Optics

Markus Lippitz
January 26, 2024

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

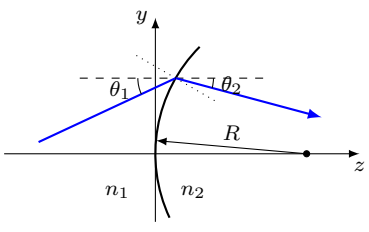

Figure 1.1: Refraction of a ray at a spherical surface

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., $n_1 = 1$.

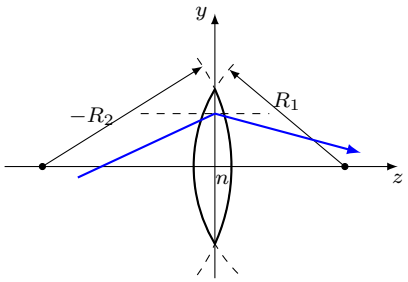

Figure 1.2: Refraction of a ray at a thin lens

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

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

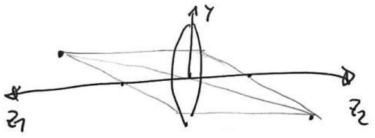

Figure 1.3: Image formation at a thin lens

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.

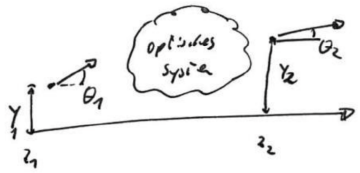

Figure 1.4: The ray-transfer matrix describes the optical system between two planes.

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.

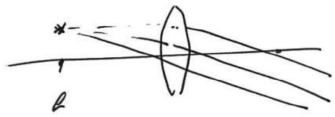

Figure 1.5: Point source in the focal plane of a lens

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

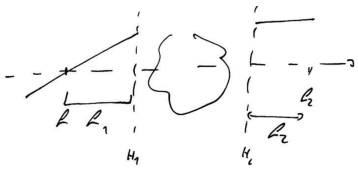

Figure 1.6: The two principal planes simplify many optical systems.

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.

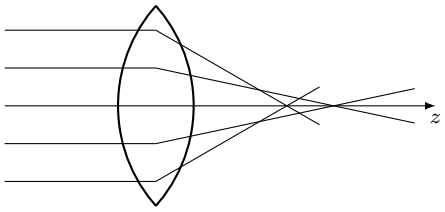

Figure 1.7: Spherical aberration: focal position depends on beam height.

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.

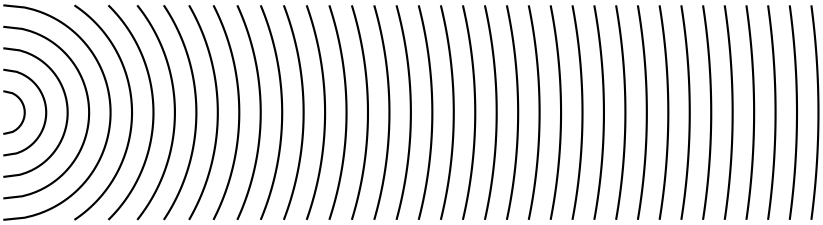

Figure 2.1: Wave fronts of a spherical wave, showing the transition to plane waves far from the origin.

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

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

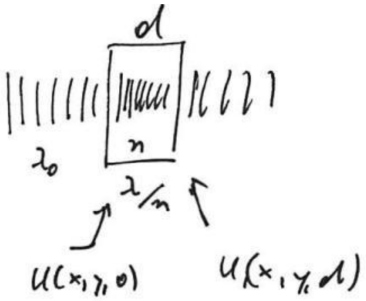

Figure 2.2: A plate

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.

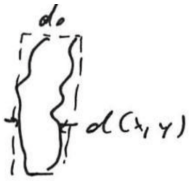

Figure 2.3: A plate of variable thickness

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

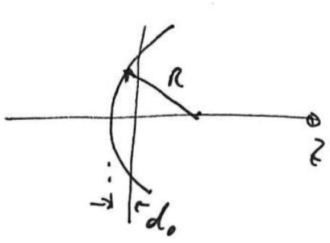

Figure 2.4: A lens as plate of variable thickness

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.

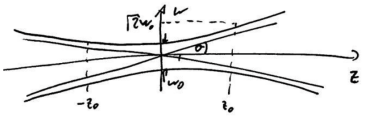

Figure 2.5: Divergence of a Gaussian beam

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 $\Theta_{\text{app}} = 0.61\lambda/R$.

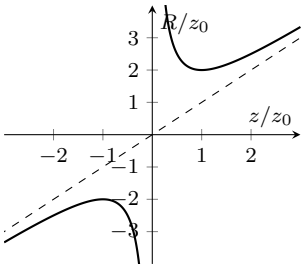

Figure 2.6: Radius of curvature R around the beam waist. Dashed: spherical wave

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.

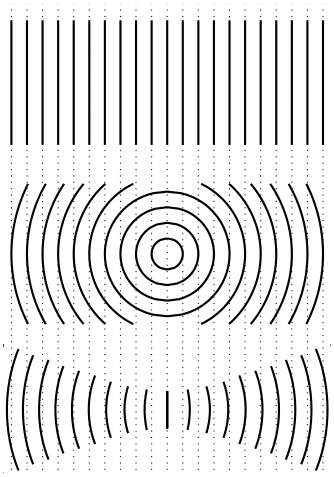

Figure 2.7: Wavefronts of a plane wave, a spherical wave and a Gaussian wave.

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

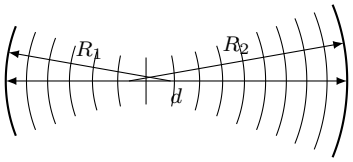

Figure 2.8: Eigenmodes of a laser cavity

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

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

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

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

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

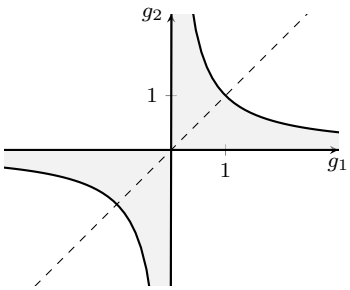

Figure 2.9: A laser cavity is stable inside the shaded region ($g_i = 1 + d/R_i$). Symmetric cavities are along the diagonal, flat mirrors at $g = 1$.

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.

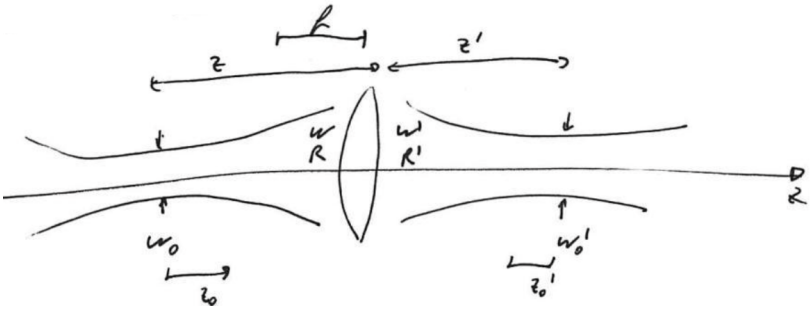

Figure 2.10: A lens acting on a Gaussian beam

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.

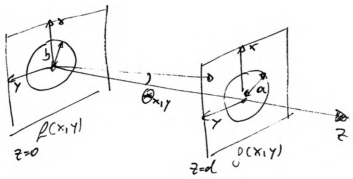

Figure 3.1: Fraunhofer condition

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

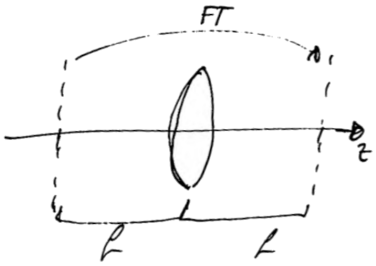

Figure 3.2: Optical Fourier transform by a lens

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]

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

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

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.

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.

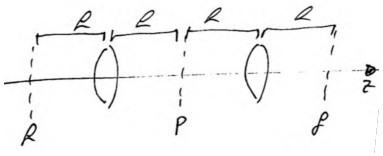

Figure 3.3: A $4f$ system can be used as spatial filter.

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

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

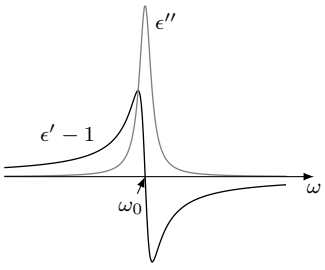

Figure 4.1: Frequency dependence of the real and imaginary parts of the Lorentz oscillator. The real and imaginary parts of the complex-valued refractive index $\tilde{n}$ look qualitatively the same.

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 4.1, 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'.

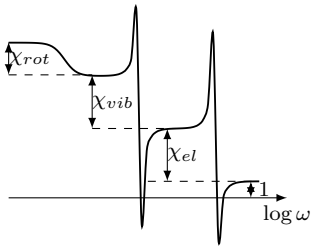

Figure 4.2: The visible spectral range lies between two resonances.

The Lorentz-shaped resonance can be shown in a demonstration experiment. The imaginary part of the dielectric function Fig. 4.1 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.

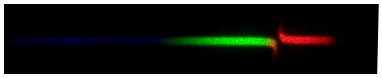

Figure 4.3: Anomalous dispersion in sodium vapor.

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

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

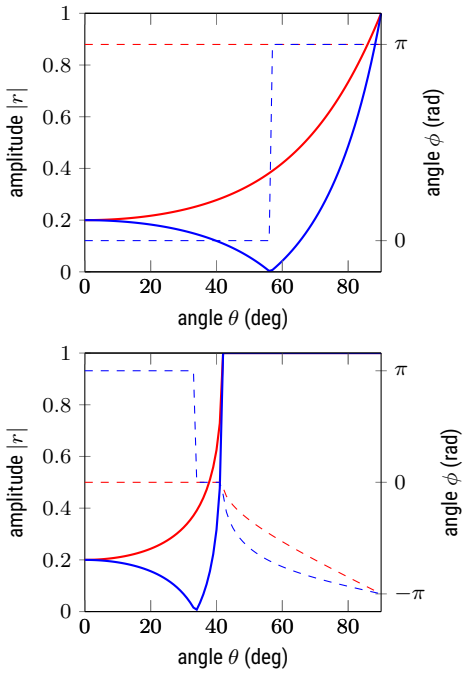

Figure 4.4: Fresnel coefficients $r = |r|e^{i\phi}$ for external (top) and internal (bottom) reflection at an air–glass interface. red: s-polarized, blue: p-polarized. dashed: phase

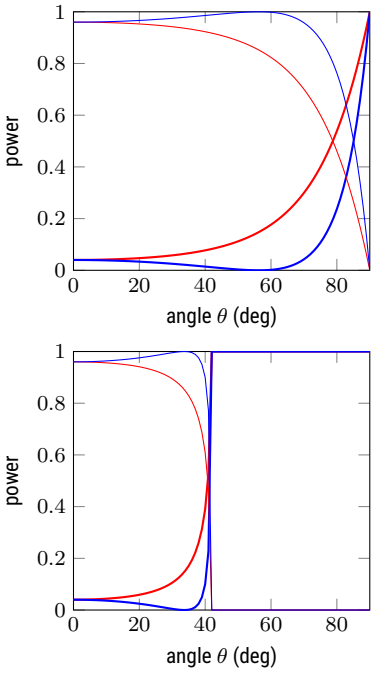

Figure 4.5: Reflected (thick) and transmitted (thin) power for external (top) and internal (bottom) reflection at an air–glass interface. red: s-polarized, blue: p-polarized

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.

Chapter 5
Polarization and anisotropic media

Markus Lippitz
December 8, 2023

By the end of this chapter you should be able to explain and experimentally demonstrate the operation of a wave plate.

Overview

Polarization is the most important property of light. It makes the difference to scalar waves. It gives an additional degree of freedom in optical devices. We first introduce the Jones formalism to describe the state of polarization and then discuss the propagation of light in anisotropic media. The most surprising effect is probably birefringence. The technical applications are components of polarization optics such as waveplates and polarizing prisms. Spectroscopic applications include the magneto-optical Kerr effect and circular dichroism.

Introduction

We start by looking at a monochromatic plane wave again. This time, however, we pay more attention to the vectorial amplitude by writing

$$\boldsymbol{E}(z,t) = \Re\left\{(A_x\hat{\boldsymbol{x}} + A_y\hat{\boldsymbol{y}})\ e^{-i\omega(t-z/c)}\right\} \quad , \tag{5.1}$$

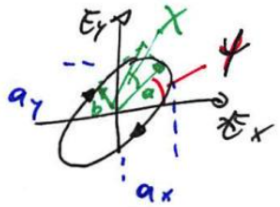

Figure 5.1: Electric field in the xy-plane.

where $A_{x,y}$ are complex-valued amplitudes. We separate them intro amplitude a and phase ϕ (and write down only the x components for simplicity)

$$A_x = a_x e^{-i\phi_x} \tag{5.2}$$

so that the x components of the field is

$$E_x = a_x \cos\left(\omega(t-z/c) + \phi_x\right) \quad . \tag{5.3}$$

In the xy-plane this describes as function of time t an ellipse with ellipticity χ and orientation ψ:

$$\tan 2\psi = \frac{2R}{1-R^2}\cos(\phi_y - \phi_x) \tag{5.4}$$

$$\sin 2\psi = \frac{2R}{1-R^2}\sin(\phi_y - \phi_x) \tag{5.5}$$

with $R = a_y/a_x$.

Test yourself

1. Draw the field $\boldsymbol{E}$ for a fixed t in the xyz-cube and for a fixed z in the xyt-cube.

Polarization states

Linear polarization is described by

$$\phi_y - \phi_y = 0 \text{ or } \pi \quad . \tag{5.6}$$

In these cases, the ellipticity vanishes ($\chi = 0$) and the orientation ψ is

$$\tan 2\psi = \frac{2R}{1-R^2} \quad \text{or} \quad \tan\psi = \frac{a_y}{a_x} \quad . \tag{5.7}$$

Circular polarization is described by

$$\phi_y - \phi_y = \pm\frac{\pi}{2} \quad \text{and} \quad a_x = a_y \quad . \tag{5.8}$$

The electric field vector describes a circle in the xy-plane. The plus sign corresponds to right-circularly polarized (RCP), the minus sign to left-circularly polarized (LCP). A snapshot in time gives in the xyz cube a helix of the corresponding[1] handedness.

[1] This sign differs in some books!

All other polarization states can be depicted as point on the *Poincaré sphere*. We need only two parameters to describe the state of polarization, χ and ψ, or R and $\phi_y - \phi_x$. For the Poincaré sphere, we set

$$r = 1 \qquad \theta = \frac{\pi}{2} - 2\chi \qquad \psi = 2\chi \quad . \tag{5.9}$$

North and south pole correspond to RCP and LCP, respectively. The equator describes linearly polarized light ($\chi = 0$) in the different directions.

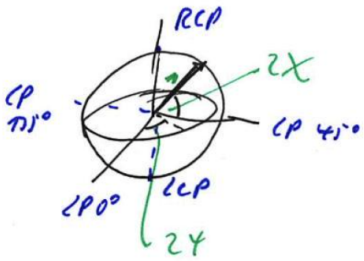

Figure 5.2: Polarization states on the Poincaré sphere.

