1

Introduction to Optics

Rainer Heintzmann[1,3] *and Ulrich Kubitscheck*[2]

[1] *Friedrich Schiller-Universität, Institut für Physikalische Chemie und Abbe Center of Photonics, Helmholtzweg 4, 07743 Jena, Germany*
[2] *Rheinische Friedrich-Wilhelms-Universität Bonn, Institut für Physikalische & Theoretische Chemie, Wegelerstr. 12, 53115 Bonn, Germany*
[3] *Leibniz Institute of Photonic Technology, Albert-Einstein Str. 9, 07745 Jena, Germany*

In this chapter, we introduce the wave nature of light by discussing interference, which is then used to explain the laws of refraction, reflection, and diffraction. We then discuss light propagation in the form of rays, which leads to the laws of lenses and ray diagrams of= optical systems. Finally, the working principles of the most common optical elements are outlined.

1.1 A Short History of Theories about Light

For a long time, scientists like Pierre Gassendi (1592–1655) and Sir Isaac Newton (1643–1727) believed that light consisted of particles named *corpuscles* traveling along straight lines, the so-called *light rays.* This concept explains brightness and darkness, and effects such as shadows or even the fuzzy boundary of shadows due to the extent of the sun in the sky. However, in the sixteenth century, it was discovered that light can sometimes “bend” around sharp edges by a phenomenon called *diffraction*. This phenomenon was not compatible with the predictions of the ray theory, but rather, light must be described as a *wave.* In the beginning, the wave theory of light – based on Christiaan Huygens’ (1629–1695) work and expanded later by Augustin Jean Fresnel (1788–1827) – was not accepted. Siméon Denis Poisson, one of the judges for evaluating Fresnel’s work in a science competition, tried to ridicule it by showing that Fresnel’s theory would predict a bright spot in the center of the dark shadow of a round, illuminated obstacle. Poisson considered this to be nonsense. Another judge, Arago, however, then demonstrated that this spot does indeed exist and can be observed when measurements are done very carefully. This was a phenomenal success of the wave description of light, and the spot was named *Poisson’s spot*.

Only at the beginning of the twentieth century quantum mechanics fused both theories and suggested that light has a dual character: a wave nature and a particle nature. Since then, the description of light has maintained this dual view.

Fluorescence Microscopy: From Principles to Biological Applications, Second Edition.
Edited by Ulrich Kubitscheck.

When it is interacting with matter, one often has to consider the quantum or particle nature of light. However, the propagation of these particles is described by equations written down by James Clerk Maxwell (1831–1879). The famous Maxwell wave equations of electrodynamics identify oscillating electromagnetic fields as the waves responsible for what we call *light* and still serve today as an extremely precise description of most aspects of light.

The wave concept is required to understand the behavior of light in the context of microscopy. Light propagation in the form of rays can explain refraction, reflection, and even aberrations, but fails to explain diffraction and interference. Therefore, we start out by introducing interference, an effect that is observed only when experiments are designed very carefully.

1.2 Properties of Light Waves

1.2.1 An Experiment on Interference

Suppose that we perform the following experiment: We construct the instrument shown in Figure 1.1 consisting of only two ordinary mirrors and two 50/50 beam splitters. These beam splitters reflect 50% of the incoming light and transmit 50%. Now we use a laser of any color to illuminate beam splitter BS1. It is absolutely crucial that the distances along the two different light paths between the two beam splitters are exactly equal with a precision better than 1/10 000 of a millimeter. Then we observe that something surprising is happening: the light entering the device will leave only through exit 1. Exit 2 will be completely dark. This is very strange, as one would expect 50% of the light exiting on either side. Even more surprising is what happens if one blocks one of the two light paths inside the instrument. Now, 25% of the light will emerge from both exits.

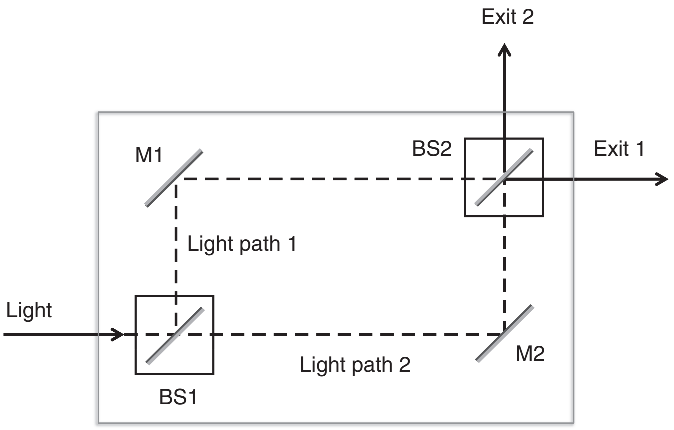


Figure 1.1 Interference experiment. In a simple arrangement of mirrors M1 and M2 and 50/50 beam splitters BS1 and BS2, an incoming light beam is split by BS1 and after reflection at M1 or M2 passes through BS2. Constructive interference occurs in direction of exit 1 if the optical path lengths of the two beams are exactly equal. But the light in exit 2 cancels by destructive interference.

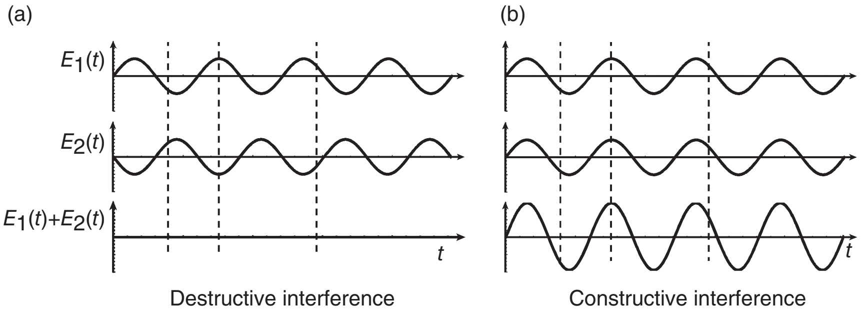


Figure 1.2 Destructive and constructive interference. (a) Two waves cancel each other if the electric fields of two interfering light waves $E_1(t)$ and $E_2(t)$ are always of opposite value, that is, they have a phase shift of π. (b) Constructive interference occurs if the two waves are completely in phase.

The explanation for this effect of interference lies in the wave nature of light. According to the wave description, light is an electromagnetic field oscillating with time. If two electromagnetic waves with identical amplitudes spatially superimpose but oscillate in exactly opposite directions, they will completely cancel each other (Figure 1.2a). Exactly this happens at exit 2 of our instrument. At exit 1, constructive interference of the two beams coming along light path 1 and 2 occurs. In constructive interference, the two waves oscillate in phase and therefore add up (Figure 1.2b). That is why all light is exiting at exit 1. Of course, the exact reasons for the asymmetric interference processes are not immediately obvious, since they are a consequence of subtle asymmetries in light path 1 and 2 for the two beams reaching the different exits. The effect will be explained later in detail after we covered some basic features of light (see Box 1.4).

The explanation for the second part of the experiment – the effect of blocking one of the light paths – is simple. Blocking one light path will prevent interference and yield 25% brightness at either exit, as expected when splitting 50% of the total input light again in two equal amounts.

The discussed instrument is called a *Mach–Zehnder interferometer*. Such interferometers are extremely sensitive instruments capable of detecting extremely small path length differences in the two arms of the interferometer.

1.2.2 Physical Description of Light Waves

In wave-optical terms, a light ray is an oscillating and propagating electromagnetic field. What does that mean?

