OPTICS

1. INTRODUCTION

Optics is the sub-field of physics that studies light and the interaction of light with matter. Optical effects provide the basis for several technologies relevant in today's society. The data traffic throughout the internet is transmitted through optical fibers around the world. CD-, DVD, and Blu-Ray discs store data using optical methods. Lasers are used for materials processing, ranging from welding of automobiles to eye surgery. Image capture, projection, and displays also rely on optical technologies. Optical spectroscopy is used to identify compounds in chemistry and environmental monitoring.

These lecture notes were written by **Prof. Martti Kauranen** for the optics course he taught (also to me) at the Tampere University of Technology (TUT) for many years. The course focuses on the most fundamental optical effects. Light is treated as an electromagnetic wave in order to understand the basic properties of its propagation as well as other important effects, such as interference and diffraction, which arise from the superposition principle of waves. Methods of geometrical optics, where light is treated as rays travelling in space, are used to describe some of the most traditional optical instruments. For the current implementation (2020) I have only slightly changed the material. Mostly, the changes have been done to help navigating the material using mobile devices.

The material is mostly based on the book Optics by E. Hecht (4th edition, Addison-Wesley, 2002). However, the book is naturally much more extensive than the present notes with regard to its content. On the other hand, the book is quite elementary in its mathematical approach in the sense that it uses quite extensively trigonometric functions to represent harmonic waves. Certain aspects are much more straightforward to discuss using the complex notation for harmonic waves. While the complex notation is mathematically more advanced, it is appropriate for the target students at TAU. In addition, the gain for any future work is enormous. The notes are therefore based on the complex notation. The lecture notes differ from the book also in the sense that an attempt has been made to follow the standard conventions within the optics community.

These notes have been improved by the constant feedback from the teaching assistants of the related courses in 2014 – 2019: Eero Halonen, Kalle Koskinen, Mikko Närhi, Kim Patokoski, Lauri Palmolahti, Jussi Rossi, Juha Tiihonen, Heidi Tuorila, Ossi Tuominen, Jan Viljanen and Matti Virkki.

– Mikko J. Huttunen

2. DESCRIPTIONS OF LIGHT

2.1 Electromagnetic waves

Light is electromagnetic radiation. In terms of classical physics, light can therefore be described by starting from the laws that govern electricity and magnetism. In 1865, Maxwell collected the existing information about these two interconnected fields and found out that they can be expressed in the form of a set of equations. By proper manipulation, Maxwell's equations can be turned into a ***wave equation*** for both the electric field **E** and the magnetic induction field **B**. The solutions of these equations describe perturbations that move in space as a function of time, i.e., wave motion.

A very specific solution to the wave equation is the ***harmonic wave***. Assuming propagation along z axis at speed v and scalar description, a harmonic wave can be written as

$$\Psi(z,t) = A\sin k(z - vt)\,. \tag{2.1.1}$$

Here, A is the ***amplitude***, which tells how strong the oscillations are. The argument of the sine function $k(z - vt)$ is known as the ***phase*** of the harmonic function, and the parameter k tells how rapidly the oscillations occur. The value of the phase thus determines which value the wave itself assumes at a given location at a given time. Note that, instead of the sine function, we could equally well use the cosine function.

For time $t = 0$, the harmonic function depends only on the z coordinate

$$\Psi(z,0) = A\sin kz\,, \tag{2.1.2}$$

which is easy to draw as a figure (Fig. 2.1). Due to the periodicity of the sine function, the wave repeats itself at spatial locations that are separated from each other by one ***wavelength*** λ (Fig. 2.1), a key quantity describing harmonic waves.

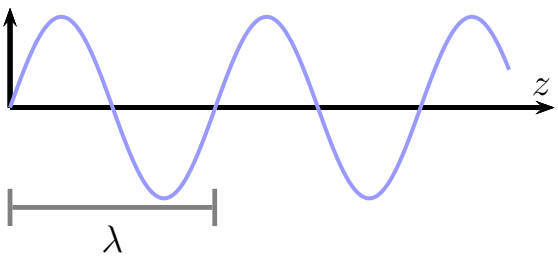

Fig. 2.1: Harmonic wave with wavelength λ.

The wave also repeats itself in time and this interval is known as the ***period*** τ of the wave. In the time domain, it is also common to talk about the ***frequency*** ν of the wave, which tells how many oscillations occur in a given time. Thus, $\nu = 1/\tau$. It is also common to use the quantity ***angular frequency*** $\omega = 2\pi\nu$, which leads to the very common form for the harmonic wave as

$$\Psi(z,t) = A\sin(kz - \omega t)\,. \tag{2.1.3}$$

We will have a more detailed discussion of these various quantities in later chapters, where wave motion and electromagnetic radiation are discussed.

Due to the wave character of light, it exhibits all the effects associated with all kinds of waves, e.g., ***interference*** and ***diffraction***, which will also be discussed in later chapters. Diffraction gives rise to important limitations to optical effects. In particular, it limits the resolutions of optical instruments to the order of one wavelength. This means that imaging optical instruments cannot resolve much finer details of objects than one wavelength.

The whole electromagnetic spectrum covers a very wide range of wavelengths and frequencies, starting from radio waves (long wavelengths) and extending to gamma rays (short wavelengths). In the strictest sense, optics deals only with the spectral regime that can be seen by the human eye, wavelengths from 390 nm (violet) to 780 nm (red). This range is thus only a very small fraction of the total electromagnetic spectrum. Techniques very similar to optical ones, however, can also be used in near-ultraviolet (approximately in the range of 200 – 400 nm) and in the near- and mid-infrared (approximately between 780 nm and 6 μm).

2.2 Rays

Geometrical optics is the approach where the wave character of light is neglected. In essence, it is based on the assumption that the wavelength is zero, so that diffraction does not limit the operation of an optical instrument. This is often sufficient to analyze at least the basic operation of optical instruments. In this approach, the propagation of light is described by rays that correspond to the direction of electromagnetic energy flow.

When the ray approach is taken to its simplest interpretation, it is sufficient to know that when light encounters an interface between two materials with different optical properties, part of the light is reflected and part refracted (see Fig. 2.2). For this purpose, the materials are described by their ***index of refraction***. At the interface, the direction of a ray is characterized by its angle with respect to the interface normal $\hat{\mathbf{n}}$. Regarding the effects of ***reflection*** and ***refraction***, it is known that the incident, reflected, and refracted rays stay in the same plane, known as the plane of incidence. In addition, the reflec-

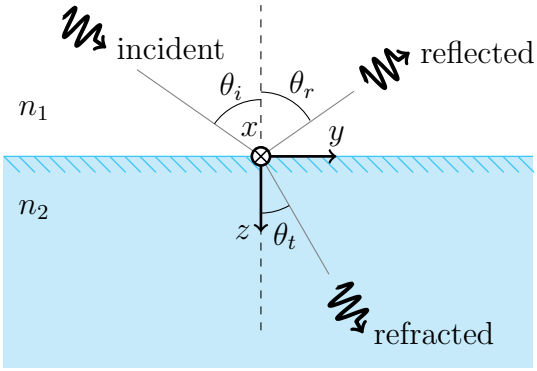

Fig. 2.2: Reflection and refraction at an interface between two materials.

tion law states that the angles of incidence θ_i and reflection θ_r are equal

$$\theta_i = \theta_r \,. \tag{2.2.4}$$

In addition, the angles of incidence and refraction are related by the refraction law, also known as ***Snell's law***

$$n_i \sin\theta_i = n_t \sin\theta_t \,, \tag{2.2.5}$$

where n_i and n_r are the indices of refraction of the two materials. These laws will be derived in a later chapter using electromagnetic theory.

2.3 Particles

When quantum-mechanical effects become important, it is often necessary to treat light as particles or quanta. In particular, when considering interaction between radiation and matter, the radiation is emitted and absorbed as quanta. The particles associated with radiation are known as ***photons***. It is known that the energy of a photon is proportional to its frequency

$$E = h\nu = \hbar\omega \,, \tag{2.3.6}$$

where $h = 6.6626 \times 10^{-34}$ Js is the Planck's constant and $\hbar = h/2\pi$ is the reduced Planck's constant.

Photons also have other properties associated with particles. They carry momentum and have spin. However, they have no rest mass nor charge. We will not discuss the quantum mechanical properties of light and light-matter interactions much during this course.

3. WAVE MOTION

3.1 Moving perturbation

We start by considering a function $f(z)$ that depends only on the z coordinate and that is centered in some sense about the location $z = 0$ (Fig. 3.1a). In order to center the same function at location $z = z_0$, we need to shift the argument so that the function becomes of the form $f(z - z_0)$ (Fig. 3.1b). Finally, we can get this function moving in the positive z direction at velocity v by taking the shifted location to be time-dependent $z_0 = vt$, where $v > 0$.

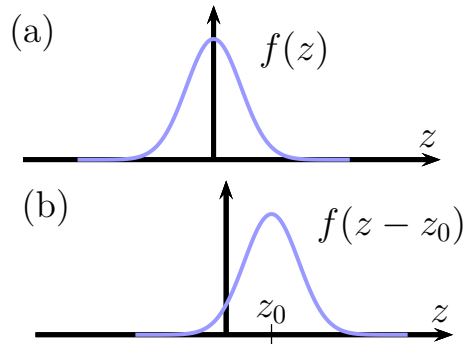

Fig. 3.1: Function f centered (a) at $z = 0$ and (b) at $z = z_0$.

We may therefore conclude that an arbitrary function of space and time $\Psi(z, t)$ that is of the form

$$\Psi(z, t) = f(z - vt)\,, \tag{3.1.1}$$

represents a perturbation that moves at constant velocity v along the z axis and maintains its shape (functional form).

3.2 Wave equation

We know from experience that physical phenomena are often described by differential equations that connect the relevant quantities to each other. We next seek possible differential equations that would lead to solutions of the form of Eq. (3.1.1). However, we want to be a bit more general and look for equations that allow motion in both positive and negative z directions, because we expect the same physical laws to govern both cases. Our possible solutions are therefore of the form

$$\Psi = f(z \pm vt) = f(z')\,, \tag{3.2.2}$$

where we have defined a dummy variable $z' = z \pm vt$. We next differentiate Eq. (3.2.2) with respect to space (z) and time (t) in order to obtain

$$\frac{\partial \Psi}{\partial z} = \frac{\partial f}{\partial z'}\frac{\partial z'}{\partial z} = \frac{\partial f}{\partial z'}\,, \tag{3.2.3}$$

$$\frac{\partial \Psi}{\partial t} = \frac{\partial f}{\partial z'}\frac{\partial z'}{\partial t} = \pm v\frac{\partial f}{\partial z'}\,. \tag{3.2.4}$$

By combining the above equations, we may conclude that

$$\frac{\partial \Psi}{\partial t} = \pm v\frac{\partial \Psi}{\partial z}\,. \tag{3.2.5}$$

This equation depends on the direction of propagation and is therefore not acceptable for us. It would indeed be a strange physical phenomenon that needs to be described in different ways depending on the direction of propagation.

In order to overcome this problem, we calculate the second partial derivatives

$$\frac{\partial^2\Psi}{\partial z^2} = \frac{\partial^2 f}{\partial z'^2}\frac{\partial z'}{\partial z} = \frac{\partial^2 f}{\partial z'^2}\,, \tag{3.2.6}$$

$$\frac{\partial^2\Psi}{\partial t^2} = \pm v\frac{\partial^2 f}{\partial z'^2}\frac{\partial z'}{\partial t} = (\pm v)^2\frac{\partial^2 f}{\partial z'^2}\,. \tag{3.2.7}$$

From here, again combining the above two equations we can conclude that a second-order differential equation of the form

$$\frac{\partial^2\Psi}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2}\,, \tag{3.2.8}$$

allows solutions of the form $\Psi(z,t) = f(z \pm vt)$. This equation is known as the ***wave equation***. More specifically, this form of the wave equation is one-dimensional, i.e., it leads only to solutions that propagate along the z axis. In general, solutions to this equation are known as waves. Note that, in order for a function of the form $\Psi(z,t) = f(z \pm vt)$ to be a wave, it must be differentiable twice with respect to space and time.

3.3 Harmonic waves

We next consider harmonic waves, which are by far the most important special cases of waves. For such cases, the wave is represented by a harmonic function, e.g., sine or cosine function, which oscillates periodically. Specifically, we consider a wave of the form

$$\Psi(z,t) = f(z - vt) = A\sin k(z - vt)\,. \tag{3.3.9}$$

Here, A is the ***amplitude*** describing us how strong the wave oscillations are. The argument of the sine function $k(z - vt)$ is known as the ***phase*** of the harmonic function, and the parameter k tells how rapidly the oscillations occur. The value of the phase thus determines which value the wave itself assumes at a given location at a given time.

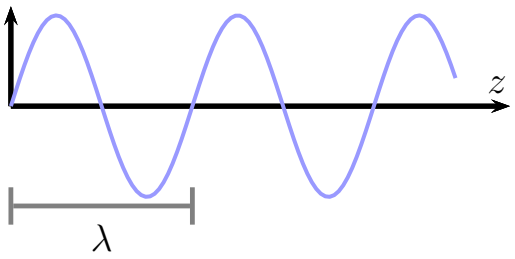

Fig. 3.2: Harmonic wave with wavelength λ.

For time $t = 0$, the harmonic function depends only on the z coordinate

$$\Psi(z, 0) = A \sin kz \,, \tag{3.3.10}$$

which is easily visualized (see Fig. 3.2).

We next define certain basic quantities that are important for harmonic waves. First, the sine function is known to be periodic, i.e., it repeats itself at spatial locations that are separated from each other by one wavelength λ (Fig. 3.2). The requirement for this is

$$\Psi(z, t) = \Psi(z \pm \lambda, t) \,, \tag{3.3.11}$$

$$A \sin k(z - vt) = A \sin k(z \pm \lambda - vt) \,. \tag{3.3.12}$$

This is true provided that $|k\lambda| = 2\pi$. By choosing the parameter k to be positive, we find that this parameter is related to wavelength by

$$k = \frac{2\pi}{\lambda} \,. \tag{3.3.13}$$

This parameter is known as the ***propagation constant*** or ***wave number*** of the harmonic wave, i.e., it determines how rapidly the wave oscillates in space.

Similarly, the wave repeats itself also in time and this interval is known as the ***period*** τ of the wave, resulting in the requirement

$$\Psi(z, t) = \Psi(z, t \pm \tau) \,. \tag{3.3.14}$$

From here we obtain a condition $|kv\tau| = 2\pi$. By choosing the period τ to be positive and using Eq. (3.3.13), we find the period to be

$$\tau = \lambda / v \,. \tag{3.3.15}$$

In the time domain, it is also common to talk about the ***frequency*** ν of the wave, which tells how many oscillations occur in a given unit of time. Thus, $\nu = 1/\tau = v/\lambda$, which is usually expressed as

$$v = \nu\lambda \,, \tag{3.3.16}$$

i.e., the velocity of the wave v is obtained as the product of the wavelength λ and the frequency ν. Note here that the symbols for velocity v (letter "v") and frequency ν (Greek symbol "nu") look similar. We will, however, use these standard symbols because they will rarely be used together later on.

It is also common to use the quantities ***angular frequency***

$$\omega = 2\pi\nu\,, \tag{3.3.17}$$

and ***spatial frequency***

$$\kappa = 1/\lambda\,, \tag{3.3.18}$$

which is the spatial analogue of frequency ν. Unfortunately, depending on the subfield of optical sciences, this quantity is also sometimes called as wave number.[1]

With all these definitions, the phase of the harmonic wave can be written as

$$k(z - vt) = kz - kvt = kz - \frac{2\pi vt}{\lambda} = kz - \frac{2\pi\nu\lambda t}{\lambda} = kz - \omega t\,, \tag{3.3.19}$$

which leads to a very common way of expressing the harmonic wave in the form

$$\Psi(z,t) = A\sin(kz - \omega t)\,. \tag{3.3.20}$$

The form given by Eq. (3.3.20) requires that the origins of space and time have been chosen just right, and the resulting wave for time $t = 0$ is shown in Fig. 3.3a. In practice, however, such choices are almost impossible to make and we wish to have more freedom in choosing how we measure space and time. This can be arranged by adding a constant ***initial phase*** ϵ to the argument of the sine function, which results in the form

$$\Psi(z,t) = A\sin(kz - \omega t + \epsilon)\,, \tag{3.3.21}$$

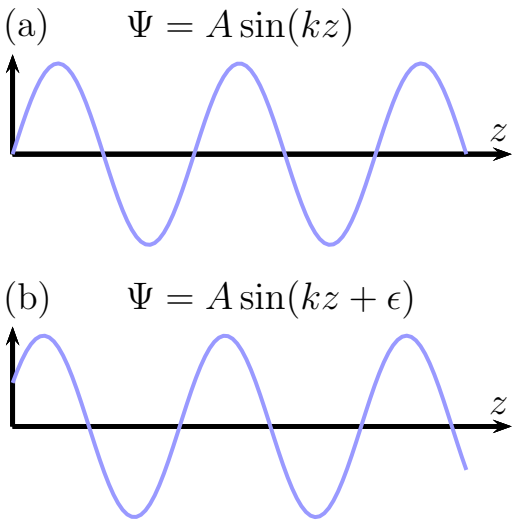

Fig. 3.3: Harmonic waves at time $t = 0$ and for intial phases of (a) 0 and (b) ϵ.

for the harmonic wave, as shown for $t = 0$ in Fig. 3.3b.

Finally, by recalling that $\cos\alpha = \sin(\alpha + \pi/2)$, we can represent exactly the same harmonic wave also by using the cosine function as

$$\Psi(z,t) = A\cos(kz - \omega t + \epsilon')\,, \tag{3.3.22}$$

where the initial phase ϵ' has now been chosen differently compared to Eq. (3.3.21).

[1] Especially in field of spectroscopy this quantity κ, often denoted as $\tilde{\nu}$, is called the wave number. In order to distinguish these different wave numbers from each other, $k = 2\pi/\lambda$ is sometimes called the ***angular wave number***.

3.4 Phase and phase velocity

We started our treatment by considering perturbations that move at a given velocity and maintain their shape, and then specialized to the case of harmonic waves.[2] In order to investigate this case in more detail, we determine a separate function for the phase of the wave

$$\phi(z,t) = kz - \omega t + \epsilon\,. \tag{3.4.23}$$

Recall that this function, together with the harmonic function, also determines the value of the wave itself.

We next want to understand at which velocity a certain point of a harmonic wave propagates. This point is determined by the phase function and maintains its height as the wave propagates (Fig. 3.4). If we now imagine ourselves riding with this point on the wave, it is evident that, from our point of view, the spatial location z is now dependent on time $z = z(t)$, and the phase function can be written as $\phi(z(t),t)$, i.e., it depends on time both explicitly and implicitly. By recalling the ***chain rule***, the total derivative of the phase function with respect of time is written as

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = \frac{\partial\phi}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial\phi}{\partial t}\frac{\partial t}{\partial t} = \frac{\partial\phi}{\partial z}\frac{\partial z}{\partial t} + \frac{\partial\phi}{\partial t}\,. \tag{3.4.24}$$

From the perspective of the traveling point that we are considering, however, the phase must be constant, i.e. total change in phase must be zero, because the point maintains its height as the wave propagates. We therefore obtain the requirement that Eq. (3.4.24) must vanish resulting in

$$\left(\frac{\partial z}{\partial t}\right)_{\phi} = -\frac{\partial\phi/\partial t}{\partial\phi/\partial z} = -\frac{(\,\mathrm{d}\phi/\,\mathrm{d}t)_z}{(\,\mathrm{d}\phi/\,\mathrm{d}z)_t}\,. \tag{3.4.25}$$

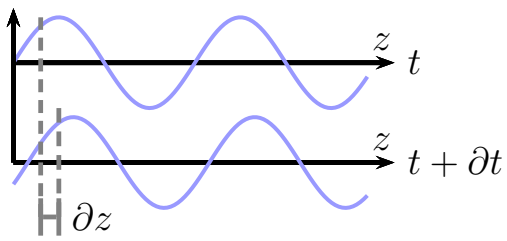

Fig. 3.4: A harmonic wave at times t and $t + \mathrm{d}t$.

In this notation, the subscripts emphasize which quantity is taken to be constant for each differentiation, i.e., the approach is very similar to the one used in thermodynamics. Furthermore, the left-hand side of Eq. (3.4.25) clearly represents the velocity for a point that is assumed to maintain constant phase. This quantity is known as the ***phase velocity***.

[2] We learn later that harmonic waves form a special family of waves because they are so-called ***eigensolutions*** of the wave equation in homogeneous (and linear) Cartesian space.

Let's next apply this result to the case of the phase function of Eq. (3.4.23). We obtain

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = -\omega\,, \qquad \frac{\mathrm{d}\phi}{\mathrm{d}z} = k\,, \tag{3.4.26}$$

and further, using Eqs. (3.4.25), (3.4.26) and (3.3.15), that the phase velocity is

$$\left(\frac{\partial z}{\partial t}\right)_{\phi} = \frac{\omega}{k} = v\,. \tag{3.4.27}$$

This result is naturally expected because our whole approach started from a perturbation propagating at velocity v.

Here, we applied this approach to a point with a given constant phase. However, we could also have applied it to the actual wave $\Psi(z,t)$. For an arbitrary wave this approach determines the instantaneous velocity of a given point at a given time.

Note also that the harmonic wave propagates as a whole, maintaining its shape. The sinusoidal oscillations are seen only at a given location as a function of time or at a given time as a function of space.

3.5 Superposition principle

We will now consider the case where we have two different solutions Ψ_1 and Ψ_2 to the wave equation, Eq. (3.2.8), allowing us to write

$$\frac{\partial^2\Psi_1}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\Psi_1}{\partial t^2}\,, \qquad \frac{\partial^2\Psi_2}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\Psi_2}{\partial t^2}\,. \tag{3.5.28}$$

Due to the fact that the derivatives are ***linear operators***, we can directly multiply these solutions by scalars α and β and add the two equations together resulting in equation

$$\frac{\partial^2}{\partial z^2}(\alpha\Psi_1 + \beta\Psi_2) = \frac{1}{v^2}\frac{\partial^2}{\partial t^2}(\alpha\Psi_1 + \beta\Psi_2)\,. \tag{3.5.29}$$

The above equation implies that the ***superposition*** $\alpha\Psi_1 + \beta\Psi_2$ of the two waves is also a solution to the same wave equation.

This result has the following extremely important implications: 1) The two perturbations $\alpha\Psi_1$ and $\beta\Psi_2$ maintain their independent properties and propagate independent of each other; 2) The total perturbation at a given location z and at a given time t is obtained as the superposition $\alpha\Psi_1 + \beta\Psi_2$ of the two individual perturbations.

When calculating the superposition, the total perturbation depends on the relative phase of the two individual waves. If the phase difference between the waves is 0°, the

two waves are ***in phase*** and amplify each other (Fig. 3.5a). This is an example of ***constructive interference***. If the phase difference is 180°, the two waves are ***out of phase*** and the superposition is weaker than either of the waves alone (Fig. 3.5b). This is an example of ***destructive interference***. In the general case, the phase difference can be arbitrary and the superposition can take on various intermediate cases between perfect constructive and destructive interference. More work is needed to calculate these general cases properly.

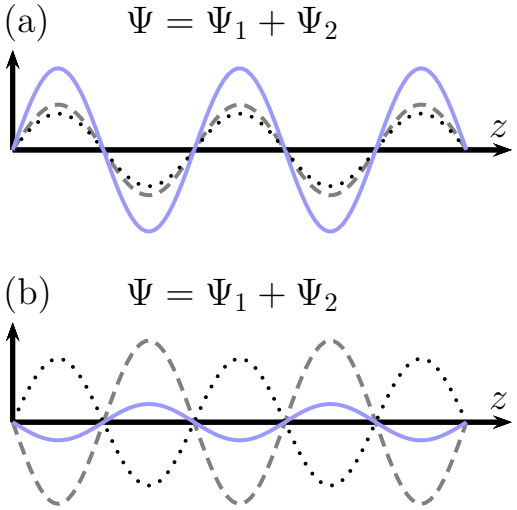

Fig. 3.5: In-phase (a) and out-of-phase (b) superpositions (blue) between two harmonic waves (dotted and dashed gray).

The ***superposition principle*** is one of the most important principles in optics.[3] Of course, the present case provides only the most elementary example of this principle. Nevertheless, its generalizations can be used to explain a wealth of very different phenomena in optics.

3.6 Complex representation of waves

Complex numbers

So far we have used sine and cosine functions to represent the harmonic waves. While this is reasonable for very simple situations, their use becomes increasingly cumbersome when dealing with more advanced cases. These problems can be overcome by using complex numbers to represent harmonic waves. We will now review some of the basic properties of complex numbers.

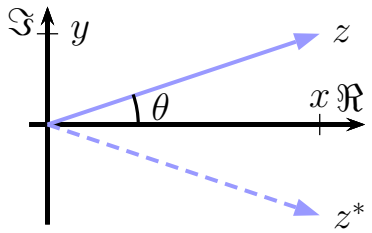

Fig. 3.6: A complex number $z = x + \mathrm{i}y$ and its complex conjugate $z^* = x - \mathrm{i}y$ in the complex plane.

A complex number z consists of a real part x and imaginary part y and is of the form

$$z = x + \mathrm{i}y\,, \tag{3.6.30}$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. Such a number can also be interpreted as a

[3] Note that this may be an understatement. The superposition principle may even be one of the most important principles in physics in general.

point (or vector) in the complex plane (Fig. 3.6) with the coordinates

$$x = r\cos\theta\,, \quad y = r\sin\theta\,, \tag{3.6.31}$$

where r is the length of the vector and angle θ is given by

$$\theta = \arctan\frac{y}{x}\,. \tag{3.6.32}$$

The horizontal axis of the complex plane thus corresponds to the real part of the complex number and the vertical axis to the imaginary part. Using ***Euler's formula***, the complex number can therefore be also represented as

$$z = r(\cos\theta + \mathrm{i}\sin\theta) = r\mathrm{e}^{\mathrm{i}\theta}\,. \tag{3.6.33}$$

A complex number has a complex conjugate z^*, which is obtained by changing the sign of the imaginary part, i.e.,

$$z^* = r(\cos\theta - \mathrm{i}\sin\theta) = r\mathrm{e}^{-\mathrm{i}\theta}\,. \tag{3.6.34}$$

The sum of and difference between two complex numbers $z_1 = x_1 + \mathrm{i}y_1 = r_1\mathrm{e}^{\mathrm{i}\theta_1}$ and $z_2 = x_2 + \mathrm{i}y_2 = r_2\mathrm{e}^{\mathrm{i}\theta_2}$ are easily calculated by adding or subtracting the real and imaginary parts separately

$$z_1 \pm z_2 = (x_1 \pm x_2) + \mathrm{i}(y_1 \pm y_2)\,. \tag{3.6.35}$$

The multiplication and division, however, are easier to calculate using the other (polar) form

$$z_1 z_2 = r_1 r_2 \mathrm{e}^{\mathrm{i}(\theta_1+\theta_2)}\,, \quad \frac{z_1}{z_2} = \frac{r_1}{r_2}\mathrm{e}^{\mathrm{i}(\theta_1-\theta_2)}\,. \tag{3.6.36}$$

The magnitude of the complex number is the length of the vector in the complex plane. In consequence,

$$|z|^2 = r\mathrm{e}^{\mathrm{i}\theta} r\mathrm{e}^{-\mathrm{i}\theta} = zz^* = x^2 + y^2\,. \tag{3.6.37}$$

Finally, the real and imaginary parts can be calculated directly as

$$\mathrm{Re}(z) = x = \frac{z + z^*}{2}\,, \qquad \mathrm{Im}(z) = y = \frac{z - z^*}{2\mathrm{i}}. \tag{3.6.38}$$

Complex representation of waves

A harmonic wave, Eq. (3.3.22), can be written as

$$\Psi(z,t) = A\cos(kz - \omega t + \epsilon)\,, \tag{3.6.39}$$

$$= \mathrm{Re}\left[A\mathrm{e}^{\mathrm{i}(kz-\omega t+\epsilon)}\right]\,, \tag{3.6.40}$$

$$= \frac{1}{2}A\mathrm{e}^{\mathrm{i}(kz-\omega t+\epsilon)} + \frac{1}{2}A\mathrm{e}^{-\mathrm{i}(kz-\omega t+\epsilon)}\,. \tag{3.6.41}$$

In the complex representation, the wave is often represented as

$$\Psi(z,t) = A\mathrm{e}^{\mathrm{i}(kz-\omega t+\epsilon)}\,, \tag{3.6.42}$$

$$= A\mathrm{e}^{\mathrm{i}\epsilon}\mathrm{e}^{\mathrm{i}(kz-\omega t)}\,, \tag{3.6.43}$$

$$= A'\mathrm{e}^{\mathrm{i}(kz-\omega t)}\,, \tag{3.6.44}$$

where the ***complex amplitude*** $A' = A\mathrm{e}^{\mathrm{i}\epsilon}$ accounts for the initial phase of the wave, thus making the notation a bit more compact.

Note that the complex representation, Eq. (3.6.44), contains an implicit assumption that the real part should be taken at the end of the calculation. In practice, however, this is usually not necessary even though the physical waves must be real-valued quantities.[4] In fact, the formalism based on the complex representation of waves has been so well-developed that the physically meaningful properties can be directly deduced from the form of the result. It is also important to note that the complex representation makes it much easier to do calculations that would be extremely difficult to do using the sine or cosine representation.

Note also that it is customary in optical physics to choose the complex time dependence to be $\mathrm{e}^{-\mathrm{i}\omega t}$, i.e, with a minus sign. In electrical engineering, however, and for optics researchers with this background, the time dependence is often of the form $\mathrm{e}^{\mathrm{j}\omega t}$, i.e., the imaginary unit is represented by j and the time dependence has a plus sign. This choice will influence the detailed form of the complex-valued quantities. When using results from existing literature, one therefore needs to first verify the choice of the sign for the time dependence.

[4] Recall that the **E**-field is defined through the (real-valued) force that acts on charged particles. From another point of view, the requirement for a real-valued electric field **E** complies also with the fact that the electric field $\mathbf{E}(\mathbf{r},t)$ is defined to be an even function under time reversal.

3.7 Plane waves

So far, we have assumed that the wave propagates along the z direction, i.e., that the spatial dependence of the wave is one-dimensional. The space, however, is three-dimensional, and we need to generalize our treatment to allow for the possibility that the wave propagates in an arbitrary direction specified by the ***wave vector*** $\mathbf{k}$.

To do this, we imagine that a wave propagating only along the z direction is rotated to propagate in the direction specified by the wave vector $\mathbf{k}$ (Fig. 3.7). To maintain all other properties of the wave, the phase must be constant in planes perpendicular to the wave vector $\mathbf{k}$. In addition, the phase must grow linearly as the position vector $\mathbf{r}$ moves along the direction of propagation.

We first choose an arbitrary plane that is perpendicular to the wave vector and a fixed point $\mathbf{r}_0$ on this plane. This plane is then defined by all other points $\mathbf{r}$ that fulfill the condition[5]

$$(\mathbf{r} - \mathbf{r}_0) \cdot \mathbf{k} = 0\,, \tag{3.7.45}$$

or by re-arranging the terms

$$\mathbf{k} \cdot \mathbf{r} = \mathbf{k} \cdot \mathbf{r}_0 = \text{constant}\,. \tag{3.7.46}$$

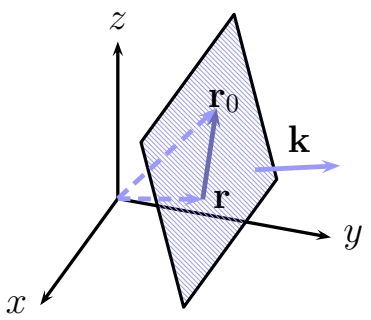

Fig. 3.7: Surface of a constant phase (i.e. a wavefront) for a plane wave propagating in the direction specified by the ***wave vector*** $\mathbf{k}$.

In addition, the constant must grow linearly along the direction of propagation $\mathbf{k}$. By choosing the spatial dependence of the ***plane wave*** to be of the form

$$\Psi(\mathbf{r}) = A\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\,, \tag{3.7.47}$$

we guarantee the wave to have the desired properties.

We next consider the periodicity of a plane wave. Because of its properties, the wave must repeat itself after propagating a distance of one wavelength in the direction of the wave vector $\mathbf{k}$. This imposes the wave the condition

$$\Psi(\mathbf{r}) = \Psi(\mathbf{r} + \lambda\hat{\mathbf{k}}) = \Psi(\mathbf{r} + \lambda\mathbf{k}/k)\,, \tag{3.7.48}$$

where $\hat{\mathbf{k}} = \mathbf{k}/\|\mathbf{k}\| = \mathbf{k}/k$ is the unit $\mathbf{k}$-vector pointing in the direction of the wave vector. By combining this requirement with Eq. (3.7.47), we immediately obtain the condition

$$\mathbf{k} \cdot (\lambda\hat{\mathbf{k}}) = \lambda\mathbf{k} \cdot \mathbf{k}/k = \lambda k = 2\pi\,. \tag{3.7.49}$$

[5]Recall the dot product's ***geometric definition***, $\|\mathbf{a}\|\,\|\mathbf{b}\|\cos\theta$ where θ is the angle between the vectors.

The wave number, or propagation number, of a plane wave k is thus equal to the length of its wave vector and, as before, obtained from

$$k = \frac{2\pi}{\lambda}\,. \tag{3.7.50}$$

By finally including the time dependence, we obtain the following general representation for a time-harmonic plane wave

$$\Psi(\mathbf{r},t) = A\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{r}-\omega t)}\,. \tag{3.7.51}$$

For such a wave, the ***wavefront*** is defined to be a surface of constant phase at a given time. Note that this definition relies on the phase function rather than the wave itself.[6] The reason for this is that the wave may not remain constant if the amplitude A is also a spatially varying function, which is possible in more general cases.[7]

We finally calculate the phase velocity for a plane wave using a different approach based on differentials with respect to space and time. The phase function for a plane wave is

$$\phi(\mathbf{r},t) = \mathbf{k}\cdot\mathbf{r} - \omega t = kr_k - \omega t\,, \tag{3.7.52}$$

where r_k is the projection of the position vector along the wave vector $\mathbf{k}$. For a point of constant phase, the total differential of this function must vanish, i.e.,

$$\mathrm{d}\phi(\mathbf{r},t) = k\,\mathrm{d}r_k - \omega\,\mathrm{d}t = 0\,, \tag{3.7.53}$$

from where we obtain the phase velocity as before in Eq. (3.4.27)

$$v = \frac{\mathrm{d}r_k}{\mathrm{d}t} = \frac{\omega}{k}\,. \tag{3.7.54}$$

3.8 Wave equation in three dimensions

We next wish to derive a wave equation that is valid in three spatial dimensions by following an approach similar to that of Section 3.4. In order to do this, we represent the wave vector $\mathbf{k}$ and position vector $\mathbf{r}$ in terms of their components

$$\mathbf{k} = \hat{\mathbf{x}}k_x + \hat{\mathbf{y}}k_y + \hat{\mathbf{z}}k_z\,, \tag{3.8.55}$$

$$\mathbf{r} = \hat{\mathbf{x}}x + \hat{\mathbf{y}}y + \hat{\mathbf{z}}z\,. \tag{3.8.56}$$

[6] Control of the phase function of the field is extremely important for example when one is working with very short pulses.

[7] Note that by taking use of the superposition principle, such more general cases could be treated, for example, as sums of plane waves.

Using these relations we can represent a plane wave in the form

$$\Psi(\mathbf{r},t) = A\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{r}-\omega t)} = A\mathrm{e}^{\mathrm{i}(k_x x+k_y y+k_z z-\omega t)}\,. \tag{3.8.57}$$

We next differentiate this equation twice with respect to each of the spatial coordinates and obtain

$$\frac{\partial^2\Psi}{\partial x^2} = -k_x^2\Psi\,, \qquad \frac{\partial^2\Psi}{\partial y^2} = -k_y^2\Psi\,, \qquad \frac{\partial^2\Psi}{\partial z^2} = -k_z^2\Psi\,. \tag{3.8.58}$$

By adding up these quantities, we find

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2} = -(k_x^2+k_y^2+k_z^2)\Psi = -k^2\Psi\,. \tag{3.8.59}$$

By also calculating the second derivative with respect to time, we obtain

$$\frac{\partial^2\Psi}{\partial t^2} = -\omega^2\Psi = -k^2v^2\Psi\,. \tag{3.8.60}$$

By finally combining the results from Eqs. (3.8.59) and (3.8.60), we obtain the wave equation for three-dimensional space

$$\frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} + \frac{\partial^2\Psi}{\partial z^2} = \frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2}\,. \tag{3.8.61}$$

This result can be written much more compactly by using the vector differential operators. In Cartesian coordinate system, the ***gradient*** is written as

$$\nabla = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\,, \tag{3.8.62}$$

while the ***Laplacian*** is given by

$$\nabla^2 = \nabla\cdot\nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\,. \tag{3.8.63}$$

By taking use of the Laplacian, the wave equation in three dimensions becomes

$$\nabla^2\Psi = \frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2}\,. \tag{3.8.64}$$

Although Eqs. (3.8.61) and (3.8.64) appear to be just two different ways of writing the same equation, there is an important difference between them. Equation (3.8.61) is based on the assumption that the problem can be treated using the Cartesian

coordinate system. While this is always possible, there are situations that are much easier to handle using a different coordinate system.[8] Equation (3.8.64), on the other hand, is independent of the coordinate system. The Laplacian can then be written out in the form appropriate for the chosen coordinate system, with the choice of Eq. (3.8.63) being only one particular possibility.

3.9 Spherical and cylindrical waves

When treating a given problem, it is useful to look at the possible symmetries present in the situation under consideration. One common case is that we expect the problem have spherical symmetry about the origin. It is then advisable to treat the problem using ***spherical coordinates*** (r, θ, ϕ) shown in Fig. 3.8a. The spherical coordinates (r, θ, ϕ) describing a point $\mathbf{r}$ can then be related to the Cartesian coordinates (x, y, z) through equations

$$x = r\sin\theta\cos\phi\,, \qquad y = r\sin\theta\sin\phi\,, \qquad z = r\cos\theta\,, \tag{3.9.65}$$

where r is the distance of the point $\mathbf{r}$ from the origin, θ is the polar angle measured from the z axis (acting as the zenith), and ϕ is the azimuthal angle in the x-y plane and measured from the x axis (acting as the azimuth).[9]

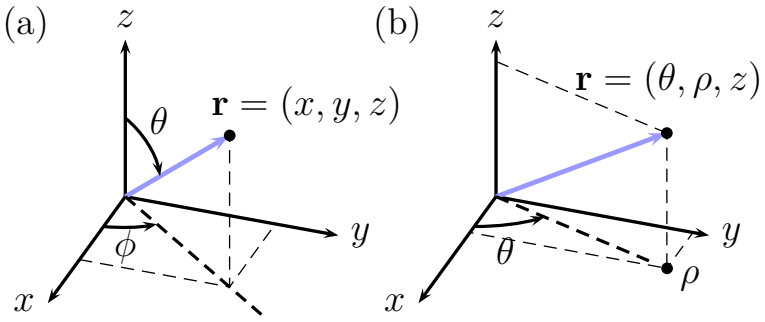

Fig. 3.8: (a) Spherical and (b) cylindrical coordinates.

The representation of the Laplacian operator in spherical coordinates is a bit cumbersome and we need not consider it in detail. However, it leads to solutions in the form of spherical waves given by

$$\Psi(\mathbf{r}, t) = \frac{A_0}{r}\mathrm{e}^{\mathrm{i}(\pm kr - \omega t)}\,, \tag{3.9.66}$$

where the (+) sign represents an outgoing (diverging) wave, the (–) sign represents an incoming (converging) spherical wave, and A_0 is a constant known as the strength of the source. This wave does have spherical symmetry, i.e., it depends only on the distance from the origin. Its wavefronts are therefore concentric spheres. In addition,

[8] Please remember that the 'smart' choice for the used coordinate system depends on the symmetries of the problem at hand.

[9] Please note that other conventions for the zenith and azimuth exist.

the amplitude of the spherical wave depends on position, i.e., $A(r) = A_0/r$. We may also conclude that far away from the origin, where the variations in the amplitude are relatively weak and the radius of curvature of the wavefronts large, a small fraction of a spherical wave looks very much like a plane wave (Fig. 3.9).

