Chapter 1
Geometrical Optics

1.1 General Comments

A light wave is an electromagnetic wave, and the wavelength that optics studies ranges from the ultraviolet ($\sim$0.2 μm) to the middle infrared ($\sim$10 μm). The spatial scales involved in most optical applications are much larger than the light wavelength. In these cases, a light wave can be approximately described by a bundle of straight optical rays. The science of studying optical rays that travel through optical media is called "geometrical optics" and is the most widely used field of optics. If the spatial scales involved are not much larger than the light wavelength, the wave nature of light must be considered; the science of studying the wave nature of light is called "wave optics."

The technique to trace an optical ray through various optical media is called "sequential raytracing" and is the main way to study geometrical optics. This chapter briefly introduces basic geometrical optics using a sequential raytracing technique. (Smith[1] is a widely cited optical-engineering reference book, and Hecht[2] is a popular optics textbook. Both are recommended here as additional information sources.)

1.2 Snell's Law

When an optical ray travels from one optical medium to another that has a different refractive index, the ray is split in two by the interface of the two media. One ray is reflected back to the first medium. Another ray is refracted and enters the second medium, as shown in Fig. 1.1. The equation describing such a refraction is called Snell's law:[3]

$$n_1 \sin(\theta_1) = n_2 \sin(\theta_2), \tag{1.1}$$

where θ_1 is the angle between the incident ray and the normal of the interface of the media at the point where the ray hits, θ_2 is the angle between the refracted ray and the same normal of the interface, and n_1 and n_2 are the refractive indices of the two media, respectively.

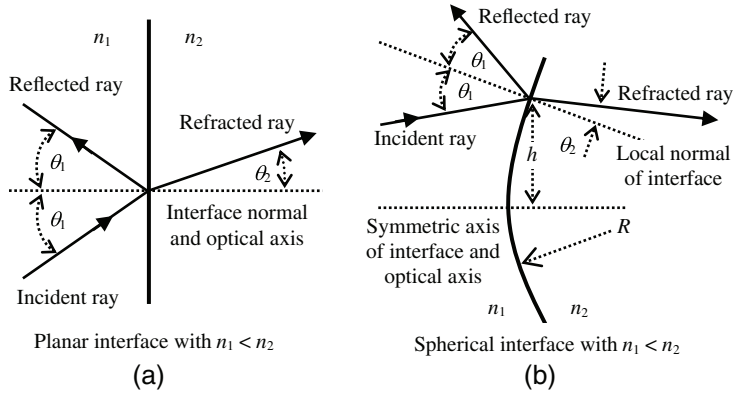

Figure 1.1 (a) An optical ray is incident on a planar interface of two optical media with refractive index n_1 and n_2, respectively, and $n_1 < n_2$. (b) An optical ray is incident on a convex interface of two media with refractive index n_1 and n_2, respectively, and $n_1 < n_2$. The symmetric axis of the interface is the optical axis. The incident angle varies as the ray height h varies. For $h \ll R$, where R is the interface radius of curvature, the interface can be considered planar.

Figure 1.1(a) illustrates a planar interface. The normal of the interface is the normal of the point at which the ray is reflected and refracted, and it is the optical axis. The reflection angle always equals the incident angle, which is the "reflection law" and will be discussed in Section 1.10.1. For $n_1 < n_2$, $\theta_1 > \theta_2$, and vice versa, according to Snell's law.

Figure 1.1(b) illustrates a curved interface. The symmetric axis of the interface is the optical axis. The incident ray height h is the vertical distance between the point on the interface where the ray hits and the optical axis. The figure shows that the incident angle θ_1 is a function of h and the interface radius of curvature R.

Snell's law is the foundation of geometrical optics, which is why every optics book mentions it.

1.3 Total Internal Reflection

Equation (1.1) shows that if $n_1 \sin(\theta_1)/n_2 > 1$, there is no solution for θ_2 because the maximum possible value of $\sin(\theta_2)$ is 1. Thus, no ray will be refracted and enter the second medium. Instead, the incident ray will be totally reflected at the interface. This phenomenon is called total internal reflection (TIR).[4]

For example, if medium 1 is N-BK7 glass with an index $n_1 \approx 1.52$ and medium 2 is air with an index $n_2 \approx 1.0$, then condition $n_1 \sin(\theta_1)/n_2 \geq 1$ leads to $\theta_1 \geq 41.1°$. Any incident ray with incident angle >41.1° will be totally reflected (41.1° is the critical angle in this case). The condition for TIR to happen is $n_1 > n_2$. A different ratio of n_1/n_2 would have a different critical angle.

Note that when a ray is incident on an interface of two media with an angle smaller than the TIR angle, the refracted ray will carry most of the energy of the incident ray, and the reflected ray will carry a relatively much smaller portion of the energy. The reflected energy can be calculated using wave optics theory, as will be discussed in Section 2.3. In many applications, the reflected energy is unwanted and is reduced by applying an antireflection (AR) coating on the interface.

1.4 Paraxial Approximation

Before the invention of computers, tracing a ray using Snell's law [Eq. (1.1)] through several curved optical interfaces required a lot of calculation and drawing. Efforts were taken to simplify the situation whenever possible. When $\theta_1 < 1$, the calculation can be simplified by expanding the sine function in Snell's law into a series:

$$\begin{aligned}
\theta_2 &= \sin^{-1}\left[\frac{n_1}{n_2}\sin(\theta_1)\right] \\
&\approx \frac{n_1}{n_2}\sin(\theta_1) + \frac{1}{6}\frac{{n_1}^3}{{n_2}^3}\sin(\theta_1)^3 + \frac{3}{40}\frac{{n_1}^5}{{n_2}^5}\sin(\theta_1)^5 + \cdots \\
&\approx \frac{n_1}{n_2}\left(\theta_1 - \frac{1}{6}{\theta_1}^3 + \frac{1}{120}{\theta_1}^5 \cdots\right) + \frac{1}{6}\frac{{n_1}^3}{{n_2}^3}\left(\theta_1 - \frac{1}{6}{\theta_1}^3 + \frac{1}{120}{\theta_1}^5 \cdots\right)^3 \\
&\quad + \frac{3}{40}\frac{{n_1}^5}{{n_2}^5}\left(\theta_1 - \frac{1}{6}{\theta_1}^3 + \frac{1}{120}{\theta_1}^5 \cdots\right)^5 + \cdots \\
&\approx \frac{n_1}{n_2}\theta_1 + \frac{1}{6}\left(\frac{{n_1}^3}{{n_2}^3} - \frac{n_1}{n_2}\right){\theta^3}_1 + \frac{1}{4}\left(\frac{1}{30}\frac{n_1}{n_2} + \frac{1}{3}\frac{{n_1}^3}{{n_2}^3} + \frac{3}{10}\frac{{n_1}^5}{{n_2}^5}\right){\theta_1}^5 \cdots .
\end{aligned} \tag{1.2}$$

The calculation accuracy requirement determines how many terms to keep in Eq. (1.2).

"Paraxial approximation"[5] means $\theta_1 \ll 1$ so that only the first-order terms of θ_1, $\theta_2 = (n_1/n_2)\theta_1$ must be kept for a fairly accurate calculation. Paraxial approximation is an ambiguous concept; there is no simple line to determine whether a θ_1 value is small enough to qualify for paraxial approximation. Rather, it depends on how accurate the calculation result must be. Figure 1.1(b) shows that $h \ll R$ leads to $\theta_1 \ll 1$. If the terms of the third order or higher of θ_1 are used, the calculation will be quite complex.

With the calculation power of modern computers, paraxial approximation is not as important as before. However, paraxial approximation is still useful for qualitative and manual analysis, and it is still used in much of the literature, partially due to tradition.

1.5 Lenses

1.5.1 Lens types

Lenses are the most commonly used optical component and can be separated into two categories: positive and negative.

Any lenses with a central thickness larger than their edge thickness are positive lenses. A positive lens can convergently refract rays that pass through the lens, i.e., focus the rays passing through the lens.

Any lenses with a central thickness smaller than their edge thickness are negative lenses. A negative lens can divergently refract the rays passing through the lens.

Figure 1.2 shows several shapes of positive lenses and negative lenses. The symmetric axis of a lens is its optical axis. The functionality of a lens can be determined by tracing rays through the lens using Snell's law.

1.5.2 Positive lenses

Figure 1.3 shows how raytracing is performed through an equi-convex positive lens to analyze the lens function. The first ray usually traced is parallel to the optical axis of the lens, traced left to right through the lens, as indicated by

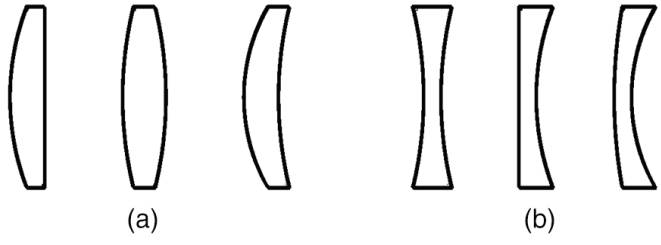

Figure 1.2 Three shapes of a (a) positive lens and (b) negative lens.

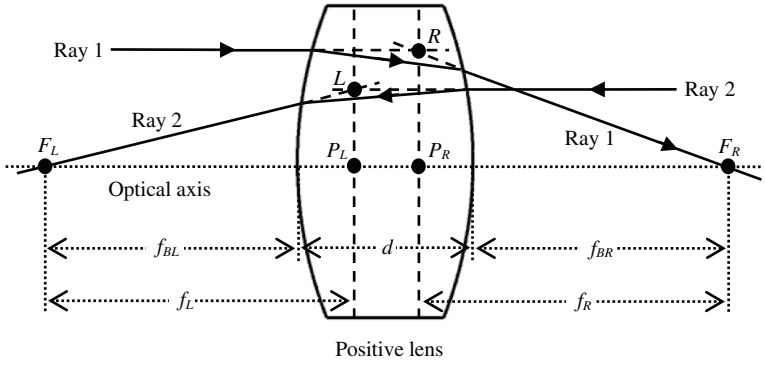

Figure 1.3 Two rays traced through an equi-convex positive lens. The rays are convergently refracted. The two cross-points of the rays and the optical axis are the two focal points of the lens, marked F_L and F_R, respectively.

Ray 1 in Fig. 1.3. The left surface of the lens convergently refracts the ray, and the right surface of the lens further convergently refracts the ray. As the ray travels forward, it eventually crosses the optical axis at point F_R, which is the right focal point of the lens. The ray is said to be focused.

The second ray traced is parallel to the optical axis of the lens and is traced from right to left through the lens, as marked by Ray 2 in Fig. 1.3. The right surface of the lens convergently refracts the ray, and the left surface of the lens further convergently refracts the ray. The ray eventually crosses the optical axis at point F_L, which is the left focal point of the lens. The ray is also focused.

