Optics

Markus Lippitz
November 24, 2023

Contents

Preface

These are the lecture notes for my lecture on optics. The lecture is aimed at students in the third year of the bachelor programme. It follows the idea of Saleh and Teich, 1991: we start with very simple models to describe light and gradually increase the complexity but also the power of the model. We start with ray optics and include geometrical optics and lens aberrations. We then move on to scalar waves, introducing Gaussian beams and Fourier optics. The next step is vectorial electromagnetic waves, which allow us to take into account material properties and birefringence. Finally, we come to quantum optics and describe light as a stream of photons.

These notes are 'work in progress', and probably never really finished. If you find mistakes, please tell me. I am also always interested in other sources covering these topics. The most current version of the lecture notes can be found at github[1]. There you also find the material for the tasks. I have put everything under a CC-BY-SA license (see footer). In my words: feel free to do with it whatever you like. If you make your work available to the public, mention me and use a similar license.

The lecture notes are typeset using the LaTeX class 'tufte-book' by Bil Kleb, Bill Wood, and Kevin Godby[2], which approximates the work of Edward Tufte[3]. I applied many of the modifications introduced by Dirk Eddelbuettel in the 'tint' R package[4]. For the time being, the source is LaTeX, not markdown.

[1] https://github.com/MarkusLippitz

[2] tufte-latex

[3] edwardtufte.com

[4] tint: Tint is not Tufte

Markus Lippitz
Bayreuth, September 18, 2023

Part I

Rays and beams

Chapter 1
Ray optics

Markus Lippitz
October 11, 2023

By the end of this chapter, you should be able to draw, calculate and align a ray's path through an optical system.

Overview

I assume that you have seen a little bit of geometrical optics in your studies, but we will briefly review it. We will introduce the postulates of ray optics and discuss rays at a mirror and a lens as an example. I will also introduce the matrix method of ray optics, which is a very convenient way of calculating the path of a ray through a system of optical elements. More details on these topics can be found in chapter 1 of Saleh and Teich, 1991, chapter 2 of Hering and Martin, 2017, chapters 5 and 6 of E. Hecht, 2017, chapter 2 of Konijnenberg, Adam, and Urbach, 2021.

Postulates of ray optics

Straight rays The propagation of light is described by straight rays that emerge from a source and end at a detector

Index of refraction A medium is described by its index of refraction n. The optical path length in a medium is given by the index of refraction n times the geometric distance d. If $\boldsymbol{r}(s)$ describes a path in 3D space as a function of the path element ds, then the total optical path from A to B is

$$\text{path length} = \int_A^B n(\boldsymbol{r}(s))ds \quad . \tag{1.1}$$

Fermat's Principle Of all the possible paths between points A and B, the light will take the one with the extremal (maximum or minimum) optical path length. This can be written as

$$\delta \int_A^B n(\boldsymbol{r}(s))ds = 0 \tag{1.2}$$

and is Fermat's Principle. The δ means 'variation', i.e. you try to modify $\boldsymbol{r}(s)$ to find shorter (or longer) paths. If several paths have the same optical path

length, then all of them are taken. Since the path length together with the velocity of light gives a travel time, and since one usually finds a minimum as an extremum, one can say that light travels along the path with the shortest travel time.

Consequences of Fermat's Principle

Shadow In an homogeneous medium the straight path is the shortest. A point source thus leads to a perfect projection of an aperture on a screen.

Mirror At a mirror, the angle of incidence equals the angle of reflection, as this gives the shortest path. We can see this when we fold the reflected beam to the side behind the mirror. Then the point of reflection is the point where the ray would cross the mirror surface.

Snell's law At a boundary between two media ($i = 1, 2$), the shortest path is such that

$$n_i \sin \Theta_i = \text{const} \quad , \tag{1.3}$$

where n is the index of refraction and Θ the angle to tge surface normal. With our current model of ray optics, we can not say anything about the amplitude ratio of reflection and transmission at such an interface.

Paraxial rays

Before we look at some optical elements, we need to introduce the idea of a paraxial ray. All the optical elements we are going to look at have an axis of high symmetry, usually rotational symmetry. And in almost all cases the individual elements are placed one after the other, but on a common axis of symmetry. This axis is called the optical axis. The optical axis has a direction, which is typically the direction of the optical ray.

Paraxial rays are those that form only a small angle with the optical axis. This allows us to use the small angle approximation $\sin\theta \approx \theta$, which we will call paraxial approximation in this context. Optics under paraxial approximation is called Gaussian optics. Under paraxial approximation, spherical surfaces are good enough for imaging and focusing. Otherwise one would need aspheric surfaces, for example parabolic or elliptic shapes.

Spherical boundary

Before we come to a (spherical) lens, lets have a look at half a lens, i.e., a single spherical surface of radius R. For convenience, we encode in the sign of the R the direction of curvature: a positive radius describes a convex surface, as seen when looking in the direction of the optical axis.

We start with a ray of angle θ_1 towards the optical axis in a medium of refractive index n_1. It hits the spherical surface at a height y above the optical axis. Here we apply Snell's law and calculate the new direction of the

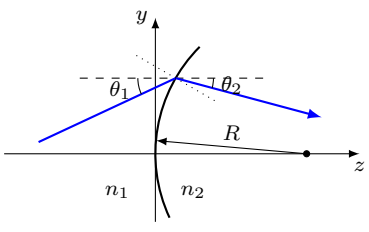

Figure 1.1: Refraction of a ray at a spherical surface

ray. Using the paraxial approximation, we get

$$\theta_2 \approx \frac{n_1}{n_2}\theta_1 = \frac{n_2 - n_1}{n_2}\,\frac{y}{R} \quad . \tag{1.4}$$

Negative angles θ_i describe ray pointing towards below the optical axis.

We can do the same for many rays originating under different angles θ_1 from a point $P_1 = (y_1, z_1)$ in medium 1. We find that they all cross in a point $P_2 = (y_2, z_2)$ in medium 2. Point P_1 is thus imaged on point P_2. For convenience, the sign convention is thus that z is measured from the intersection of the optical axis and the surface, i.e., both z_i are positive. We get

$$\frac{n_1}{z_1} + \frac{n_2}{z_2} \approx \frac{n_2 - n_1}{R} \quad \text{and} \quad y_2 = -\frac{n_1}{n_2}\,\frac{z_2}{z_1}\,y_1 \quad . \tag{1.5}$$

The position z_2 along the optical axis of the image point does not depend on y_1, i.e., every point in the plane $z = z_1$ will have its image in the plane $z = z_2$. These two planes are *conjugate planes*.

Thin lens

We combine two spherical surfaces of radius R_1 and R_2. In the sketch 1.2 R_2 is negative, as this is a concave surface when seen along the optical axis. The two surfaces enclose a medium of refractive index n, while the outside is air, i.e., $n_1 = 1$.

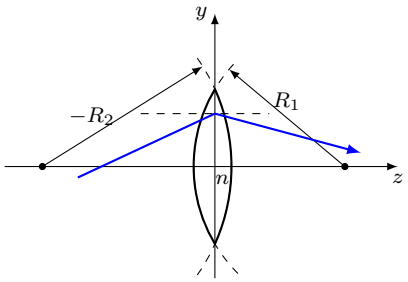

Figure 1.2: Refraction of a ray at a thin lens

We make the approximation that this is a *thin lens*, i.e., that the width Δ of the lens on the optical axis is so small that we can neglect the change in height y of the ray across the lens. We apply twice eq. 1.4 and get

$$\theta_2 = \theta_1 - \frac{y}{f} \quad \text{with} \quad \frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{1.6}$$

with the *focal length* f. The coordinates of the image points are

$$\frac{1}{z_1} + \frac{1}{z_2} = \frac{1}{f} \quad \text{and} \quad y_2 = -\frac{z_2}{z_1}\,y_2 \quad . \tag{1.7}$$

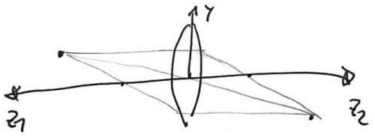

Figure 1.3: Image formation at a thin lens

Again, this holds only in the paraxial approximation. When the rays make a too large angle with the optical axis, they will not be focused ideally. A spherical lens shows aberrations.

For three special rays the action of a lens becomes very simple:

- a ray that arrives parallel to the optical axis ($\theta_1 = 0$) will leave such that it passes through the focal point $(0, f)$ on the other side
- a ray that arrives passing the focal point will leave parallel to the optical axis
- a ray that passes through the center of the lens ($y = 0$) will remain unchanged

These rules have been formulated assuming a positive focal length f. When f is negative, the same rules apply, but it appears that the ray would have passed the focal point on the other side of the lens. Additionally, it helps to remember that parallel rays will intersect in the focal plane.

Matrix method

When tracing a ray through an optical system, all you have to do is apply Snell's law at each interface. This is possible, but a bit tedious. A simpler approach is the idea of matrix optics. We describe a ray at a given position z along the optical axis by two parameters: its angle θ with the optical axis and its height y above the axis. We assume rotational symmetry so that $x = y$ and combine θ and y into one vector. The effect of each optical element can then be written as a matrix acting on the vector, since in the paraxial approximation everything becomes linear.

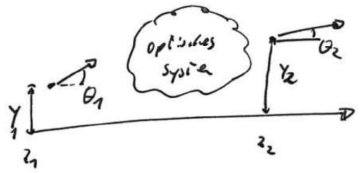

Figure 1.4: The ray-transfer matrix describes the optical system between two planes.

Propagation When a ray travels a distance d through a homogeneous medium, its angle does not change. The height y changes by $d \cdot \theta$. We write this as a ray-transfer matrix

$$\begin{pmatrix} y_2 \\ \theta_2 \end{pmatrix} = \boldsymbol{M}_{\text{prop}} \cdot \begin{pmatrix} y_1 \\ \theta_1 \end{pmatrix} \quad \text{with} \quad \boldsymbol{M}_{\text{prop}} = \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \quad . \tag{1.8}$$

Planar interface Refraction at a planar interface does not change the height, but the angle

$$\boldsymbol{M}_{\text{planar}} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{n_1}{n_2} \end{pmatrix} \quad . \tag{1.9}$$

Spherical interface Refraction at a spherical interface also does not change the height. The change in angle depends on ray height y

$$\boldsymbol{M}_{\text{spherical}} = \begin{pmatrix} 1 & 0 \\ -\frac{n_2 - n_1}{n_2 R} & \frac{n_1}{n_2} \end{pmatrix} \quad . \tag{1.10}$$

Thin lens The action of a thin lens in paraxial apprixaltion is

$$\boldsymbol{M}_{\text{lens}} = \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix} \quad . \tag{1.11}$$

A sequence of optical elements is modelled as a product of ray transfer matrices. Note that the order is reversed. We typically propagate a ray from left to right, but mathematics is of Arabic origin, i.e. reads from right to left. The very first optical element is therefore represented by the rightmost matrix in the matrix product.

Example: Point source in the focal plane of a lens

As an example, let us calculate the effect of a point source that is placed in the focal plane of a thin lens. We start with a ray vector

$$\boldsymbol{v}_{\text{in}} = \begin{pmatrix} y \\ \theta \end{pmatrix} \quad , \tag{1.12}$$

let it propagate by a distance $d = f$ and then pass through a lens. In total we have

$$\boldsymbol{v}_{\text{out}} = \boldsymbol{M}_{\text{lens}} \cdot \boldsymbol{M}_{\text{prop}} \cdot \boldsymbol{v}_{\text{in}} = \begin{pmatrix} y + f\theta \\ -\frac{y}{f} \end{pmatrix} \quad . \tag{1.13}$$

We see that the outgoing angle $\theta_{\text{out}} = -y/f$ does not depend on the direction θ in which the ray leaves the point source. All these rays are thus parallel, as expected.

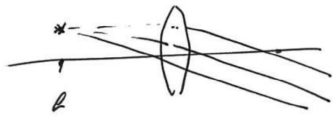

Figure 1.5: Point source in the focal plane of a lens

A thick lens with principal planes

The approximation of a thin lens can be removed by introducing principal planes[1]. It can be shown that the action of a thick lens (i.e, to spherical surfaces separated on the optical axis by a larger distance) and even the action of a sequence of lenses can be described as a single thin lens plus two principal planes. The rays enter the first principal plane and then immediately leave the second principal plane as if they would have passed an effective thin lens of focal length f. The position of the planes and the effective focal length are the only free parameters.

[1] German: Hauptebenen

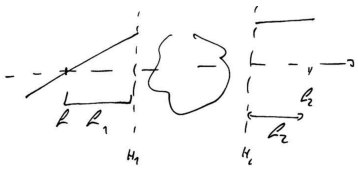

Figure 1.6: The two principal planes simplify many optical systems.

Test yourself

1. A telephoto lens consists of many lenses, but can still be described by a single focal length. This focal length can be longer than the distance between the front lens and the film. The effective single lens is therefore outside the telephoto lens. This can be described by introducing principal planes in the matrix method of ray optics.

   Consider an optical element that can be described by a transfer matrix $\boldsymbol{M}$ with $\det(\boldsymbol{M}) = 1$. The two principal planes are located at distances d_1 before and d_2 after this element. These distances may also be negative. Assume that the refractive index of these domains is one. Show that these three domains together act like a thin lens and calculate d_1 and d_2.

2. Show that any arrangement of thin lenses and distances between these lenses satisfies the above requirement $\det(\boldsymbol{M}) = 1$.
   Hint: $\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$

Aberrations

For[2] designing advanced optical systems Gaussian geometrical optics is not sufficient. Instead non-paraxial rays, and among them also non-meridional[3] rays, must be traced using software based on Snell's Law with the sine of the angles of incidence and refraction. Often many thousands of rays are traced to evaluate the quality of an image. It is then found that in general the non-paraxial rays do not intersect at the ideal Gaussian image point. Instead of a single spot, a spot diagram is found which is more or less confined. The deviation from an ideal point image is quantified in terms of *aberrations*. One distinguishes between monochromatic and chromatic aberrations. The latter are caused by the fact that the refractive index depends on wavelength. The first are a consequence of the small angle approximation in paraxial optics. If instead one retains the first two terms of the Taylor series of the sine, the errors in the image can be quantified by five monochromatic aberrations, the so-called *primary* or *Seidel aberrations* (see, for example, wikipedia[4]). The best known is *spherical aberration*, which is caused by the fact that for a convergent spherical lens, the rays that makes a large angle with the optical axis

[2] This section is adapted from Konijnenberg, Adam, and Urbach, 2021

[3] meridional rays run in a plane than contains the optical axis

[4] https://en.wikipedia.org/wiki/Optical_aberration

are focused closer to the lens than the paraxial rays (see Fig. 1.7). *Distortion* causes deformation of images due to the fact that the magnification depends on the distance of the object point to the optical axis.