Another special case is cylindrical symmetry, which is best solved using the ***cylindrical coordinates*** (ρ, θ, z) shown in Fig. 3.8b), which are related to the Cartesian components through the equations

$$x = \rho\cos\theta\,, \qquad y = \rho\sin\theta\,, \qquad z = z\,, \tag{3.9.67}$$

where ρ is the distance of a point from the z axis in the x-y plane and θ is the azimuthal angle in the x-y plane and measured from the x axis.

The detailed form of a cylindrical wave is even more complicated than that of a spherical wave. However, far away from the z axis, it can be approximated by the form

$$\Psi(\mathbf{r}, t) \approx \frac{A_0}{\sqrt{\rho}}\mathrm{e}^{\mathrm{i}(\pm k\rho - \omega t)}\,. \tag{3.9.68}$$

The qualitative statements regarding spherical waves are also applicable for the cylindrical waves.

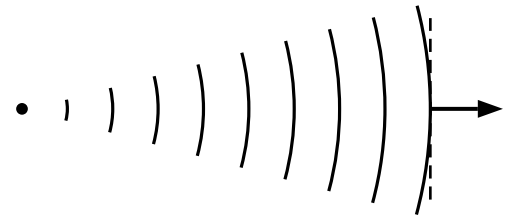

Fig. 3.9: Far away from its source, a fraction of a spherical or cylindrical wave resembles a plane wave.

3.10 Vector waves

So far, we have assumed that the amplitude of the harmonic wave is a scalar quantity. However, it can equally well be a vector quantity,[10] which leads to the following form for a plane wave

$$\mathbf{\Psi}(\mathbf{r}, t) = \mathbf{A}_0\mathrm{e}^{\mathrm{i}(\mathbf{k}\cdot\mathbf{r} - \omega t)}\,. \tag{3.10.69}$$

Of course, the vector amplitude $\mathbf{A}_0$ can also be represented in terms of its components, e.g., in the Cartesian basis as[11]

$$\mathbf{A}_0 = \hat{\mathbf{x}}A_x + \hat{\mathbf{y}}A_y + \hat{\mathbf{z}}A_z\,. \tag{3.10.70}$$

Each component of a vector wave is therefore a harmonic wave as such.

[10] This generalization allows to transport, in addition to energy, also e.g. the spin from point A to point B.

[11] As we later learn, there is some subtlety here. In fact already two components are enough to fully describe a time-harmonic vector plane wave, the final third component being dependent on the other two.

3.11 Tutorial on vector calculus

Next, we revise some of the basic properties of vectors and ***vector calculus***.[12] For simplicity and conceptual clarity, we restrict our treatment to the three-dimensional Cartesian coordinate system.[13] We start by defining two complex-valued scalars a and b and two (complex-valued) vectors $\mathbf{u} = [u_1,\, u_2,\, u_3]^\intercal$ and $\mathbf{v} = [v_1,\, v_2,\, v_3]^\intercal$. These vectors have the following basic algebraic operations

$$\mathbf{u} + \mathbf{v} = [u_1 + v_1, u_2 + v_2, u_3 + v_3]^\intercal\,, \qquad \textbf{\textit{vector addition}} \tag{3.11.71}$$

$$(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}\,, \qquad \textbf{\textit{scalar multiplication}} \tag{3.11.72}$$

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{3} u_i v_i = (u_1 v_1 + u_2 v_2 + u_3 v_3)\,, \qquad \textbf{\textit{dot product}} \tag{3.11.73}$$

$$\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\|\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{n}}\,, \qquad \textbf{\textit{cross product}} \tag{3.11.74}$$

where θ is the angle between $\mathbf{u}$ and $\mathbf{v}$ and $\hat{\mathbf{n}}$ is the unit vector perpendicular to the plane spanned by the vectors $\mathbf{u}$ and $\mathbf{v}$.[14] The dot and cross products of vectors $\mathbf{u}$ and $\mathbf{v}$ can be written compactly using the ***Einstein summation convention*** as

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^{3} u_i v_i = u_i v_i\,, \tag{3.11.75}$$

$$\mathbf{u} \times \mathbf{v} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \varepsilon^{ijk} u_j v_k \hat{\mathbf{e}}_i = \varepsilon^{ijk} u_j v_k \hat{\mathbf{e}}_i\,, \tag{3.11.76}$$

where the repeated indices in the Einstein notation correspond to summations over dummy indices. The ε^{ijk} is the ***Levi–Civita parity symbol*** defined in three dimensions commonly as

$$\varepsilon^{ijk} = \begin{cases} +1 & \text{if } (i,j,k) \text{ is } (1,2,3),\ (2,3,1), \text{or } (3,1,2) \\ -1 & \text{if } (i,j,k) \text{ is } (3,2,1),\ (2,1,3), \text{or } (1,3,2) \\ 0 & \text{if } \quad i = j,\ \text{or } j = k,\ \text{or } k = i\,. \end{cases} \tag{3.11.77}$$

---

[12] Further subtlety exists in the treatment of vectors, namely quantities known as ***pseudoscalars*** and ***pseudovectors***. These quantities behave differently upon certain coordinate transformations from their counterparts of normal scalars and vectors.

[13] By choosing a different metric/coordinate system, we could also force the treatment to be explicitly compatible e.g. with the ***special theory of relativity***. This we could do by generalizing the treatment into the four-dimensional ***Minkowski space***.

[14] The specific direction (up or down) of $\mathbf{u} \times \mathbf{v}$ is given by the ***right-hand rule***.

In addition to the above, operations known as the ***triple products*** exist

$$\mathbf{v}_1 \cdot (\mathbf{v}_2 \times \mathbf{v}_3) = \det(\mathbf{v}_1,\, \mathbf{v}_2,\, \mathbf{v}_3)\,, \qquad \textbf{\textit{scalar triple product}} \quad (3.11.78)$$

$$\mathbf{v}_1 \times (\mathbf{v}_2 \times \mathbf{v}_3) = (\mathbf{v}_1 \cdot \mathbf{v}_3)\mathbf{v}_2 - (\mathbf{v}_1 \cdot \mathbf{v}_2)\mathbf{v}_3\,. \qquad \textbf{\textit{vector triple product}} \quad (3.11.79)$$

These products are often encountered when using vector calculations to describe perceived physical reality.[15]

Next, we define the basic operations that are often encountered in ***vector calculus***. We start by defining the differential operation notations by taking a scalar-valued function $\phi(x, y, z)$ and a vector-valued function $\mathbf{E}(x, y, z)$.[16] In order for the differential operations to exist and be well behaving, we further require the functions to be ***continuously differentiable***. Working with such functions, the following differential operators are often encountered:

$$\nabla\phi = \frac{\partial\phi}{\partial x}\hat{\mathbf{x}} + \frac{\partial\phi}{\partial y}\hat{\mathbf{y}} + \frac{\partial\phi}{\partial z}\hat{\mathbf{z}}\,, \qquad \textbf{\textit{gradient}} \text{ (of a scalar function)} \quad (3.11.80)$$

$$\nabla\mathbf{E} = \mathbf{J}_{\mathbf{E}} = \left(\frac{\partial E_i}{\partial x_j}\right)_{ij}\,, \qquad \textbf{\textit{gradient}} \text{ (of a vector function)} \quad (3.11.81)$$

$$\nabla \cdot \mathbf{E} = \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\,, \qquad \textbf{\textit{divergence}} \quad (3.11.82)$$

$$\nabla \times \mathbf{E} = \varepsilon^{ijk}\frac{\partial E_k}{\partial x_j}\hat{\mathbf{e}}_i\,, \qquad \textbf{\textit{curl}} \quad (3.11.83)$$

where the Levi–Civita parity symbol ε^{ijk} and the Einstein summation convention allow to write the curl of the function $\mathbf{E}$ very compactly.[17] Note that the gradient of a vector-valued function is a 3×3 matrix, also known as the ***Jacobian matrix***.

Finally, we list some of the most often encountered ***second derivative identities***,

[15] For example, application of the scalar triple product results in the known vector calculus identity for the cross product of two vectors $\mathbf{A}$ and $\mathbf{B}$, given by $\nabla \cdot (\mathbf{A} \times \mathbf{B}) = (\nabla \times \mathbf{A}) \cdot \mathbf{B} - \mathbf{A} \cdot (\nabla \times \mathbf{B})$.

[16] Or alternatively scalar-valued function f and/or vector-valued function $\mathbf{F}$.

[17] The conventional notation based on the determinant may naturally also be used.

which will be found useful in our treatment:

$$\begin{aligned}
\Delta\phi = \nabla^2\phi &= (\nabla\cdot\nabla)\,\phi \\
&= \frac{\partial^2\phi}{\partial x^2} + \frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2}\,, && \textbf{\textit{Laplacian}} && (3.11.84)\\
\Delta\mathbf{E} = \nabla^2\mathbf{E} &= \nabla(\nabla\cdot\mathbf{E}) - \nabla\times\nabla\times\mathbf{E}\,, && \textbf{\textit{vector Laplacian}} && (3.11.85)\\
\nabla\times(\nabla\phi) &= \mathbf{0}\,, && \textbf{\textit{curl of gradient}} && (3.11.86)\\
\nabla\cdot(\nabla\times\mathbf{E}) &= 0\,. && \textbf{\textit{divergence of curl}} && (3.11.87)
\end{aligned}$$

The Laplacian of a scalar function can be also understood as the divergence of the gradient of the function. A slightly different form of the vector Laplacian is also known as the ***Lagrange's formula***.[18]

3.12 Electrostatic approximation and potentials

In order to familiarize with the above operations, let's next dwell briefly on a situation where the ***electrostatic approximation*** is valid. In this case, the electric field $\mathbf{E}$ occurs solely due to non-moving (i.e. stationary) electric charges described by the charge density ρ, prompting us to describe their connection using the ***Gauss' law***:

$$\oiint_S \mathbf{E}\cdot\,\mathrm{d}\mathbf{A} = \iiint_V \frac{\rho}{\epsilon_0}\,\mathrm{d}^3r\,, \tag{3.12.88}$$

where the left-hand side describes a surface integral over the surface $S = \partial V$ enclosing volume V (i.e. $\mathrm{d}\mathbf{A}$ is a differential surface element) while the right-hand side of the equation concerns an integration over a volume V (i.e. $\mathrm{d}^3r = \mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ is a differential volume element).[19] The above equation is the integral form of the Gauss' law. By taking use of the ***divergence theorem***[20] we can write the Gauss' law in a so-called differential form

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0}\,, \tag{3.12.89}$$

which is mathematically equivalent with the integral form. Therefore, in principle it is a matter of taste which form one uses. However, in the scope of this course we prefer to use the differential forms for conceptual clarity as they allow slightly cleaner notation.

---

[18] Namely the double curl equation $\nabla\times\nabla\times\mathbf{E} = \nabla(\nabla\cdot\mathbf{E}) - \nabla^2\mathbf{E}$.

[19] In Cartesian coordinate system, $\mathrm{d}\mathbf{A} = \mathrm{d}y\,\mathrm{d}z\,\hat{\mathbf{x}} + \mathrm{d}x\,\mathrm{d}z\,\hat{\mathbf{y}} + \mathrm{d}x\,\mathrm{d}y\,\hat{\mathbf{z}}$.

[20] Other useful properties of surface–volume and curve–surface integrals are found from ***here***.

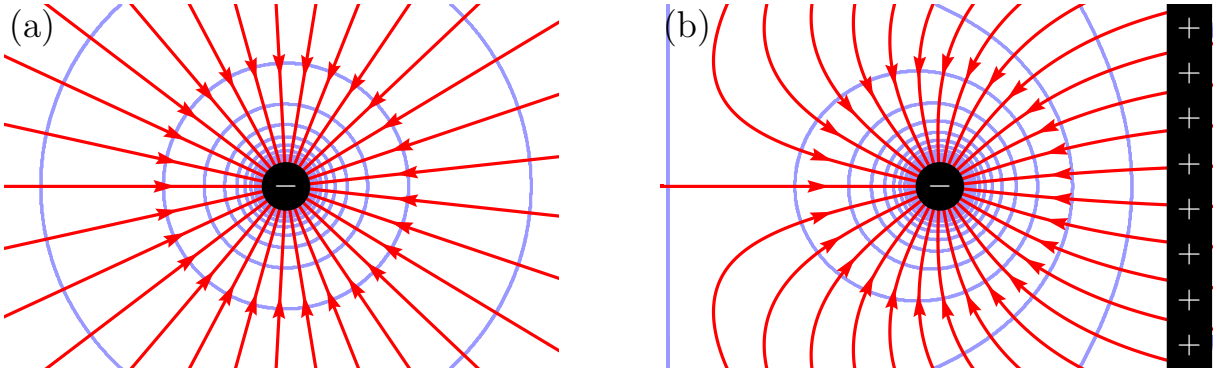

Fig. 3.10: Electrostatic electric field **E** (red lines) and the electric scalar potential ϕ (blue lines denote equipotential surfaces) due to a negative charge in (a) free-space and (b) near a flat positively charged surface.

Next, we note that in the considered situation of stationary charges ($\mathbf{J} = \mathbf{0}$), time-varying magnetic fields cannot exist. By now introducing the Maxwell–Faraday equation

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = \mathbf{0}\,, \tag{3.12.90}$$

we see that the curl of the field **E** is zero.[21] This means that the electrostatic electric field **E** is ***irrotational***. By recalling the above Eq. (3.11.86) regarding the curl of the gradient, we see that the electrostatic field **E** can be described as the gradient of some scalar field ϕ defined by the equation[22]

$$\mathbf{E} = -\nabla\phi\,. \tag{3.12.91}$$

The above equation, on the other hand, is nothing but the definition for a ***conservative vector field***, stating that **E** is in fact a conservative vector field, as it is described as the gradient of some function ϕ.

In terms of electromagnetism, this scalar function ϕ is better known as the ***electric scalar potential***. This observation leads us to note that any line integral over the electrostatic field **E** is path independent.[23] Finally, we use the above Eq. (3.12.91) to

[21] Rigorous derivation of this statement follows by noting that in this case the electric field **E** can be described using the ***generalized Coulomb's law***. Such a field can be written as a gradient of a scalar function ϕ ($\mathbf{E} = -\nabla\phi$). Recalling the identity of Eq. (3.11.86) we see that **E** has to be irrotational ($\nabla \times \mathbf{E} = \mathbf{0}$).

[22] The minus sign is inserted in order to comply with the conventional notation based on the ***Helmholtz decomposition*** of the field **E**. **This fundamental theorem of vector calculus** states that **any** twice continuously differentiable vector field **F** can be decomposed into curl-free and divergence-free components as given by $\mathbf{F} = -\nabla\Phi + \nabla \times \mathbf{A}$.

[23] Much of high-school physics and classical mechanics concerns conservative vector fields. For example, to a ***very good approximation*** the gravitational field on the surface of Earth is conservative equipping us with the convenient notion of gravitational potential energy ($E = mgh$).

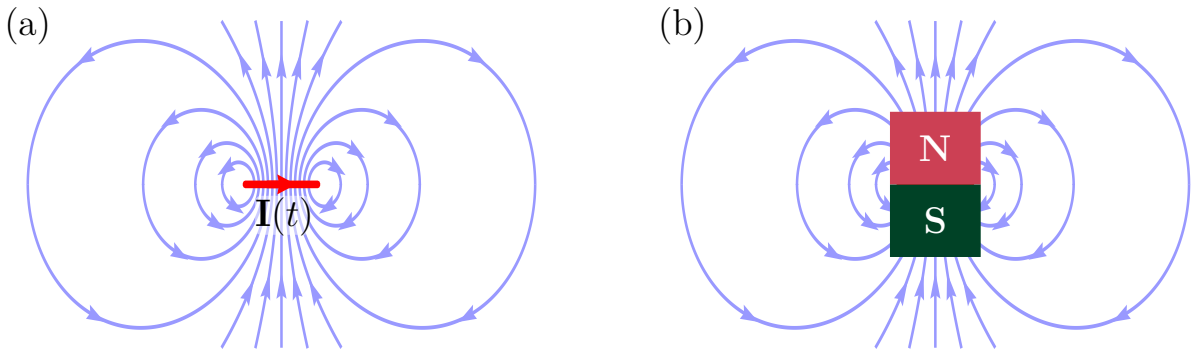

Fig. 3.11: Magnetostatic magnetic field **B** (blue lines) generated by a (a) loop of an electric current **I** (red) and (b) a permanent magnet. Note the similarity of the generated **B**-fields and their solenoidal nature.

transform the Gauss' law as given by Eq. (3.12.89) into

$$\nabla \cdot \mathbf{E} = -(\nabla \cdot \nabla)\phi = -\nabla^2\phi = -\Delta\phi = \frac{\rho}{\epsilon_0}\,, \tag{3.12.92}$$

where the last equality is a second-order partial differential equation known as the ***Poisson's equation***. Solutions of this equation are known allowing us to describe the electrostatic electric field **E** after calculation of the gradient of the ϕ (see Fig. 3.10). Therefore, it seems a matter of taste whether the electric scalar potential ϕ or the electrostatic field **E** is used to describe the related phenomena. Next we note, that solutions of the above Poisson's equation do not result in propagating waves, in other words light.[24] Such propagating fields (i.e. waves) require moving charges and time-varying magnetic fields, which situation will be considered in the next Chapter.

We end this discussion by stating that similar conditions known as the ***magnetostatics*** can also occur, and be utilized to understand electromagnetic behavior under such conditions (see Fig. 3.11). In this case, a new potential field, known as the ***magnetic vector potential*** **A** needs to be introduced. This introduction changes also the description of the electric field **E** and magnetic field **B** in terms of the two potentials, which are defined to fulfil the following equations

$$\mathbf{B} = \nabla \times \mathbf{A}\,, \qquad \mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial t}\,. \tag{3.12.93}$$

These equations form the basis for treating electrodynamics in the potential form.[25]

[24] To clarify, the electrostatic field **E** exhibits capability to do work, e.g. to move/accelerate charges, but does not transfer energy or other attributes from point A to B.

[25] The potential formulation is often used for example in treatment of ***relativistic electrodynamics***.

4. ELECTROMAGNETIC WAVES

4.1 Microscopic Maxwell's equations

Light is known to be electromagnetic radiation. We can therefore expect that the existence of electromagnetic waves can be predicted starting from the laws that govern electricity and magnetism. In 1865, Maxwell collected the existing information about these two interconnected fields and found out that they can be expressed in the form of a set of equations. In fact, the ***Maxwell's equations*** can be written in a number of different forms, depending on what type of material is considered, on what level (microscopic versus macroscopic) the material is described, and what types of electromagnetic quantities are explicitly considered. For conceptual clarity, let's start by taking the ***microscopic equations*** given by

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}\,, \qquad \text{(Gauss's law)} \tag{4.1.1}$$

$$\nabla \cdot \mathbf{B} = 0\,, \qquad \text{(Gauss's law for magnetism)}\,, \tag{4.1.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\,, \qquad \text{(Maxwell} - \text{Faraday equation)}\,, \tag{4.1.3}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\,, \qquad \text{(Ampère's circuital law)}\,, \tag{4.1.4}$$

where $\mathbf{E}$ is the electric field, $\mathbf{B}$ is the magnetic field,[1] ρ is the total electric charge density (total charge per unit volume), $\mathbf{J}$ is the total electric current density (total current per unit area), $\epsilon_0 = 8.854 \times 10^{-12}$ F/m is the permittivity (dielectric constant) of vacuum, and $\mu_0 = 1.257 \times 10^{-5}$ H/m is the permeability (magnetic constant) of vacuum.

The simplest possible case to consider is vacuum. There are no charges in vacuum ($\rho = 0$, and $\mathbf{J} = \mathbf{0}$)[2] simplifying the above Maxwell's equations into

$$\nabla \cdot \mathbf{E} = 0\,, \qquad \nabla \cdot \mathbf{B} = 0\,, \tag{4.1.5}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\,, \qquad \nabla \times \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}\,. \tag{4.1.6}$$

On the basis of Chapter 3, a wave equation will involve second derivatives with respect to space and time. We therefore operate with $\nabla\times$ on Eqs. (4.1.6). By assuming that $\mathbf{B}$ and $\mathbf{E}$ are sufficiently well-behaved functions that they can be differentiated twice (which is always true for physical quantities), we can change the order of spatial and

[1] Different terminology exist, such as the magnetic flux density, magnetic induction field. I prefer to use the simplest and the one making most sense using the microscopic formulation.

[2] In other words, we are going to treat next only the propagation of light.

temporal derivatives to obtain

$$\nabla \times (\nabla \times \mathbf{E}) = -\frac{\partial}{\partial t}(\nabla \times \mathbf{B})\,, \tag{4.1.7}$$

$$\nabla \times (\nabla \times \mathbf{B}) = \mu_0\epsilon_0\frac{\partial}{\partial t}(\nabla \times \mathbf{E})\,. \tag{4.1.8}$$

In these equations, the left-hand sides can be manipulated using known identities regarding the nabla operator ∇,[3] whereas application of Eqs. (4.1.6) the second time the right-hand sides can be re-written to yield

$$\nabla(\nabla \cdot \mathbf{E}) - \nabla^2\mathbf{E} = -\mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}\,, \tag{4.1.9}$$

$$\nabla(\nabla \cdot \mathbf{B}) - \nabla^2\mathbf{B} = -\mu_0\epsilon_0\frac{\partial^2\mathbf{B}}{\partial t^2}\,, \tag{4.1.10}$$

Finally, the first terms on the left-hand sides disappear by applying Eqs. (4.1.5), resulting in two equations given by

$$\nabla^2\mathbf{E} = \mu_0\epsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2}\,, \tag{4.1.11}$$

$$\nabla^2\mathbf{B} = \mu_0\epsilon_0\frac{\partial^2\mathbf{B}}{\partial t^2}\,. \tag{4.1.12}$$

These equations are exactly of the form of the wave equation (Eq. (3.2.8)) for the vector fields $\mathbf{E}$ and $\mathbf{B}$, respectively. The implication is that the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$ propagate as waves. We further note that the fields $\mathbf{E}$ and $\mathbf{B}$ are intimately connected through the (curl) Eqs. (4.1.6). In other words, they are only different manifestations of a single entity known as the ***electromagnetic field tensor***.[4] The fields $\mathbf{E}$ and $\mathbf{B}$ thus propagate together at the speed that can be directly read from Eqs. (4.1.11) and (4.1.12)

$$c = 1/\sqrt{\mu_0\epsilon_0} \approx 3 \times 10^8 \text{ m/s}\,, \tag{4.1.13}$$

where c emphasizes that the quantity is the ***speed of light***.

4.2 Transverse waves and charge conservation

We next derive some of the basic properties of electromagnetic waves by considering a simple case of a time-harmonic plane wave as an example. Although this

[3] Recall the ***vector triple product*** identity we also know as the Lagrange's formula $\nabla \times \nabla \times \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla^2\mathbf{F}$.

[4] A glimpse of this can be also seen by recalling the potential formulation of the Maxwell's equations.

situation may seem simplistic, it is by no means a limitation, because recalling ***Fourier synthesis*** any solution of the wave equation Eq. (4.1.11) can be constructed as a sum of time-harmonic waves.[5] We thus take an electric field of form $\mathbf{E}(\mathbf{r},t) = \mathbf{E}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t} = \mathbf{P}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\mathrm{e}^{-\mathrm{i}\omega t}$. This wave propagates along $\mathbf{k}$ and by recalling that the field is ***solenoidal***, i.e., divergence-free (see Eq. (4.1.5)) we obtain the condition[6]

$$\mathbf{k}\cdot\mathbf{P} = 0\,. \tag{4.2.14}$$

It therefore follows that the electric field $\mathbf{E}$ is a ***transverse wave***.[7] In other words, the field can only have components of $\mathbf{P}$ that are perpendicular to the direction of propagation $\mathbf{k}$. Similar conclusion can be easily derived for the magnetic field $\mathbf{B}$. The possible directions of the electric component of the electromagnetic field are known as ***polarizations***. Note also that the polarization components can be complex-valued.

Next, we show that the $\mathbf{E}$ and $\mathbf{B}$ fields are perpendicular to each other. We start by inserting the above considered field into Eq. (4.1.6) resulting in[8]

$$\nabla\times\mathbf{E}(\mathbf{r},t) = \mathrm{i}\mathbf{k}\times\mathbf{E}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t} = -\frac{\partial\mathbf{B}}{\partial t}. \tag{4.2.15}$$

Now by integrating both sides over time and taking use of the fact that $\mathbf{k} = \|\mathbf{k}\|\hat{\mathbf{k}} = k\hat{\mathbf{k}}$ where $k = \omega/c$, we get an equality[9]

$$\hat{\mathbf{k}}\times\mathbf{E}(\mathbf{r},t) = c\mathbf{B}(\mathbf{r},t). \tag{4.2.16}$$

The fields $\mathbf{E}$ and $\mathbf{B}$, and the direction of propagation $\mathbf{k}$ thus form a right-handed vector triplet and all three vectors are perpendicular to each other (Fig. 4.1). Note also that the fields $\mathbf{E}$ and $\mathbf{B}$ oscillate in phase and that the proportionality factor between the electric and magnetic components is c. These results hold in general for electromagnetic waves in vacuum.

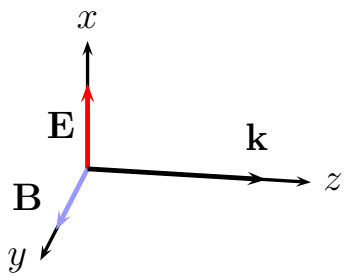

Fig. 4.1: The relative orientations of $\mathbf{E}$ and $\mathbf{B}$ fields for x-polarized plane wave propagating in z direction.

The conservation of charge also follows implicitly from the Maxwell's equations. We see this by first calculat-

[5] Similarly, an arbitrary spatial profile can be constructed as a sum of many plane waves propagating in slightly different directions.

[6] This straightforward result is derived as a homework. Additional information is found by considering the ***Helmholtz decomposition*** of the field.

[7] Recall the geometric definition of ***dot product***, $\mathbf{k}\cdot\mathbf{P} = \|\mathbf{k}\|\|\mathbf{P}\|\cos\theta$ where θ is the angle between $\mathbf{k}$ and $\mathbf{P}$.

[8] Again, this result is derived in a homework.

[9] Remember that $\int f'(t)\mathrm{e}^{f(t)}\,\mathrm{d}t = \mathrm{e}^{f(t)}$. So that here $\int \frac{-\mathrm{i}\omega}{-\mathrm{i}\omega}\mathbf{E}(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}\,\mathrm{d}t = \frac{\mathbf{E}(\mathbf{r},t)}{-\mathrm{i}\omega}$.

ing the divergence of Eq. (4.1.4) and recalling that a solenoidal field is also a transverse field resulting in

$$\nabla \cdot (\nabla \times \mathbf{B}) = \nabla \cdot \left(\mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) ,$$
$$0 = \mu_0 \left(\nabla \cdot \mathbf{J} + \frac{\partial}{\partial t} \epsilon_0 \nabla \cdot \mathbf{E} \right) . \tag{4.2.17}$$

By using the Gauss's law (Eq. (4.1.1)), the above equation simplifies into condition

$$\nabla \cdot \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 , \tag{4.2.18}$$

which is known as the equation for the ***charge conservation***.[10]

4.3 Energy of electromagnetic field

We have now shown that the laws of electromagnetism give rise to transverse waves where the electric field **E** and the magnetic field **B** propagate together. We may next consider which other important physical quantities can be associated with electromagnetic waves. One essential quantity, of course, is energy. Because we have no other guidelines to calculate the energy of waves, which oscillate in space and time, we assume that we can apply the results from ***electrostatics*** and ***magnetostatics*** to calculate the energy content of the field. More specifically, the ***energy density***, i.e., energy per unit volume of the electric field is given by[11]

$$u_E = \frac{\epsilon_0}{2} E^2 . \tag{4.3.19}$$

Similarly, the energy density of the magnetic field is written as

$$u_B = \frac{1}{2\mu_0} B^2 , \tag{4.3.20}$$

By recalling Eq. (4.1.13) for the speed of light c and Eq. (4.2.16) for the interrelation between the electric and magnetic quantities, we immediately conclude that the magnetic energy density can be written as

$$u_B = \frac{1}{2} \epsilon_0 c^2 B^2 = \frac{\epsilon_0}{2} E^2 = u_E . \tag{4.3.21}$$

[10]Note here that the Maxwell's equations are ***lorentz invariant*** implying their compatibility with the theory of ***special relativity***. Therefore, this result is alternatively and neatly derived by setting the divergence of the Lorentz invariant ***four-current*** to be zero. See also the connection between the ***classical electromagnetism and special relativity***.

[11]The derivation is beyond our scope, but follows straightforwardly from the Gauss's law, the properties of conservative fields and the ***electric potential energy***.

The energy density of the electromagnetic field is thus equally distributed to the electric and magnetic components of the field. This result is very satisfactory because we know that the two components can propagate only together as a wave and we expect the wave therefore to exhibit high symmetry between the two components. The total energy density of the wave can thus be written as

$$u = u_E + u_B = \epsilon_0 E^2 = \frac{1}{\mu_0} B^2 \,. \tag{4.3.22}$$

As the wave propagates in some direction, we expect that the wave also carries electromagnetic energy in the same direction.[12] In order to understand how much energy crosses a given area A in a given time t, we first consider the situation for a plane wave. We then imagine a volume V with cross-sectional area A and length l along the wave (Fig. 4.2). The total amount of electromagnetic energy in this volume is thus

$$U = uV = uAl \,. \tag{4.3.23}$$

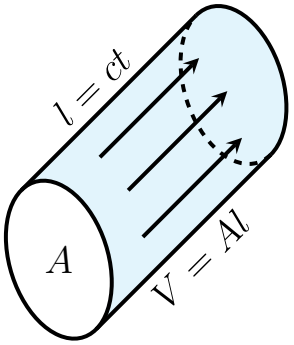

Fig. 4.2: Field volume for energy considerations.

Because this volume propagates at the speed of light c, the associated energy U crosses the cross-sectional area A in time $t = l/c$. The quantity we are interested is known as the ***irradiance*** of the wave and is found to be

$$S = \frac{U}{At} = uc = c\epsilon_0 E^2 = c^2 \epsilon_0 EB \,, \tag{4.3.24}$$

where the last form is used to emphasize the symmetry between the electric and magnetic components of the field. In everyday life, irradiance is often called ***intensity***. However, in strict terminology intensity has a different meaning.

If our medium is isotropic, i.e., it looks similar in all possible directions, the electromagnetic energy must propagate in the same direction as the wave. This assumption is certainly true for vacuum. We can therefore associate a vector with the energy flow of the electromagnetic field. This vector is known as the ***Poynting vector***, and is obtained from Eq. (4.3.24), by writing it in the vector

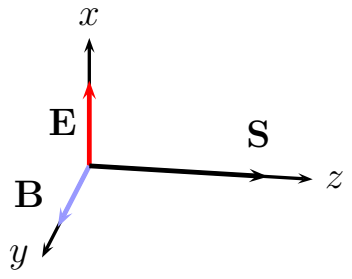

Fig. 4.3: Poynting vector **S**.

[12] Interestingly, in some materials this is not the case.

form (Fig. 4.3)

$$\mathbf{S} = c^2\epsilon_0 \mathbf{E} \times \mathbf{B}\,. \tag{4.3.25}$$

The irradiance of the wave, Eq. (4.3.24) is thus obtained from the magnitude of the Poynting vector $S = \|\mathbf{S}\|$.

We will next apply these results explicitly to the case of a time-harmonic plane wave. To take all possible factors into account, we first treat the wave using the cosine function. The electric component $\mathbf{E}$ and magnetic component $\mathbf{B}$ of the wave are thus

$$\mathbf{E} = \mathbf{E}_0 \cos(\mathbf{k}\cdot\mathbf{r} - \omega t)\,, \tag{4.3.26}$$

$$\mathbf{B} = \mathbf{B}_0 \cos(\mathbf{k}\cdot\mathbf{r} - \omega t)\,. \tag{4.3.27}$$

This immediately yields the Poynting vector to be of form

$$\mathbf{S} = c^2\epsilon_0 \mathbf{E}_0 \times \mathbf{B}_0 \cos^2(\mathbf{k}\cdot\mathbf{r} - \omega t)\,, \tag{4.3.28}$$

which depends on space and time through the square of the cosine function.

In order to understand how the result of Eq. (4.3.28) should be interpreted, we need to consider how we could try to measure the irradiance. This is done using photodetectors, which convert the incident irradiance to an electric signal through the photoelectric effect (Fig. 4.4). If we assume that the photodetector is placed at location $\mathbf{r} = 0$, the signal to be measured oscillates in time as $\cos^2(\omega t)$. For optical frequencies ($\sim 10^{15}$ Hz), however, these oscillations occur so rapidly that no detector can follow them in real time. The detector is therefore sensitive to a time average of the Poynting vector measured over appropriately long time. It is evident that the time average of the cosine function vanishes over sufficiently long times

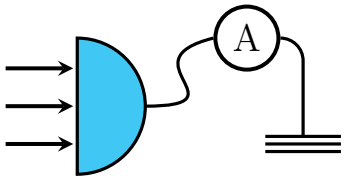

Fig. 4.4: A photodetector converts incident light to an electrical signal.

$$\langle \cos \omega t \rangle_T = 0\,. \tag{4.3.29}$$

The average of the square of the cosine function, however, yields

$$\left\langle \cos^2 \omega t \right\rangle_T = \left\langle \frac{1}{2} + \frac{1}{2}\cos 2\omega t \right\rangle_T = \frac{1}{2}\,, \tag{4.3.30}$$

By taking all of the above into account, the time average of the Poynting vector is

$$\langle \mathbf{S} \rangle_T = \frac{1}{2} c^2 \epsilon_0 \mathbf{E}_0 \times \mathbf{B}_0\,, \tag{4.3.31}$$

and its magnitude can be written as

$$\langle \|\mathbf{S}\| \rangle_T = \frac{1}{2} c^2 \epsilon_0 E_0 B_0 = \frac{1}{2} c \epsilon_0 E_0^2 \,. \tag{4.3.32}$$

The last form of this equation is written only in terms of the electric field. The reason for this is that the interaction of matter with the electric component of the electromagnetic field is usually much stronger than the interaction with the magnetic component.

By taking all these considerations into account, it is reasonable to define the irradiance in the optical regime through the time average as obtained from the magnitude of the Poynting vector given by Eq. (4.3.32). By using the cosine representation for the electric component of the electromagnetic field, the irradiance is thus

$$I = \frac{1}{2} c \epsilon_0 \|\mathbf{E}_0\|^2 = \left\langle c \epsilon_0 \|\mathbf{E}\|^2 \right\rangle_T \,, \quad \text{when} \qquad \mathbf{E} = \mathbf{E}_0 \cos(\mathbf{k} \cdot \mathbf{r} - \omega t) \,. \tag{4.3.33}$$

The irradiance can thus be calculated from the amplitude of the wave, in which case there is no need to calculate the time average. Alternatively, the irradiance can be obtained from the space- and time-dependent field in which case the time average needs to be calculated.

From now on, we prefer to use the complex notation for the electromagnetic field. For example at location $\mathbf{r} = 0$, the time-dependent complex field is

$$\mathbf{E} = \mathbf{E}_0 \mathrm{e}^{-\mathrm{i}\omega t} \,, \quad \text{and} \qquad \mathbf{E}^* = \mathbf{E}_0^* \mathrm{e}^{\mathrm{i}\omega t} \,. \tag{4.3.34}$$

The square of the magnitude of this quantity is thus

$$\|\mathbf{E}\|^2 = \mathbf{E} \cdot \mathbf{E}^* = \mathbf{E}_0 \cdot \mathbf{E}_0^* \,. \tag{4.3.35}$$

The irradiance is therefore found to be

$$I = \frac{1}{2} c \epsilon_0 \|\mathbf{E}\|^2 = \frac{1}{2} c \epsilon_0 \mathbf{E}_0 \cdot \mathbf{E}_0^* \,, \tag{4.3.36}$$

i.e., there is no need to calculate the time average explicitly. This, of course, arises from the fact that our convention for the complex notation does not treat the two terms of Eq. (3.6.44) explicitly. The result of Eq. (4.3.36) is nevertheless extremely useful as long as one remembers the assumptions are behind the complex notation.[13]

[13]Recall that the time-varying electric field $\mathbf{E}(\mathbf{r}, t)$ is a real-valued quantity.

4.4 Radiation pressure and momentum of electromagnetic field

Another important quantity familiar from many other contexts is the momentum.[14] Maxwell already argued that the electromagnetic field must give rise to a ***radiation pressure*** on a material body. He found out that this pressure is equal to the energy density of the field, i.e.,

$$P = u = S/c\,. \tag{4.4.37}$$

Due to the rapid oscillations of the optical fields, this quantity is also better to treat through its time average

$$\langle P \rangle_T = I/c\,. \tag{4.4.38}$$

Because the field can exert a pressure on a material body, it must also carry momentum. We are interested in ***momentum density***, i.e., momentum per unit volume. We next consider a situation where a plane wave strikes a piece of material that fully absorbs the entire field striking the material (Fig. 4.5). As in the previous section, we consider a volume V of the plane wave with cross-sectional area A and length l. The force exerted by this volume on the material can be calculated in two different ways as

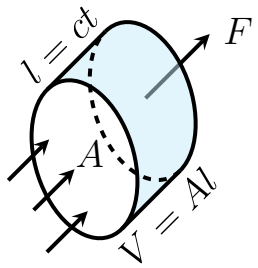

Fig. 4.5: Light striking an object exerts a force F on the object.

$$F = PA = \frac{p}{t}\,, \tag{4.4.39}$$

where p is the total momentum in volume V and $t = l/c$ is the time it takes for this volume to be absorbed. In terms of the momentum density p_V, we obtain

$$PA = \frac{p_V V}{t} = \frac{p_V A l}{l/c}\,. \tag{4.4.40}$$

The pressure is thus found to be $P = p_V c$ and using Eq. (4.4.38) we find that the time average of the momentum density is

$$p_V = I/c^2\,. \tag{4.4.41}$$

It is important to note that this treatment was made for a material that absorbs all the incident radiation. For the case of a reflecting material, the light after reflection

[14] Interestingly, this might e.g. provide a superior alternative to traditional rocket propulsion technology used in today's spacecrafts. E.g. project ***breakthrough starshot*** aims to use a phased array of lasers to propel and steer a spacecraft to alpha centauri. Travel time is estimated to be only 20 years implying the spacecraft to move at a speed >10% of the speed of light.

would be propagating in a direction opposite to the original, i.e., the direction of its momentum flow would be reversed.

4.5 Dipole radiation

It is known from the electromagnetic theory that electric charges under accelerating motion act as sources of radiation.[15] Depending on the wavelength range under consideration, this can be accomplished in different ways. For example, linear particle accelerators, cyclotrons, and synchrotrons are all used to generate radiation at very short wavelengths.[16] On the other hand, electric currents oscillating along metal wires provide the basis for controlling electromagnetic waves at radio frequencies.

In the optical regime, by far the most important source of waves is ***dipole radiation***. More specifically, this term refers to radiation from ***electric dipoles***. Such dipoles consist of a positive and a negative charge separated by some distance $d = \|\mathbf{d}\|$ (Fig. 4.6). The ***dipole moment*** of such a system is defined to be[17]

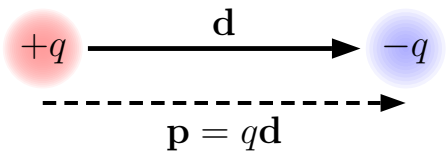

Fig. 4.6: Opposite charges of equal magnitude q separated by $\mathbf{d}$ give rise to a dipole moment $\mathbf{p}$.

$$\mathbf{p} = q\mathbf{d}\,, \tag{4.5.42}$$

where q is the magnitude of the two charges and $\mathbf{d}$ is the separation vector between the charges as measured from the negative one to the positive one. At distances sufficiently far away from the two charges, the system can be approximately described as a point-like vector quantity.

In order for such a system to radiate, the two charges need to be in accelerating motion. This is conveniently achieved by setting the charges to oscillate in time (Fig. 4.7), which gives rise to a time-dependent dipole moment of the form

$$\mathbf{p} = \mathbf{p}_0 \mathrm{e}^{-\mathrm{i}\omega t}\,, \tag{4.5.43}$$

where $\mathbf{p}_0$ is the amplitude of the dipole oscillation.