Figure 1.3 provides a graphical representation of such a light wave in the simplest case – a plane wave in vacuum. The wave comprises an electric and a magnetic field component. Both oscillate perpendicular to each other and also to the propagation direction of the wave. Therefore, light is called a *transverse wave*. Since most of the effects of light on matter are caused by its electric field, we often neglect for simplicity the magnetic field altogether. In the figure, several important parameters describing such waves are indicated. The wavelength, λ, describes the spatial distance between two electric field maxima. The direction

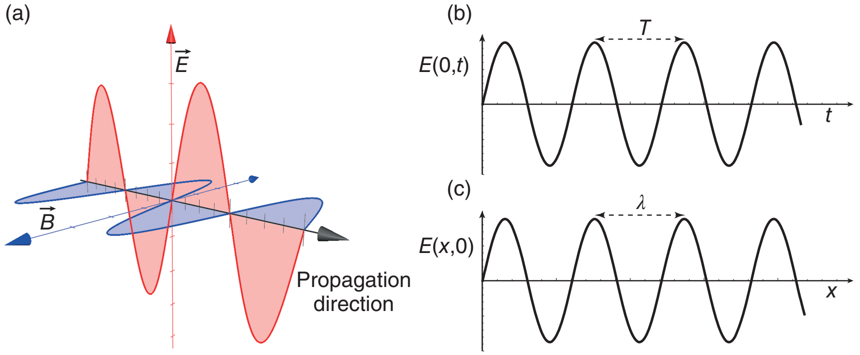


Figure 1.3 Sketch of a linearly polarized electromagnetic wave. (a) Wave with electric and magnetic field components, $\vec{E}$ and $\vec{B}$. (b) Temporal oscillation at a fixed place in space. (c) Still image of the wave.

and strength of the oscillating electrical field is given by the vector $\vec{E}$, and the propagation direction is characterized by the vector $\vec{k}$. The oscillation direction of the electric field in Figure 1.3 is constant in space. We call this the *polarization direction* of the light wave and such a wave *linearly polarized*. The polarization direction is not necessarily constant, for example, it may also rotate around the propagation direction. Such waves are called *elliptically* or *circular polarized* (see Box 1.1 for details).

The oscillation of the electric field as a function of time at a specific position in space is shown in Figure 1.3b. The time period until the electric field assumes again an identical profile is designated as oscillation duration T. The frequency $\nu = 1/T$ at which the electric field oscillates at a given position defines the *color* of the light wave. Electromagnetic waves exist over a vast range of frequencies, out of which only a very small range is perceived as "light" and detected by our eyes or cameras (Figure 1.4). Yet, the same wave theory of light governs

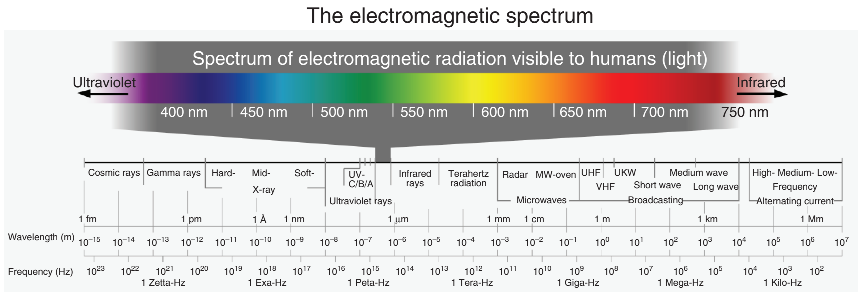


Figure 1.4 Electromagnetic spectrum. Different types of radiation are essentially electromagnetic waves with oscillation frequencies or vacuum wavelengths ranging over many orders of magnitude. English version of a graphic by Horst Frank (https://de.wikipedia.org/wiki/Elektromagnetisches_Spektrum, https://en.wikipedia.org/wiki/GNU_Free_Documentation_License).

all wavelength ranges of the electromagnetic spectrum, from cosmic waves to gamma rays. However, light microscopy uses only the visible and near-infrared range. Blue light has a higher frequency ν and higher energy $E = h\nu$ per photon than green, yellow, red, and infrared light. Here, h denotes Planck's constant. Figure 1.3c shows a still image of the wave. The spatial distance between two positions in which the electric field assumes identical values is designated as wavelength λ. In vacuum, the oscillation frequency ν and wavelength λ are related to each other as follows:

$$\lambda\nu = c \tag{1.1}$$

with c denoting the speed of light in vacuum. The wavelength λ is short for blue (about 450 nm), and longer for green (~520 nm), yellow (~580 nm), red (~630 nm), and infrared (~800 nm) light.

A short summary of the mathematical description of waves in space and time using trigonometric functions is given in Box 1.1. A mathematically more advanced description uses complex numbers and functions, a notation that is used in several later chapters of this book and considerably simplifies calculations.

Box 1.1 Mathematical Description of Waves

A harmonic, linearly polarized plane wave in space traveling in the x-direction is described as

$$\vec{E}(x,t) = \vec{E_0} \sin\left(\frac{2\pi}{\lambda}x - \frac{2\pi}{T}t\right) \tag{1.2}$$

where $\vec{E_0}$ describes the amplitude and direction of the oscillating electric field, λ is the wavelength, T is the oscillation duration, and t the time. A still image of the wave, for example, at time $t = 0$, showing the spatial profile of the wave

$$\vec{E}(x,0) = \vec{E_0} \sin\left(\frac{2\pi}{\lambda}x\right) \tag{1.3}$$

is shown in Figure 1.3c.

After the distance $x = \lambda$, the wave profile is repeated. The ratio $k = 2\pi/\lambda$ is designated as the wave number. At a specific position in space, for example, at $x = 0$, we see a periodic change of the electric field with the oscillation duration T (Figure 1.3b):

$$\vec{E}(0,t) = \vec{E_0} \sin\left(-\frac{2\pi}{T}t\right) \tag{1.4}$$

The frequency of the oscillation is given by $\nu = 1/T$, whereas the ratio $\omega = 2\pi/T$ is called the *angular frequency*. We see that $\omega = 2\pi\nu$. Inserting the respective definitions into Eq. (1.2) yields

$$\vec{E}(x,t) = \vec{E_0} \sin(kx - \omega t) = \vec{E_0} \sin\left[k\left(x - \frac{\omega}{k}t\right)\right] \tag{1.5}$$

We see that the wave moves in the direction of the positive x-axis. When time t advances, x must correspondingly advance such that argument and function

(Continued)

Box 1.1 (Continued)

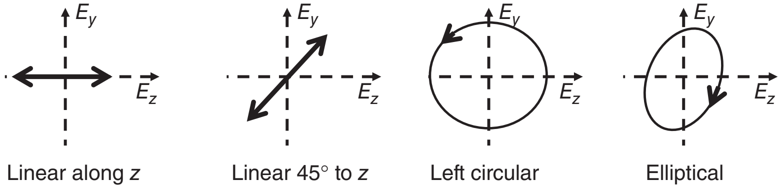
value remain constant. Thus the speed c of, for example, a wave crest is given by

$$c = \frac{\omega}{k} = \frac{2\pi\nu}{2\pi/\lambda} = \nu\lambda \tag{1.6}$$

For electromagnetic waves in a homogeneous isotropic medium, the oscillation direction $\vec{E_0}$ is perpendicular to the propagation direction. In this case, the direction $\vec{E_0}$ is constant in space and time. Such a wave is designated as *linearly polarized*, and the polarization direction is identical to the direction of $\vec{E_0}$. However, there are other cases possible. In the most general case, the wave can be described by

$$\vec{E}(x,t) = \begin{pmatrix} E_y \sin(kx - \omega t + \phi_y) \\ E_z \sin(kx - \omega t + \phi_z) \end{pmatrix} \tag{1.7}$$

where E_y, E_z, ϕ_y, and ϕ_z denote the amplitudes and phases with regard to the y- and z-directions. There are several cases possible. If the phase difference between both wave components is 0, that is, $\phi_y = \phi_z$, the light is linearly polarized for arbitrary choices of E_y and E_z. Then the polarization direction is given by the vector $\begin{pmatrix} 0 \\ E_y \\ E_z \end{pmatrix}$.