For high-quality imaging the aberrations have to be reduced by adding more lenses and optimizing the curvatures of the surfaces, the thicknesses of the lenses and the distances between them. For high quality systems, a lens with an aspherical surface is sometimes used. Systems with very small aberrations are extremely expensive, in particular if the field of view is large, as is the case in lithographic imaging systems used in the manufacturing of integrated circuits.

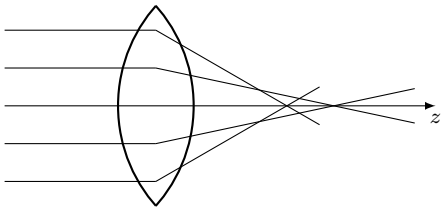

Figure 1.7: Spherical aberration: focal position depends on beam height.

Hands on: Aligning optical elements I

In the practical part of this chapter, you should practice to align optical elements as mirrors, lenses and beam splitters.

Laser beam We will discuss in the next chapter the idea of a laser beam more iin detail. For now, it is sufficient to think of a bundle of rays that run more or less parallel to each other. So the beam is described by the diameter of the ray bundle and the angle of divergence within the bundle. Without lenses, we can for now assume that the rays remain parallel so that a laser beam can be approximated by a ray of geometrical optics.

Degrees of freedom It is useful to have in mind the degrees of freedom that a ray of light has. Above we have used y and θ to describe a ray in a single plane. In three dimensions, we need two spatial coordinates x and y, and two angles with the optical axis, say θ_x and θ_y. If we want to align and thus define a beam, we need to fix these four degrees of freedom.

Alignment tip A three-dimensional volume above an optical table is quite huge. In almost all cases it is not needed. It is sufficient to restrict the beam to one plane parallel to the table surface. The height of the beam above the table is defined by tje alignment tip. The center of the tip should coincide with the center of the beam.

Defining the first leg The laser beam leaves the laser typically in an not too well defined direction and position. We then need two mirrors to define the four degrees of freedom of the beam as each mirror is described by two angles. We place the tip first after the second mirror and use the first mirror to bring the beam on the tip. Then we place the tip far from the second mirror and use the second mirror for alignment. Iterating this procedure will converge.[5]

[5] The other way round does not converge!

Defining all further legs The advantage of a fixed beam hight comes with all further legs of the beam path. At a third mirror, the beam has already the correct height. We place the mirror such that its surface sits at the intersection oif the first two legs. Then we use the two angles of the mirror to define the new direction of tge beam.

Lenses A lens should be centered on the beam. We first align the beam without lens, then place the tip after the intended position of the lens. We put the lens into the beam and translate perpendicular to the beam until it again goes over the tip. The angle of the lens relative to the beam can be check by back reflections. When translating the lens in beam direction, one needs to pay attention that the lens does not leave its centered position.

Beam splitter At a beam splitter, three rotational degrees of freedom come into play. The translational degrees are identical with those of a mirror.

References

Hecht, Eugene (2017). *Optics*. Fifth edition, global edition. Boston: Pearson.

Hering, Ekbert and Rolf Martin (2017). *Optik für Ingenieure und Naturwissenschaftler*. München: Fachbuchverlag Leipzig im Carl Hanser Verlag.

Konijnenberg, Sander, Aurèle J.L. Adam, and Paul Urbach (2021). *BSc Optics*. TU Delft Open.

Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley.

Chapter 2
Gaussian Beams

Markus Lippitz
October 26, 2023

By the end of this chapter you should be able to explain the electric field in a Gaussian focus. You can construct a Gaussian beam 'by hand' for typical lens systems and calculate it using the ABCD law.

Overview

We extend our model to describe light. In this and the following chapters we will use wave optics and assume that light is a scalar wave. We will introduce typical wave functions as plane and spherical waves. Of particular importance are Gaussian beams: as eigenmodes of a laser resonator, they are ubiquitous in optical experiments. We will discuss how these waves and beams are transmitted through optical elements and how we can determine their properties. More details on these topics can be found in chapter 2 and 3 of Saleh and Teich, 1991, chapter 4.6 of Hering and Martin, 2017, chapter 13 of E. Hecht, 2017.

Postulates of Wave Optics

The wave function $u(\boldsymbol{r},t)$ is complex-valued and fulfills the wave equation

$$\nabla^2 u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \tag{2.1}$$

with $c = c_0/n$ the velocity of light in the medium of refractive index n. We do not yet assign a physical meaning to the wave function $u(\boldsymbol{r},t)$. But since you have seen Maxwell's equations elsewhere, you might think of it as one component of the electric field, for example. At interfaces between media, the index of refraction n changes and thus also $1/c$, but we still do not discuss the physics of such interfaces and partial reflection is beyond our scope. The only connection we make to observable physical quantities is by defining the *intensity* I of the wave as

$$I(\boldsymbol{r}) = \langle |u(\boldsymbol{r},t)|^2 \rangle \tag{2.2}$$

where the pointed brackets indicate a time average over a period long compared to the wave period.

A consequence of the linear wave equation is the superposition principle. If u and v are solutions to the wave equation, then also $\alpha u + \beta v$ is a solution. This also means that light beams cross themselves without interaction.

Monochromatic waves

The solutions of the wave equation can be written as harmonic functions

$$u(\boldsymbol{r}, t) = \tilde{u}(\boldsymbol{r})\, e^{-i\omega t} \tag{2.3}$$

with an angular frequency $\omega = 2\pi\nu$. The spatial part $\tilde{u}(\boldsymbol{r})$ fulfils the Helmholtz equation

$$\nabla^2 \tilde{u} + k^2 \tilde{u} = 0 \quad \text{with} \quad k = \frac{\omega}{c} \quad . \tag{2.4}$$

k is called the *wavenumber* and becomes the *wavevector* when going to three dimensions. The intensity is then given by $\tilde{u}(\boldsymbol{r})$

$$I(\boldsymbol{r}) = \langle |u(\boldsymbol{r}, t)|^2 \rangle = |\tilde{u}(\boldsymbol{r})|^2 \quad , \tag{2.5}$$

i.e., the intensity of a monochromatic wave is constant in time.

Lets discuss a few typical examples

Plane wave The amplitude $\tilde{u}$ is given by

$$\tilde{u}(\boldsymbol{r}) = A\, e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \tag{2.6}$$

with $\boldsymbol{k}$ the wavevector and $|\boldsymbol{k}| = k$. The *wavefronts*, i.e., surfaces of constant phase $\phi = q\, 2\pi = \arg \tilde{u}(\boldsymbol{r})$, are parallel and equidistant planes. The distance is the wavelength $\lambda = c/\nu = 2\pi/k$.

When the index of refraction n changes at an interface, the frequency ω remains the same, but the wavelength λ, the velocity of light c and the wavenumber k change

$$\lambda = \frac{\lambda_0}{n} \qquad c = \frac{c_0}{n} \qquad k = n\, k_0 \quad . \tag{2.7}$$

Spherical wave Here the amplitude $\tilde{u}$ is given by

$$\tilde{u}(\boldsymbol{r}) = \frac{A}{r}\, e^{ikr} \quad \text{with} \quad r = |\boldsymbol{r}| \quad . \tag{2.8}$$

Note that the right side of the equation does only use scalar variables. The wavefunction depends thus only on the distance to the origin and has spherical symmetry. The wavefronts are concentric spheres of distance λ.

Paraboloidal wave Close to the optical axis, we can approximate the spherical wave by a paraboloidal wave. We call θ

$$\theta^2 = \frac{x^2 + y^2}{z^2} \ll 1 \tag{2.9}$$

and write r as a Taylor expansion on θ

$$r = \sqrt{x^2 + y^2 + z^2} = z\sqrt{1 + \theta^2} = z\left(1 + \frac{\theta^2}{2} - \frac{\theta^4}{8} + \cdots\right) \tag{2.10}$$

$$\approx z\left(1 + \frac{\theta^2}{2}\right) = z + \frac{x^2 + y^2}{2z} \quad . \tag{2.11}$$

This is called the *Fresnel approximation*. We put it into eq. 2.8 and approximate in the amplitude term even $r \approx z$. We get

$$\tilde{u}(\boldsymbol{r}) = \frac{A}{z}\, e^{ikz}\, e^{ik\,\frac{x^2+y^2}{2z}} \quad . \tag{2.12}$$

For points close to the optical axis but far from the origin, a spherical wave approaches a planar wave. In between, the paraboloidal wave is a useful approximation.

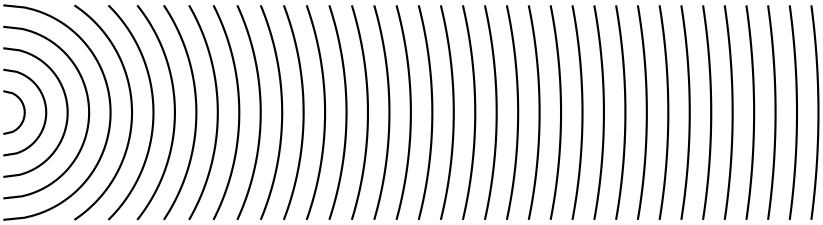

Figure 2.1: Wave fronts of a spherical wave, showing the transition to plane waves far from the origin.

A transparent plate

As most simple optical element, we consider a transparent plate of thickness d and index of refraction n in air. We transmit a plane wave. The wavefunction is continuos at the interface. We are interested in the complex-valued transmission function $t(x, y)$

$$t(x, y) = \frac{\tilde{u}(x, y, d)}{\tilde{u}(x, y, 0)} \quad . \tag{2.13}$$

For perpendicular incidence, the phase advances by $nk_0 d$ from left to right. The transmission function is thus

$$t(x, y) = e^{ink_0 d} \quad . \tag{2.14}$$

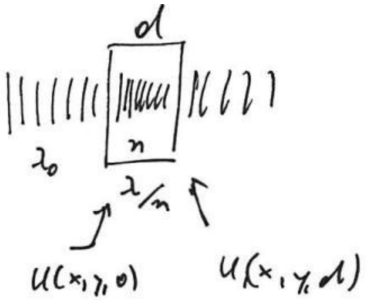

Figure 2.2: A plate

When the plane wave approaches the plate under angle θ, then Snell's law gives the internal angle θ_i as $\sin\theta = n \sin\theta_i$. The wavevector makes this angle θ_i with the optical axis, so that the z-component of the term $\boldsymbol{k} \cdot \boldsymbol{r}$ at the right side gives $nk \cos\theta_i$ and the total transmission function is

$$t(x, y) = e^{ink_0 d \cos\theta_i} \quad . \tag{2.15}$$

This is always against my intuition. The geometrical path in the plate gets longer by tilting it, but the phase difference becomes smaller. The point is that we only take the component along z into account, as shifting a plane wave perpendicular to its direction of travel does not change anything.

We of course make again the approximation that the angle θ is small enough so that we can ignore the $\cos\theta_i$ part.

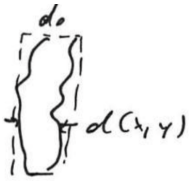

Figure 2.3: A plate of variable thickness

If the plate has a variable thickness $d(x, y)$, we enclose it in a box of thickness d_0. Then part of the phase progression goes with n, part with air ($n = 1$). In total this is

$$t(x, y) \approx e^{ink_0 d(x,y)}\, e^{ik_0(d_0 - d(x,y))} = h_0 e^{i(n-1)k_0 d(x,y)} \tag{2.16}$$

with $h_0 = e^{ik_0 d_0}$ a constant phase factor. This makes the approximation that all angles are small enough and neighboring parts of the plate do not 'mix' at the output.

Conversion of a plane wave to a spherical wave by a lens

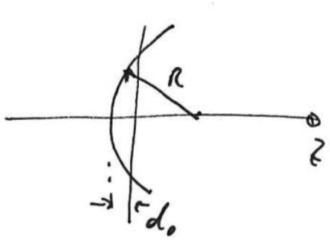

Figure 2.4: A lens as plate of variable thickness

The most interesting thin plate of variable thickness is a lens. For simplicity, we use a plane convex lens, i.e, set one radius of curvature to infinity. The thickness $d(x, y)$ of this plate is then

$$d(x,y) = d_0 - \left(R - \sqrt{R^2 - (x^2+y^2)}\right) \quad . \tag{2.17}$$

We again use the Fresnel approximation $x^2 + y^2 \ll R^2$ and approximate the square-root term

$$\sqrt{R^2 - (x^2+y^2)} = R\sqrt{1 - \frac{x^2+y^2}{R^2}} \approx R\left(1 - \frac{x^2+y^2}{2R^2}\right) \tag{2.18}$$

so that

$$d(x,y) \approx d_0 - \frac{x^2+y^2}{2R^2} \quad . \tag{2.19}$$

The transmission function is then

$$t(x,y) = h_0\, e^{-ik_0 \frac{x^2+y^2}{2f}} \quad \text{with} \quad f = \frac{R}{n-1} \tag{2.20}$$

and $h_0 = e^{ink_0 d_0}$ another constant phase factor that we ignore.

A spherical lens thus transforms a plane wave into a paraboloidal wave centered around $z = f$.

Gaussian beams as a paraxial solution of the wave equation

When we have discussed typical solutions to the wave equation above, we started from the full wave equation, found spherical waves as solution, and then made the paraxial approximation to arrive at the paraboloidal waves. We could also have gone a different route. We can apply the paraxial approximation to the wave equation directly. This leads to the paraxial Helmholtz equation

$$\nabla_T^2 A + i2k\frac{\partial A}{\partial z} = 0 \quad \text{and} \quad \tilde{u}(\boldsymbol{r}) = A(\boldsymbol{r})\, e^{ikz} \tag{2.21}$$

with ∇_T acting only on the transverse coordinates only. The envelop $A(\boldsymbol{r})$ modulates the carrier $\exp(ikz)$. A needs to be *slowly varying*, i.e., on a wavelength length scale it should not change much.