For Ray 1, the forward extension of the incident ray and the backward extension of the exit ray meet at point R. For Ray 2, the forward extension of the incident ray and the backward extension of the exit ray meet at point L. Points L and R determine the axial locations of the two principal planes, which are shown by the two vertical dashed lines. The cross-points of the two principal planes and the optical axis are the two principal points, marked P_L and P_R, respectively, in Fig. 1.3.

The axial distance between the left (right) principal plane and the left (right) focal points is the focal length denoted by f_L (f_R), as marked in Fig. 1.3. f_L always equals f_R. Lens thickness d is defined as the axial distance between the two vertices of the lens surfaces.

The "back focal length" is the axial distance between the left (right) focal point and the vertex of the lens left (right) surface, as marked by f_{BL} and f_{BR}, respectively, in Fig. 1.3. f_{BL} equals f_{BR} only when the two surfaces of the lens have the same shape but opposite orientations. The equi-convex lens shown here is such a lens.

Positive lenses of any shape can focus rays, and their focal lengths are, by definition, positive.

1.5.3 Negative lenses

Figure 1.4 shows how rays are traced through an equal-concave negative lens. Again, the first ray traced is parallel to the optical axis of the lens and is traced from left to right through the lens, as marked by Ray 1 in Fig. 1.4. The left surface of the lens divergently refracts the ray, and the right surface of the lens further divergently refracts the ray. The ray never crosses the optical axis as it travels forward. However, a virtual ray that is the leftward extension of the exit Ray 1 can be conceived, as shown by the dashed line in Fig. 1.4. This virtual ray will cross the optical axis at point F_L, which is the left focal point of the lens. Such a focal point is a virtual focal point; the ray is never actually focused.

Similarly, the second ray traced is parallel to the optical axis of the lens and is traced from right to left through the lens, as marked by Ray 2 in Fig. 1.4. The right-side surface of the lens divergently refracts the ray, and the left-side surface of the lens further divergently refracts the ray. The ray

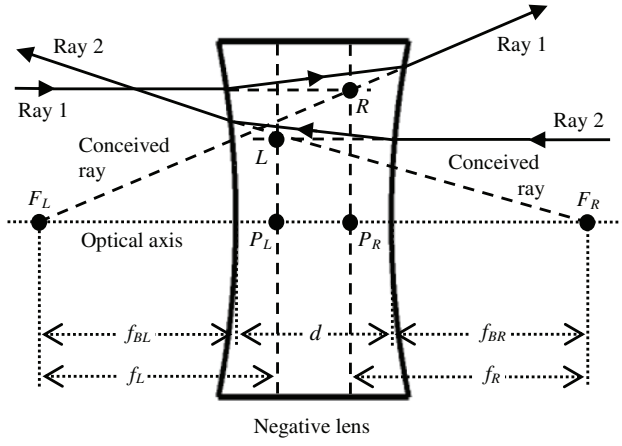

Figure 1.4 Two rays traced through an equal-concave negative lens. The rays are divergently refracted. The two cross-points *L* and *R* of the backward extension of the exit rays and the optical axis are the two focal points of the lens, marked by F_L and F_R, respectively. Points *L* and *R* determine the axial positions of the two principal points P_L and P_R.

never crosses the optical axis. The conceived virtual ray that is extended rightward from the exit Ray 2 will cross the optical axis at point F_R, which is the right-side focal point of the lens. Such a focal point is also a virtual focal point since the ray is never focused.

The two principal points and planes of this negative lens can be found by using the same raytracing technique as described earlier for the equi-convex lens. The focal length and back focal length of a negative lens are measured the same way as for the equi-convex positive lens. However, the focal length of any negative lens is defined as negative. No negative lens will ever focus rays.

1.5.4 Cardinal points

Any lens has three pairs of cardinal points:[6] a pair of principal points, a pair of focal points, and a pair of nodal points. The principal point pair P_L and P_R and the focal point pair f_L and f_R were discussed earlier and marked in Figs. 1.3 and 1.4, respectively. This subsection further discusses the optical meaning of the principal points and planes.

In Figs. 1.3 and 1.4, if the lens is viewed from the right side, the refracted Ray 1 appears to be emitted from point *R*, and P_R is the optical position of the lens when viewed from the right side. If the lens is viewed from the left side, the refracted Ray 2 appears to be emitted from point *L*, and P_L is the optical position of the lens when viewed from that side.

Figure 1.5 shows the positions of the two principal planes for various shapes of lenses. The principal plane positions of some lenses can be outside

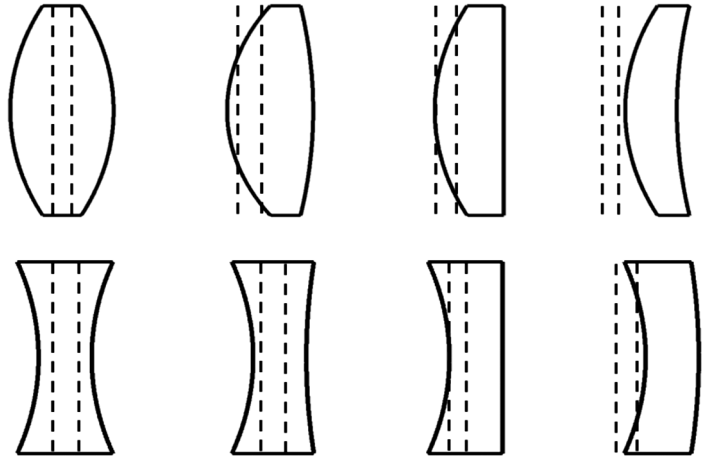

Figure 1.5 Positions of the two principal planes for various shapes of lenses.

the lenses. If the distance between the two principal points is much smaller than the focal length, the two principal points can be considered to coincide, and the lens is called a "thin lens."

The raytracings shown in Figs. 1.3 and 1.4 did not consider the effects of spherical aberrations and are therefore only approximations. The actual positions of the two principal points and the focal points, and thus the focal length, vary as the incident ray height changes.

Figure 1.6 shows accurate raytracing diagrams obtained using the optical design software Zemax. The diagrams illustrate how the position of the principal point changes as the incident ray height changes. The positions of the principal point in the paraxial approximation (i.e., the incident ray height approaches zero), marked by the larger dots in Fig. 1.6, are usually used. The positions of the focal points and the value of focal length usually used are also paraxial approximation values.

1.5.5 Nodal points

Every lens has a pair of nodal points. In most cases, including the case shown in Fig. 1.7, the medium at the left and right side of the lens are the same, such as air. Then the two nodal points coincide with the two principal points. Nodal points have two unique properties:

1. A ray aimed at one of the nodal points will be refracted by the lens such that it appears to come from the other nodal point and have the same angle with respect to the optical axis.
2. The right (left) focal point position of a lens is not shifted when the lens is rotated about its right (left) nodal point.

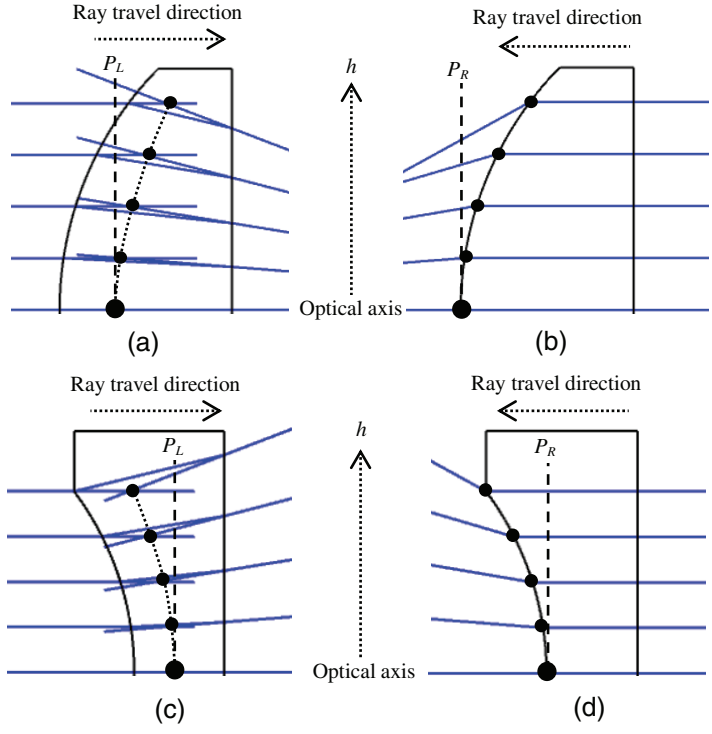

Figure 1.6 Zemax-generated accurate raytracing diagram of the principal point positions as a function of incident ray height *h*. The principal points are marked by dots. (a) and (b) Rays are traced right and left, respectively, to determine the positions of the two principal points for a positive lens with convex-planar surfaces. (c) and (d) Rays are traced right and left, respectively, to determine the positions of the two principal points for a negative lens with concave-planar surfaces.

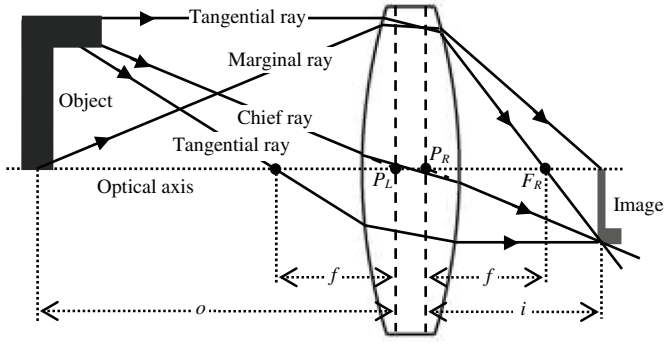

Figure 1.7 An equi-convex positive lens illustrates the nodal points, tangential plane and rays, and the formation of an image. Since the object is placed outside the focal point of the lens, the image formed is a smaller, inverse, and real image. The image is thinner than the object because the axial magnification (see Section 1.9) is less than 1.

Figure 1.7 illustrates the first property of the nodal points by tracing the chief ray through the lens. (Figures 14.11 and 14.12 in Section 14.6 demonstrate the second property of nodal points with two Zemax-generated, accurate raytracing diagrams.)

1.6 Rays and Planes

Raytracing is a geometrical optics technique widely used in the design and analysis of optical imaging systems. For a given object and lens, the approximate location, size, and orientation of the image formed by the lens can be found by tracing a few rays from the object through the lens. Although raytracing is now mainly performed by computers and optical software, familiarity with raytracing is still helpful.

To effectively trace rays, some special rays and planes must first be defined. Each of them has a special name and meaning. Figures 1.7 and 1.8 use the simplest one-lens example to illustrate these rays and planes.