If $E_y = E_z$ and the phase difference $\phi_y - \phi_z = \pi/4$, we have a *circular polarized* wave. This means that the polarization vector of constant length is rotating around the propagation direction clockwise or counterclockwise depending on the sign of the phase difference (see sketch and online material). The so-called elliptically polarized light is obtained if $E_y \neq E_z$, or in the case where $E_y = E_z$ but $\phi_y - \phi_z \neq \pi/4$.

Water waves are well-known everyday examples that illustrate many properties of waves quite plainly. They provide, however, only a two-dimensional analogy to electromagnetic waves. However, it is merely an analogy to compare light to waves in a medium, such as water, where the particles of the medium actually are displaced. In the case of light, there is really no displacement of matter necessary for its description, as the basic oscillating quantity is the electric field, which can even propagate in vacuum. The moving line of a wave crest is a good conception of a *phase front.* A phase front of a light wave is the surface in space formed by the local maxima of the electric field or, more generally, the surface formed of any equal phase of the wave. Such phase fronts travel with the propagation speed of the wave. The waves we observe in water close to the shore that form straight lines approaching the beach are an analogy to what is called a *plane wave* as described in Box 1.1, whereas the waves seen in a pond, when we throw a stone into the water, are a two-dimensional analogy to a spherical wave.

Finally it should be noted that when discussing the properties of light, we often neglect the direction of the oscillating electric field, that is, the direction of the vector $\vec{E}$, and rather just use the scalar amplitude of the wave. This is just a bit careless, but convenient way of describing light when polarization effects do not matter for the experiment under consideration.

1.3 Four Effects of Interference

The wave nature of light can explain four fundamental aspects of light: diffraction, the refractive index, refraction, and reflection. Reflection at a mirror has the famous property of the incident angle corresponding to the angle of the light leaving the mirror. Refraction refers to the effect where light *rays* seem to change their direction when they pass from one medium to another. This is, for example, seen when trying to look at a scene through a glass full of water. It is connected with the fact that different media often show different refractive indices. Finally, diffraction is a phenomenon occurring when light interacts with very fine structures or openings, which are often quasi-periodic.

Even though these effects may seem very different at a first glance, all of these effects are ultimately based on interference. Diffraction is most prominent when light illuminates structures of a feature size similar to the wavelength of light. In contrast, reflection and refraction, for example, the bending of light rays caused by a lens, occur when different media such as air and glass comprise elements – molecules – that are far smaller than the wavelength of light but cover isotropic domains that are much more extended than the wavelength.

1.3.1 Diffraction

To describe diffraction, it is useful to first look at the light emitted by a point-like source. Let us consider an idealized source, which is infinitely small and emits only a single light color. This source then emits a spherical wave.

Christiaan Huygens (1629–1695) had a clever idea: to find out how a wave will spread in space, he proposed choosing an arbitrary border surface – meaning a surface of equal phase – then placing virtual point emitters everywhere on this surface, and letting the light of these emitters interfere. The resulting interference pattern will reconstitute the original wave beyond that surface. This “Huygens’ principle” can well explain that parallel waves remain parallel, because we find constructive interference only in the propagation direction of the wave. Strictly speaking, one would also find a backward-propagating wave. However, when Huygen’s idea is formulated in a mathematically rigorous way, the backward wave is not obtained. Huygens’ principle is very useful when predicting the spreading of a wave hitting a structure with feature sizes comparable to its wavelength, for example, a small slit aperture or a periodic diffraction grating. In Figure 1.5, we consider the example of diffraction at a grating of transmitting

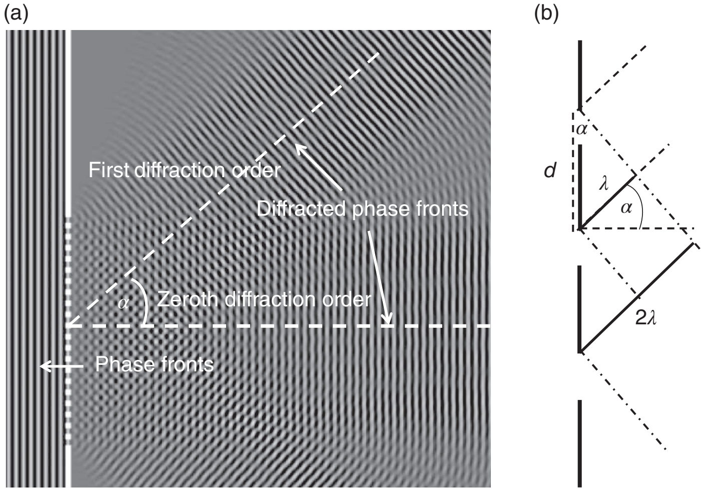


Figure 1.5 Diffraction at a grating. (a) A plane wave hits perpendicularly on a grating. The directions of constructive interference where maxima and minima of one wave interfere constructively with the maxima and minima of the second wave are shown for the zeroth- and first-order diffraction. (b) Magnified view of the grating geometry with the first-order diffraction direction. The condition for constructive interference is that the pathlength difference is equal to multiples of the wavelength λ.

slits. The distance between the slits is designated as the grating constant *d*. The figure shows how circular waves corresponding to Huygens' wavelets originate at each aperture and join to form new wave fronts, finally giving rise to plane waves traveling in various distinct directions. At most places, they interfere with a great diversity of phases, which means that they cancel in those locations, but in certain directions the wavelets add up constructively. This is seen best at large distances from the grating, where the individual waves join to form uniform plane phase fronts again (see the online supplemental material for an animated version of Figure 1.5). The directions of constructive interference can be characterized by their angle α with regard to the propagation direction of the incident wave. Then α fulfills the condition (Figure 1.5)

$$\sin \alpha = N\frac{\lambda}{d} \tag{1.8}$$

with N denoting an integer multiple of wavelengths λ to yield the same phase for creating constructive interference. N is called the *diffraction order*. Note that the angle α of the diffracted waves depends on the wavelength and thus on the color of light. In addition, note that the crests of the waves form connected lines, which are called *phase fronts* or *wave fronts*, whereas the dashed lines perpendicular to these phase fronts can be thought of as corresponding to the light rays of geometrical optics.

A compact disk is a good example of such a diffractive structure. White light is a mixture of light of many visible wavelengths. Thus, illuminating a compact disk with a white light source from a large distance will cause only certain colors to be diffracted from certain places on the disk into our eyes. This leads to the observation of the beautiful rainbow-like color effect when looking at it.

1.3.2 The Refractive Index

Understanding the behavior of light inside materials is not at all trivial. For a detailed treatment on an introductory level, we refer the reader to [1].

