow \quad q_2 = q_1 + z = \frac{1 \cdot q_1 + z}{0 \cdot q_1 + 1} = \frac{Aq_1 + B}{Cq_1 + D} \tag{6.6.3}$$

So, the free space propagation fulfills the ABCD matrix law of Gaussian beams (see equation (6.6.2)).

6.6.2 Thin lens

For the transformation of a Gaussian beam at a thin lens equation (6.5.2) is valid. The paraxial matrix of a thin lens with focal length f is:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \quad \Rightarrow \quad q_2 = \frac{f q_1}{f - q_1} = \frac{1 \cdot q_1 + 0}{-\frac{q_1}{f} + 1} = \frac{Aq_1 + B}{Cq_1 + D} \tag{6.6.4}$$

So, also the transformation of a Gaussian beam at a thin lens fulfills the ABCD matrix law of equation (6.6.2).

6.6.3 A sequence of lenses and free space propagation

We assume that M_1 and M_2 are the paraxial matrices for two subsequent operations like free space propagation or transformation at a thin lens. The Gaussian beam has the q–parameters q_0 before the first operation, q_1 after the first operation and q_2 after the second operation. Both equations (6.6.3) and (6.6.4) fulfill equation (6.6.2). Therefore, we have the relations:

$$q_1 = \frac{A_1 q_0 + B_1}{C_1 q_0 + D_1} \quad \text{and} \quad q_2 = \frac{A_2 q_1 + B_2}{C_2 q_1 + D_2} \tag{6.6.5}$$

Substitution of q_1 in q_2 gives:

$$q_2 = \frac{A_2 \frac{A_1 q_0 + B_1}{C_1 q_0 + D_1} + B_2}{C_2 \frac{A_1 q_0 + B_1}{C_1 q_0 + D_1} + D_2} = \frac{(A_2 A_1 + B_2 C_1) q_0 + (A_2 B_1 + B_2 D_1)}{(C_2 A_1 + D_2 C_1) q_0 + (C_2 B_1 + D_2 D_1)} \tag{6.6.6}$$

However, the paraxial matrix M of both operations is:

$$\begin{aligned} M &= \begin{pmatrix} A & B \\ C & D \end{pmatrix} = M_2 M_1 = \begin{pmatrix} A_2 & B_2 \\ C_2 & D_2 \end{pmatrix} \cdot \begin{pmatrix} A_1 & B_1 \\ C_1 & D_1 \end{pmatrix} = \\ &= \begin{pmatrix} A_2 A_1 + B_2 C_1 & A_2 B_1 + B_2 D_1 \\ C_2 A_1 + D_2 C_1 & C_2 B_1 + D_2 D_1 \end{pmatrix} \end{aligned} \tag{6.6.7}$$

Summarizing, the relation between q_2 and q_0 is:

$$q_2 = \frac{A q_0 + B}{C q_0 + D} \tag{6.6.8}$$

This shows that the ABCD matrix law is valid for two subsequent operations of free space propagation or transformation at a thin lens. Therefore, it has also to be valid for an arbitrary number of those operations. Geometrical optics shows that a thick lens can be replaced by a thin lens and free space propagation. Therefore, the ABCD matrix law can also be applied for thick lenses or a system consisting of many lenses. We assume that the paraxial ABCD matrix of such a system is known describing the propagation between two planes with the optical system in between. Then, the transformation of the Gaussian beam parameter q_1 at the first plane to the parameter q_2 at the second plane is described by equation (6.6.2). Of course, it is always assumed that no apertures are in the system and that the paraxial approximation is valid, i.e. the optical system is ideal and does not introduce any aberrations.