1.6.1 Tangential planes and tangential rays

For a given lens and object, two orthogonal planes and several types of rays are defined. The tangential plane is defined by the optical axis and the object point from which the ray originated. In Fig. 1.7, the tangential plane is the plane of the page. A tangential plane is also called a meridional plane.

A ray that travels in the tangential plane is a tangential ray or meridional ray. All four rays shown in Fig. 1.7 are tangential rays or meridional rays. There are different types of tangential rays, such as chief and marginal rays.

The chief ray travels from the top of the object through the center of the aperture stop (see Section 1.7). In Fig. 1.7, the lens aperture is the aperture stop. A chief ray is also called a principal ray and is frequently used in raytracing.

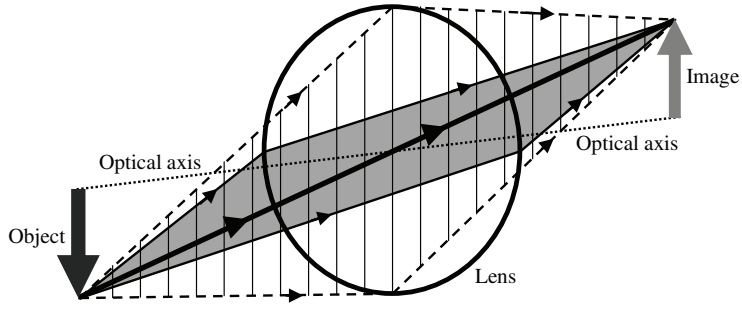

Figure 1.8 Sagittal plane and sagittal rays. The plane marked by the thin vertical lines is the tangential plane. Dashed lines indicate tangential rays. The shaded plane is the sagittal plane. All of the rays in the sagittal plane are sagittal rays, drawn with solid lines. The intersection line of the two planes is the chief ray, denoted by the thick solid line.

A marginal ray travels from the point where the object and the optical axis cross the edge of the aperture stop. In Fig. 1.7, the lens edge is the aperture stop edge. Tracing at least two tangential rays from an object point through a lens can determine the location of the image point, as shown in the figure.

1.6.2 Sagittal planes, sagittal rays, and skew rays

There are an infinite number of planes that are perpendicular to the tangential plane. Among these planes, only one plane contains the chief ray; such a plane is called the sagittal plane, as shown in Fig. 1.8.

Rays traveling in the sagittal plane are sagittal rays. The chief ray is the intersecting line of the tangential and sagittal planes. Technically, it is also a sagittal ray, but it is usually considered as a tangential ray. Sagittal rays are also called transverse rays. Sagittal rays are rarely used in manual raytracing because it is difficult to conceive and draw rays perpendicular to the plane of the page. However, modern computers and optical software can easily trace sagittal rays.

Skew rays are those rays that neither travel in the tangential plane nor cross the optical axis anywhere, and they are not parallel to the optical axis. Sagittal rays, except the chief ray, are special skew rays.

1.7 Stops and Pupils

1.7.1 Definitions

Stops are important optical components in an optical system. Pupils are the images of stops. Most image lenses have an aperture stop and a field stop. The aperture stop sets the largest cone angle the lens imposes on the object. The field stop sets the largest field angle the lens can see. These stops can significantly affect the lens characteristics and should be understood. The stops can be either lens aperture edges or some other structures.

Most image lenses also have an entrance pupil and an exit pupil. These two pupils are the images of the aperture stop formed by the lens at the object side and image side, respectively. Manually locating the stops and pupils and determining their sizes can be complex, whereas computer and optical design software can easily perform these tasks. Three examples are included here to explain the process.

1.7.2 Example 1: a schematic microscope

Figure 1.9 includes a microscope that consists of an objective and an eyepiece to illustrate the concepts of an aperture stop, field stop, entrance pupil, and exit pupil. In Fig. 1.9(a), the aperture of the objective restrains the cone angle of the microscope imposing on the object and is the aperture stop. The aperture of the eyepiece restrains the maximum allowed field angle. For example, the rays from the solid square on the object plane are completely blocked by the edge of the

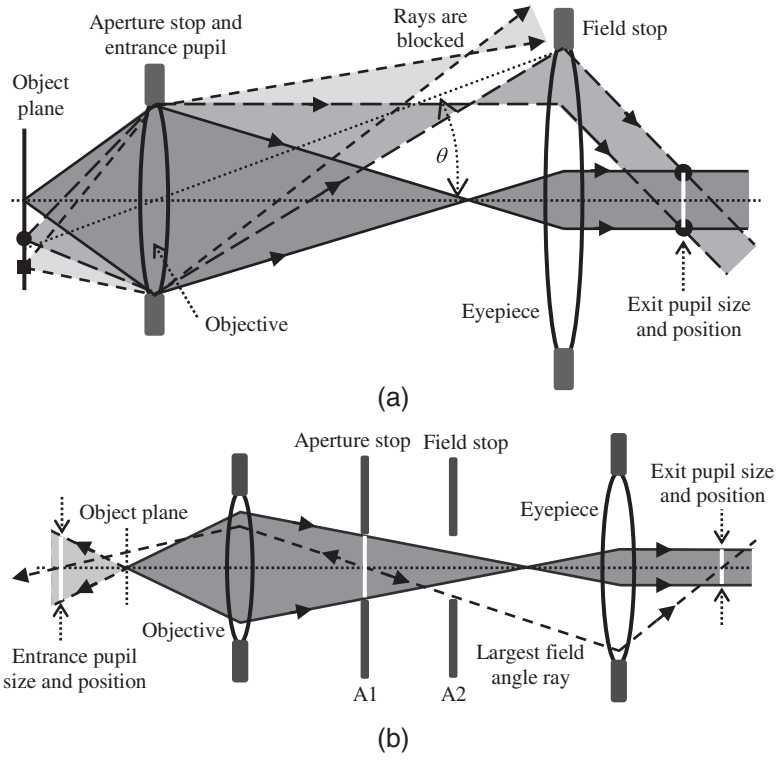

Figure 1.9 Schematics of two microscopes: (a) The microscope consists of an objective and an eyepiece. The aperture of the objective is the aperture stop and the entrance pupil. The aperture of the eyepiece is the field stop. The image of the objective aperture at the image space (right side of the microscope) is the exit pupil. (b) A similar microscope with two additional apertures A1 and A2 in the optical path. The image of A1 at the image space is the exit pupil.

eyepiece, and all of the rays from the solid dot on the object plane can travel through the eyepiece. The aperture of the eyepiece is the field stop. The largest field angle allowed by this field aperture is θ, marked in Fig. 1.9(a).

The entrance pupil of a lens is defined as the image of the aperture stop at the object space. In Fig. 1.9(a), there is no lens at the left side of the aperture stop. The aperture stop itself is the entrance pupil.

The exit pupil of a lens is defined as the image of the aperture stop at the image space. In Fig. 1.9(a), the exit pupil position and size can be found by tracing four rays rightward. The two rays from the top of the aperture stop (objective) meet at the point marked by a dot. The two rays from the bottom of the aperture stop meet at the point marked by another dot. These two dots determine the exit pupil location and size, as shown by the white vertical line linking the two dots.

Most lenses that generate an image for direct viewing have their exit pupil outside of the lenses. If a viewer places an eyeball at the exit pupil, they only

need to rotate their eyeball to view the entire scene; otherwise, they must move their head to view the entire scene.

1.7.3 Example 2: a schematic microscope with two additional apertures

For other lenses with different structures, the aperture stop and field stop locations and sizes can be very different from those shown in Fig. 1.9(a). Figure 1.9(b) shows a microscope similar to the one in Fig. 1.9(a) but with two additional apertures A1 and A2 in the optical path. Aperture A1, rather than the aperture of the objective, restrains the cone angle of rays from the object. A1 is the aperture stop.

A more standard approach to determine the field stop and the pupils traces rays from the center of the aperture stop both right and left, as shown by the dashed line in Fig. 1.9(b). Aperture A2, rather than the aperture of the eyepiece, restrains the largest field angle. A2 is the field stop.

By tracing the dashed-line ray left, the ray will cross the optical axis. The cross-point is the location of the entrance pupil. The size of the ray bundle at the entrance pupil location is the entrance pupil size, as marked by the white line.

By tracing this dashed-line ray right, it will cross the optical axis again. The cross-point is the location of the exit pupil. The size of the ray bundle at the exit pupil location is the exit pupil size, as marked by the white line.

In real lenses, the entrance and exit pupils may have severe aberrations caused by lens aberrations. The pupils can be curved. The locations of the pupils can be far away or at infinity. The pupil sizes can be large or even infinite. Minimizing pupil aberration is one of the optical design goals.

1.7.4 Example 3: a real double Gauss lens

Figure 1.10 presents a Zemax-generated raytracing diagram of a real double Gauss lens to further illustrate stops, pupils, and some other lens parameters. Two notes to make here:

1. Both the aperture stop and field stop are real physical components. While both the entrance and exit pupils are images, either real or virtual, they can appear anywhere or disappear (no image can be formed).
2. The positions and sizes of stops and pupils obtained here by using manual raytracings are not accurate but are sufficient for illustration. Any optical design software can provide much more accurate results.

Most real-image lenses use a size-adjustable iris to limit the amount of light that passes through the lens or adjust the *F*-number of the lens. This iris is placed at the aperture stop and is marked in Figs. 1.10(a–c).