The paraboloidal waves

$$\tilde{u}(\boldsymbol{r}) = \frac{A}{z}\, e^{ikz}\, e^{ik\frac{x^2+y^2}{2z}} \tag{2.22}$$

i.e.,

$$A(\boldsymbol{r}) = \frac{A_1}{z}\, e^{ik\frac{x^2+y^2}{2z}} \tag{2.23}$$

fulfil this paraxial Helmholtz equation. The interesting point is that we can come to other solutions of the paraxial Helmholtz equation by replacing z by $q(z) = z - iz_0$, i.e.

$$A(\boldsymbol{r}) = \frac{A_1}{q(z)}\, e^{ik\frac{x^2+y^2}{2q(z)}} \quad . \tag{2.24}$$

These are *Gaussian beams*. We call q the q-parameter and z_0 the *Rayleigh range*. We separate the complex function $1/q(z)$ into its real and imaginary part

$$\frac{1}{q(z)} = \frac{1}{z - iz_0} = \frac{1}{R(z)} + i\frac{\lambda}{\pi W^2(z)} \quad . \tag{2.25}$$

We will see that R and W give the wavefront radius of curvature and the beam width, respectively. Putting everything together, the wavefunction reads

$$\tilde{u}(\boldsymbol{r}) = A_0 \frac{W_0}{W(z)} \exp\left(-\frac{\rho^2}{W^2(z)}\right) \exp\left(+ikz + ik\frac{\rho^2}{2R(z)} - i\zeta(z)\right) \tag{2.26}$$

with

$$W(z) = W_0\sqrt{1+\left(\frac{z}{z_0}\right)^2} \tag{2.27}$$

$$R(z) = z\left[1+\left(\frac{z_0}{z}\right)^2\right] \tag{2.28}$$

$$\zeta(z) = \arctan\frac{z}{z_0} \tag{2.29}$$

$$W_0 = \sqrt{\frac{\lambda z_0}{\pi}} \quad . \tag{2.30}$$

Note that there are only two independent parameters next to the wavelength λ, namely the amplitude A_0 and the Rayleigh range z_0.

Parameters and Properties of Gaussian Beams

Let us discuss some properties of a Gaussian beam. The *intensity* I is

$$I(\rho, z) = |\tilde{u}(\rho, z)|^2 = I_0\left(\frac{W_0}{W(z)}\right)^2 e^{-\frac{2\rho^2}{W(z)^2}} \tag{2.31}$$

i.e, the transversal profile of the intensity is a Gaussian. Along the z axis

$$I(0, z) = I_0\left(\frac{W_0}{W(z)}\right)^2 = \frac{I_0}{1+\left(\frac{z}{z_0}\right)^2} \approx I_0\left(\frac{z_0}{z}\right)^2 \tag{2.32}$$

where we assumed $z \gg z_0$ in the last approximation. This means that the intensity drops as $1/z^2$, as a spherical wave.

At the *beam waist* ($z = 0$), the width $W(z = 0) = W_0$ describes the radial distance ρ at which the intensity has dropped to $1/e^2$ of the peak value. When moving one Rayleigh range z_0 out of the focus, this radius increases by $\sqrt{2}$, as $W(z_0) = \sqrt{2}W_0$. At this distance, the intensity on the axis has dropped by a factor $1/2$. When integrating over any surface perpendicular to the optical axis, the integrated intensity or power remains the same.

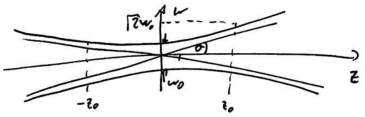

Figure 2.5: Divergence of a Gaussian beam

We can define a *divergence*, or opening angle Θ of the Gaussian beam. Far away from the beam waist, i.e. $z \gg z_0$ we have

$$W(z) \approx W_0\frac{z}{z_0} = \Theta\, z \quad \text{with} \quad \Theta = \frac{W_0}{z_0} = \frac{\lambda}{\pi W_0} \quad . \tag{2.33}$$

Note that next to the wavelength λ only the Rayleigh range z_0 *or* the beam waist W_0 is a free parameter, not both. The divergence of the beam is fully contained in the beam waist. One can interpret this as diffraction of the Gaussian beam at its own waist. For comparison, diffraction at a circular aperture of radius R would lead to an angle $\Theta_{\text{app}} = 0.61\lambda/R$.

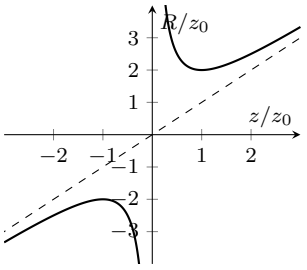

Figure 2.6: Radius of curvature R around the beam waist. Dashed: spherical wave

The phase of the Gaussian beam is given by the imaginary factors of the exponential function, i.e.

$$\phi(\rho, z) = kz + k\frac{\rho^2}{2R(z)} - \zeta(z) \tag{2.34}$$

and the curvature of the phase fronts by

$$R(z) = z\left[1 + \left(\frac{z_0}{z}\right)^2\right] \quad . \tag{2.35}$$

At the focus ($z = 0$) and for $z \to \infty$ we find a diverging radius of curvature, i.e., a plane wave. For large distances this is the same as for a circular wave. Around the focus, the Gaussian beam differs as the radius of curvature changes such that we also get a plane wave exactly at $z = 0$. Another peculairity of Gaussian beams is the *Gouy phase* $\zeta(z)$

$$\zeta(z) = \arctan\frac{z}{z_0} \quad . \tag{2.36}$$

When passing through the focus, the wave undergoes a pi phase shift. An intuitive picture could be the following[1]: In geometrical optics, the ray would go through the focus. In a Gaussian beam, the path along the $1/e$ contour stays on the same side of the optical axis and in thus around the focus a bit shorter. This is compensated by the Gouy phase shift.

[1] Boyd, 1980.

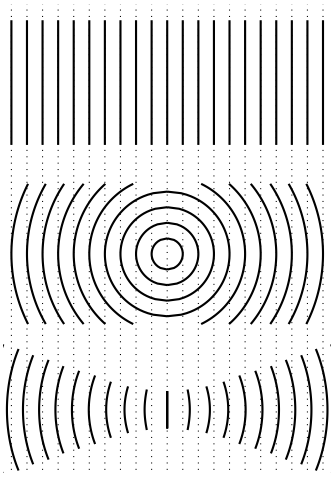

Figure 2.7: Wavefronts of a plane wave, a spherical wave and a Gaussian wave.

Gaussian beams as eigenmodes of a resonator

The importance of Gaussian beams comes from the laser as a ubiquitous light source. A laser produces Gaussian beams because these wave functions are the eigenmodes of a resonator formed by two spherical mirrors.

In a laser, we are interested in eigenmodes, i.e. optical wave functions that do not change as they bounce back and forth in the resonator. The mirrors in a laser cavity are typically so highly reflective that there are many round trips before the field leaves the cavity.

For an eigenmode to occur, the wavefront of the mode at the position of the mirror must match the shape of the mirror, otherwise it will reflect back into itself. The design of the cavity gives the radius of curvature R_1 and R_2 and the distance d between the mirrors. We now show that under certain conditions a Gaussian beam is an eigenmode of such a cavity.

We search for the positions z_1 and z_2 of the mirrors and the Rayleigh range z_0 if the mean. We have the equation system

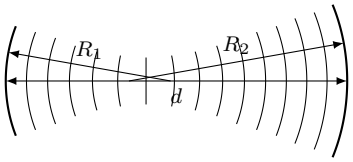

Figure 2.8: Eigenmodes of a laser cavity

$$z_2 = z_1 + d \tag{2.37}$$

$$R_1 = z_1\left[1 + \left(\frac{z_0}{z_1}\right)^2\right] \tag{2.38}$$

$$R_2 = z_2\left[1 + \left(\frac{z_0}{z_2}\right)^2\right] \quad . \tag{2.39}$$

The solution is

$$z_1 = \frac{-d(R_2+d)}{R_1+R_2+2d} \tag{2.40}$$

$$z_2 = z_1 + d \tag{2.41}$$

$$z_0^2 = \frac{-d(R_1+d)(R_2+d)(R_1+R_2+d)}{(R_1+R_2+2d)^2} \quad . \tag{2.42}$$

For a Gaussian beam z_0 must be real (or $z_0^2 > 0$). Otherwise $q = z - iz_0$ would be real and we would get a paraboloidal wave. This results in the *stability condition* of a spherical cavity

$$0 \leq \left(1 + \frac{d}{R_1}\right)\left(1 + \frac{d}{R_2}\right) \leq 1 \tag{2.43}$$

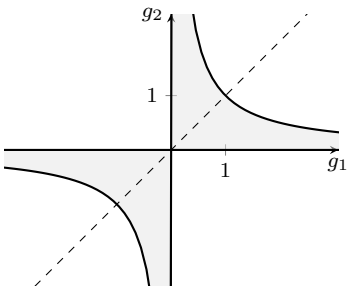

Figure 2.9: A laser cavity is stable inside the shaded region ($g_i = 1 + d/R_i$). Symmetric cavities are along the diagonal, flat mirrors at $g = 1$.

Thin lens

What happens when a Gaussian beam passes through a thin lens? We assume a thin lens, so z does not change. The radial amplitude distribution $\tilde{u}(\rho, z)$ also does not change, which means that the width parameter W remains the same, i.e,

$$W^{(L)} = W^{(R)} \quad . \tag{2.44}$$

The phase needs a bit more attention. Just before the lens, the phase of the Gaussian beam is

$$\phi^{(L)} = kz + k\frac{\rho^2}{2R} - \zeta(z) \tag{2.45}$$

The phase effect of a lens is (see eq. 2.20)

$$\Delta\phi_{\text{lens}} = -k\frac{\rho^2}{2f} \tag{2.46}$$

so that after the lens we have in total

$$\phi^{(R)} = \phi^{(L)} + \Delta\phi_{\text{lens}} = kz - \zeta(z) + k\left(\frac{\rho^2}{2R} - \frac{\rho^2}{2f}\right) \tag{2.47}$$

i.e.

$$\frac{1}{R^{(R)}} = \frac{1}{R^{(L)}} - \frac{1}{f} \quad . \tag{2.48}$$

What does this mean for the other properties of a Gaussian beam? How are Rayleigh range z_0 and beam waits W_0 modified by a lens? Knowing λ, $W(z)$ and R, i.e, the beam properties at the lens, we can use eqs. 2.27–2.30 to calculate z_0, W_0 and z, i.e. the focal parameter and the distance z of focus and lens. We get

$$W_0 = \frac{W(z)}{\sqrt{1 + \left(\frac{\pi W(z)}{\lambda R(z)}\right)^2}} \tag{2.49}$$

$$z = \frac{R(z)}{1 + \left(\frac{\pi W(z)}{\lambda R(z)}\right)^2} \quad . \tag{2.50}$$

We thus can connect the left and right beam parameters.

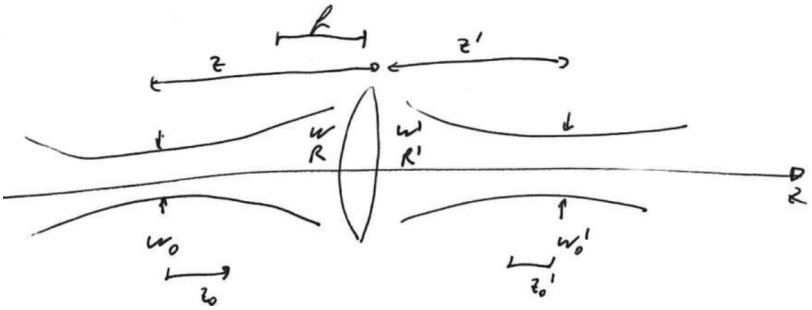

Figure 2.10: A lens acting on a Gaussian beam

When the left beam waist is far from the lens, then we can see this arrangement as imaging of the left beam waist to a position $z^{(R)}$ on the right side of the lens, which will magnify the beam waist radius. The magnification factor in ray optics is

$$M_r = \left| \frac{f}{z^{(L)} - f} \right| \quad \text{and} \quad W_0^{(R)} \approx M_r W_0^{(L)} \quad . \tag{2.51}$$

The requirement 'far enough' means $z^{(L)} - f \gg z_0^{(L)}$, or

$$r = \frac{z_0^{(L)}}{z^{(L)} - f} \ll 1 \quad . \tag{2.52}$$

We can calculate the beam parameters without this approximation using a general magnification factor M and obtain

$$M = \frac{M_r}{\sqrt{1 + r^2}} \tag{2.53}$$

$$W_0^{(R)} = M W_0^{(L)} \tag{2.54}$$

$$(z^{(R)} - f) = M^2 (z^{(L)} - f) \tag{2.55}$$

$$z_0^{(R)} = M^2 z_0^{(L)} \tag{2.56}$$

$$\Theta_0^{(R)} = \frac{\Theta_0^{(L)}}{M} \quad . \tag{2.57}$$

ABCD Law and q parameter

Things become simpler when we realize that the q parameter is governing the Gaussian beam. We introduced above the Gaussian beams by

$$q(z) = z - i z_0 \quad \text{and} \quad \frac{1}{q(z)} = \frac{1}{R(z)} + i \frac{\lambda}{\pi W^2(z)} \quad . \tag{2.58}$$

As soon as we know q at a single position along the beam, we can calculate all the rest. When q_1 and q_2 describe the q parameters left and right of an optical element, both are connected by the *ABCD law*

$$q_2 = \frac{A q_1 + B}{C q_1 + D} \tag{2.59}$$

where

$$\boldsymbol{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{2.60}$$

is the 2×2 matrix of the matrix method in ray optics, as introduced in the last chapter.[2]

[2] see Brooker, 2008, chapter 7.9, for a justification.

Propagation by a distance d thus leads to

$$q_2 = q_1 + d \quad . \tag{2.61}$$

The action of a *lens* is described by

$$q_2 = \frac{f q_1}{f - q_1} \quad . \tag{2.62}$$

Technique: Knife Edge Test

A common tool for determining the properties of a Gaussian beam is a knife edge or razor blade. We mount it so that it can be moved perpendicular to the optical axis, cutting out a variable part of the beam. If we measure the power in the beam after the knife edge as a function of its position, we get a partial integral over the cross-section of the beam.

$$P(x_0) = \int_{x=x_0}^{\infty} \int_{y=-\infty}^{\infty} I(x,y)\, dx\, dy \quad . \tag{2.63}$$

Taking the numerical derivative yields the beams intensity profile.

$$I(x) \propto \frac{\partial P(x_0)}{\partial x_0} \quad . \tag{2.64}$$

We can also observe the shadow of the knife edge to determine the position of the beam waist. We place a screen far from the waist and the knife edge at the estimated waist position. If the knife is between the waist and the screen, the shadow will start from the same side as the knife enters the beam. If the knife is further away than the waist, the directions are reversed. If the knife is exactly at the waist, the image on the screen will turn dark with no apparent direction. This is the *Foucault knife edge test* to determine the position and quality of a focus, originally of a spherical mirror.