A more general method to describe polarized light is the four-component *Stokes vector*. The first component is the intensity of the light field that was not contained in the Poincaré sphere. The other three are the three cartesian components of the point on the Poincaré sphere, i.e.

$$S_0 = a_x^2 + a_y^2 \tag{5.10}$$
$$S_1 = a_x^2 - a_y^2 \tag{5.11}$$
$$S_2 = 2a_x a_y \cos(\phi_y - \phi_x) \tag{5.12}$$
$$S_3 = 2a_x a_y \sin(\phi_y - \phi_x) \quad . \tag{5.13}$$

As we have introduced the $a_{x,y}$ above, we always get $S_1^2 + S_2^2 + S_3^2 = 1$, which would not justify an additional component. However, the scheme can be extended to describe partially polarized or unpolarized light, for which the last component is needed.

Jones formalism

The simplest way to describe the effect of optical elements such as quarter and half wave plates or polarizers on the polarization state of light is the Jones formalism. We write the complex amplitudes A_x and A_y as the two components of a Jones vector

$$\boldsymbol{J} = \begin{pmatrix} A_x \\ A_y \end{pmatrix} \quad . \tag{5.14}$$

In most cases we normalize $|\boldsymbol{J}| = 1$. Examples are

$$\begin{pmatrix}1\\0\end{pmatrix} \qquad \begin{pmatrix}\cos\theta\\\sin\theta\end{pmatrix} \qquad \frac{1}{\sqrt{2}}\begin{pmatrix}1\\i\end{pmatrix} \qquad \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-i\end{pmatrix} \tag{5.15}$$

for linear polarization in x direction, under an angle θ to the x-direction, and right and left-handed circular polarization, respectively.

A 2×2 matrix describes the effect of an optical element. A *linear polarizer* is

$$T = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} \quad . \tag{5.16}$$

A *wave plate* delays the slow y component by a phase ϕ compared to the fast x component, i.e.

$$T = \begin{pmatrix}1 & 0\\0 & e^{-i\phi}\end{pmatrix} \quad . \tag{5.17}$$

As light travels faster when polarized along the x axis, it is called the 'fast axis' of the wave plate. Important cases are the quarter-wave plate with $\phi = \pi/2 = 2\pi/4$ and the half-wave plate with $\phi = \pi = 2\pi/2$. A *quarter wave plate* converts linear polarized light under $45°$ into LCP, and RCP in linear polarized light under $45°$:

$$\begin{pmatrix}1\\-i\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & e^{-i\pi/2}\end{pmatrix} \cdot \begin{pmatrix}1\\1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & e^{-i\pi/2}\end{pmatrix} \cdot \begin{pmatrix}1\\i\end{pmatrix} \quad . \tag{5.18}$$

A *half wave plate converts* turns the polarization direction of linear polarized light from $+45°$ to $-45°$, i.e., by $90°$. It also converts RCP into LCP:

$$\begin{pmatrix}1\\-1\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & e^{-i\pi}\end{pmatrix} \cdot \begin{pmatrix}1\\1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1\\-i\end{pmatrix} = \begin{pmatrix}1 & 0\\0 & e^{-i\pi}\end{pmatrix} \cdot \begin{pmatrix}1\\i\end{pmatrix} \quad . \tag{5.19}$$

It is important to note that these conversions only work for the indicated relative orientation of the waves plates x and y axis and the incoming polarization state. In almost all cases, one needs to rotate the optical elements relative to the lab-frame xy coordinate system. The effect of a rotated element $T'(\theta)$ is obtained by rotation matrices

$$T'(\theta) = R(\theta) \cdot T \cdot R(-\theta) \quad \text{with} \quad R(\theta) = \begin{pmatrix}\cos\theta & \sin\theta\\-\sin\theta & \cos\theta\end{pmatrix} \quad . \tag{5.20}$$

Test yourself

2. Find orientations of quarter- and half-wave plate that do not change a linearly polarized beam!

3. Plot the total power in a beam as function of angle θ of a linear polarizer relative to a linear polarized beam and to a left-circular polarized beam.

Anisotropic media

The refractive index of anisotropic media depends on the polarization direction of the light wave. This statement defines anisotropic media. This

relationship makes anisotropic media of special interest for polarization optics, and polarization optics of importance for the study of anisotropic media. Before we get to the propagation of light in anisotropic media, let us discuss the microscopic patterns of anisotropy. On the nanoscale, molecules could be ordered in position and / or orientation.

gases, liquids, amorphous solids Both position and orientation are random. The media are isotropic

polycrystallline solids On a short length scale we find order in position and orientation, leading to anisotropy. On a longer length scale, the crystallites are disordered, averaging out the anisotropy.

crystalls Both position and direction are ordered. In general case, crystals are anisotropic.

liquid crystals The position is random, but the orientation is ordered. This is enough to find anisotropy.

Figure 5.3: Order in position and / or orientation.

Index ellipsoid

When we deal with anisotropic media, we have to take the tensorial nature of the electric permittivity $\boldsymbol{\epsilon}$ into account:

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{\epsilon} \boldsymbol{E} \quad \text{or} \quad D_i = \epsilon_0 \sum_j \epsilon_{ij} E_j \quad \text{with} \quad i,j = x,y,z \tag{5.21}$$

where ϵ_{ij} are the components of the 3×3 tensor $\boldsymbol{\epsilon}$. For most materials, the tensor is symmetric, i.e., $\epsilon_{ij} = \epsilon_{ji}$. This holds for nonmagnetic dielectrics that do not show optical activity, and in absence of magnetic fields. The symmetry reduces the originally 9 elements to 6 independent ones.[2]

We can depict such a symmetric tensor as ellipsoidal surface defined by

$$\sum_{i,j} \epsilon_{ij} x_i x_j = 1 \quad . \tag{5.22}$$

When we change the coordinate system, then both the x_i and the ϵ_{ij} change and the ellipsoid remains unchanged. There is one coordinate system in which the ϵ_{ij} matrix is diagonal. This are the principal axes. In these directions, $\boldsymbol{E}$ amd $\boldsymbol{D}$ are parallel. In the following, my spatial coordinate system is always this principal system, labeled as x_1, x_2, x_3. The semi-axes of the ellipsoid, i.e., where the ellipsoid crosses the axes, have the values $1/\sqrt{\epsilon_i} = 1/n_i$. Here n_i is the index of refraction along one of the principal axes.

In the following, it is more useful to have the inverse tensor $\boldsymbol{\eta} = \boldsymbol{\epsilon}^{-1}$, i.e.

$$\boldsymbol{E} = \frac{1}{\epsilon_0} \boldsymbol{\epsilon}^{-1} \boldsymbol{D} = \frac{\boldsymbol{\eta}}{\epsilon_0} \boldsymbol{D} \quad . \tag{5.23}$$

When depicted as ellipsoid, called *index ellipsoid*, its principal axes agree with the permittivity tensor. The intersections with the axes are at n_i and we can write

$$\sum_{i=1,2,3} \frac{x_i^2}{n_i^2} = 1 \quad . \tag{5.24}$$

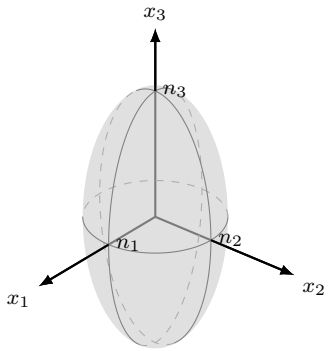

Figure 5.4: The index ellipsoid for an uniaxial crystal.

[2] see chapter 15.4 of Brooker, 2008 for an interpretation of the coefficients

We distinguish crystals by their set of n_i:

$$n_1 = n_2 = n_3 \qquad \text{isotropic} \tag{5.25}$$

$$n_o = n_1 = n_2 \neq n_3 = n_{eo} \qquad \text{uniaxial} \tag{5.26}$$

$$n_1 \neq n_2 \neq n_3 \qquad \text{biaxial.} \tag{5.27}$$

In this chapter, we will only discuss uniaxial crystals. Their two index values are called 'ordinary' and 'extraordinary'. The principal axis of the third, extraordinary value is called 'optics axis', not to be confused with the optical[3] axis ($= z$).

[3] In German, both is 'optische Achse'!

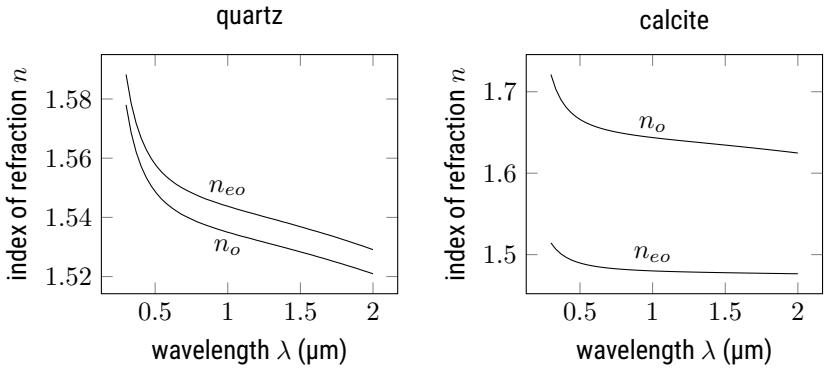

Figure 5.5: Ordinary (n_o) and extraordinary (n_{eo}) index of refraction for quartz (SiO_2) and calcite ($CaCO_3$). As for quartz, $n_{eo} > n_0$, quartz is called positive-uniaxial, while calcite is negative-uniaxial.

Propagation along a principal axis

We are interested in how a wave propagates in anisotropic media. In which direction does the energy flow, how are the wave fronts oriented, and what is the polarization state. We are especially interested in "normal modes", i.e. situations that do not change. In this case, this means polarization states that do not change, that travel unperturbed through the crystal.

Let us start discussing the propagation of light in anisotropic media by assuming that we travel along a principal axis. This means that the wavevector $\boldsymbol{k}$ is parallel to one of the x_i. When the electric field $\boldsymbol{E}$ is parallel to another principal x_j, then also $\boldsymbol{D}$ remains parallel to x_j. The wave travels as in an homogenous medium with index of refraction n_j and the state of polarization remains unchanged.

When the direction of the electric field does not coincide with one of the x_i, then we can write it as superposition of fields along several x_i. Each of these fields travels as above, but experiences a different index of refraction. After a distance d they have acquired a phase difference of

$$\Delta\phi = \Delta n \, k_0 \, d \quad . \tag{5.28}$$

This is what a wave plate is doing. A uniaxial crystal is cut such that the in-plane directions have a different index of refraction.

Propagation along an arbitrary direction

Now the plane wave is allowed to travel in any direction $\boldsymbol{k}$. It is not constrained to a principal axis as above. We draw[4] the wave vector $\boldsymbol{k}$ in the index

[4] see Saleh and Teich, 1991 for a proof that this works

ellipsoid and construct a plane perpendicular to it that contains the origin of the coordinate system. The ellipsoid crosses this plane, forming an *index ellipse*. The principal axes of this index ellipse are perpendicular to $\boldsymbol{k}$ and describe the two linear polarization directions that propagate unperturbed, i.e. are normal modes. The lengths of the semi-axes of the ellipse give the index of refraction experienced by these two polarization states.

In the special case of an uniaxial crystal, one wave, the ordinary wave, will always experience the ordinary index of refraction n_o. The other, extraordinary one will experience an index that is a mixture of ordinary and extraordinary. The mixing ratio depends on the angle θ that the wave vector makes with the optics axis:

$$\frac{1}{n(\theta)} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_{eo}^2} \quad . \tag{5.29}$$

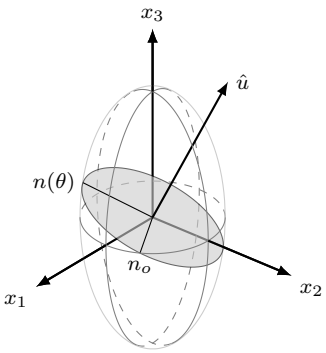

Figure 5.6: When traveling in direction $\hat{\boldsymbol{u}}$, the refractive index of the eigen-modes are found as semi-axes of the ellipse perpendicular to $\hat{\boldsymbol{u}}$.

Energy flow and wave fronts

The extraordinary wave is really extraordinary. We will see this when looking at the direction of the Poynting vector. To describe an optical plane wave, we have the wave vector $\boldsymbol{k}$, the fields $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{B}$, and $\boldsymbol{H}$ and the Poynting vector $\boldsymbol{S}$. When the spatial variation of all fields is proportional to $\exp(i\boldsymbol{k}\cdot\boldsymbol{r})$, then Maxwells equations lead to

$$\boldsymbol{k}\times\boldsymbol{H} = -\omega\boldsymbol{D} \tag{5.30}$$

$$\boldsymbol{k}\times\boldsymbol{E} = \omega\mu_0\boldsymbol{H} \quad . \tag{5.31}$$

This means that $\boldsymbol{k}$, $\boldsymbol{H}$, and $\boldsymbol{D}$ are perpendicular to each other, and $\boldsymbol{k}$, $\boldsymbol{E}$, and $\boldsymbol{H}$. The Poynting vector $\boldsymbol{S} = \frac{1}{2}\boldsymbol{E}\times\boldsymbol{H}^\star$ is perpendicular on $\boldsymbol{E}$ and $\boldsymbol{H}$.

For the extraordinary wave, the tensorial nature of $\boldsymbol{\epsilon}$ does comes into play: $\boldsymbol{E}$ and $\boldsymbol{D}$ are not necessarily parallel. Luckily, this is also not requited by any relation in the last paragraph. Fig. 5.7 sketches how the required perpendicularities are resolved with non-parallel $\boldsymbol{E}$ and $\boldsymbol{D}$. This means that the wave vector $\boldsymbol{k}$ is not parallel anymore to the Poynting vector $\boldsymbol{S}$. As the wave fronts are perpendicular to $\boldsymbol{k}$, in the case of extraordinary waves, the energy transport ist not perpendicular to the wavefronts anymore! This justifies the label 'extraordinary'. In the following, I will call this object of wavefronts not perpendicular to energy flow a 'beam', in contrast to a plane wave, and similar to a Gaussian beam, in which also the wave fronts are curved and thus not everywhere perpendicular to the energy flow.

Lets look at the direction of energy flow and wave fronts a bit more in detail. We combine the equations 5.30 and 5.31 with 5.21 and get

$$\boldsymbol{k}\times(\boldsymbol{k}\times\boldsymbol{E}) + \omega^2\mu_0\epsilon_0\boldsymbol{\epsilon}\boldsymbol{E} = 0 \quad . \tag{5.32}$$

At a given frequency ω, this is a linear equation system for the thee components of $\boldsymbol{E}$. It has a solution if the determinant vanishes, which gives a condition for the components of $\boldsymbol{k}$. In the case of uniaxial crystals, this reads

$$(k^2 - n_o^2\,k_0^2)\left(\frac{k_1^2+k_2^2}{n_{eo}^2} + \frac{k_3^2}{n_o^2} - k_0^2\right) = 0 \quad . \tag{5.33}$$

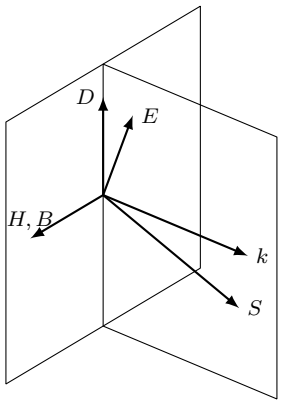

Figure 5.7: For an extraordinary wave, Poynting vector $\boldsymbol{S}$ and wave vector $\boldsymbol{k}$ point in different directions.

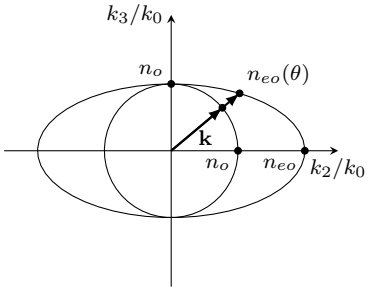

Figure 5.8: k surface cutting the yz-plane for an uniaxial crystal with $n_{eo} > n_o$.