Somewhat simplified, Huygens' idea can also explain what happens when light traverses a homogeneous medium. Here, the Huygens emitters correspond to the real molecules inside the material. The incoming wave with its electric field induces an oscillation of the electrons with respect to their atomic nuclei. These oscillating electrons constitute accelerated charges that radiate a new electromagnetic wave. It turns out that the emission from each molecule is slightly phase-shifted with respect to the incoming wave. Notably, the magnitude of the phase shift as well as the amplitude of the emitted wave depends on the nature of the material (see Box 1.4). Even though each scattering molecule generates a spherical wave, the superposition of all the scattered waves from atoms at random positions will interfere constructively only in the propagation direction of the incoming wave. Thus, each very thin layer of molecules generates another parallel wave, which differs in phase from the original wave. The sum of the original sinusoidal wave and the interfering sinusoidal wave of scattered light results in a forward-propagating parallel wave with sinusoidal modulation, but typically lagging slightly in phase. In a dense medium, this phase delay is continuously occurring in every new layer of material throughout the medium, giving the impression that the wave has "slowed down" in this medium compared to vacuum. This reduction in the phase propagation speed is described by the refractive index n

$$c_{\text{medium}} = \frac{c_{\text{vacuum}}}{n} \tag{1.9}$$

According to Eq. (1.6), it can also be interpreted as an effectively reduced wavelength λ_{medium} inside the medium:

$$\lambda_{\text{medium}} = \frac{\lambda_{\text{vacuum}}}{n} \tag{1.10}$$

with λ_{vacuum} denoting the wavelength in vacuum. Note that the oscillation frequency of the electric field does not depend on the medium. Usually, n is dependent on the wavelength of the incoming light, $n = n(\lambda)$. This wavelength dependence of the refractive index is called *dispersion*.

1.3.3 Refraction

Now we analyze what happens when a plane light wave hits the interface between two materials with different refractive indices n_1 and n_2 at an angle α_1 to the surface normal, as shown in Figure 1.6. The wave directly at the interface is the incident wave, and the first layer of molecules of material 2 still "feels" this wave. This means that the phases of the light wave along the boundary must be identical in both materials. However, the wavelength of the light wave is $\lambda_1 = \lambda_{\text{vacuum}}/n_1$ in material 1 and $\lambda_2 = \lambda_{\text{vacuum}}/n_2$ in material 2. Both conditions can be fulfilled only if

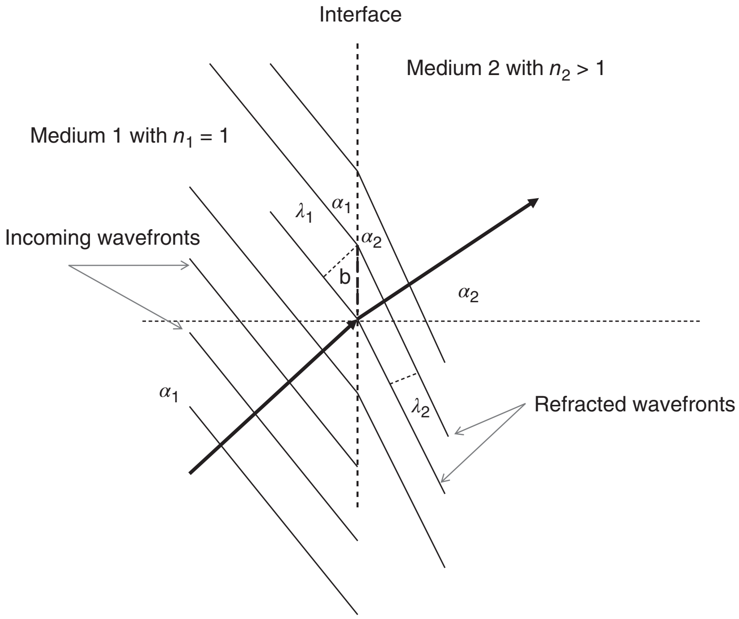


Figure 1.6 Snell's law of refraction. The phases of the electric field along the interface between the two materials must be identical. The wavelengths inside the materials are given by $\lambda_{\text{vacuum}}/n_1$ and $\lambda_{\text{vacuum}}/n_2$. We note that $\sin\alpha_1 = \lambda_1/b$, where b denotes the distance between two wave crests at the interface, and also that $\sin\alpha_2 = \lambda_2/b$. Eliminating b yields Eq. (1.12).

the wave changes its propagation direction in the second material. We denote the angles between the direction of propagation of the plane waves in material 1 and 2 and the perpendicular line onto the medium's surface by α_1 and α_2, respectively. We find (Figure 1.6) that

$$\frac{\lambda_1}{\lambda_2} = \frac{n_2}{n_1} = \frac{\sin\alpha_1}{\sin\alpha_2} \tag{1.11}$$

or

$$n_1 \sin\alpha_1 = n_2 \sin\alpha_2 \tag{1.12}$$

This is Snell's famous law of refraction, which forms the foundation of geometrical optics. One important use is to trace rays on their way through an optical system with multiple transitions between optical glasses and air.

1.3.4 Reflection

Light waves hitting surfaces are also reflected. There is an immediate consequence of considering the behavior of the Huygens wavelets at an interface. The law of reflection states that the angle of the incoming light is equal to the angle of the outgoing light. This can immediately be deduced by applying the reasoning used for the derivation of Snell's law by assuming that the wave is continuous along the interface and the reflected wave is traveling back into medium 1. It turns out that this simplified picture is a little bit too crude when it comes to fully understanding reflection at different materials such as metals and dielectrics, since a little bit of scattering volume in medium 2 is needed to provide enough power for the reflected wave. Even though the law of reflection is always valid, the phase of the reflected wave may change, which has to do with the penetration depth of the wave into the material. For a more detailed discussion on polarization and reflection, see Box 1.2.

1.3.5 Light Waves and Light Rays

How can we connect the concept of light rays as commonly sketched in geometrical optics with the concept of light as a wave phenomenon? A light ray represents a plane wave with a lateral extension small enough to be looked upon as a ray but extended enough not to show substantial broadening by diffraction. The light beam emitted by a laser pointer is a good example for such a ray. When light rays encounter interfaces between two media with different refractive indices, we can apply Snell's law to calculate what happens to the ray when it hits the interface at different angles. The ray bends toward the surface normal at the transition from an optically thin medium such as air with $n_1 = 1$ to an optically thicker medium such as glass with $n_2 = 1.52$ and $n_2 > n_1$. The opposite happens at a glass–air interface. Lenses have curved surfaces, and therefore the fate of the incoming beams depends on the position at which they hit the interface. Such calculations can rapidly be performed with high spatial resolution for complete lens surfaces as a function of the wavelength using computers. Such ray-tracing computations serve to exactly predict the effect of lenses of any shape and are a key tool to design modern optical systems.

Box 1.2 Polarization and Reflection

Light in a homogeneous isotropic medium is a transverse electromagnetic wave. This means that the vector of the electric field is perpendicular to the direction of propagation. Sometimes we can simplify this view by ignoring the direction of the electric field vector and treating the amplitude as a scalar. However, there are situations in which the direction of the electric field vector, that is, its polarization direction, is important. Ultimately polarization-dependent effects can be traced back to the fact that scattered light emitted from individual molecules ("dipole emission") has an angular dependence rather than being a spherical wave with uniform strength.