Test yourself

1. Explain why the image turns dark with no apparent direction.

References

Boyd, Robert W. (1980). “Intuitive explanation of the phase anomaly of focused light beams”. In: *Journal of the Optical Society of America* 70, pp. 877–880.

Brooker, Geoffrey (2008). *Modern classical optics*. 1. publ., repr. with corr. Oxford master series in physics. Oxford [u.a.]: Oxford Univ. Press.

Hecht, Eugene (2017). *Optics*. Fifth edition, global edition. Boston: Pearson.

Hering, Ekbert and Rolf Martin (2017). *Optik für Ingenieure und Naturwissenschaftler*. München: Fachbuchverlag Leipzig im Carl Hanser Verlag.

Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley.

Part II

Fourier optics

Chapter 3
Fourier Optics

Markus Lippitz
November 8, 2023

By the end of this chapter you should be able to explain and experimentally demonstrate the filtering of spatial frequencies.

Overview

Fourier transformation simplifies the description of light, especially when it passes through obstacles, as in diffraction. The action of a lens also involves a Fourier transform. This is the field of *Fourier optics*. I will follow chapter 4 of Saleh and Teich, 1991 here. Another good source is Goodman, 2005. Note that books (as Saleh & Teich) from the engineering side of optics use $j = -i = -\sqrt{-1}$ instead of i. Sometimes this j is even written as i, so engineering is the complex conjugate of physics.

We will briefly lay the foundations of Fourier optics and then discuss diffraction and optical Fourier transform through a lens. For our purposes it is sufficient to consider scalar waves, i.e. we ignore the vectorial nature of the electric (or magnetic) field of light and use only a complex scalar value at each point in space to describe light. We will need fundamental properties of the Fourier transformation, as de scribed in Appendix A. For numerical applications Appendix B might be useful.

Spatial frequencies

Let us start with a plane wave

$$U(\boldsymbol{r}) = A e^{i\boldsymbol{k}\cdot\boldsymbol{r}} \quad \text{with} \quad k = |\boldsymbol{k}| = \frac{2\pi}{\lambda} \quad . \tag{3.1}$$

We assume that all three components of $\boldsymbol{k}$ are real (*far-field optics* in contrast to *near-field optics*), but the amplitude A might be complex. The wave vector $\boldsymbol{k}$ makes the angles $\Theta_{x,y}$ with the x–z and the y–z plane, respectively, with

$$\sin\Theta_x = \frac{k_x}{k} \quad . \tag{3.2}$$

In the $z = 0$ plane, the field is

$$U(x,y,0) = f(x,y) = A\, e^{2\pi\, i(\nu_x x + \nu_y y)} \tag{3.3}$$

with the *spatial frequencies* ν_x and ν_y

$$\nu_{x,y} = \frac{k_{x,y}}{2\pi} = \frac{1}{\Lambda_{x,y}} \tag{3.4}$$

and the period of the field $\Lambda_{x,y}$ in the x and y direction. And of course all this is related, i.e.,

$$\sin\Theta_x = \frac{k_x}{k} = \lambda\nu_x = \frac{\lambda}{\Lambda_x} \tag{3.5}$$

and similar for the y direction. The assumption of all-real $\boldsymbol{k}$ components makes sure that for all combinations of k_x,k_y, k_z an angle Θ can be found, i.e., the right side of the equation is real and below one in absolute value.

We will almost always make the *paraxial approximation* assuming that the wave vector is roughly parallel to the z-direction, the angles $\Theta_{x,y}$ are thus small, and $k_{x,y} \ll k$. Then we can omit the sine in the last equation and get

$$\Theta_x \approx \frac{k_x}{k} = \lambda\nu_x = \frac{\lambda}{\Lambda_x} \quad . \tag{3.6}$$

What happened here? The combination of all-real $\boldsymbol{k}$ components, i.e., optical far-field, and fixed wavelength λ removes one degree of freedom in the three components of the wave vector. As long as we know the wavelength and we know that the plane wave is nicely propagating, only two real values are enough to fully describe it. These two values could be the angles $\Theta_{x,y}$, or the spatial frequencies $\nu_{x,y}$ or the $\Lambda_{x,y}$.

Transmittance function

A plane wave of amplitude one is traveling in $+z$ direction. At $z = 0$ it is transmitted through a thin optical element with the complex transmittance function $f(x, y)$ with

$$f(x, y) = e^{2\pi\, i(\nu_x x + \nu_y y)} \quad . \tag{3.7}$$

Directly after this plate, the optical field is $U(x, y, 0) = f(x, y)$, i.e., the field is modulated by the transmittance function. We know from above that such a field is traveling in the direction given by the $\Theta_{x,y}$ or equally by the spatial frequencies $\nu_{x,y}$. The field is thus diffracted in this direction.[1]

[1] This is not an optical grating yet, as this would change the amplitudes only, i.e., have a real-valued tranmittance function.

In general, if the transmittance function f would have an arbitrary shape, it could be decomposed into a sum of harmonic functions. Each harmonic component would diffract a part of the plane wave into its direction. So when we express f by its Fourier transform F

$$f(x, y) = \mathcal{FT}\left\{F(\nu_x, \nu_y)\right\} = \iint F(\nu_x, \nu_y)\, e^{2\pi\, i(\nu_x x + \nu_y y)}\, d\nu_x d\nu_y \tag{3.8}$$

then we get

$$U(x, y, 0) = \iint F(\nu_x, \nu_y)\, e^{2\pi\, i(\nu_x x + \nu_y y)}\, d\nu_x d\nu_y \quad . \tag{3.9}$$

This becomes useful when calculating the field *at any point in space*, i.e., by including the z coordinate:

$$U(x, y, z) = \iint F(\nu_x, \nu_y)\, e^{2\pi\, i(\nu_x x + \nu_y y)}\, e^{ik_z z}\, d\nu_x d\nu_y \quad , \tag{3.10}$$

where k_z now depends on the integrating variables

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2} = 2\pi\sqrt{\frac{1}{\lambda^2} - \nu_x^2 - \nu_y^2} \quad . \tag{3.11}$$

Again the requirement of propagating waves entails $\nu_x^2 + \nu_y^2 < 1/\lambda^2$, so not all Fourier components of F play a role.

Transfer function and impulse response

Let us first introduce the concepts with electric circuits such as an RC-filter. One can define a transfer function $H(\omega)$ that relates the frequency spectrum $F(\omega)$ at the input (of the filter) with that at the output

$$G(\omega) = F(\omega) \cdot H(\omega) \quad . \tag{3.12}$$

In time domain, the impulse response $h(t)$ is another description. The signal $f(t)$ at the input results in an output $g(t)$

$$g(t) = \int h(\tau) f(t-\tau) d\tau \quad , \tag{3.13}$$

where causality requires that $h(t)$ is zero for $t < 0$. The interesting point is that not only the signals f and g are connected to their Fourier transforms F and G, but also the transfer function H is the Fourier transform of the impulse response h. A Fourier transform converts a product into a convolution, and vice versa.

Transfer function of free space

We now apply this scheme to spatial frequencies describing a superposition of plane waves. Letting the wave propagate by a distance d from a source plane $f(x,y) = U(x,y,0)$ to a target plane $g(x,y) = U(x,y,d)$, how do the spatial amplitudes F and G relate? Looking at eq. 3.10, we see that it is just the last exponential function that depends on z, but we need to take eq. 3.11 into account. Together we find

$$H(\nu_x, \nu_y) = \exp\left(2\pi\, i\, d \sqrt{\frac{1}{\lambda^2} - \nu_x^2 - \nu_y^2}\right) \quad . \tag{3.14}$$

For spatial frequencies $\nu_x^2 + \nu_y^2 < 1/\lambda^2$, i.e., within a circle of radius $1/\lambda$, the magnitude does not change ($|H| = 1$), only the phase changes. Outside this circle, the magnitude drops exponentially width d, as the square-root becomes imaginary. These waves are called *evanescent waves*, as they do not propagate and only exist in the near-field.

High spatial frequencies ν near $1/\lambda$ are far away from the paraxial approximation. In most cases it is sufficient to restrict ourself to low spatial frequencies $\ll 1/\lambda$. In this case, we can use the *Fresnel approximation* of the transfer function

$$H(\nu_x, \nu_y)_{\text{Fresnel}} = H_0 \exp\left(-\pi\, i\, d\lambda\, (\nu_x^2 + \nu_y^2)\right) \quad \text{with} \quad H_0 = e^{ikd} \quad . \tag{3.15}$$

The term H_0 factors out the trivial phase evolution due to propagation along the optical axis.

When we know the spatial frequency amplitudes F at $z = 0$, then we obtain G at $z = d$ by

$$G(\nu_x, \nu_y) = F(\nu_x, \nu_y) \cdot H(\nu_x, \nu_y) \quad . \tag{3.16}$$

We can Fourier transform the equation to obtain

$$g(x, y) = f(x, y) \otimes h(x, y) \tag{3.17}$$

where $\otimes$ signals a convolution. The impulse response of free space is in the Fresnel approximation

$$h(x, y)_{\text{Fresnel}} \approx h_0 \, \exp\left(ik\frac{x^2 + y^2}{2d}\right) \quad \text{with} \quad h_0 = -\frac{i}{\lambda d}\, e^{ikd} \quad . \tag{3.18}$$

Eq. 3.17 means that we get from one plane to the other by convolving each source point with a wave of shape h. This is equivalent to the Huygens principle, where each point should be a source of a spherical wave. When we take the paraxial approximation of a spherical wave we obtain $h(x, y)_{\text{Fresnel}}$.

Optical Fourier transform by propagation

Up to now we used the Fourier transform to simplify description of optical fields. In this section, we will show that the propagation of an optical field by a long enough distance allows to optically 'compute' the Fourier transform. We will find that the field in the target plane $g(x, y)$ is proportional to the Fourier transform F of the field in the source plane.

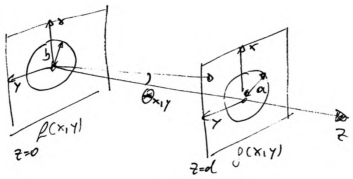

Figure 3.1: Fraunhofer condition

The Fourier components F of the field f in the source plane determine the direction of travel of the plane waves, as we have seen above. The problem is that a plane wave is everywhere in space. We need thus to find a condition for 'far enough' so that the individual pieces of the plane wave have separated enough. We do not only employ the paraxial approximation, i.e., that the wave vectors are not too inclined on the optical axis. The key point is that we also require the size of the source plane to be limited. This leads to the two conditions of the Fraunhofer approximation

$$N_F = \frac{a^2}{\lambda d} \ll 1 \quad \text{and} \quad N'_F = \frac{b^2}{\lambda d} \ll 1 \tag{3.19}$$

where the two N_F are the Fresnel numbers, and a,b are the radius of the relevant and allowed regions in the target and source planes, respectively. d is again the distance between the planes. The Fraunhofer approximation is a more severe restriction than the Fresnel approximation.

We start by writing down the convolution integral of eq. 3.17 in the Fresnel approximation

$$g(x, y) = f(x, y) \otimes h(x, y)_{\text{Fresnel}} \tag{3.20}$$

$$= h_0 \iint f(x', y') \, \exp\left(ik\frac{(x - x')^2 + (y - y')^2}{2d}\right) dx' dy' \quad . \tag{3.21}$$

The term $(x - x')^2$ in the exponent of the exponential function is multiplied out into three terms. We keep the mixed terms. Both squared terms can be neglected due to the Fraunhofer approximation. For example we get

$$\exp\left(i\pi\frac{x'^2 + y'^2}{\lambda d}\right) \approx 1 \tag{3.22}$$

as $N'_F \ll 1$. The terms without prime vanish due to $N_F \ll 1$. So we have

$$g(x,y) \approx h_0 \iint f(x',y') \exp\left(-i2\pi\frac{xx' + yy'}{\lambda d}\right) dx'dy' \quad . \tag{3.23}$$

We now identify the factor $x/\lambda d$ with the spatial frequency ν_x (y similar) and write

$$g(x,y) \approx h_0 F(\nu_x, \nu_y) = h_0 F\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right) \quad . \tag{3.24}$$

When we place a screen g at a distance fulfilling the Fraunhofer condition after a diffracting obstacle f, the interference pattern visible on the screen will be described by the Fourier transform F of f. This simplifies a lot the calculation of single slit, double slit and grating, as typically presented in the introductory optics lecture.

Test yourself

1. Convince yourself that the textbook solution, for example in Demtröder, can be obtained by a Fourier transform.
2. Estimate the required distance so that a typical diffraction grating fulfils the Fraunhofer condition.

Optical Fourier transform by a lens

The distance d required to stay within the Fraunhofer approximation can be prohibitively large. We will see here that a lens is able to shorten the distance between the grating and the screen and still keep the Fourier relation. This explains why spectrometers are not too long, but contain a lens or curved mirror.