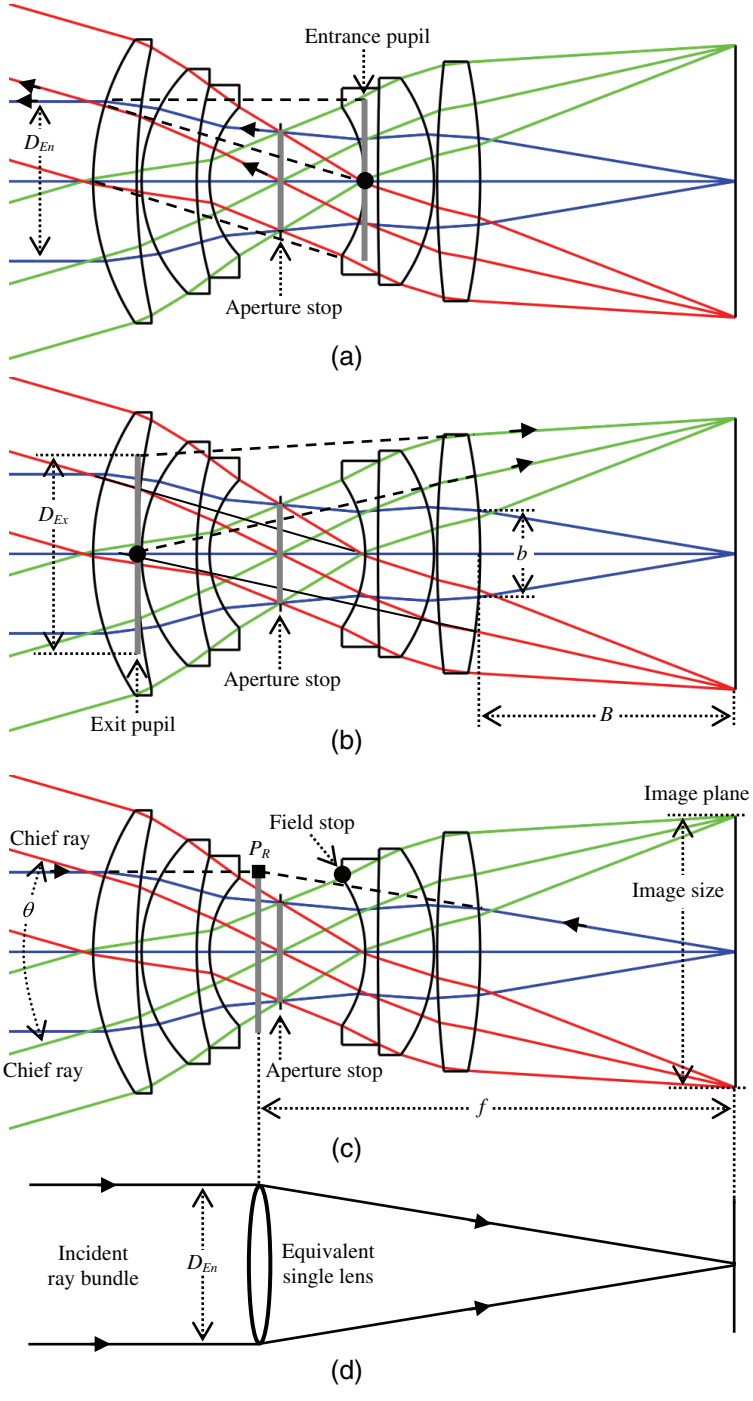

Figure 1.10 (a)–(c) Zemax-generated raytracing diagram of a double Gauss lens. (d) A single-lens model for the double Gauss lens.

In Fig. 1.10(a), where a ray is traced from the center of the aperture stop leftward, as indicated by the arrow, the ray never crosses the optical axis. However, the backward extension of the ray marked by the dashed line crosses the optical axis at the point marked by the black dot. This point is the location of the entrance pupil. If a ray were traced from the top of the aperture stop leftward, as indicated by another arrow, the ray again never crosses the optical axis. However, the backward extension of the ray marked by the dashed line passes through the plane where the entrance pupil is located and determines the entrance pupil size D_{En}.

In Fig. 1.10(b), where a ray is traced from the center of the aperture stop rightward, as indicated by the arrow, the ray never crosses the optical axis. However, the backward extension of the ray marked by the dashed line crosses the optical axis at the point marked by the black dot. This point is the location of the exit pupil. If a ray were traced from the top of the aperture stop rightward, as indicated by the arrow, the ray again never crosses the optical axis, but the backward extension of the ray marked by the dashed line passes through the plane where the exit pupil is located and determines the exit pupil size D_{Ex}. It is not rare that the exit pupil is located at the left side of the entrance pupil.

The field stop is relatively easy to find, as marked by the black dot in Fig. 1.10(c), because this point limits the largest field angle.

The right principal plane of this double Gauss lens can be found by extending forward the incident ray and extending backward the focused ray, as indicated and marked by the arrows and dashed lines, respectively, in Fig. 1.10(c). The cross-point of these two extended rays, marked by the solid black square, determines the principal-plane location and size. The distance between the principal plane and the image plane is the focal length f. (The determination of the left principal plane is omitted.)

Figures 1.10 also illustrate a few other lens parameters. The back working distance B and image space F-number B/b are marked in Fig. 1.10(b). The angle between the two chief rays that have the largest incident angles is the maximum field angle θ of this lens, as marked in Fig. 10(c); the image size and plane are also marked.

1.7.5 Single-lens model of a complex lens

This double Gauss lens and any other complex lenses can be approximately represented by a single lens model, as shown in Fig. 1.10(d). This single lens is located at the right principal plane location of the double Gauss lens, and its size is the same as the entrance pupil size of the Gauss lens. The focal length, the image plane location, and the size of this single lens are the same as those of the double Gauss lens.

The single-lens model can provide insight into complex lenses and significantly simplify the qualitative analysis.

1.8 Analytical Modeling of a Lens and Rays

1.8.1 Some comments

Manually tracing rays through a lens takes a lot of effort and is rarely performed. People either use computers and optical design software to trace rays for accurate results or solve analytical equations for approximate results. Analytical modeling is a simple and approximate tool that has a unique advantage and is still widely used today for fast estimations.

A few equations about lens focal lengths and how a lens manipulates rays have been derived for analytical modeling. This section discusses the derivation and applications of these equations. Note that it is more important to be able to skillfully use these equations than to derive them. Readers who do not have enough time and/or interest in the mathematical details can skip the derivation process without affecting their ability to work through the book.

1.8.2 Rules of sign for lens focal length and surface radius of curvature

Before proceeding, the rules of sign for the lens focal length and surface radius of curvature must be clearly stated, using the lens shown in Fig. 1.11 as an example:

1. Any positive lenses that focus rays have a positive focal length.
2. Any negative lenses that divert rays have a negative focal length.
3. When the vertex of a lens surface is to the left of the rest of the surface, the surface radius of curvature is positive. According to this rule, R_1 in Fig. 1.11 is positive.
4. When the vertex of a lens surface is to the right of the rest of the surface, the surface radius of curvature is negative. According to this rule, R_2 in Fig. 1.11 is negative.

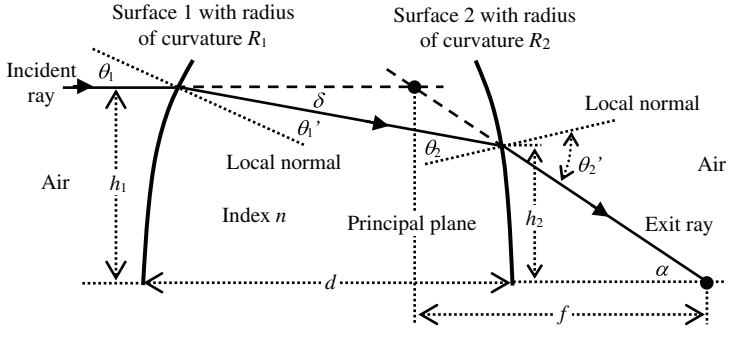

Figure 1.11 Derivation of the paraxial focal length for a thick lens.

1.8.3 Derivation of the paraxial focal length for a lens

The lens in Fig. 1.11 has a central thickness d, two surface radii of curvature R_1 and R_2, respectively, and a lens material index of n. The following process derives the paraxial focal length of this lens. Consider a ray parallel to the optical axis of the lens incident on surface 1 of the lens with ray height h_1 and incident angle θ_1, as shown in Fig. 1.11; the paraxial condition leads to

$$\theta_1 = \frac{h_1}{R_1}. \tag{1.3}$$

According to Snell's law, with paraxial approximation and Eq. (1.3), the ray refracted by surface 1 has an angle θ_1', given by

$$\begin{aligned} \theta_1' &= \frac{\theta_1}{n} \\ &= \frac{h_1}{R_1 n}. \end{aligned} \tag{1.4}$$

The difference between θ_1 and θ_1' is given by Eqs. (1.3) and (1.4):

$$\begin{aligned} \delta &= \theta_1 - \theta_1' \\ &= \frac{h_1}{R_1}\left(1 - \frac{1}{n}\right). \end{aligned} \tag{1.5}$$

The height of the refracted ray hitting surface 2 is found by using Eq. (1.5) and the paraxial condition

$$\begin{aligned} h_2 &= h_1 - d\delta \\ &= h_1\left[1 - \frac{d}{R_1}\left(1 - \frac{1}{n}\right)\right]. \end{aligned} \tag{1.6}$$

The incident angle of this ray on surface 2 is

$$\begin{aligned} \theta_2 &= \frac{h_2}{-R_2} + \delta \\ &= h_1\left[\frac{1}{-R_2} - \frac{d}{R_1(-R_2)}\left(1 - \frac{1}{n}\right) + \frac{1}{R_1}\left(1 - \frac{1}{n}\right)\right], \end{aligned} \tag{1.7}$$

where the negative sign of R_2 is from the rule of sign stated earlier. According to Snell's law, the ray refracted by surface 2 has an angle θ_2', given by

$$\theta_2' = n\theta_2. \tag{1.8}$$

This ray crosses the optical axis with angle α, which is given with paraxial approximation by

$$\begin{aligned}\alpha &= \theta_2' - \frac{h_2}{-R_2} \\ &= n\theta_2 + \frac{h_2}{R_2} \\ &= nh_1\left[\frac{1}{R_2} + \frac{d}{R_1R_2}\left(1-\frac{1}{n}\right) + \frac{1}{R_1}\left(1-\frac{1}{n}\right)\right] + \frac{h_1}{R_2}\left[1 - \frac{d}{R_1}\left(1-\frac{1}{n}\right)\right] \\ &= h_1(n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)d}{nR_1R_2}\right].\end{aligned} \tag{1.9}$$

The last step to derive Eq. (1.9) requires some rearranging and cleaning up of the equation. The paraxial focal length f is related to h_1 and α by the following equation, as shown in Fig. 1.11:

$$\begin{aligned}\frac{1}{f} &= \frac{\alpha}{h_1} \\ &= (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)d}{nR_1R_2}\right].\end{aligned} \tag{1.10}$$

Equation (1.10) is the final result: the widely used paraxial focal length of a thick lens.

For any lens, once the two surface radii of curvature R_1 and R_2, the central thickness d, and the material index n are known, the paraxial focal length f of the lens can be easily calculated using Eq. (1.10).

When $d \ll (R_1R_2)^{0.5}$, Eq. (1.10) reduces to the widely used thin lens form

$$\frac{1}{f} \approx (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{1.11}$$

1.8.4 Derivation of the thin lens equation for a thin lens

A thin lens is a simplified concept. A thin lens has properties such that its central thickness is much less than its focal length and the two principal planes of the lens can be considered to coincide. The thin lens equation is an analytical tool to analyze the effects of a lens on rays that travel through the lens. The derivation process of the thin lens equation for one lens is explained in this subsection.

Consider a positive lens with focal length f that focuses rays from an object with height h and a distance o from the lens, as shown in Fig. 1.12.

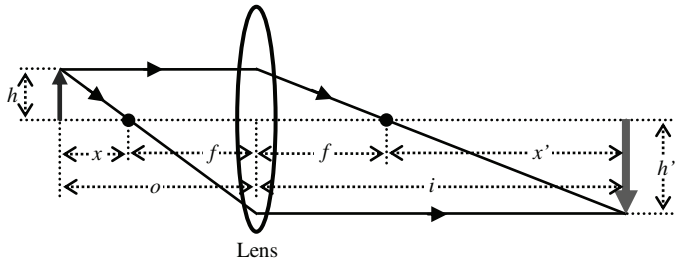

Figure 1.12 Derivation of the thin lens equation. *f* is the focal length of the lens, and *o* and *i* are the object and image distances, respectively. The lens position is its center position.