6.7 Some examples for the propagation of Gaussian beams

6.7.1 Transformation in the case of geometrical imaging

A Gaussian beam with beam parameter q_1 at the distance d_1 in front of the first principal plane of a lens (or a lens system) with focal length f is examined at the distance d_2 behind the second principal plane of the lens. There, the Gaussian beam has the beam parameter q_2. Additionally, it is assumed that the distances d_1 and d_2 and the focal length of the lens fulfill the imaging equation of paraxial geometrical optics:

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f} \quad \Rightarrow \tag{6.7.1}$$

$$1 - \frac{d_2}{f} = -\frac{d_2}{d_1} = \beta \tag{6.7.2}$$

$$d_1 + d_2 - \frac{d_1 d_2}{f} = 0 \tag{6.7.3}$$

$$1 - \frac{d_1}{f} = -\frac{d_1}{d_2} = \frac{1}{\beta} \tag{6.7.4}$$

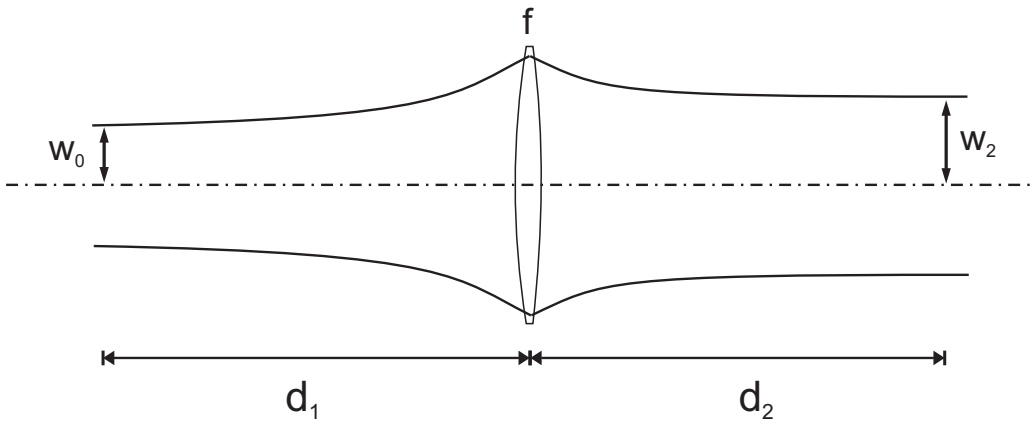


Figure 6.6: Scheme showing the parameters to calculate the position and the size of the beam waist of a Gaussian beam behind a lens.

By using equation (6.5.3) we obtain:

$$q_2 = \frac{\beta q_1}{-\frac{q_1}{f} + \frac{1}{\beta}} \;\Rightarrow\; \frac{1}{q_2} = -\frac{1}{\beta f} + \frac{1}{\beta^2 q_1} \tag{6.7.5}$$

Using equation (6.1.11) to split the complex q–parameters into their real variables delivers:

$$\begin{aligned} &\frac{1}{R_2} + i\frac{\lambda_n}{\pi w_2^2} = -\frac{1}{\beta f} + \frac{1}{\beta^2 R_1} + i\frac{\lambda_n}{\beta^2 \pi w_1^2} \\ \Rightarrow\quad &\frac{1}{R_2} = \frac{1}{\beta^2 R_1} - \frac{1}{\beta f} \quad \text{and} \quad w_2 = |\beta| w_1 \end{aligned} \tag{6.7.6}$$

The result is that the beam radius transforms from one plane to another plane with the lateral magnification β if the imaging equation is fulfilled for these two planes, i.e. if the first plane is imaged by the lens onto the second plane.

6.7.2 Position and size of the beam waist behind a lens

It is assumed that a Gaussian beam has its beam waist w_0 at the distance d_1 in front of a lens with focal length f so that the beam parameter in the first plane is according to equation (6.1.12) $q_1 = -i\pi w_0^2/\lambda_n$. The position and size of the beam waist behind the lens has to be calculated. The wanted parameters are the size w_2 of the beam waist and its distance d_2 from the lens (see fig. 6.6).
Using equations (6.1.11) and (6.5.3) results in:

$$\frac{1}{q_2} = \frac{1}{R_2} + i\frac{\lambda_n}{\pi w_2^2} = \frac{i\frac{\pi w_0^2}{\lambda_n f} + \left(1 - \frac{d_1}{f}\right)}{-i\left(1 - \frac{d_2}{f}\right)\frac{\pi w_0^2}{\lambda_n} + \left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)}$$

$$\Rightarrow\; \frac{1}{R_2} + i\frac{\lambda_n}{\pi w_2^2} = \frac{i\frac{\pi w_0^2}{\lambda_n} - \left(1 - \frac{d_2}{f}\right)\frac{\pi^2 w_0^4}{\lambda_n^2 f} + \left(1 - \frac{d_1}{f}\right)\left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)}{\left(1 - \frac{d_2}{f}\right)^2 \frac{\pi^2 w_0^4}{\lambda_n^2} + \left(d_1 + d_2 - \frac{d_1 d_2}{f}\right)^2} \tag{6.7.7}$$