From geometrical optics in the paraxial approximation we know already that a lens focuses a beam (angles Θ_x, Θ_y to the optical axis) on a point

$$(x,y) = (f\Theta_x, f\Theta_y) \tag{3.25}$$

in the focal plane, where f describes the focal length of the lens. A lens thus separates plane waves by their propagation direction. As in the beginning of the chapter, we can convert angles into optical frequencies and thus find that the field in the target plane g is proportional the the Fourier amplitude F

$$g(x,y) = \tilde{h}\, F(\nu_x, \nu_y) = \tilde{h}\, F\left(\frac{x}{\lambda d}, \frac{y}{\lambda d}\right) \quad . \tag{3.26}$$

The remaining question is the prefactor $\tilde{h}$. If it would depend of the spatial coordinates x and y, this would destroy the Fourier transform. To obtain $\tilde{h}$, we multiply the transfer functions of free space for the distance source plane to lens (length d) and lens to target plane (length f). And we need to multiply

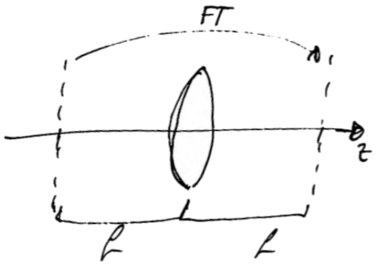

Figure 3.2: Optical Fourier transform by a lens

a transfer function for the lens, as the lens has a thickness profile $t(x, y)$ of a material with a certain index of refraction. All together one obtains[2]

$$\tilde{h}(x,y) = \tilde{h}_0 \exp\left(-i\pi \frac{(x^2+y^2)(d-f)}{\lambda f^2}\right) \quad \text{with} \quad \tilde{h}_0 = \frac{-i}{\lambda f} e^{ik(d+f)} \quad . \tag{3.27}$$

This factor becomes spatially constant when the condition $d = f$ is met. A lens thus performs an optical Fourier transform between is two focal planes. In a spectrometer, the grating sits in the front focal plane of the curved mirror (acting as a lens), the detector in its back focal plane.

[2] details in Saleh and Teich, 1991, chapter 4

Spatial filter

In addition to spectrometers, the spatial filter is another important application of a lens as a Fourier transform device. We consider a so-called 4f-system, see Saleh and Teich, 1991. All components are separated by one focal length f: a source plane f, a first lens, a filter plane p, a second lens and a target plane g. Both lenses are identical.

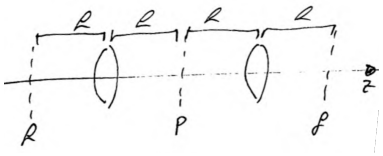

Figure 3.3: A $4f$ system can be used as spatial filter.

Let the transfer function p of the filter plane be $p(x, y) = 1$ for the beginning. Then the first lens Fourier transforms f into F in the filter plane. The filter does nothing and the second lens transforms back F into f, so that we get in the target plane what we started with , i.e., $f = g$. Of course this make the assumption that all plane waves nicely propagate, i.e., the spatial frequencies in f are small enough to cause only propagating plane waves.

The filter plane can be used to modify the Fourier components F. At position x in the p plane, only the Fourier component $\nu_x = x/(\lambda f)$ is present. We can put a mask $p(x, y)$, either just absorbing or with a complex transfer function in the filter plane. The overall transfer function of the 4f-system is then

$$H(\nu_x, \nu_y) = p(\lambda f \nu_x, \lambda f \nu_y) \quad , \tag{3.28}$$

ignoring an overall phase factor for the propagation.

An often used transfer function is a circular aperture. It removes all spatial frequencies above a certain threshold. In this way, one can clean up a laser beam, so that is follows the expected Gaussian profile even after transmission though many non-ideal optical elements.

The inverse filter, i.e. a opaque disc, acts as high-pass filter, increasing the edges in an optical image. A vertical slit lets only pass horizontal features in the image.

Resolution of a microscope

In an optical microscope, a sample is imaged on a detector by a system of lenses. Not all spatial frequencies are transmitted equally well. A fundamental limit is the transfer function of free space (eq. 3.15), which limits the spatial frequencies to $\nu_{x,y} \leq n/\lambda_0$. A short vacuum wavelength λ_0, i.e. blue or ultraviolet light, or a high refractive index n, i.e. immersion oil instead of air, shifts the limit to higher spatial frequencies. The highest observable spatial frequency is the wavelength of light in the medium.

A technical limitation sets in earlier. High spatial frequencies correspond to plane waves with a large angle θ to the optical axis. In the limiting case, the plane wave propagates perpendicular to the optical axes, i.e. parallel to the sample surface. The microscope objective has to capture as many of these rays as possible, i.e. it has to have a large aperture angle $\theta_{\text{objective}}$. This is quantified in the *numerical aperture*

$$NA = n \sin \theta_{\text{objective}} \tag{3.29}$$

so that the maximum spatial frequencies are $\nu_{x,y} \leq NA/\lambda_0$. Numerical apertures greater than one require the use of an immersion medium. Glycerin ($n = 1.47$) allows to obtain NA $\approx 1.3 \ldots 1.45$.

Another approach to the resolution of a microscope is the *point spread function* (PSF). It is the impulse response of the imaging system, i.e., the image of a point source in the sample plane. Diffraction at the circular aperture of the microscope objective leads to the Airy-pattern of the Bessel function J_1 (see A)

$$PSF(\rho) = \left(\frac{J_1(k_0\, NA\, \rho)}{k_0\, NA\, \rho} \right)^2 \tag{3.30}$$

where ρ is the radial coordinate in the sample plane, i.e., the point placed there appears to be larger. Two points are said to be resolvable if they are separated so that the image of one point falls into the first zero of the PSF of the second point. This results in the value

$$\Delta x = 0.61 \frac{\lambda_0}{NA} \quad . \tag{3.31}$$

Hands on: Align a pinhole

A spatial filter in the form of a pinhole at the combined focus of two lenses is often used to clean up a beam. Optical surfaces of lenses, mirrors, beam splitters and crystals are never perfectly flat, but contain more or less strong deviations from the ideal shape. This results in a more or less distorted shape of the beam, which deviates from the ideal Gaussian shape. These deviations cause high spatial frequencies in the beam. They distort the optical resolution because, for example, we would excite emitters that would otherwise not be in the ideal laser focus. It is therefore common to clean the beam at one or more points in an optical setup.

The design is that of a spatial filter as described above: two lenses with a combined focus. The focal lengths of the lenses can be different to adjust the beam diameter. At the position of the focus we place a pinhole, a metal foil with a hole of a few tens of microns in diameter. The smaller the diameter of the hole, the better the high spatial frequency is filtered, but also the overall transmission is reduced and laser power is lost. So you have to balance filtering and transmission. A good starting value is a radius of Δx as above.

The alignment is rather sensitive in a direction perpendicular to the beam and rather insensitive in a direction parallel to the beam. In the parallel direction, only distances on the scale of the ray range matter. In the perpendicular direction, it is often advisable to start with a larger diameter, optimize the

alignment, and then move to a smaller diameter. However, this assumes that we can change the pinhole without losing much of the alignment. Another approach is to center the pinhole first at a position away from the focus where the beam diameter is larger and the overlap of beam and pinhole is easier to achieve.

References

Goodman, Joseph W. (2005). *Introduction to Fourier optics*. 3. ed. Roberts.

Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley.

Part III

Light in matter

Chapter 4
Dielectric Materials

Markus Lippitz
November 24, 2023

By the end of this chapter you should be able to explain and experimentally demonstrate total internal reflection and the Brewster effect.

Overview

With this chapter we begin to consider the optical properties of media beyond their refractive index. The physics of the medium will play a role and have consequences for the propagation of light. To be able to do this, we also have to describe light as an electromagnetic wave, with three components for the electric and magnetic field, and not only as a scalar wave as in the last chapters. This will lead to the phenomenon of absorption and dispersion. We will also be able to assign a value to the amplitude of the reflected and transmitted waves at an interface. These topics are described in chapter 5 and 6 of Saleh and Teich, 1991 and chapter 3 of E. Hecht, 2017.

Maxwells equations

For completeness, let us start with the Maxwell equations in their macroscopic form

$$\nabla \boldsymbol{D} = \rho \tag{4.1}$$

$$\nabla \boldsymbol{B} = 0 \tag{4.2}$$

$$\nabla \times \boldsymbol{E} = -\dot{\boldsymbol{B}} \tag{4.3}$$

$$\nabla \times \boldsymbol{H} = \dot{\boldsymbol{D}} + \boldsymbol{j} \quad . \tag{4.4}$$

Matter comes in by the respective material equations

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \boldsymbol{P} = \epsilon\epsilon_0 \boldsymbol{E} \tag{4.5}$$

$$\boldsymbol{H} = \frac{1}{\mu_0}\boldsymbol{B} - \boldsymbol{M} = \frac{1}{\mu\mu_0}\boldsymbol{B} \tag{4.6}$$

$$\boldsymbol{j} = \sigma \boldsymbol{E} \quad . \tag{4.7}$$

Note that I use a unit-free dielectric function ϵ. In literature, one finds different other methods to write the term $\epsilon\epsilon_0$. At the second equal sign we have assumed in each case a linear and isotropic medium. Let us define these and similar terms:

linear The relation between the electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ and the polarization $\boldsymbol{P}(\boldsymbol{r}, t)$ is linear.

isotropic The relation between $\boldsymbol{E}$ and $\boldsymbol{P}$ is independent of the direction of $\boldsymbol{E}$. This also means that $\boldsymbol{E}$ and $\boldsymbol{P}$ are parallel.

homogeneous The relation between $\boldsymbol{E}$ and $\boldsymbol{P}$ is independent of the position $\boldsymbol{r}$.

nondispersive The relation between $\boldsymbol{E}$ and $\boldsymbol{P}$ is instantaneous, i.e., it depends only on the value of $\boldsymbol{E}$ at time t, but not on earlier times. As we will see, this is equivalent to saying that the relation does not depend on the frequency ω of light. This is a thought model and is only approximated by real materials.

local The relation between $\boldsymbol{E}$ and $\boldsymbol{P}$ depends only on the value of $\boldsymbol{E}$ at one point $\boldsymbol{r}$, not at other points. This is also called *spatially nondispersive*. Optical active media (next chapter) are nonlocal.

Wave equations

When we assume a source-free medium ($\boldsymbol{j} = 0, \rho = 0$), one can derive the wave equation for an isotropic and linear medium

$$\nabla^2 \boldsymbol{E} = \frac{n^2}{c_0^2} \ddot{\boldsymbol{E}} \quad \text{with} \quad c_0^2 = \frac{1}{\mu_0 \epsilon_0} \quad ; \quad n^2 = \epsilon \quad ; \quad \mu \approx 1 \tag{4.8}$$

A similar equation exists for $\boldsymbol{H}$. The individual vector components of the electrical and magnetic field fulfil thus the scalar wave equation of chapter 2.

The flow of electromagnetic energy is described by the Poynting vector[1]

[1] John Henry Poynting, 1852–1914

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{4.9}$$

The intensity I of a wave on a surface with normal $\boldsymbol{n}$ is the temporal average of the Poynting vector, i.e.

$$I = \langle \boldsymbol{S} \cdot \boldsymbol{n} \rangle_T = \frac{c\, n\, \epsilon_0}{2} |\boldsymbol{E}_0|^2 = \frac{1}{2\eta} |\boldsymbol{E}_0|^2 \tag{4.10}$$

where $\boldsymbol{E}_0$ is the amplitude of the electrical field and $\eta = \sqrt{\mu\mu_0/(\epsilon\epsilon_0)}$ the impedance of the medium. For vacuum, $\eta_0 \approx 377\,\Omega$. An intensity of 10 W/cm^2 corresponds to an electric field of about 87 V/m.

The Poynting vector fulfills the Poynting theorem: the flow of energy through a surface enclosing a volume either changes the energy density within that volume or performs work on magnetic or electric dipoles. As equation:

$$\nabla \boldsymbol{S} = -\frac{\partial}{\partial t} \left(\frac{1}{2} \epsilon\epsilon_0 \boldsymbol{E}^2 + \frac{1}{2} \mu\mu_0 \boldsymbol{H}^2 \right) + \boldsymbol{E} \cdot \frac{\partial \boldsymbol{P}}{\partial t} + \mu_0 \boldsymbol{H} \cdot \frac{\partial \boldsymbol{M}}{\partial t} \tag{4.11}$$

As with scalar waves, we find different solutions to the wave equation. The plain wave also exists as electromagnetic wave:

$$\boldsymbol{H}(\boldsymbol{r}, t) = \boldsymbol{H}_0\, e^{i(\boldsymbol{k}\boldsymbol{r} - \omega t)} \tag{4.12}$$

$$\boldsymbol{E}(\boldsymbol{r}, t) = \boldsymbol{E}_0\, e^{i(\boldsymbol{k}\boldsymbol{r} - \omega t)} \tag{4.13}$$

with $|\boldsymbol{k}| = k = 2\pi n/\lambda_0$ and $\boldsymbol{H}_0$, $\boldsymbol{B}_0$ and $\boldsymbol{k}$ orthogonal on each other. The electromagnetic wave is thus a *transversal* wave.

The vectorial electromagnetic forms of paraboloidal wave and Gaussian beams can be constructed by vectorizing the scalar waves $u(\boldsymbol{r})$ of the preceding chapters:

$$\boldsymbol{E}(\boldsymbol{r}) = \mathcal{E}_0 \left(-\hat{\boldsymbol{x}} + \frac{x}{z + iz_0}\hat{\boldsymbol{z}} \right) u(\boldsymbol{r}) \tag{4.14}$$

where $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are unit vectors pointing in x and z direction, respectively, and $\mathcal{E}_0$ is a scalar amplitude. z_0 is set to zero for a paraboloidal wave.

Phenomenological approach to absorption

Let us begin by describing absorption in media without attributing a microscopic origin. The susceptibility χ is complex-valued , i.e. $\chi = \chi' + i\chi''$ and thus the dielectric function

$$\epsilon = 1 + \chi = 1 + \chi' + i\chi'' = \epsilon' + i\epsilon'' \tag{4.15}$$

This means that the wave number k will become complex, too

$$k = \frac{\omega}{c} = k_0\sqrt{\epsilon} = k_0\sqrt{1 + \chi' + i\chi''} = \beta + i\frac{\alpha}{2} \tag{4.16}$$

The meaning of the real-valued α and β will become clear when we use this definition in a plane wave:

$$\mathcal{E}(z,t) = \mathcal{E}_0\, e^{i(kz-\omega t)} = \mathcal{E}_0\, e^{-i\omega t}\, e^{i\beta z}\, e^{-\alpha z/2} \tag{4.17}$$

The intensity of this waves thus drops as

$$I(z) \propto |\mathcal{E}(z,t)|^2 = |\mathcal{E}_0|^2\, e^{-\alpha z} \tag{4.18}$$

α is thus the absorption coefficient[2]. Positive α means a decay of intensity, negative α would mean a gain, as in a laser. β describes the progression of the phase or wave fronts. It is related to the real part n of the refractive index[3] by $\beta = nk_0$. Everything together we have

[2] or attenuation or extinction coefficient

[3] I use the form $\tilde{n} = n + i\kappa$.

$$n + i\kappa = \frac{\beta}{k_0} + i\,\frac{1}{2}\,\frac{\alpha}{k_0} = \pm\sqrt{1 + \chi' + i\chi''} \tag{4.19}$$

The sign of the square root is chosen such that a positive (absorbing) χ'' leads to a positive (absorbing) α, independent of the sign of χ'. As we will see below $\chi' < 0$ is possible, e.g., near resonances.