Transparent materials such as glass reflect a certain but small amount of light at their surface, even though they are 100% transparent once the light is inside the material. At perpendicular incidence, ~4% of the incident light is reflected. Under oblique incidence, the amount and the phase of the reflected light strongly depend on the polarization of the incident light. This is quantified by the so-called Fresnel reflection coefficients. These coefficients are plotted in Figure 1.7 for light hitting an air/glass and a glass/air interface. The plane that is formed by the propagation direction of the incident beam and the surface normal is called the *plane of incidence*. The Fresnel coefficients differ largely for the polarization component oscillating within the plane of incidence, the p-component (from the

(Continued)

Box 1.2 (Continued)

German word “parallel”), and the polarization component perpendicular to it, the so-called s-polarization (from the German word “senkrecht” for perpendicular). There is a specific angle at which the p-component has a zero reflectivity coefficient. At this so-called Brewster angle, this component is entirely transmitted into the glass without any reflection (Figure 1.7a). Let us assume that we illuminate the surface with unpolarized light precisely at the Brewster angle. Then the reflected light will be completely polarized in the direction perpendicular to the incidence plane, the s-direction. The p-component completely enters the glass. Interestingly, the Fresnel coefficients for a glass–air interface predict a range of total reflection when the inclination angle of the incoming light is beyond a certain angle, the so-called critical angle (Figure 1.7b). This effect is employed in total internal reflection microscopy, which is abbreviated as TIRF microscopy (see Chapter 9).

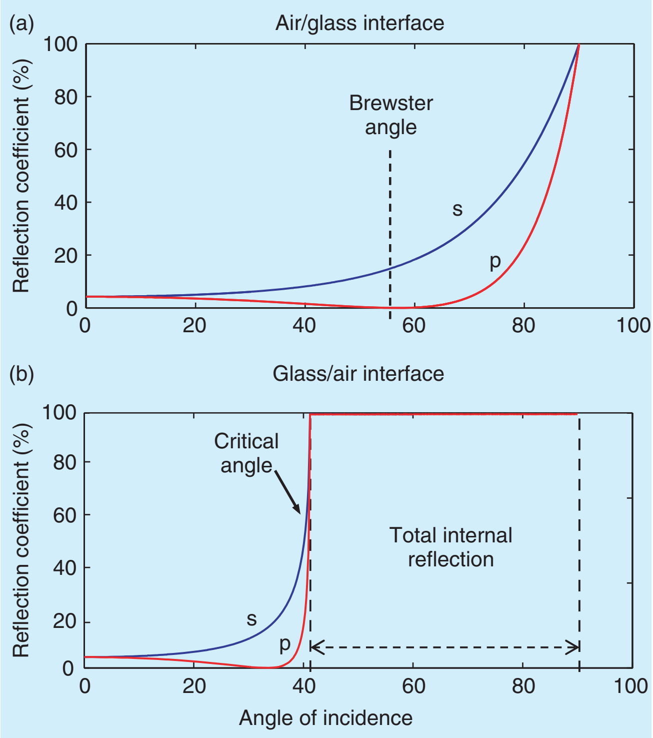


Figure 1.7 Fresnel coefficients quantify the reflectivity of interfaces. (a) Reflectivity of an air/glass interface for p (red) and s (blue) polarization directions of the incoming light. “p” means that the polarization vector is in the plane of incidence, whereas “s” means that the polarization vector is perpendicular to this plane. The angles are given with respect to the surface normal of the interface. The refractive indices for air and glass are $n_{\text{air}} = 1.0$ and $n_{\text{glass}} = 1.52$. (b) Reflectivity of a glass/air interface. Note the total reflection at supercritical angles.

There are crystalline materials that are not isotropic: that is, their unit cells have an asymmetry that leads to different refractive indices for different polarization directions. This effect is designated as *birefringence*. A beam entering such a material will usually be split into two beams within the birefringent crystal that travel into different directions. By cutting crystal wedges along different directions and joining them together, one can create the so-called Wollaston or Normaski prisms. The p- and s-polarized light components leaving the crystal will be slightly tilted with respect to each other. Such prisms are used for differential interference contrast (DIC) microscopy (Chapter 2).

1.4 Optical Elements

With the knowledge that light should ultimately be described as a wave, we can now move to more practical aspects of what can be done with light. In the context of this book, we need to understand various optical elements that are used in microscopes: lenses, mirrors, pinholes, filters, and chromatic reflectors. Although the wave picture is essential in microscopy, it can nevertheless be sometimes useful to approximate light propagation using the ray picture. This is the realm of "geometrical optics," which will be used for some of the following considerations.

1.4.1 Lenses

Here we will analyze a few situations to understand the general behavior of lenses. In principle, we should use Snell's law to calculate the shape of an ideal lens with perfect focusing ability. However, this would be beyond the scope of this chapter. Rather, we assume that a spherical lens made from glass with refractive index n and radii of curvature R_1 and R_2 – both positive for a convex surface – focuses parallel incoming light rays into a point at the focal distance f behind the lens. Then f is given by [2]

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} + \frac{1}{R_2} - \frac{(n-1)d}{nR_1R_2}\right] \tag{1.13}$$

where d denotes the thickness of the lens measured at its center on the optical axis. The above equation is called the *lensmaker's equation* for air. If the lens is thin and the radii of curvature are large, the term containing $d/(R_1R_2)$ can be neglected, yielding the equation for "thin" lenses:

$$f = \frac{1}{(n-1)\left({}^{1}/_{R_1} + {}^{1}/_{R_2}\right)} \tag{1.14}$$

This approximation is usually made, and the rules of geometrical optics as stated below apply.

The beauty of geometrical optics is that one can construct *ray diagrams* with pencil and paper and can graphically work out what happens in an optical system. Figure 1.8a shows how all the rays parallel to the optical axis are focused

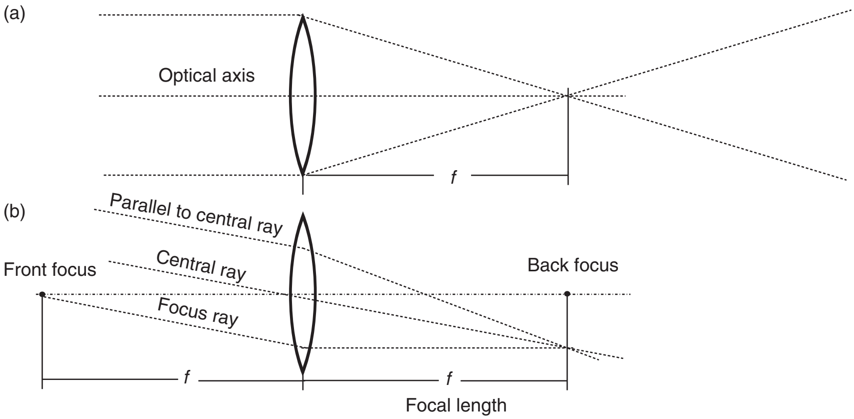


Figure 1.8 Focus of a lens under parallel illumination. (a) Illumination parallel to the optical axis. (b) Parallel illumination by beams inclined with respect to the optical axis, yielding a focus in the same plane but at a distance from the optical axis.

onto the focus of the lens. The optical axis is the symmetry axis of the lens and describes the general direction of the rays. Using the wave picture, we can say that a plane incoming wave is transformed by the lens to a spherically converging wave behind the lens, which converges at the lens focus. Figure 1.8b shows that this is also true for parallel rays entering the lens at an oblique angle. They are also focused onto the same plane. There are two more basic rays used for the geometrical construction of ray diagrams. The ray traversing the center of a thin lens is always unperturbed; it is called the *center ray*. This is easily understood, as the material is oriented at the same angle on its input side and exit side. For a thin lens, this “slab of glass” is considered to be infinitely thin. Thus, we can follow the ray right through the center of the lens.