---

[15] Note that this classical picture leads to controversies at very small dimensions, e.g. when dealing with atomic systems. In such cases one needs to resort to ***quantum mechanics***.

[16] Interestingly to some, more compact table-top sources, that are based in nonlinear optical processes, are under development.

[17] Recall that the dipole moment of a molecular system is often defined to be $\mathbf{p} = \epsilon_0 \alpha \mathbf{E}$ where α is the ***molecular polarizability***.

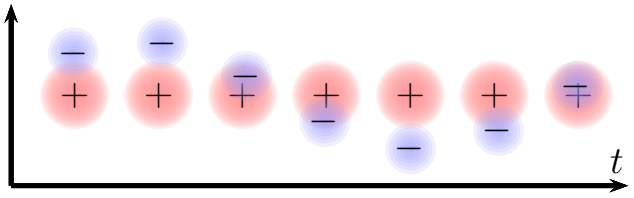

Fig. 4.7: Charges of an oscillating dipole moment at different times.

The detailed theory of dipole radiation is quite involved, especially when the radiation field is considered at distances from the dipole comparable to the separation between the charges. The reason for this is that at such short distances, the two charges need to be considered separately. In such ***near field***, all the field components do not even correspond to propagating waves. On the other hand, at distances sufficiently far away from the two charges, the system looks like a point-like vectorial quantity. Such a case is often described using terms like ***far field***, ***radiation field***, or ***radiation zone***. In this regime, the field only consists of propagating waves.

The radiation field can be approximated with a particularly simple form, where the scalar electric component of the wave is

$$E(r,\theta,t) = \frac{p_0 k^2 \sin\theta}{4\pi\epsilon_0}\frac{\mathrm{e}^{\mathrm{i}(kr-\omega t)}}{r}\,. \tag{4.5.44}$$

Here $p_0 = \|\mathbf{p}_0\|$ is the amplitude of the dipole oscillation, wave number $k = \omega/c$, r is the distance from the dipole, and θ is the angle between the direction of observation and the dipole moment $\mathbf{p}$ (Fig. 4.8). This result has a very intuitive physical interpretation. First, as a point-like source, the dipole emits a spherical wave as described by the second factor of Eq. (4.5.44). Second, an electric dipole tends to produce an electric field that oscillates in the same direction as the dipole. A propagating field, however, is constrained by the fact that the field must be transverse (recall Eq. (4.2.14)). The best solution is then obtained by projecting the dipole moment into a direction that is transverse with respect to the direction of observation. This is described by the $\sin\theta$ factor. The remaining factors are required for proper normalization of the field generated by the dipole.

By using Eq. (4.3.36), we find that the irradiance of dipole radiation is

$$I(r,\theta) = \frac{p_0^2\omega^4}{32\pi^2 c^3\epsilon_0}\frac{\sin^2\theta}{r^2}\,. \tag{4.5.45}$$

The most important qualitative factor to note here is that dipole radiation scales with the fourth power of frequency ($\propto \omega^4$) or, alternatively, with the fourth

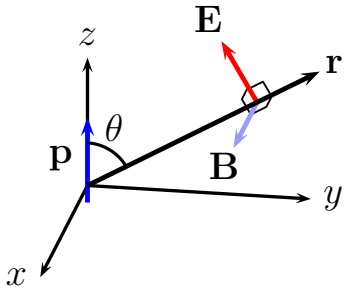

Fig. 4.8: Far field radiation from an oscillating dipole.

power of the inverse of wavelength ($\propto 1/\lambda^4$). This
scaling is very strong, with short wavelengths being radiated/scattered much more strongly than long ones.

The strong scaling of radiation with wavelength even affects the world as we see it around us in everyday life. More specifically, the blue end of the spectrum scatters much more efficiently than the red end. This simple fact can explain the color of the sky at different times of the day. During the daytime, light from the Sun is directly above and passes only through a relatively thin layer of atmosphere. The blue light is scattered very efficiently several times. In consequence, blue light is everywhere and propagating in all possible directions, making the sky appear blue for any viewer. During the sunset, on the other hand, we are viewing directly at the Sun. The light reaching our eyes from the Sun must then propagate through a thick layer of air along the surface of the Earth. In this case, majority of blue light has been scattered away into all directions and the dominant component reaching our eyes is orange and/or red.

4.6 Light in matter (macroscopic Maxwell's equations)

The above discussion provides a good basis for starting to understand how light propagates in a macroscopic medium. As the incident wave enters the medium, it starts interacting with the charges of its atomic or molecular constituents.[18] Each elementary building unit then acts as a dipole source of radiation oscillating at the same frequency as the incident field. Because all the scattered fields due to the radiating sources affect the behavior of adjacent sources of radiation, we face a ***many-body problem*** of a truly grand scale.[19] It is therefore next to impossible to use the microscopic Maxwell's equations to rigorously describe propagation of light in a solid medium.

The solution is to take use of the ***macroscopic Maxwell's equations***, that are

[18] In the microscopic Maxwell's this interaction enters the equations naturally through ρ and $\mathbf{J}$.
[19] There are already around $N_A \approx 6 \times 10^{23}$ atoms in a 1 cm^3 of ordinary matter.

written as[20]

$$\nabla \cdot \mathbf{D} = \rho_\mathrm{f}\,, \qquad \nabla \cdot \mathbf{B} = 0\,, \tag{4.6.46}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}\,, \qquad \nabla \times \mathbf{H} = \mathbf{J}_\mathrm{f} + \frac{\partial \mathbf{D}}{\partial t}\,, \tag{4.6.47}$$

where ρ_f is the free electric charge density and $\mathbf{J}_\mathrm{f}$ is the free electric current density. We have also taken use of the ***auxiliary fields*** $\mathbf{D}$ and $\mathbf{H}$ defined through equations[21]

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}\,, \qquad \mathbf{H} = \frac{1}{\mu_0}\mathbf{B} - \mathbf{M}\,, \tag{4.6.48}$$

where $\mathbf{D}$ is the electric displacement, $\mathbf{P}$ is material polarization, $\mathbf{H}$ is the magnetizing field and $\mathbf{M}$ is the magnetization. Finally, we connect the macroscopic bound charges ρ_b and currents $\mathbf{J}_\mathrm{b}$ to the material polarization $\mathbf{P}$ and magnetization $\mathbf{M}$ by defining them as

$$\rho_\mathrm{b} = -\nabla \cdot \mathbf{P}\,, \qquad \mathbf{J}_\mathrm{b} = \nabla \times \mathbf{M} + \frac{\partial \mathbf{P}}{\partial t}\,, \tag{4.6.49}$$

where the total, free and bound charges (currents) are related by $\rho = \rho_\mathrm{f} + \rho_\mathrm{b}$ ($\mathbf{J} = \mathbf{J}_\mathrm{f} + \mathbf{J}_\mathrm{b}$). At optical frequencies the magnetic properties of materials are often negligible. Therefore, we will focus here on the equation for the electric displacement $\mathbf{D}$.

Macroscopically, the effect of the material is described by the ***polarization*** of the material.[22] From the microscopic point of view, the polarization $\mathbf{P}$ is defined as the average dipole moment per unit volume

$$\mathbf{P} = \mathbf{p}/V = N\mathbf{p} = N\alpha\epsilon_0\mathbf{E} = \epsilon_0\chi\mathbf{E}\,, \tag{4.6.50}$$

where $\mathbf{p} = \epsilon_0\alpha\mathbf{E}$ is the dipole moment of each microscopic building unit of the material (atoms/molecules/etc.), V is the unit volume, and N is the number density associated with the unit volume. Looking at the above Maxwell's equations, it becomes clear that the electromagnetic wave exiting the medium has somehow been modified by the presence of the medium, because the exiting field is a superposition of the incident wave and all the wavelets scattered by the building units of the medium.[23]

[20] In fact these are closer to the original Maxwell's equations. The microscopic equations were actually formulated by Lorentz in order to calculate the macroscopic properties of materials from their microscopic properties.

[21] More general definition exists for ***bianisotropic*** materials, but is beyond our scope.

[22] Note here that the term polarization has two completely different meanings, the other being the direction of the electric component of the electromagnetic field.

[23] This is quite generally the case how nature works. However, rigorous calculations of such sums is often next to impossible. In some cases, the ***principle of least action*** can be utilized to find the most probable solution.

For reasons to be discussed later, the strongest component of the exiting wave still propagates in the same direction as the incident wave.

The macroscopic polarization of the medium $\mathbf{P}$ tells how much, on the average, the medium deviates from vacuum. On the other hand, the polarization is built up in response to the incident optical field. We therefore define the ***dielectric constant*** or ***permittivity*** of the medium ϵ through the equation[24]

$$\mathbf{P} = (\epsilon - \epsilon_0)\mathbf{E}\,, \tag{4.6.51}$$

where, as before, ϵ_0 is the permittivity of vacuum.

It can be shown that one way to account for the propagation of light in a medium is thus to replace the vacuum permittivity ϵ_0 by the material permittivity ϵ in Maxwell's equations. Similarly, if the medium has magnetic properties, we would replace the vacuum permeability μ_0 by the permeability of the material μ in the Maxwell's equations. It turns out, however, that the magnetic response of most materials at optical frequencies can be neglected. By making these replacements into Maxwell's equations, we can immediately conclude that the speed of light in matter is given by

$$v = 1/\sqrt{\epsilon\mu}\,. \tag{4.6.52}$$

The ***index of refraction*** of the medium is defined as the ratio of the speed of light in vacuum and in the material. It is thus given by

$$n = \frac{c}{v} = \sqrt{\frac{\epsilon\mu}{\epsilon_0\mu_0}} \approx \sqrt{\frac{\epsilon}{\epsilon_0}}\,, \tag{4.6.53}$$

where the last form holds provided that the magnetic properties can be neglected (i.e. $\mu \approx \mu_0$). It is known that the index of refraction of most materials at optical frequencies is larger than unity. However, there are no fundamental reasons why this should always be true. In fact, the index of refraction can be smaller than unity or even negative. These unconventional cases are important research topics of today.

We argued earlier that the frequency of light cannot change when light interacts with materials. From Eq. (3.3.16), the velocity of a harmonic wave is $v = \nu\lambda$. By applying this result for vacuum and a medium, we conclude that the index of refraction n must

[24]Recall that $\mathbf{D} = \epsilon_0\mathbf{E} + \mathbf{P} = \epsilon_0\epsilon_r\mathbf{E} = \epsilon\mathbf{E}$, where ϵ_r is a dimensionless *relative permittivity* while $\epsilon = \epsilon_0\epsilon_r$ is the (material) permittivity. Note also that some authors confusingly drop out the sub-index from the relative permittivity.

change the wavelength of light. Specifically, the wavelength in the medium is

$$\lambda = \lambda_0/n\,, \tag{4.6.54}$$

where λ_0 is the wavelength in vacuum. By recalling from Eq. (3.7.50) that the wave number of a wave is $k = 2\pi/\lambda = \omega/c$ we find the wave number in the material to be

$$k = nk_0 = \frac{n\omega}{c}\,. \tag{4.6.55}$$

Similarly, the wave vector is found to be

$$\mathbf{k} = n\mathbf{k}_0 = \frac{n\omega}{c}\hat{\mathbf{k}}\,, \tag{4.6.56}$$

where k_0 and $\mathbf{k}_0$ are the respective quantities in vacuum.

By applying these results to a harmonic plane wave that propagates in z direction, we find that it is of the form

$$\mathbf{E} = \mathbf{A}\mathrm{e}^{\mathrm{i}(kz-\omega t)} = \mathbf{A}\mathrm{e}^{\mathrm{i}(nk_0 z-\omega t)}\,. \tag{4.6.57}$$

The index of refraction n is therefore seen to influence the spatial evolution of the phase of the propagating wave. We can also define a distance which corresponds to the equivalent distance propagated in vacuum. This quantity is known as the ***optical path length***, which is thus

$$\mathrm{OPL} = nz\,. \tag{4.6.58}$$

Finally we note that the index of refraction of a material is not a constant quantity. Quite the contrary, the index of refraction depends on frequency (wavelength), i.e., $n = n(\omega)$ although we have not explicitly indicated this dependence in the above discussion. This effect is known as ***dispersion***. It is also important to note that if some effect is said to occur at a given wavelength, the reference is made to the vacuum wavelength λ_0, because it is the only constant concept regarding the wavelength. The actual wavelength in a material λ can be very different from λ_0, determined by the index of refraction of the material. As we will see later, dispersion at the optical wavelengths arises from the electronic response of the material to the incident electromagnetic wave. At other wavelength regimes, other mechanisms, such as the motion of ions in a crystal or the tendency of certain molecules to orient along the direction of the electric field, become important.

5. LIGHT-MATTER INTERACTIONS

5.1 Radiation from atoms and molecules

In the optical regime, the dipole radiation is associated with the interaction of light with atoms and molecules. We know that the electrons in atoms and molecules can reside in different energy levels (Fig. 5.1). In most cases, such a system is in the level of lowest energy, known as the ***ground state*** $|g\rangle$.[1] However, in certain situations, the system can end up being in one of the levels with higher energy, so-called ***excited states*** $|i\rangle$. The energy difference between the states can also be expressed in terms of frequency or angular frequency as

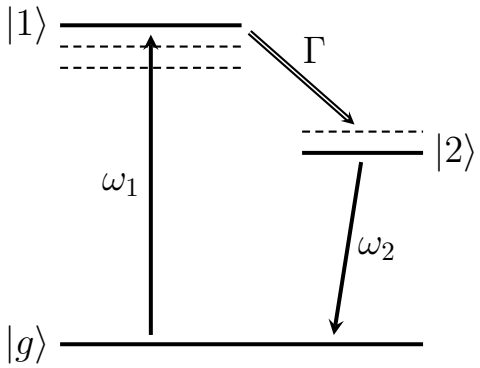

Fig. 5.1: Ground state $|g\rangle$ and excited states $|1\rangle$ and $|2\rangle$ of an atomic system, associated with energies E_1 and E_2.

$$\Delta E = E_i - E_g = h\nu_0 = \hbar\omega_0\,, \tag{5.1.1}$$

where the (angular) frequency ω_0 is known as the ***resonance frequency*** between the two states. It is also important to remember that a system in an excited state tends to relax back to it ground state in one way or another. Such relaxation is usually exponential and characterized by the ***lifetime*** of the excited state. Depending on the system, the lifetime can vary over several orders of magnitude. However, for typical atoms, the lifetime is on the order of 10 ns.

5.2 Basic light-matter interactions

When an atomic or molecular system interacts with light whose frequency is close to the resonance frequency ω_0, there is a possibility that the system becomes excited through ***absorption*** (Fig. 5.2). In this case, one photon from the optical field is destroyed and its energy is used to move the system from the ground state to the excited state. After some time, the system will relax back to the ground state. One way for the relaxation is that the excess energy of the excited state is released through the ***emission*** of a photon with a frequency very close to the resonance frequency. This process is also often described by the term ***spontaneous emission***, because it starts by itself from the need of the excited state to relax.

This process can be qualitatively understood using both the concepts of photons and electromagnetic waves.[2] A typical system has no preferred direction, and the photon

[1] We assume the reader to be familiar with the basics of quantum mechanics and subsequently with the ***bra–ket notation***.

[2] The detailed treatment of these processes is beyond our scope and is left to the more advanced courses focusing on laser physics.

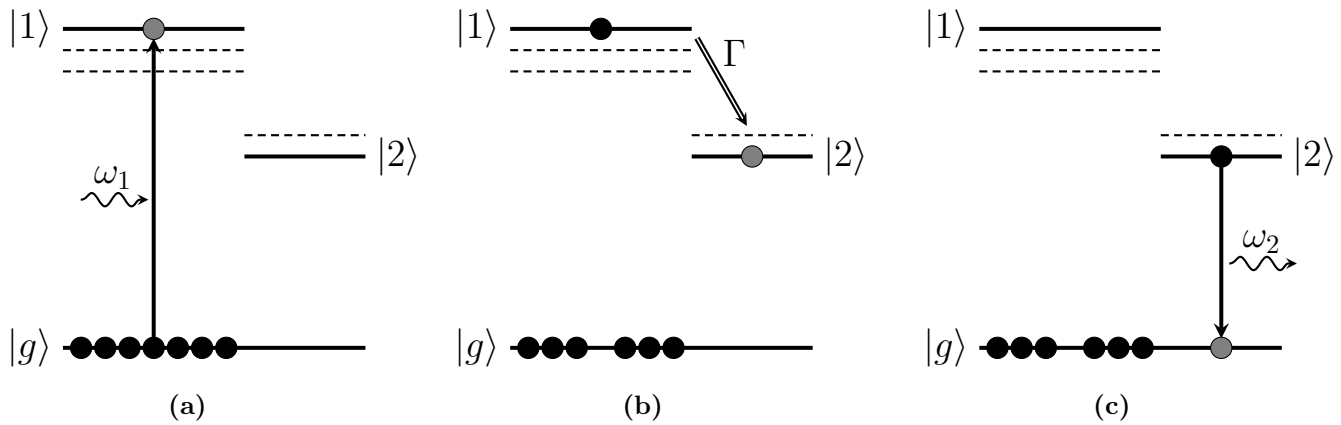

Fig. 5.2: (a) Absorption of a photon from an optical field (ω_1) by the system. (b) The excited state $|1\rangle$ may undergo some internal dynamics, such as relax to a lower-energy excited state $|2\rangle$. (c) After a short time the excited state $|2\rangle$ relaxes back into the ground state $|g\rangle$ by spontaneously emitting another photon oscillating at ω_2.

is therefore emitted in a random direction. In the wave picture, we can assume that a photon corresponds to an elementary wavelet of radiation. In addition, during the relaxation of the system, we assume it to act as a source of dipole radiation. On average, the system will therefore emit a dipole pattern of radiation. Due to the fact that the radiation starts spontaneously, the elementary wavelet is emitted with a random initial phase. Subsequently, after such an absorption–emission cycle, the emitted elementary wavelet thus has no phase correlation with the field that originally excited the system.

It is also possible that when the atomic (or molecular) system is in the excited state, an incident photon at the resonance frequency interacts with the system. In such a case, a process of ***stimulated emission*** may occur, where the incident photon drives the system back to the ground state (Fig. 5.3). In this process, the energy of the excited state is released as a photon, which is now an exact copy of the incident photon with the same frequency, phase, and direction of propagation. After stimulated emission, one incident photon has thus become two identical photons, so that the process can amplify light.

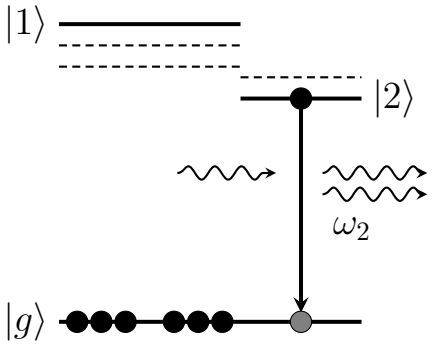

Fig. 5.3: In stimulated emission an incident photon triggers the system to relax from the excited state to the ground state. An identical copy of the incident photon is generated during the process.

A different situation is encountered when the frequency of the incident field is very

different from any of the resonance frequencies of the system. The system cannot then become excited by the field.[3] We now treat the system as consisting of a positively-charged nucleus surrounded by negatively-charged electrons. It is clear that these charged particles must interact somehow with the electromagnetic field. The electrons, being much lighter in weight than the nucleus, are much easier to move by the field, which results in some perturbation in the motion of the electrons (Fig. 5.4). Because the field is oscillating, the perturbation is also oscillating, resulting in a small oscillating dipole moment in the atomic or molecular system. Finally, this dipole moment acts as a source of radiation, giving rise to a spherical elementary wavelet. This whole process is known as ***scattering*** of light by the atomic or molecular system.

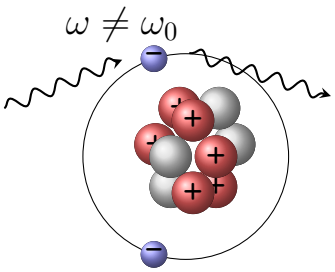

Fig. 5.4: Non-resonant scattering of light by an atom.

It is important to note that in scattering the electrons follow the incident field instantaneously. The phase of the scattered field is therefore correlated with the phase of the incident field although there may be a constant phase shift between the two fields. This phase correlation is the key difference between scattering and the absorption–emission cycle. Furthermore, the direction of the electron oscillation is determined by the polarization of the incident field, giving rise to the angular $\sin\theta$ factor in dipole radiation. For the case of excited atoms or molecules, the dipole can often (but not always) be assumed to have a random orientation, leading to a more spherical emission pattern.

5.3 Lorentz model of an atom (and an electron)

The discussion of the previous section suggests that the optical response of materials is very different depending on whether the optical frequency is close to or far from the resonance frequency of the atomic system. In order to understand how the index of refraction behaves as a function of frequency, we will next introduce the ***Lorentz model*** for the atomic medium. This model is classical and in many ways the simplest possible model that can be used to describe certain optical properties of materials.

[3] Strictly speaking, the system cannot be excited due to a single photon process. However, excitation may occur via ***multiphoton processes***.

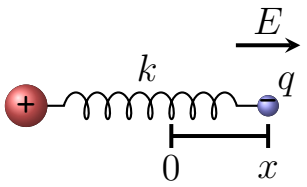

Fig. 5.5: Lorentz model of an atom.

The Lorentz model assumes that a system of an electron bound to an atom is equivalent to a ***harmonic oscillator*** (Fig. 5.5), i.e. to a system where a mass (electron) is attached with a spring to a much heavier mass (nucleus). The simplest possible version of this model is one-dimensional where the electron is taken to move only in x direction. This model is sufficient to understand the behavior of isotropic materials, where the electron moves only in the direction of the electric component of the optical field driving the system. The electric field is thus assumed to be x-polarized.[4]

Since only one direction (here x) is important, we can treat all quantities as scalars. We further assume that at the location of the atom, the electric component of the electromagnetic field oscillates harmonically with respect to time and is of the form

$$E(t) = E_0 \mathrm{e}^{-\mathrm{i}\omega t} \,. \tag{5.3.2}$$

Recalling the ***Lorentz force***, the field gives rise to a force on the electron of form

$$F_E(t) = qE(t) = qE_0 \mathrm{e}^{-\mathrm{i}\omega t} \,, \tag{5.3.3}$$

where q is the charge of the electron. This force will displace the electron from its equilibrium position, which we set to be at $x = 0$.

We further assume that the binding of the electron to the nucleus is described by a spring with spring constant k. When the electron is displaced to location x, the spring will thus exert a restoring force on the electron given by

$$F_R(t) = -kx(t) \,. \tag{5.3.4}$$

By invoking the Newton's laws, the ***equation of motion*** for the electron can be written to be

$$m\ddot{x} + kx = qE_0 \mathrm{e}^{-\mathrm{i}\omega t} \,, \tag{5.3.5}$$

where m is the mass of the electron and dots above the symbol denote derivation over time. By recalling the properties of harmonic oscillators from mechanics and rearranging the terms, we obtain the equation

$$\ddot{x} + \omega_0^2 x = \frac{qE_0}{m} \mathrm{e}^{-\mathrm{i}\omega t} \,, \tag{5.3.6}$$

[4]Note that we neglect here the effect the magnetic field component **B** has on the displacement of the electron.

where $\omega_0 = \sqrt{k/m}$ is the resonance frequency of the harmonic oscillator. This equation can be solved in a number of different ways.[5] However, in the present case we know that the electron is forced to oscillate by the electric field. We therefore expect that the electron will follow the oscillations of the field and therefore use a trial solution

$$x(t) = x_0 \mathrm{e}^{-\mathrm{i}\omega t}\,, \tag{5.3.7}$$

where x_0 is the amplitude of the electronic oscillations. This trial solution is easy to differentiate with respect to time resulting in

$$-\omega^2 x_0 \mathrm{e}^{-\mathrm{i}\omega t} + \omega_0^2 x_0 \mathrm{e}^{-\mathrm{i}\omega t} = \left(qE_0/m\right) \mathrm{e}^{-\mathrm{i}\omega t}\,, \tag{5.3.8}$$

which directly yields the solution for the amplitude of the electronic oscillations as

$$x_0 = \frac{q/m}{\omega_0^2 - \omega^2} E_0\,, \tag{5.3.9}$$

and for the displacement of the electron at time t as

$$x(t) = \frac{q/m}{\omega_0^2 - \omega^2} E_0 \mathrm{e}^{-\mathrm{i}\omega t} = \frac{q/m}{\omega_0^2 - \omega^2} E(t)\,. \tag{5.3.10}$$

This result clearly shows that, as expected, the electron tends to follow the oscillations of the optical field. In addition, the amplitude of the oscillations depends on the difference between the frequency of the field and the resonance frequency of the atom. When the field is close to the resonance of the atom, the oscillations can become very strong.

We next calculate the dipole moment p of the atom. It is known that atoms left at rest possess no dipole moment. Hence, we can assume that the oscillating dipole moment arises from the displacement of the electron from its equilibrium position. The dipole moment of a single atom is thus $p(t) = qx(t)$. Furthermore, the polarization P of a collection of atoms depends on the number density of atoms N (number of atoms per unit volume) and is given by

$$P(t) = Nqx(t) = \frac{(q^2 N/m)}{\omega_0^2 - \omega^2} E(t)\,. \tag{5.3.11}$$

[5] In the case of an arbitrary excitation field (such as a few-cycle ultra-short pulse), e.g. the ***Green's function method*** might provide superior tools for solving the problem.

By recalling Eq. (4.6.51), the above should be equal to $P(t) = (\epsilon - \epsilon_0)E(t)$, allowing us to write the dielectric constant of the medium as

$$\epsilon(\omega) = \epsilon_0 + \frac{q^2 N}{m\left(\omega_0^2 - \omega^2\right)}\,, \tag{5.3.12}$$

where we have emphasized the fact that the dielectric constant depends on frequency. By finally using Eq. (4.6.53), we find that the index of refraction of the material fulfils

$$n^2(\omega) = 1 + \frac{Nq^2}{\epsilon_0 m}\frac{1}{\omega_0^2 - \omega^2}\,. \tag{5.3.13}$$

This result shows that the index of refraction $n(\omega)$ is larger than unity when the frequency of the optical field ω is smaller than the resonance frequency ω_0 of the material. On the other hand, the index is smaller than unity when the frequency of the optical field is larger than the resonance frequency. Finally, it is clear that the result is unphysical on exact resonance ($\omega = \omega_0$) as the result tends to infinity. Our simple model and the equation of motion given by Eq. (5.3.5) does therefore not include all the effects that are important for the physical system.

The problem of Eq. (5.3.5) is that it would leave the atom oscillating forever even when the driving field E is turned off. This is not possible because any realistic system will have some losses. This problem can be resolved by adding to Eq. (5.3.5) a damping (friction) term proportional to the velocity of the electron. The equation of motion then becomes

$$m\ddot{x} + m\gamma\dot{x} + kx = qE_0 \mathrm{e}^{-\mathrm{i}\omega t}\,, \tag{5.3.14}$$

where γ is the damping constant for the electronic oscillations. By solving this equation in exactly the same way as before, we find that the index of refraction follows

$$n^2(\omega) = 1 + \frac{Nq^2}{\epsilon_0 m}\frac{1}{\omega_0^2 - \omega^2 - \mathrm{i}\gamma\omega}\,. \tag{5.3.15}$$

In the presence of damping, the index of refraction becomes thus a complex-valued quantity. It is straightforward to show that the imaginary part of the refractive index leads to absorption of light that propagates in the medium. The real part, on the other hand, is still associated with the harmonic oscillations of the optical field. The Lorentz model contains all the essential effects that are known to occur when light interacts with atoms. In spite of this, it is too simple to describe real atoms. In

particular, real atoms are known to contain an infinite number of possible energy levels and resonance frequencies. Furthermore, transitions between different energy levels vary in strength, i.e., some transitions are much more likely to happen than others.[6] The Lorentz model, on the other hand, assumes that the atom has only one electron and one resonance frequency.

In order to generalize the Lorentz model to more realistic cases, we assume that the electron is distributed over several different transitions, each with a certain probability, and that the index of refraction is obtained by summing the contributions from all possible transitions.[7] This results in the following relation for the refractive index

$$n^2(\omega) = 1 + \frac{Nq^2}{\epsilon_0 m} \sum_j \frac{f_j}{\omega_{0,j}^2 - \omega^2 - \mathrm{i}\gamma_j \omega}\,, \quad (5.3.16)$$

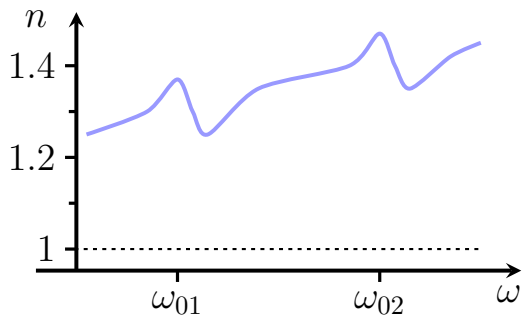

Fig. 5.6: General behavior of the real part of the index of refraction as a function of angular frequency ω.

where j labels the transitions and f_j is the ***oscillator strength*** describing the probability of the corresponding transition. This equation is also known as ***Helmholtz–Ketteler formula*** or Kramers–Heisenberg dielectric function.[8] Note that both the resonance frequencies $\omega_{0,j}$ and damping constants γ_j of the various transitions can depend on j. Note also that

$$\sum_j f_j = 1\,, \quad (5.3.17)$$

because the oscillator strengths are related to probabilities.

The general behavior of (the real part of) the index of refraction as a function of frequency or wavelength can be approximately understood by considering Eq. (5.3.16).[9] First, when the frequency of the field is much smaller than any resonance frequency,

[6] These are dictated by the symmetry properties of the quantum states associated to the transitions giving rise to ***selection rules***.

[7] Appreciate here, that our classical approach is quickly drifting towards quantum mechanics. The state of the electron is no longer a single state, but in fact a superposition of many possible states.

[8] Interestingly, this approach can be further generalized to include many electrons that are coupled to each other forming a ***solid state system***. This generalization explains why some materials are spatially dispersive, i.e., their refractive indices depend also on $\mathbf{k}$. This is clearly explained in the book *Semiconductor optics* by Klingshirn.

[9] Closely related method to treat the chromatic material dispersion is to use the ***Sellmeyer equation***. However, the Sellmeyer equation fails for materials close to their material resonances, whereas the Helmholtz-Ketteler formula often remains valid.

the index deviates from unity only very little. When the field frequency approaches the lowest resonance frequency, it first grows. However, after passing the resonance, the index becomes lower. Similarly, in the neighborhood of any resonance frequency, its contribution to the index of refraction is the most important. By taking all these factors into account, we find that the general trend in the index of refraction is as shown in Fig. 5.6.

In traditional optics, we are mainly interested in materials that are transparent for optical wavelengths, especially the visible ones (380–740 nm). The transparency arises from the fact that all the possible transitions of such materials occur at wavelengths shorter than the optical wavelengths. Usually, the first transitions occur at the ultraviolet wavelength range but these transitions affect the index even in the visible range. In consequence, the index of refraction of transparent materials tends to grow for shorter wavelengths (Fig. 5.7). This effect is known as ***normal dispersion***. In contrast, the opposite tendency near a resonance is known as ***anomalous dispersion*** (e.g. resonance regions in Fig. 5.6).

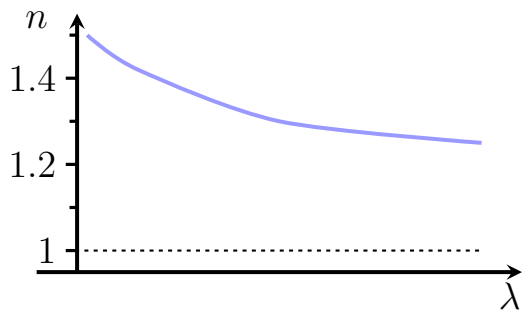

Fig. 5.7: Normal dispersion of the index of refraction in transparent materials as a function of wavelength.

5.4 Laser principle

One of the most spectacular results of detailed understanding of light-matter interactions are ***lasers***. In order to understand their basic operation principle, we return to the processes described in Section 5.1. More specifically, we focus on absorption and stimulated emission, which are the two processes that can occur in response of the atomic medium to incident light.

It is clear that absorption is a loss mechanism for light because it removes photons from the incident light beam. Stimulated emission, on the other hand, is a ***gain*** mechanism because it adds photons to the incident beam. The question as to whether loss or gain dominates depends on whether we have more atoms in the ground or excited state.

In thermal equilibrium, the ***populations*** (i.e., number densities) of the ground state

N_1 and excited state N_2 obey the ***Boltzmann distribution***, which yields

$$\frac{N_2}{N_1} = \exp\left(-h\nu_0/k_BT\right) , \tag{5.4.18}$$

where h is the Planck constant, $k_B = 1.380 \times 10^{-23}$ J/K is the Boltzmann constant, ν_0 is the oscillation frequency related to the energy difference of the respective populations and T is the temperature. It is evident that at room temperature (300 K) and for optical frequencies (10^{15} Hz), most of the atoms reside in the ground state (Fig. 5.8a). The probability of absorption is therefore orders of magnitude higher than that of stimulated emission and loss dominates.

In order to achieve gain, we therefore need to break the thermal equilibrium. This is done by ***pumping*** the system, i.e., providing energy from outside to the atomic medium in order to achieve ***population inversion***, during which the majority of atoms populate the excited state (Fig. 5.8b). Such a medium can then act as an ***amplifier*** for an optical beam passing through the medium.[10]

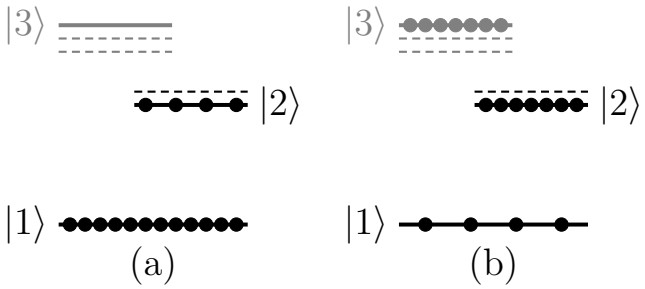

Fig. 5.8: Schematic of the ground and excited state populations in (a) thermal equilibrium and (b) under population inversion.

In order to achieve laser oscillation, we need to provide feedback to the system. This is achieved by placing the gain medium into an optical cavity, which is a Fabry–Pérot resonator (see Section 8.6). Inside the cavity, light is forced to pass through the gain medium multiple times, which provides additional amplification of light oscillating at the resonance frequencies of the Fabry–Pérot cavity (Fig. 5.9).

With sufficient gain, the system can become self-starting, i.e., a single photon spontaneously emitted along the axis of the cavity can initiate the amplification process. The laser output is provided by the small fraction of light that is transmitted through one of the cavity mirrors.

[10] Amplifiers themselves are also extremely useful instruments. They e.g. provide means to make ***more powerful lasers***, act as ***repeaters*** for amplifying communication signals propagating in optical fibers or enable development of light sources oscillating at new frequencies known as ***optical parametric oscillators***.

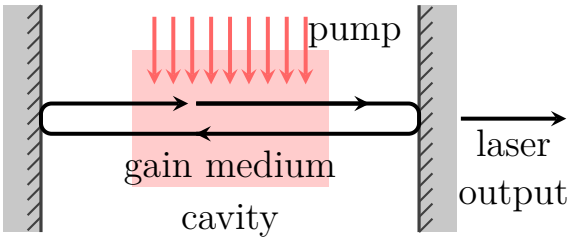

Fig. 5.9: Laser (oscillator) principle. Pumping provides population inversion and optical gain. The light inside the cavity is amplified by propagating back and forth through the gain medium.

6. PROPAGATION

6.1 Wave fronts and rays

In the following, we will use two complementary concepts to describe the propagation of light in various cases. As already defined in Section 3.7, wavefronts are surfaces of constant phase, which propagate at the speed of c/n, i.e., the vacuum speed c is modified by the index of refraction n. More specifically, wavefronts are surfaces whose normal is defined by the wave vector.

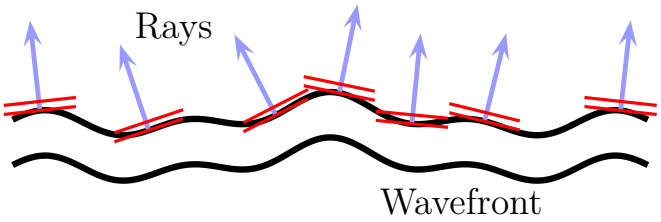

Fig. 6.1: In isotropic materials, rays are locally normal to wavefronts.

It is often cumbersome to draw figures where all the wavefronts are shown. This problem can be avoided by using the concept of a ***ray***, as we will later see in the discussion of geometrical optics in Chapter 10. By definition, rays describe the direction of propagation of radiative energy. In isotropic materials, the energy must propagate in the direction of the wave vector. Consequently, rays are orthogonal to the wave fronts (Fig. 6.1).

6.2 Phenomenology of transmission and reflection

In Section 4.6, we argued that the propagation of light in matter can be described by the index of refraction n, which modifies the spatial dependence of a harmonic wave. Furthermore, it can be shown that in dense and/or homogeneous materials, this is essentially the only effect that needs to be considered. In dilute and/or inhomogeneous materials, the situation is more complicated and one needs to pay attention to how individual atoms and molecules scatter light in different directions, as we discussed in the context of the color of the sky.

We also made a simple link between the atomic-level properties of the material (molecular dipole moment $\mathbf{p} = \epsilon_0 \alpha \mathbf{E}$) and the macroscopic properties (polarization $\mathbf{P}$, dielectric constant ϵ, and index of refraction n). A more detailed analysis between the atomic and macroscopic properties in any material must be treated with great care in order to account for all the relevant effects properly, but this is not a topic of the present discussion.

An interface between two materials is not homogeneous even if the two materials separately are. We therefore must consider at least the possibility that the interface transmits only part of an incident wave and reflects the remaining part. In fact, this is what does happen when light is incident on an interface between two materials with different indices of refraction. However, there are several different approaches to

understand even the basic properties of refraction and reflection.[1]

To make our approach specific, we consider the situations of Figs. 6.2 and 6.3. We assume that an incident plane wave encounters a planar interface between materials i (incident) and t (transmitted) with respective indices of refraction n_i and n_t. The wave fronts of the incident wave make an angle θ_i with the plane of the interface. In consequence, the rays of the wave make the same angle with respect to the surface normal. This angle is known as the ***angle of incidence***. We will now follow wave front AB in time. Different parts of this wave front encounter the interface at different times. We assume that the interaction with the interface gives rise to reflected and transmitted waves, which propagate at the speeds determined by the refractive indices of the two materials. In addition, we require that the wave fronts remain continuous at the interface.

We will first consider the reflected wave and focus on its wave front DC, whose direction of propagation is determined by the angle of reflection θ_r. A front is defined by constant phase. In order for DC to be AB at a later time, the propagation times from A to C and from B to D must be equal. Both the incident and reflected waves propagate in the same material and with the same speed. We therefore obtain the requirement that the line segments AC and BD must be of equal length. These segments can be connected through the segment AD, yielding the lengths

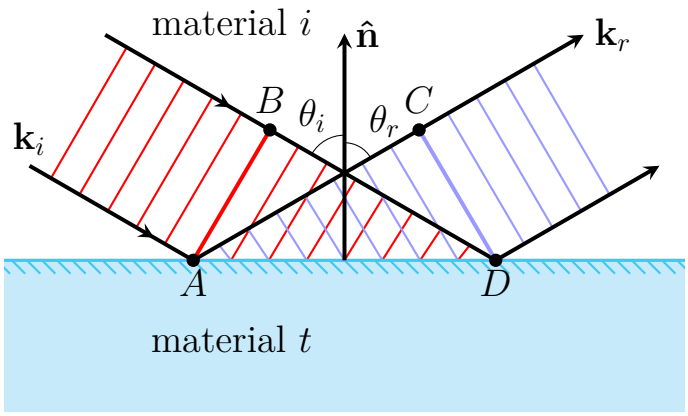

Fig. 6.2: Generation of the reflected wave at the interface between two materials. For clarity, the incident wave is shown by red and the reflected wave by blue.

$$|BD| = |AD| \sin\theta_i \,, \tag{6.2.1}$$

$$|AC| = |AD| \sin\theta_r \,. \tag{6.2.2}$$

By requiring these two lengths to be equal, we obtain the reflection law

$$\theta_i = \theta_r \,. \tag{6.2.3}$$

[1] These basic properties are beautifully described using the QED and path integral framework by R. Feynman in some of his popular books on the ***quantum electrodynamics (QED)***.