It is convenient to have approximate forms of eq. 4.19 for the limiting cases of weak and strong absorption

$$\chi'' \ll 1 + \chi' \rightarrow \qquad n \approx \sqrt{1 + \chi'} \qquad \alpha \approx \frac{k_0}{n}\chi'' \tag{4.20}$$

$$\chi'' \gg |1 + \chi'| \rightarrow \qquad n \approx \sqrt{\chi''/2} \qquad \alpha \approx 2k_0\sqrt{\chi''/2} \tag{4.21}$$

The Kramers-Kronig relations

So far, we have only discussed the relationship between the applied external field $E(t)$ and the resulting polarization $P(t)$ for 'monochromatic' fields of the type $\exp(-i\omega t)$, i.e. for a precisely defined frequency ω:

$$P(t) = \chi(\omega)\epsilon_0 E(t) \quad \text{for} \quad E(t) = E_0 e^{-i\omega t} \quad . \tag{4.22}$$

This gave the frequency dependence of $\chi(\omega)$. We can generalize this for any time evolution of the field $E(t)$. The susceptibility is the *impulse response* of the material, the memory so to speak:

$$P(t) = \epsilon_0 \int_{-\infty}^{+\infty} \chi(\Delta t = t - t')\, E(t')\, dt' \quad \text{for} \quad E(t) = \text{any} \quad . \tag{4.23}$$

The polarization P now, i.e. at time t, depends on the electric field at all other times t'. How strong the fields are depends only on the time interval Δt. Causality requires that the polarization 'now' does not depend on the field amplitudes in the future. Therefore $\chi(\Delta t = t - t' < 0)$ must be zero. This means that the susceptibility $\chi(\Delta t)$ is complex, but known over half of the time ray as fixed to zero. This has consequences for the Fourier transform, i.e. for $\chi(\omega)$.

These consequences can be derived with the help of function theory[4] and are the Kramers-Kronig relations. The following relationship exists between the real (χ') and imaginary (χ'') parts of the susceptibility if they obey causality:

[4] see also Appendix A of Yariv, 1989

$$\chi'(\nu) = \frac{2}{\pi} P \int_0^\infty \frac{s\, \chi''(s)}{s^2 - \nu^2}\, ds \tag{4.24}$$

$$\chi''(\nu) = \frac{2}{\pi} P \int_0^\infty \frac{\nu\, \chi'(s)}{\nu^2 - s^2}\, ds \quad . \tag{4.25}$$

P denotes the Cauchy principal value integral. Similar relationships also exist for $\chi(\omega)$ and $\epsilon(\omega)$ as well as for all other variables that are subject to causality.

In principle, it is therefore sufficient to measure the real part of the susceptibility $\chi(\omega)$ in order to determine the imaginary part and thus the complete complex-valued function. Unfortunately, however, the integrals in Eq 4.25 run over the entire frequency range from zero to infinity, which is of course not accessible experimentally. The Kramers-Kronig relations can still be used sensibly by making appropriate assumptions about the course outside the measured interval.

Lorentz oscillator model

The response of matter to an electric field is governed by the charged ions and electrons. Restoring forces lead to resonances depending on the frequency of the optical field. In the infrared, bound ions resonate, while in the ultraviolet, bound electrons dominate.

The Lorentz oscillator model is a simple model that can be used to describe the frequency dependence of the dielectric function in the vicinity of resonances. In a damped harmonic oscillator (mass m, damping constant γ, natural frequency ω_0), the mass is deflected by a periodic electric field (amplitude E_0, frequency ω) by x because the mass carries a charge e. All together

$$m\ddot{x} + \gamma\dot{x} + m\omega_0^2 x = eE_0 e^{-i\omega t} \quad . \tag{4.26}$$

The stationary solution of this differential equation is

$$x(t) = \frac{e\, E_0}{m(\omega_0^2 - \omega^2) - i\gamma\omega}\, e^{-i\omega t} \quad . \tag{4.27}$$

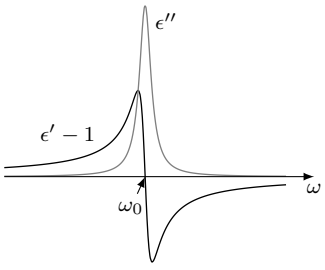

Figure 4.1: Frequency dependence of the real and imaginary parts of the Lorentz oscillator. The real and imaginary parts of the complex-valued refractive index $\tilde{n}$ look qualitatively the same.

The macroscopic polarization P is the sum of all microscopic polarizations, i.e.

$$P(t) = N\,e\,x(t) = (\epsilon - 1)\epsilon_0\,E_0 e^{-i\omega t} = \chi\epsilon_0 E(t) \quad . \tag{4.28}$$

This results in the dielectric function

$$\epsilon(\omega) = 1 + N\alpha = 1 + \frac{Ne^2}{\epsilon_0}\frac{1}{m(\omega_0^2 - \omega^2) - i\gamma\omega} = \epsilon' + i\epsilon'' \quad . \tag{4.29}$$

Explicit real and imaginary parts are

$$\epsilon' = 1 + \frac{Ne^2}{\epsilon_0}\frac{m(\omega_0^2 - \omega^2)}{m^2(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} \tag{4.30}$$

$$\epsilon'' = \frac{Ne^2}{\epsilon_0}\frac{\gamma\omega}{m^2(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} \quad . \tag{4.31}$$

Test yourself

1. Analogous to Figure 4.2, show the frequency dependence of the components of the refractive index, i.e. of n and k.

2. Approximate the real and imaginary parts of ϵ near resonance at ω_0 as a function of $\Delta\omega = \omega - \omega_0$. In the case of the real part, only the range $|\Delta\omega| \gg \gamma/m$ is of interest.

Normal and anomalous dispersion

The visible spectral region is at a higher frequency than the resonance of the bound ions in the infrared, but at a lower frequency than that of the bound electrons in the ultraviolet. The real part n of the refractive index increases with frequency, i.e. $n(\text{blue}) > n(\text{red})$ (see Fig. 4.2). This is called 'normal' dispersion. It causes red light to deviate less than blue light in a prism and to be focused by a lens at a greater distance. On the energetically 'other' side of a resonance, the opposite behavior can be observed, 'anomalous dispersion'.

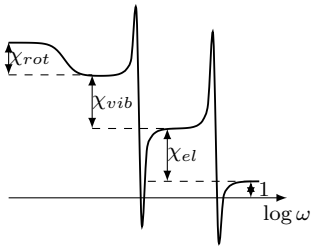

Figure 4.2: The visible spectral range lies between two resonances.

The Lorentz-shaped resonance can be shown in a demonstration experiment. The imaginary part of the dielectric function Fig. 4.2 determines the absorption and thus the line shape in the absorption spectrum of atoms or molecules. The real part determines the dispersion, i.e. the refractive index of a medium. A simple method of determining the refractive index is to use a prism made of the material to be examined. In a prism, the deflection of the light beam is proportional to the difference of the refractive index inside compared to outside (actually always air $\approx$ vacuum). However, the electronic resonance must also be shifted from the ultraviolet to the visible. In the experiment, a prism made of sodium vapor is used for this purpose. The strong absorption of the sodium D lines at a wavelength of around 589 nm produces a highly visible effect.

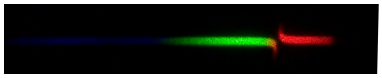

Figure 4.3: Anomalous dispersion in sodium vapor.

Sodium vapor is generated in an evacuated tube by strongly heating solid sodium. The tube is heated from below and cooled from above so that the vapor density decreases towards the top. This corresponds to a prism with its tip pointing upwards. Here, too, the effective glass thickness decreases towards the top when averaged over the entire beam path. The light beam is then passed through a glass prism with a vertical axis to create a horizontal

wavelength axis. The result is a spectrum as shown in the adjacent figure. The horizontal axis is proportional to the wavelength, the vertical axis to the deviation of the refractive index from unity. The spectrum is interrupted at the absorption line itself because the sodium vapor completely absorbs the light there. It can be seen that the refractive index falls below unity at the higher energy side of the resonance.

Reflection and transmission

Now that we can describe matter, we want to know how much of a wave is transmitted through an interface between two media and how much is reflected. Let us assume that the ray travels in the xz-plane. The surface is an xy plane. As we will see in the next chapter on polarization optics, it is sufficient to examine the response for linearly polarized light, where the direction of polarization is either in the plane defined by the rays (xz) or perpendicular to it (y). The first case is called p-polarized (p for parallel) or transverse magnetic (TM), since the magnetic field is orthogonal to the xz-plane of incidence. The second case is called s-polarized (s as 'senkrecht', perpendicular) or transverse electric (TE) because the electric field is perpendicular to the xz-plane of incidence.

At the interface, the sum of incident and reflected wave on each side has to match the transmitted wave on the other side. Matching means that the phases need to agree and the amplitudes need to follow the continuity conditions of electromagnetic fields. The phase argument defines the angles of reflected and transmitted waves, luckily identical to geometrical optics. The continuity argument defines the amplitudes as reflection r_{12} and transmission t_{12} coefficients for the electric field. These are called the *Fresnel equations*. This calculation can be found in textbooks on electrodynamics, e.g, Nolting, 2016.

We follow here Novotny and B. Hecht, 2012, who follow Born and Wolf, 2002, especially in the direction of the field vectors, see Fig. 2.2 in Novotny and B. Hecht, 2012. In this definition, r^s and r^p differ at normal incidence by a factor of -1. We assume non-magnetic materials ($\mu = 1$) and describe the propagation direction by the z component k_z of the wave vector $\boldsymbol{k}$. For a wave traveling from medium 1 towards medium 2 we get

$$r_{12}^s = \frac{k_{z,1} - k_{z,2}}{k_{z,1} + k_{z,2}} = -r_{21}^s \tag{4.32}$$

$$t_{12}^s = \frac{2\,k_{z,1}}{k_{z,1} + k_{z,2}} = \frac{k_{z,1}}{k_{z,2}}\,t_{21}^s \tag{4.33}$$

$$r_{12}^p = \frac{\epsilon_2 k_{z,1} - \epsilon_1 k_{z,2}}{\epsilon_2 k_{z,1} + \epsilon_1 k_{z,2}} = -r_{21}^p \tag{4.34}$$

$$t_{12}^p = \frac{2\sqrt{\epsilon_1\epsilon_2}\,k_{z,1}}{\epsilon_2 k_{z,1} + \epsilon_1 k_{z,2}} = \frac{k_{z,1}}{k_{z,2}}\,t_{21}^p \quad . \tag{4.35}$$

We could also write these coefficients in terms of angle of incidence θ with

$$\theta = \arcsin\frac{k_x}{nk_0} = \arcsin\sqrt{1 - \left(\frac{k_z}{nk_0}\right)^2} \quad . \tag{4.36}$$

This would also hold in the case of evanescent waves ($k_x > nk_0$) when we

allow complex angles θ. We nowhere need that θ is a geometrical angle. We only need that $n \sin\theta$ is the same on both sides.

Figure 4.4 shows the amplitude and phase of the reflection coefficient r for a reflection at an air–glass and a glass–air interface. Coming from the less dense medium, we find a zero reflectivity for p-polarized light. This is the Brewster effect: the incident field induces oscillating dipoles at the surface of the medium, aligned with the polarization direction of the field. A dipole emits radiation in all directions, but not in the direction of its oscillation. If this direction of oscillation is in the expected direction of the outgoing wave, the amplitude of that wave must be zero.

When light is incident from the dense medium, we observe total internal reflection for both polarization directions above a critical angle. As we have already seen with Fourier optics, in these cases the in-plane component of the wave vector on the glass side is larger than the total length of the wave vector on the air side. So we have evanescent waves on the air side. Note that even though the reflectivity $|r|$ is always one above the critical angle, the reflected field acquires a phase that depends on the angle of incidence and the direction of polarization.

These coefficients are for the fields. The reflected power is the fraction $R = |r|^2$ of the incident power. The transmitted power is the fraction T with

$$T = 1 - R \neq |t|^2 \quad . \tag{4.37}$$

This inequality is caused by the differing impedance on both sides of the interface and the differing direction of travel due tue refraction. The orientation of the 'power meter surface' needs to change. Both can be corrected so that we get for a wave traveling from 1 to 2

$$T = \frac{n_2 \cos\theta_2}{n_1 \cos\theta_1} |t|^2 \quad . \tag{4.38}$$

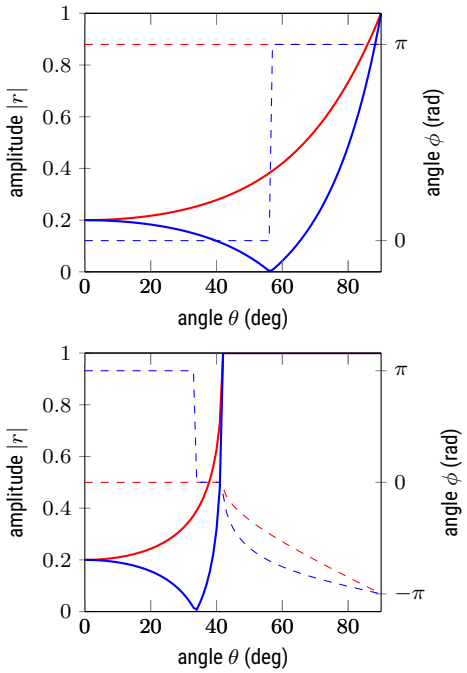

Figure 4.4: Fresnel coefficients $r = |r|e^{i\phi}$ for external (top) and internal (bottom) reflection at an air–glass interface. red: s-polarized, blue: p-polarized. dashed: phase

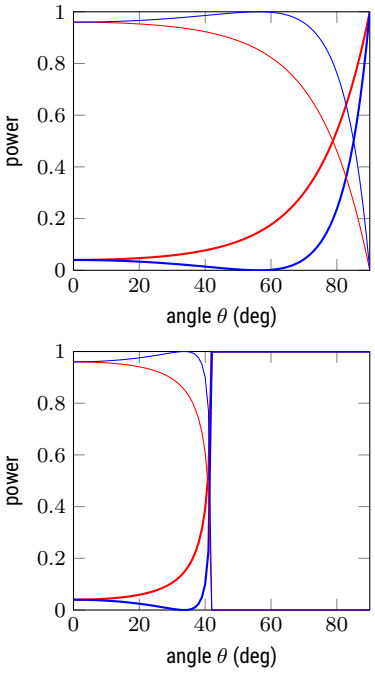

Figure 4.5: Reflected (thick) and transmitted (thin) power for external (top) and internal (bottom) reflection at an air–glass interface. red: s-polarized, blue: p-polarized

References

Born, Max and Emil Wolf (2002). *Principles of optics*. 7. (expanded) ed., reprinted with corr. Cambridge [u.a.]: Cambridge Univ. Press.