This equation describes a surface in the 3d space of $\boldsymbol{k}$, the k-surface. In fact, we find two solutions: the first factor becomes zero, leading to a spherical surface related to the ordinary wave

$$|\boldsymbol{k}| = k = \sqrt{k_1^2 + k_2^2 + k_3^2} = n_o\, k_0 \quad . \tag{5.34}$$

The second factor describes an ellipsoid and the extraordinary wave

$$\frac{k_1^2 + k_2^2}{n_{eo}^2} + \frac{k_3^2}{n_o^2} = k_0^2 \quad . \tag{5.35}$$

How does one use these k-surfaces? We start by defining the direction $\hat{\boldsymbol{u}}$ of the wavevector $\boldsymbol{k}$. The distance of the surface from the origin in the direction of $\hat{\boldsymbol{u}}$ gives the index of refraction of that wave. This recover its dependence on angle θ between $\boldsymbol{k}$ and the optics axis x_3, as discussed in the last section. Additionally, one can show[5] that the Poynting vector $\boldsymbol{S}$ is perpendicular to the surface at these points. And the wave fronts are perpendicular to $\boldsymbol{k}$ and thus $\hat{\boldsymbol{u}}$ as usual.

[5] Saleh and Teich, 1991.

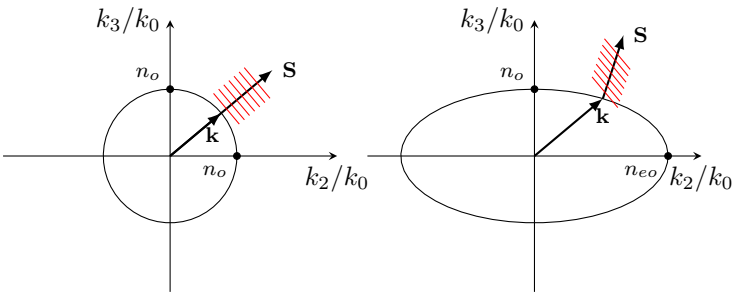

Figure 5.9: Ordinary (left) and extraordinary (right) wave traveling in direction $\boldsymbol{k}$. The electric field $\boldsymbol{E}$ point out of the plane for the ordinary wave, and is in the plane for the extraordinary wave. For the extraordinary wave, the wave fronts are perpendicular to $\boldsymbol{k}$, but not to $\boldsymbol{S}$.

Birefringence

The existence of both an ordinary and an extraordinary wave leads to birefringence. A plane wave is diffracted into two different directions when entering an anisotropic material from the air side. Depending on the polarization direction, the wave sees the ordinary index of refraction n_o, or the extraordinary n_{eo}, where the latter depends on the angle relative to the optics axis of the crystal. For both polarization directions, Snell's law has to hold, i.e.

$$\sin\theta_{air} = n_o \sin\theta_o \quad \text{and} \quad \sin\theta_{air} = n_{eo} \sin\theta_{eo} \quad . \tag{5.36}$$

When θ_a is the angle of the optics axis with the surface normal, then we can calculate $n_{eo} = n(\theta_a + \theta_{eo})$, using eq. 5.29 above.

Snell's law describes the direction of the wave vector $\boldsymbol{k}$ and the wave fronts, but it does not describe the energy flow or the direction of the Poynting vector $\boldsymbol{S}$, at least not in the case of anisotropic media. We see this when letting the plane wave impinge perpendicular on the crystal surface, i.e., set $\theta_{air} = 0$. So also inside the crystal, the wave fronts remain parallel to the crystal surface, all angles remain equal to zero. But the Poynting vector is perpendicular to the k-surface. When the optics axis of the crystal is neither parallel nor perpendicular to the crystal surface, the Poynting vector for the extraordinary beam will deviate. A beam with the extraordinary polarization direction will be displaced in an anisotropic plate. This is used in different types of polarizing beam-splitter prisms.

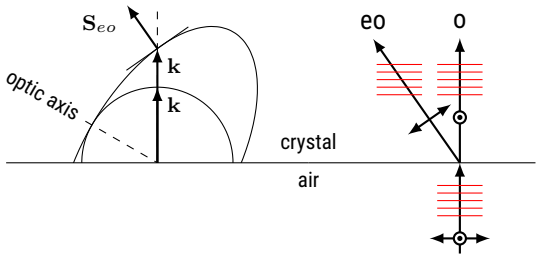

Figure 5.10: When the optics axis is tilted relative to the crystal surface, the extraordinary beam will be refracted, even when the incoming beam is perpendicular to the sample surface!

Optical activity and magneto-optics

I would like to briefly discuss another aspect of anisotropic media. Until now we have assumes that the dielectric tensor $\boldsymbol{\epsilon}$ is symmetric, i.e. $\epsilon_{ij} = \epsilon_{ji}$. This has led to linear polarization states as eigen-modes, i.e. linearly polarized light travels unperturbed for some polarization directions and directions of travels. Now we do something else. We assume that the dielectric tensor $\boldsymbol{\epsilon}$ is hermitian, i.e. $\epsilon_{ij} = \epsilon^{\star}_{ji}$. This leads to circular polarization as eigen-modes. RCP and LCP light travels unperturbed, but with a different index of refraction.

We can write linear polarized light (angie θ) as superposition of RCP and LCP:

$$\begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} = \frac{1}{2} e^{-i\theta} \begin{pmatrix} 1 \\ +i \end{pmatrix} + \frac{1}{2} e^{+i\theta} \begin{pmatrix} 1 \\ -i \end{pmatrix} \quad . \tag{5.37}$$

After traveling a distance d, RCP and LCP waves have accumulated a different phase factor, ϕ_+ and ϕ_-

$$\phi_{\pm} = 2\pi n_{\pm} \frac{d}{\lambda} \quad . \tag{5.38}$$

The difference $\phi = \phi_- - \phi_+$ leads to a rotation of the linear polarization state. The resulting Jones vector is

$$e^{-i(\phi_+ + \phi_-)/2} \begin{pmatrix} \cos\theta - \phi/2 \\ \sin\theta - \phi/2 \end{pmatrix} \quad . \tag{5.39}$$

The 'rotary power' ρ is, for small $\zeta \ll n = \sqrt{\epsilon_{ii}}$

$$\rho = \frac{\phi/2}{d} \approx -\frac{\pi\zeta}{n\lambda_0} \quad . \tag{5.40}$$

Such a tensor is connected with two effects: Optical activity and magneto-optics. Optical active media can be described by

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{\epsilon} \boldsymbol{E} + i\epsilon_0 \boldsymbol{G} \times \boldsymbol{E} \quad \text{with} \quad \boldsymbol{G} = \zeta \boldsymbol{k} \tag{5.41}$$

with the gyration vector $\boldsymbol{G}$. When the wave travels in z-direction, then only the off-diagonal elements $\epsilon_{12} = i\zeta = \epsilon^{\star}_{21}$ are different from zero and the diagonal tensor elements are all the same. This can be found in materials with a helical structure, for example quartz, or liquids of chiral molecules, for example amino-acids or sugars.

The Faraday effect is a similar effect in magneto-optics. The dielectric tensor has the same structure, except that the gyration vector is

$$\boldsymbol{G} = \gamma \boldsymbol{B} \tag{5.42}$$

with the gyromagnetic ratio γ. The fundamental difference is that an optically active medium rotates the direction of polarization as a function of the direction of propagation, while the Faraday effect is a function of the direction of the magnetic field. Consequently, when a wave is reflected and travels back through the same medium, the effect of optical activity is canceled, while the Faraday effect is doubled. Normally, in optics, the direction of travel can be reversed and everything happens in the opposite direction, as we have done when tracing light rays through lenses. When the Faraday effect is involved, this does not work. Running back does not give the starting condition anymore. In this way, one can build an optical diode, the *Faraday isolator*, which allows light to pass only in one direction and prevents, for example, light from being reflected back into a laser cavity.

References

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

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

Part IV

Interference and Coherence

Chapter 6
Interference

Markus Lippitz
December 20, 2023

By the end of this chapter you should be able to explain and experimentally demonstrate the operation of an interferometer.

Overview

Interference describes phenomena resulting from the superposition of two or more waves. We have already seen such phenomena in previous chapters: diffraction of light at apertures and Fourier optics in general, or rotation of polarization states in optically active media. In this chapter we will discuss interference phenomena in more detail. A prerequisite for interference is coherence, which is the topic of the next chapter. Here we assume that all waves are monochromatic and coherent, and only in the next chapter we discuss what this means.

This chapter consists of two parts: first we discuss the most common types of interferometers and their applications, then we move on to interference in planar stacks of dielectric materials, as used in dielectric filters.

Interference in general

In chapter 4 we have defined the intensity I of a wave as the temporal average of the Poynting vector projected on a reference surface:

$$I = \langle \boldsymbol{S} \cdot \boldsymbol{n} \rangle_T = \frac{c\, n\, \epsilon_0}{2} |E_0|^2 \tag{6.1}$$

where E_0 is the amplitude of the electrical field in one of the two orthogonal polarization eigen-states[1] $\hat{\boldsymbol{p}}$, i.e.

[1] horizontal and vertical linear or RCP and LCP

$$\boldsymbol{E}(\boldsymbol{r},t) = E_0\, \hat{\boldsymbol{p}}\, e^{i(\boldsymbol{kr}-\omega t)} \quad . \tag{6.2}$$

Interference occurs when E_0 results from the superposition of several waves, i.e. $E_0 = \sum E_i$. These waves are summed *coherently*

$$I \propto \left| \sum_i E_i \right|^2 \tag{6.3}$$

in contrast to an incoherent sum

$$I \propto \sum_i |E_i|^2 \quad . \tag{6.4}$$

Waves in the same polarization eigen-state are summed coherently, waves in different polarization eigen-states are summed incoherently. In the latter case, no interference occurs. When only two (complex-valued) wave amplitudes contribute, we can write

$$\begin{aligned} I =& \frac{c\, n\, \epsilon_0}{2} |E_1 + E_2|^2 && (6.5)\\ =& \frac{c\, n\, \epsilon_0}{2} \left\{ |E_1|^2 + |E_2|^2 + 2|E_1|\,|E_2|\, \cos\Delta\phi \right\} && (6.6)\\ =& I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Delta\phi && (6.7) \end{aligned}$$

where $\Delta\phi$ is the phase difference $\phi_2 - \phi_1$ between the waves with $E_i = |E_i| e^{i\phi_i}$. The detected intensity changes periodically with the phase difference $\Delta\phi$. When both waves have the same intensity, then

$$I = 2I_0(1 + \cos\Delta\phi) = 4I_0 \cos^2 \Delta\phi/2 \quad . \tag{6.8}$$

Michelson interferometer

In a Michelson interferometer one and the same beam splitter splits and recombines the waves that travel along the two arms. After recombination, they have acquired a phase difference $\Delta\phi$ due to a path length difference or a difference in index of refraction of the medium in the interferometer arm.

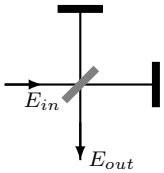

Figure 6.1: Michelson interferometer

The Michelson interferometer has two outputs: one through the fourth facet of the beam splitter, and another one in reverse direction of the incoming beam. The beam-splitter is reciprocal and loss-free, i.e. its (complex valued) refection and transmission coefficients r and t do not depend on the direction of travel, and theirs squares add up to one: $|r|^2 + |t|^2 = 1$. The fields at the outputs are

$$\begin{aligned} E_a =& E_0 \left(r\, t + t\, r\, e^{i\Delta\phi} \right) && (6.9)\\ E_b =& E_0 \left(r\, r + t\, t\, e^{i\Delta\phi} \right) \quad . && (6.10) \end{aligned}$$

Energy conservation requires that

$$|E_a|^2 + |E_b|^2 = |E_0|^2 \tag{6.11}$$

independent of $\Delta\phi$. This is fulfilled when

$$\frac{r}{t} = -\frac{r^\star}{t^\star} \quad . \tag{6.12}$$

This means that the phases of refection and transmission coefficients differ by 90 degrees, i.e.

$$r = |r| e^{i\alpha} \quad \text{and} \quad t = |t| e^{i(\alpha \pm \pi/2)} \quad . \tag{6.13}$$

The Fresnel coefficients of a single interface fulfill this requirement, as do all combinations of interfaces and materials. When one output of the interferometer is at its maximum, the other is at its minimum. Total power is conserved.

Mach-Zehnder interferometer

In a Mach-Zehnder interferometer, one beam splitter separates the waves and a second beam splitter recombines them. Again, we have two output ports of the interferometer. In a Mach-Zehnder interferometer, they are more easily accessible than in a Michelson interferometer because the incoming beam takes a different path. When the path length in both arms is the same ($\Delta\phi = 0$) the symmetric output takes all power, i.e., each beam is transmitted and reflected once.

Mach-Zehnder interferometers are used in optical telecommunications to modulate the intensity of a beam in an optical fiber. By slightly changing the refractive index in one arm of the interferometer, a phase difference is built up that switches one output from bright to dark. The power of the second output is then dumped.

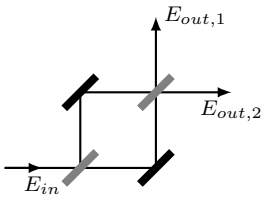

Figure 6.2: Mach-Zehnder interferometer

Sagnac interferometer

The Sagnac interferometer is an example of a common-path interferometer. Both beams travel the same path, but in reverse direction. When both beams have seen the same index of refraction, the total power is reflected back through the input port, as in this direction, each beam is reflected and transmitted once by the only beam splitter of the interferometer.

Only a few effects are capable of causing an effective path length difference for the counterpropagating beams: The interferometer could rotate, magneto-optical effects of non-reciprocal media could be involved, or short laser pulses could interact with a rapidly changing medium. The rotating Sagnac interferometer is the basis of a laser gyroscope. The phase difference between the beams depends on the angular velocity of rotation ω, the area A of the interferometer, and the dot product between their directions:

$$\phi = \frac{8\pi}{\lambda c}\,\boldsymbol{\omega}\cdot\boldsymbol{A} \quad . \tag{6.14}$$

A fiber-loop increases the effect.

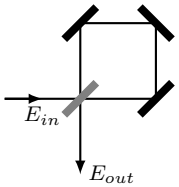

Figure 6.3: Sagnac interferometer

Fabry-Perot interferometer

All of the examples above were two-wave interference. The Fabry-Perot interferometer is an example of multiwave interference. It consists of two parallel, partially reflecting mirrors. The transmitted beam is composed of several waves: the directly transmitted wave, a wave with one round trip, a wave with two round trips, and so on. Compare to the direly transmitted wave E_1, all other waves acquire for each round trip a complex factor h that contains two reflections at the mirrors and a phase lag due to the distance between the mirrors. As equation, this reads

$$E_{\text{out}} = E_1\left(1 + h + h^2 + \dots\right) = \frac{E_1}{1-h} \quad . \tag{6.15}$$

The intensity is thus

$$I = \frac{I_0}{|1-h|^2} = \frac{I_0}{(1-|h|)^2 + 4|h|\sin^2\phi/2} \tag{6.16}$$

Figure 6.4: Fabry-Perot interferometer. The mirrors reflect less than 100% of the light.

with $h = |h|e^{i\phi}$. This can be written as

$$I = \frac{I_{\text{max}}}{1 + (2\mathcal{F}/\pi)^2 \sin^2 \phi/2} \tag{6.17}$$

with the *finesse* $\mathcal{F}$

$$\mathcal{F} = \frac{\pi\sqrt{|h|}}{1 - |h|} \quad . \tag{6.18}$$

The higher the quality of the mirrors, the better their reflectivity, the closer $|h|$ gets to one, and the higher the finesse.