At the position of the beam waist the real part $1/R_2$ of this equation has to vanish. This gives a condition for calculating d_2:

$$\begin{aligned} &-\left(1-\frac{d_2}{f}\right)\frac{\pi^2 w_0^4}{\lambda_n^2 f}+\left(1-\frac{d_1}{f}\right)\left(d_1+d_2-\frac{d_1 d_2}{f}\right)=0 \\ \Rightarrow\ & d_2=\frac{\frac{\pi^2 w_0^4}{\lambda_n^2 f}-d_1\left(1-\frac{d_1}{f}\right)}{\frac{\pi^2 w_0^4}{\lambda_n^2 f^2}+\left(1-\frac{d_1}{f}\right)^2} \end{aligned} \tag{6.7.8}$$

In the limiting case of geometrical optics, i.e. $w_0 \to 0$, this equation is equal to the paraxial imaging equation.
The beam waist w_2 can be determined using the imaginary part of equation (6.7.7) and replacing then the term $d_1 + d_2 - d_1 d_2/f$ with the help of the first part of equation (6.7.8):

$$\begin{aligned} \frac{\lambda_n}{\pi w_2^2} &= \frac{\frac{\pi w_0^2}{\lambda_n}}{\left(1-\frac{d_2}{f}\right)^2 \frac{\pi^2 w_0^4}{\lambda_n^2}\left[1+\frac{\frac{\pi^2 w_0^4}{\lambda_n^2 f^2}}{\left(1-\frac{d_1}{f}\right)^2}\right]} \\ \Rightarrow\ w_2^2 &= w_0^2\left(1-\frac{d_2}{f}\right)^2\left[1+\frac{\frac{\pi^2 w_0^4}{\lambda_n^2 f^2}}{\left(1-\frac{d_1}{f}\right)^2}\right] \end{aligned} \tag{6.7.9}$$

Equation (6.7.8) delivers

$$1-\frac{d_2}{f}=\frac{1-\frac{d_1}{f}}{\frac{\pi^2 w_0^4}{\lambda_n^2 f^2}+\left(1-\frac{d_1}{f}\right)^2}, \tag{6.7.10}$$

so that the final result for w_2 is:

$$w_2^2=\frac{w_0^2}{\frac{\pi^2 w_0^4}{\lambda_n^2 f^2}+\left(1-\frac{d_1}{f}\right)^2} \tag{6.7.11}$$

Of special interest is the case that the beam waist w_0 in front of the lens lies in the front focal plane of the lens, i.e. $d_1 = f$. The two equations (6.7.8) and (6.7.11) reduce in this special case to:

$$d_2 = f \quad \text{and} \quad w_2 = \frac{\lambda_n f}{\pi w_0} \tag{6.7.12}$$

So, if the beam waist of the incident Gaussian beam lies in the front focal plane of the lens the beam waist of the transformed Gaussian beam lies in the back focal plane of the lens (see fig. 6.7). Additionally, its size w_2 will be the product of the focal length f of the lens and the far field angle (equation (6.3.8)) of the incident Gaussian beam. This result shows that the transformation of Gaussian beams should not be confused with the transformation of paraxial spherical waves of geometrical optics.