Hecht, Eugene (2017). *Optics*. Fifth edition, global edition. Boston: Pearson.

Nolting, Wolfgang (2016). *Theoretical Physics 3 Electrodynamics*. Springer.

Novotny, Lukas and Bert Hecht (2012). *Principles of nano-optics*. 2. ed. Cambridge Univ. Press.

Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley.

Yariv, Amnon (1989). *Quantum electronics*. 3. ed. New York: Wiley.

Appendices

Appendix A
Fourier transformation

Markus Lippitz
September 18, 2023

Overview

It is useful and helpful to have an intuitive approach to the Fourier transform. The bottom line is that in experimental physics one rarely needs to actually calculate a Fourier transform. Very often it is sufficient to know a few frequently occurring Fourier pairs and to combine them with simple rules. This is what I want to present here. A very nice and much more detailed presentation can be found in Butz, 2015. I will follow his notation here.

Before we get to Fourier pairs, however, we need to lay down some foundations.

Fourier series: a periodic function and its Fourier coefficients

We first consider everything here in one dimension in time or frequency space with the variables t and $\omega = 2\pi\nu$. Let the function $f(t)$ be periodic in time with period T, i.e.

$$f(t) = f(t+T) \quad . \tag{A.1}$$

Then this can be written as a Fourier series

$$f(t) = \sum_{k=-\infty}^{\infty} C_k \, e^{i\,\omega_k\,t} \quad \text{with} \quad \omega_k = \frac{2\pi\,k}{T} \tag{A.2}$$

and the Fourier coefficients

$$C_k = \frac{1}{T} \int_{-T/2}^{T/2} f(t)\; e^{-i\,\omega_k\,t}\, dt \quad . \tag{A.3}$$

Note the negative sign in the exponential function in contrast to the equation before. For reel-valued functions $f(t)$, 'opposite' C_k are conjugate-complex, so $C_k = C^{\star}_{-k}$. For $k < 0$ the frequencies ω_k are negative, but this is not a problem.[1] Thus, the zeroth coefficient C_0 is just the time average of the function $f(t)$.

[1] One could alternatively require $k \geq 0$ and apply a $\sin$ and $\cos$ series.

An arbitrary function and its Fourier transform

Now we remove the restriction to periodic functions $f(t)$ by letting the period T go to infinity. This turns the sum into an integral and the discrete ω_k become continuous. Thus

$$F(\omega) = \int_{-\infty}^{+\infty} f(t)\, e^{-i\omega\, t}\, dt \tag{A.4}$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(\omega)\, e^{+i\omega\, t}\, d\omega \quad . \tag{A.5}$$

Here, the first equation is the forward transformation (minus sign in the exponent), and the second is the reverse transformation (plus sign in the exponent). The symmetry is broken by the 2π. But this is necessary if one wants to keep $F(\omega = 0)$ as mean[2]. Alternatively, we could formulate all this with ν instead of ω, but then we would have a 2π in many more places, though not before the integral.

[2] $F(0) = \int f(t)\, dt$ without $1/T$ in front of it is meant here by Butz as mean!

Sidenote: Delta Function

The delta function can be written as

$$\delta(x) = \lim_{a\to 0} f_a(x) \quad \text{with} \quad f_a(x) = \begin{cases} a & \text{if } |x| < \frac{1}{2a} \\ 0 & \text{other} \end{cases} \tag{A.6}$$

or as

$$\delta(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{+i\, x\, y}\, dy \quad . \tag{A.7}$$

An important property is that the delta function selects a value, i.e.

$$\int_{-\infty}^{+\infty} \delta(x)\, f(x)\, dx = f(0) \quad . \tag{A.8}$$

Important Fourier pairs

It is very often sufficient to know the following pairs of functions and their Fourier transforms. I write them here, following Butz, as pairs in t and ω (not $\nu = \omega/(2\pi)$). In the same way, one could have written pairs in x and k. The important question is whether a 2π appears in the exponential function of the plane wave or not. So

$$e^{i\omega t} \quad \text{and} \quad e^{ikx} \quad \text{, but} \quad e^{i2\pi\nu t} \quad . \tag{A.9}$$

Further, I follow here the convention made above about the asymmetric distribution of the 2π between forward and reverse transformations. If you distribute them differently, then of course the prefactors change. A good overview of many more Fourier pairs in various '2π' conventions can be found in the English Wikipedia under 'Fourier transform'. In their nomenclature, the Butz convention used here is 'non-unitary, angular frequency'.

constant and delta function $f(t) = a$ becomes $F(\omega) = a\, 2\pi\, \delta(\omega)$ and $f(t) = a\, \delta(t)$ becomes $F(\omega) = a$. This is again the asymmetric 2π.

rectangle and sinc The rectangle function of width b becomes a sinc[3], the sinus cardinalis. So from

$$f(t) = \mathrm{rect}_b(t) = \begin{cases} 1 & \text{for} \quad |t| < b/2 \\ 0 & \text{other} \end{cases} \tag{A.10}$$

we get

$$F(\omega) = b\,\frac{\sin \omega b/2}{\omega b/2} = b\,\mathrm{sinc}(\omega b/2) \quad . \tag{A.11}$$

Gaussian The Gaussian function is preserved under Fourier transform. Its width changes into the reciprocal value. So from a Gauss function of area one

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{t}{\sigma}\right)^2} \tag{A.12}$$

we get

$$F(\omega) = e^{-\frac{1}{2}(\sigma\,\omega)^2} \quad . \tag{A.13}$$

(two-sided) exponential decay and Lorentz curve From a curve decaying exponentially at both positive and negative times

$$f(t) = e^{-|t|/\tau} \tag{A.14}$$

we obtain the Lorentz curve

$$F(\omega) = \frac{2\tau}{1 + \omega^2\,\tau^2} \quad . \tag{A.15}$$

one-sided exponential decay As a side note, here the one-sided exponential decay

$$f(t) = \begin{cases} e^{-\lambda t} & \text{for} \quad t > 0 \\ 0 & \text{other} \end{cases} \quad . \tag{A.16}$$

It will become

$$F(\omega) = \frac{1}{\lambda + i\,\omega} \tag{A.17}$$

and it is therefore complex-valued. Its magnitude squared is again a Lorentz function

$$|F(\omega)|^2 = \frac{1}{\lambda^2 + \omega^2} \tag{A.18}$$

and the phase is $\phi = -\omega/\lambda$.

One-dimensional point lattice An equidistant chain of points or delta functions remains an equidistant chain under Fourier transform. The distances take the reciprocal value. So from

$$f(t) = \sum_n \delta(t - \delta t\, n) \tag{A.19}$$

we get

$$F(\omega) = \frac{2\pi}{\delta t} \sum_n \delta\left(\omega - n\frac{2\pi}{\Delta t}\right). \quad . \tag{A.20}$$

Three-dimensional cubic lattice A three-dimensional primitive cubic lattice of side length a makes the transitions to a primitive cubic lattice of side length $2\pi/a$. A face-centered cubic lattice with lattice constant a of conventional unit cell is converted to a space-centered cubic lattice with lattice constant $4\pi/a$ and vice versa.

[3] sometimes $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$ is defined, especially when ν and not ω is used as conjugate variable.

Theorems and properties of the Fourier transform

In addition to the Fourier pairs, we need a few properties of the Fourier transform. In the following, let $f(t)$ and $F(\omega)$ be Fourier conjugates and likewise g and G.

linearity The Fourier transform is linear

$$a\,f(t) + b\,g(t) \quad \leftrightarrow \quad a\,F(\omega) + b\,G(\omega) \quad . \tag{A.21}$$

shift A shift in time implies a modulation in frequency and vice versa.

$$f(t-a) \quad \leftrightarrow \quad F(\omega)\,e^{-i\omega a} \tag{A.22}$$

$$f(t)\;e^{-i\omega_0 t} \quad \leftrightarrow \quad F(\omega+\omega_0) \quad . \tag{A.23}$$

scaling

$$f(a\,t) \quad \leftrightarrow \quad \frac{1}{|a|}\,F\left(\frac{\omega}{a}\right) \quad . \tag{A.24}$$

convolution and multiplication Convolution is converted into a product, and vice versa

$$f(t) \otimes g(t) = \int f(\zeta)g(t-\zeta)d\zeta \quad \leftrightarrow \quad F(\omega)\,G(\omega) \tag{A.25}$$

and

$$f(t)\,g(t) \quad \leftrightarrow \quad \frac{1}{2\pi}\,F(\omega) \otimes G(\omega) \quad . \tag{A.26}$$

Parseval's Theorem The total power is the same in both time and frequency domain

$$\int |f(t)|^2\,dt = \frac{1}{2\pi}\int |F(\omega)|^2\,d\omega \tag{A.27}$$

time derivatives

$$\frac{d\,f(t)}{dt} \quad \leftrightarrow \quad i\omega\,F(\omega) \quad . \tag{A.28}$$

Example: Diffraction at a double slit

As an example, we consider the Fourier transform of a double slit, which describes its diffraction pattern. The slits have a width b and a center distance d. Thus the slit is described by a convolution of the rectangular function with two delta functions at the distance d

$$f(x) = \mathrm{rect}_b(x) \otimes (\delta(x-d/2) + \delta(x+d/2)) \quad . \tag{A.29}$$

The Fourier transform of the rectangular function is the sinc, that of the delta functions a constant. However, the shift in position causes a modulation in k-space. Thus, the sum of the two delta functions becomes

$$\mathcal{FT}\left\{\delta(x-d/2) + \delta(x+d/2)\right\} = e^{-ikd/2} + e^{+ikd/2} = 2\cos(kd/2) \quad . \tag{A.30}$$

The convolution with the rectangular function passes into a multiplication with the sinc. Together we get

$$\mathcal{FT}\left\{f(x)\right\} = b\frac{\sin(kb/2)}{kb/2}\,2\cos(kd/2) = \frac{4}{k}\,\sin(kb/2)\,\cos(kd/2) \quad . \tag{A.31}$$

The intensity in direction k is then the squared magnitude of this.

Test yourself

1. *Temporal shift* Sketch the amplitude and phase of the FT of a temporal square pulse pulse centred on time zero! What changes if the pulse is shifted to positive times?

2. *Pulse sequence* You wonder what the Fourier transform (magnitude squared) of an infinite sequence of square pulses looks like and start searching for it on the internet. Your fellow student replies that you can "see" it immediately. Sketch the Fourier transform! Explain why you could derive it directly or why you should "see" it!

3. *Light pulse* Think of a "light pulse" as a mathematical construction of an infinitely long cosine oscillation corresponding to the frequency of light. The "pulse" is obtained by multiplying the wave by a time-limited Gaussian pulse envelope (e.g. half-width of 10 light oscillations). Sketch the construction of the Fourier transform in the spectral domain.

Two-dimensional Fourier transformation

We can extend the definition of the Fourier transform to two and more dimensions. The conjugated variables are (x, y) and (k_x, k_y) instead of t and ω. The wave vector $k_i = 2\pi/\lambda_i$ contains the factor 2π as in the angular frequency ω. We define

$$F(k_x, k_y) = \iint_{-\infty}^{+\infty} f(x,y)\, e^{-i(k_x\, x + k_y\, y)}\, dx\, dy \tag{A.32}$$

$$f(x,y) = \frac{1}{(2\pi)^2} \iint_{-\infty}^{+\infty} F(k_x, k_y)\; e^{+i(k_x\, x + k_y\, y)}\, dk_x\, dk_y \quad . \tag{A.33}$$

When we can separate the function $f(x, y)$ into a product of one-dimensional functions, then the Fourier transform is simply the product of the individual Fourier transforms

$$f(x,y) = g(x)\cdot h(y) \quad \leftrightarrow \quad F(k_x, k_y) = G(k_x)\cdot H(k_y) \quad . \tag{A.34}$$

A rectangle of size $a \times b$ is transformed into a product of sinc functions

$$(x,y) = \mathsf{rect}_a(x)\cdot \mathsf{rect}_b(y) \tag{A.35}$$

$$\leftrightarrow \quad F(k_x, k_y) = ab\,\mathsf{sinc}(k_x a/2)\,\mathsf{sinc}(k_y b/2) \quad . \tag{A.36}$$

A special case of this is the rotational symmetric two-dimensional Gaussian function

$$f(x,y) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{x^2+y^2}{2\sigma^2}} \quad \leftrightarrow \quad F(k_x, k_y) = e^{-\frac{\sigma^2}{2}\left(k_x^2 + k_y^2\right)} \quad . \tag{A.37}$$

One important function can not be separated into a product of one-dimensional functions: a disc of radius a

$$f(x,y) = \begin{cases} 1 & \text{for} \quad x^2 + y^2 < a \\ 0 & \text{other} \end{cases} \tag{A.38}$$

is transformed into

$$F(k_x, k_y) = a\,\frac{J_1(\pi\, a\, \rho)}{\rho} \quad \text{width} \quad \rho = \sqrt{k_x^2 + k_y^2} \tag{A.39}$$

and the (cylindrical) Bessel function of the first kind $J_1(x)$

$$J_1(x) = \frac{1}{\pi}\int_0^\pi \cos(\tau - x\sin\tau)\, d\tau \quad , \tag{A.40}$$

which is the cylindrical analogue of a sinc function.

References

Butz, Tilman (2015). *Fourier Transformation for Pedestrians*. 2. ed. Springer. .