The Fabry-Perot interferometer is a spectral filter. Let us assume that the reflectivity r of the mirrors does not depend on the wavelength. The distance d between the mirrors leads to a wavelength-dependent phase shift ϕ

$$\phi = 2\,d\,k = \frac{4\pi d}{\lambda} = \frac{4\pi d\nu}{c} \tag{6.19}$$

with the optical frequency ν. We get a peak in transmission when $\phi/2 = n\pi$ or when the frequency of the optical wave changes by

$$\Delta\nu_{\text{FSR}} = \frac{c}{2d} \tag{6.20}$$

which is called the *free spectral range*. The FWHM of a peak is as phase $\delta\phi$ or frequency $\delta\nu$

$$\delta\phi = \frac{2\pi}{\mathcal{F}} \qquad \delta\nu = \frac{c}{2d\mathcal{F}} \quad . \tag{6.21}$$

The finesse $\mathcal{F}$ can be written as

$$\mathcal{F} = \frac{\Delta\nu_{\text{FSR}}}{\delta\nu} \tag{6.22}$$

and defines the usefulness of the interferometer by free span between the peaks per peak width. The Fabry-Perot interferometer is a narrow spectral filter, but its transmission function has repeated maxima. One must therefore know in advance that the expected signal is spectrally narrower than the free spectral range. Or you can cascade several interferometers of different fineness, with the first one acting as a pre-filter for the others. By tuning the distance between the mirrors, the Fabry-Perot interferometer can be used as a spectrometer.

Optics of layered media: Transmission and scattering matrix

Multiple beam interference is tedious because you have to keep track of the multiple reflection and transmission paths. In the case of the Fabry-Perot interferometer, we could identify the series and have an analytical solution. With three or more layers, this becomes difficult. One way out is the T-matrix method. It is a versatile technique for studying the optical properties of layered media, which are stacks of unstructured films of materials with different dielectric functions. These can be dielectric materials leading to e.g. Bragg reflections and dielectric filters, or metal films leading to surface plasmons. We will discuss transmission and reflection of such stacks.

In a layered medium, a wave traveling through the stack of layers is partially reflected and partially transmitted at each interface. The multiple reflections interfere with each other. To keep track of this, we use in each layer

a combined wave traveling in the $+z$ direction and one traveling in the $-z$ direction. These waves mix at interfaces. This formalism is described in chapter 7 of Saleh and Teich, 1991 and Yeh, 2005. A similar formalism with an E and B field traveling in the same direction is described in Pedrotti et al., 2008 and Macleod, 2001.

Let us assume that we have left of the interface a wave traveling to the right ($+z$ direction) of amplitude U_1^+, and one wave traveling to the left of amplitude U_1^-. On the right side of the interface, we get the amplitudes $U_2^\pm$ by multiplication with a *transmission* or *transfer* matrix $\mathbf{M}$

$$\begin{pmatrix} U_2^+ \\ U_2^- \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \cdot \begin{pmatrix} U_1^+ \\ U_1^- \end{pmatrix} = \mathbf{M} \begin{pmatrix} U_1^+ \\ U_1^- \end{pmatrix} \quad . \tag{6.23}$$

Below we will derive transmission matrices $\mathbf{M}_i$ for every interface and the homogeneous space in between. The full stack can then be described by a product matrix, multiplying together all partial matrices $\mathbf{M}_i$ along the stack

$$\mathbf{M}_{\text{total}} = \mathbf{M}_n \cdot \mathbf{M}_{n-1} \cdots \mathbf{M}_2 \cdot \mathbf{M}_1 \quad . \tag{6.24}$$

This is a very convenient feature of the transmission matrix. Note that we label the interactions from left to right with 1 to n, but the matrices are multiplied from right to left, as mathematics has it origin in Arabic culture.

An inconvenient feature of the transmission matrix is that its matrix element have no direct physical meaning. The problem is that we multiply on the matrix a vector that is half an input, half an output of this interface. We know what comes out (travels to the left), and the matrix should tell us what comes in from the other side. In this sense, the related *scattering* matrix $\mathbf{S}$ is closer to physical meaning:

$$\begin{pmatrix} U_2^+ \\ U_1^- \end{pmatrix} = \begin{pmatrix} t_{12} & r_{21} \\ r_{12} & t_{21} \end{pmatrix} \cdot \begin{pmatrix} U_1^+ \\ U_2^- \end{pmatrix} = \mathbf{S} \begin{pmatrix} U_1^+ \\ U_2^- \end{pmatrix} \quad . \tag{6.25}$$

The scattering matrix connects waves traveling towards the interface with those traveling away from the interface. The entries t_{ij} and r_{ij} are the transmission and reflection coefficients for the amplitudes of the waves traveling from i to j (i.e. 12 is traveling towards the right, $+z$ direction). However, for the scattering matrix $\mathbf{S}$, the full stack can not be calculated by multiplying together all partial matrices.

It is therefore convenient to switch between both representations, derive the scattering matrix $\mathbf{S}$ for each situation, and then convert into a transmission matrix $\mathbf{M}$. The relations are[2]

$$\mathbf{M} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \frac{1}{t_{21}} \begin{pmatrix} t_{12}t_{21} - r_{12}r_{21} & r_{21} \\ -r_{12} & 1 \end{pmatrix} \tag{6.26}$$

$$\mathbf{S} = \begin{pmatrix} t_{12} & r_{21} \\ r_{12} & t_{21} \end{pmatrix} = \frac{1}{D} \begin{pmatrix} AD - BC & B \\ -C & 1 \end{pmatrix} \tag{6.27}$$

as long as D or t_{21} are not zero.

The transmission in backward direction t_{21} is thus the reciprocal of the D-element of $\mathbf{M}_{\text{total}}$. The transmission in forward direction is

$$t_{12} = \frac{\det \mathbf{M}_{\text{total}}}{D} \tag{6.28}$$

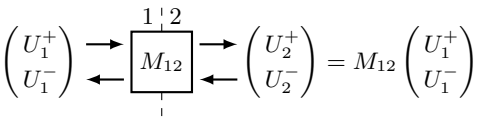

Figure 6.5: The operation of the transmission matrix

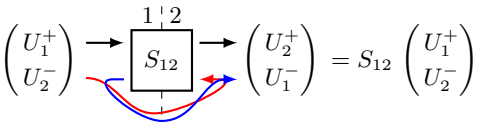

Figure 6.6: The operation of the scattering matrix

[2] Saleh and Teich, 1991 eq. 7.7

and similar for the reflection from the front side

$$r_{12} = -\frac{C}{D} \quad . \tag{6.29}$$

Electrical fields

We need to define the physical meaning of the amplitudes $U_i^{\pm}$ to be able to calculate the reflection (r_{ij}) and transmission (t_{ij}) coefficients. We assume plane waves

$$\mathbf{E}\, e^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} = \hat{\mathbf{E}}\, U\, e^{i\,k_z z}\, e^{i\,k_x x}\, e^{-i\omega t} \tag{6.30}$$

where the wave vector $\mathbf{k}$ lies in the xz-plane, U defines the amplitude of the wave and $\hat{\mathbf{E}}$ the polarization direction. With the full length of the wave vector in vacuum $k_0 = 2\pi/\lambda$ and the refractive index n of the medium we get

$$k_z^2 + k_x^2 = n^2\, k_0^2 \quad . \tag{6.31}$$

The polarization directions are

$$\hat{\mathbf{E}}^{(s)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \hat{\mathbf{E}}^{(p)} = \frac{1}{n\,k_0} \begin{pmatrix} \pm k_z \\ 0 \\ k_x \end{pmatrix} \quad . \tag{6.32}$$

The $\pm$-sign takes the sign of the direction of travel, see Fig. 2.2 in Novotny and B. Hecht, 2012. Note that with this definition we have $|\hat{\mathbf{E}}| = 1$, which differs from problem 12.4 in Novotny and B. Hecht, 2012.

The left and right traveling waves are thus

$$\mathbf{E}^{+} = \hat{\mathbf{E}}\, U^{+}\, e^{+i\,k_z z} \quad \text{and} \quad \mathbf{E}^{-} = \hat{\mathbf{E}}\, U^{-}\, e^{-i\,k_z z} \tag{6.33}$$

where we have split off the global term $e^{i\,(k_x x - \omega t)}$.

Propagation matrix

Before we come to interfaces, let us discuss the transmission matrix of a homogeneous material layer j of thickness d_j and (complex) refractive index n_j. Relevant is the z-component of the (complex) wave vector $k_{z,j}$. Note that we do *not* use the sign of $k_{z,j}$ to describe the direction of travel. Independent of the propagation direction, each wave sees a reflection coefficient $r = 0$ and a (complex) transmission coefficient t

$$t = t_{12} = t_{21} = e^{+i\,k_{z,j}\,d_j} \quad . \tag{6.34}$$

The transmission matrix of a homogeneous medium is thus

$$\mathbf{M} = \begin{pmatrix} e^{+i\,k_{z,j}\,d_j} & 0 \\ 0 & e^{-i\,k_{z,j}\,d_j} \end{pmatrix} \quad . \tag{6.35}$$

Interface matrix

The transmission and reflection coefficients of an interface are the Fresnel coefficients r and t for s and p polarization, as defined in chapter 4, especially r^s and r^p differ at normal incidence by a factor of -1. We assume

non-magnetic materials ($\mu = 1$). Independent of polarization dirction, the forward and backward coffeicients are related:

$$r_{12} = -\, r_{21} \tag{6.36}$$

$$t_{12} = \frac{k_{z,1}}{k_{z,2}}\, t_{21} \tag{6.37}$$

so that

$$t_{12}t_{21} - r_{12}r_{21} = t_{21}^2 \frac{k_{z,1}}{k_{z,2}} + r_{21}^2 = 1 \quad . \tag{6.38}$$

With eq. 6.26 we get for both polarization directions the transmission matrix

$$\mathbf{M}_{12} = \frac{1}{t_{21}} \begin{pmatrix} 1 & r_{21} \\ r_{21} & 1 \end{pmatrix} \quad . \tag{6.39}$$

Note that the transmission matrix from medium 1 to medium 2 uses the Fresnel coefficients of the backwards direction! We can write this without referring to Fresnel coefficients as[3] (see also appendix at the end of this chapter)

$$\mathbf{M}_{12} = \frac{1}{2\eta} \begin{pmatrix} 1+\kappa & 1-\kappa \\ 1-\kappa & 1+\kappa \end{pmatrix} \tag{6.40}$$

with

$$\kappa = \eta^2 \frac{k_{z,1}}{k_{z,2}} \quad \text{and} \quad \eta^s = 1 \quad \text{or} \quad \eta^p = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \quad . \tag{6.41}$$

The factors η in front of the transmission matrix $\mathbf{M}_{12}$ can be collected in front of the total transmission matrix $\mathbf{M}_{\text{total}}$, in case one is not interested in the distribution of the fields inside the stack. Then, all η^p collapse into $\sqrt{\epsilon_{\text{first}}/\epsilon_{\text{last}}}$, which is equal to one in case the terminating half-spaces of the layered medium have both the same dielectric constant.

[3] In problem 12.4 in Novotny and B. Hecht, 2012 the leading $1/\eta$ seems to be missing!

Example: Fabry-Perot etalon

A simple way to construct a Fabry-Perot interferometer is a glass plate with parallel surfaces. Each surface acts as (weak) mirror. Let us use the T-matrix method to calculate its transmission properties. We have three media, where medium 2 is glass and media 1 and 3 are air. The total transmission matrix is thus

$$\mathbf{M}_{\text{total}} = \mathbf{M}_{23}\, \mathbf{M}_{\text{prop},2}\, \mathbf{M}_{12} \quad . \tag{6.42}$$

Inserting the definition from above and using $r_{12} = r_{32} = -r_{21}$ and $t_{12} = t_{32} = (k_{z,3}/k_{z,2})\, t_{21}$ we get

$$\mathbf{M}_{\text{total}} = \frac{k_{z,1}}{k_{z,2}} \begin{pmatrix} 1 & -r_{21} \\ -r_{21} & 1 \end{pmatrix} \begin{pmatrix} e^{+i\,\phi_2} & 0 \\ 0 & e^{-i\,\phi_2} \end{pmatrix} \begin{pmatrix} 1 & r_{21} \\ r_{21} & 1 \end{pmatrix} \tag{6.43}$$

with $\phi_2 = k_{z,2}\, d_2$. The overall transmission is

$$t = \frac{\det \mathbf{M}_{\text{total}}}{D} = \frac{(1-r_{21}^2)^2 e^{+i\,\phi_2}}{1 - r_{21}^2\, e^{+i\,2\phi_2}} \tag{6.44}$$

which agrees with eq. 6.15 using $h = r_{21}^2\, e^{+i\,2\phi_2}$.

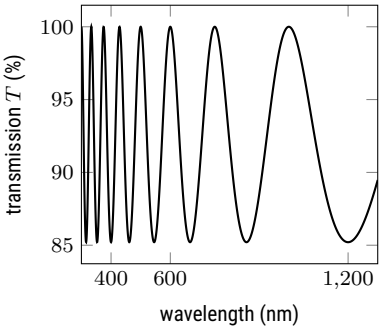

Figure 6.7: Transmission of a 1 µm thick glass layer in air.

Quarter-wave film as anti-reflection coating

The most simple anti-refection coating for an interface between two media of refractive index n_1 and n_3 is a film of refractive index $n_2 = \sqrt{n_1\, n_3}$ of thickness $d = \lambda/4 = \lambda_0/(4n_2)$. In this case

$$\mathbf{M}_{\text{total}} = \frac{k_{z,1}}{k_{z,2}} \begin{pmatrix} 1 & r_{32} \\ r_{32} & 1 \end{pmatrix} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} 1 & r_{21} \\ r_{21} & 1 \end{pmatrix} \quad . \tag{6.45}$$

The total reflection vanishes when the B and C elements of the transmission matrix vanish, which is the case, when

$$r_{32} = r_{21} \quad . \tag{6.46}$$

This holds for perpendicular incidence at the design wavelength. At other wavelength the reflection increases, up to the value of an uncoated n_1–n_3 interface.

A spectrally broader antireflective coating requires multiple layers with at least two different refractive indices. Designing such multilayer stacks is an art in itself. Some details are discussed in CVI Melles Griot, 2009.

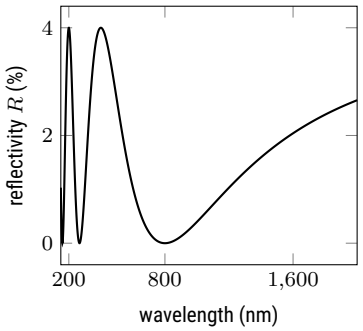

Figure 6.8: Quarter-wave AR coating on air-glass interface.

Distributed Bragg reflector (DBR)

An example of a multilayer coating is the distributed Bragg reflector (DBR). It consists of alternative layers of refractive indices n_1 and n_2, with thicknesses d_1 and d_2, embedded in a medium of index n_1. The difference between n_1 and n_2 will be small because we have to stack similar materials. Each surface will reflect only a little. But we let the many single reflections add up coherently, in phase, but with a good choice of the distances d_i. In this way a very large total reflectivity can be achieved. These DBRs are used as cavity mirrors in semiconductor lasers.