The angles of incidence and reflection are thus equal. In addition, we may conclude that the incident and reflected rays must be in the same plane with the surface normal. This plane is known as the ***plane of incidence***.

The case of the transmitted wave can be treated in a very similar way by focusing on wave fronts AB and DE (Fig. 6.3). The key difference compared to the previous is that the incident and transmitted waves propagate in materials with different indices of refraction. The speeds of the waves are thus $v_i = c/n_i$ and $v_t = c/n_t$, respectively. The requirement of equal times from A to E and from B to D therefore gives the condition

$$|BD|n_i = |AE|n_t\,. \tag{6.2.4}$$

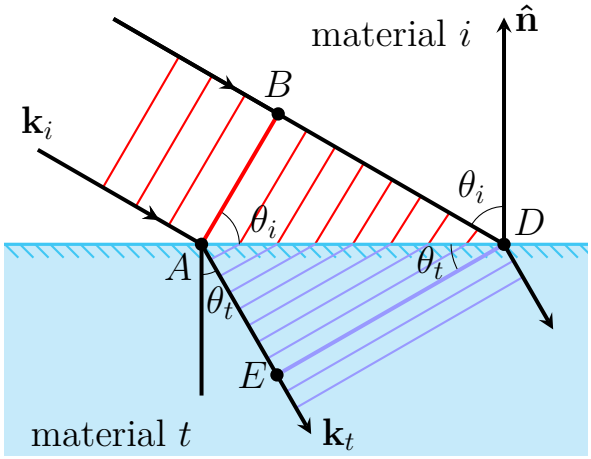

Fig. 6.3: Refraction at the interface between two materials. For clarity, the incident (transmitted) wave is shown by red (blue).

In addition, the line segment AE has the length

$$|AE| = |AD| \sin\theta_t\,. \tag{6.2.5}$$

By combining these results we obtain the refraction law

$$n_i \sin\theta_i = n_t \sin\theta_t\,. \tag{6.2.6}$$

The word refraction here refers to the fact that the transmitted wave is characterized by a different direction of propagation than the incident wave. Consequently, the transmitted wave can also be called as the ***refracted wave***. The refraction law is also known as ***Snell's law***.

As an example, we may consider the case where the material 't' has a higher index of refraction than material 'i' ($n_t > n_i$). This implies that the angle of refraction θ_t is smaller than the angle of incidence θ_i. This is what happens, e.g., when light enters glass from air and leads to the familiar statement that the rays are refracted towards the surface normal as light enters an optically denser medium. This is also the example shown in Fig. 6.3.

The laws of reflection and refraction can be derived also in other ways. One way is based on the requirement of momentum conservation of photons. A planar interface

can account for momentum mismatch only in the direction of its normal but not along the interface which is translationally invariant. Consequently, the momentum of the photons must be conserved along the interface.

For a completely different way, we first have to generalize the concept of the optical path length. As discussed in Section 4.6, the optical path length is the vacuum equivalent path between two points. For example, if a ray of light propagates a length s_1 in a medium with the index of refraction n_1 and length s_2 in a medium with the index of refraction n_2, the total time required for the trip is

$$t = \frac{s_1}{v_1} + \frac{s_2}{v_2} = \frac{n_1 s_1}{c} + \frac{n_2 s_2}{c} = \frac{n_1 s_1 + n_2 s_2}{c}\,. \tag{6.2.7}$$

The optical path length (OPL) is thus

$$\mathrm{OPL} = n_1 s_1 + n_2 s_2\,. \tag{6.2.8}$$

This is straightforward to generalize for more materials or even for the case where the index of refraction varies continuously along the path S taken by the ray. These results can be expressed as

$$\mathrm{OPL} = \sum_i n_i s_i = \int_S n(s)\,\mathrm{d}s\,. \tag{6.2.9}$$

According to ***Fermat's principle***,[2] from all the possible paths between two points, a ray of light takes the one for which the optical path length is stationary, i.e.,

$$\frac{\partial(\mathrm{OPL(S)})}{\partial S} = 0\,. \tag{6.2.10}$$

Usually, the stationary optical path is also the shortest one, but there are very special cases where this is not true. This principle is also a common way to derive the laws of reflection and refraction.

6.3 Electromagnetic theory of reflection and refraction

In the previous section, we discussed several ways to derive the laws of reflection and refraction given by Eqs. (6.2.3) and (6.2.6), respectively. These laws relate the angles of incidence, reflection, and refraction to each other. However, these laws only tell in which directions the various waves propagate but do not tell how much of the incident light is reflected or transmitted. In order to understand these issues, we need to carry out a full electromagnetic calculation.

[2]Or relatedly by the ***principle of least action***.

We consider the situation shown in Fig. 6.4. A plane wave $\mathbf{E}_i$ is incident on the interface between two isotropic materials with the indices of refraction as n_i and n_t. On the basis of the previous discussion, we expect that this gives rise to reflected and refracted (transmitted) waves $\mathbf{E}_r$ and $\mathbf{E}_t$, respectively. The directions of propagation of the fields, as shown by rays, are characterized by the angles of incidence (θ_i), reflection (θ_r), and refraction (θ_t). The interface normal $\hat{\mathbf{n}}$ points in z direction and the interface is located at $z = b$. We assume that the two materials are non-magnetic and thus characterized by the vacuum permeability μ_0. The three waves are therefore of the forms

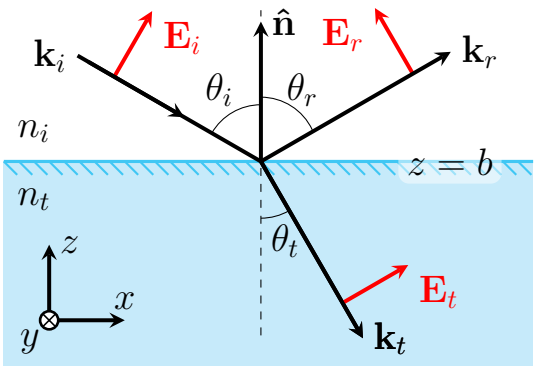

Fig. 6.4: Geometry for the electromagnetic theory of reflection and refraction.

$$\mathbf{E}_i = \mathbf{A}_i e^{i(\mathbf{k}_i\cdot\mathbf{r}-\omega_i t)}\,, \qquad \mathbf{E}_r = \mathbf{A}_r e^{i(\mathbf{k}_r\cdot\mathbf{r}-\omega_r t)}\,, \qquad \mathbf{E}_t = \mathbf{A}_t e^{i(\mathbf{k}_t\cdot\mathbf{r}-\omega_t t)}\,, \tag{6.3.11}$$

where the amplitudes can be complex-valued quantities. Note that so far, we have not made any strong assumptions about the waves. They can propagate in arbitrary directions and their electric fields can point in any direction. In addition, we allow for the possibility that the waves oscillate at different frequencies.

We can now derive the interface conditions for the above fields at the interface. Starting from the integral form of the Maxwell–Faraday law[3]

$$\oint_{\partial S} \mathbf{E}\cdot \mathrm{d}\boldsymbol{\ell} = -\iint_S \frac{\partial \mathbf{B}}{\partial t}\cdot\hat{\mathbf{s}}\,\mathrm{d}A\,, \tag{6.3.12}$$

where we take the surface S to be a small square across the interface, $\hat{\mathbf{s}}$ to be a normal vector of S, and $\partial S = \boldsymbol{\ell}_1 + \boldsymbol{\ell}_2 + \boldsymbol{\ell}_3 + \boldsymbol{\ell}_4$ to be the boundary of the surface S (Fig. 6.5). By shrinking the line segments piercing the interface ($\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_3$) to infinitesimal lengths, the integration area of the right-hand side equation approaches zero

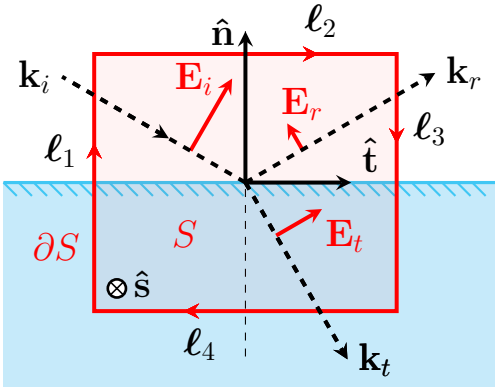

Fig. 6.5: Interface conditions for the electromagnetic field.

$$\lim_{\|\boldsymbol{\ell}_1\|,\|\boldsymbol{\ell}_3\|\to 0} S = 0\,. \tag{6.3.13}$$

[3]Recall that the differential form of the Maxwell–Faraday law, $\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$, can be transformed into the integral form by using the ***Kelvin–Stokes theorem***.

Noting that the integrand $\frac{\partial \mathbf{B}}{\partial t} \cdot \hat{\mathbf{s}}$ remains finite when approaching this limit, the whole right-hand side vanishes leaving us with the line integral

$$\oint_{\partial S} \mathbf{E} \cdot d\boldsymbol{\ell} = 0\,. \tag{6.3.14}$$

Because we took the line segments $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_3$ to be infinitesimally small, this simplifies further into

$$\int_{\ell_2} \mathbf{E} \cdot \,\mathrm{d}\boldsymbol{\ell} + \int_{\ell_4} \mathbf{E} \cdot \,\mathrm{d}\boldsymbol{\ell} = 0\,. \tag{6.3.15}$$

By noting that our square S is small, we can assume the field $\mathbf{E}$ to be piece-wise constant-valued over the line segments $\boldsymbol{\ell}_2$ and $\boldsymbol{\ell}_4$, in other words $\mathbf{E}(\boldsymbol{\ell}_2) = \mathbf{E}_i + \mathbf{E}_r$ and $\mathbf{E}(\boldsymbol{\ell}_4) = \mathbf{E}_t$. By then defining the unit direction of the line segment $\boldsymbol{\ell}_2$ to be $\hat{\mathbf{t}}$, we see the unit direction of $\boldsymbol{\ell}_4$ is respectively $-\hat{\mathbf{t}}$. We also note that the above line segments are equally long ($\|\boldsymbol{\ell}_2\| = \|\boldsymbol{\ell}_4\| = l$). Using these assumptions, the line integral turns into

$$(\mathbf{E}_i + \mathbf{E}_r) \cdot \hat{\mathbf{t}} \int_{\ell_2} \mathrm{d}\ell - \mathbf{E}_t \cdot \hat{\mathbf{t}} \int_{\ell_4} \mathrm{d}\ell = \Big((\mathbf{E}_i + \mathbf{E}_r) \cdot \hat{\mathbf{t}}\Big)\, l - \Big(\mathbf{E}_t \cdot \hat{\mathbf{t}}\Big)\, l = 0\,. \tag{6.3.16}$$

Dividing the above equation by l and re-arranging terms we find the condition

$$(\mathbf{E}_i + \mathbf{E}_r) \cdot \hat{\mathbf{t}} = \mathbf{E}_t \cdot \hat{\mathbf{t}}\,. \tag{6.3.17}$$

This condition holds for an arbitrary tangential direction $\hat{\mathbf{t}}$. Therefore, the tangential components of the fields have to be equal, in other words, the tangential components of the electric field $\mathbf{E}$ are continuous across the interface.

The above condition is often described by using the surface normal $\hat{\mathbf{n}}$. By recalling the geometric definition of the dot product, the above equality implies that the sum of the fields has to point along the surface normal $\hat{\mathbf{n}}$, resulting in the preferred representations of the condition

$$(\mathbf{E}_i + \mathbf{E}_r) \times \hat{\mathbf{n}} = \mathbf{E}_t \times \hat{\mathbf{n}}\,, \tag{6.3.18}$$

$$(\mathbf{E}_i + \mathbf{E}_r - \mathbf{E}_t) \times \hat{\mathbf{n}} = \mathbf{0}\,. \tag{6.3.19}$$

The above condition must hold at plane $z = b$ at all times. This can only be true if all the exponents are equal at plane $z = b$ at all times, i.e.,

$$\mathbf{k}_i \cdot \mathbf{r} - \omega_i t = \mathbf{k}_r \cdot \mathbf{r} - \omega_r t = \mathbf{k}_t \cdot \mathbf{r} - \omega_t t\,, \tag{6.3.20}$$

Suppose that this condition is valid at a given point on plane $z = b$ at a given time t. It is then evident that this condition can remain valid for later times only if all frequencies are equal

$$\omega_i = \omega_r = \omega_t \,. \tag{6.3.21}$$

This result is in full agreement with our earlier discussions, which showed that the frequency of light cannot change when light interacts with matter.

We thus obtain a new condition that must hold at plane $z = b$,

$$\mathbf{k}_i \cdot \mathbf{r} = \mathbf{k}_r \cdot \mathbf{r} = \mathbf{k}_t \cdot \mathbf{r} \,. \tag{6.3.22}$$

We will first consider the requirement

$$(\mathbf{k}_i - \mathbf{k}_r) \cdot \mathbf{r} = 0 \,, \tag{6.3.23}$$

for any point $\mathbf{r}$ on plane $z = b$. This condition implies that the vector $\mathbf{k}_i - \mathbf{k}_r$ is orthogonal to plane $z = b$, i.e., the vector is parallel to the surface normal $\hat{\mathbf{n}}$. Consequently, the vectors $\mathbf{k}_i$, $\mathbf{k}_r$, and $\hat{\mathbf{n}}$ are in the same plane, which is the ***plane of incidence***.

The above condition also implies that[4]

$$\hat{\mathbf{n}} \times (\mathbf{k}_i - \mathbf{k}_r) = \mathbf{0} \,. \tag{6.3.24}$$

By calculating the above vector products, we then obtain

$$k_i \sin(\pi - \theta_i) = k_r \sin\theta_r \,, \tag{6.3.25}$$

The incident and reflected wave are in the same medium, hence, their wave numbers k_i and k_r are equal. We thus obtain the condition

$$\sin\theta_i = \sin\theta_r \,, \tag{6.3.26}$$

and finally the reflection law as before

$$\theta_i = \theta_r \,. \tag{6.3.27}$$

We next consider the requirement for the incident and transmitted waves, which is

$$(\mathbf{k}_i - \mathbf{k}_t) \cdot \mathbf{r} = 0 \,, \tag{6.3.28}$$

[4]Recall the geometric definition of the cross product.

for any point $\mathbf{r}$ on plane $z = b$. This calculation proceeds exactly the same way as before until the condition

$$k_i \sin(\pi - \theta_i) = k_t \sin(\pi - \theta_t)\,, \tag{6.3.29}$$

However, the incident and transmitted waves are in materials with different indices of refraction. Their wave numbers are thus $k_i = n_i\omega/c$ and $k_t = n_t\omega/c$, respectively. This finally yields the refraction law as before

$$n_i \sin\theta_i = n_t \sin\theta_t\,. \tag{6.3.30}$$

In order to obtain further information, we have to treat separately the cases where the electric fields are in the plane of incidence or orthogonal to it. Since the harmonic parts of all the waves are by now guaranteed to be equal at plane $z = b$, we will apply the continuity conditions to the wave amplitudes.

6.4 Fresnel coefficients

We start by considering the case where the electric field $\mathbf{E}$ is normal to the plane of incidence (see Fig. 6.6a). The tangential components of the electric field $\mathbf{E}$ and the magnetic field $\mathbf{B}$ for the incident, reflected and transmitted waves are then written as

$$E_i\,, \qquad E_r\,, \qquad E_t\,, \tag{6.4.31}$$

$$-B_i\cos\theta_i\,, \qquad B_r\cos\theta_i\,, \qquad -B_t\cos\theta_t\,. \tag{6.4.32}$$

Due to the fact that we have assumed non-magnetic materials, we can apply the continuity requirements directly to the magnetic field $\mathbf{B}$. The conditions are thus

$$E_i + E_r = E_t\,, \tag{6.4.33}$$

$$-B_i\cos\theta_i + B_r\cos\theta_i = -B_t\cos\theta_t\,. \tag{6.4.34}$$

In general $B = (n/c)E$ and Eq. (6.4.34) becomes

$$-n_iE_i\cos\theta_i + n_iE_r\cos\theta_i = -n_tE_t\cos\theta_t\,. \tag{6.4.35}$$

We therefore need to solve from Eqs. (6.4.33) and (6.4.35) the reflected and transmitted fields as functions of the incident field. The solution is

$$E_r = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t}E_i\,, \tag{6.4.36}$$

$$E_t = \frac{2n_i\cos\theta_i}{n_i\cos\theta_i + n_t\cos\theta_t}E_i\,. \tag{6.4.37}$$

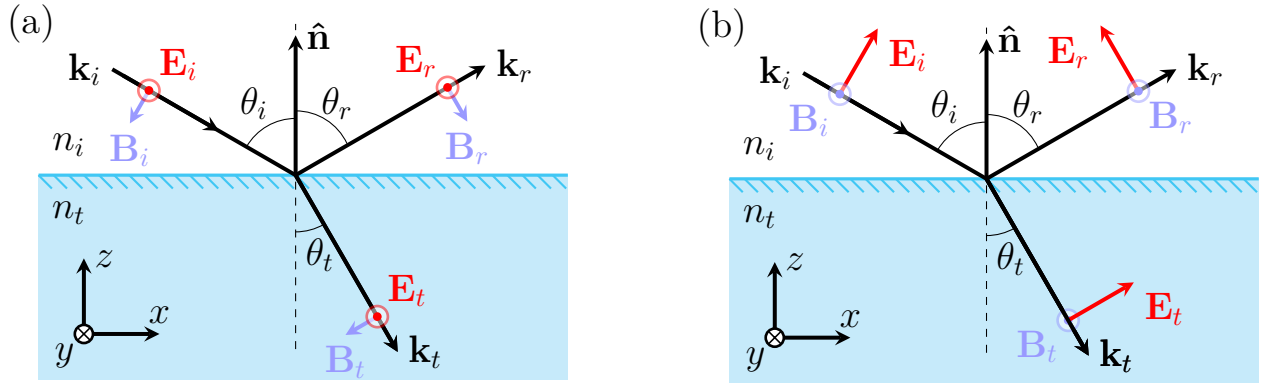

Fig. 6.6: Chosen field directions for **E** and **B** (a) normal to the plane of incidence and (b) in the plane of incidence.

As expected, the reflected and transmitted fields scale linearly with the incident field. This allows us to define the reflection and transmission coefficients for the field amplitudes as follows

$$r_\perp = \frac{E_r}{E_i} = \frac{n_i \cos\theta_i - n_t \cos\theta_t}{n_i \cos\theta_i + n_t \cos\theta_t}\,, \tag{6.4.38}$$

$$t_\perp = \frac{E_t}{E_i} = \frac{2n_i \cos\theta_i}{n_i \cos\theta_i + n_t \cos\theta_t}\,, \tag{6.4.39}$$

where the subscript $\perp$ emphasizes that the results are valid for the present choice of the field polarization.[5]

We next perform a similar calculation for the case where the electric field is in the plane of incidence (Fig. 6.6b). The tangential components of the fields are now

$$E_i \cos\theta_i\,, \qquad -E_r \cos\theta_i\,, \qquad E_t \cos\theta_t\,, \tag{6.4.40}$$

$$B_i\,, \qquad B_r\,, \qquad B_t\,, \tag{6.4.41}$$

so that the continuity equations become

$$E_i \cos\theta_i - E_r \cos\theta_i = E_t \cos\theta_t\,, \tag{6.4.42}$$

$$B_i + B_r = B_t\,, \tag{6.4.43}$$

These equations can be solved exactly in the same way as the previous ones, yielding

[5] Another common notation is to use letter s (*senkrecht*) as a sub-index, as in r_s or t_s.

the reflection and transmission coefficients for the ||-polarization[6]

$$r_{||} = \frac{E_r}{E_i} = \frac{n_t \cos\theta_i - n_i \cos\theta_t}{n_t \cos\theta_i + n_i \cos\theta_t}\,, \tag{6.4.44}$$

$$t_{||} = \frac{E_t}{E_i} = \frac{2n_i \cos\theta_i}{n_t \cos\theta_i + n_i \cos\theta_t}\,. \tag{6.4.45}$$

Eqs. (6.4.38), (6.4.39), (6.4.44) and (6.4.45) are known as the ***Fresnel equations***. It is important to note that the form of these equations depends on the choice of the field directions in Fig. 6.6. Our choice, however, is the most natural one in the sense that we have chosen the fields (**E** or **B**) that are normal to the plane of incidence to point in the same direction. Note also that all quantities in the Fresnel equations can be complex, leading to complex-valued transmission and reflection coefficients. The coefficients thus also contain information about possible phase shifts between the fields. We will see in a while that some of the quantities can become complex even when the refractive indices are real.

As a simple example, we consider the common interface between air and glass. Air can be taken to have unity refractive index, whereas typical glasses have a refractive index of about 1.5. For now, we limit ourselves to cases where all quantities in Eqs. (6.4.38), (6.4.39), (6.4.44) and (6.4.45) remain real.

The Fresnel reflection and transmission coefficients are shown in Fig. 6.7 as functions of the angle of incidence for the interface and for light incident from either side of the interface. The index of refraction of glass n is taken to be 1.5, which is a good approximation for several typical glasses. When light is incident from air (air–glass interface), we see a general trend that both transmission coefficients become smaller when the angle of incidence is increased.

However, both maintain their positive sign for all angles of incidence. The reflection coefficient $r_\perp$ is always negative and its magnitude increases as a function of the angle of incidence. The reflection coefficient $r_{||}$, on the other hand starts from a positive value, passes through zero, and then further becomes more negative for large angles of incidence. When light is incident from the glass side (glass–air interface), all curves end at the angle of incidence of 41.81°, implying that something unexpected happens for larger angles. It turns out that in this regime, the coefficients become complex-valued and this special situation will be discussed in upcoming Section 6.6. However,

[6]The above-mentioned alternative notation would use the letter p (*parallel*) as a sub-index, as in r_p or t_p.

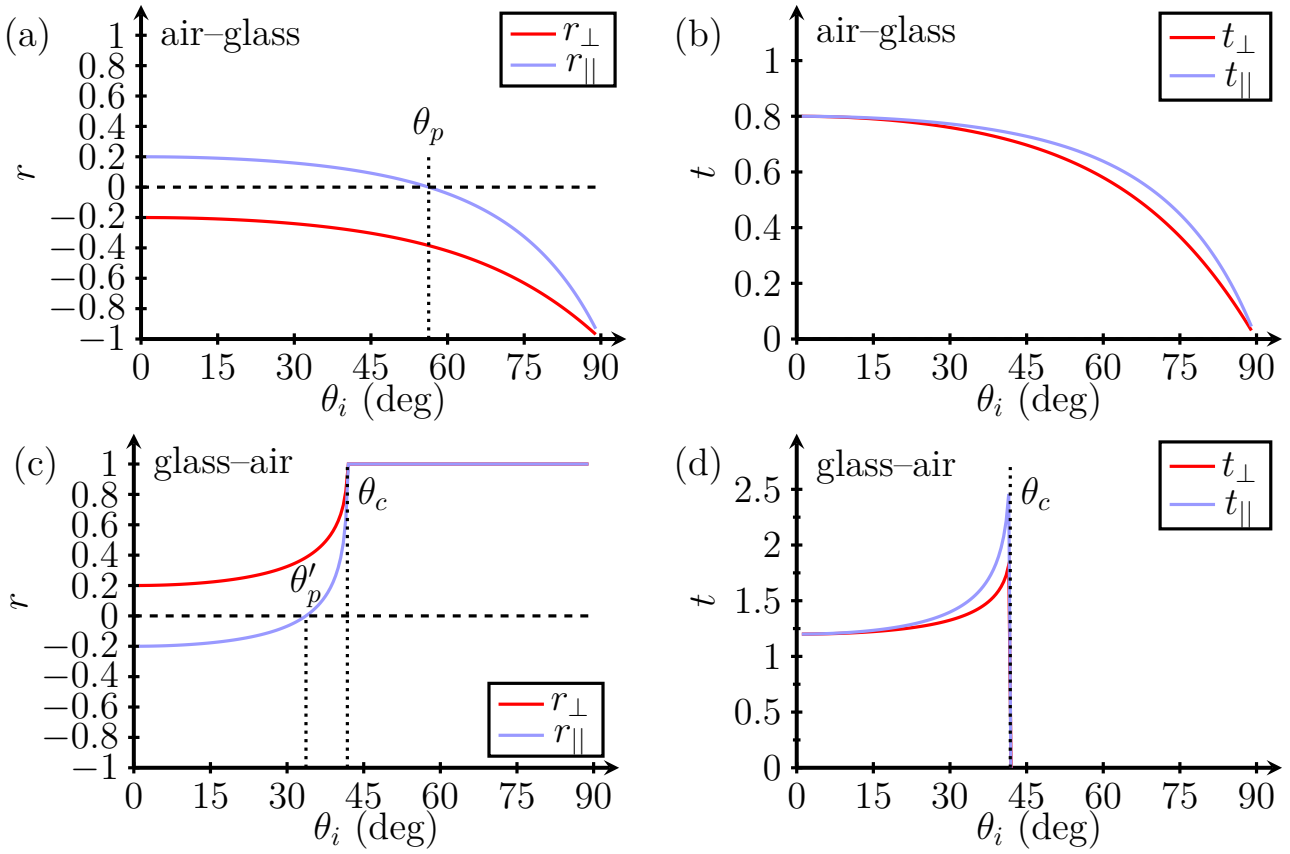

Fig. 6.7: The Fresnel reflection and transmission coefficients for the interface between air and glass. Reflection (a) and transmission (b) coefficients for light incident from air (air–glass interface). Reflection (c) and transmission (d) coefficients for light incident from glass (glass–air interface). The glass is assumed to have refractive index of 1.5. Note the very different vertical scales for each plot.

for the angles of incidence shown, with the exception of $r_\parallel$, which again changes sign, all quantities are positive, and both the reflection and transmission coefficients increase as the angle of incidence grows.

All these results can be understood qualitatively by considering the form of the Fresnel coefficients. It is clear that the transmission coefficients $t_\perp$ and $t_\parallel$ are always positive. Hence, the fields experience no phase shifts as they pass through the interface. To understand how the reflection coefficients behave, we first need to consider the implications of the refraction law for the angles of incidence and refraction. The result is that the angle is small in the medium with high index of refraction. Consequently, the cosine of the angle is large when the index is large. The numerator of coefficient $r_\perp$ thus combines a large index with a large cosine and a small index with a small cosine. This coefficient will therefore not change its sign for any angle

of incidence.

The coefficient $r_{||}$ (see Eq. (6.4.44)), on the other hand, mixes large and small quantities and can therefore change sign. This occurs when the denominator vanishes

$$n_t \cos\theta_i - n_i \cos\theta_t = 0\,. \tag{6.4.46}$$

From here we obtain a different condition

$$n_t \sin\theta_i \cos\theta_i = n_i \sin\theta_i \cos\theta_t = n_t \sin\theta_t \cos\theta_t\,, \tag{6.4.47}$$

where Snell's law was used between the two last forms. The requirement is thus

$$\sin\theta_i \cos\theta_i = \sin\theta_t \cos\theta_t\,. \tag{6.4.48}$$

This is only possible when the angles of incidence and refraction are each others' complement angles, i.e.,

$$\theta_i + \theta_t = 90^\circ\,. \tag{6.4.49}$$

When this condition is fulfilled, the angle of incidence is known as ***Brewster's angle*** or polarization angle θ_p. The implication is that the $||$-polarization of the incident field is not reflected at all, i.e., the reflected field is entirely $\perp$-polarized.[7]

6.5 Reflectivity and transmissivity

The Fresnel coefficients were derived by requiring that the fields fulfil the continuity conditions at each point of the interface. It is also important to know how electromagnetic energy is reflected and transmitted through the interface. To do this properly, we need to consider the amount of energy that crosses a unit area in unit time (Fig. 6.8). We first need to adapt the expression for the irradiance, Eq. (4.3.36), for materials. This is done by replacing $c \to v$, $\epsilon_0 \to \epsilon$ and $\mu_0 \to \mu$. However, we have assumed that our materials have no magnetic properties so that $\mu = \mu_0$. In addition, $v = c/n$ and $\epsilon = n^2\epsilon_0$. We therefore find that the irradiance within a material with the refractive index of n is

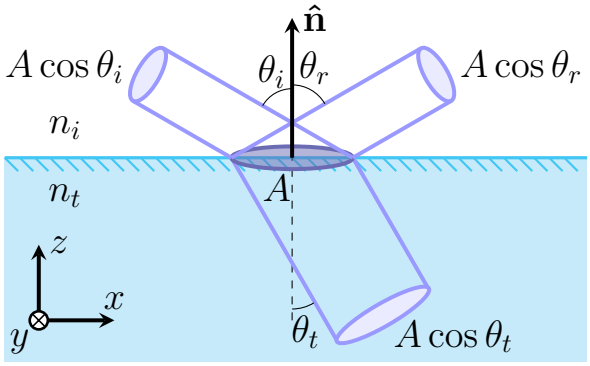

Fig. 6.8: Cross-sectional areas to calculate reflectivity and transmissivity.

$$I = \frac{1}{2} nc\epsilon_0 |E_0|^2\,, \tag{6.5.50}$$

[7] In lasers, the gain material's end facets are often either anti-reflection coated or cut at the Brewster's angle in order to minimize undesired effects due to reflection.

where E_0 is the amplitude of the field. Compared to vacuum, Eq. (4.3.36), the irradiance is thus increased by a factor n.[8] This is related to the fact that within the material, light travels slower and a given energy density is therefore compressed into a smaller volume.

We next take a given area A at the interface between two media and consider the amount of power falling onto that area, which is then either reflected or transmitted (Fig. 6.8). We also need to relate this area to the cross-sectional areas of each of the beams (bunch of rays) by proper projection. The ***reflectivity*** R of the interface is defined as the ratio of the reflected power to the incident power, which is

$$R = \frac{I_r A \cos\theta_r}{I_i A \cos\theta_i} = \frac{I_r}{I_i} = \frac{n_i(c\epsilon_0/2)|E_{0r}|^2}{n_i(c\epsilon_0/2)|E_{0i}|^2} = |r|^2 \,, \tag{6.5.51}$$

where we have relied on the fact that the incident and reflected beams are in the same medium and the final form is valid even when the reflection coefficient is complex-valued.

Similarly, the ***transmissivity*** T of the interface is defined as the ratio of the transmitted power to the incident power as

$$T = \frac{I_t A \cos\theta_t}{I_i A \cos\theta_i} = \frac{n_t(c\epsilon_0/2)|E_{0t}|^2 \cos\theta_t}{n_i(c\epsilon_0/2)|E_{0i}|^2 \cos\theta_i} = \frac{n_t \cos\theta_t}{n_i \cos\theta_i}|t|^2 \,. \tag{6.5.52}$$

The transmissivity therefore has additional factors compared to $|t|^2$, because the incident and transmitted beams exist in different materials. The reflectivities R and transmissivities T for two often encountered situations (air–glass and glass–air interfaces) are plotted in Fig. 6.9.

It is a straightforward exercise to show that the energy is conserved for both polarization cases, i.e.,

$$R + T = 1 \,. \tag{6.5.53}$$

[8] This is quite interesting result, if light could be dramatically ***slowed down*** the intensity would also increase dramatically. Some seriously smart people have been investigating this a lot because of its implications on e.g. ***nonlinear optics***.

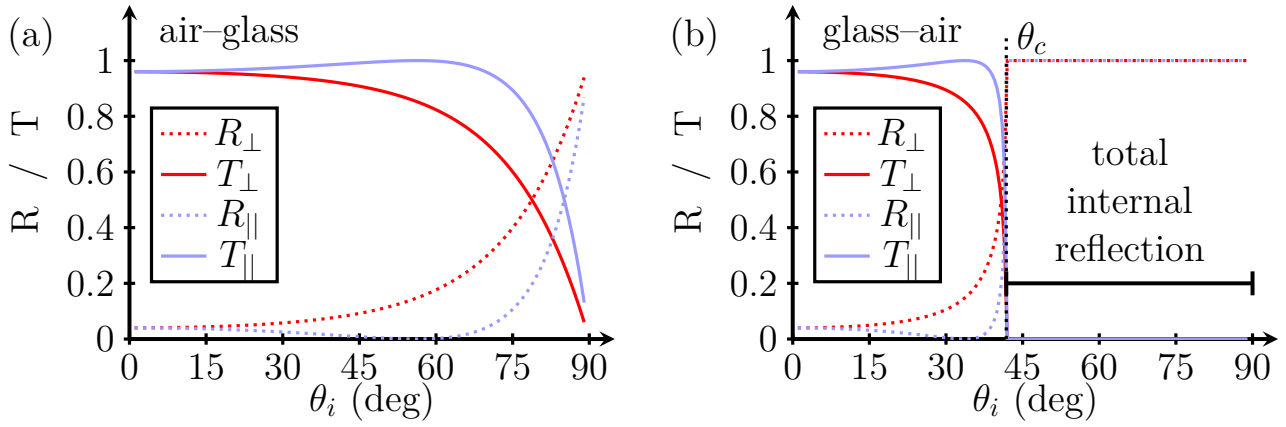

Fig. 6.9: Reflectivities R and transmissivities T for (a) air–glass and (b) glass–air interfaces for the two linear input polarizations. (b) The total internal reflection occurs when the incident angle exceeds the critical angle θ_c.

6.6 Total internal reflection

We already discussed in Figs. 6.7 and 6.9b that when the angle of incidence at the glass–air interface exceeds 41.81°, the reflection and transmission coefficients become complex. This implies that something peculiar happens for large angles of incidence when light arrives to the interface from the optically denser medium.

This effect can be traced back to the law of refraction, Eq. (6.3.30). Because the angle of incidence can take on any value up to 90° and the refractive indices have the relation $n_i > n_t$, the law of refraction can lead to the requirement that $\sin\theta_t > 1$, which is not possible for real-valued angles. This occurs if θ_i exceeds the ***critical angle*** corresponding to the case where θ_t has the highest possible value of 90° (Fig. 6.10). The critical angle is thus

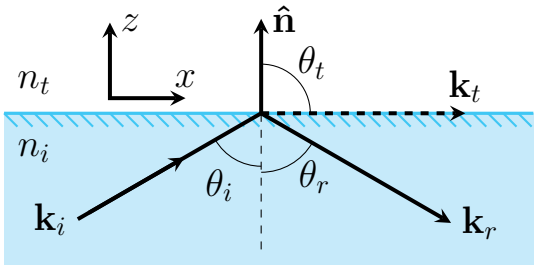

Fig. 6.10: Critical angle θ_c ($\theta_t = 90°$) for total internal reflection.

$$\sin\theta_c = n_t/n_i\,. \tag{6.6.54}$$

For angles of incidence exceeding the critical angle, the transmitted field cannot anymore propagate in a direction that can be described by real-valued angle of refraction. We may therefore expect that no transmitted field exists and that all light must be reflected from the interface. In order to understand this better, we assume that our mathematical formalism is valid also for cases where the sine of the angle exceeds unity. This, of course, is possible if the angles can be complex valued.

By applying the law of refraction for angles of incidence exceeding the critical angle, we obtain

$$\sin\theta_t = \frac{n_i}{n_t}\sin\theta_i\,, \tag{6.6.55}$$

which indeed requires complex-valued angle of refraction. In order to calculate the Fresnel coefficients for the present case, we express the cosine of the angle of refraction as

$$\cos\theta_t = \pm\sqrt{1-\sin^2\theta_t} = \pm\mathrm{i}\sqrt{\sin^2\theta_t - 1}\,. \tag{6.6.56}$$

By using this, e.g., in the Fresnel coefficient for the case where the electric field is normal to the plane of incidence, we obtain

$$r_\perp = \frac{n_i\cos\theta_i - n_t\cos\theta_t}{n_i\cos\theta_i + n_t\cos\theta_t} = \frac{n_i\cos\theta_i \pm \mathrm{i}n_t\sqrt{\sin^2\theta_t - 1}}{n_i\cos\theta_i \mp \mathrm{i}n_t\sqrt{\sin^2\theta_t - 1}}\,. \tag{6.6.57}$$

Independent of the choice of the sign, this expression is a ratio between a complex number and its complex conjugate.[9] The reflectivity of the interface is thus

$$R_\perp = |r_\perp|^2 = 1\,, \tag{6.6.58}$$

which proves that all energy is indeed reflected by the interface.[10] A similar result applies also for the other polarization case.

It is important to realize, however, that this result does not imply that the Fresnel transmission coefficients vanish. This fact is easy to verify from Eqs. (6.4.39) and (6.4.45). We therefore need to have a closer look on the transmitted field in the geometry shown in Fig. 6.11. The wave vector of the transmitted field is now

$$\mathbf{k}_t = k_t(\hat{\mathbf{x}}\sin\theta_t + \hat{\mathbf{z}}\cos\theta_t)\,, \tag{6.6.59}$$

[9] Recall that for a complex number z, $|z| = |z^*|$.

[10] Interestingly, communication via long-wave radio waves utilizes this phenomenon, where the charged particles in the ***ionosphere*** act as the culprits. More earthly example are the ***optical fibers*** and the numerous ***reflective prisms*** such as the ***Fresnel rhomb***. Similar wave behavior occurs with acoustic waves, and is also possibly utilized by ***fin whales***.

so by invoking the Snell's law we see the field has the spatial dependence of the form

$$\begin{aligned} \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}} &= \mathrm{e}^{\mathrm{i}k_t(x\sin\theta_t + z\cos\theta_t)} = \mathrm{e}^{\mathrm{i}k_t x\sin\theta_t}\mathrm{e}^{\mathrm{i}k_t z\cos\theta_t} \\ &= \mathrm{e}^{\mathrm{i}k_t x(n_i/n_t)\sin\theta_i}\mathrm{e}^{\mp k_t z(1/n_t)\sqrt{n_i^2\sin^2\theta_i - n_t^2}}, \end{aligned} \quad (6.6.60)$$

where the replacement of the cosine term follows from Eq. (6.6.56). This form shows that the transmitted wave has harmonic oscillations along the interface (x direction). However, along the surface normal, the wave is either exponentially growing or decaying function. It is clear that the only physically possible solution is exponential decay as the wave moves away from the interface. This way, no energy is transported away from the interface, which agrees with the earlier result that all energy is reflected by the interface. Taken together, we have a ***surface wave*** that propagates along the surface, but the wave is ***evanescent***, meaning that its amplitude decays as the distance from the surface grows.

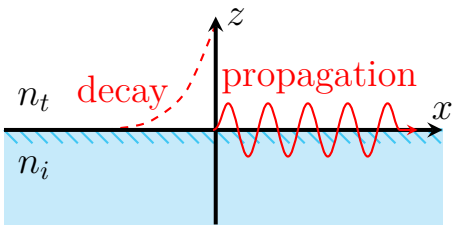

Fig. 6.11: Evanescent surface wave.

This discussion shows, however, that the transmitted wave has a non-vanishing value for some distance away from the interface. If we now bring another interface very close to the original one (Fig. 6.12) and choose the third medium in such a way that the angle of refraction in the third medium fulfils the condition $\sin\theta_3 < 1$, the field there is again a propagating wave (with a component away from the interface). This effect is known as the ***frustrated total internal reflection***, which allows one to transfer electromagnetic energy across small gaps even when the gap material itself does not support propagating waves. This effect can also give rise to detrimental losses in optical waveguides/fibers by allowing the propagating waves to scatter out.

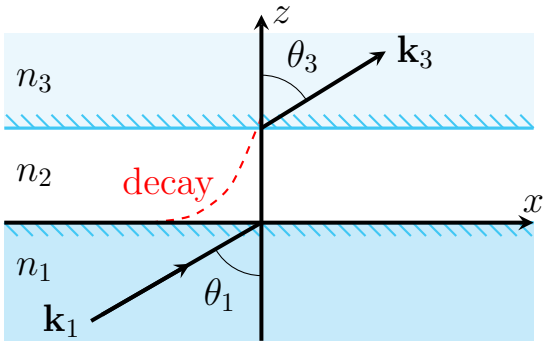

Fig. 6.12: Frustrated total internal reflection.