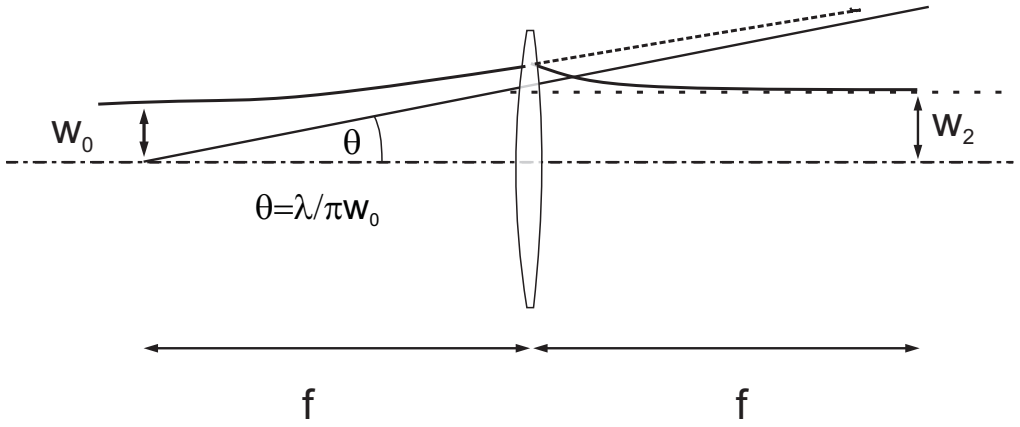


Figure 6.7: Special case of the transformation of a Gaussian beam at a lens where the beam waist lies in the front focal plane of the lens.

Chapter 7

Holography

7.1 History

The optical holography [47],[48],[49],[50] was invented by Dennis Gabor in 1948 (Nobel price for the invention of holography in 1971). But, at this time the laser has not yet been invented (1960: invention of the ruby laser by Theodore H. Maiman) and so Gabor had to use a mercury–vapor lamp. To increase the temporal coherence a wavelength filter was used and with the help of a pinhole the spatial coherence was increased. Gabor also used the method of the so called in–line holography where the object and the reference wave propagate in the same direction. So, a separation of the different diffraction orders was not possible. But, in 1959 Emmett Leith and Juris Upatnieks used two tilted waves as object and reference wave. So, it was the first time possible to separate the different diffraction orders.

7.2 The principle of holography

7.3 Computer generated holograms

Computer–generated holograms (CGHs) or computer–generated diffractive optical elements (DOEs) [51],[52],[53],[54],[55],[56],[57],[58] are used to generate quite arbitrary optical waves.

Chapter 8

Thin films and the Fresnel equations

Text !!!

Bibliography

[1] M. Born, E. Wolf: *Principles of Optics*, 6. edn (Cambridge University Press, Cambridge New York Oakleigh 1997)

[2] R.W. Boyd: *Nonlinear Optics*, (Academic Press, San Diego 2003)

[3] A. Yariv, P. Yeh: *Optical Waves in Crystals*, (John Wiley & Sons, New York 1984)

[4] D.R. Lide: *CRC Handbook of Chemistry and Physics*, 79. edn (CRC Press, Boca Raton 1999)

[5] V.G. Veselago: The Electrodynamics of Substances with Simultaneously Negative Values of ϵ and μ, SOVIET PHYSICS USPEKHI **10** (1968) 509–514

[6] J.B. Pendry: Negative Refraction makes a Perfect Lens, Phys. Rev. Lett. **85** (2000) 3966–3969

[7] A.E. Siegman: Propagating modes in gain–guided optical fibers, J. Opt. Soc. Am. A **20** (2003) 1617–1628

[8] J.W. Goodman: *Introduction to Fourier optics*, 2. edn (McGraw–Hill, New York 1996)

[9] H. Haferkorn: *Optik*, 4. edn (Wiley–VCH, Weinheim 2003)

[10] E. Hecht: *Optics*, 3. edn (Addison–Wesley, Reading 1998)

[11] M.V. Klein, Th.E. Furtak: *Optics*, 2. edn (Wiley, New York 1986)

[12] V.N. Mahajan: *Optical Imaging and Aberrations, Part II: Wave Diffraction Optics*, (SPIE Press, Bellingham 2001)

[13] D. Marcuse: *Light transmission optics*, 2. edn (Van Nostrand, New York 1982)

[14] A.E. Siegman: *Lasers*, (Univ. Science Books, Mill Valley 1986)

[15] J.D. Jackson: *Classical Electrodynamics*, 2. edn (John Wiley & Sons, New York 1975)

[16] R.A. Chipman: Polarization analysis of optical systems, Opt. Eng. **28** (1989) 90–99