The transmission matrix of one pair of layers is

$$\mathbf{M}_{\text{pair}} = \mathbf{M}_{\text{prop},1}\mathbf{M}_{2,1}\mathbf{M}_{\text{prop},2}\mathbf{M}_{1,2} \quad . \tag{6.47}$$

The full stack consists of N pairs

$$\mathbf{M}_{\text{full}} = \mathbf{M}_{\text{pair}}^{N} \quad . \tag{6.48}$$

Saleh and Teich, 1991 show steps of the analytical solution in chapter 7.1. One finds *stop bands*, i.e., spectral regions of high reflectivity, around the Bragg frequency ν_B at which the round-trip phase in a pair is 2π, i.e.

$$2\,\frac{2\pi}{c_0}\,\nu_B\,(n_1 d_1 + n_2 d_2) = 2\pi \quad . \tag{6.49}$$

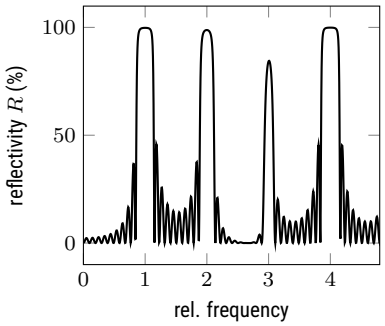

Figure 6.9: Transmission of 10 pairs of air and glass.

Appendix: derivation of eq. 6.40

We start from

$$\mathbf{M}_{12} = \frac{1}{t_{21}} \begin{pmatrix} 1 & r_{21} \\ r_{21} & 1 \end{pmatrix} \tag{6.50}$$

and abbreviate the Fresnel coefficients as

$$r_{21}^s = \frac{k_{z,2} - k_{z,1}}{k_{z,1} + k_{z,2}} = \frac{b-a}{a+b} \tag{6.51}$$

$$t_{21}^s = \frac{2\,k_{z,2}}{k_{z,1} + k_{z,2}} = \frac{2b\eta}{a+b} \tag{6.52}$$

$$r_{21}^p = \frac{\epsilon_1 k_{z,2} - \epsilon_2 k_{z,1}}{\epsilon_2 k_{z,1} + \epsilon_1 k_{z,2}} = \frac{b-a}{a+b} \tag{6.53}$$

$$t_{21}^p = \frac{2\sqrt{\epsilon_1 \epsilon_2}\,k_{z,2}}{\epsilon_2 k_{z,1} + \epsilon_1 k_{z,2}} = \frac{2b\eta}{a+b} \tag{6.54}$$

with $a = \epsilon_2 k_{z,1}$, $b = \epsilon_1 k_{z,2}$ and $\eta = \sqrt{\epsilon_2/\epsilon_1}$. In the case of s-polarization, the ϵ_i are ignored / set to one. With this we get

$$\mathbf{M}_{12} = \frac{a+b}{2b\eta} \begin{pmatrix} 1 & (b-a)/(a+b) \\ (b-a)/(a+b) & 1 \end{pmatrix} = \frac{1}{2b\eta} \begin{pmatrix} b+a & b-a \\ b-a & b+a \end{pmatrix} \tag{6.55}$$

$$= \frac{1}{2\eta} \begin{pmatrix} 1+\frac{a}{b} & 1-\frac{a}{b} \\ 1-\frac{a}{b} & 1+\frac{a}{b} \end{pmatrix} = \frac{1}{2\eta} \begin{pmatrix} 1+\kappa & 1-\kappa \\ 1-\kappa & 1+\kappa \end{pmatrix} \tag{6.56}$$

with

$$\kappa = \frac{a}{b} = \eta^2 \frac{k_{z,1}}{k_{z,2}} \quad \text{and} \quad \eta^s = 1 \quad \text{or} \quad \eta^p = \sqrt{\frac{\epsilon_2}{\epsilon_1}} \quad . \tag{6.57}$$

References

CVI Melles Griot (2009). *Interference Filter Coatings - CVI Melles Griot Technical Guide, Vol 2, Issue 2*. (visited on 12/19/2023).

Macleod, H. Angus (2001). *Thin film optical filters*. 3. ed. Bristol [u.a.]: Inst. of Physics Publ..

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

Pedrotti, Frank L. et al. (2008). *Optik für Ingenieure*. 4., bearb. Aufl. Berlin [u.a.]: Springer..

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

Yeh, Pochi (2005). *Optical waves in layered media*. Hoboken, NJ: Wiley-Interscience.

Chapter 7
Coherence

Markus Lippitz
January 12, 2023

By the end of this chapter you should be able to identify requirements for intereference contrast.

Overview

So far we have always assumed perfectly deterministic light. We wrote something like

$$u(\boldsymbol{r},t) = U(\boldsymbol{r})\, e^{-i\omega t} \quad \text{with} \quad U(\boldsymbol{r}) = \frac{A}{r}\, e^{i\boldsymbol{k}\boldsymbol{r}} \tag{7.1}$$

and could thus determine the field amplitude at each point in space $\boldsymbol{r}$ for all times t. This is what we call 'coherent'. It is in contrast to a light field which contains some randomness. In this chapter we will discuss ways to describe the statistical nature of a partially random light field. As always when randomness is involved, we will not be able to predict every realization of a random process, but only the average and some other statistical properties of an ensemble of realizations. We will define a measure of the average intensity and the 'randomness' of the light field. The latter is called the coherence function or autocorrelation.

We will discuss a relationship between the coherence function and the spectrum of light (the Wiener-Khinchin theorem), and a method for measuring the diameter of a distant star by analyzing the coherence of the detected light (the Michelson stellar interferometer).

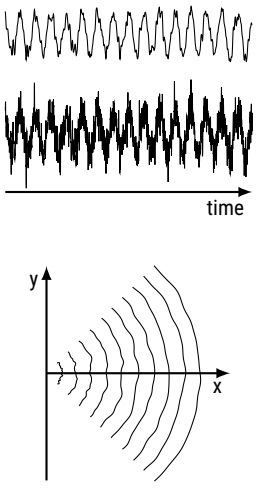

Figure 7.1: Partially coherent waves in space and time.

Intensity

We have written the intensity of a scalar wave (and similar for a vectorial electromagnetic wave) as

$$I(\boldsymbol{r},t) = |u(\boldsymbol{r},t)|^2 \quad . \tag{7.2}$$

This assumes perfect coherence. If we allow randomness, we have to average over many realizations of the same random process, similar to rolling the dice very often and averaging over the result. We write this as

$$I(\boldsymbol{r},t) = \left\langle |u(\boldsymbol{r},t)|^2 \right\rangle \tag{7.3}$$

where the pointed brackets denote the ensemble average. From now on, we will call I the (average) intensity and $|u|^2$ the instantaneous intensity.

The intensity I can be time-dependent or time-independent. The latter case is a statistically stationary process, which means that the average is independent of time, although each individual realization may still vary with time. The light from a light bulb is an example. In the case of a *stationary process*, we can write

$$I(\boldsymbol{r}) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} |u(\boldsymbol{r},t)|^2 \, dt \quad . \tag{7.4}$$

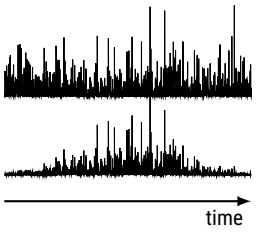

Figure 7.2: Stationary (top) and non-stationary (bottom) wave. Plotted is $|u(t)]^2$.

Temporal coherence

Let us drop the spatial dependence on $\boldsymbol{r}$ for a moment and just look at a single point in space. We want to describe the 'randomness' of the complex-valued field u, assuming a stationary process. The question is how similar are the amplitudes $u(t)$ now and a time τ later, i.e. $u(t+\tau)$. The more similar they are, the more memory the process has, the less random it is. We quantify this by the *temporal coherence function* or *autocorrelation function* $G(\tau)$:

$$G(\tau) = \langle u^\star(t)\, u(t+\tau)\rangle = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} u^\star(t)\, u(t+\tau)\, dt \tag{7.5}$$

with the properties

$$G(\tau) = G^\star(-\tau) \quad \text{and} \quad I = G(0) \quad . \tag{7.6}$$

As an example, consider a field u that is either $+1$ or -1. On average, it should be as often positive as negative, i.e. its average $\langle u(t)\rangle = 0$. The expression $u^\star(t)\, u(t+\tau)$ is positive if u does not change its sign, otherwise it is negative. If the relation between $u(t)$ and $u(t+\tau)$ is completely random, the mean of $u^\star(t)\, u(t+\tau)$ will be zero. On the other hand, if u does not flip within the time τ, G is positive. If u preferentially flips the sign within the time delay τ, $G(\tau)$ will be negative.

It is convenient to remove the intensity from the definition of the coherence function. We define the *degree of temporal coherence* $g(\tau)$ as

$$g(\tau) = \frac{G(\tau)}{G(0)} = \frac{\langle u^\star(t)\, u(t+\tau)\rangle}{\langle u^\star(t)\, u(t)\rangle} \tag{7.7}$$

so that $g(0) = 1$ and $|g(\tau)| \leq 1$. For a coherent wave such as eq. 7.2, we have $|g(\tau)| = 1$.

For partially coherent light, $|g(\tau)|$ decreases with increasing delay τ. The *coherence time* τ_c is a characteristic time of this decrease, describing the width of $|g(\tau)|$. One defines

$$\tau_c = \int_{-\infty}^{\infty} |g(\tau)|^2 \, d\tau \quad . \tag{7.8}$$

Together with the speed of light we get a coherence length $l_c = c_0\, \tau_c$.

Power spectral density

Lets discuss Fourier transformations of random fields $u(t)$. We could just obtain the Fourier transform $v(\nu)$ of $u(t)$ by

$$v(\nu) = \int_{-\infty}^{\infty} u(t)\, e^{i2\pi\nu t}\, dt \quad . \tag{7.9}$$

Since $I = \langle |u|^2 \rangle$ is the intensity (energy per time and area), $\langle |v(\nu)|^2 \rangle$ is an energy spectral density (energy per frequency interval and area). If the process is really stationary, it lasts forever, so its total energy is infinite. It thus makes more sense to truncate the Fourier integral

$$v_T(\nu) = \int_{-T/2}^{T/2} u(t)\, e^{i2\pi\nu t}\, dt \tag{7.10}$$

and defining the *power spectral density* or spectrum $S(\nu)$ (energy per frequency interval, area and time) as

$$S(\nu) = \lim_{T\to\infty} \frac{1}{2T} \langle |v_T(\nu)|^2 \rangle \quad . \tag{7.11}$$

The interesting point for us in this chapter is that the spectrum $S(\nu)$ is the Fourier transform of the autocorrelation function $G(\tau)$.

$$S(\nu) = \int_{-\infty}^{\infty} G(\tau)\; e^{i2\pi\nu\tau}\, d\tau \quad . \tag{7.12}$$

This relation is the *Wiener-Khinchin theorem*. It is rather trivial if the Fourier transform 7.9 exists, i.e. the integral eq. 7.9 converges, because the field $u(t)$ is some kind of non-stationary pulse. The important point of the theorem is that it also works for stationary processes.

As spectrum and autocorrelation function are related by a Fourier transform, also the spectral width $\Delta\nu$ and coherence time τ_c are related by

$$\Delta\nu = \frac{1}{\tau_c} \tag{7.13}$$

using a suitable definition of $\Delta\nu$ (see Saleh and Teich, 1991, eq.11.1-18). The narrower the spectrum of a light source is, the longer is its coherence length.

Interference and temporal coherence

Let us consider a partially coherent wave $u(t)$ described by its autocorrelation function $g(\tau)$, traveling through a Michelson interferometer. At the symmetric output of the interferometer, the total field is $u(t) + u(t+\tau)$, dropping the reflection and transmission coefficients rt in the prefactor. The arm length difference d defines the time lag $\tau = 2d/c_0$. We measure the *interferogram*, the relation between I and τ

$$\begin{aligned} I(\tau) &= \langle |u(t) + u(t+\tau)|^2 \rangle && (7.14)\\ &= 2\,\langle |u|^2 \rangle + \langle u^\star(t) u(t+\tau) \rangle + \langle u(t) u^\star(t+\tau) \rangle && (7.15)\\ &= 2I_0 + 2\Re\{G(\tau)\} && (7.16)\\ &= 2I_0\,(1 + |g(\tau)| \cos[\arg\{g(\tau)\}]) && (7.17) \end{aligned}$$

where $\arg\{|a|e^{i\phi}\} = \phi$. For a perfectly coherent wave with frequency ω, we have $g(\tau) = e^{-i\omega\tau}$ and recover the usual result. For a partially coherent wave, $|g(\tau)|$ drops in amplitude with delay τ, so that the oscillation fringes reduce in amplitude. We call the *visibility* $\mathcal{V}$ (or modulation contrast or depth)

$$\mathcal{V} = \frac{I_{\text{max}} - I_{\text{min}}}{I_{\text{max}} + I_{\text{min}}} = |g(\tau)| \quad . \tag{7.18}$$

The magnitude of the autocorrelation gives the contrast of the interference fringes; its phase shifts the peaks relative to the perfectly coherent case.

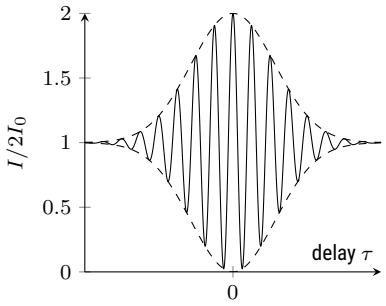

Figure 7.3: Interferogram of a partially coherent wave.

Fourier-transform spectroscopy

Using the Wiener-Khinchin theorem in reverse direction

$$G(\tau) = I_0 g(\tau) = \int_0^\infty S(\nu) e^{-i2\pi\nu\tau}\, d\nu \tag{7.19}$$

and noting that $S(\nu)$ is real, we can write eq. 7.17 as

$$I(\tau) = 2\int_0^\infty S(\nu)\left[1 + \cos(2\pi\nu\tau)\right]\, d\nu \quad . \tag{7.20}$$

This can be reversed to obtain the power spectral density $S(\nu)$ from the interferogram $I(\tau)$

$$S(\nu) = 2\int_0^\infty \left[I(\tau) - \frac{1}{2}I(0)\right]\cos(2\pi\nu\tau)\, d\tau \quad . \tag{7.21}$$

Measuring the spectrally integrated transmitted power through a Michelson interferometer as a function of delay τ and roughly Fourier transforming the result gives the spectrum of the light source.

How does this happen? The spectral transmission function of a Michelson interferometer is similar to that of a Fabry-Perot interferometer, a $\cos^2$ function where the distance between the peaks increases with decreasing path length difference d. When we transmit our unknown light spectrum $S(\nu)$ through the Michelson interferometer, we multiply it by this transmission function. Measuring the total power is a spectral integration. This is what a cosine transform does. Eq. 7.21 simply reverses the operation of the interferometer.

Fourier transform spectroscopy is particularly useful in cases where no other spectrometer can be used, such as in the infrared spectral range. This is called Fourier Transform Infrared (FTIR) spectroscopy. In this spectral region, it is technically challenging to build bright light sources and efficient detectors. The advantage of FTIR is that you only need one detector (not an array of detectors) and about half of the light power is always incident on that detector. In a grating spectrometer, on the other hand, the total power is distributed over many pixels. However, controlling the delay τ accurately enough is a challenge.

Spatial coherence

Similar to two points t and $t + \tau$ in time we can also compare two points $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ in space. We define spatial autocorrelation functions

$$G(\boldsymbol{r}_1, \boldsymbol{r}_2) = \langle u^\star(\boldsymbol{r}_1, t)\, u(\boldsymbol{r}_2, t)\rangle \tag{7.22}$$

and

$$g(\boldsymbol{r}_1, \boldsymbol{r}_2) = \frac{G(\boldsymbol{r}_1, \boldsymbol{r}_2)}{\sqrt{I(\boldsymbol{r}_1)\, I(\boldsymbol{r}_2)}} \quad . \tag{7.23}$$

Again, the coherence or autocorrelation typically decreases with the distance between $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. The *coherence area* A_c is an area within which $g(\boldsymbol{r}_1, \boldsymbol{r}_2)$ has only dropped to a critical value. For a hot emitting surface, the coherence area is about λ^2, so in most cases this source can be assumed to be incoherent. If the coherence area is larger than, say, the size of a pinhole, the beam can be assumed to be fully spatially coherent.