An image with height h' is formed by the lens at a distance i from the lens. Equations (1.12) and (1.13) can be written using some triangular geometrics:

$$\frac{h}{x} = \frac{h'}{f}, \tag{1.12}$$

$$\frac{h}{f} = \frac{h'}{x'}. \tag{1.13}$$

Equation (1.14) can be found by combining Eqs. (1.12) and (1.13) to eliminate h and h':

$$xx' = f^2. \tag{1.14}$$

Equation (1.14) is one form of the thin lens equation, but a more popular form can be found by inserting $x = o - f$ and $x' = i - f$ into Eq. (1.14); the result after some rearrangements is

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}, \tag{1.15}$$

which is the widely used thin lens equation.

The object is conventionally placed to the left of the lens, and the rays travel rightward from the object through the lens to form an image at the right of the lens a distance i away from the lens. In such a case, $o > 0$ by definition. By definition, $i > 0$ means the image is on the right of the lens, and $i < 0$ means a virtual image is formed to the left of the lens. f is the focal length of the lens; $f > 0$ and $f < 0$ mean a positive lens and a negative lens, respectively. The thin lens equation can be used to conveniently analyze the location and size of an image formed by a lens. The results are accurate enough for many applications.

1.8.5 Derivation of the focal length of two thin lenses

When there are two thin lenses with focal length f_1 and f_2, respectively, and a small distance $d \ll f_1$ and $d \ll f_2$ between them, the focal length of the two lenses combined can be derived by the following process.

Consider a ray emitted by an on-axis object point located a distance o_1 from Lens 1, as shown in Fig. 1.13. If there is no Lens 2, this ray will be focused by Lens 1 at a point marked by the grey dot with a distance of i_1 from Lens 1. For Lens 1, the thin lens equation [Eq. (1.15)] takes the form

$$\frac{1}{o_1} + \frac{1}{i_1} = \frac{1}{f_1}$$

or

$$i_1 = \frac{f_1 o_1}{o_1 - f_1}. \tag{1.16}$$

For Lens 2, the grey point is a virtual object point with a distance

$$o_2 = i_1 - d \tag{1.17}$$

from Lens 2. The thin lens equation for Lens 2 takes the form

$$\frac{1}{-o_2} + \frac{1}{i_2} = \frac{1}{f_2}, \tag{1.18}$$

where the negative sign of o_2 comes from the fact that the virtual object point is located to the right of Lens 2. Equations (1.17) and (1.18) are combined to eliminate o_2 and solve for i_1:

$$i_1 = \frac{i_2 f_2}{f_2 - i_2} + d. \tag{1.19}$$

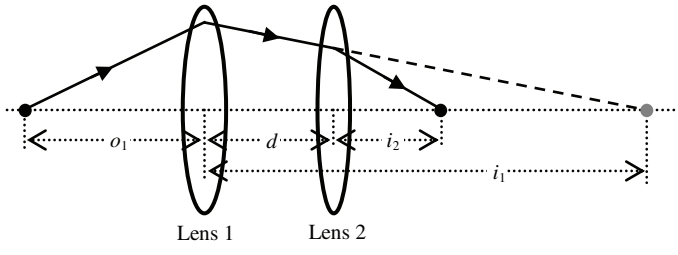

Figure 1.13 Derivation of the focal length of two thin lenses with distance *d* between them.

Equations (1.16) and (1.19) are combined to illuminate i_1:

$$\frac{i_2 f_2}{f_2 - i_2} + d = \frac{o_1 f_1}{o_1 - f_1}. \tag{1.20}$$

In the case of $o_1 \to \infty$, i_2 becomes the right focal length f_r of the two lenses. With some rearrangements, Eq. (1.20) reduces to

$$\frac{1}{f_r} = \frac{f_1 + f_2 - d}{f_2(f_1 - d)}. \tag{1.21}$$

For $d \to 0$, Eq. (1.21) reduces to the well-known form

$$\frac{1}{f_r} = \frac{1}{f_1} + \frac{1}{f_2}. \tag{1.22}$$

In the case of $i_2 \to \infty$, o_1 becomes the left focal length f_l of the two lenses. With some rearrangements, Eq. (1.20) reduces to

$$\frac{1}{f_l} = \frac{f_1 + f_2 - d}{f_1(f_2 - d)}. \tag{1.23}$$

For $d \to 0$, Eq. (1.23) also reduces to the well-known Eq. (1.22).

1.8.6 Application examples of the thin lens equation for a positive lens

Five cases about a positive lens are discussed in this section with the help of Fig. 1.14, where the focal point is marked by F, and the focal length f is not marked. In all of these cases, the object is on the lens optical axis, the rays are symmetric about the optical axis, and the image point is on the optical axis. These case studies are particularly useful for analyzing a laser beam propagating through a lens.

- Case 1: A convergent ray bundle is incident on the lens, as shown in Fig. 1.14(a). No point light source can emit such a ray bundle, and therefore there is no object point. However, the forward extensions of the incident rays cross the optical axis of the lens at a point marked by the open square. This point is a virtual object point. In such a case, $o < 0$. The focused spot of the rays is the image point. From Eq. (1.15) it can be found that $i < f$, and the rays will be focused at a point inside the focal length, as marked by the solid square in Fig. 1.14(a).
- Case 2: $o \to \infty$, the rays are from infinity and parallel to the lens optical axis when they reach the lens, as shown in Fig. 1.14(b). Equation (1.15) shows that $i = f$. The rays are focused at the right focal point of the lens.

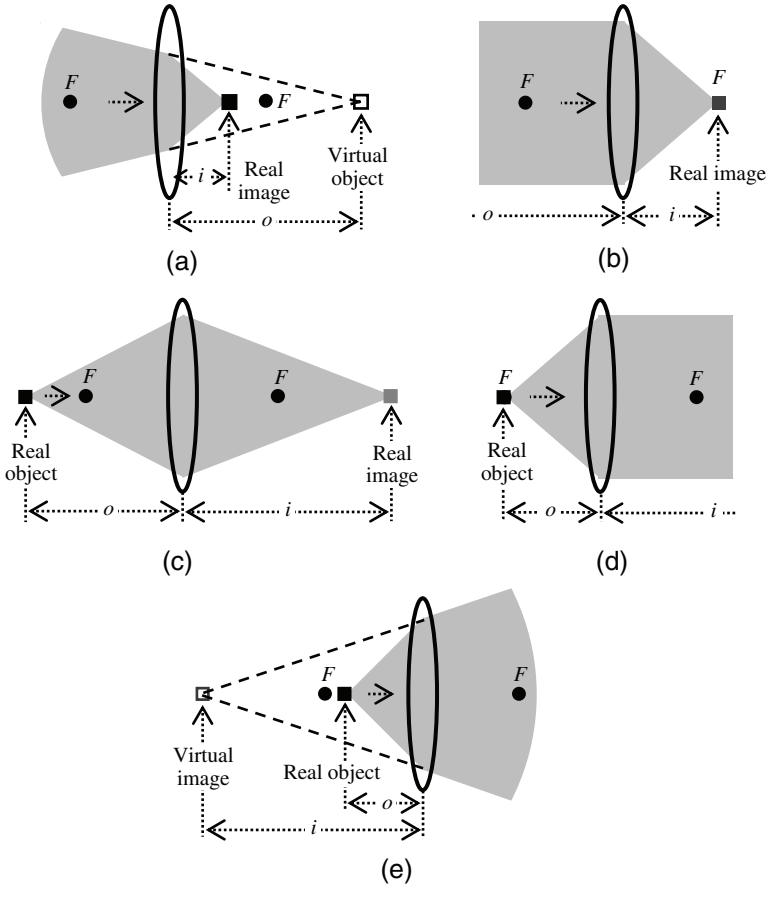

Figure 1.14 Illustration of the thin lens equation [Eq. (1.15)] for a positive lens with focal points marked by F and focal length $f > 0$. (a) When the ray bundle is convergently incident on the lens, the forward extension of the incident rays crosses the optical axis at a point marked by the open square. (b) When the object point is infinitely away, the rays are parallel to the optical axis as they reach the lens and will be focused by the lens on its focal point. (c) When the object point is outside the focal length, the rays will be focused by the lens at a point beyond the focal point. (d) When the object point is at a focal point, the rays will be collimated by the lens, which means the image is at infinity. (e) When the object point is inside the focal length, the rays passing through the lens will still be divergent. A virtual image marked by the grey open square can be found by extending the rays backward.

- Case 3: A real object is marked by the solid square shown in Fig. 1.14(c). The object distance is $o > f$. According to Eq. (1.15), $i > f$. The rays are focused at a point beyond the right focal point, as marked by the grey solid square.
- Case 4: A real object marked by the solid square is at the focal point, that is $o = f$, and $i \to \infty$, according to Eq. (1.15). The situation is shown in Fig. 1.14(d). The image appears at infinity, which means rays are collimated.

- Case 5: The real object marked by the solid square is inside the focal length, as shown in Fig. 1.14(e), i.e., $o < f$ and $i < 0$, according to Eq. (1.15). The rays passing through the lens are divergent. No real image is formed. But the backward extensions of the rays cross the optical axis at a point marked by the grey open square; this point is the virtual image of the object. Since the virtual image is to the left of the lens, $i < 0$.

1.8.7 Application examples of the thin lens equation for a negative lens

Four cases about a negative lens are discussed in this subsection with the help of Fig. 1.15, where the focal point is marked by F, and the focal length f is not marked. In all these cases, the object is on the lens optical axis, the rays are symmetric about the optical axis, and the image point is on the optical axis. Note that for a negative lens, the focal length f in Eq. (1.15) is negative.