7. SUPERPOSITION

7.1 Superposition principle

We will now return to the wave-optical treatment of light. More specifically, we will have a closer look on different aspects of the superposition principle. This is needed to discuss the important topics of interference and diffraction.

The superposition principle was already mentioned in Section 3.5. The general form of the wave equation is [cf. Eq. (3.8.64)]

$$\nabla^2\Psi = \frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2}\,. \tag{7.1.1}$$

The derivatives are linear operators. In consequence, if we have a set of different solutions to Eq. (7.1.1), then an arbitrary superposition of the solutions

$$\Psi(\mathbf{r},t) = \sum_i^n c_i\Psi_i(\mathbf{r},t)\,, \tag{7.1.2}$$

is also a solution.

In the following, we will use the complex notation for the fields and, when necessary, take into account the vectorial character of the fields.

7.2 Harmonic waves

It was already mentioned (Section 3.5) that the superposition of two waves depends on the phase difference between the waves. In order to understand this properly, we need to have a closer look on how the wave is expressed in complex notation.

Consider first a harmonic (scalar) wave propagating in the z direction, which can be expressed in a number of different forms

$$\tilde{E}(z,t) = E(z)\mathrm{e}^{-\mathrm{i}\omega t} = A\mathrm{e}^{\mathrm{i}\epsilon}\mathrm{e}^{\mathrm{i}(kz-\omega t)}\,. \tag{7.2.3}$$

Here the tilde (~) on the left-hand form indicates that the quantity includes the rapid temporal oscillations (now only at frequency ω). For a monochromatic wave, the time dependence is very simple and it is useful to define a quantity $E(z)$ that includes only the spatial dependence. Finally, the form on the right-hand side shows explicitly even the initial phase of the wave, so that the amplitude A is real valued. The spatial dependence is thus

$$E(z) = A\mathrm{e}^{\mathrm{i}\epsilon}\mathrm{e}^{\mathrm{i}kz}\,. \tag{7.2.4}$$

Formally, this equation represents a plane wave propagating along z direction. In superposition experiments, however, the various fields may have traced very different

paths to the point of observation. It is therefore crucial that we generalize the spatial coordinate z to account for the whole optical path length [Eq. (6.2.9)] from a given reference location (with known phase) to the observation point even when the path includes sections with propagation along directions different from z. In order to shorten the notation, we write

$$E(z) = A\mathrm{e}^{\mathrm{i}\alpha}\,, \tag{7.2.5}$$

where $\alpha = \epsilon + kz$.

We are now ready to consider the superposition of two waves at the same frequency. In order to allow for the possibility that their initial phases are different and that they may have traced different paths to the observation point, we write them as

$$\tilde{E}_1(z_1,t) = E_1(z_1)\mathrm{e}^{-\mathrm{i}\omega t}\,, \qquad \tilde{E}_2(z_2,t) = E_2(z_2)\mathrm{e}^{-\mathrm{i}\omega t}\,, \tag{7.2.6}$$

where the spatial parts are

$$E_1(z_1) = A_1(z_1)\mathrm{e}^{\mathrm{i}\epsilon_1}\mathrm{e}^{\mathrm{i}kz_1} = A_1\mathrm{e}^{\mathrm{i}\alpha_1}\,, \qquad E_2(z_2) = A_2(z_2)\mathrm{e}^{\mathrm{i}\epsilon_2}\mathrm{e}^{\mathrm{i}kz_2} = A_2\mathrm{e}^{\mathrm{i}\alpha_2}\,. \tag{7.2.7}$$

The phases $\alpha_1 = \epsilon_1 + kz_1$ and $\alpha_2 = \epsilon_2 + kz_2$ thus include the initial phases of the two waves at their respective reference locations as well as their paths from these locations to the observation point.

The superposition of the two waves is

$$\begin{aligned}\tilde{E}(z,t) = \tilde{E}_1(z_1,t) + \tilde{E}_2(z_2,t) &= \left[E_1(z_1) + E_2(z_2)\right]\mathrm{e}^{-\mathrm{i}\omega t}\,,\\ &= \left[A_1\mathrm{e}^{\mathrm{i}\alpha_1} + A_2\mathrm{e}^{\mathrm{i}\alpha_2}\right]\mathrm{e}^{-\mathrm{i}\omega t}\,,\\ &= \left[A_1 + A_2\mathrm{e}^{\mathrm{i}(\alpha_2-\alpha_1)}\right]\mathrm{e}^{\mathrm{i}\alpha_1}\mathrm{e}^{-\mathrm{i}\omega t}\,.\end{aligned} \tag{7.2.8}$$

The irradiance of the total field is proportional to the square of its magnitude

$$\begin{aligned}I(z) \propto |\tilde{E}(z,t)|^2 &= \left[A_1 + A_2\mathrm{e}^{\mathrm{i}(\alpha_2-\alpha_1)}\right]\left[A_1 + A_2\mathrm{e}^{-\mathrm{i}(\alpha_2-\alpha_1)}\right]\,,\\ &= A_1^2 + A_2^2 + 2A_1A_2\cos(\alpha_2-\alpha_1)\,.\end{aligned} \tag{7.2.9}$$

The two first terms of this expression represent the irradiances of the two waves independently. The last term, on the other hand, is the interference term, which depends on the phase difference $\delta = \alpha_2 - \alpha_1$ between the waves. It is clear that an interference maximum is obtained when $\delta = m2\pi$ and a minimum when $\delta = m2\pi + \pi$. The phase difference is also

$$\delta = \epsilon_2 - \epsilon_1 + k(z_2 - z_1)\,. \tag{7.2.10}$$

The phase difference thus depends on the initial phases of the waves and on their optical paths to the observation point. This can of course be generalized to the case where the two waves pass through several different materials with different refractive indices, resulting in optical path lengths given by Eq. (6.2.9) for each wave separately.

The above treatment is straightforward to generalize for an arbitrary number N of waves, resulting in the total field

$$\begin{aligned}\tilde{E}(z,t) &= \sum_{j=1}^{N} E_j(z_j)\mathrm{e}^{-\mathrm{i}\omega t} = \left(\sum_{j=1}^{N} A_j \mathrm{e}^{\mathrm{i}\alpha_j}\right)\mathrm{e}^{-\mathrm{i}\omega t}\,, \\ &= A\mathrm{e}^{\mathrm{i}\alpha}\mathrm{e}^{-\mathrm{i}\omega t}\,. \end{aligned} \tag{7.2.11}$$

The total field is thus obtained by adding the complex quantities $A_j\mathrm{e}^{\mathrm{i}\alpha_j}$. In the complex notation, it is customary to include the initial phases of the source in the complex amplitude, i.e., by replacing $A_j\mathrm{e}^{\mathrm{i}\epsilon_j} \to A_j$.

7.3 Standing waves

The wave equation allows solutions that propagate either in the positive or negative z direction (Fig. 7.1). As an example of this, we consider the situation where an incident field $\tilde{E}_I(z,t) = A_I\mathrm{e}^{\mathrm{i}(kz-\omega t)}$ is reflected by a mirror located at $z = 0$, resulting in a reflected wave $\tilde{E}_R(z,t) = A_R\mathrm{e}^{\mathrm{i}(-kz-\omega t)}$. The superposition is thus

$$\tilde{E}(z,t) = \left(A_I\mathrm{e}^{\mathrm{i}kz} + A_R\mathrm{e}^{-\mathrm{i}kz)}\right)\mathrm{e}^{-\mathrm{i}\omega t}\,. \tag{7.3.12}$$

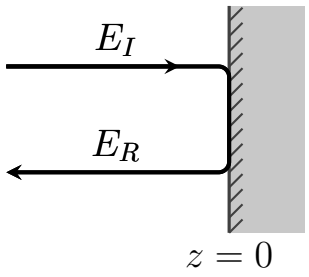

Fig. 7.1: wave reflecting from a mirror.

For the particular case of a perfect metallic mirror, the total field at $z = 0$ must vanish[1], and therefore $A_R = -A_I$. The superposition then becomes

$$\tilde{E}(z,t) = A_I\left(\mathrm{e}^{\mathrm{i}kz} - \mathrm{e}^{-\mathrm{i}kz}\right)\mathrm{e}^{-\mathrm{i}\omega t} = 2\mathrm{i}A_I\sin(kz)\mathrm{e}^{-\mathrm{i}\omega t}\,. \tag{7.3.13}$$

The superposition therefore oscillates harmonically in time. However, the total field stays at zero at certain points given by the zeros of the sine function. These points are known as ***nodal points***. The points between the nodal points, on the other hand, oscillate back and forth in time, as shown in Fig. 7.2.

[1]Note that the ***boundary conditions*** for dielectric–metal interface are different than for two dielectric materials.

It is evident from this discussion that the total field does not propagate. The wave is thus known as the ***standing wave***. It is important to note, however, that the two original waves do propagate as such independent of each other.

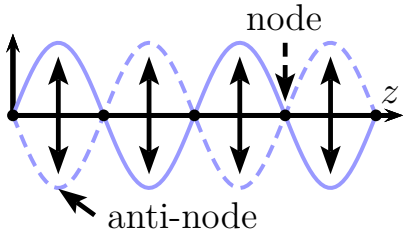

Fig. 7.2: Standing wave with nodes and antinodes.

7.4 Superposition of several frequencies and coherence

For a basic understanding how wave packets or pulses can be built from several different frequencies, we focus on temporal signals. Their connection to propagating waves is obtained by assuming that the properties of the wave packets are characterized at some fixed point in space.

The ***Fourier inversion theorem*** states that an arbitrary (but well behaving) function of time t can be represented as a superposition of several different frequency components. This is known as the Fourier integral theorem

$$f(t) = \int_{-\infty}^{\infty} F(\omega) \mathrm{e}^{-\mathrm{i}\omega t} \, \mathrm{d}\omega \, . \tag{7.4.14}$$

The frequency-dependent function $F(\omega)$ is known as the ***Fourier transform*** of the time dependent function $f(t)$, and is given by[2]

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) \mathrm{e}^{\mathrm{i}\omega t} \, \mathrm{d}t \, , \tag{7.4.15}$$

and describes the frequency content of the time-dependent function. This function $F(\omega)$ is also known as the ***spectrum*** of the time-dependent function $f(t)$, and the above process of decomposing $f(t)$ into its spectral components is known as ***Fourier analysis***. Eq. (7.4.14), on the other hand, is also known as the ***inverse transform***, and the process of forming the time-dependent function $f(t)$ from its frequency components is called ***Fourier synthesis***.

[2]Note that many other ***integral transforms*** exist that greatly help the mathematical treatment of many problems. Recall also that using our previous notation we could have also further clarified the basis (t versus ω) by using the tilde notation $\tilde{f}(t)$. E.g. $\tilde{E}(t) = \int_{-\infty}^{\infty} E(\omega) \mathrm{e}^{-\mathrm{i}\omega t} \, \mathrm{d}\omega$. Some authors also use the hat notation [$\hat{E}(\mathbf{k})$ vs $E(\mathbf{r})$] to further clarify the basis (now $\mathbf{r}$ versus $\mathbf{k}$) that is used to describe the function.

The Fourier transform and the inverse transform can be defined in a number of different ways.[3] In particular, the factor of $1/2\pi$ in Eq. (7.4.15) can be divided arbitrarily between the transform and inverse transform. The present choice emphasizes the fact that $F(\omega)$ as such is the amplitude of the wave at a given frequency.

As a simple but important example, we consider a harmonic wave packet (Fig. 7.4a)

$$f(t) = \begin{cases} \mathrm{e}^{-\mathrm{i}\omega_0 t}, & |t| < T/2 \\ 0 \quad , & \text{otherwise} \end{cases} . \tag{7.4.16}$$

The Fourier transform of this function is easy to calculate and is:[4]

$$\begin{aligned} F(\omega) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t = \frac{1}{2\pi} \int_{-T/2}^{T/2} \mathrm{e}^{-\mathrm{i}\omega_0 t} \mathrm{e}^{\mathrm{i}\omega t}\, \mathrm{d}t\,, \\ &= \frac{1}{2\pi \mathrm{i}(\omega - \omega_0)} \left[\mathrm{e}^{\mathrm{i}(\omega-\omega_0)T/2} - \mathrm{e}^{-\mathrm{i}(\omega-\omega_0)T/2} \right] , \\ &= \frac{1}{\pi(\omega - \omega_0)} \sin\left[(\omega - \omega_0)T/2\right] , \\ &= \frac{T/2}{\pi(\omega - \omega_0)T/2} \sin\left[(\omega - \omega_0)T/2\right] , \\ &= \frac{T}{2\pi} \operatorname{sinc}\left[(\omega - \omega_0)T/2\right] . \end{aligned} \tag{7.4.17}$$

The sinc-function, is defined here as $\operatorname{sinc}(x) = \sin(x)/x$, which is also known as the unnormalized sinc-function. Alternatively, the sinc-function could be defined as $\operatorname{sinc}(x) = \sin(\pi x)/\pi x$, which is also known as the normalized sinc-function. These alternative definitions have been plotted in Fig. 7.3. During this course, we mainly use the unnormalized sinc-function convention (Fig. 7.3a).

The resulting function of Eq. (7.4.17) is plotted in Fig. 7.4b as a function of the parameter $x = (\omega - \omega_0)T/2$. The function is seen to have a strong maximum at $x = (\omega - \omega_0)T/2 = 0$, i.e., at $\omega_0 = \omega$. However, this maximum extends on both sides to

[3] Please see further details e.g. from ***here***.

[4] Alternatively, this can be calculated by taking use of the known properties of Fourier transforms. First, the Fourier transform of a ***box car function*** $f(at) = \operatorname{rect}(at)$ is $F(\omega) = \frac{1}{2\pi|a|} \operatorname{sinc}\left(\frac{\omega}{2a}\right)$, where for us $a = 1/T$. Then, we take use of the ***frequency shifting property***, stating that if $h(t) = \mathrm{e}^{\mathrm{i}\omega_0 t} f(t)$, for real ω_0, then $H(\omega) = F(\omega - \omega_0)$. Therefore, our function turns into $F(\omega) = \frac{T}{2\pi} \operatorname{sinc}\left[(\omega - \omega_0)T/2\right]$. In practise, this approach is both more general and powerful as it provides means to quickly calculate also quite complicated transforms. Knowledge of these properties also often provides deeper insights to the problem at hand.

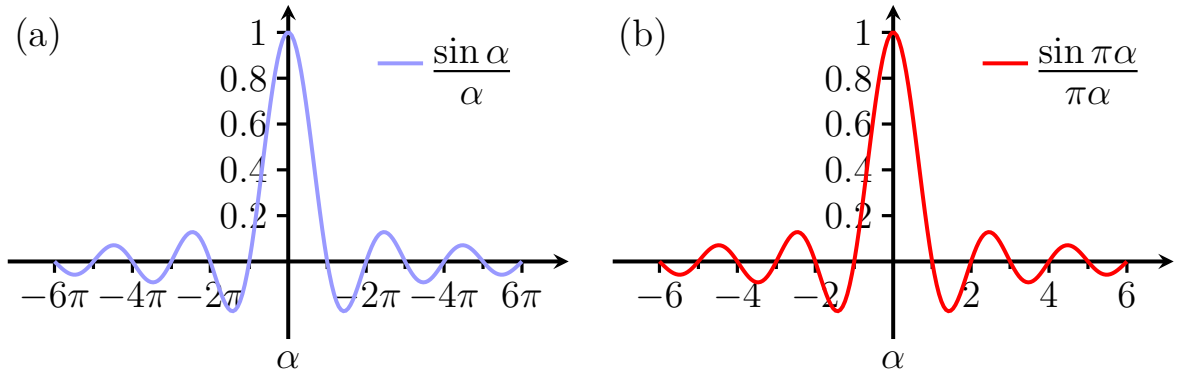

Fig. 7.3: (a) The unnormalized sinc-function and (b) the normalized sinc-function.

the first zero of the $\sin(x)$ and exhibits further weak oscillations for even higher values of x. In terms of frequency, the original frequency ω_0 makes the strongest contribution to the wave packet but nearby frequencies also make appreciable contributions.

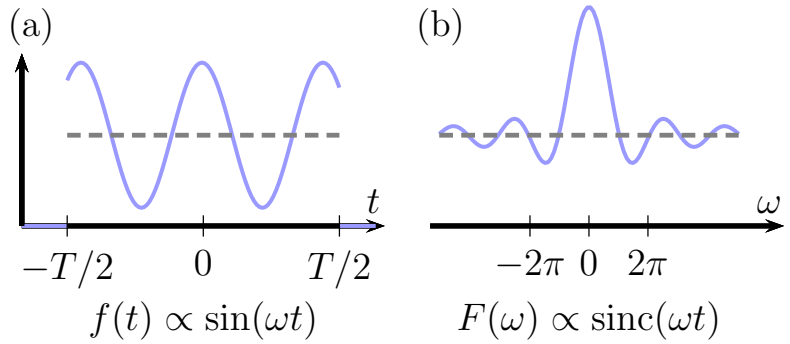

Fig. 7.4: Harmonic wave packet $f(t)$ of length T and its Fourier transform $F(\omega)$.

The Fourier transform itself can be interpreted as the amplitude of the wave at a certain frequency component. In general, the transform is complex valued. In order to understand how much power each frequency component carries, it is useful to define the ***power spectrum*** of the function $f(t)$ as

$$S(\omega) = |F(\omega)|^2 \,. \tag{7.4.18}$$

It is evident that the power spectrum further emphasizes the role of the central maximum in Fig. 7.4b.

We are often interested in understanding which frequencies need to be taken into account to obtain a reasonable representation of the original function. For the case of the harmonic wave packet, we can choose rather arbitrarily that the important frequencies are contained between the first zeros of Fig. 7.4b on both sides, i.e.,

$$(\omega - \omega_0)T/2 = \pm\pi \,. \tag{7.4.19}$$

The important frequencies thus cover the range between $\omega_0 \pm 2\pi/T$ and we can define that the ***bandwidth*** of the spectrum is

$$\Delta\omega = 4\pi/T\,. \tag{7.4.20}$$

For our example, T clearly represents the ***pulse length*** of the wave packet $\Delta t = T$. We therefore obtain the relation between the bandwidth and the pulse length as

$$\Delta\omega\Delta t = 4\pi\,. \tag{7.4.21}$$

This expression further simplifies by considering the bandwidth for real frequency (rather than angular frequency) $\nu = \omega/2\pi$. We thus obtain

$$\Delta\nu\Delta t = 2\,. \tag{7.4.22}$$

It is important to realize however, that the choice for the width of the spectrum given by Eq. (7.4.20) is quite arbitrary.[5] We could equally well have chosen that the representative range is between $(\omega - \omega_0)T/2 = \pm\pi/2$. In addition, depending on the mathematical form of the spectrum in each specific case, the most natural choice may be different. In each case, however, the orders of magnitude of the bandwidth and pulse length are approximately related by[6]

$$\Delta\nu\Delta t \sim 1\,. \tag{7.4.23}$$

The bandwidth and pulse length are therefore not precise concepts and there is easily at least a factor of two between different equally justifiable choices for these quantities.

For real light sources, we can often assume that radiation is emitted as harmonic wave packets whose duration varies. During this period, the wave behaves as a harmonic wave at some frequency ω_0. However, after some time, which is on the average Δt, the phase of the wave jumps to a new and completely random value (Fig. 7.4). In consequence, the phase correlation of the wave is maintained quite well between two points of time separated by less than the average

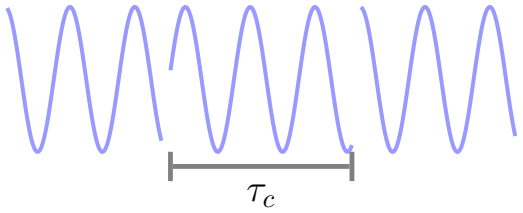

Fig. 7.5: Light sources emit radiation as wave packet with average length of coherence time τ_c after which the phase jumps to a random value.

[5] Many conventions exist, such as taking the ***full width at half maximum (FWHM)*** or so-called ***3 dB bandwidth***.

[6] In general, this kind of ***uncertainty*** in the conjugate variables (here t and ω) arises from the properties for the Fourier transform.

duration. However, if the two points of time are separated by much more than the average duration, the phase correlation is completely lost. The average duration is also known as the ***coherence time*** of the wave, often denoted by τ_c.

We can also define a distance $l_c = \tau_c c$ travelled by light in coherence time, known as the ***coherence length***. The implication for superposition experiments is that if light from a given source is divided to follow two different paths and then brought together again, no interference can be observed if the optical path difference between the paths exceeds the coherence length. This is because the phase difference of the interference term in Eq. (7.2.9) will become random and average to zero. Again, it is important to note that the coherence time and coherence length are not precisely defined quantities.[7] Rather, theyrepresent typical scales for the phase jumps to occur. In consequence, the significance of the interference term becomes gradually less and less important when the path difference is somewhat less or more than the coherence length.

[7]Proper treatment is found e.g. from book *Optical Coherence and Quantum Optics* by Mandel and Wolf.

8. INTERFERENCE

8.1 Conditions for interference

We are now prepared to discuss superposition effects in more detail. We will start with ***interference***. This term is usually used when the superposition occurs between a number of distinct waves, each with a finite amplitude, as opposed to a continuous range of infinitesimal waves.

We first consider two plane waves at frequency ω with vector amplitudes, which are of the form

$$\tilde{\mathbf{E}}_1(\mathbf{r},t) = \mathbf{A}_1 \mathrm{e}^{\mathrm{i}\epsilon_1} \mathrm{e}^{\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r}-\omega t)} \qquad \text{and} \qquad \tilde{\mathbf{E}}_2(\mathbf{r},t) = \mathbf{A}_2 \mathrm{e}^{\mathrm{i}\epsilon_2} \mathrm{e}^{\mathrm{i}(\mathbf{k}_2\cdot\mathbf{r}-\omega t)} \,. \tag{8.1.1}$$

Note that we have again explicitly shown the initial phases of the waves, because they play an important role in the superposition wave. In spite of this, the vector amplitudes can still be complex-valued if their polarization unit vectors are complex.

The superposition wave is thus

$$\begin{aligned} \tilde{\mathbf{E}}(\mathbf{r},t) &= \tilde{\mathbf{E}}_1(\mathbf{r},t) + \tilde{\mathbf{E}}_2(\mathbf{r},t) \\ &= \mathbf{A}_1 \mathrm{e}^{\mathrm{i}\epsilon_1} \mathrm{e}^{\mathrm{i}(\mathbf{k}_1\cdot\mathbf{r}-\omega t)} + \mathbf{A}_2 \mathrm{e}^{\mathrm{i}\epsilon_2} \mathrm{e}^{\mathrm{i}(\mathbf{k}_2\cdot\mathbf{r}-\omega t)} \\ &= \left(\mathbf{A}_1 \mathrm{e}^{\mathrm{i}\epsilon_1} \mathrm{e}^{\mathrm{i}\mathbf{k}_1\cdot\mathbf{r}} + \mathbf{A}_2 \mathrm{e}^{\mathrm{i}\epsilon_2} \mathrm{e}^{\mathrm{i}\mathbf{k}_2\cdot\mathbf{r}} \right) \mathrm{e}^{-\mathrm{i}\omega t} = \mathbf{E}(\mathbf{r}) \mathrm{e}^{-\mathrm{i}\omega t} \,, \end{aligned} \tag{8.1.2}$$

whose irradiance is, as before,

$$\begin{aligned} I \propto |\tilde{\mathbf{E}}|^2 = \mathbf{E}\cdot\mathbf{E}^* &= (\mathbf{E}_1 + \mathbf{E}_2)\cdot(\mathbf{E}_1^* + \mathbf{E}_2^*) \\ &= |\mathbf{E}_1|^2 + |\mathbf{E}_2|^2 + \mathbf{E}_1\cdot\mathbf{E}_2^* + \mathbf{E}_1^*\cdot\mathbf{E}_2 \,. \end{aligned} \tag{8.1.3}$$

Here, the first two terms represent the irradiances of the two independent waves and the two last terms the interference between the waves. The interference terms become

$$\mathbf{E}_1\cdot\mathbf{E}_2^* = \mathbf{A}_1\cdot\mathbf{A}_2^* \mathrm{e}^{\mathrm{i}(\epsilon_1-\epsilon_2)} \mathrm{e}^{\mathrm{i}(\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r}} \,, \tag{8.1.4}$$

$$\mathbf{E}_1^*\cdot\mathbf{E}_2 = \mathbf{A}_1^*\cdot\mathbf{A}_2 \mathrm{e}^{-\mathrm{i}(\epsilon_1-\epsilon_2)} \mathrm{e}^{-\mathrm{i}(\mathbf{k}_1-\mathbf{k}_2)\cdot\mathbf{r}} \,. \tag{8.1.5}$$

The results of Eqs. (8.1.4) and (8.1.5) show a new aspect that the interference terms vanish if the polarizations of the two waves are orthogonal $\mathbf{A}_1\cdot\mathbf{A}_2^* = 0$.

We will next assume that the two waves have identical polarizations, so that $\mathbf{A}_1\cdot\mathbf{A}_2^* = \mathbf{A}_1^*\cdot\mathbf{A}_2 = A_1 A_2$. The superposition wave can then be written as

$$I \propto |\tilde{\mathbf{E}}|^2 = |A_1|^2 + |A_2|^2 + 2A_1A_2\cos\delta \,, \tag{8.1.6}$$

where the phase difference is

$$\delta = \epsilon_1 - \epsilon_2 + (\mathbf{k}_1 - \mathbf{k}_2)\cdot\mathbf{r} \,. \tag{8.1.7}$$

In terms of the irradiances, Eq. (8.1.6) becomes

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\delta\,. \tag{8.1.8}$$

For the special case of equal irradiances $I_1 = I_2 = I_0$ of the two waves, this can be written as

$$I = 2I_0 + 2I_0\cos\delta = 4I_0\cos^2(\delta/2)\,. \tag{8.1.9}$$

Independent of the form and as before, interference maxima are obtained when $\delta = n2\pi$ and minima when $\delta = n2\pi + \pi$.

We can now collect the earlier and present results as the basic requirements for the observation of interference effects. In Section 7.4, we pointed out that the interference term vanishes if there is no phase correlation between the two waves. This requirement is related to the coherence properties of the waves.

Our discussion in Section 7.4 was related to ***temporal (longitudinal) coherence*** of the light source. More specifically, if light from a single source is divided into two waves that follow different paths with optical lengths s_1 and s_2 to the observation point (Fig. 8.1a), interference can only be observed if their travel times $t_1 = s_1/c$ and $t_2 = s_2/c$ are within the coherence time of the sources $|t_1 - t_2| < \tau_c$. Equivalently, the optical path difference must be less than the coherence length of the source $|s_1 - s_2| < l_c = c\tau_c$.

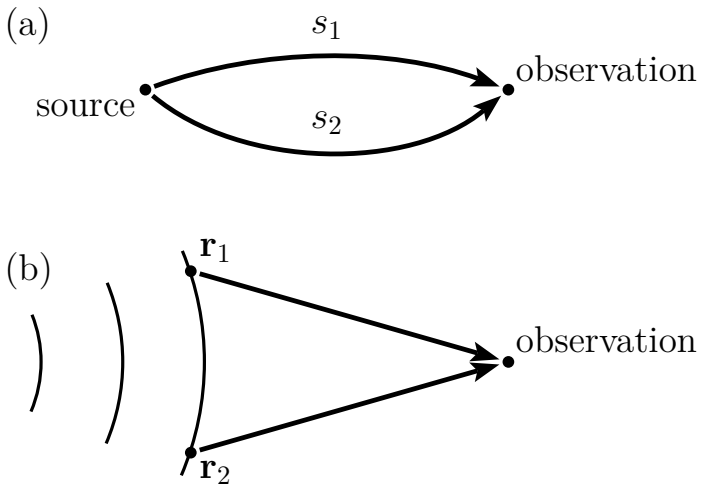

Fig. 8.1: Relevant quantities for temporal (a) and spatial (b) coherence.

In some cases, we need to consider the ***spatial (transverse) coherence*** of the optical field. This occurs when light from different parts of the same wave front is brought to a single observation point (Fig. 8.1b). It is often the case that phase correlation within the wave front is only maintained over a distance known as ***coherence width*** w_c. The distance between the two source points must then be less than this width $|\mathbf{r}_1 - \mathbf{r}_2| < w_c$.

We also noted after Eqs. (8.1.4) and (8.1.5) that the interference term vanishes if

the polarizations of the two waves are orthogonal. On the other hand, light from conventional sources is unpolarized in the sense that its any two orthogonal states of polarization are mutually incoherent. In consequence, two waves originating from the orthogonal states of polarization do not interfere even if their polarizations are rotated to point in the same direction. Historically, these rules are known as the ***Fresnel–Arago laws***, whose original formulation is quite awkward.

8.2 Wavefront splitting interferometers

Any instrument whose operation is based on interference is known as an ***interferometer***. They are classified into two main types. We will first consider ***wavefront*** splitting interferometers. In this case, light from two points of the same wave front is brought to the same observation point. By varying the location of the observation point, one can then observe interference maxima and minima, also known as ***interference fringes***.

A classic example of a wavefront splitting case is Young's ***double slit experiment*** (Fig. 8.2a). Here, an incident wave is first transmitted through a narrow slit (or a small hole). The cylindrical or spherical wave thus generated is incident on screen with two slits. These secondary waves then interfere to produce a fringe pattern on a screen located at some distance from the screen with the two slits. One may ask why the incident wave is not directly applied on the screen with two slits. The reason is historical: at the time of the original experiment by Young, no light sources with high coherence were available. The light from the narrow slit, however, can be seen as an elementary source with good coherence properties. The wavefront striking the two secondary slits therefore possesses sufficient spatial coherence for the fringe pattern to be observed.

In order to model Young's experiment, we consider Fig. 8.2b. The distance between the screen with two slits and the observation screen is s. This distance is much larger than the separation between the slits a. In addition, the distance of the observation point P from the center of the observation screen y, is assumed small. The optical path lengths from the two slits to the observation point are the lengths $r_1 = |S_1P|$ and $r_2 = |S_2P|$. Within our assumptions, the path difference is approximately

$$r_1 - r_2 \approx |S_2B| = a \sin\theta \approx a\theta \,. \tag{8.2.10}$$

In addition, $\theta \approx y/s$. The phase difference between the two waves at the observation

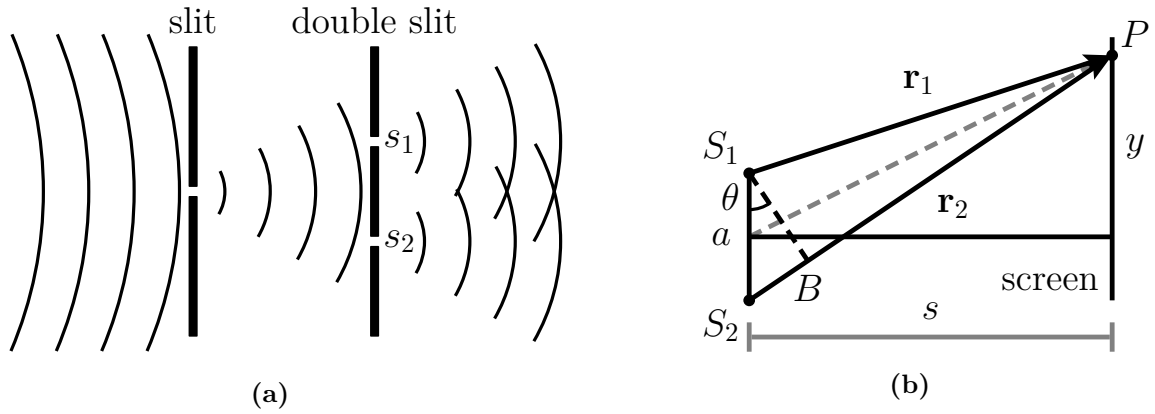

Fig. 8.2: (a) Young's double slit experiment. (b) Simple model of Young's double slit experiment.

point is thus

$$\delta = k(r_1 - r_2) = \frac{2\pi}{\lambda}(r_1 - r_2) = \frac{2\pi}{\lambda}\frac{a}{s}y\,. \tag{8.2.11}$$

Interference maxima are observed when $\delta = m2\pi$, i.e., when $(r_1 - r_2) = m\lambda$, where m is known as the ***order of interference***. The m^{th} maximum is therefore observed at location

$$y_m = m\frac{s}{a}\lambda\,. \tag{8.2.12}$$

Alternatively, we may also characterize the location by the direction of propagation $\theta \approx y/s$, yielding

$$\theta_m = m\frac{\lambda}{a}\,. \tag{8.2.13}$$

8.3 Amplitude-splitting interferometers

The other main class of interferometers is ***amplitude-splitting interferometers***. Here, the amplitude of the incident wave is split by reflections and transmissions into two or more waves that follow different paths to the observation point.

The simplest example of amplitude-splitting interference is the case of thin films with thickness on the order of wavelength (Fig. 8.3). If the reflection coefficients of the film are sufficiently low, only two important rays propagate in the reflected direction. These rays are reflected by the front and back surfaces of the film and propagate parallel to each other to the observation point. The two rays thus interfere with each other only at infinite distance from the film. Recall, however, that such parallel rays can be forced to intercept each other at a finite distance by using a lens to focus them to the back focal plane of the lens.

We now treat this case more carefully. The film has refractive index of n_t and is surrounded by media with indices n_1 and n_2. The thickness of the film is d. A ray is incident at point A on the top surface of the film from the medium with index n_1 and angle of incidence θ_i. The reflection from point A propagates as ray 1 towards the observation point. Part of the incident wave is transmitted at A and propagates at angle θ_t within the film. This ray is only reflected at point B on the back surface of the film. This ray is transmitted by the top surface at point C to provide ray 2 propagating towards the observation point. The path difference between the two rays depends therefore on the distances AD and ABC.

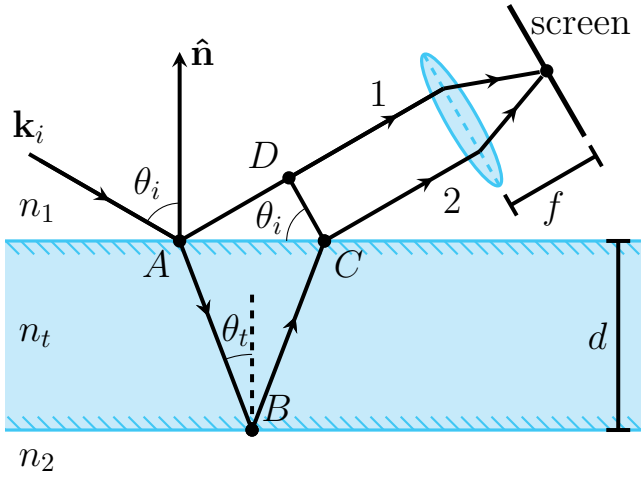

Fig. 8.3: Thin-film interference resulting in two parallel rays that interfere at infinity. The fringes can be focused on a screen by a lens.

From Fig. 8.3, we first obtain the relation $|AC| = 2d\tan\theta_t$, while application of the Snell's law results in $n_1\sin\theta_i = n_t\sin\theta_t$. The optical path lengths of the two rays are then

$$\begin{aligned}\mathrm{OPL}_1 &= n_1|AD| = n_1|AC|\sin\theta_i = 2n_1 d\tan\theta_t\sin\theta_i\,,\\ &= 2n_t d\tan\theta_t\sin\theta_t = \frac{2n_t d\sin^2\theta_t}{\cos\theta_t}\,, \qquad (8.3.14)\\ \mathrm{OPL}_2 &= 2n_t|AB| = \frac{2n_t d}{\cos\theta_t}\,. \qquad (8.3.15)\end{aligned}$$

The optical path difference between the two rays is thus

$$\Delta\mathrm{OPL}_{1,2} = \Lambda = \frac{2n_t d}{\cos\theta_t}(1-\sin^2\theta_t) = 2n_t d\cos\theta_t\,. \qquad (8.3.16)$$

As an example, we can consider the case where the media on both sides of the film are the same and have the refractive index of n. This index must be either larger or smaller than that of the film n_t. Consequently, one of the reflections, but not both, results in a π phase shift (recall the fresnel coefficients plotted in Fig. 6.7). The total

phase shift between the two reflected rays is thus

$$\delta = k\Lambda \pm \pi = \frac{2\pi}{\lambda_0}\Lambda \pm \pi = \frac{4\pi}{\lambda_0} n_t d \cos\theta_t \pm \pi\,, \tag{8.3.17}$$

where λ_0 is the vacuum wavelength, as usual.

Interference maxima are again observed when $\delta = m2\pi$ and we obtain the condition for the maxima in the form

$$d\cos\theta_t = (2m+1)\,\frac{\lambda_0}{4n_t} = (2m+1)\,\frac{\lambda_t}{4} = \left(m+\frac{1}{2}\right)\frac{\lambda_t}{2}\,, \tag{8.3.18}$$

where $\lambda_t = \lambda_0/n_t$ is the wavelength in the film. Successive maxima are observed when the quantity $d\cos\theta_t$ changes by half a wavelength. As light propagates twice through the film, this quantity can be interpreted as sort of an effective thickness of the film.

The maxima thus depend both on the physical thickness of the film d and the propagation angle within the film θ_t. For a constant thickness, we thus obtain fringes for different angles of observation, resulting in a circularly symmetric fringe pattern for a given wavelength, known as the ***Haidinger fringes***. On the other hand, for a given direction of observation, different colors lead to maxima for different film thicknesses. These fringes are known as ***Newton's rings***[1]. These, albeit not with a ring shape, are often observed in soap bubbles or oil on water puddles.

8.4 Michelson interferometer

Historically, one of the most important cases is the ***Michelson interferometer***. A version of such a configuration was used to measure the speed of light. Its schematic principle of operation is shown in Fig. 8.5.

The light incident on the interferometer first meets at point A a ***beam splitter*** that reflects part of the incident wave and transmits the remaining part. The beam splitter is usually designed in such a way that its one surface reflects and transmits 50 % of the incident light. The other surface, on the other hand, has ***anti-reflection coating*** that minimizes

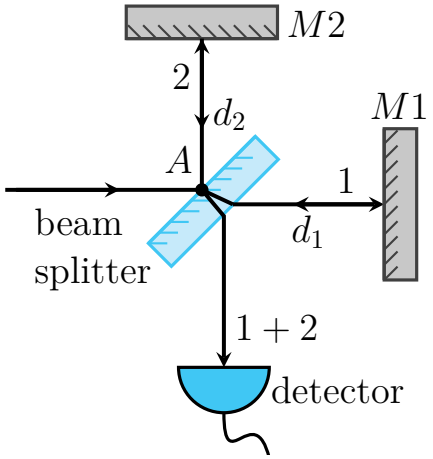

Fig. 8.4: Operation principle of the Michelson interferometer.

[1] In the strictest sense of nomenclature, Newton's rings arise from interference effects between two surfaces. The first surface is a spherical surface while the adjacent, touching surface is flat.

the reflections. Such a beam splitter is known as a 50/50 beam splitter. The transmitted wave propagates to mirror $M1$ and is reflected directly backwards. The reflected wave propagates to mirror $M2$ and is also reflected directly backwards. The waves reflected by the mirrors meet again at the beam splitter. 50 % of each wave contributes to the superposition wave that propagates to the detector, whereas the remaining parts propagate back to the original light source.

We denote the optical paths from point A to the respective mirrors by d_1 and d_2. For the superposition wave, this gives rise to a total path difference of

$$\Lambda = 2(d_2 - d_1) = 2d\,. \tag{8.4.19}$$

Mirrors M1 and M2 are usually identical so that the reflections at the mirrors do not contribute to the total phase difference. However, at the beam splitter one of the waves undergoes external reflection, whereas the other undergoes internal reflection. This results in an additional phase difference of $\pm\pi$ between the two waves. The total phase difference is thus

$$\delta = k\Lambda \pm \pi = \frac{2\pi}{\lambda_0}\Lambda \pm \pi = \frac{4\pi}{\lambda_0}d \pm \pi\,. \tag{8.4.20}$$

Interference maxima are again observed when $\delta = m2\pi$ and we obtain the condition for the maxima in the form

$$d = (2m+1)\frac{\lambda_0}{4} = \left(m + \frac{1}{2}\right)\frac{\lambda_0}{2}\,. \tag{8.4.21}$$

Interference minima are observed when $\delta = m2\pi + \pi$, resulting in the condition

$$d = m\frac{\lambda_0}{2}\,. \tag{8.4.22}$$

The successive maxima or minima are thus separated by the distance $\Delta d = \lambda_0/2$. However, a shift from a maximum to minimum occurs in a distance $\Delta d = \lambda_0/4$. In principle it is also possible to measure fractional changes in the fringes, resulting in even higher sensitivity.