Sometimes it is useful or necessary to combine spatial and temporal coherence:

$$g(\boldsymbol{r}_1, \boldsymbol{r}_2, \tau) = \frac{\langle u^\star(\boldsymbol{r}_1, t)\, u(\boldsymbol{r}_2, t+\tau)\rangle}{\sqrt{I(\boldsymbol{r}_1)\, I(\boldsymbol{r}_2)}} \quad . \tag{7.24}$$

Double-slit experiment with partially coherent waves

Above, we used a Michelson interferometer to test and demonstrate the effect of temporal coherence by interfering a wave with a time-shifted copy of it. Now we do the same with spatial coherence using a Young double-slit experiment.

A partially coherent wave u is described by its autocorrelation function $g(\boldsymbol{r}_1, \boldsymbol{r}_2, \tau)$. It hits an opaque screen with two small spherical holes at positions $\boldsymbol{r}_{1,2} = (\pm a, 0, 0)$. Each hole results in a diffracted spherical wave. On a screen at distance d we find an interference pattern which we observe along a line $\boldsymbol{r} = (x, 0, d)$. For simplicity, we assume that the intensity I_0 of both spherical waves is the same.

At point x along our line of observation $\boldsymbol{r}$, two waves interfere. We calculate the difference τ in the travel time between the hole at $\boldsymbol{r}_{1,2}$ and the screen at $\boldsymbol{r}$, since this time difference τ enters the coherence function. By simple geometry we get

$$\tau = \frac{|\boldsymbol{r} - \boldsymbol{r}_1| - |\boldsymbol{r} - \boldsymbol{r}_2|}{c_0} = \frac{(x+a)^2 - (x-a)^2}{2dc_0} = \frac{2ax}{dc_0} = \frac{\theta}{c_0}\, x \tag{7.25}$$

where θ is the angle between the holes as seen from the screen. Following the same scheme as above with the Michelson interferometer, we get

$$I(x) = 2I_0\,(1 + |g(\boldsymbol{r}_1, \boldsymbol{r}_2, \tau)|\cos[\arg\{g(\boldsymbol{r}_1, \boldsymbol{r}_2, \tau)\}]) \quad . \tag{7.26}$$

As above, the visibility $\mathcal{V}$ of the fringes drops with deceasing coherence. The characteristic decay length is coherence length divided by the opening angle θ. The period of the spatial oscillation is $\bar{\lambda}/\theta$.

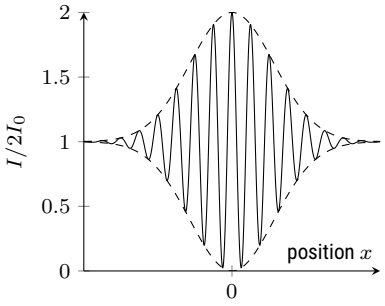

Figure 7.4: Young double slit interference of a partially coherent wave, assuming very small holes.

Gain of spatial coherence by propagation

A spatially incoherent light source can result in spatially coherent light! One way to achieve this is simply to let it propagate long enough. We can divide the light source into patches of the size of the spatial coherence area, A_c. Within this area (which can be just a point) the source is coherent. Since the area has a finite size, the light is diffracted as it propagates, covering a larger and larger area. At a distance, neighboring points on a screen will see a very

similar mixture of fields from the individual coherence areas of the source. Spatial coherence thus has increased.

Lets look at this more formally. We use the impulse response h to connect the field u_1 in the source plane with the field u_2 in the target plane, as introduced in the chapter 3 on Fourier optics (see, e.g. eq. 3.18)

$$u_2(\boldsymbol{r}) = \int h(\boldsymbol{r}, \boldsymbol{r}')\, u_1(\boldsymbol{r}')\, d\boldsymbol{r}' \quad . \tag{7.27}$$

Using the definition of spatial coherence function $G(\boldsymbol{r}_1, \boldsymbol{r}_2)$ (eq. 7.22), we can calculate the spatial coherence in the target plane as function of the coherence in the source plane

$$G_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = \iint h^\star(\boldsymbol{r}_1, \boldsymbol{r}_1')\, h(\boldsymbol{r}_2, \boldsymbol{r}_2')\, G_1(\boldsymbol{r}_1', \boldsymbol{r}_2')\, d\boldsymbol{r}_1' d\boldsymbol{r}_2' \quad . \tag{7.28}$$

When the light source is fully incoherent, then G_1 is different from zero only for $\boldsymbol{r}_1' = \boldsymbol{r}_2'$ and at these points it is identical to the intensity I. Things thus simplify to

$$G_2(\boldsymbol{r}_1, \boldsymbol{r}_2) = \int h^\star(\boldsymbol{r}_1, \boldsymbol{r}')\, h(\boldsymbol{r}_2, \boldsymbol{r}')\, I(\boldsymbol{r}')\, d\boldsymbol{r}' \quad . \tag{7.29}$$

Now we can use the same approximations that led to the optical Fourier transform by propagation in the Fraunhofer approximation (eq. 3.24). When propagating in free space over a long distance, each impulse response effectively contributes an exponential function identical to that of a Fourier transform. The normalized coherence function g_2 at the target plane is thus the spatial 2D Fourier transform $\mathcal{I}_1$ of the intensity at the source plane.

$$|g_2(x_1, y_1, x_2, y_2)| = \frac{\left|\mathcal{I}_1\left(\frac{x_1 - x_2}{\lambda d}, \frac{y_1 - y_2}{\lambda d}\right)\right|}{\mathcal{I}_1(0, 0)} \quad . \tag{7.30}$$

Radiation of an incoherent circular source

As an example, lets look at the sun. We model it as incoherent circular source. Within the circle, it should have a constant intensity. It is convenient to combine the radius a and the distance d to the apparent opening angle θ_s of the source as seen from the target plane

$$\theta_s = \frac{2a}{d} \quad . \tag{7.31}$$

The 2D Fourier transform of a circle is similar to a sinc-function, replacing the sine by a Bessel function J_1 (see Appendix A). In total we get

$$|g_2(x_1, y_1, x_2, y_2)| = \left|\frac{2J_1(\pi\rho\theta_s/\lambda)}{\pi\rho\theta_s/\lambda}\right| \tag{7.32}$$

using $\rho^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2$. The first zero of the Bessel function defines a characteristic radius in the target plane

$$\rho_c = 1.22\frac{\lambda}{\theta_s} \quad . \tag{7.33}$$

We see our sun under an angle of about $0.5° \approx 8.7 \cdot 10^{-3}$ rad. At a wavelength of 500 nm, the coherence radius ρ_c is about 70 µm.

To generate spatially coherent light it is thus sufficient to let sunlight shine through about 140 µm diameter pinhole, without any lens or similar. The sun's intensity is about 1 kW/m^2 on earth on a bright day, so that about 15 µW pass through the pinhole. To also obtain temporally coherent light, we need to spectrally filter to a narrow enough spectral range (eq. 7.12), which reduces the power further.

Michelson stellar interferometer

We can also turn the argument around. By measuring the coherence radius ρ_c of a distant star, we can determine its apparent angle θ_s. Together with a known distance d we can thus measure its radius a. This is the idea behind the Michelson stellar interferometer. The star Betelgeuse[1] in the constellation Orion (α-Orionis) is one of the brightest stars in our northern sky. At a wavelength of 570 nm a coherence radius of $\rho_c = 3.1$ m was measured, corresponding to a source angle of $\theta_s = 22.6 \cdot 10^{-8}$ rad. The distance to Betelgeuse is 548 light-years ($5.2 \cdot 10^{18}$ m), so its diameter is $1.2 \cdot 10^{12}$ m, which is within 20 % of the Wikipedia value. Betelgeuse is a red supergiant. Its diameter is 1500 times larger than the Sun, or about 7 times larger than the Earth's orbit around the Sun.

[1] dt. Beteigeuze

How do you do this in practice? We need to collect light from the star at two points on Earth, separated by a variable distance ρ, and determine the visibility $\mathcal{V}$ of the interference fringes when the light beams overlap. The advantage is that one does not need an optical telescope of diameter ρ. Even the optical telescopes at distance ρ need not be of exceptionally high quality, since we only need to collect light, not resolve the star. As long as the star is much brighter than its immediate surroundings, even simple searchlights have been used.

The collected light must be directed by mirrors to a central position and interfered. Fluctuations in the atmosphere will lead to a phase difference between the two telescopes, which will randomly shift the position of the fringes. But we are not interested in the position of the fringes, only in their contrast $\mathcal{V}$, so even these fluctuations can be tolerated.[2]

[2] see Brooker, 2008 for a discussion

Intensity autocorrelation

To detect interference fringes in the Michelson stellar interferometer, the light from the two remote telescopes must be brought to a common point without losing the phase relationship. This is possible, but inconvenient. Here the intensity (or second order) autocorrelation comes in handy, as R. Hanbury Brown and R. Q. Twiss demonstrated in 1956 (Hanbury Brown and Twiss, 1956). We define

$$G^{(2)}(\tau) = \langle u^\star(t)\, u^\star(t+\tau)\, u(t) u(t+\tau)\rangle = \langle I(t)\, I(t+\tau)\rangle \tag{7.34}$$

and label all our old (amplitude) correlations as $G^{(1)}$ or $g^{(1)}$. So we are correlating intensities, not fields. You have to think a bit about the term 'intensity' here. It means that we average over some time to get the envelope modulated by $e^{i\omega t}$, but that we do not average too much to preserve the fluctuations. Only the brackets average over a long time.

Since second-order autocorrelation is also based on the same fields as first-order autocorrelation, there is a relationship between the two, at least for classical (chaotic) light[3].

$$g^{(2)}(\tau) - 1 = \left| g^{(1)}(\tau) \right|^2 \quad . \tag{7.35}$$

[3] light from a laser or an atom does not follow this relationship

It is therefore sufficient to measure the intensity of the light at the two telescopes with sufficient time resolution and then calculate $g^{(2)}(\tau)$ to determine the diameter of the stars. This effectively replaces the light beam between the telescopes with a cable.

References

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

Hanbury Brown, R and R Q Twiss (1956). “A Test of a new type of stellar interferometer on Sirius”. In: *Nature* 178.4541, pp. 1046–1048.

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

Part V

Quantum optics

Chapter 8
Quantum Optics

Markus Lippitz
January 26, 2024

By the end of this chapter, you should be able to explain the results of experiments by the properties of photons.

Overview

Optics is the study of the interaction of light and matter. We can use different types of models to describe the light and the matter part of it. In all previous chapters we treated light as a scalar or electromagnetic wave. We have also described matter as a classical Lorentz oscillator. Although we have not used it in this lecture, you have seen in other places how to describe matter (especially electrons) with quantum mechanics. In most cases it is sufficient to use quantum mechanics only for the matter part and to keep a classical description for the light part. For example, the external photoelectric effect requires only that the electron be quantized. Light can be described classically.

In this chapter we will go beyond this. We will discuss the so-called 'second quantization'. We now use quantum mechanics to describe the light part by introducing the photon. We will use this formalism to describe photon anti-bunching in the emission of atoms, the Hong-Ou-Mandel experiment of identical photons at a beam splitter, and entanglement between photons.

The photon

A photon is the quantum of energy in a single optical mode. This idea goes back to Max Planck and the description of black-body radiation. We assume a resonator cavity that forms the optical mode at an angular frequency ω. The mode is characterized by the associated electric field $\boldsymbol{E}_{\text{mode}}(\boldsymbol{r},t)$, the wave vector $\boldsymbol{k}$ and a polarization state. The central idea is to quantize the energy of each mode:

$$E_{\text{mode}} = \left(n + \frac{1}{2}\right) \hbar\omega \tag{8.1}$$

where n is the number of photons in this mode.

We can connect the photon description of electromagnetic fields to the classical description by defining a vacuum field amplitude. Independent of

the description, a mode should contain the same energy. In the dark, i.e. in the state $n = 0$, quantum mechanics gives an eigen-energy $E_0 = 1/2\,\hbar\omega$. This is what we require also from classical electrodynamics[1]

[1] Fox, 2007, chap. 7.5.

$$E_0 = \int_{\text{cavity}} \frac{1}{2} \left(\boldsymbol{H} \cdot \boldsymbol{B} + \boldsymbol{E} \cdot \boldsymbol{D} \right) \, d\boldsymbol{r} = \int_{\text{cavity}} \epsilon_0 \boldsymbol{E}^2 \, d\boldsymbol{r} = \frac{1}{2}\,\hbar\omega \tag{8.2}$$

so that

$$E_{vac} = \sqrt{\frac{\hbar\omega}{2\epsilon_0\, V}} \tag{8.3}$$

is the amplitude of the field in the dark vacuum, with V being the volume of the cavity.[2]

[2] This is the reason we require a cavity. Otherwise the integral would diverge.

A photon carries a *momentum* of

$$\boldsymbol{p} = \hbar \boldsymbol{k} \tag{8.4}$$

leading to, e.g. recoil when emitting or reflecting a photon. Position and momentum are connected by an uncertainty relation, which also motivates diffraction of photons at apertures.

The photon is a *Boson*. All photons within a mode are identical, but the decomposition of a given electric field into modes is not unique. As usual for Bosons, the spin is an integer multiple of $\hbar$. For a photon

$$S = \pm\hbar \quad \text{but not } S = 0. \tag{8.5}$$

The reason for the exclusion of $S = 0$ is that the photon has no rest mass. The two spin states correspond to right and left circular polarized light. In these cases, the spin is oriented parallel (RCP) and anti-parallel (LCP) to the wave vector. Linear polarized light corresponds to a superposition of $S = +\hbar$ and $S = -\hbar$. Photons may also have orbital angular momentum. This is the case when the electromagnetic field depends like $\exp(i\ell\phi)$ on the angle ϕ in the cylindrical coordinate system, for example in Laguerre-Gaussian modes. Then ℓ is the quantum number and $L = \ell\hbar$ is the orbital angular momentum.

The *position* of a photon is only known when a detectors reports an event. Before that, we only know a probability density to find a photon, which is proportional to the intensity of the optical wave at this position, i.e.

$$p(\boldsymbol{r}) \propto I(\boldsymbol{r}) \propto |u(\boldsymbol{r})|^2 \quad . \tag{8.6}$$

A beam splitter thus does not split a photon. It only splits the probability density to detect it at one output or the other.

Photon stream

For visible light, 1 W corresponds to about $3 \cdot 10^{18}$ photons per second. So in a beam of 1 nW, we have *on average* 3 photons per nanosecond. Fast photodetectors have a time resolution of about 1 ps. Such a detector would see 0.003 photons per picosecond, i.e. most of the time it would detect nothing and every now and then a photon. The intensity of a beam only determines the average photon rate. The properties of the light source determine the photon statistics, which can be described by the probability of finding n photons in a time interval T or by the intensity correlation function $g^{(2)}(\tau)$ (see

end of last chapter). In the following, we discuss first thermal light of low coherence time and coherent laser light, and later the emission of a single atom or dye molecule.