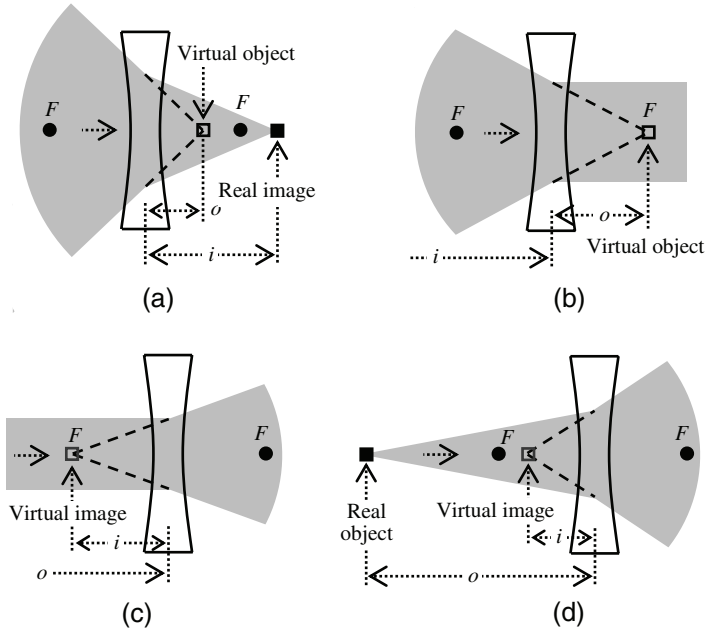

Figure 1.15 Illustration of thin lens equation [Eq. (1.15)] for a negative lens with focal points marked by F and focal length $f < 0$. (a) When a convergent ray bundle is incident on the lens, the forward extensions of the incident rays cross the optical axis at a point marked by the open square, which is a virtual object point. This virtual object point is inside the focal length, rays are focused. (b) Similar to the situation shown in (a), but the virtual object is at the focal point, and the rays are collimated. (c) When the object point is infinitely away, the rays passing through the lens are divergent. The backward extensions of the rays will cross the optical axis on the focal point and form the virtual image. (d) When the object point is not infinitely away, the extension of the exit rays will cross the optical axis at a point inside the focal length. This point is the virtual image.

- Case 1: A convergent ray bundle is incident on the lens, as shown in Fig. 1.15(a). No point light source can emit such a ray bundle, and therefore there is no object point. However, the forward extensions of the incident rays cross the optical axis of the lens at a point marked by the open square. This point is a virtual object point. In such a case, $o < 0$. For $o < 0$ and $f < 0$, i can be either positive or negative, based on Eq. (1.15). In Fig. 1.15(a), the virtual object point is inside the focal length; it can be found from Eq. (1.15) that $i > 0$. The rays will be focused at a point to the right of the lens marked by a solid square.
- Case 2: Similar to Case 1, but the virtual object point is at the focal point, as shown in Fig. 1.15(b). Equation (1.15) leads to $o = f$ and $i \to \infty$. The rays are collimated by the lens.
- Case 3: The object is at infinity, i.e., $o \to \infty$. Then, $i = f < 0$ according to Eq. (1.15). $i < 0$ means that there is a virtual image at the left focal point of the lens, as shown by the backward-extended rays and the grey open square in Fig. 1.15(c).
- Case 4: The object is at finite $o > 0$. Then, $0 > i > f$, according to Eq. (1.15). $i < 0$ means that there is a virtual image to the left of the lens, and $i > f$ means that $|i| < |f|$ since f is negative; the virtual-image point position is inside the focal plane, as shown by the backward-extended rays and the grey open square in Fig. 1.15(d).

1.9 Lateral and Axial Magnifications of Lenses

So far, the object being considered is a point. For a real object of a certain size, the image formed by a lens often has a size and orientation different than those of the object because the lens has a lateral magnification M_L and axial magnification M_A. Note that all of the lens imaging drawings in this section neglect spherical aberration to avoid unnecessary complexity.

1.9.1 Definition of lateral magnifications and axial magnifications

In Fig. 1.12, the lateral magnification M_L is defined as the ratio of the image height and object height. Based on the left side of Fig. 1.12 and some triangular geometry, M_L can be found to be

$$\begin{aligned} M_L &= -\frac{h'}{h} \\ &= \frac{f}{f-o}. \end{aligned} \tag{1.24}$$

Equations (1.15) and (1.24) are combined to eliminate f, and M_L is found to be

$$M_L = -\frac{i}{o}. \tag{1.25}$$

There are two notes to make about M_L:

1. A negative M_L means that the image has an orientation opposite to the object, which is the case shown in Fig. 1.12. Since h and h' in Fig. 1.2 are defined as positive values, a negative sign must be placed in front of M_L in Eq. (1.24).
2. Equation (1.24) states that for a given lens, f is fixed, and changing o can change the value of M_L.

Lateral magnification is also called transverse magnification. Axial magnification is defined by $M_A = \Delta i/\Delta o$, which can be found by differentiating Eq. (1.15). The result is

$$\begin{aligned} M_A &= \frac{\Delta i}{\Delta o} \\ &= -\frac{i^2}{o^2} \\ &= -{M_L}^2. \end{aligned} \tag{1.26}$$

Since the value of ${M_L}^2$ is always positive, M_A is always negative, which means that when an object is axially extended towards the lens, the corresponding image is extended away from the lens, as shown in Fig. 1.16. Axial magnification is also called longitudinal magnification.

1.9.2 Schematic examples of an image formed by a positive lens

Figure 1.16 is basically a re-plot of Fig. 1.7 but with the intention to illustrate the lateral and axial magnifications. The object is placed outside the focal length of this positive lens. The image found by some raytracings has an orientation opposite that of the object and has a height approximately half the object height. Thus, $M_L = -h'/h \approx -0.5$, and $M_A = -{M_L}^2 \approx -0.25$. Figure 1.16

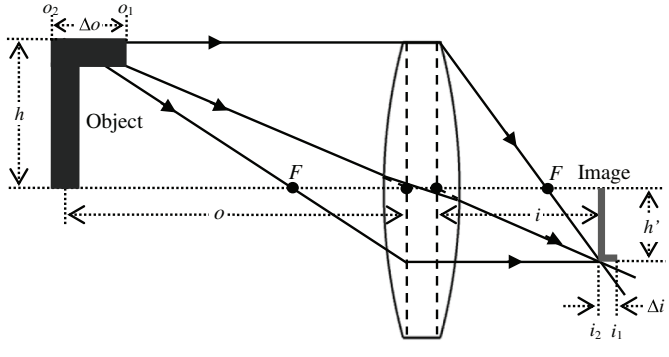

Figure 1.16 A positive lens forms a real and negative image of an object placed outside the focal length. Here, $M_L = h'/h \approx -0.5$ and $M_A = -{M_L}^2 \approx -0.25$. The image is thinner than the object.

shows that $o_2 > o_1$, while $i_2 < i_1$, which explains the negative value of M_A. The image is drawn approximately to the right proportion and is thinner than the object because M_A and M_L have different values.

Figure 1.17 shows another example of a positive lens forming an image. The object is placed inside the focal length of the lens. The image formed has the same orientation as the object and has a height about three times larger than the object height. This h' in Fig. 1.17 is defined as negative (opposite to h' in Fig. 1.16), so $M_L = -h'/h \approx 3$, and $M_A = -M_L{}^2 \approx -9$. Again, the image is much thicker than the object because M_A and M_L have different values.

1.9.3 Schematic examples of an image formed by a negative lens

Figure 1.18 shows an example of a negative lens forming an image. The object is placed outside the focal length of the lens. The image formed has the same orientation as the object and has a height approximately one-third of the object height: $M_L = -h'/h \approx 1/3$, and $M_A = -M_L{}^2 \approx -1/9$. The image is much thinner than the object because M_A and M_L have different values.

Figure 1.19 shows another example of a negative lens forming an image. The object is placed right at the focal point of the lens. The image formed has the same orientation as the object and has a height approximately half the object height. Thus, $M_L = -h'/h \approx 1/2$, and $M_A = -M_L{}^2 \approx -1/4$. The image is thinner than the object because M_A and M_L have different values.

There are two unique phenomena about negative lenses that have a negative focal length:

1. $f < 0$ and $f - o < 0$ (since the object is at the left of the lens $o > 0$) produce an M_L that is always positive, according to Eq. (1.24) and shown in Figs. (1.18) and (1.19).

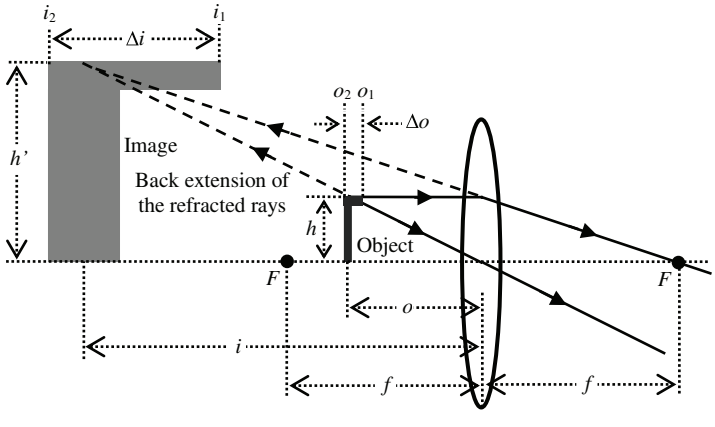

Figure 1.17 A positive lens forms a virtual and positive image of an object placed inside the focal length. Here, $M_L = -h'/h \approx 3$ and $M_A = -M_L{}^2 \approx -9$. The image is much thicker than the object.

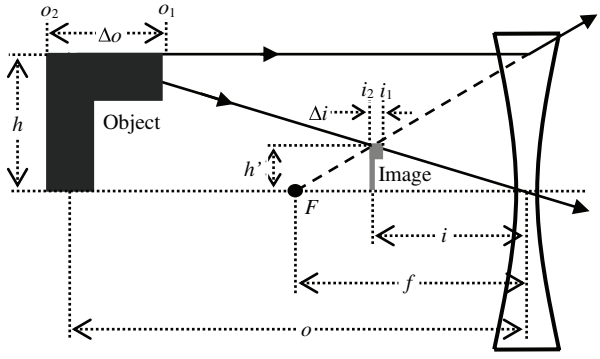

Figure 1.18 A negative lens forms a virtual and positive image of an object placed outside the focal length. In this setup, $M_L = -h'/h \approx 1/3$, and $M_A = -M_L^2 \approx -1/9$. The image is much thinner than the object.

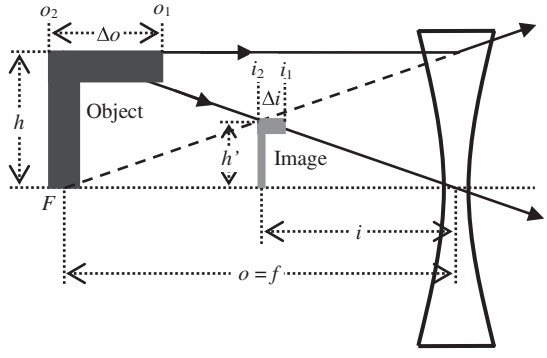

Figure 1.19 A negative lens forms a virtual and positive image of an object placed at the focal point. Here, $M_L = -h'/h \approx 1/2$, and $M_A = -M_L^2 \approx -1/4$. The image is thinner than the object.

2. $i = fo/(o - f)$ is always negative, and the images formed by a negative lens are always virtual images at the left of the lens, as shown in Figs. (1.18) and (1.19).