[17] R.A. Chipman: Mechanics of polarization ray tracing, Opt. Eng. **34** (1995) 1636–1645

[18] E. Waluschka: Polarization ray trace, Opt. Eng. **28** (1989) 86–89

[19] R.C. Jones: A new calculus for the treatment of optical systems, J. Opt. Soc. Am. **31** (1941) 488–503

[20] J.W. Goodman: *Statistical optics*, (Wiley, New York 1985)

[21] M. Françon: *Optical Interferometry*, (Academic Press, New York 1984)

[22] P. Hariharan: *Optical Interferometry*, (Academic Press Australia, Sydney 1985)

[23] D. Malacara (ed): *Optical Shop Testing*, 2. edn (John Wiley and Sons, New York 1991)

[24] D.W. Robinson, G.T. Reid (ed): *Interferogram Analysis*, (IOP Publishing Ltd, Bristol 1993)

[25] C.M. Haaland: Laser Electron Acceleration in Vacuum, Opt. Comm. **114** (1995) 280–284

[26] Y.C. Huang, R.L. Byer: A proposed high–gradient laser–driven electron accelerator using crossed cylindrical laser focusing, Appl. Phys. Lett. **69** (1996) 2175–2177

[27] R. Dändliker: Two–reference–beam holographic interferometry. In: *Holographic Interferometry*, ed. by P.K. Rastogi (Springer Series in Optical Sciences, Berlin, 1994) Vol. 68, pp. 75–108

[28] R. Dändliker: Heterodyne Holographic Interferometry. In: *Prog. in Optics*, vol XVII, ed by E. Wolf (Elsevier Science Publishing, New York 1980) pp 1–84

[29] G. Schulz, J. Schwider: Interferometric testing of smooth surfaces. In: *Prog. in Optics*, vol XIII, ed by E. Wolf (Elsevier Science Publishing, New York 1976) pp 93–167

[30] J. Schwider: Advanced evaluation techniques in interferometry. In: *Prog. in Optics*, vol XXVIII, ed by E. Wolf (Elsevier Science Publishing, New York 1990) pp 271–359

[31] W.H. Steel: Two–Beam Interferometry. In: *Prog. in Optics*, vol V, ed by E. Wolf (Elsevier Science Publishing, New York 1966) pp 145–197

[32] H.J. Tiziani: Optical Metrology of Engineering Surfaces – Scope and Trends. In: *Optical Measurement Techniques and Applications*, ed by P.K. Rastogi (Artech House Inc., Norwood 1997) pp 15–50

[33] O. Bryngdahl: Applications of Shearing Interferometry. In: *Prog. in Optics*, vol IV, ed by E. Wolf (Elsevier Science Publishing, New York 1965) pp 37–83

[34] H. Sickinger, O. Falkenstörfer, N. Lindlein, J. Schwider: Characterization of microlenses using a phase–shifting shearing interferometer, Opt. Eng. **33** (1994) 2680–2686

[35] H. Schreiber, J. Schwider: A lateral shearing interferometer based on two Ronchi gratings in series, Appl. Opt. **36** (1997) 5321–5324

[36] K. Creath: Phase–Measurement Interferometry Techniques. In: *Prog. in Optics*, vol XXVI, ed by E. Wolf (Elsevier Science Publishing, New York 1988) pp 349–393

[37] A. Hettwer, J. Kranz, J. Schwider: Three channel phase–shifting interferometer using polarization–optics and a diffraction grating, Opt. Eng. **39** (2000) 960–966

[38] J. Schwider, O. Falkenstörfer, H. Schreiber, A. Zöller, N. Streibl: New compensating four–phase algorithm for phase–shift interferometry, Opt. Eng. **32** (1993) 1883–1885

[39] M. Takeda, H. Ina, S. Kobayash: Fourier–transform method of fringe–pattern analysis for computer–based topography and interferometry, J. Opt. Soc. Am. **72** (1982) 156–160

[40] S. Quabis, R. Dorn, M. Eberler, O. Glöckl, G. Leuchs: Focusing light to a tighter spot, Opt. Comm. **179** (2000) 1–7