The other key ray is traversing the front focal point of the lens at any angle and leaving the lens as a ray parallel to the optical axis. In geometrical optics of thin lenses, lenses are assumed to be symmetrical. Thus, the front focal distance of a thin lens is the same as the back focal distance.

The principle of *optical reciprocity* of geometrical optics states that we can always retrace the direction of light rays and yield identical paths of the rays. Obviously, this is not strictly true for any optical component. For example, absorption filters will not lead to amplification when the rays are propagated backward! However, it follows from this principle that any ray traversing a lens focus on the optical axis will leave the lens parallel to the optical axis.

Parallel incoming light is often referred to as coming from sources *at infinity*, as this is the limiting case when moving a source further and further away from the lens. Let us now consider what happens to the light emitted from an object located at “infinity.” The starry night sky is a good example of light sources at practically infinite distance. A lens will map such an object to its focal plane because each object point source at an “infinite distance” generates a parallel wave with its unique direction. A lens will “image” such sources to unique positions in its focal plane. Therefore, telescopes produce images of the night sky in their focal plane.

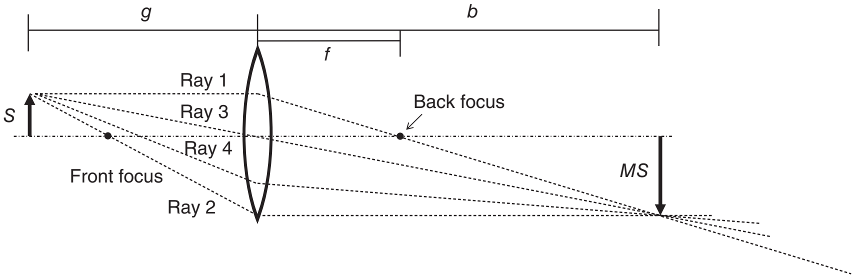


Figure 1.9 A single lens imaging an object as an example for drawing optical ray diagrams. For discussion of the ray paths, see text.

In Figure 1.9 we consider what happens to light emitted from an object of height S located at a distance g from a lens with focal length f. We assume that $g > f$. The trick now is to follow those key light rays for which we know their path through the lens:

- A ray parallel to the optical axis will pass through the back focus on the optical axis (Ray 1).
- A ray traversing the front focal point on the optical axis will end up being parallel to it behind the lens (Ray 2).
- A ray going through the center of the lens will pass through it unperturbed (Ray 3).

We know that lenses generate images, that is, if two special rays emerging from the same object point cross each other at a certain point on the image side, all other rays emerging from that object point will cross at this same point. Thus, if one such crossing is found, we are free to draw further rays from the same source object point to the same crossing on the image side, for example, Ray 4 in Figure 1.9.

We see that the image is flipped over and has generally a different size MS compared to the original size S, with M denoting the *magnification factor*. We find two conditions for similar triangles: $MS/b = S/g$ and $S/(g - f) = MS/f$.

This leads to the description of the imaging properties of a single lens:

$$\frac{1}{f} = \frac{1}{g} + \frac{1}{b} \quad \text{with the magnification} \quad M = \frac{b}{g} \tag{1.15}$$

where b is the distance between the lens and the image.

Using the aforementioned steps of optical construction, even complicated optical setups can be treated with ease.

1.4.2 Metallic Mirrors

In the discussion of refraction, we considered materials in which the electrons are bound to their nuclei. This leads to a phase shift of the scattered wave with respect to the incident wave. The situation is slightly different when an electromagnetic wave encounters a metal because the valence electrons of the metal are essentially free and very mobile. At the metal surface, the wave induces oscillations of the electrons, which oscillate such that they emit a wave that is phase-shifted by 180° (or π radians) with respect to the incoming one. This causes destructive interference along the propagation direction, and there is no transmitted wave. The reason for the 180° phase shift is that the conducting metal will always compensate for any electric field in its interior by moving charges such that the field vanishes. Thus, there is only a reflected wave from the surface, which means that very good mirrors can be made using planar metal surfaces. Application of Huygens' principle then leads to the law of reflection as discussed previously.

However, there is a small current induced by the electric field of the light wave wiggling at the electrons. Also, the material of the mirror has some electrical resistance. This will then lead to absorption losses. The reflectivity of a good-quality front-silvered mirror is typically 97.5% for wavelengths between 450 nm and 2 μm. Such unprotected, front-silvered mirrors are quite delicate and easy to damage, which often leads to much higher losses in praxis.

1.4.3 Dielectric Mirrors

Nonmetallic materials also show a certain amount of reflection. For example, glass with a refractive index of $n = 1.52$ shows about 4% reflection from its surface at perpendicular incidence of light owing to the interference effects of the waves generated in the material. Through careful deposition of multiple layers of materials with different refractive indices at well-defined thicknesses, it is possible to reach a reflectivity of almost 100% for very defined spectral ranges or, alternatively, over large wavelength bands and specific incidence angles.

The working principle of dielectric mirrors is sketched in Figure 1.10. A glass substrate is coated with a carefully designed series of layers of materials with different refractive indices. The layer thickness can be controlled such that only a

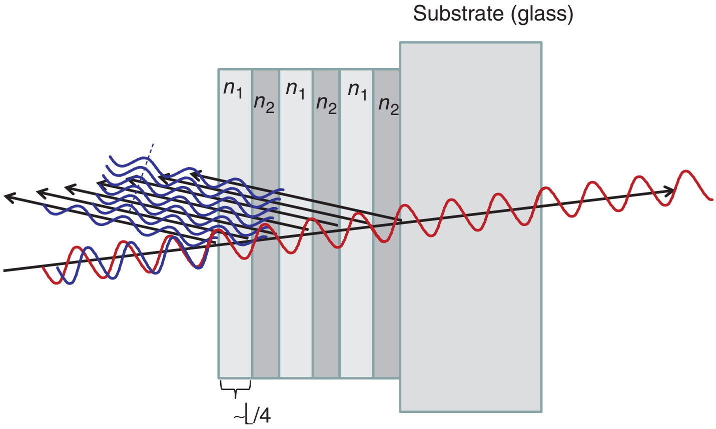


Figure 1.10 Principle of reflection by multilayer coatings. Dielectric mirrors, interference filters, and chromatic reflectors are all based on this principle. The exact order, thickness, and refractive index of the various layers of dielectric materials lead to a wavelength-dependent constructive or destructive interference for the reflected or transmitted light. In this manner, from very narrow-band notch filters to broadband dielectric mirrors with >99.9% reflectivity over a considerable spectral range can be constructed.

specific wavelength is reflected or transmitted. By varying the thickness of the layers, it is also possible to design the coating such that it reflects very well over a broad range of wavelengths. These layers yield multiple reflections and build up standing optical waves inside the material, which leads to the name "cavities" of these layers. In wavelength-selective mirrors, typically a thin layer of a material with a high index of refraction alternates with a thicker layer of a material with a low index of refraction such that the optical path for a reflected wave is an integer multiple of the wavelength, leading to constructive interference for the reflected light (Figure 1.10).

Such mirrors with selective or optimized reflectivity are employed in microscopes to steer the beam, to keep the microscope compact, to allow precise optical adjustment, and to switch between its various modes of operation. Dielectric mirrors are also used in lasers to form their cavity for amplifying the light. Such laser mirrors can have a reflectivity well above 99.999%.