1.9.4 Distortion of a 3D image formed by a perfect lens

Figures (1.16)–(1.19) are drawn with the assumption that the lateral and axial magnifications are constant for a given lens and object. This is true only when the axial size of the object $\Delta o = o_2 - o_1$ meets the condition $\Delta o \ll f$, so that o and i can be treated as constant numbers, and the resulting lateral and axial magnifications are constants too. However, in some cases, $\Delta o \sim f$, and the different axial portion of the object has different o values and different magnifications, both lateral and axial.

Figure 1.20 illustrates such a case. The front position of the object is at $o_2 = 2f$. According to Eqs. (1.15), (1.25), and (1.26), $i_2 = 2f$, $M_L = -1$, and $M_A = -1$, respectively. The image of the front part of the object is just an inverse of the object, nothing else. The back position of the object is at

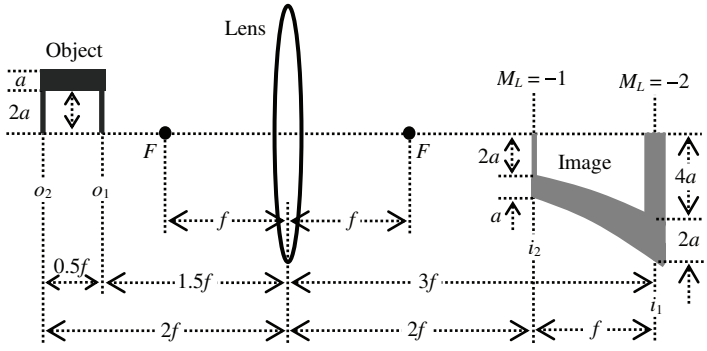

Figure 1.20 A detailed analysis and illustration of image distortion caused by lateral and axial magnifications varying as a function of the object position, even though the lens is perfect.

$o_1 = 1.5f$, which leads to $i_1 = 3f$, $M_L = -2$, and $M_L = -4$. The back part of the object has much larger magnification than the front part. The image formed in such a case will have an approximate size and shape; as shown in Fig. 1.20, the distortion is severe.

The images shown in Figs. (1.16)–(1.19) should contain severe distortions too, but these distortions are neglected for the purpose of simplification. The distortion discussed here is not caused by defects or aberrations of the lens. Even a perfect lens has such a distortion.

1.10 Mirrors

Mirrors are probably the second-most-used optical components next to lenses. Mirrors are similar and different to lenses in several aspects.

1.10.1 Reflection law

When a ray is incident on a mirror, most of the energy of the ray is reflected. The reflection law states that

1. The incident ray, the reflected ray, and the normal of the reflection surface at the point of incidence lie in the same plane.
2. The angle that the incident ray makes with the normal equals the angle that the reflected ray makes to the same normal.
3. The reflected ray and the incident ray are on the opposite sides of the normal.

The reflection that occurs at an interface of two optical media obeys the same reflection law so long as the incident angle does not exceed the total reflection angle (see Section 1.3).

Figure 1.21 illustrates the reflection law using three differently shaped interfaces: planar, convex, and concave. The optical axis of an interface is its

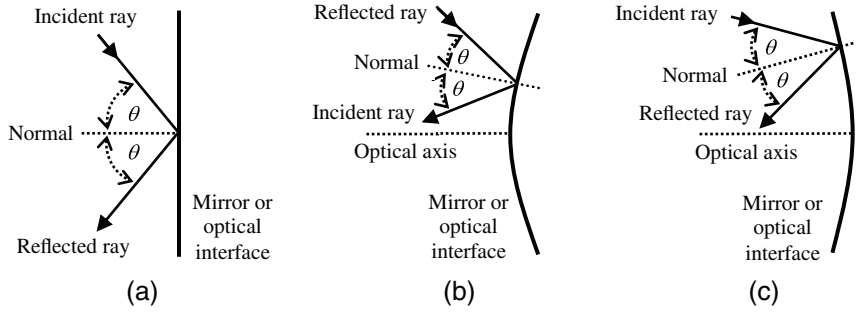

Figure 1.21 Reflection law using three differently shaped mirrors or optical interfaces: (a) planar, (b) convex, and (c) concave.

axis of symmetry. For a planar interface, the normal of the point at which the ray hits is the optical axis.

1.10.2 Mirror equation: lateral and axial magnifications

The thin lens equation [Eq. (1.15)] is also effective for mirrors and is rewritten here for convenience:

$$\frac{1}{o}+\frac{1}{i}=\frac{1}{f}, \tag{1.27}$$

where o is the object distance, i is the image distance, and f is the mirror focal length.

Only the rules of sign for mirrors are different from those for lenses:

1. A concave mirror is equivalent to a positive lens and has a positive focal length.
2. A convex mirror is equivalent to a negative lens and has a negative focal length.
3. The object is conventionally assumed at the left of the mirror with a positive object distance o.
4. The image can be either to the left or right side of the mirror with a positive or negative image distance i, respectively.
5. For spherical mirrors, their focal length always equals half their surface radius of curvature.

The lateral and axial magnifications for a mirror can be defined similar to those for a lens, only with different signs:

$$\begin{aligned} M_L &= -\frac{i}{o} \\ &= \frac{f}{f-o}, \end{aligned} \tag{1.28}$$

$$M_A = \frac{i^2}{o^2} = {M_L}^2. \tag{1.29}$$

M_A is always positive, which is an indication of the horizontal orientation of the image relative to the object, as will be shown in Fig. 1.22. It can be proven that for convex mirrors the image is always virtual.

1.11 Mirror Imaging

By tracing rays according to the reflection law and with Eqs. (1.27)–(1.29) in mind, the image of a mirror can be analyzed. Various rays can be traced from an object in front of a mirror to the mirror. Only two rays from the top of the object are usually traced. One ray is parallel to the optical axis of the mirror, and the other ray hits the cross-point of the mirror and the optical axis. The two rays are reflected by the mirror. Any other two rays can be selected to trace so long as the raytracing is convenient.

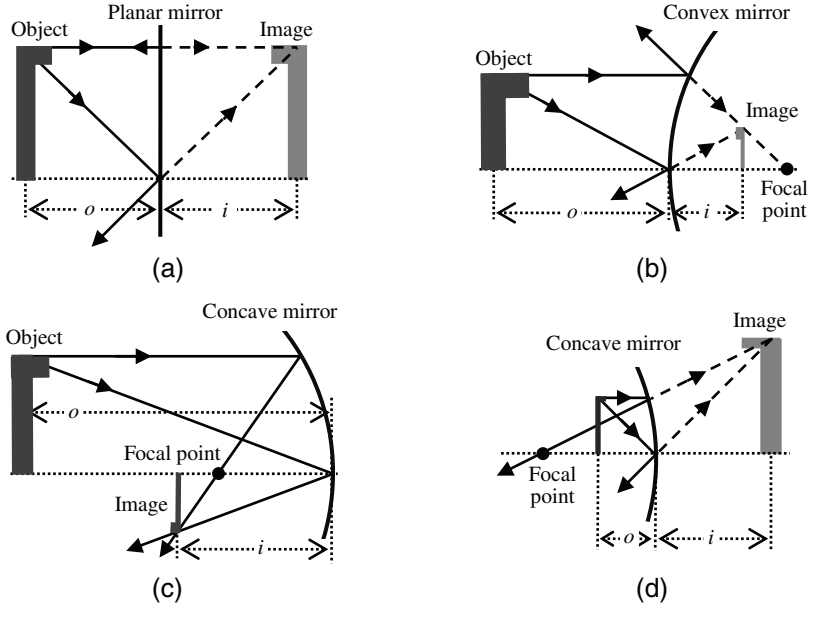

Figure 1.22 The raytracing technique is used to find the image of an object formed by three differently shaped mirrors. The solid lines represent rays. The dashed lines represent the backward extension of the rays. The spherical aberration is neglected. (a) Planar mirror. The image formed is positive, virtual, and the same size as the object. (b) Convex mirror. The image formed is positive, virtual, and smaller than the object. (c) Concave mirror with the object being placed outside the focal length. The image formed is negative, real, and smaller than the object. (d) Concave mirror with the object being placed inside the focal length. The image formed is positive, virtual, and larger than the object.

1.11.1 Planar mirror imaging

Figure 1.22 shows four examples. Figure 1.22(a) shows the raytracing for a planar mirror. Two rays are traced from the top of the object towards the mirror. The two reflected rays do not cross, but their back extensions do. The cross-point is the image of the top point of the object. Using the same raytracing technique, the image of any point on the object can be found, and thereby the whole image is found.

The image of a mirror can also be analyzed using Eqs. (1.27)–(1.29). For a planar mirror $f \to \infty$, Eqs. (1.27)–(1.29) lead to $i = -o < 0$, $M_L = 1$, and $M_A = 1$. The image is a virtual positive image at the right of the mirror. The horizontal orientation of the image shown here is $M_A > 0$.

1.11.2 Convex mirror imaging

Figure 1.22(b) shows the image formed by a convex lens using a raytracing technique with $o = -1.5f > 0$ (since for a convex lens $f < 0$ and o is always positive). Two rays are drawn from the top of the object toward the mirror. The two reflected rays do not cross. The backward extension of any ray parallel to the optical axis should pass through the focal point of the mirror if the spherical aberration of the mirror is neglected. The two backward extensions of these two rays cross and determine the position of the image of the object top. The whole image can be found using this raytracing technique and is positive, virtual, and smaller than the object.

The image can also be found analytically. Equations (1.27)–(1.29) lead to $i = 0.6f < 0$, the image is at the right side of the mirror, $M_L = -0.6f/(-1.5f) = 0.4f > 0$. The image is positive and 0.4 times the height of the object. $M_A = 0.16$, i.e., the image is much thinner than the object.

The image shown in Fig. 1.22(b) is approximately the right proportion, with M_L and M_A assumed to be constant to simplify the analysis. Both M_L and M_A are a function of o and can be treated as constant only when $\Delta o \ll f$. This situation is the same as the M_L and M_A for lenses.

1.11.3 Concave mirror imaging

Figure 1.22(c) shows the imaging of a concave mirror. The object is placed outside the focal length of the mirror with $o = 3f$. The raytracing is similar to those shown in Figs. 1.22(a) and (b), but the two reflected rays cross and determine that the image is real and negative.

Analytically, Eqs. (1.27)–(1.29) lead to $i = 1.5f$, $M_L = -0.5$, and $M_A = 0.25$. The image is thinner than the object. The drawing in Fig. 1.22(c) is approximately the right proportion.

Figure 1.22(d) shows the imaging of the concave mirror same as the mirror shown in Fig. 1.22(c). The difference is now the object is placed inside the focal length of the mirror with $o = 0.5f$. The two reflected rays do not

cross, but their backward extensions cross and determine that the image is virtual and positive. From Eqs. (1.27)–(1.29), it can be found that $i = -f$, $M_L = 2$, and $M_A = 4$. The image is fatter than the object. The drawing in Fig. 1.22(d) is also approximately the right proportion.