Photon stream of thermal light

Thermal light of a black body in a cavity is in thermal equilibrium. The probability p to find the energy E_n in the optical mode follows the Boltzmann distribution

$$p(E_n) \propto \exp\left(-\frac{E_n}{k_B T}\right) \tag{8.7}$$

at a temperature T. With quantized photons, the energy of a mode is determined by the number of photons in it

$$E_n = \left(n + \frac{1}{2}\right) h\nu \tag{8.8}$$

so that we get for the probability to find n photons in a mode

$$p(n) \propto \exp\left(-\frac{nh\nu}{k_B T}\right) = \left[\exp\left(-\frac{h\nu}{k_B T}\right)\right]^n \tag{8.9}$$

where the $1/2$ hides in the proportionally sign. We require that the $p(n)$ are normalized, i.e., their sum equals one. We define the mean photon number $\bar{n}$ by

$$\bar{n} = \frac{1}{\exp(h\nu/k_b T) - 1} \quad . \tag{8.10}$$

Everything together we get a geometric distribution or Bose-Einstein distribution

$$p(n) = \frac{1}{\bar{n}+1}\left(\frac{\bar{n}}{\bar{n}+1}\right)^n \quad . \tag{8.11}$$

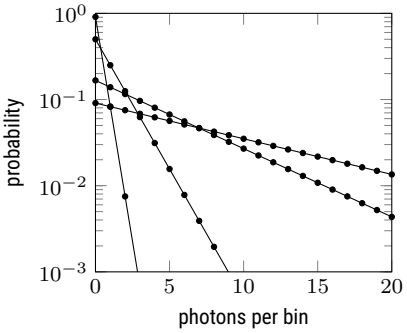

Figure 8.1: Thermal light. Probability $p(n)$ to find n photons in a time interval T when on average we have $\bar{n} = 0.1$, 1, 5, or 10 photons per time bin.

Photon stream of coherent laser light

A coherent light beam is what we have been assuming all along, until we came to the last chapter's discussion of (in)coherence. It is described by an electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ with angular frequency ω. If P is the power of the beam, the photon flux Φ (units: photons per second) can be calculated as

$$\Phi = \frac{P}{\hbar\omega} \quad . \tag{8.12}$$

We now imagine a segment of length L of such a beam. It contains

$$\bar{n} = \frac{\Phi L}{c} = \frac{P\,l}{\hbar\omega c} \tag{8.13}$$

photons, where we have assumed that L is so large that $\bar{n}$ is well defined, i.e., the granularity of the photons has been averaged out.

Now we divide the length L into a large number N of segments. We make N so large that the average number of photons per segment is far less than one. In each sub-segment we find either no photons or one photon. The probability of finding a photon is $p = \bar{n}/N$.

Let us reconstruct the whole segment from these sub-segments by taking N sub-segments. What is the probability of ending up with n photons? This is the probability of finding n subsegments with a photon (with probability p) and $N-n$ subsegments without a photon (with probability $1-p$). The order does not matter, so we end up with a binomial distribution

$$P(n) = \frac{N!}{n!(N-n)!}\, p^n (1-p)^{N-n} = \frac{N!}{n!(N-n)!}\, \left(\frac{\bar{n}}{N}\right)^n \left(1-\frac{\bar{n}}{N}\right)^{N-n} \quad . \tag{8.14}$$

We let $N \to \infty$ and obtain after some math the *Poisson distribution*

$$P(n) = \frac{\bar{n}^n}{n!} e^{-\bar{n}} \quad \text{for} \quad n = 0, 1, 2, \ldots \quad . \tag{8.15}$$

The same distribution also describes the number of events per time interval in a Geiger-Müller counter. The Poisson distribution has the important property that its variance is equal to its mean $\bar{n}$, or its standard deviation is $\sqrt{\bar{n}}$. For $\bar{n} \gtrsim 10$ the Poisson distribution approaches a normal distribution with the same mean and standard deviation. In the logarithmic plot, this appears as inverted parabola.

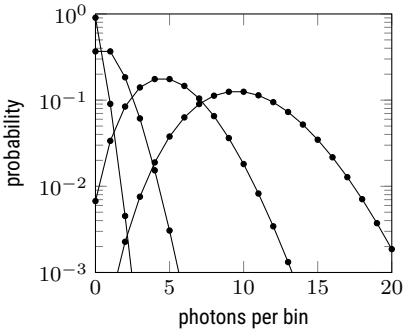

Figure 8.2: Coherent light. Same as above.

Photon stream of an atom

The fluorescence emission of a single atom or dye molecule differs in its photon statistics from both thermal light and coherent laser light. This was first observed experimentally by Kimble et al. in 1977 for sodium ions[3] . Let us examine the processes surrounding photon emission in a single atom or molecule described by quantum mechanics. The atom has a ground state and one or more excited states. Optical excitation brings the atom from the ground state to the excited state. This is a statistical process, so it takes some time after the excitation light source is turned on. The excited state decays with some probability to the ground state by emitting a photon. It may also decay without photon emission or into other states. However, after emission of the photon, the atom is certainly in the ground state. And by definition, a ground state cannot decay further, especially it cannot emit another photon. So it takes some time for the excitation light source to bring the atom back to the excited state where a second photon can be emitted. The average time between two emission events is thus given by the excitation rate and the emission rate, and under no circumstances can our quantum mechanical two-level system emit two photons simultaneously. This phenomenon is called *anti-bunching*.

[3] Kimble, Dagenais, and Mandel, 1977.

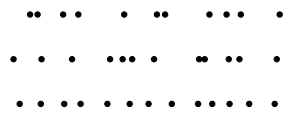

Figure 8.3: Sketch of photon detection events over time for thermal (top), coherent (mid), and anti-bunched (bottom) light.

Hanburry Brown-Twiss experiment for photons

We have introduced in the last chapter the intensity autocorrelation function or second-order correlation function that compared intensities $I(t)$ with those shifted in time

$$G^{(2)}(\tau) = \langle u^\star(t)\, u^\star(t+\tau)\, u(t) u(t+\tau) \rangle = \langle I(t)\, I(t+\tau) \rangle \quad . \tag{8.16}$$

We can replace the intensity $I(t)$ with the number $n(t)$ of photons detected in the time interval $(t, t+T)$, with the bin integration time T

$$G^{(2)}(\tau) = \langle n(t)\, n(t+\tau) \rangle \quad . \tag{8.17}$$

When T approaches the time resolution of the detector, then $n(t)$ is either zero or one.

To check for anti-bunching, we need to test whether two photons can be detected simultaneously. Typical photodetectors cannot distinguish between one or more photons at the same time. The experimental trick is the Hanburry Brown-Twiss experiment: one splits the photon stream from the atom into two streams and detects each stream with a separate detector. When two photons hit the beam splitter, in some cases they would separate into the two beams. In other cases, they would stay in the same beam, but that does no harm. So we look for coincidences, i.e. both detectors clicking at the same time. This is what $G^{(2)}(0)$ shows. Anti-bunching leads to a dip in the autocorrelation function around $\tau = 0$. The slope of the dip depends on the sum of the excitation and emission rates.

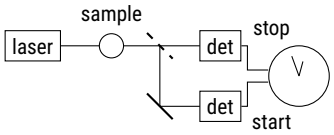

Figure 8.4: Hanbury Brown–Twiss experiment. The time interval τ between two photons is determined.

We can determine $I(t)$ for light by counting photons within a short interval T and then writing $g^{(2)}(\tau)$ with the counting rate $n(t)$. However, to detect antibunching, T must be very small (about 100 ps). At the same time, averaging requires a long total time, i.e. a large amount of data. A more data-efficient approximation is not register the full $n(t)$ trace to calculate $G(\tau)$ (equivalent to all pairs of photons), but to register only the time between *successive* pairs of photons, which we call $C(\tau)$. Finding two photons at distance τ can happen with a varying number of other photons in between. No photon in between is described by $C(\tau)$. A single photon in between can occur at any time τ' with $0 < \tau' < \tau$. We could integrate over these possibilities. So $C(\tau)$ and $G(\tau)$ are related:

$$G(\tau) = C(\tau) + \int_0^\tau C(\tau')C(\tau - \tau')d\tau' + \dots \quad . \tag{8.18}$$

Further double, triple, etc. integrals would then describe two, three, etc. photons in between. If the photon rate is small enough or the time of interest τ is short enough, we can assume $G(\tau) \approx C(\tau)$. So effectively we do not measure $n(t)$ for two detectors in the Hanburry Brown–Twiss experiment, but we start a clock by one detector and stop it via the other, so that we measure $C(\tau)$.

The important point is that anti-bunching only is visible when a single atom or molecule is the source of the photon stream. Averaging over many emitters removes the effect, because sooner or later the emission from one emitter would surely coincide with the emission from another emitter, and the dip would disappear.

Anti-Bunching in semiconductor quantum dots

An emitter that exhibits antibunching, i.e. emits only a single photon at a time, is called a single-photon emitter or single-photon source. Such light sources are needed for quantum key distribution, as discussed below. Single atoms or dye molecules are single photon emitters. A technologically interesting alternative are quantum dots. These are droplets of a low bandgap semiconductor embedded in a matrix of a higher bandgap semiconductor. In essence, a 3D particle-in-a-box is formed for the electrons and holes. As the size of the droplet approaches the de Broglie wavelength of the electrons,

the band structure disappears and the electron states become quantized and discrete, as in an atom or two-level system in general.

For a two-level system, $C(\tau)$ is easy to determine. Immediately after the first photon you are in the ground state with absolute certainty. The excited state is reached with the excitation rate W_P, from there back to the ground state with the rate Γ of the spontaneous emission. The characteristic time t_d is the reciprocal of the total rate for one cycle.

$$t_d = \frac{1}{W_P + \Gamma} \quad \text{and so} \quad g^{(2)}(\tau) \approx 1 - ae^{-\tau/t_d} \quad . \tag{8.19}$$

The amplitude a is $a = 1$ in the ideal case. In reality, dark noise and background photons cause $a < 1$. However, the case $a > 0.5$ can only be generated by a single photon source or a single two-level system.

Such an experiment is shown in the figure 8.5 for a GaAs quantum dot. The optical excitation here was via the surrounding semiconductor, not directly via the exciton. This leads to the 'overshoots' with $g > 1$, which are taken into account in the model.

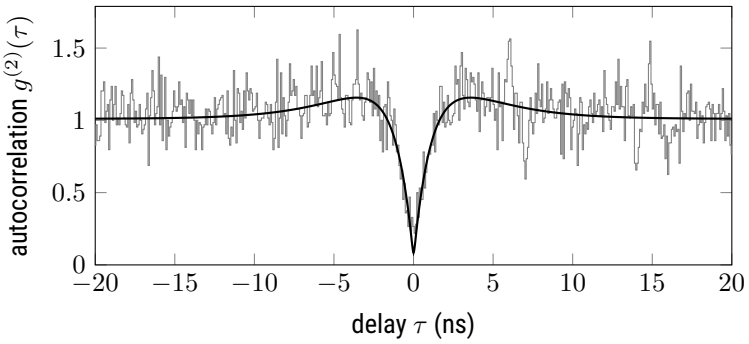

Figure 8.5: Anti-bunching in a GaAs quantum dot. Data from Wu et al., 2017.

Quantum Key Distribution

An increasingly important technological application of single-photon sources is the quantum-mechanically secure transmission of an encryption key. There are several ways to encrypt messages. For example, it is possible to exploit the fact that a large number can only be broken down into its prime factors with great effort, but the reverse is easy. However, the complexity of breaking the encryption depends on the available technologies and may be difficult to predict for the future. An encryption that can never be broken is the exclusive-or relation (XOR) of the source text with a key that is the same length as the message and is never used again. This key is called a one-time pad. The recipient needs the same key, performs another XOR with the encrypted message, and receives the plaintext. However, this only transforms the problem of encryption into the problem of transmitting the key. For example, you could distribute disks with the (very long) key in advance using a trusted messenger.

This is where quantum key distribution comes in. The key is distributed in the form of individual photons so that both the sender (Alice) and the receiver (Bob) can later use the same key. The encrypted message can then be transmitted over a normal channel.

I describe here the BB84 protocol[4] by Charles Bennett and Gilles Brassard. We use four linear polarization states of light: horizontal ($|h\rangle$) and vertical ($|v\rangle$), diagonal ($|+\rangle$) and anti-diagonal ($|-\rangle$). They are not independent of each other, but $|h\rangle$ and $|v\rangle$ as well as $|+\rangle$ and $|-\rangle$ form a basis.

[4] see also chapter 11.8.2 in Gerry and Knight, 2005, original in Bennett and Brassard, 2014

Alice sends photons to Bob and randomly chooses one of the four states for each photon. Bob is unaware of this, chooses one of the two bases at random, and measures the polarization state of the incoming photon in that basis, for example using a polarization beam splitter and two photodetectors. The bases could be selected using an appropriately rotated waveplate. When Alice and Bob have finished transmitting and measuring, Alice transmits the basis she has chosen over an open channel, not the polarization state. Bob compares this to his list and also tells Alice via an open channel if they have chosen the same basis. Both delete the other photons. But now they both have a list of polarization states that Bob measured in the same basis that Alice transmitted in. The polarization state is now a bit of a key.

Individual photons are essential. Otherwise, the eavesdropping Eve could intercept part of the beam and measure it herself, possibly even in both bases at the same time. Only if it is a single photon is it certain that the measurement destroys the state and it cannot be measured a second time.

Ladder operators

Let us look at the quantum mechanics of a photon. We have n photons in our mode and an energy

$$E_n = \left(n + \frac{1}{2}\right) \hbar\omega \tag{8.20}$$

as in a harmonic oscillator. We have a ladder of equidistant states, bounded at the bottom by E_0, but open at the top, since n can be arbitrarily large. In quantum mechanics it is convenient to use ladder operators for the creation ($\hat{a}^\dagger$) and annihilation ($\hat{a}$) of an energy quantum, i.e.

$$\hat{a}^\dagger |n\rangle = \sqrt{n+1}\ |n+1\rangle \quad \text{and} \quad \hat{a}\,|n\rangle = \sqrt{n}\ |n-1\rangle \quad . \tag{8.21}$$

Useful properties are

$$\hat{a}\,|0\rangle = |0\rangle \quad \text{and} \quad \hat{a}^\dagger \hat{a}\,|n\rangle = n\,|n\rangle \quad . \tag{8.22}$$

The electrical field of a single optical mode in a cavity can then be written[5,6] using the vacuum field amplitude (eq. 8.3)

[5] Gerry and Knight, 2005, chap. 2.1 and 2.4.
[6] Rand, 2016, chap. 6.1.

$$\hat{\boldsymbol{E}}(z,t) = \hat{\boldsymbol{x}}\, E_{vac} \left(\hat{a}\, e^{i(kz-\omega t)} + \hat{a}^\dagger\, e^{-i(kz-\omega t)}\right) \quad , \tag{8.23}$$

where $\hat{\boldsymbol{x}}$ is a unit vector defining the direction of polarization. The field in the cavity is a superposition of right and left propagating plane waves. The amplitude of each is related to the operators $\hat{a}$ and $\hat{a}^\dagger$, respectively.