The light losses of dielectric mirrors can be much smaller than for metallic mirrors, leading to overall higher performances. In addition, scratches on the mirror surface have less impact on the performance of these mirrors. However, one should be aware of the fact that waves with different polarizations penetrate the mirror material to different extents when they hit the mirror surface not perpendicularly. This leads to different phase shifts for p- and s-polarized waves. A linear polarization, when oriented at 45° to the plane containing the incident and reflected beams, for instance, is typically not conserved but is converted to elliptical polarization. This effect is particularly noticeable when the wavelength of the incident wave is close to the edge of the specified wavelength range of the coating.

Finally, such multilayer coatings can also be used to reduce the reflectivity. The usual 4% reflection at each air–glass and glass–air surface of optical elements can be significantly reduced by coating the lenses with layers of well-defined thicknesses and refractive indices. These are the so-called antireflection coatings. When viewed at oblique angles, antireflection-coated surfaces often display a blue, oily shimmer as can be noticed on photographic camera lenses.

1.4.4 Filters

There are two different types of optical filters and combinations of these. Absorption filters consist of a thick piece of glass in which a material with strong absorption in a certain wavelength band is embedded. Such filters have the advantage that a scratch will not significantly influence the filtering characteristics. A problem is that the spectral edge of the transition between absorption and transmission is usually not very steep and the transmission in the transmitted band of wavelengths is not very high. Note that when one uses the term *wavelengths* a bit carelessly, as in this case, one usually refers to the corresponding wavelength in vacuum and not to the wavelength inside the material.

The so-called interference filters are always coated on at least one side with a multilayer structure of dielectric materials as discussed previously. These coatings can be tailored precisely to reflect exactly a range of wavelength while effectively transmitting another well-defined range. Such coatings function by interference, and therefore there is an inherent angular and wavelength dependence.

This means that a filter placed at a 45° angle will transmit a different range of wavelengths than when placed at perpendicular incidence of the incoming light. This has consequences for the practical use of such filters in microscopy, as outlined in Box 1.3.

Box 1.3 Practical Aspects of Using Filters in Microscopy

We will see later that in a fluorescence microscope the exact object position in the front focal plane of the objective lens will define the specific angle at which the light from this point leaves the objective. Fluorescence filters are typically placed in the space between the objective and the tube lens. Therefore, this angle is also the angle of incidence on the filter. Theoretically, there is therefore a position-dependent color sensitivity. However, the fluorescence spectra are rather broad, and the angular spread for a typical field of view is in the range of only ±2°. Therefore, this effect can be completely neglected for fluorescence microscopy.

In microscopic imaging, it is important to reduce the background light as much as possible. Background can stem from the generation of residual fluorescence or Raman excitation even in glass. For this reason, optical filters always need to have their coated side facing the incident light. Because the coatings usually never reach completely the edge of the filter, one can determine the coated side by careful inspection. When building setups in-house, one also has to ensure that no light can possibly pass through an uncoated portion of a filter, as this would have disastrous effects on the suppression of unwanted scattering.

1.4.5 Chromatic Reflectors

Chromatic reflectors are designed to reflect a specific wavelength range, which is usually shorter than a critical wavelength, and to transmit another range. Like normal interference filters, chromatic reflectors are manufactured by multilayer dielectric coating. The comments about the wavelength and angular dependence given in Section 1.4.4 also apply here. Actually, they are especially valid for chromatic filters, which are mostly used at an incidence angle of 45°. In this case, the angular dependence is much stronger and the spectral steepness of the edges is much softer than for angles closer to perpendicular incidence.

Chromatic reflectors are often designated as *dichroic mirrors.* This is a confusing and potentially misleading term because the physical effect of “dichroism”, which refers to a polarization-dependent absorption, has nothing to do with the functioning of these special mirrors.

Box 1.4 Beam Paths and Interference in the Mach–Zehnder Interferometer

Finally, we can examine the two beam paths in the Mach–Zehnder interferometer shown in Figure 1.1. Of course, the key to the working of the instrument is the phases of the waves traveling through various paths. Having discussed some basic properties of light in materials and at surfaces, we first summarize some detailed facts about the used components.

BS1 and BS2 reflect half the incident light and refract the other half through them. The speed of light in air is approximately the speed of light (*c*) in vacuum. However, the speed of light inside glass is significantly smaller than that in vacuum, namely about (2/3)*c*. Thus, when a light ray traverses a medium such as a glass plate, its phase will be altered by an amount that depends on the index of refraction of the medium and the path length within the medium. When a light ray hits an interface, and the material on the other side of the interface has a *higher* index of refraction (i.e., a *lower* speed of light) than the medium the light is traveling in, then the reflected light ray is shifted in its phase by exactly π. This is the case when light is reflected by a metallic mirror. When a light ray hits an interface and the material on the other side of the interface has a *lower* index of refraction, the reflected light ray does not change its phase. When a light ray passes from one medium into another, its direction changes as a result of Snell's law of refraction but there is no phase change at the interface (Figure 1.6).

Now we can consider the two light paths leading to exit 1 (see Figure 1.1).

Light path 1:

1. Reflection at the front of BS1 yields a phase change of π.
2. Reflection by M1 yields a further phase change of π.
3. Transmission through BS2 yields some constant phase change $\Delta\varphi$.

Light path 2:

1. Transmission through BS1 yields some constant phase change $\Delta\varphi$.
2. Reflection at M2 yields a phase change of π.
3. Reflection at the front of BS2 yields a phase change of π.

Summing up the phase shifts for the two paths shows that they are identical, $2\pi + \Delta\varphi$. The light leaving the instrument at exit 1 via the two paths is in phase, which results in constructive interference.

Now we consider the two light paths leading to exit 2:

Light path 1:

1. Reflection at the front of BS1 yields a phase change of π.
2. Reflection by M1 yields a further phase change of π.
3. Transmission through BS2 yields some constant phase change $\Delta\varphi$. Reflection at the inner surface of BS2 yields no phase change. Transmission through BS2 a second time yields again some constant phase change $\Delta\varphi$.

Light path 2:

1. Transmission through BS1 yields some constant phase change $\Delta\varphi$.
2. Reflection at M2 yields a phase change of π.
3. Transmission through BS2 yields some constant phase change $\Delta\varphi$.

The sum of the phase shifts for light path 1 yields $2\pi + 2\Delta\varphi$. The sum of the phase shifts for light path 2, however, is $\pi + 2\Delta\varphi$. Thus, they differ exactly by π, which results in destructive interference in the direction of exit 2. Notably, this result does not depend on the wavelength! This explanation followed the presentation of David M. Harrison [3].

1.5 Optical Aberrations

High-resolution microscopes are difficult to manufacture and usually come with quite tight demands on experimental parameters (e.g., the specified temperature and the refractive index of the mounting medium) that should be used. Thus, in the real world, the microscope will almost always show imperfections because the demands are not perfectly fulfilled. Usually, this leads to aberrations in the optical path. For a microscopist, it is very useful to be able to recognize such optical aberrations and classify them. One can then try to remove them through modifications to the setup or, in the extreme case, by using *adaptive optics*, where such aberrations are corrected by using special elements such as deformable mirrors that can modify and adjust the wave front before it passes through the optics and the sample.

The most basic modes of aberration called *tip*, *tilt*, and *defocus* are usually not noticed because they simply displace the image slightly along the x-, y-, and/or z-direction. They are not a problem because the image quality remains unaffected.