[41] B. Richards, E. Wolf: Electromagnetic diffraction in optical systems II. Structure of the image field in an aplanatic system, Proc. R. Soc. A **253** (1959) 358–379

[42] M. Françon: *Diffraction*, (Pergamon Press, Oxford 1966)

[43] J.E. Harvey: Fourier treatment of near-field scalar diffraction theory, Am. J. Phys. **47** (1979) 974–980

[44] E. Lalor: Conditions for the Validity of the Angular Spectrum of Plane Waves, J. Opt. Soc. Am. **58** (1968) 1235–1237

[45] U. Vokinger: Propagation, modification and analysis of partially coherent light fields. Dissertation (University of Neuchatel (UFO), Allensbach 2000)

[46] J.J. Stamnes: *Waves in Focal Regions*, (Hilger, Bristol 1986)

[47] H.J. Caulfield: *Handbook of optical Holography*, (Academic Press, New York 1979)

[48] P. Hariharan: *Basics of Holography*, (Cambridge Univ. Press, Cambridge 2002)

[49] E.N. Leith, J. Upatnieks: Recent Advances in Holography. In: *Prog. in Optics*, vol VI, ed by E. Wolf (Elsevier Science Publishing, New York 1967) pp 1–52

[50] G. Saxby: *Practical Holography*, (Prentice Hall, New York 1988)

[51] B.R. Brown, A.W. Lohmann: Computer generated binary holograms, IBM Journal **13** (1969) 160–168

[52] O. Bryngdahl, F. Wyrowski: Digital Holography — Computer–Generated Holograms. In: *Prog. in Optics*, vol XXVIII, ed by E. Wolf (Elsevier Science Publishing, New York 1990) pp 1–86

[53] H.P. Herzig: *Micro–Optics*, (Taylor & Francis, London 1997)

[54] B. Kress, P. Meyrueis: *Digital Diffractive Optics*, (Wiley, Chicester 2000)

[55] W.–H. Lee: Computer–Generated Holograms: Techniques and Applications. In: *Prog. in Optics*, vol XVI, ed by E. Wolf (Elsevier Science Publishing, New York 1978) pp 119–232

[56] D. Maystre: Rigorous Vector Theories of Diffraction Gratings. In: *Prog. in Optics*, vol XXI, ed by E. Wolf (Elsevier Science Publishing, New York 1984) pp 1–67

[57] M. Nevière, E. Popov: *Light Propagation in Periodic Media; Differential Theory and Design*, (Marcel Dekker, New York 2003)

[58] S. Sinzinger, J. Jahns: *Microoptics*, (Wiley–VCH, Weinheim 1999)

[59] I.N. Bronstein, K.A. Semendjajew: *Taschenbuch der Mathematik*, 23. edn (Thun, Frankfurt/Main 1987)

[60] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling: Fourier transform spectral methods. In: *Numerical Recipes in C*, (Cambridge University Press, Cambridge 1991) pp 398–470

[61] B. Besold, N. Lindlein: Fractional Talbot effect for periodic microlens arrays, Opt. Eng. **36** (1997) 1099–1105

[62] S. Quabis, R. Dorn, M. Eberler, O. Glöckl, G. Leuchs: The focus of light-theoretical calculation and experimental tomographic reconstruction, Appl. Phys. B **B72** (2001) 109–113

[63] M. Mansuripur: Distribution of light at and near the focus of high–numerical–aperture objectives, J. Opt. Soc. Am. A **3** (1986) 2086–2093

[64] M. Mansuripur: Distribution of light at and near the focus of high–numerical–aperture objectives: erratum, J. Opt. Soc. Am. A **10** (1993) 382–383

[65] R. Dorn, S. Quabis, G. Leuchs: The focus of light – linear polarization breaks the rotational symmetry of the focal spot, J. of Mod. Opt. **50** (2003) 1917–1926

[66] H. Kogelnik, T. Li: Laser Beams and Resonators, Appl. Opt. **5** (1966) 1550–1567

[67] W. Brouwer: *Matrix Methods in Optical Instrument Design*, (W.A. Benjamin, Inc., New York 1964)

[68] A. Gerrard, J.M. Burch: *Introduction to Matrix Methods in Optics*, (John Wiley & Sons, London 1975)

Index