Appendix B
Numerical Fourier Transformation

Markus Lippitz
September 18, 2023

Discrete FT: a periodic sequence of values

In particular, if one collects and evaluates measurement data with a computer, then one does not know the measured function $f(t)$ on a continuous axis t, but only at discrete times $t_k = k\,\delta t$, nor does one know the function from $t = -\infty$ to $t = +\infty$. So we have only a finite sequence of numbers f_k as a starting point. Because we do not know the sequence of numbers outside the measured interval we make the assumption that it is periodic. With N measured values the period is $T = N\Delta t$. For simplicity, we also define $f_k = f_{k+N}$ and thus $f_{-k} = f_{N-k}$ with $k = 0, 1, \ldots, N-1$. Thus the Fourier transform becomes[1]

$$F_j = \frac{1}{N} \sum_{k=0}^{N-1} f_k \, e^{-k\,j\,2\pi i/N} \tag{B.1}$$

[1] see Butz, 2015 chap. 4, Horowitz and Hill, 2015, chap. 1.08, 7.20, 15.18

and its inverse transform

$$f_k = \sum_{j=0}^{N-1} F_j \, e^{+k\,j\,2\pi i/N} \quad . \tag{B.2}$$

The definition is again such that F_0 corresponds to the mean. Because of $f_{-k} = f_{N-k}$, the positive frequencies are in the first half of F_j as the frequency increases. After that come the negative frequencies, starting at the 'most negative' frequency and increasing to the last frequency before zero. So the maximum frequency that can be represented is the Nyquist (angular) frequency

$$\Omega_{\text{Nyquist}} = \frac{\pi}{\delta t} \quad . \tag{B.3}$$

This frequency is such that we take two samples per period of the oscillation. Faster oscillations or fewer samples per period cannot be represented. Even with f_{Nyquist} the imaginary part is always zero, because we always sample the sine at the zero crossing.

FFTW

The most used package for numerical Fourier transform is probably FFTW[2].

[2] https://www.fftw.org/

You have to pay attention to the details of the definition. In particular, the prefactors may differ between different packages. In FFTW, the prefactor $1/N$ changes from the forward to the backward transformation, i.e.

$$F_j = \sum_{k=0}^{N-1} f_k \, e^{-k\,j\,2\pi i/N} \tag{B.4}$$

and the inverse Fourier transform

$$f_k = \frac{1}{N} \sum_{j=0}^{N-1} F_j \, e^{+k\,j\,2\pi i/N} \quad . \tag{B.5}$$

In equations, I (and Butz) use mathematical indices (starting from zero). Some programming languages count from one (e.g., Julia).

One helpful thing of FFTW is that is supplies also a frequency axis. As mentioned above, first come the positive frequencies, starting from zero to the maximum, then the most negative frequency, again rising until just before zero. Depending wether the number of samples N is even or odd, it is a little bit of a hassle to calculate the respective frequencies, but FFTW does this for us:

```
fftfreq(5) # gives [0.0, 0.2, 0.4, -0.4, -0.2]
fftfreq(6) # gives [0.0, 0.166, 0.333, -0.5, -0.333, -0.1.66]
```

Test yourself

1. Try yourself the FFT in a language of your choice. The FFT of, say, $[1111]$ should give something like $[4000]$.
2. The inverse FFT is IFFT. Check that it inverts and test how the pre-factors are distributed.

Wrapping & fftshift

Now lets look at the Fourier transform of a cosine. We evaluate the cosine at 8 points:

$$x_n = n\frac{2\pi}{8} \quad \text{with} \quad n = 0 \ldots 7 \tag{B.6}$$

$$f_n = \cos x_n \tag{B.7}$$

$$F = \mathcal{FT}(f) \quad . \tag{B.8}$$

We find that only F_1 and F_7 are different from zero and have the same, real value. Two values must be different from zero because

$$\cos(x) = \frac{1}{2}\left(e^{ix} + e^{-ix}\right) \quad . \tag{B.9}$$

In general, for real values f_n we have

$$F_{N-j} = F_j^\star \quad . \tag{B.10}$$

The position of these two non-zero values is a consequence of the definition of F_k: first come all positive frequencies and then all negative. For a nicer representation it is often better if the frequency zero is not the first element but in the middle between the positive and negative frequencies. frequencies. This we get by `fftshift` or backwards by `ifftshift`.

Test yourself

3. Convince yourself that you understand why it is element 1 and 7 that differs from zero in the example above.

4. Replace the cosine with a sine in this example and explain the result.

Sampling theorem

We need at least two samples per period to describe a function by its Fourier coefficients. The frequencies must be below the Nyquist frequency f_{Nyquist}

$$f_{\text{Nyquist}} = \frac{1}{2\Delta t} \quad . \tag{B.11}$$

The *sampling theorem* states that this is then also sufficient, i.e., we do not lose any detail by sampling. Let $f(t)$ be a bandwidth-limited function, i.e. $F(\omega)$ is different from zero only in the interval $|\omega| \leq \Omega_{\text{Nyquist}}$. Then the sampling theorem[3] applies and gives

[3] for a proof see Butz, 2015, chap. 4.4

$$f(t) \stackrel{!}{=} \sum_{k=-\infty}^{\infty} f(k\Delta t)\, \text{sinc}\left(\Omega_{\text{Nyquist}} \cdot [t - k\Delta t]\right) \quad . \tag{B.12}$$

So it is enough to sample f all Δt. At the times in between, f is completely described by the (infinitely long) sum of the neighbouring values times the sinc.

In measurement technology, therefore, all we need to do is ensure, for example by means of an electrical filter, that all the frequencies of a signal are below Ω_{Nyquist}, and then our digital acquisition of the signal will be is identical to the signal itself. However, if we sample too infrequently, or if there are higher frequencies present, then these too high frequency components will be reflected at the Nyquist frequency and end up at seemingly lower frequencies. This 'aliasing' distorts the signal.

Zero padding

We began with a repeating pattern of numerical values and their Fourier transform. We always picked the length of the sequence in the examples to match an integer multiple of the period. But of course, this isn't feasible in reality. We lack accurate knowledge of the signal's duration. Or sometimes, multiple signals with varying frequencies are important.

The problem is then a truncation error, which leads to artefacts in the Fourier transform. Fig. B.1 shows an example. 12 data points of a cosine with period 8 are sampled. The FFT assumes periodic continuation (thick) which is not the 'true' signal (thin). In this case, the FFT of the data is far from a peak at the original frequencies. The real part is even spectrally constant (see below Fig. B.2)

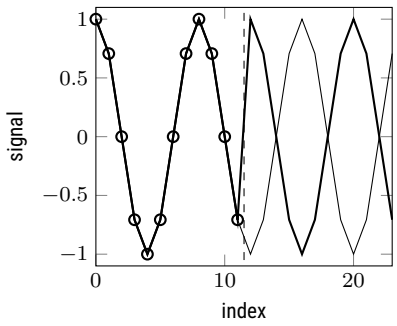

Figure B.1: Clipping a cosine after 1.5 periods

The way out is *zero-padding*. Let our actual measured signal sequence $f(t)$, which we know in the interval $[-T, T]$. Now we pretend that we measured instead

$$g(t) = f(t) \cdot w(t) \tag{B.13}$$

with the window function $w(t)$

$$w(t) = 1 \quad \text{for} \quad -T < t < T \quad \text{other} = 0 \quad . \tag{B.14}$$

Thus we can 'measure' $g(t)$ over arbitrarily long times, because it is is quasi always zero. But the Fourier transform is

$$G(\omega) = F(\omega) \otimes W(\omega) \tag{B.15}$$

with

$$W(\omega) = 2T \, \frac{\sin \omega T}{\omega T} = 2T \, \text{sinc}(\omega T). \quad . \tag{B.16}$$

So we extend our data set on both sides with zeros. The effect is that we convolve the actual Fourier transform of our data set with a sinc whose characteristic width is determined by the actual measurement duration. The frequency resolution does not increase. Rather, a kind of interpolation in Fourier space occurs, which just eliminates the artefacts of the truncation error.

We consider the same data set as above, only we 'extend' it to 10 times the length. This means that the clipping error has less influence and the peak is always at 1 Hz in frequency space. But this does not give more resolution, of course. Peaks that are close to each other cannot be separated by zero-padding, only the position of a peak can be be determined better .

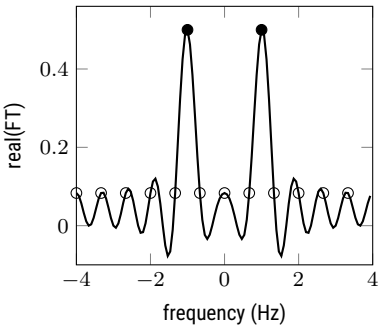

Figure B.2: Zeropadding (line) approaches better the real spectrum (filled symbols) compared to the clipped FT (open symbols).

Windowing

The oscillations in the spectrum in the last example are still artefacts. Actually, one would expect two delta functions at ± 1Hz. They are a consequence of the rectangular window $w(t)$, which leads to the sinc in frequency space. The square-wave window is natural in the sense that we always start and stop measuring. Other window functions[4], however, may be better.They differ the width of the peak and the steepness of the slopes. Unfortunately one must trade one against the other. Interesting parameters are the width of the central peak in frequency space, measured as a -3dB bandwidth, as well as the sideband suppression in [5] dB or its drop in dB/octave.

[4] https://en.wikipedia.org/wiki/Window_function

[5] dB = decibel = $10 \log_1 0x$

Typical window functions are (with $|x| = |t/T| < 1/2$)

$$\text{cosine} = \cos \pi x \tag{B.17}$$

$$\text{triangle} = 1 - 2|x| \tag{B.18}$$

$$\text{Hanning} = \cos^2 \pi x \tag{B.19}$$

$$\text{Hamming} = a + (1-a) \cos^2 \pi x \tag{B.20}$$

$$\text{Gauss} = \exp\left(-\frac{1}{2}\frac{x^2}{\sigma^2}\right) \tag{B.21}$$

$$\text{Kaiser-Bessel} = \frac{I_0(\pi\alpha\sqrt{1-4x^2})}{I_0(\pi\alpha)} \tag{B.22}$$

with the modified Bessel function I_0.

With a window, the measured values are reduced, but the Fourier transform is smoother, because the transition to the zero padding becomes smoother. This makes it possible to recognize in the example the peaks at ± 1Hz even with very few sampled points.

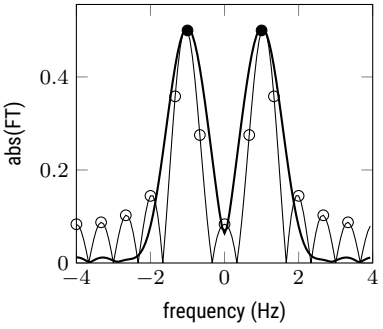

Figure B.3: Zeropadding after windowing (thick) removes the fringes of the un-windowed data (thin) and approaches the true spectrum (solid symbols).

We consider as example[6] a sum of 6 cosine functions with partly very different amplitudes A_l and frequencies f_l:

$$f(t) = \cos\omega t + 10^{-2}\cos 1.15\omega t + 10^{-3}\cos 1.25\omega t + 10^{-3}\cos 2\omega t + 10^{-4}\cos 2.75\omega t + 10^{-5}\cos 3\omega t \tag{B.23}$$

We sample 256 data points at intervals of $\Delta t = 1/8$, i.e. only $8/3 \approx 3$ data points per oscillation of the highest occurring frequency, which is 5 orders of magnitude weaker than the lowest frequency. Nevertheless, this peak can be found with a suitable window and zero-padding.

[6] from Butz, 2015, chapter 3.10

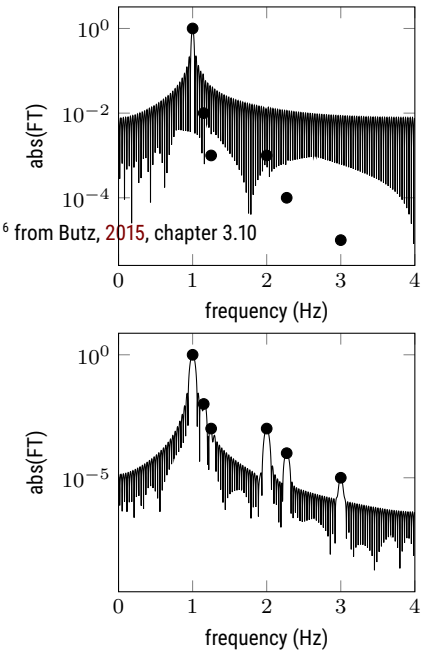

Figure B.4: Without windowing (top), only the main signal component is recovered. A Hanning window (bottom) allows to find even signals 10^{-5} below the main component.

References

Butz, Tilman (2015). *Fourier Transformation for Pedestrians*. 2. ed. Springer. .

Horowitz, Paul and Winfield Hill (2015). *The art of electronics*. Third edition. New York, NY: Cambridge University Press.

Bibliography

Born, Max and Emil Wolf (2002). *Principles of optics*. 7. (expanded) ed., reprinted with corr. Cambridge [u.a.]: Cambridge Univ. Press.

Boyd, Robert W. (1980). "Intuitive explanation of the phase anomaly of focused light beams". In: *Journal of the Optical Society of America* 70, pp. 877–880. .

Brooker, Geoffrey (2008). *Modern classical optics*. 1. publ., repr. with corr. Oxford master series in physics. Oxford [u.a.]: Oxford Univ. Press.

Butz, Tilman (2015). *Fourier Transformation for Pedestrians*. 2. ed. Springer. .

Goodman, Joseph W. (2005). *Introduction to Fourier optics*. 3. ed. Roberts.

Hecht, Eugene (2017). *Optics*. Fifth edition, global edition. Boston: Pearson.

Hering, Ekbert and Rolf Martin (2017). *Optik für Ingenieure und Naturwissenschaftler*. München: Fachbuchverlag Leipzig im Carl Hanser Verlag.

Horowitz, Paul and Winfield Hill (2015). *The art of electronics*. Third edition. New York, NY: Cambridge University Press.

Konijnenberg, Sander, Aurèle J.L. Adam, and Paul Urbach (2021). *BSc Optics*. TU Delft Open. .

Nolting, Wolfgang (2016). *Theoretical Physics 3 Electrodynamics*. Springer. .

Novotny, Lukas and Bert Hecht (2012). *Principles of nano-optics*. 2. ed. Cambridge Univ. Press. .

Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley. .

Yariv, Amnon (1989). *Quantum electronics*. 3. ed. New York: Wiley.