The Michelson interferometers can be used in a number of ways. Often one of the mirrors is at fixed location but the other moves with an object. By calculating fringes using a light source of known wavelength, one can then measure distances. Alternatively by counting fringes for two different wavelengths, one can calibrate an

unknown wavelength against a known one. The other possible application is to keep both mirrors fixed and vary the medium in one of the arms. The changes in fringes can then be correlated with the refractive index of the material.

It is also possible to use the Michelson interferometer using rays that propagate at some angle θ with respect to the axis. The distance d must then be replaced by $d\cos\theta$, resulting in fringes that depend on angle θ. With both mirrors fixed, one then gets fringes on a screen.

The configuration of Fig. 8.4 is sufficient if the light source has good coherence properties such as lasers. For more traditional sources, however, more care must be given to the configuration. In Fig. 8.5, for example, where the back surface of the beam splitter is 50 % reflecting, the ray in the horizontal arm passes once through the beam splitter, whereas the ray in the vertical arm passes three times through the beam splitter. The difference can be balanced by inserting a ***compensator*** in the horizontal arm in order to make both paths as identical as possible. The compensator is identical with the beam splitter with the exception that its neither surface is reflecting. In addition, a lens close to an extended source may be necessary to generate parallel beams of rays propagating in various angles to enter the interferometer. In such a case, another lens is needed to focus the fringe pattern on the screen.

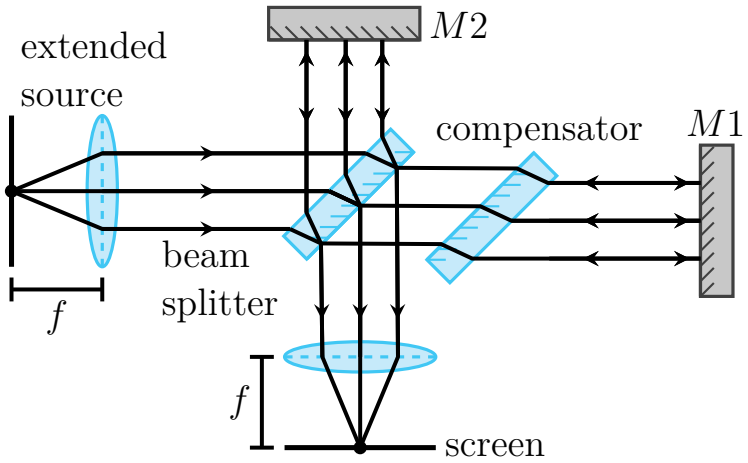

Fig. 8.5: Configuration of Michelson interferometer for extended and possibly incoherent sources.

8.5 Multiple-beam interference

We next move to the case where several waves need to be taken into account in calculating the superposition. The most common way of achieving this is to have a system

of two surfaces with high reflectivities. Conceptually, this can be achieved in the thin-film geometry of Fig. 8.6 by taking in account several reflected and transmitted rays. In practice, however, one needs to pay special attention on how to achieve the high reflectivities.

As an example of this situation, we consider Fig. 8.6, where a field with amplitude E_0 is incident on a system consisting of two parallel surfaces with high reflectivity. The material between the surfaces has refractive index of n_t. The material on the top and bottom sides is assumed to be identical with refractive index of n_1. We take the reflection and transmission coefficients of the interfaces for the field incident from the external medium as r and t, respectively. Similarly, the coefficients for the field incident from the internal medium are r' and t'. We further assume that the top and bottom interfaces are identical, although in several cases they are different.

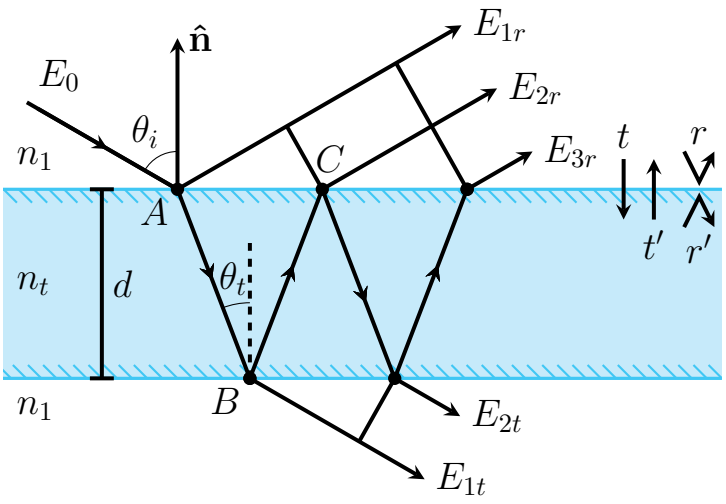

Fig. 8.6: Geometry for multiple-beam interference.

The total reflected and transmitted fields now consists of several partial waves E_{nr} and E_{nt}. It is evident on the basis of our former treatment of the thin film that the path difference between the reflected waves E_{2r} and E_{1r} is given by Eq. (8.3.16)

$$\Lambda_{2,1} = 2n_t d\cos\theta_t\,. \tag{8.5.23}$$

Similarly, by using point C as reference, the path difference between waves E_{3r} and E_{2r} is also $\Lambda_{3,2} = \Lambda_{2,1}$ and the path difference between waves E_{3r} and E_{1r} is thus $\Lambda_{3,1} = 2\Lambda_{2,1}$. In the same way, by first using point B as reference, we may conclude that the path difference between the transmitted waves E_{2t} and E_{1t} is also $\Lambda_{2,1}$. All of this can be summarized by noting that the phase difference arising from propagation between any two successive partial rays in either direction is given by the quantity

$$\delta = k\Lambda = \frac{2\pi}{\lambda_0}\Lambda = \frac{4\pi}{\lambda}n_t d\cos\theta_t\,, \tag{8.5.24}$$

where λ_0 is the vacuum wavelength.

Beyond this fundamental phase difference, there is an additional phase shift between the transmitted and reflected waves that arises from propagation between points A and B. We label this phase shift by ϕ and, as we will see, need not be concerned about its detailed form. In principle, the reflection and transmission coefficients, as given by the Fresnel coefficients [Eqs. (8.5.35), (8.5.36), (8.5.41) and (8.5.42)] could also be complex-valued quantities giving rise to additional phase shifts.

The total reflected field, referenced to point A, is thus

$$E_r = E_{1r} + E_{2r} + E_{3r} + \dots , \tag{8.5.25}$$

where the partial waves are

$$E_{1r} = rE_0 , \tag{8.5.26}$$
$$E_{2r} = t'r't'E_0 \mathrm{e}^{\mathrm{i}\delta} , \tag{8.5.27}$$
$$E_{3r} = (r')^2 t'r'tE_0 \mathrm{e}^{\mathrm{i}2\delta} . \tag{8.5.28}$$

There is thus a repeating factor $(r')^2\mathrm{e}^{\mathrm{i}\delta}$ between any two successive waves and the total reflected field becomes

$$E_r = rE_0 + t'r'tE_0\mathrm{e}^{\mathrm{i}\delta}\left[1 + (r')^2\mathrm{e}^{\mathrm{i}\delta} + (r')^4\mathrm{e}^{\mathrm{i}2\delta} + \dots + (r')^{2n}\mathrm{e}^{\mathrm{i}n\delta}\right] . \tag{8.5.29}$$

The second term in this equation is a converging geometric sequence up to order $n+1$, allowing the total field to be approximated ($n \to \infty$) as[2]

$$E_r = rE_0 + \frac{t'r'tE_0\mathrm{e}^{\mathrm{i}\delta}\left(1 - (r')^{2n}\mathrm{e}^{\mathrm{i}n\delta}\right)}{1 - (r')^2\mathrm{e}^{\mathrm{i}\delta}} \approx rE_0 + \frac{t'r'tE_0\mathrm{e}^{\mathrm{i}\delta}}{1 - (r')^2\mathrm{e}^{\mathrm{i}\delta}} . \tag{8.5.30}$$

Finally, the reflection and transmission coefficients for dielectric materials can be shown to fulfil the conditions $r = -r'$ and $tt' = 1 - r^2$. By using these relations, the total reflected field can be manipulated into the form

$$E_r = E_0 \frac{r\left(1 - \mathrm{e}^{\mathrm{i}\delta}\right)}{1 - r^2\mathrm{e}^{\mathrm{i}\delta}} . \tag{8.5.31}$$

We are usually interested in the irradiance of the field, and the reflected irradiance becomes

$$I_r = I_0 \frac{r^2\left(1 - \mathrm{e}^{\mathrm{i}\delta}\right)\left(1 - \mathrm{e}^{-\mathrm{i}\delta}\right)}{\left(1 - r^2\mathrm{e}^{\mathrm{i}\delta}\right)\left(1 - r^2\mathrm{e}^{-\mathrm{i}\delta}\right)} = I_0 \frac{2r^2\left(1 - \cos\delta\right)}{\left(1 + r^4 - 2r^2\cos\delta\right)} . \tag{8.5.32}$$

[2]Particularly, this ***geometric sequence*** is of form $\sum_{k=1}^{n} ar^{k-1} = \frac{a(1-r^n)}{1-r}$, when $r \neq 1$.

The total transmitted field, also referenced to point A, is

$$E_t = E_{1t} + E_{2t} + E_{3t} + \dots , \tag{8.5.33}$$

with the partial waves

$$E_{1t} = t't E_0 \mathrm{e}^{\mathrm{i}\phi} , \tag{8.5.34}$$

$$E_{2t} = (r')^2 t't E_0 \mathrm{e}^{\mathrm{i}\phi}\mathrm{e}^{\mathrm{i}\delta} . \tag{8.5.35}$$

The repeating factor is again $(r')^2\mathrm{e}^{\mathrm{i}\delta}$ and the total transmitted field then becomes

$$E_t = tt' E_0 \mathrm{e}^{\mathrm{i}\phi}\left[1 + (r')^2\mathrm{e}^{\mathrm{i}\delta} + \dots\right] \approx \frac{tt' E_0 \mathrm{e}^{\mathrm{i}\phi}}{1-(r')^2\mathrm{e}^{\mathrm{i}\delta}} = \frac{\left(1-r^2\right) E_0 \mathrm{e}^{\mathrm{i}\phi}}{1-r^2\mathrm{e}^{\mathrm{i}\delta}} . \tag{8.5.36}$$

The irradiance of this field is

$$I_t = I_0 \frac{(1-r^2)^2}{\left(1-r^2\mathrm{e}^{\mathrm{i}\delta}\right)\left(1-r^2\mathrm{e}^{-\mathrm{i}\delta}\right)} = I_0 \frac{\left(1-r^2\right)^2}{1+r^4-2r^2\cos\delta} . \tag{8.5.37}$$

It is straightforward to verify that energy is conserved, i.e.,

$$I_r + I_t = I_0 . \tag{8.5.38}$$

In the following, we will focus on the irradiance of the transmitted field, because this is the quantity that is usually detected in optical components and instruments relying on multiple-beam interference. In order to do this, we will manipulate the denominator of Eq. (8.5.37) by using the relation $\cos\delta = 1 - 2\sin^2(\delta/2)$, resulting in the form

$$1 + r^4 - 2r^2 + 4r^2\sin^2(\delta/2) = \left(1-r^2\right)^2 + 4r^2\sin(\delta/2) . \tag{8.5.39}$$

With this, the transmitted irradiance becomes

$$I_t = I_0 \frac{1}{1+\left(\frac{2r}{1-r^2}\right)^2\sin^2(\delta/2)} = I_0\mathcal{A} , \tag{8.5.40}$$

which is based on the definitions of the ***finesse coefficient***[3]

$$F = \left(\frac{2r}{1-r^2}\right)^2 , \tag{8.5.41}$$

and the ***Airy function*** defined compactly using the finesse coefficient F as

$$\mathcal{A} = \frac{1}{1+F\sin^2(\delta/2)} . \tag{8.5.42}$$

[3]Not to be confused with the related quantity ***finesse*** $\mathcal{F} = \frac{\pi\sqrt{F}}{2}$, defined later on.

Note that the Airy function $\mathcal{A}$ is unity whenever the sine function vanishes, i.e., when $\delta/2 = m\pi$. On the other hand, when the surfaces have high reflectivities, the finesse coefficient can be very large. In consequence, the Airy function becomes small when the sine function deviates just a little bit from zero. Multiple-beam interference can therefore give rise to very sharp transmission peaks as a function of the phase difference δ. Two examples of this are shown in Fig. 8.7. For low finesse coefficient ($F = 5$), the fringes are not very sharp and the transmission maxima and minima differ by about a factor of 10. For high coefficient ($F = 200$), the contrast is much better and the transmission maxima are quite narrow. It is important to note that much sharper fringes than shown in Fig. 8.7b can also be achieved.

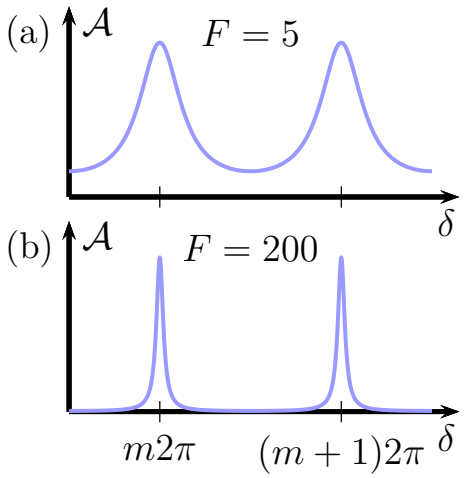

Fig. 8.7: $\mathcal{A}$ for (a) $F = 5$ and (b) $F = 200$.

8.6 Fabry–Pérot instruments

Instruments that are based on multiple-beam interference are known as ***Fabry–Pérot*** instruments. On the basis of Eq. (8.5.42), the transmission of a Fabry–Pérot instrument depends sensitively on the phase difference δ between two successive rays given by Eq. (8.5.24). The transmission thus depends on the vacuum wavelength λ_0, index of refraction between the two reflecting surfaces n_t, the distance between the surfaces d, and the angle of propagation between the surfaces θ_t. One can therefore use Fabry–Pérot instruments to study various phenomena as a function of any of these parameters. These dependences lead to two basic approaches for implementing a Fabry–Pérot instrument:

A) A ***Fabry–Pérot interferometer*** typically consists of two highly-reflecting mirrors for which the quantity $n_t d$ is varied (Fig. 8.8), usually by scanning the distance between the mirrors. The transmitted signal is then detected in the direction $\theta_t \approx 0°$ as a function of the quantity $n_t d$.

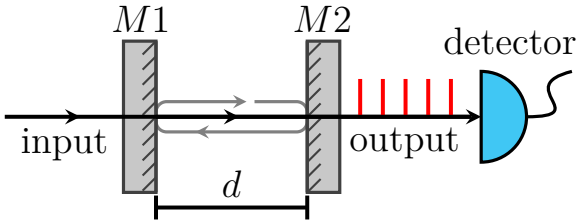

Fig. 8.8: Typical use of a Fabry–Pérot interferometer, where the distance d between two mirrors is varied.

B) An ***etalon*** consists of a solid piece of, e.g., glass whose both surfaces have been

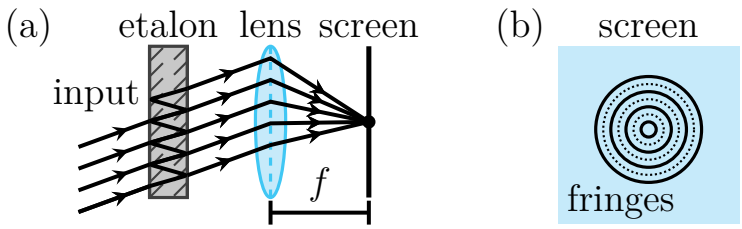

Fig. 8.9: In a Fabry–Pérot etalon light propagates at different angles (a), resulting in a fringe pattern that can be focused on a screen (b).

made highly reflecting (Fig. 8.9). The quantity $n_t d$ is therefore fixed and the interference fringes are measured at different angles θ_t. For a fixed wavelength, this results on a set of concentric fringes that can be visualized on a screen with the aid of a focusing lens.

In either case, it is important to understand how sharp the fringes are (Fig. 8.10). For Fabry–Pérot instruments, the width of the transmission maxima is usually characterized by their ***full width at half maximum (FWHM)*** γ, i.e., when the Airy function reaches the value 1/2. This is clearly obtained when $\sin^2(\delta/2) = 1/F$ or $\sin(\delta/2) = \pm 1/\sqrt{F}$. For large F, the deviation of the phase difference γ from $m2\pi$ can be only very small and we can measure the phase difference as this deviation. For such cases, the following condition is true $\sin(m\pi \pm \delta/2) \approx \pm\delta/2$. The FWHM value is therefore

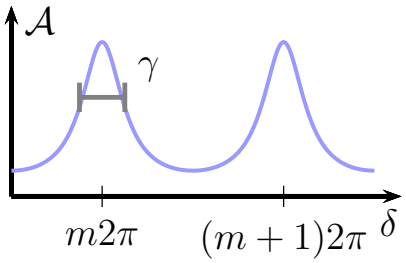

Fig. 8.10: Sharpness of Fabry–Pérot fringes is characterized by their FWHM width and separation.

$$\gamma = \frac{4}{\sqrt{F}}\,. \tag{8.6.43}$$

The relative sharpness of the fringes can be characterized by comparing their width to the separation between the maxima for orders m and $m+1$, which is 2π. The ratio between the separation and width is known as the ***finesse*** of the instrument, given by

$$\mathcal{F} = \frac{2\pi}{\gamma} = \frac{\pi\sqrt{F}}{2}\,. \tag{8.6.44}$$

A typical finesse $\mathcal{F}$ of an etalon is of the order of 10. For interferometers, however, the finesse $\mathcal{F}$ can be much higher, even reaching values of 100 000 and more.

8.7 Fabry–Pérot spectroscopy

Due to the possibility of achieving very sharp transmission fringes, Fabry–Pérot interferometers have important applications in ***spectroscopy***, i.e., in measuring how different quantities depend on wavelength, where the proper quantity is the vacuum wavelength λ_0 or frequency $\nu = c/\lambda_0$. The fundamental phase difference given by Eq. (8.5.24) can also be written as

$$\delta = \frac{4\pi}{\lambda_0} n_t d \cos\theta_t = \frac{4\pi\nu}{c} n_t d \cos\theta_t = \frac{4\pi\nu}{c} L \,, \tag{8.7.45}$$

where the quantity $L = n_t d \cos\theta_t$ can be interpreted as the effective length of the interferometer. A transmission resonance is obtained when $\delta = m2\pi$.

It is clear that when the effective length is varied, different frequencies (or wavelengths) become resonant for different effective lengths. In order m, the resonance frequency is obtained from

$$\frac{4\pi\nu}{c} L = m2\pi \,, \tag{8.7.46}$$

yielding the resonant frequency

$$\nu_m = \frac{mc}{2L} \,. \tag{8.7.47}$$

Alternatively, we can say that a given frequency ν becomes resonant in order m when the effective length is

$$L_m = \frac{mc}{2\nu} \,. \tag{8.7.48}$$

The interferometer is often used by first choosing a nominal value for the effective length L. The resonant frequency is then tuned by making small changes ΔL in the length. By differentiating Eq. (8.7.47), we obtain

$$\Delta\nu_m = -\frac{mc}{2L^2} \Delta L = -\nu_m \frac{\Delta L}{L} \,. \tag{8.7.49}$$

The subscripts here refer to the fact that we assume the frequency to be determined in order m.

The first thing we need to realize from Eq. (8.7.48) is that the resonance for a given frequency ν in orders m and $m+1$ corresponds to a length difference

$$\Delta L = L_{m+1} - L_m = \frac{c}{2\nu} \,. \tag{8.7.50}$$

In terms of frequency difference, this becomes

$$\Delta\nu_m = \frac{c}{2\nu} \frac{\nu}{L} = \frac{c}{2L} \,. \tag{8.7.51}$$

Note that this very result is obtained also from Eq. (8.7.47) by considering the resonance frequencies for orders m and $m+1$ and a fixed nominal length L. These results imply that there is a fundamental limitation to the range of frequencies that can be measured by the interferometer before the successive orders start overlapping each other. This range is known as the ***free spectral range*** of the instrument (Fig. 8.11)

$$\Delta\nu_{\text{fsr}} = \frac{c}{2L}\,, \tag{8.7.52}$$

which is very easy to remember as a rule of thumb. Note that when the material between the reflecting surfaces is air ($n_t \approx 1$) and detection is at small angle ($\theta_t \approx 0°$), even the effective and physical lengths are equal, i.e., $L = d$.

In doing spectroscopy, another key issue is to understand how small differences in frequency can be resolved, assuming that all frequencies are equally bright. For the case of two different frequencies ν_1 and ν_2, we thus have two different requirements (Fig. 8.11). First, the frequencies need to be sufficiently separated in a given order m. Second, we need to avoid overlap between ν_1 in order m and ν_2 in order $m+1$.

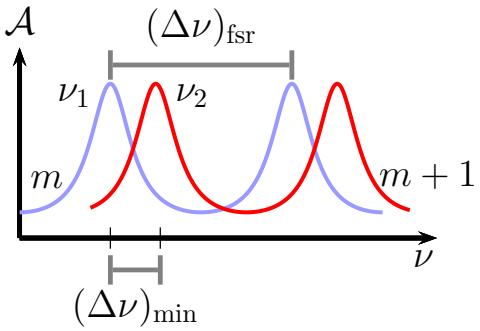

Fig. 8.11: Separation of two different frequencies using a Fabry–Pérot instrument.

We can argue that the two frequencies are resolved in order m, when they are separated by at least the FWHM of the transmission maximum given by Eq. (8.6.43), i.e., the following condition is satisfied

$$\Delta\delta_{\text{min}} = \frac{4\pi\Delta\nu_{\text{min}}}{c}L \approx \gamma \approx \frac{4}{\sqrt{F}}\,. \tag{8.7.53}$$

This implies that the frequency separation must be at least

$$\Delta\nu_{\text{min}} \approx \frac{c}{L\pi\sqrt{F}}\,. \tag{8.7.54}$$

The ***resolving power*** of a spectroscopic instrument is defined as the ratio between the measured quantity and its smallest observable change. In the present case, it is

$$\frac{\nu_m}{\Delta\nu_{\text{min}}} = m\frac{\pi\sqrt{F}}{2} = m\mathcal{F}\,. \tag{8.7.55}$$

This result is not a big surprise because we know from Eq. (8.6.44) that the finesse is related to the sharpness of the Fabry–Pérot fringes.

9. DIFFRACTION

9.1 Basic theory

When the superposition principle is applied to an infinite number of infinitesimal sources, the resulting effects are known as ***diffraction*** effects. These results can be understood on the basis of the ***Huygens–Fresnel principle***, where each point of a wavefront is considered as a source of secondary radiation. The new wave front is then obtained as a superposition of all the elementary wavelets, i.e., their amplitude and phase needs to be taken into account.

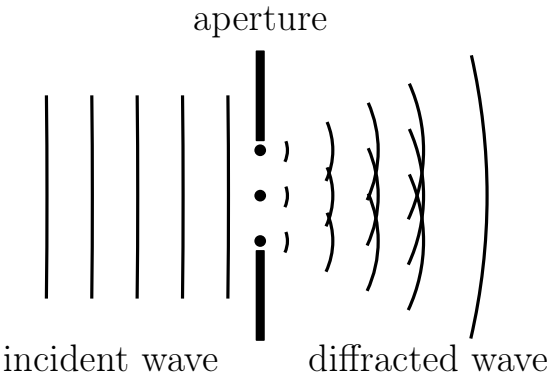

Fig. 9.1: Huygens–Fresnel principle of diffraction by an aperture.

This principle is illustrated in Fig. 9.1, where a plane wave is incident on a screen with an aperture. In a more general case, the form of the incident wave could be even more arbitrary. Each point of the aperture is taken to be a source of spherical waves, whose amplitude is proportional to that of the incident wave. The general form of the spherical wave was of form

$$E(r) = \frac{E_A(r)}{r} \mathrm{e}^{\mathrm{i}kr} \,, \tag{9.1.1}$$

where we have assumed that the field can be treated within the scalar approximation and have omitted the harmonic time dependence $\mathrm{e}^{\mathrm{i}\omega t}$. Furthermore, r is the distance from a given source point to the observation point $\mathbf{R}$. The total field at the observation point $\mathbf{R}$ is then obtained by integrating over all the possible spherical waves within the aperture S, i.e.,

$$E_{\mathrm{tot}}(\mathbf{R}) = \int\limits_S \frac{E_A(r)}{r} \mathrm{e}^{\mathrm{i}kr} \, \mathrm{d}S \,, \tag{9.1.2}$$

Of course, this form as such is rather useless as it only expresses a general principle.

A more detailed treatment of diffraction requires that Eq. (9.1.2) be formulated in a proper mathematical way. The general case is quite complicated but it can be simplified by using appropriate approximations. ***Fresnel diffraction*** is applicable as long as the observation point is at least some distance away from the aperture, but this case is still quite complicated. The simplest case is obtained by ***Fraunhofer diffraction***, which requires that the observation point is sufficiently far from the aperture. This regime is known as the ***far field*** or ***far zone***. In the following, we will only consider Fraunhofer diffraction.

9.2 Fraunhofer diffraction

In order to treat Fraunhofer diffraction properly, we consider the geometry of Fig. 9.2. A wave propagating mostly in the positive z direction illuminates an aperture in an opaque screen, which is in the x-y plane. A characteristic dimension of the aperture is a and it is in some sense centered near the origin.

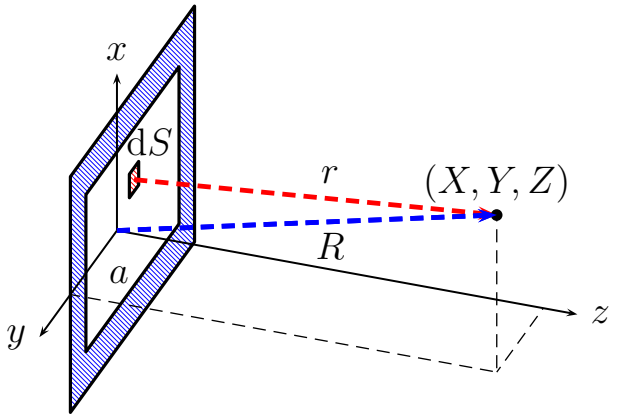

Fig. 9.2: Geometry for Fraunhofer diffraction.

We first consider a source element $\mathrm{d}S = \mathrm{d}x\,\mathrm{d}y$ at an arbitrary location $\mathbf{r} = (x, y, 0)$ in the aperture. Our observation point is at plane $z = Z$ and we label this point by upper case letters, i.e., the point is $\mathbf{R} = (X, Y, Z)$. The distance from the source point to the observation point is then

$$r = \sqrt{(X - x)^2 + (Y - y)^2 + Z^2}\,, \tag{9.2.3}$$

By inserting this into Eq. (9.1.2), it is easy to realize that the integration over the aperture plane becomes extremely difficult.

In order to simplify the treatment, we first assume that the observation point is far away from the aperture. In this case, the factor $1/r$ can be assumed to be almost constant and can be replaced to good accuracy by $1/R$. The superposition, however, is very sensitive to phase effects. Hence, more careful treatment is required for the phase factor $\mathrm{e}^{\mathrm{i}kr}$. For this purpose, we expand Eq. (9.2.3) into the form

$$\begin{aligned} r^2 &= X^2 - 2Xx + x^2 + Y^2 - 2Yy + y^2 + Z^2\,, \\ &= R^2 + x^2 + y^2 - 2(Xx + Yy)\,, \\ &= R^2 \left[1 + \frac{x^2 + y^2}{R^2} - \frac{2(Xy + Yy)}{R^2}\right]\,. \end{aligned} \tag{9.2.4}$$

By assuming that the source dimensions are much smaller than the distance to the observation point ($x, y \ll R$), the square root of Eq. (9.2.4) can be approximated by

($\sqrt{1+a} \approx 1 + a/2$, when $a \ll 1$)

$$r = R\left[1 + \frac{x^2+y^2}{2R^2} - \frac{2(Xy+Yy)}{2R^2}\right],$$
$$= R + \frac{x^2+y^2}{2R} - \frac{(Xy+Yy)}{R}. \tag{9.2.5}$$

Within the present approximation, the second term of this equation has small quantities of second-order in the numerator. We will next estimate under which conditions this term could be neglected. For our aperture, the transverse size is of the order a, hence the second term is at most of the order $a^2/2R$. Its contribution to the phase of the spherical wave is then at most

$$k\frac{a^2}{2R} = \frac{2\pi}{\lambda}\frac{a^2}{2R} = \pi\frac{a^2}{\lambda R}. \tag{9.2.6}$$

The contribution to the phase is certainly negligible when this term is much smaller than unity. The Fraunhofer or far-field diffraction occurs therefore in the regime where

$$R > a^2/\lambda. \tag{9.2.7}$$

When this condition is fulfilled, the approximation for the distance between the source and observation points becomes

$$r = R - \frac{Xx+Yy}{R}. \tag{9.2.8}$$

The elementary spherical wave at the observation point $(X,\, Y)$ emitted from the differential source element $\mathrm{d}S$ is then

$$E(X,Y) = \frac{E_A(x,y)}{R}\mathrm{e}^{\mathrm{i}kR}\mathrm{e}^{-\mathrm{i}k(Xx+Yy)/R}, \tag{9.2.9}$$

and the total field at the observation point $(X,\, Y)$ becomes[1]

$$E_{\mathrm{tot}}(X,Y) = \iint\limits_S \frac{E_A(x,y)}{R}\mathrm{e}^{\mathrm{i}kR}\mathrm{e}^{-\mathrm{i}k(Xx+Yy)/R}\,\mathrm{d}x\,\mathrm{d}y,$$
$$\approx \frac{\mathrm{e}^{\mathrm{i}kR}}{R}\iint\limits_S E_A(x,y)\mathrm{e}^{-\mathrm{i}k(Xx+Yy)/R}\,\mathrm{d}x\,\mathrm{d}y. \tag{9.2.10}$$

[1]Note that this integral should be repeated for other observation points $(X',\, Y')$ of interest in order to calculate the actual diffraction pattern.

Note that, mathematically, Eq. (9.2.10) corresponds to a two-dimensional spatial Fourier transform of the aperture S.[2]

The results of Eq. (9.2.10) can be interpreted in two different ways. The first interpretation is that the quantities X/R and Y/R are linked to the propagation angles of the rays from the aperture to the observation point. More specifically, we can define $\sin\theta_X = X/R$ and $\sin\theta_Y = Y/R$. As usual, such rays propagating at a given angle can be focused by a lens to its back focal plane. The lens thus brings the far field to a more practical distance. The second interpretation is useful when the transverse coordinates X and Y of the observation point are small compared to the distance between the aperture and observation point. We then have $R \approx Z$ and we can place a screen at plane $z = Z$ to observe the diffraction pattern.

9.3 Diffraction from basic aperture shapes

We first consider diffraction from a narrow slit with a width a in x direction (Fig. 9.3a). The slit extends from $x = -a/2$ to $x = a/2$. The integral of Eq. (9.2.10) is straightforward to calculate using the Cartesian coordinate system where $\mathrm{d}S = \mathrm{d}x\,\mathrm{d}y$. The integrals in the x and y directions are calculated separately.[3] We take the field at the aperture to be a constant field normalized to unity $E_A(x, y) = 1$. The result in y direction becomes then

$$E_{\mathrm{tot}}(Y) = \frac{\mathrm{e}^{\mathrm{i}kR}}{R} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}kYy/R}\,\mathrm{d}y = \frac{\mathrm{e}^{\mathrm{i}kR}}{R}\delta(kY/R)\,. \tag{9.3.11}$$

This result states that the field in the observation point Y is affected only by the ray originating from the aperture plane propagating towards that point ($k' = kY/R = k\sin\theta_Y$). The δ-function essentially picks out this ray from the collection of rays that form the original spherical wave. In other words, no diffraction occurs in y direction. This occurs because the aperture does not limit the transmission of light in this direction.

[2] This is a very powerful result, as it connects the source of the diffracted fields to its Fourier transform which could even be measured in the far field. In addition, it allows to utilize the known properties of the Fourier transforms in order to calculate such diffraction integrals.

[3] This remains valid as long as the function of interest is ***multiplicatively separable function*** in x and y. Here, we restrict our treatment to these kinds of functions.

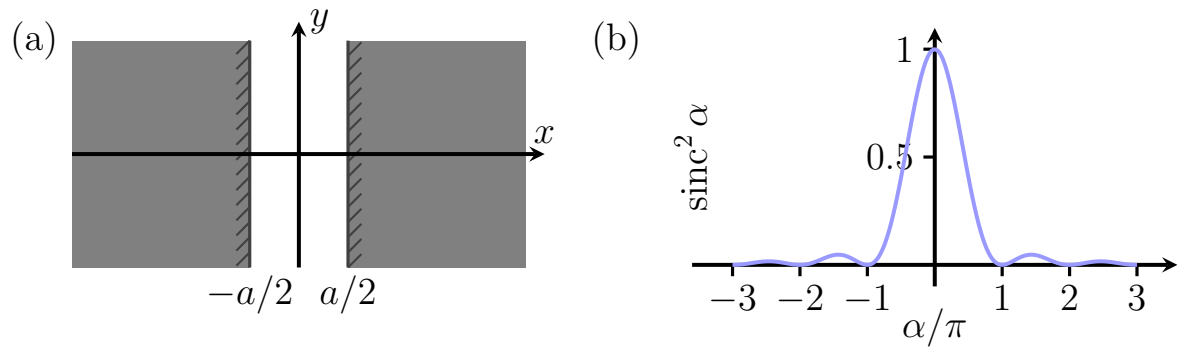

Fig. 9.3: Diffraction from a narrow slit of width a. (a) Geometry of slit. (b) Diffraction pattern of the slit in x direction as a function of the normalized variable $\alpha = kXa/2R$.

The integral in x direction becomes[4]

$$\int_{-a/2}^{a/2} \mathrm{e}^{-\mathrm{i}kXx/R}\,\mathrm{d}x = -\frac{R}{\mathrm{i}kX}\left(\mathrm{e}^{-\mathrm{i}kXa/2R} - \mathrm{e}^{\mathrm{i}kXa/2R}\right),$$

$$= -\frac{a}{2\mathrm{i}}\frac{2R}{kXa}\left(\mathrm{e}^{-\mathrm{i}kXa/2R} - \mathrm{e}^{\mathrm{i}kXa/2R}\right). \qquad (9.3.12)$$

In order to manipulate this further, we define a variable $\alpha = kXa/2R$, which can be interpreted as a renormalized direction of propagation or location at the observation plane. Using this variable, Eq. (9.3.12) becomes

$$\int_{-a/2}^{a/2} \mathrm{e}^{-\mathrm{i}kXx/R}\,\mathrm{d}x = -\frac{a}{2\mathrm{i}}\frac{2R}{kXa}\left(\mathrm{e}^{-\mathrm{i}kXa/2R} - \mathrm{e}^{-\mathrm{i}kXa/2R}\right),$$

$$= a\,\mathrm{sinc}\,\alpha\,. \qquad (9.3.13)$$

The irradiance at location X (or direction X/R) can therefore be written as

$$I(X) = I(0)\frac{\sin^2\alpha}{\alpha^2} = I(0)\,\mathrm{sinc}^2\,\alpha\,. \qquad (9.3.14)$$

This is shown in Fig. 9.3b as a function of the normalized coordinate $\alpha = kXa/2R$. It is clear that as the width of the slit a becomes narrower, the diffraction pattern becomes broader as a function of X or X/R, as expected because of the Fourier transform relation between the aperture and the diffraction pattern.

[4]Alternatively, one could extend the limits of integration to infinities and calculate the ***Fourier transform*** of a properly-sized ***rectangular function***.

The results for a slit are easily generalized for a rectangular aperture with dimensions a and b in the x and y directions, respectively (Fig. 9.4). Using normalized variables $\alpha = kXa/2R$ and $\beta = kYb/2R$, the result is

$$I(X,Y) = I(0,0)\,\mathrm{sinc}^2\,\alpha\,\mathrm{sinc}^2\,\beta\,. \qquad (9.3.15)$$

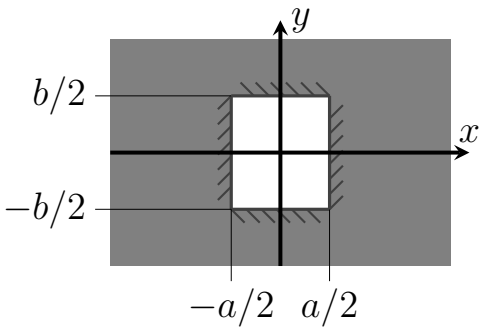

Fig. 9.4: Geometry of rectangular aperture.

We next consider the case of a circular aperture with radius a, i.e., diameter $D = 2a$ (Fig. 9.5a). This case is difficult to treat using Cartesian coordinates, prompting us to instead use cylindrical coordinates. For this purpose, we define radius q at the observation plane. Because of the cylindrical symmetry of the problem, the diffraction pattern must have circular symmetry at the observation plane. Without going into the details of the calculation, the irradiance as a function of the radius at the observation plane is[5]

$$I(q) = I(0)\left[\frac{2J_1(kaq/R)}{kaq/R}\right]^2, \qquad (9.3.16)$$

where $J_1(kaq/R)$ is a ***Bessel function***. The result of Eq. (9.3.16) is plotted in Fig. 9.5b as a function of a renormalized variable $u = kaq/R$.

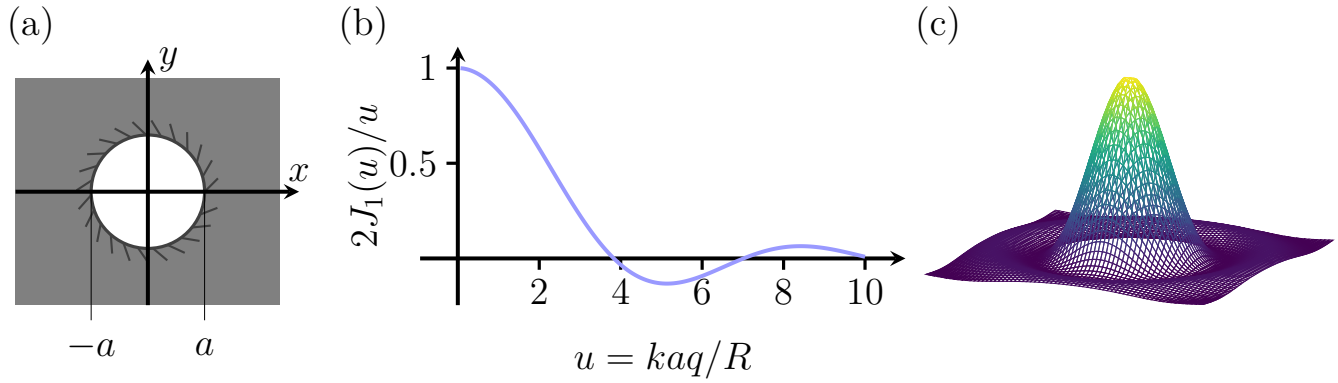

Fig. 9.5: Diffraction from a circular aperture with diameter $D = 2a$. (a) Geometry of aperture. (b) Diffraction pattern of the slit as a function of the normalized variable $u = kaq/R$. (c) Two-dimensional illustration of the diffraction profile.

The width of the central maximum of the diffraction pattern can be characterized by the first zero of $J_1(kaq/R)$, which occurs for $kaq/R \approx 3.83$. This results in the

[5]This is rather straightforward calculation by using the known Fourier transform of a ***circ function***.

condition

$$\frac{kaq}{R} = \frac{2\pi}{\lambda}\frac{aq}{R} = 3.83\,, \tag{9.3.17}$$

for the minimum. The minimum occurs thus at the location of the observation plane

$$q \approx \frac{3.83}{\pi}\frac{\lambda R}{2a} \approx 1.22\frac{\lambda R}{D}\,, \tag{9.3.18}$$

or in terms of the direction of propagation

$$\sin\theta_r = \frac{q}{R} \approx 1.22\frac{\lambda}{D}\,. \tag{9.3.19}$$

Of course, for small angles, we can use the approximation $\sin\theta_r \approx \theta_r$, which is often useful.