Beam splitter

From the point of view of classical electrodynamics, a beam splitter seems to be a rather trivial device. But quantum optics will surprise you. Let us recall

what electromagnetic waves do at a beam splitter: We shine in with a field $\boldsymbol{E}_1$. Part of it is reflected (coefficient r), part is transmitted (coefficient t). The two output fields are

$$\boldsymbol{E}_2 = r\,\boldsymbol{E}_1 \quad \text{and} \quad \boldsymbol{E}_3 = t\,\boldsymbol{E}_1 \tag{8.24}$$

where conservation of energy requires that

$$|r|^2 + |t|^2 = 1 \quad . \tag{8.25}$$

In quantum optics we describe the modes for the input, reflected and transmitted beam by three ladder operators $\hat{a}_i$ with $i = 1, 2, 3$. Operators on different modes commute, i.e. have a commutator of zero. Only the pair of operators on the same mode have a commutator of one. As equations

$$\left[\hat{a}_i, \hat{a}_j^\dagger\right] = \delta_{ij} \tag{8.26}$$

$$\left[\hat{a}_i, \hat{a}_j\right] = 0 \tag{8.27}$$

$$\left[\hat{a}_i^\dagger, \hat{a}_j^\dagger\right] = 0 \quad . \tag{8.28}$$

Writing eq. 8.24 in the formalism of quantum optics gives

$$\hat{a}_2 = r\,\hat{a}_1 \quad \text{and} \quad \hat{a}_3 = t\,\hat{a}_1 \quad . \tag{8.29}$$

The problem is that these definitions do not satisfy the commutators. We can check several combinations and all of them give the required result of zero only if we set either $r = 0$ or $t = 0$, i.e. if we remove the beam splitter

$$\left[\hat{a}_2, \hat{a}_2^\dagger\right] = |r|^2 \left[\hat{a}_1, \hat{a}_1^\dagger\right] = |r|^2 \neq 1 \tag{8.30}$$

$$\left[\hat{a}_3, \hat{a}_3^\dagger\right] = |t|^2 \left[\hat{a}_1, \hat{a}_1^\dagger\right] = |t|^2 \neq 1 \tag{8.31}$$

$$\left[\hat{a}_2, \hat{a}_3^\dagger\right] = rt^\star \left[\hat{a}_1, \hat{a}_1^\dagger\right] = rt^\star \neq 0 \quad . \tag{8.32}$$

So in quantum optics there seems to be something more to the beam splitter than in classical electrodynamics. The point is the fourth side of the beam splitting cube (which we label $i = 0$ for convenience). Classically, we assumed that no light would enter here, so we set $\boldsymbol{E}_0 = 0$ and ignored it in our calculations. But in quantum optics we have to take into account that $|\boldsymbol{E}_0| = E_{vac}$. The vacuum fluctuations are shining into the empty port of the beam splitter.

We expand eq. 8.29 to take the fourth port into account. In matrix form

$$\begin{pmatrix} \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \begin{pmatrix} t' & r \\ r' & t \end{pmatrix} \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \end{pmatrix} \quad . \tag{8.33}$$

This definition satisfies all commutator relations when the reflection and transmission coefficients satisfy energy conservation ($|r|^2 + |t|^2 = 1$), reciprocity ($|r'| = |r|$, $|t'| = |t|$), and a phase condition as discussed with the Michelson interferometer. For a single dielectric layer, the transmitted and reflected beams are 90 degrees out of phase. A 50:50 beamsplitter is thus

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 & i \\ i & 1 \end{pmatrix} \quad . \tag{8.34}$$

A single photon at a 50:50 beam splitter

As an example, let us examine what happens to a single photon in a 50:50 beam splitter. I label the eigenstates with n photons in beam i as $|n\rangle_i$. If nothing (a vacuum) shines into a beam splitter, nothing comes out, i.e. trivially

$$|0\rangle_0\ |0\rangle_1 \xrightarrow{BS} |0\rangle_2\ |0\rangle_3 \quad . \tag{8.35}$$

A single photon at port 1 is written by the creation operator acting on the vacuum state

$$|0\rangle_0\ |1\rangle_1 = \hat{a}_1^\dagger\ |0\rangle_0\ |0\rangle_1 \tag{8.36}$$

and $\hat{a}_1^\dagger$ can be written with the port 2 and 3 operators of a 50:50 beam splitter

$$\hat{a}_1^\dagger = \frac{1}{\sqrt{2}}\left(i\hat{a}_2^\dagger + \hat{a}_3^\dagger\right) \tag{8.37}$$

so that

$$|0\rangle_0\ |1\rangle_1 \xrightarrow{BS} \frac{1}{\sqrt{2}}\left(i\,|1\rangle_2\ |0\rangle_3 + |0\rangle_2\ |1\rangle_3\right) \quad . \tag{8.38}$$

A single photon entering port 1 together with a vacuum at port 0 leaves either port 2 or port 3, but not both! The final state is a *entangled state*[7], since it cannot be written as a product of states of the form $|n\rangle_2\,|m\rangle_3$.

[7] dt: verschränkter Zustand

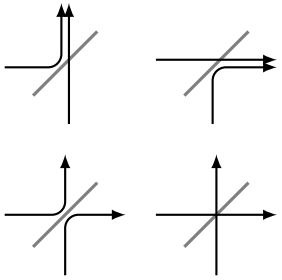

Figure 8.6: Two identical photons entering a beam splitter leave together (top), because the two processes below cannot be distinguished and interfere destructively.

Hong-Ou-Mandel experiment: two photons on a beam splitter

What we can do with one photon, we can do with two! Let us shine a photon on each input port of the beam splitter, i.e. start with the state $|1\rangle_0\ |1\rangle_1$. We can create it from vacuum with two creation operators

$$|1\rangle_0\ |1\rangle_1 = \hat{a}_1^\dagger\hat{a}_0^\dagger\,|0\rangle_0\ |0\rangle_1 \tag{8.39}$$

and write additionally to $\hat{a}_1^\dagger$ as above also $\hat{a}_0^\dagger$ in terms of the port 2 and 3 operators

$$\hat{a}_0^\dagger = \frac{1}{\sqrt{2}}\left(\hat{a}_2^\dagger + i\hat{a}_3^\dagger\right) \tag{8.40}$$

where only the position of the i has swapped. Everything together is

$$|1\rangle_0\ |1\rangle_1 \xrightarrow{BS} \frac{1}{2}\left(i\hat{a}_2^\dagger + \hat{a}_3^\dagger\right)\left(\hat{a}_2^\dagger + i\hat{a}_3^\dagger\right)|0\rangle_2\ |0\rangle_3 \tag{8.41}$$

$$= \frac{i}{2}\left(\hat{a}_2^\dagger\hat{a}_2^\dagger + \hat{a}_3^\dagger\hat{a}_3^\dagger\right)|0\rangle_2\ |0\rangle_3 \tag{8.42}$$

$$= \frac{i}{\sqrt{2}}\left(|2\rangle_2\ |0\rangle_3 + |0\rangle_2\ |2\rangle_3\right) \quad . \tag{8.43}$$

When multiplying out the two brackets, the cross terms $\hat{a}_2^\dagger\hat{a}_3^\dagger$ cancel out and only the symmetric terms remain. Two photons entering at different ports of the beam splitter will exit through the same port! Either both at port 2 or both at port 3. This is a quantum interference effect. The case where both photons are transmitted interferes destructively with the case where both photons are reflected. Since the photons are indistinguishable, we cannot separate the two cases and must add the probability amplitudes before taking the square modulus to get the power. This was demonstrated in 1987 by Hong, Ou, and Mandel[8].

[8] Hong, Ou, and L. Mandel, 1987.

The word 'indistinguishable' is important here. The two photons must agree in every conceivable way. They must arrive at the beam splitter at the same time. They must have the same spectrum/color, the same polarization state, and the same mode/wavefront or beam profile. As soon as one photon is a little different from the other, we could distinguish the 'both reflected' case from the 'both transmitted' case.

The Hong-Ou-Mandel experiment shows that photons are bosons. Bosons want to be in the same state. The two photons want to leave the beam splitter together. This is in stark contrast to electrons, which are fermions. Fermions want to be in different states. A beam splitter for electron waves would favor the $|1\rangle_2 |1\rangle_3$ result over a $|2\rangle_2 |0\rangle_3$ result.

Down-conversion source of entangled photons

Experiments in quantum optics require a single-photon source or a source of single entangled photon pairs. A single-photon source such as an atom, a molecule or a quantum dot can be studied with current technology even in an undergraduate lab. But if it should be the source for the following actual experiment, it is a bit too complicated. A technically simpler device is a source of entangled photon pairs (but not 'single' pairs). The process is called down-conversion.

In nonlinear optics, there are processes that split a photon into two. The energy of the incoming photon is split into two parts, leaving as two photons. This is called down conversion. There is also the reverse process where the energy of two photons is combined into one. This is called sum frequency generation, since the optical frequency of the outgoing photon is the sum of the frequencies of the incoming photons.

Both processes require that the polarization P induced by a field E is to some extent quadratically dependent on the field. In chapter 4 we wrote

$$P(t) = \chi \epsilon_0 E(t) \quad . \tag{8.44}$$

Now we see this as first order of a Taylor expansion and write

$$P(t) = \epsilon_0 \left(\chi^{(1)} E(t) + \chi^{(2)} E^2(t) + \chi^{(3)} E^3(t) + \dots \right) \quad . \tag{8.45}$$

The expression with $\chi^{(2)}$ is important for us. If

$$E(t) = E_1 e^{i\omega_1 t} + E_2 e^{i\omega_2 t} \tag{8.46}$$

then $P(t)$ will contain frequency components at ω_1 and ω_2 due to the $\chi^{(1)}$ term and at 0, $2\omega_1$, $2\omega_2$ and $\omega_1 \pm \omega_2$ due to the $\chi^{(2)}$ term. The latter is sum and difference frequency generation.

In a down-conversion source, a blue laser beam hits a material with a rather high value of $\chi^{(2)}$: BBO (Beta Barium Borate).The energy of the 400 nm photon is split into two photons of 800 nm wavelength. This conserves energy. Conservation of momentum would be easy if the index of refraction at both wavelengths were identical. However, this is not the case, as we saw in chapter 4. This is where birefringence comes in. The ordinary and the extraordinary ray have slightly different refractive indices. In the chosen material, this difference helps to compensate for the difference due to

dispersion. As a result, one of the 800 nm photons leaves as the ordinary beam and the other as the extraordinary beam. Both photons are polarized orthogonal to each other. By carefully aligning the source, one can generate polarization entangled photons, i.e. a state in which one does not know how each photon is polarized, except that one knows they are orthogonal to each other.

References

Bennett, Charles H. and Gilles Brassard (2014). "Quantum cryptography: Public key distribution and coin tossing [reprint]". In: *Theoretical Computer Science* 560, pp. 7–11.

Fox, Mark (2007). *Quantum optics*. Oxford University Press.

Gerry, Christopher C. and Peter L. Knight (2005). *Introductory quantum optics*. Cambridge Univ. Press.

Hong, C. K., Z. Y. Ou, and L. Mandel (1987). "Measurement of subpicosecond time intervals between two photons by interference". In: *Phys. Rev. Lett.* 59 (18), pp. 2044–2046.

Kimble, H J, M Dagenais, and L Mandel (1977). "Photon Antibunching in Resonance Fluorescence". In: *Phys. Rev. Lett.* 39.11, pp. 691–695.

Rand, Stephen C. (2016). *Lectures on light. nonlinear and quantum optics using the density matrix*. Second edition. Oxford University Press.

Wu, Xiaofei et al. (2017). "On-Chip Single-Plasmon Nanocircuit Driven by a Self-Assembled Quantum Dot". In: *Nano Letters* 17.7, pp. 4291–4296.

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)

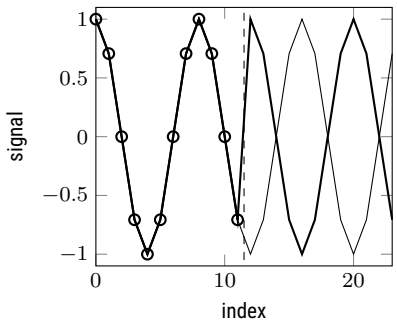

Figure B.1: Clipping a cosine after 1.5 periods

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 .

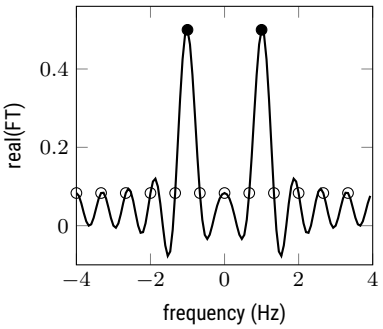

Figure B.2: Zeropadding (line) approaches better the real spectrum (filled symbols) compared to the clipped FT (open symbols).

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.

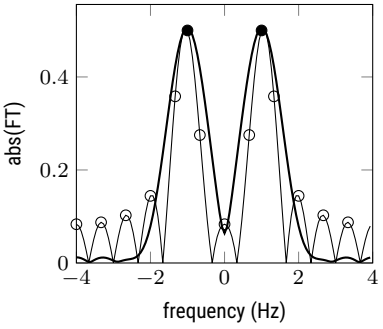

Figure B.3: Zeropadding after windowing (thick) removes the fringes of the un-windowed data (thin) and approaches the true spectrum (solid symbols).

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

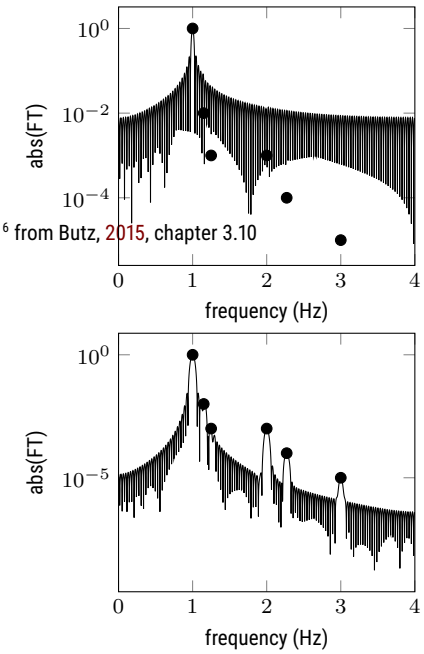

Figure B.4: Without windowing (top), only the main signal component is recovered. A Hanning window (bottom) allows to find even signals 10^{-5} below the main component.

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

Bennett, Charles H. and Gilles Brassard (2014). “Quantum cryptography: Public key distribution and coin tossing [reprint]”. In: *Theoretical Computer Science* 560, pp. 7–11.

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.

CVI Melles Griot (2009). *Interference Filter Coatings - CVI Melles Griot Technical Guide, Vol 2, Issue 2*. (visited on 12/19/2023).

Fox, Mark (2007). *Quantum optics*. Oxford University Press.

Gerry, Christopher C. and Peter L. Knight (2005). *Introductory quantum optics*. Cambridge Univ. Press.

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

Hanbury Brown, R and R Q Twiss (1956). “A Test of a new type of stellar interferometer on Sirius”. In: *Nature* 178.4541, pp. 1046–1048.

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.

Hong, C. K., Z. Y. Ou, and L. Mandel (1987). “Measurement of subpicosecond time intervals between two photons by interference”. In: *Phys. Rev. Lett.* 59 (18), pp. 2044–2046.

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

Kimble, H J, M Dagenais, and L Mandel (1977). “Photon Antibunching in Resonance Fluorescence”. In: *Phys. Rev. Lett.* 39.11, pp. 691–695.

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

Macleod, H. Angus (2001). *Thin film optical filters*. 3. ed. Bristol [u.a.]: Inst. of Physics Publ.

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

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

Pedrotti, Frank L. et al. (2008). *Optik für Ingenieure*. 4., bearb. Aufl. Berlin [u.a.]: Springer.
Rand, Stephen C. (2016). *Lectures on light. nonlinear and quantum optics using the density matrix*. Second edition. Oxford University Press.
Saleh, Bahaa E. A. and Malvin C. Teich (1991). *Fundamentals of photonics*. New York, NY [u.a.]: Wiley.
Wu, Xiaofei et al. (2017). “On-Chip Single-Plasmon Nanocircuit Driven by a Self-Assembled Quantum Dot”. In: *Nano Letters* 17.7, pp. 4291–4296.
Yariv, Amnon (1989). *Quantum electronics*. 3. ed. New York: Wiley.
Yeh, Pochi (2005). *Optical waves in layered media*. Hoboken, NJ: Wiley-Interscience.