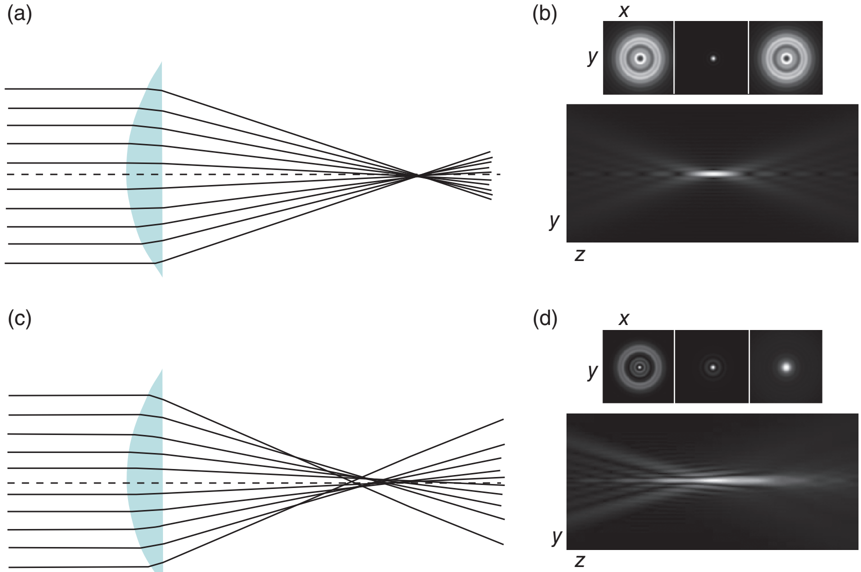


Figure 1.11 Spherical aberration. (a) Perfect lens without spherical aberration. (b) Lateral (top) and axial (bottom) beam profiles. The lateral profiles show the intensity before (left), at (center), and behind the focus (right). (c) An uncorrected lens showing positive spherical aberration: off-axis rays are refracted too much and miss the nominal focus. (d) This causes a focus asymmetry: overly sharp rings and a fine peak appear before the nominal focus, while at equal distance behind the focus a fluffy spot is seen. To improve visibility, the axial distributions show the square-root of the intensity and the lateral distributions show the intensity normalized to the brightest point. Negative spherical aberration occurs under certain conditions, and then the effect is reversed. (The graphics was inspired by http://en.wikipedia.org/wiki/Spherical_aberration.)

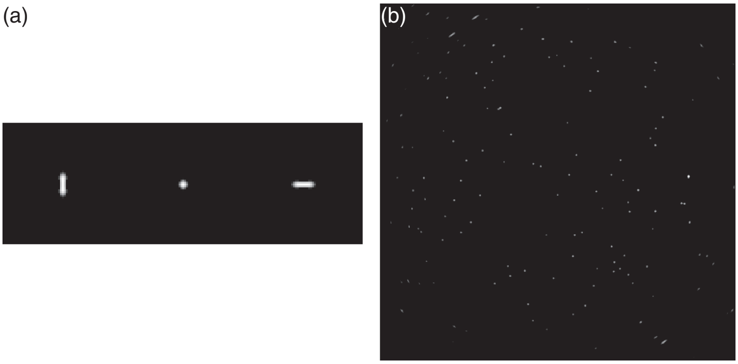


Figure 1.12 Astigmatism. (a) Image of a point object above, at, and below the focus displaying astigmatism. A vertical ellipse, a cross, and a horizontal ellipse can be observed. (b) Astigmatism as a field-dependent aberration in a microscopic image.

The most important aberration that indeed affects microscopic imaging is *spherical aberration.* In fact, a normal thick lens with spherical surfaces to which the lensmaker's equation would apply shows spherical aberration. Rays from the outer rim of the lens focus to a slightly different position than rays from the inner area, as shown in Figure 1.11. This blurs the focus and decreases the brightness when pointlike objects are imaged. A modern microscope objective can achieve an image free of spherical aberration when its optics is carefully designed. However, even for quality objectives, spherical aberration is commonly observed in praxis. This is because of a mismatch between the room temperature and the design temperature, or because the objective is not used according to its specifications. This happens, for example, when a sample is embedded in water with a refractive index of $n = 1.33$, and imaged by an oil-immersion objective designed for samples embedded in materials having a refractive index similar to that of oil, namely $n = 1.52$. Another common reason for spherical aberrations is the use of coverslips with a different thickness than that for which the objective was designed. A convenient way to notice spherical aberration is to pick a small, bright object and manually defocus up and down. If both sides of defocus essentially show the same fuzzy pattern of rings, the setup is free of spherical aberrations. However, if an especially nice pattern of rings is seen on one side and no rings or very fuzzy rings are seen on the other side, there are strong spherical aberrations present.

Another common aberration is *astigmatism.* When defocusing a small pointlike object up and down, one sees an ellipse oriented in a particular direction above the plane of focus and a perpendicular orientation of the ellipse below the focus (Figure 1.12a). Astigmatism is often seen in images at positions far away from the center of the ideal field of view (Figure 1.12b). A close observation of the in-focus point objects shows a star- or cross-shaped appearance. In astigmatism, there is essentially a disparity between the focus along the x-direction and the focus along the y-direction.

The last aberration to mention here is *coma.* Coma can arise from misaligned – tilted – lenses. At the position of best focus, a point object imaged with a coma aberration looks like an off-center dot, as opposed by an asymmetric quarter-moon-shaped comet when in focus. The farther one goes from the focus, the more prominent is the asymmetry. What should look like a fluffy ring seems to be missing a part.

Aberrations can also result in a position-dependent displacement of image points called *radial distortion*. Depending on the way that the positions are distorted, one discriminates between a barrel distortion, in which x- and y-axes are pushed to the outside, or a pin-cushion distortion, where the x- and y-axes are squeezed to the center.

A final monochromatic distortion is designated *curvature of field* or *Petzval field curvature.* The image of a flat, planar object cannot be sharply projected onto a plane in the image space. A perfectly sharp image can be produced only on a curved surface. This generates problems when objects are imaged with a flat two-dimensional image sensor such as a charge-coupled device (CCD) camera chip. Image points near the optical axis will be perfectly in focus, but off-axis rays come into focus before the image sensor. Effectively, the image cannot be sharp over the complete field of view.

We often use different colors for microscopic imaging, especially when in fluorescence microscopy (Chapter 3). Therefore, it is useful to examine how the light paths through lenses depend on the specific light color. The index of refraction of most materials depends on the wavelength. According to the lensmaker's equation, a single lens made out of one type of glass will therefore show a wavelength-dependent focal length f. As a consequence, the images of an object seen in different colors will have slightly different sizes and also different positions in the image space. The shift of the image along the optical axis is called *axial chromatic aberration.* The wavelength-dependent magnification is called *lateral chromatic aberration.* By combining different lenses of different materials into a single optical element, it is possible to at least partially compensate for this dispersion effect. A typical example is the readily available achromatic doublet lenses, where two lenses of different materials and curvatures are combined to yield a focus largely free of chromatic aberrations. Nevertheless, all microscopes show a fair amount of – predominantly axial – chromatic aberration. This can be demonstrated by imaging the same object in red and blue light. Usually, there is a significant shift along the axial direction between the respective positions of the two images.

References

1 Feynman, R.P. (2011) *The Feynman Lectures on Physics*, The New Millennium Edition: Mainly Mechanics, Radiation and Heat, vol. **1**, Basic Books, New York.

2 Hecht, E. (2002) *Optics*, 4th edn, Chapter 6.2, Addison-Wesley. ISBN: 0-321-18 878-0.

3 Harrison, D.M. (1999) *The Physics Virtual Bookshelf*, University of Toronto, http://www.upscale.utoronto.ca/GeneralInterest/Harrison/MachZehnder/MachZehnder.html (accessed 25 October 2016).