9.4 Diffraction from multiple slits

We finally consider a system that consists of several narrow slits periodically ordered along the x direction (Fig. 9.6a). The width of each slit is a and they are centered at locations Nb. The integral of Eq. (9.2.10) is again evaluated separately in the x and y directions with the y direction resulting in a δ-function. The integral in the x direction, on the other hand, becomes

$$\begin{aligned}\sum_{n=0}^{N-1}\int_{nb-a/2}^{nb+a/2} \mathrm{e}^{-\mathrm{i}kXx/R}\,\mathrm{d}x &= \sum_{n=0}^{N-1} -\frac{R}{\mathrm{i}kX}\left[\mathrm{e}^{-\mathrm{i}kX(nb+a/2)/R} - \mathrm{e}^{-\mathrm{i}kX(nb-a/2)/R}\right]\,,\\ &= \sum_{n=0}^{N-1} \mathrm{e}^{-\mathrm{i}kXnb/R}\left(-\frac{R}{\mathrm{i}kX}\right)\left[\mathrm{e}^{-\mathrm{i}kXa/2R} - \mathrm{e}^{\mathrm{i}kXa/2R}\right]\,. \end{aligned}\tag{9.4.20}$$

The two last factors under the sum, which are in brackets, represent the diffraction from a single slit and can be evaluated as before for the case of a single slit. The result is $a\sin\alpha/\alpha = a\,\mathrm{sinc}\,\alpha$, where $\alpha = kXa/2R$.[6]

[6]Here, we use the so-called unnormalized definition for the ***sinc-function***.

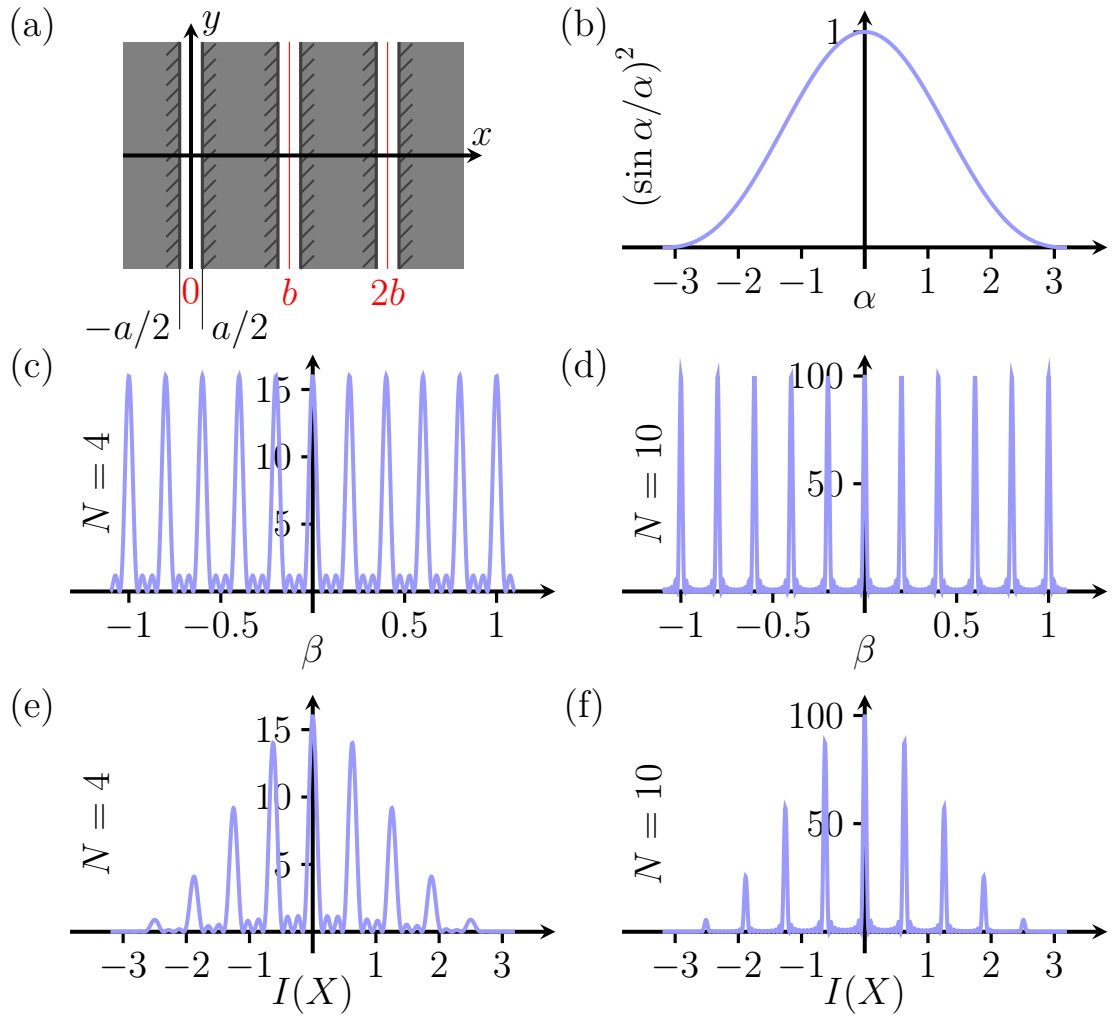

Fig. 9.6: (a) Periodic structure with several narrow slits of width a. The period is $b = 5a$. (b) The factor representing diffraction from a single slit. (c) The factor representing interference of $N = 5$ slits. (d) The factor representing interference of $N = 10$ slits. (e) Diffracted intensity pattern caused by interference of $N = 5$ slits. (f) Diffracted intensity pattern caused by interference of $N = 10$ slits.

We are thus left with the evaluation of the sum in Eq. (9.4.20). By defining a variable $\beta = kXb/2R$, the sum becomes

$$\sum_{n=0}^{N-1} \mathrm{e}^{-\mathrm{i}n2\beta} = \frac{1-\mathrm{e}^{-\mathrm{i}2N\beta}}{1-\mathrm{e}^{-\mathrm{i}2\beta}} = \frac{\mathrm{e}^{-\mathrm{i}N\beta}}{\mathrm{e}^{-\mathrm{i}\beta}} \frac{\mathrm{e}^{\mathrm{i}N\beta}-\mathrm{e}^{-\mathrm{i}N\beta}}{\mathrm{e}^{\mathrm{i}\beta}-\mathrm{e}^{-\mathrm{i}\beta}} \,.$$
$$= \mathrm{e}^{-\mathrm{i}(N-1)\beta} \frac{\sin N\beta}{\sin\beta} \tag{9.4.21}$$

The total diffracted field is thus proportional to

$$E(X) \propto a\mathrm{e}^{-\mathrm{i}(N-1)\beta}\frac{\sin\alpha}{\alpha}\frac{\sin N\beta}{\sin\beta}\,, \tag{9.4.22}$$

and its irradiance can be written as

$$I(X) = I(0)\left(\frac{\sin\alpha}{\alpha}\right)^2\left(\frac{\sin N\beta}{\sin\beta}\right)^2\,. \tag{9.4.23}$$

Here, the first angular factor corresponds to diffraction from a single slit. The second angular factor, on the other hand, corresponds to interference between the diffraction patterns from several slits. These factors and their product are shown in Fig. 9.6b–9.6d for the case $b = 5a$ and $N = 5$ and $N = 10$. Note that the factor representing diffraction (Fig. 9.6b) exhibits only weak modulation for the parameters chosen. The factor representing interference (Fig. 9.6c and 9.6d), on the other hand, gives rise to sharp maxima when $\beta = m\pi$. In addition, the maxima become narrower and more intense as the number of slits is increased. The sharp maxima thus arise from interference between several similar sources, whereas the diffraction of a single slit results only in weak modulation of the overall pattern. The overall diffracted intensity patterns corresponding to Eq. (9.4.23) are shown in Figs. 9.6e and 9.6f for the cases of $N = 5$ and $N = 10$, respectively.

9.5 Diffraction gratings

The discussion of the previous section showed that the diffraction from several identical slits is dominated by interference between the elementary diffraction patterns produced by each slit. This provides the basis for ***diffraction gratings***. Such gratings are components that lead to a periodic modulation of the amplitude and/or phase of an incident wave. We can then argue that each period acts as a source of its own diffraction pattern, whose form can be quite arbitrary. However, the overall pattern is dominated by interference between a number of elementary wavelets produced by each period, as given by Eq. (9.4.22).

The maxima of diffraction gratings are obtained by requiring that the phase difference between two successive wavelets is $m2\pi$, where m is the ***order of diffraction***. We note that although the effect is dominated by interference, it is nevertheless common to talk about diffraction orders. Diffraction gratings can operate in transmission or reflection (Fig. 9.7). In addition, the structure of the grating can be modulated in such a way as to enhance a certain diffraction order in transmission or reflection.

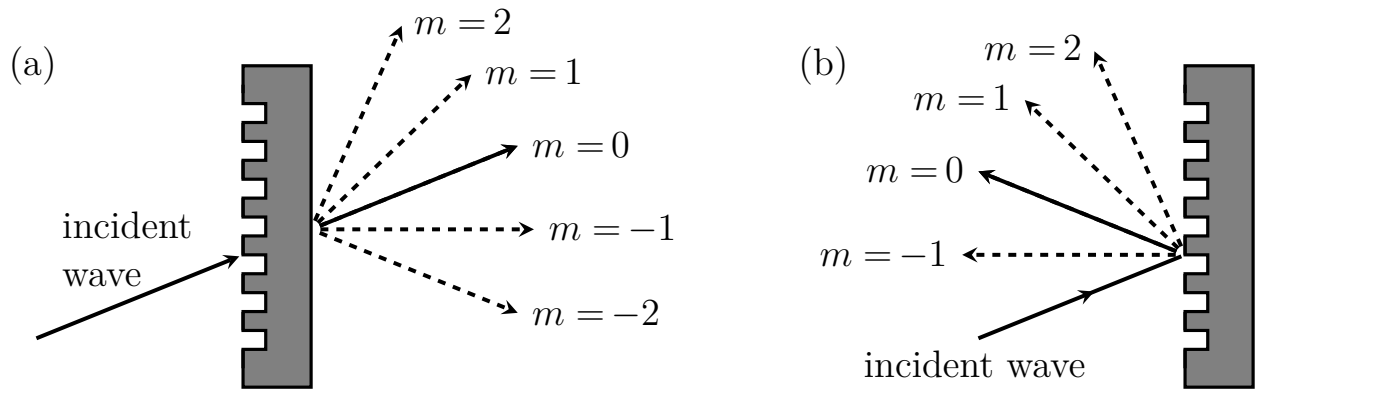

Fig. 9.7: Diffraction of incident wave to different diffraction orders m by a grating. (a) Transmission grating. (b) Reflection grating.

In order to understand how the maxima of diffraction gratings are calculated, we consider the case of a transmission grating shown in Fig. 9.7a. The simple requirement is that the rays from the respective parts of each period have a phase difference of $m2\pi$. In the geometry of Fig. 9.8, a plane wave is incident on the grating at angle θ_i and we are interested in diffracted waves that propagate in a direction given by the angle θ_m for a given order m. The total path difference between the two rays from the respective positions of successive periods is now

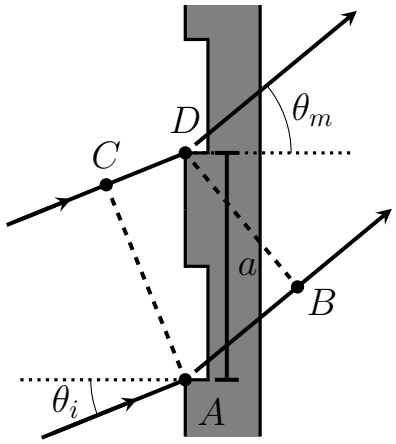

Fig. 9.8: Diffraction of incident light to order m by a transmission grating.

$$\mathrm{OPL} = AB - CD = a(\sin\theta_m - \sin\theta_i)\,, \qquad (9.5.24)$$

where a is the period of the grating. This results in a phase difference

$$\delta = \frac{2\pi}{\lambda}\mathrm{OPL} = \frac{2\pi}{\lambda}a(\sin\theta_m - \sin\theta_i)\,, \qquad (9.5.25)$$

and the condition ($\delta = m2\pi$) for a maximum of order m results in

$$a(\sin\theta_m - \sin\theta_i) = m\lambda\,. \qquad (9.5.26)$$

The sharpness of the maxima will again depend on the number of periods that give rise to the diffraction pattern.

10. GEOMETRICAL OPTICS

10.1 Basic definitions

We will next start discussing how optical systems can be used to form images of objects. Usually, the object is either illuminated with an external light source or the object emits radiation itself. In both cases, each point of the object is a source of radiation, emitting a bunch (beam) of rays that diverge from the object point (Fig. 10.1). In practice, only part of these rays can be collected by an optical system. Similarly, the optical system can manipulate the beam of rays in such a way that in the end they converge towards a point.

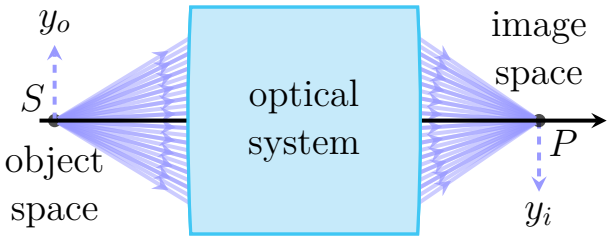

Fig. 10.1: A general optical system collects a spherical wave from the object point S and converges it to the image point P.

The goal of an ***imaging system*** is thus to collect a diverging beam from a source point S and converge it to an image point P (Fig. 10.1). When discussing such systems, we use the convention that light propagates from left to right. The ***object space*** is the part of space that is on the left-hand side of the system. Similarly, the ***image space*** is on the right-hand side.

The system is said to be ***stigmatic*** if there is one-to-one correspondence between beams of rays diverging from the object point S and converging to the image point P. In this case, P is the perfect image of S. The reciprocity present in most of light–matter interactions implies that the propagation of light between any two points is reversible, following the same path in both directions. Thus, it is possible to use a given system in the reversed direction, where P is treated as object and S as image. The distinction between the object and image points is therefore somewhat arbitrary and the points S and P are also referred to be ***conjugate*** points of each other.

So far, we have discussed the propagation of light between two points, one in the object and one in the image space. To form the image of an extended object, we need to do the same for several different object points. Ideally (Fig. 10.2a), each object point is imaged to a single image point. In practice, this is very difficult and the image of a single object point is an extended dot in the neighborhood of the ideal image point (Fig. 10.3b). The quality of the imaging system therefore greatly depends on how small this dot can be made.

The ultimate limitation to the size of the image dot arises from the wave character of light. The propagation of waves is limited by ***diffraction*** with the main result

that the image dot has a size that is comparable to wavelength. Whenever this limit is reached, the system is said to be ***diffraction-limited system*** and cannot be improved further.

Geometrical optics is the approach where the wave character of light is neglected. In essence, it is based on the assumption that the wavelength is zero.[1] It is then sufficient to consider only the propagation of geometrical rays through the optical system with no need to consider the wave propagation. In this case, the diffraction limit becomes the ideal image point instead of an extended dot. In practice, however, optical systems are also geometrically imperfect and the image of a dot again becomes an extended dot. These imperfections arise from the ***aberrations*** of the system, and the design of an optical system usually concentrates on minimizing those aberrations that are critical for a given application.We also have to make distinction between a ***real image*** and a ***virtual image***. A real image is formed when the rays in the image space converge towards a point. Such an image point can therefore be visualized on a screen. A virtual image is formed when the rays in the image space are diverging but appear to originate from a point. Although these rays occur in the image space, their point of origin may appear to be in the object space.

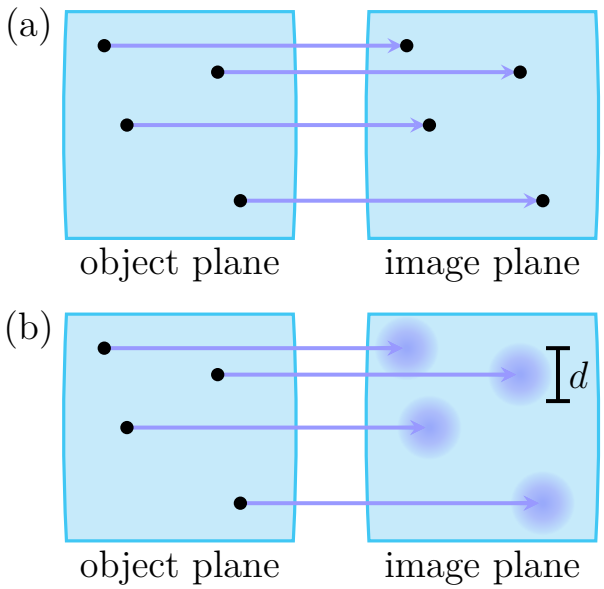

Fig. 10.2: For an ideal optical system (a), each object point is imaged to a single image point. In practice (b), the image is a dot with finite size d.

10.2 Refraction at a spherical surface

Most optical systems consist of or at least contain lenses. Traditionally, both surfaces of a lens have been ***spherical***, just because they have been easiest to make. Grinding two glass surfaces against each other naturally gives rise to a spherical form. It is important to note, however, that spherical surfaces are not ideal from the viewpoint of aberrations. As the fabrication techniques have evolved, more and more optical systems now contain ***aspherical*** surfaces.

[1] Geometrical optics can be derived from the Maxwell's equations by presenting the problem as an ***Eikonal equation***, which is intimately connected to the ***Fermat's principle***.

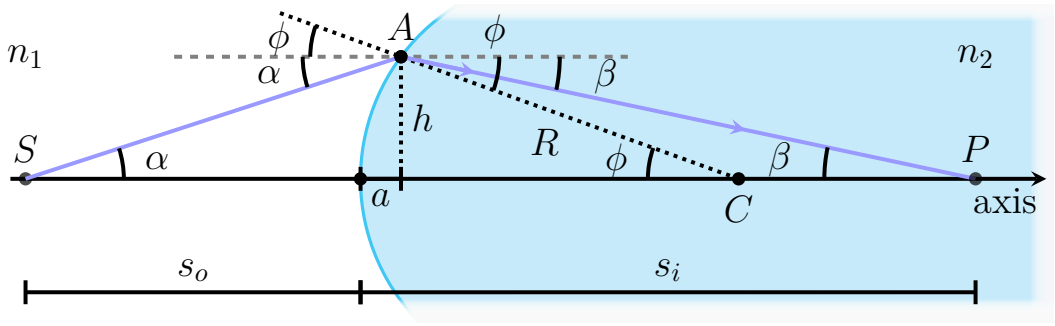

Fig. 10.3: Refraction of a ray at a spherical surface.

Nevertheless, we now proceed to discuss the refraction of light at spherical surfaces. We consider the situation shown in Fig. 10.3 where we assume that our object and image points lie on a line that is also a normal to the spherical surface. This line is known as the ***axis*** of our simple system. The spherical surface has a radius of curvature R. The object and image spaces have refractive indices of n_1 and n_2, respectively.

We consider an object point S that is located at the ***object distance*** s_o from the point where the surface and the axis intersect (apex of the surface). We follow through the system a ray that starts from the object point and makes an upward angle α with the axis. This ray meets the refracting surface at height h from the axis (point A) and is refracted in such a way that it propagates downward at angle β with respect to the axis. Consequently, the ray meets the axis at the ***image distance*** s_i from the apex. We can now hypothetically expect that this point could be the image point P of the object point S because we have two rays from S to the same point P. The other ray, of course, propagates along the axis and is not deviated by the surface.

In order to understand how our ray is refracted by the surface, we apply the law of refraction at A using the local surface normal that passes through the center of curvature C of the surface. Using the notations of Fig. 10.3, the angles of incidence and refraction are $\alpha + \phi$ and $\phi - \beta$, respectively, where ϕ is the angle between the local surface normal and the axis. The law of refraction is thus

$$n_1 \sin(\alpha + \phi) = n_2 \sin(\phi - \beta)\,. \tag{10.2.1}$$

In addition, the geometry of Fig. 10.3 gives us the following relations

$$\tan\alpha = \frac{h}{s_0 + a}\,, \quad \tan\beta = \frac{h}{s_i - a}\,, \quad \sin\phi = \frac{h}{R}\,, \quad a = R - R\cos\phi\,. \tag{10.2.2}$$

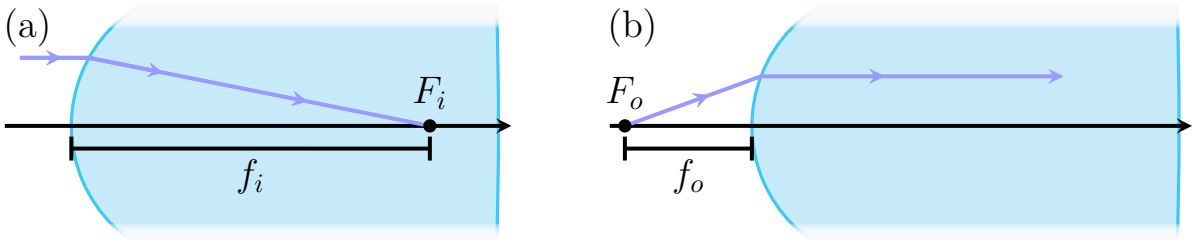

Fig. 10.4: (a) Back focal length f_i and (b) front focal length f_o of a spherical surface.

It turns out that these equations have no easy solution in general. In fact, they cannot be solved in a closed form for the relation between the object and image distances.

We therefore need to seek an approximate solution that is valid in certain important cases. In order to do this, we assume that the rays of interest propagate almost parallel to the axis. This is known as the ***paraxial approximation***, leading to the very special case of geometrical optics, known with a number of names, e.g., ***paraxial optics***, ***Gaussian optics***, and ***first-order optics***. Note that very similar terms are used also in other contexts but here we are discussing only geometrical optics.

Under this approximation, all angles in Eqs. (10.2.1) and (10.2.2) are small and we can replace them by their lowest-order approximations, e.g., $\sin\phi \approx \phi$ or $\cos\phi \approx 1$. Equations (10.2.1) and (10.2.2) then become

$$n_1(\alpha + \phi) = n_2(\phi - \beta)\,, \tag{10.2.3}$$

and

$$\tan\alpha \approx \frac{h}{s_0}\,, \quad \tan\beta \approx \frac{h}{s_i}\,, \quad \sin\phi \approx \frac{h}{R}\,, \quad a \approx 0\,. \tag{10.2.4}$$

From these, we first obtain

$$n_1\left(\frac{h}{s_0} + \frac{h}{R}\right) = n_2\left(\frac{h}{R} - \frac{h}{s_i}\right)\,, \tag{10.2.5}$$

and further, by reorganizing

$$\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}\,. \tag{10.2.6}$$

It is crucial that the result connecting the object and image distances is independent of the assumed height h. This implies that, within the paraxial approximation, all rays originating from S cross the axis in the image space at point P, i.e., P is the image of S.

Table 10.1: Sign conventions for a spherical surface.

Quantity	Sign	Location from surface	Type of object or image
s_o , f_o	positive/negative	left/right	real/virtual
s_i , f_i	positive/negative	right/left	real/virtual
R	positive/negative	right/left	

We next apply this result to two special cases (Fig. 10.4). Let's first consider the situation where the object is infinitely far so that the object distance is $s_o = \infty$. The image is then formed at the distance of the ***back focal length*** from the surface, which is found to be

$$f_i = \frac{n_2}{n_2 - n_1} R \,. \tag{10.2.7}$$

The image point is known as the back focal point F_i. Similarly, an object whose image is formed at $s_i = \infty$ is located at the distance of the front focal length from the surface, which is

$$f_o = \frac{n_1}{n_2 - n_1} R \,. \tag{10.2.8}$$

The object is then at the front focal point F_o of the surface. Note that the back and front focal lengths are not equal.

The situation shown in Fig. 10.3 represents a very special case where all quantities shown are positive. It turns out, however, that the results are equally applicable when some of the quantities are negative. The sign conventions and their interpretation for Fig. 10.3 are shown in Table 10.1.

10.3 Thin lenses

We next construct a lens by combining two spherical surfaces (Fig. 10.5). The lens material has refractive index of n_l. For simplicity, we assume that the lens is surrounded by a material with refractive index of n_m. In a more general case, the refractive indices on the two sides of the lens could be different. Finally, the lens has a thickness d as measured on axis.

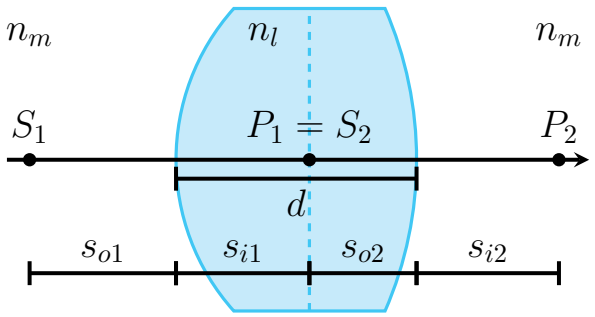

Fig. 10.5: Lens constructed from two spherical surfaces.

We first apply Eq. (10.2.8) to the first sur-

face

$$\frac{n_m}{s_{o1}} + \frac{n_l}{s_{i1}} = \frac{n_l - n_m}{R_1}\,, \tag{10.3.9}$$

where the number in the subscript refers to the surface. This equation is valid even for the case where the image distance s_{i1} extends beyond the physical thickness of the lens. It is then evident that the object distance for the second surface is $s_{o2} = d - s_{i1}$. Similarly, for the second surface, we have

$$\frac{n_l}{s_{o2}} + \frac{n_m}{s_{i2}} = \frac{n_m - n_l}{R_2}\,. \tag{10.3.10}$$

By adding Eqs. (10.3.9) and (10.3.10), we obtain

$$\frac{n_m}{s_{o1}} + \frac{n_l}{s_{i1}} + \frac{n_l}{d - s_{i1}} + \frac{n_m}{s_{i2}} = (n_l - n_m)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\,, \tag{10.3.11}$$

which can be further arranged into the form

$$\frac{n_m}{s_{o1}} + \frac{n_m}{s_{i2}} + \frac{n_l d}{s_{i1}(d - s_{i1})} = (n_l - n_m)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\,. \tag{10.3.12}$$

This general result becomes particularly simple if the lens can be assumed to be thin in the sense that its thickness d is much smaller than any of the other relevant distances. In this special case, we obtain

$$\frac{1}{s_{o1}} + \frac{1}{s_{i2}} = \left(\frac{n_l}{n_m} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\,. \tag{10.3.13}$$

In the special case where the surrounding material is air ($n_m = 1$), this equation is known as the ***lensmaker's equation***.

From this result, it is straightforward to calculate the back focal length of a thin lens as

$$\frac{1}{f_i} = \left(\frac{n_l}{n_m} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\,. \tag{10.3.14}$$

and the front focal length

$$\frac{1}{f_o} = \left(\frac{n_l}{n_m} - 1\right)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{1}{f_i} = \frac{1}{f}\,. \tag{10.3.15}$$

Hence, in the present situation, the two focal lengths are equal and we can associate a single quantity, ***focal length*** f, to the lens. It is important to realize that this is only true for a thin lens when the refractive indices on its both sides are equal. This case is thus very common but by no means general.

Of course, for a thin lens, the object distance for the whole system is $s_o = s_{o1}$ and the image distance is $s_i = s_{i2}$. Combining Eqs. (10.3.13) and (10.3.14), we then obtain imaging equation for a thin lens in the familiar form

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}\,. \tag{10.3.16}$$

Note that this is so far only valid for object and image points on axis. This equation is also known as the ***Gaussian lens formula***. The inverse of the focal length is also known as the ***refractive power*** of the lens. This quantity tells how strongly the lens tends to refract the rays.

Equation (10.3.15) implies that when the lens material has higher refractive index, a convex surface of a lens gives rise to positive focal length and refractive power and a concave to negative focal length and refractive power (Fig. 10.6). A lens with a positive focal length turns a parallel beam of rays into a beam that converges to the ***back focal point*** of the lens (Fig. 10.7a). A lens with a negative focal length, on the other hand, turns a parallel beam of rays into a diverging beam such that the continuation of the rays intercepts the axis before the lens (Fig. 10.7b). The back focal point is thus virtual and located before the lens.

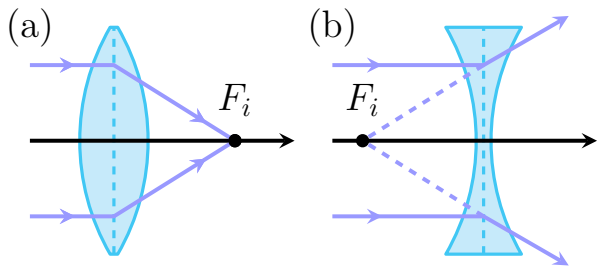

Fig. 10.6: Effect of a positive (a) and negative (b) lens of a parallel beam of rays.

When the object is at infinity, all rays arrive to the lens parallel to the axis and converge at the distance of the focal length from the lens (Fig. 10.7). This point is known as the ***back focal point*** F_i of the lens, or simply the (back) ***focus***. Similarly, a ***front focal point*** F_0 exists in the object space. For thin lenses and within the paraxial approximation, the situation remains almost unchanged for a parallel beam of rays that arrives to the lens at some angle α with respect to the axis. We can then consider the line that is parallel to the rays and passes through the center of the lens, known as the ***optical center***, as a temporary axis. The ray propagating along this axis is

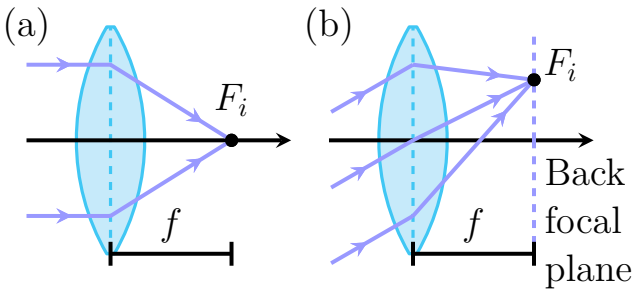

Fig. 10.7: Focusing of parallel beams of rays at the back focal plane of a lens. (a) rays arrive parallel to the axis. (b) Rays arrive at some angle with respect to the axis.

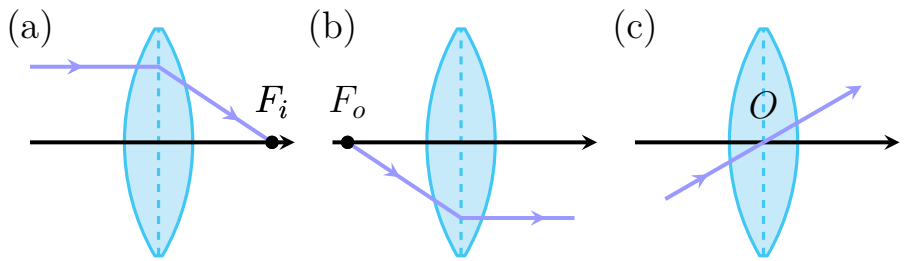

Fig. 10.8: The three special rays needed to understand the image formation of objects with off axis points.

not deviated. On the other hand, all other rays are now focused approximately at the distance of the focal length from the lens but on the temporary, instead of the true axis. As the angle α varies these points cover a plane at the distance of the focal length from the lens. This plane is the ***back focal plane*** of the lens. Similarly, the plane at the focal distance before the lens is the ***front focal plane*** of the lens. If we have an extended source of light at the front focal plane, each point on the plane leads to a parallel beam of rays in the image space propagating at an angle defined by the location of the point on the plane.

10.4 Image formation

It is important to note that so far we have considered in detail only the case where both the object and image points are located on the same axis. This is clearly not sufficient to understand how images of extended objects with off-axis object points are formed. In order to proceed with regard to these more general cases, it is sufficient to realize that the refraction of rays by the lens is independent of the origin of the rays. We can therefore focus our attention to three special rays (Fig. 10.8): a) The ray that arrives to the lens parallel to the axis must pass through the back focal point F_i; b) The ray that departs from the lens parallel to the axis must have passed through the front focal point F_o; c) The ray that passes through the optical center O is not deviated by the lens.

We next consider the situation shown in Fig. 10.9. The object point is at the distance s_o from the lens but at the height y_o from the axis. Alternatively, we can also measure the object location from the front focal plane by the distance x_o. Similarly, the image is formed at the distance s_i from the lens but at the height y_i from the axis, and could also be measured from the back focal plane by distance x_i. The sign conventions for the figure are shown in Table 10.2. In Fig. 10.9, all quantities except for y_i are thus drawn as positive.

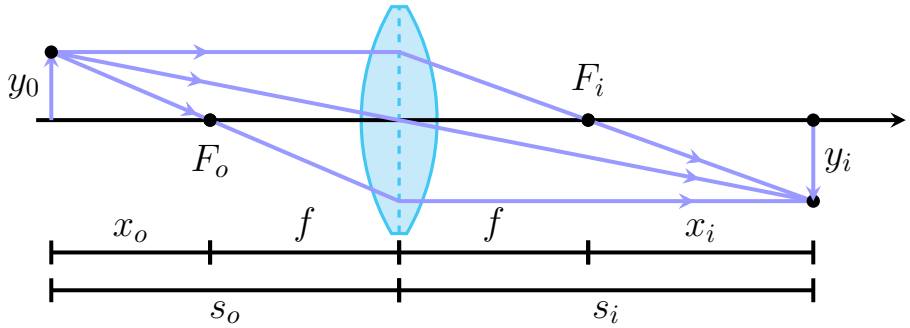

Fig. 10.9: Notation for imaging of objects with finite transverse size.

Table 10.2: Sign conventions for imaging by a thin lens.

Quantity	Sign	Location from surface	Type of object or image
s_o , f_o	positive/negative	left/right from lens	real/virtual
x_o	positive/negative	left/right from F_o	
s_i , f_i	positive/negative	right/left from lens	real/virtual
x_i	positive/negative	right/left from F_i	
R	positive/negative	above/below axis	

Comparing various triangles in Fig. 10.9, we directly see the relations

$$\frac{y_o}{|y_i|} = \frac{f}{s_i - f} = \frac{s_o}{s_i} = \frac{x_o}{f} = \frac{f}{x_i} \,. \tag{10.4.17}$$

from which we obtain

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f} \,. \tag{10.4.18}$$

The imaging equation originally derived for on-axis points is thus seen to be valid also for off-axis points.

On the other hand, Eq. (10.4.17) also yields a very simple result in terms of the distances measured form the focal points

$$x_o x_i = f^2 \,. \tag{10.4.19}$$

This is known as the ***Newtonian lens formula***. Since f^2 is always positive, the two distances x_o and x_i have the same sign. The object and image points are therefore always on different sides from their reference focal points.

The goal of imaging systems is usually to somehow manipulate the object to make it easier to view. This is obtained by affecting the size of the image compared to that of the object, as defined by the ***magnification***. More specifically, the ***transverse magnification*** of a lens is defined as the ratio between the transverse sizes of the

image and the object, i.e.,

$$M_T = \frac{y_i}{y_o} = -\frac{s_i}{s_o} = -\frac{x_i}{f} = -\frac{f}{x_o}\,. \tag{10.4.20}$$

where the minus sign follows from the fact that the object and image heights in Fig. 10.9 have different signs. This quantity therefore takes into account whether the image is on the other side of the axis than the object.

We can also define ***longitudinal magnification*** as a differential change in the position of the image for a similar change in the position of the object (Fig. 10.10)

$$M_L = \frac{dx_i}{dx_o}\,. \tag{10.4.21}$$

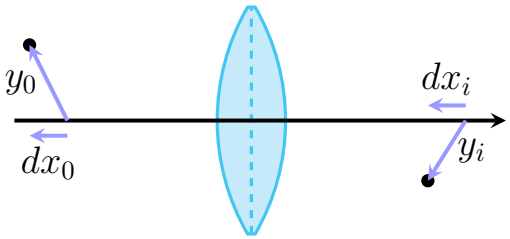

Fig. 10.10: Quantities for transverse and longitudinal magnification.

From Eq. (10.4.20), we first obtain $x_i = f^2/x_o$, which yields for the longitudinal magnification

$$M_L = -\frac{f^2}{x_o^2} = -M_T^2\,. \tag{10.4.22}$$

This result implies that, for any other magnification than unity, the image will be distorted with regard to its transverse and longitudinal dimensions.

We finally summarize additional sign conventions regarding the imaging properties of a thin lens in Table 10.3.

Table 10.3: Additional sign conventions for imaging by a thin lens.

Quantity	Sign	Location from surface	Type of object or image
s_o	positive/negative	left/right from lens	real/virtual object
s_i	positive/negative	right/left from lens	real/virtual image
f	positive/negative	right/left from lens	converging/diverging lens
y_o	positive/negative	above/below axis	
y_i	positive/negative	above/below axis	
M_T	positive/negative		object and image on the same/different side from axis

10.5 Combinations of lenses

Optical system usually consists of several lenses. One reason for this is that different lenses have different functions from the viewpoint of the operation of the whole system. Another reason is that, by using different lenses made of different materials, one

can reduce the amount of aberrations that the system possesses. This is possible even for non-paraxial rays, which is important for collecting more light while maintaining a good image quality.

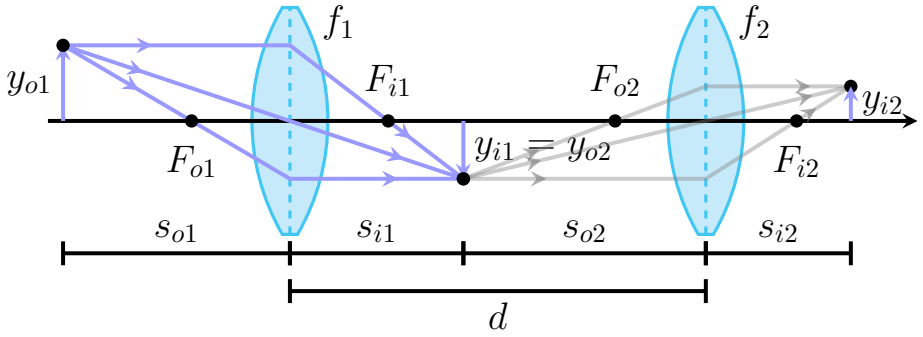

Fig. 10.11: Two thin lenses separated by a distance d.

In order to treat a system of several lenses, there is only one universal way to proceed: one must consider each lens separately. As an example, we may consider a system shown in Fig. 10.11, where lenses with focal lengths f_1 and f_2 are separated by distance d. The image formed by the first lens thus acts as the object for the second lens. The object distance for the second lens is thus $s_{o2} = d - s_{i1}$. Furthermore, the object height of the second lens equals the image height of the first lens $y_{o2} = y_{i1}$. For the magnification, we thus obtain

$$M_T = \frac{y_{i2}}{y_{o1}} = \frac{y_{i1}}{y_{o1}}\frac{y_{i2}}{y_{i1}} = \frac{y_{i1}}{y_{o1}}\frac{y_{i2}}{y_{02}} = M_{T1}M_{T2}\,. \tag{10.5.23}$$

The only exception where this general approach can be simplified is the case where the two lenses are in contact. The imaging equations for the two lenses are

$$\frac{1}{s_{o1}} + \frac{1}{s_{i1}} = \frac{1}{f_1}\,, \quad \frac{1}{s_{o2}} + \frac{1}{s_{i2}} = \frac{1}{f_2}\,. \tag{10.5.24}$$

By adding these equations and noting that now $s_{o2} = -s_{i1}$, we obtain

$$\frac{1}{s_{o1}} + \frac{1}{s_{i2}} = \frac{1}{f_1} + \frac{1}{f_2}\,. \tag{10.5.25}$$

In addition, the object and image distances for the whole system are $s_o = -s_{o1}$ and $s_i = -s_{i2}$. We can therefore conclude that the system act as a single lens with the focal length

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2}\,. \tag{10.5.26}$$

Although this result as such is interesting, it may appear useless in practice, because it should not make much difference whether a given focal length is achieved by using

one or more thin lenses. However, the indices of refraction of materials depend on wavelength, which gives rise to ***chromatic aberration***, i.e., dependence of focal length on wavelength. It turns out that, by using lenses made from different materials, one can reduce chromatic aberration.

We are therefore back to the rule that systems of several lenses should be analyzed by treating each lens separately. In practice, this is not done by using the imaging equation, Eq. (10.3.16), for each lens separately. Instead, the analysis can be programmed to follow the rays through the system in such a way that refraction at each surface is taken into account separately. Such ***ray tracing*** can easily be generalized also for thick lenses, allowing much more general systems to be analyzed.

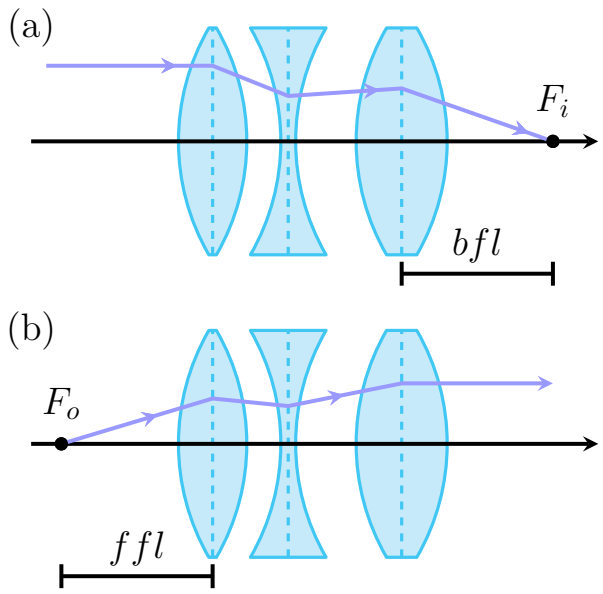

Fig. 10.12: Back (a) and front (b) focal lengths of a combination of lenses.

When the system consists of several lenses separated by some distance, the system is not thin any more. It is then not evident from which location the focal length should be measured. This leads to the result that we need to define separately the ***back focal length*** (b.f.l) and the ***front focal length*** (f.f.l) of the system as distances from the last surface of the system to the focal point in the image space and from the first surface to the focal point in the object space, respectively (Fig. 10.12). The focal points, of course, are found in the usual way by tracing through the system rays that arrive or depart parallel to the axis.

10.6 Apertures and stops

We next introduce a few more advanced concepts that play an important role in the design of optical systems. These concepts are related to the light collection capability of the system and the size of the object that can be captured by the system (Fig. 10.13).

The ***aperture stop*** (AS) of the system is the element that limits the amount of light that can be collected by the system. In real systems, all lenses have a finite diameter. In principle, the rim of any lens could then be the aperture stop and, in some cases, one needs to consider all possibilities as described below. In a well-designed case, however, the system contains a separate diaphragm whose location has been carefully considered. For example, in cameras the size of the diaphragm is adjustable, and in high-level cameras the user can choose the size according to his/her needs. In the example of Fig. 10.13, rays 1 from an on-axis object point barely make it through the aperture stop, but rays 2 are blocked.

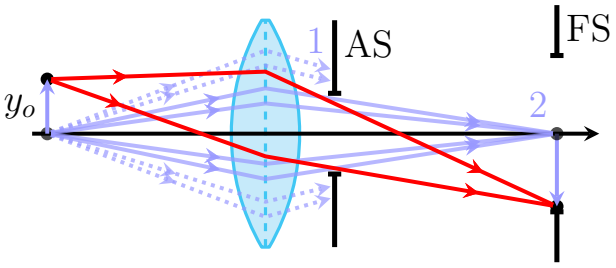

Fig. 10.13: Aperture (AS) and field (FS) stops of an optical system.

The ***field stop*** (FS) limits the size of the object that can be captured at the image plane. Often, the limitation comes from the size of the image detector. In film cameras, the standard size of the frame used to be 24 mm × 36 mm. In contemporary digital cameras, on the other hand, a number of different sizes of the CCD or CMOS detectors are in use. In other applications, the field stop may be located inside the system. In Fig. 10.13, where the dashed rays correspond to the extreme off-axis object and image points, the field stop is at the image plane, and an object any larger than shown would not fit on the available area.

In practice, the definition of the aperture stop given above is not sufficient to determine which element acts as the aperture stop. In order to understand this better, we need to introduce two additional concepts. The ***entrance pupil*** of the system is the image of the aperture stop in the object space. In consequence, all rays that pass through the entrance pupil must pass also through the aperture stop. The location and size of the entrance pupil are thus determined by imaging the aperture stop backward (to the left) through all the lenses that precede the aperture stop. In cases where the aperture stop is not known a priori, all possibilities need to be considered. For example, in Fig. 10.14, we have a system

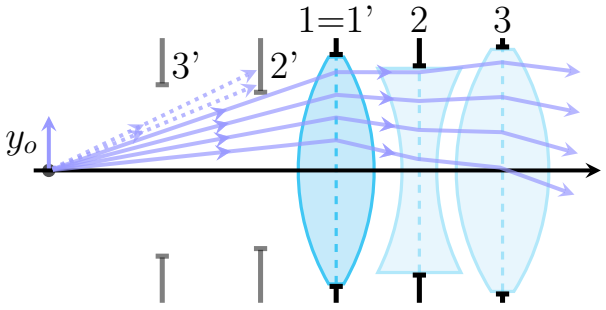

Fig. 10.14: Determination of the entrance pupil of an optical system.

of three lenses with different diameters, as indicated by the numbers 1,2, and 3. By imaging all possibilities to the object space, we obtain potential entrance pupils, as indicted by the numbers 1', 2', and 3'. The actual entrance pupil is then the one that subtends the smallest solid angle as seen from an axial object point. In the present example this is 2', which then determines lens 2 as the actual aperture stop. Similarly, the ***exit pupil*** of the system is the image of the aperture stop in the image space as imaged through all the lenses after (to the right) the aperture stop.

Here, we have discussed the aperture stop and the pupils from the viewpoint of the light collection capability of the system. However, in general, they play an important role in other properties of the system as well. When designing an optical system, one must therefore pay significant attention to the stops (especially the aperture stop) and the pupils. For example, in visual optical instruments, e.g., telescope or binoculars, the pupil of the eye should be placed at the exit pupil. As all light exits the system through the exit pupil, it is easy to see a bright spot before the eyepiece (ocular) of the telescope (Fig. 10.15). Furthermore, depending on the application, the exit pupil needs to be designed at different distances from the last lens of the system. This distance is known as the ***eye relief*** of the system.

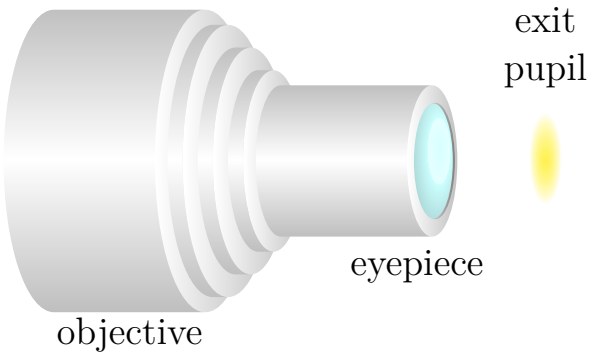

Fig. 10.15: The exit pupil can be seen as a "floating" bright spot before the eyepiece of a telescope.

Another important issue is that in some cases the size of the beam of rays collected by the system becomes smaller for off-axis object points. This is known as ***vignetting***. In consequence, the images of the off-axis object points are not as bright as those of the on-axis points. One may think that vignetting is always an undesirable property. However, in some cases, the instrument is more pleasant to use if the change from brightness in the central part to the darkness outside the field is gradual rather than abrupt. Some vignetting is

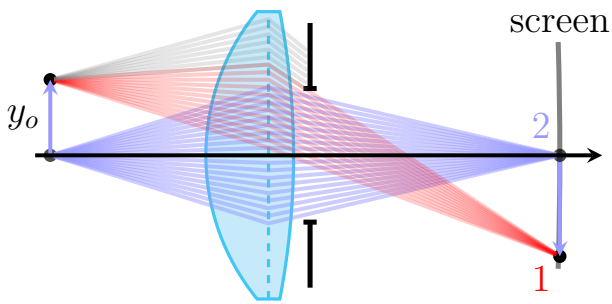

Fig. 10.16: Vignetting. The amount of light collected from off-axis object points (dashed rays) is less than that from on-axis points (solid rays).

thus often designed into the system on purpose.

We finally come to the concepts of the ***relative aperture*** and ***f-number***, which are also related to the light collection capability of the system. It is clear that the amount of light collected by the system is proportional to the area of the entrance pupil, which scales as $\propto D^2$, where D is the diameter of the entrance pupil (Fig. 10.17). At the image plane, the collected light is distributed over an area that depends on the image size. From Eq. (10.4.17), we obtain

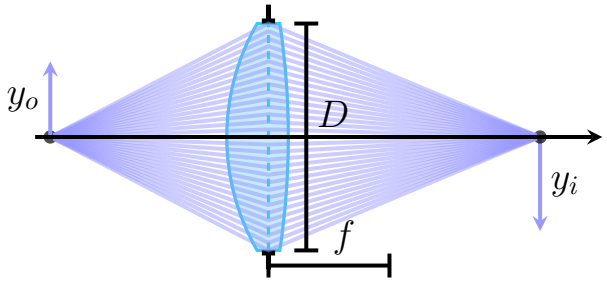

Fig. 10.17: Quantities for defining the relative aperture and f-number.

$$y_i = -\frac{y_o}{x_o} f \,. \tag{10.6.27}$$

The size of the image is therefore proportional to the focal length of the system and the collected light is distributed over an area that scales as $\propto f^2$. The brightness of the image is therefore proportional to $\propto (D/f)^2$, where D/f is the relative aperture.

The inverse of the relative aperture D/f is known as the f-number, which is often inscribed to the instrument using notations, such as $f^{\#}$ or f/#. In cameras, where the size of the diaphragm is adjustable, the f-number is indicated by a sequence of numbers 1, 1.4, 2, 2.8, 4, etc. The sequence thus goes in factors of about $\sqrt{2}$, implying that each step reduces the amount of light collected by the system by a factor of two.

10.7 Mirrors

Not all optical systems can be conveniently implemented by using only lenses. In some cases, it is necessary to use also mirrors. We will next discuss some of the key issues that need to be considered for mirrors.

We will first consider a planar mirror (Fig. 10.18). We take the surface normal of the mirror to be our axis and place our object at distance s_o from the mirror. We then follow a ray from the object point S that propagates at some angle with respect to the axis. This ray is reflected by the mirror in such a way that the angle of reflection is the same. The continuation of this ray thus intercepts the axis at point P at the image distance $s_i = s_o$ behind the mirror. The image formed is thus virtual. We therefore choose that both the object and image distances are negative on the right

from the mirror.

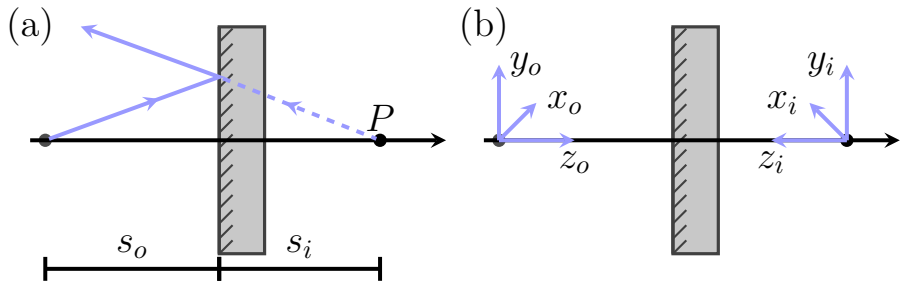

Fig. 10.18: Reflection from a planar mirror. (a) Image formation. (b) Change in handedness between the object and the image.

For the case of a planar mirror, it is evident that any surface normal can be chosen as an axis. We will next consider how a three-dimensional object, represented by the right-handed coordinate system (x_o, y_o, z_o) is imaged by the mirror. It is then easy to see that the resulting image coordinate system (x_i, y_i, z_i) is left-handed. The mirrors therefore result in images where the handedness of the object is reversed.

We next consider a ***spherical mirror*** with a concave reflecting surface and radius of curvature R (Fig. 10.19). We again choose the axis as the local normal to the surface of the mirror. We place the object point O at distance s_o from the mirror and follow the ray that propagates at some angle with respect to the axis. The reflection of this ray by the mirror is determined by the local surface normal at the point where the ray meets the surface of the mirror. The reflected ray thus intercepts the axis at the image point I.

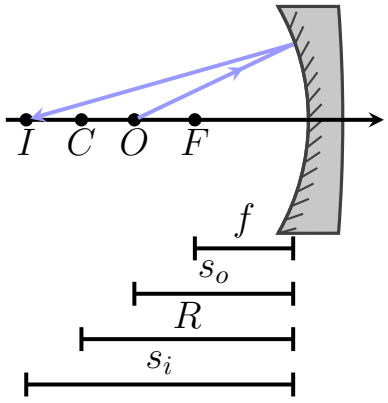

Fig. 10.19: Imaging by spherical mirror.

Without going into details, it can be shown that for paraxial rays the imaging equation for the mirror is

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}\,, \quad f = -R/2\,, \tag{10.7.28}$$

i.e., the form is exactly the same as that for a thin lens. It is important to note, however, that this form is only valid with very well defined sign conventions, as summarized in Table 10.4. These conventions show that a concave mirror behaves as a lens with positive focal length. In addition, the definition of the transverse

Table 10.4: Sign conventions for imaging by a spherical mirror.

Quantity	Sign	Location	Type of object or image
s_o	positive/negative	left/right from mirror	real/virtual object
s_i	positive/negative	left/right from mirror	real/virtual image
f	positive/negative	left/right from mirror	concave/convex
R	positive/negative	right/left from mirror	convex/concave
y_o	positive/negative	above/below axis	
y_i	positive/negative	above/below axis	

magnification is obviously also for mirrors

$$M_T = \frac{y_i}{y_o}\,. \tag{10.7.29}$$

The main advantage of mirrors is that their operation is based on the reflection law, which is independent of the refractive indices of the media. They are therefore not sensitive to the dispersion of the refractive indices with wavelength and behave in the same way for all wavelengths. Lenses, on the other hand, behave slightly differently for different wavelengths. The dependence of the imaging properties on wavelength is known as ***chromatic aberration***.

On the other hand, systems with mirrors are a bit cumbersome to analyze because each reflection reverses the direction of propagation. The sign conventions for the various quantities are therefore more ambiguous than for systems consisting of only lenses, where light propagates from left to right all the time. Systems with mirrors are therefore often analyzed by an effective system where each mirror is replaced by a thin lens with equal focal length.

10.8 Prisms

Prisms are other components often used in optical systems. They exist in a large variety of forms to achieve the desired function. However, their operation is based on either refraction (e.g., Fig. 10.20a) or total internal reflection (e.g., Fig. 10.20b).

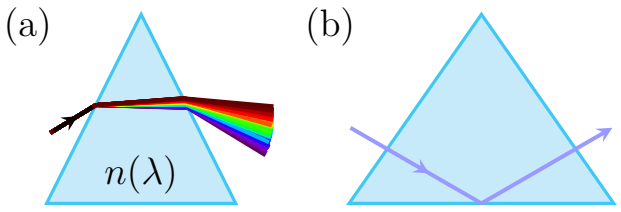

Fig. 10.20: Use of prisms is based on refraction (a) or on total internal reflection (b).

10.9 The human eye

For most people, vision is probably one of the most important senses. The visual system consists of the eye, its nerve connections to the brain, and the complicated processes within the brain that give us the visual perception. In the present discussion,

we will only consider the eye from the viewpoint of its optical function.

A cross-section of the eye from the side is shown in Fig. 10.21. The outermost surface of the eye is the ***cornea***. The opening into the eye is the ***pupil***, which is surrounded by the ***iris***, which gives the color to the eyes. This is followed by the ***lens*** of the eye. The inside of the eye is filled with ***vitreous humor***, a liquid- or gel-like material. Finally, the inner back side is known as the ***retina***.

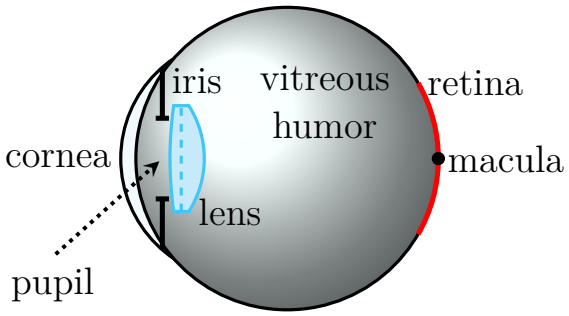

Fig. 10.21: Schematic of the human eye.

It is important to note that from the viewpoint of image formation, the strongest refraction occurs at the interface between air and cornea. The purpose of the lens is thus to provide some fine focusing capability in order to maintain a sharp image on the retina for varying object distances. The size of the iris varies with the amount of light, thus acting as a diaphragm and maintaining proper level of illumination on the retina. The retina acts as the light detector, where the sensitivity is provided by the rod and cone cells. The central area of the retina is known as the ***macula***, which is the area for sharpest vision.

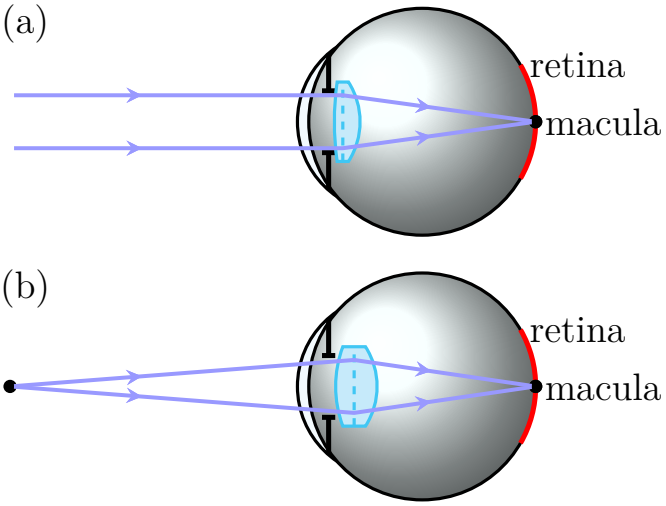

Fig. 10.22: (a) For far-away objects, the lens of the eye is relaxed. (b) For nearby objects, the lens accommodates providing more refractive power.

When the eye is looking at objects at different distances, its refractive power is adjusted by changes in the shape of the lens to maintain a sharp image on the retina. This is known as ***accommodation***. When the object is far away, the eye is relaxed, i.e., the lens has the flattest shape and lowest refractive power (Fig. 10.22a). When the object is brought closer, the eye accommodates, so that the lens becomes thicker and it provides more refractive power to maintain the focus on the retina (Fig. 10.22b).

Problems with vision are very common. One situation is that the eye is not able

to focus on objects that are brought arbitrarily close. The shortest distance where the eye can focus is known as the ***near point***. For young babies, this distance can be as short as 7 cm. This distance becomes longer with age and can be even 1 m for old people. This is clearly too much for convenient reading. The implication is that most people need to start wearing reading glasses at the age of about 40 – 50 years. The need for this occurs when the near point is further than 25 – 40 cm, where the variation arises from different standards in different countries. Another common problem is that the relaxed eye is not able to focus at objects that are arbitrarily far away. This is known as ***nearsightedness***, which is a common problem for young people. The distance where the relaxed eye focuses is known as the ***far point***.

The vision problems are traditionally corrected by eye glasses or contact lenses. In the recent years it has also become possible to reshape the surface of the cornea in an operation. Today, these operations are done using lasers with ultrashort, femtosecond pulses, which reduce the damage to the tissue around the area being operated.

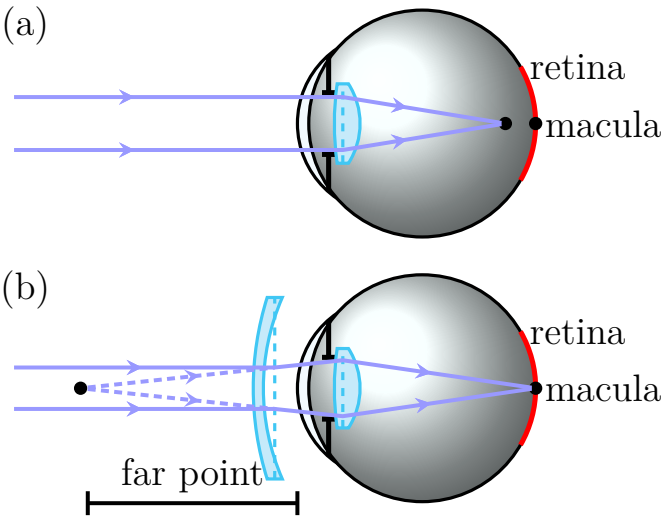

Fig. 10.23: Nearsightedness. (a) The relaxed eye has too much refractive power for far-away objects. (b) Correction by a negative lens, which brings the far-away objects to the far point.

The medical term for nearsightedness is ***myopia***. It arises from the fact that the relaxed eye still has too much refractive power for far-away objects, whose image is formed before the retina (Fig. 10.23a). This condition is therefore corrected by negative lenses that reduce the total refractive power of the lens–eye system (Fig. 10.23b). The lens thus provides a virtual image of the far-away object at the far point of the person. When wearing eye glasses, the nearby objects are seen by accommodating the eye.

The medical term for farsightedness is ***hyperopia***. It arises from the condition where the fully accommodated eye has too little refractive power and the image is formed behind the retina (Fig. 10.24a). This condition is corrected by positive lenses that add more refractive power (Fig. 10.24b). The power of the lens is chosen in such a

way that it provides a virtual image of an object at 25 cm, i.e., the near point of the person. When the object is moved further away, the accommodation of the eye is reduced.

Of course, near- and farsightednesses are not the full list of vision problems. For example, it is common that the amount of correction needed is different for two orthogonal directions. The lens then has ***cylindrical error*** or ***astigmatism***. Another common situation is that when people with nearsightedness become older, they also need reading glasses. When both corrections are combined in one lens, one wears ***bifocal*** or ***multifocal*** eye glasses. On the more medical side, most old people develop ***cataract*** where the lens of the eye starts becoming opaque. This can be treated by replacing the lens of the person by an artificial lens, which can also be optimized for the desired vision.

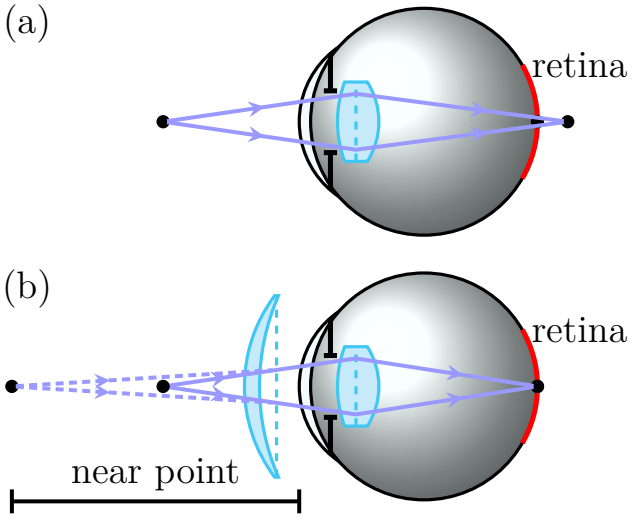

Fig. 10.24: Farsightedness. (a) The accommodated eye has too little refractive power for nearby objects. (b) Correction by a positive lens, which moves the nearby objects to the near point.

10.10 Magnifying glass

In order to understand what limits the size of the small objects that can be seen by the eye, we consider Fig. 10.25. The size of the object on the retina depends on the ***viewing angle*** α_u of the object. This angle can be made larger by bringing the object closer and closer to the eye. However, the closest distance is limited by the near point of the eye d_o, setting the limit to the viewing angle at $\alpha_u \approx -y_o/d_o$, where y_o is the object size.

A ***magnifying glass*** is a simple positive lens that increases the refractive power of the lens–eye system and allows the object to be brought closer than the original near point. In consequence, the viewing angle is increased and the object is perceived larger than originally.

In the ideal case (Fig. 10.26), the magnifying glass forms a virtual image of the object at infinity so that the person can use relaxed eye for viewing. By using the imaging equation, the celebrated Eq. (10.3.16), this situation is achieved when the object

is placed at the front focal plane of the magnifying glass. The viewing angle then becomes $\alpha_a \approx -y_o/f$, where f is the focal length of the magnifying glass.

The ***magnifying power*** of a magnifying glass is defined to be the ratio between the aided and unaided viewing angles, i.e.,

$$MP = \frac{\alpha_a}{\alpha_u} = \frac{d_o}{f} = d_o\mathcal{D}\,. \quad (10.10.30)$$

A lens with 10 cm focal length, i.e., 10 D refractive power, therefore provides a magnifying power of 2.5×. In practice, the magnifying power that can be achieved by a simple magnifying glass is limited to 2 – 3× by the aberrations of the lens.

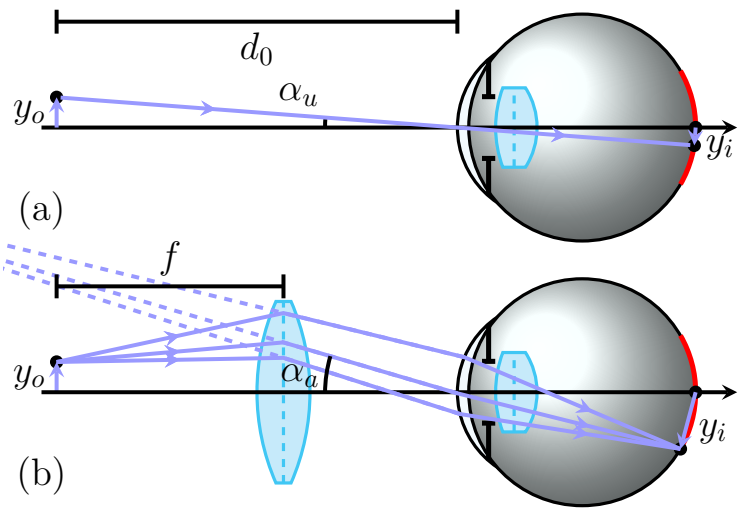

Fig. 10.25: (a) The size of the image of an object on the retina is limited by the near point distance d_0. (b) A magnifying glass increases the viewing angle of an object and forms its virtual image at infinity.

10.11 Eyepiece (ocular)

A typical optical instrument consists of two lenses. The ***objective*** is placed close to the object and provides an intermediate image of the object. This image is viewed by the ***eyepiece***, also known as the ***ocular***, which provides further magnifying power. More generally, the eyepiece is used to look at the image formed by any preceding lens system.

In practice, both the objective and the eyepiece consist of several lenses to provide high image quality. Nevertheless, they can be functionally reduced to a simple thin lens. From this viewpoint, an eyepiece is just a very advanced version of a magnifying glass.

10.12 Microscope

The first real optical instrument we consider is the ***microscope***. The main task of the microscope is to form greatly magnified images of small objects, which are very close to the objective of the microscope.

The basic layout of a conventional microscope is shown in Fig. 10.27. The objective forms a magnified real image at an intermediate plane between the objective and the eyepiece. The eyepiece is then used as a magnifying glass to increase the viewing

angle of the intermediate image for the eye.

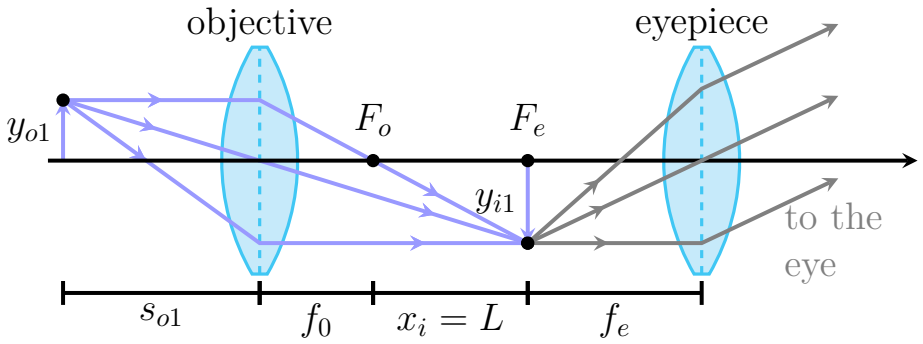

Fig. 10.26: Conventional design for a microscope.

Microscopes can be adapted to several different types of objects, each requiring a very specific objective and/or eyepiece. It is therefore important that certain parts of the microscope are standardized. In particular, any objective must form the intermediate image at the same plane. This is obtained by standardizing the distance from the back focal plane of the objective to the intermediate image plane to a fixed value. In the general theory of imaging (see Fig. 10.9 in Section 10.4), this distance is x_i. In the case of microscopes, this distance is known as the ***tube length*** of the microscope, usually labeled by L. Several manufacturers have standardized this distance to ∼160 mm.[2] The transverse magnification obtained is then simply from Eq. (10.4.20)

$$M_T = -x_i/f_o = -L/f_o\,. \tag{10.12.31}$$

where f_o is the focal length of the objective. We note that this equation works also for infinity-corrected objectives (where $L = f_t$). A focal length of, e.g., 32 mm then gives a magnification of 5×, which is usually marked on the barrel of the objective. In addition, the eyepiece typically gives a factor of, e.g., 10× to the magnifying power, leading to a total magnification of 50× for the whole microscope for the present example.

Another useful parameter of objective lenses is the ***field number*** (FN), which is defined as the diameter of the area of the intermediate image plane that is still observed through the eyepiece. By knowing the field number and the magnification of the objective, the ***field-of-view*** (FOV) of the objective can be calculated from

$$\text{FOV} = \text{FN}/M_T\,. \tag{10.12.32}$$

[2]Thorlabs, Nikon, and Leica use 200 mm, Olympus uses 180 mm and Zeiss 165 mm standards.

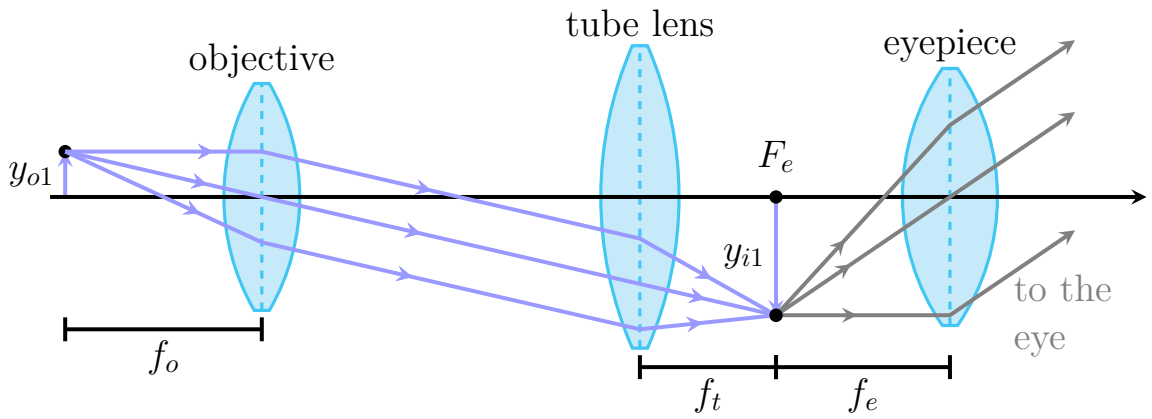

Fig. 10.27: Modern design microscope. Tube lens allows to use objectives with infinite image distance.

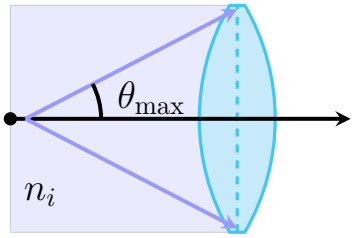

Fig. 10.28: Numerical aperture of the objective characterizes its light-collection capability.

It is important to note that the conventional microscope objectives must be designed for very specific cases where the object and image distances are essentially fixed. They therefore have very limited use in other applications. With the development of lasers, a great need has arisen for situations where one of the distances is infinite, corresponding to a collimated laser beam. In order to make it possible to use such objectives also in microscopic imaging applications, more modern microscopes often rely on the design shown in Fig. 10.27. Here the objective forms the intermediate image at infinity. In this case, the object is imaged on the desired intermediate image plane by a ***tube lens***, which is fixed on the frame of the microscope. Such design gives much more flexibility in the use of microscope objectives in varying applications.

For microscopic imaging, the object can be very close to the front surface of the objective, in some cases even touching it. Microscope objectives therefore operate in a highly non-paraxial regime, which makes their design extremely demanding. In fact, different objectives are needed for each different type of object. The most common cases are the following: 1) the object is held in place with a cover slip, a thin (e.g., 0.17 mm thick) piece of glass with a standardized refractive index (e.g., $n = 1.515$), 2) the object is in air, 3) the object is immersed in a liquid, e.g., oil. For high quality imaging, the objective must be optimized for each of these cases separately.

Another issue is that, for objects very close to the objective, the relative aperture and f-number are not good measures of the light-collection capability. Instead this is

characterized by the ***numerical aperture*** of the objective, defined as (Fig. 10.29)

$$\mathrm{NA} = n_i \sin\theta_{\max}\,, \tag{10.12.33}$$

where n_i is the refractive index of the medium in which the object is immersed and $\theta_{\max}$ is the angle, with respect to the axis, of the rays that are still collected by the objective. The numerical aperture also determines the diffraction-limited resolution of the microscope as will be discussed later on.

10.13 Telescope

Another important basic instrument is the ***telescope***. The goal of telescopes is to provide a magnified view of far-away objects as seen by the eye. The telescopes also consist of an objective and an eyepiece. As the object is very far from the telescope, the intermediate image formed by the objective is necessarily demagnified. Furthermore, the intermediate image is formed at the back focal plane of the objective, which is also the front focal plane of the eyepiece.

We will first consider the propagation of a beam of rays parallel to the axis through the telescope (Fig. 10.30). It is then evident that this beam also escapes the system parallel to the axis. In consequence, both the back and front focal lengths of the system are infinite. In other words, the system has no well-defined focal length although it clearly manipulates the size of the beam of rays. Such a system is said to be ***afocal***.

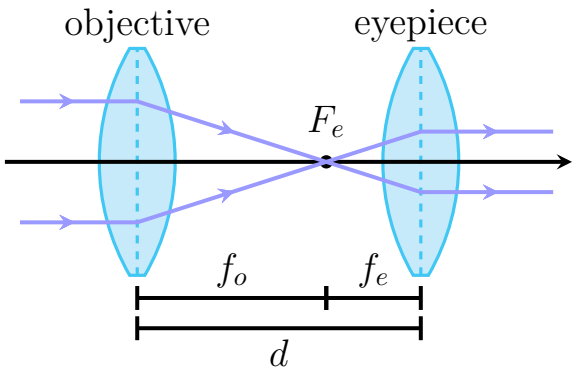

Fig. 10.29: Propagation of a beam of rays parallel to the axis through a telescope.

In order to understand how the magnified view is obtained, we consider rays that arrive from an off-axis object point (Fig. 10.31). As the object itself is far away, it is best to describe the ray by the angle α_u it makes with respect to the axis. Note that this is also the angle at which the ray would enter the eye without the telescope. This ray forms the intermediate image at the back focal plane at a distance y from the axis determined by the arrival angle $\alpha_u = y/f_o$, where f_o is the focal length of the objective.

The formed image is now the object for the eyepiece at its front focal plane. In consequence, the angle at which the rays escape the whole system and enter the eye is $\alpha_a = y/f_e$, where f_e is the focal length of the eyepiece. The magnification provided is thus angular and described by the magnifying power

$$MP = \alpha_a/\alpha_u = f_o/f_e\,. \qquad (10.13.34)$$

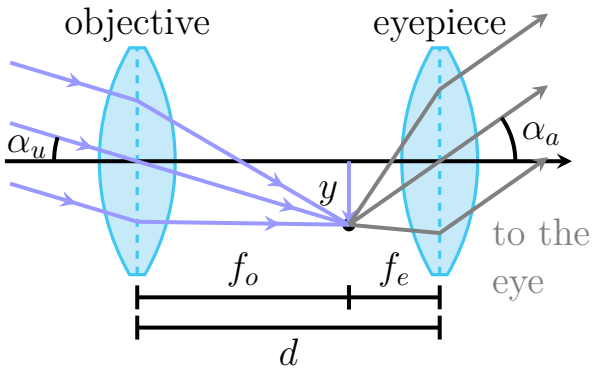

Fig. 10.30: Angular magnification of a telescope.

The magnification is therefore determined by the ratio of the focal lengths of the two lenses. In addition, the lenses must be separated by the distance $d = f_o + f_e$.

There are two basic configurations of telescopes (Fig. 10.32). Historically, the first one is the ***Galilean telescope***. In this case, the eyepiece is a negative lens and the magnifying power is positive. This implies that the objects are seen in their original orientation. Furthermore, the length of a Galilean telescope is shorter than the focal length of the objective. One may therefore think that the Galilean design has only positive attributes. Unfortunately, however, the exit pupil of a Galilean telescope is inside the system. It is therefore impossible to bring the eye to the exit pupil, which makes the use of the telescope quite difficult.

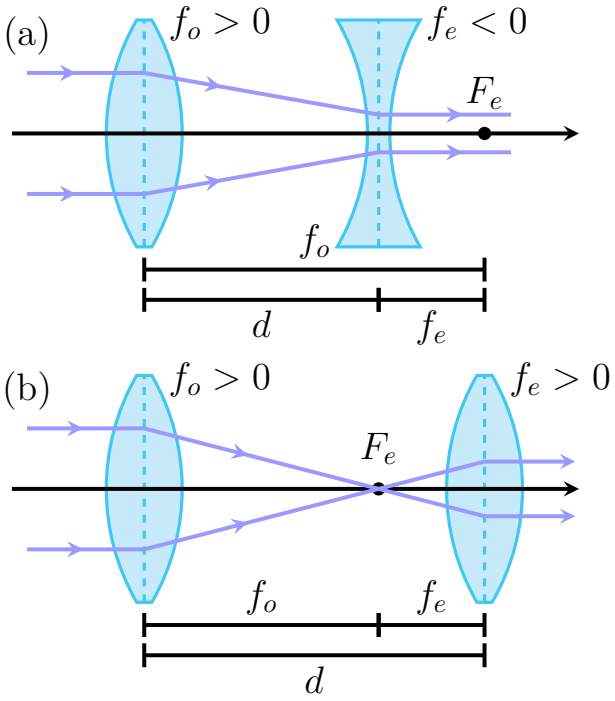

Fig. 10.31: Galilean (a) and Keplerian (b) telescopes.

The other basic design is the ***Keplerian telescope***. In this case, the eyepiece is a positive lens and the magnifying power is negative. The design therefore turns the objects upside down. This, of course, is awkward for viewing objects on earth. On the other hand, for the original purpose of viewing astronomical objects it does not really matter whether they are seen upright or upside down. ***Binoculars*** are based on the Keplerian design. However, they have a prism system between the objective

and the ocular, which turns the image in the proper upright orientation. Finally, the exit pupil of Kepler's telescope is outside of the system, which makes it much more convenient to use.

The main problem of both the Galilean and Keplerian designs is that they are susceptible to chromatic aberration, i.e., different colors follow different paths through the system. This is because the refractive indices of the lenses depend on wavelength. A major improvement with respect to this was obtained from the ***Newtonian telescope***, where the objective was replaced by a mirror (Fig. 10.32). In this design, all wavelengths are reflected in the same way.

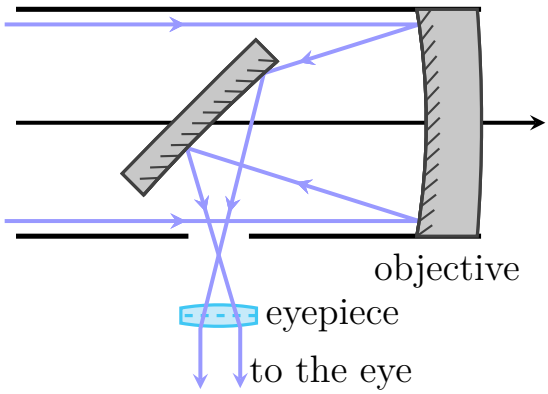

Fig. 10.32: Newtonian telescope has a reflecting objective.

INDEX