1.12 Optical Aberrations

In most cases, an imaging lens cannot form a near-perfect image because of the presence of various aberrations in the lens. Aberrations can be significantly reduced by carefully designing the lens to use more-appropriate optical elements. However, more elements means higher cost, larger size, and heavier weight. In reality, the performance, cost, etc., of a lens must be balanced. In this section, several types of the most frequently seen optical aberrations are discussed.

Optical aberration is a complex subject. Optical books traditionally study aberrations in detail with a lot of mathematics, e.g., Born and Wolf,[7] which is necessary to manually locate the root of certain aberrations and reduce it. But with the help of optical design software, such exhaustive study is no longer necessary, particularly for those engineers and scientists who have limited time to study optical engineering. Therefore, this book includes only one section to briefly describe optical aberrations.

1.12.1 Spherical aberration

Three raytracing diagrams simulated by optical design software Zemax are plotted in Fig. 1.23(a)–(c), where the object point is on the optical axis and is infinitely away from a positive lens. The rays are parallel to the optical axis when they reach the lens and are focused by the lens to cross the optical axis. These rays are traced by Zemax with high accuracy, not just schematic drawings. All of the rays in Fig. 1.23 have a 0.55-μm wavelength.

The lens shown in Fig. 1.23(a) is an equi-convex lens made of Ohara S-FPL53 glass with an index of 1.44. Rays with different heights are focused at different locations, and a sharp and clean focused spot does not exist. The focused rays are said to have severe spherical aberration. The term "focal length" usually means the focal length for paraxial rays (rays with near-zero height), as shown in Fig. 1.23(a).

The magnitude of spherical aberration decreases as the lens surface radii of curvature increase. If Ohara S-LAH79 glass with a large refractive index of 2.00 is used to make a lens of the same size and focal length, the surface radii are longer, according to Eq. (1.10) and shown in Fig. 1.23(b), and the spherical aberration is less severe than the spherical aberration shown in Fig. 1.23(a).

If the small-index Ohara S-FPL53 glass is still used to make a lens of the same size but with a longer focal length of 50 mm, as shown in Fig. 1.23(c), the spherical aberration is barely noticeable within the resolution of the diagram since the lens surfaces are even more flat.

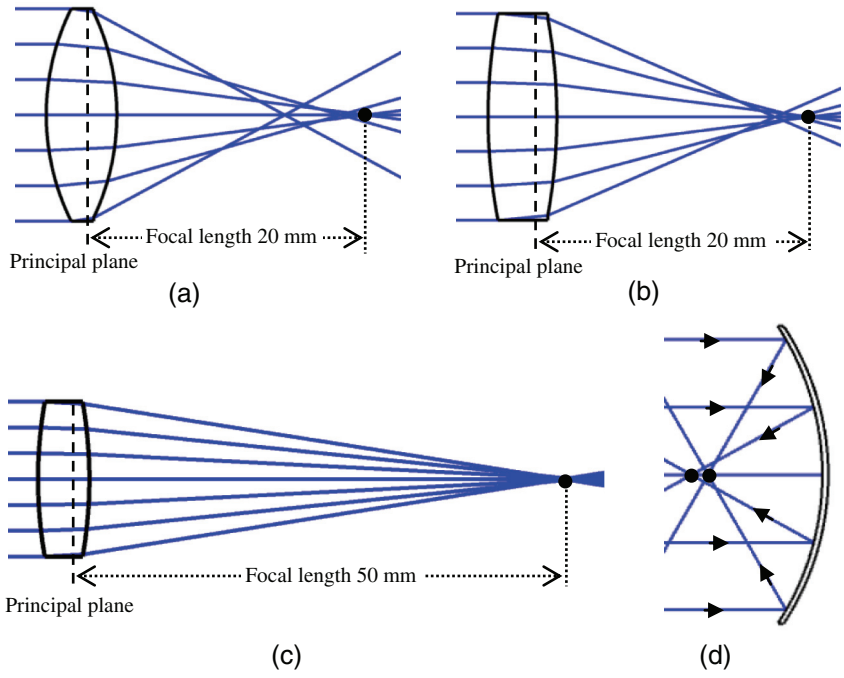

Figure 1.23 Zemax-generated raytracing diagrams to illustrate spherical aberration for (a)–(c) three lenses and (d) one mirror. The object point is on the optical axis and is infinitely distant. The rays are parallel when they reach the lens or mirror. (a) The lens has a 15-mm diameter, a 5-mm central thickness, a 20-mm focal length, and is made of Ohara S-FPL53 glass with a relatively small refractive index of 1.44. The spherical aberration is very severe. Higher rays are focused at a point closer to the lens. (b) The lens has the same size and focal length as those of the lens in (a) but is made of Ohara S-LAH79 glass with a relatively large refractive index of 2.00. The lens surfaces are flatter, and the spherical aberration is less severe. (c) The lens has a size and material the same as those of the lens in (a) except the focal length is 50 mm. The lens surfaces are flatter, and the spherical aberration is not apparent. (d) Spherical aberration also exits in the rays reflected by a mirror. The higher rays are focused at a point closer to the mirror.

The following conclusions can be drawn:

1. For a given lens size, a longer focal length will result in a smaller spherical aberration.
2. For a given lens size and focal length, a larger-refractive-index glass will result in a smaller spherical aberration. However, larger-index glasses usually have a smaller Abbe number and will lead to larger color aberration (discussed in Section 1.12.4).
3. For a given glass type and focal length, a smaller lens size will result in a smaller spherical aberration because only those rays with a smaller height can pass through the smaller lens.

Mirror reflection also has spherical aberrations; Fig. 1.23(d) shows an example. The rays are traced accurately by Zemax. Higher rays are focused by the concave spherical mirror at a point closer to the mirror, similar to lens spherical aberration. Spherical aberration is the primary aberration that degrades the image quality and must be reduced during the optical design process.

1.12.2 Coma

The spherical aberration discussed previously applies to incident rays that are parallel to the optical axis. The focused spot is on the axis. When the ray bundle incident on the lens has an angle to the optical axis, the focused spot is not on the axis, and the spherical aberration becomes "coma." Figure 1.24(a) shows the same lens as that in Fig. 1.23(a), but the ray incident angle is 20° rather than 0°. The rays are poorly focused. The focused spot has a coma-like intensity pattern. Again, only one wavelength of 0.55 μm is used to avoid color aberration.

Because coma is a type of spherical aberration, it exits the images formed by a mirror as well. Figure 1.24(b) shows a mirror the same as that in Fig. 1.23(d), but the ray incident angle is 10° instead of 0°. The focused spot has a coma-like intensity pattern too.

1.12.3 Astigmatism

The raytracing diagram shown in Fig. 1.24 is not symmetric about the optical axis. Only the rays in the plane of the page or the tangential plane are plotted.

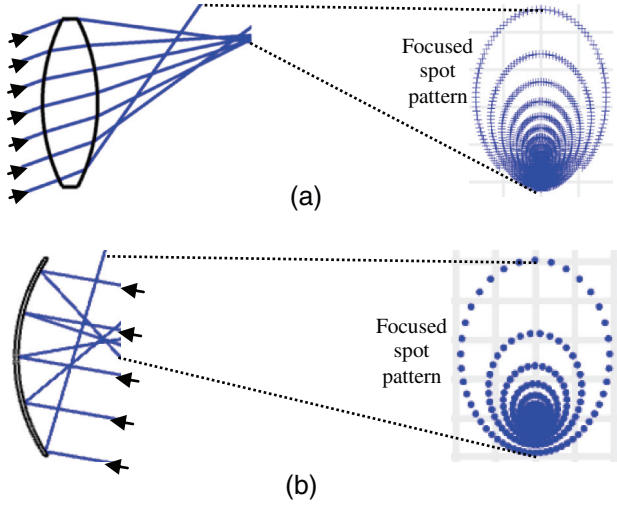

Figure 1.24 (a) The lens is the same as that in Fig. 1.23(a), but the incident ray bundle has an angle of 20° to the optical axis. (b) The mirror is the same as that in Fig. 1.23(d), but the ray bundle has an angle of 10° to the optical axis.

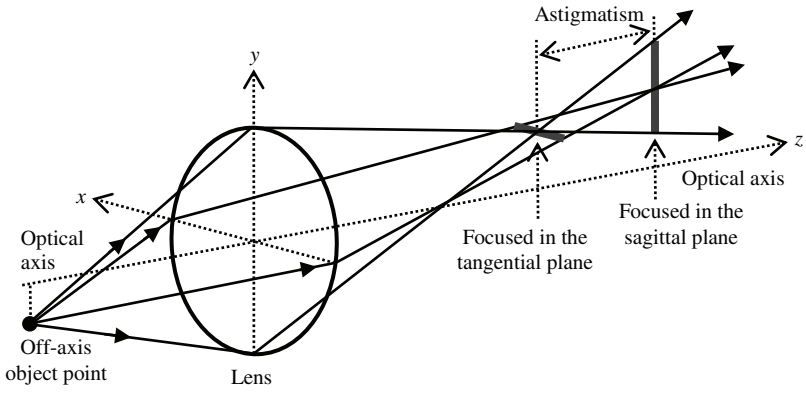

Figure 1.25 Illustration of astigmatism. When the light source is off axis in the *y-z* plane, the focusing points in the *x-z* and *y-z* planes are not at the same location. The axial distance between the two focusing points is the magnitude of astigmatism.

The behavior of the rays in the plane perpendicular to the page plane, particularly in the sagittal plane, is unknown. Figure 1.25 illustrates the situation with a 3D sketch.

The optical axis of the lens is the *z* axis, and the object point is at the *y-z* plane, which is the tangential plane. The *x-z* plane is the sagittal plane. As shown in Fig. 1.25, the focused spots in the tangential and sagittal planes are not at the same point. This phenomenon is common in optics and is called "astigmatism." In most cases, the focusing point in the tangential plane is closer to the lens. The axial distance between the two focusing points is the magnitude of the astigmatism. Just like spherical aberration, astigmatism will degrade the image quality and must be reduced during the optical design process.

1.12.4 Color aberrations

The refractive indices of optical materials are not constants; they vary as the wavelength varies, and a longer wavelength usually produces a smaller index: $dn/d\lambda < 0$, where n is the refractive index, and λ is the wavelength. This phenomenon is called "color dispersion." The magnitude of color dispersion is related to a parameter called the "Abbe number" (see Section 4.1.2 for details). A larger Abbe number means weaker color dispersion, which is usually desired.

Equation (1.10) shows that the focal length of a lens is proportional to the inverse of the refractive index of the lens material. So, the focal length of a lens is not a constant. A longer wavelength leads to a smaller index and longer focal length, and the results are called "color aberrations." There are two types of color aberration: longitudinal color and lateral color.

The lens used in Fig. 1.23(a) is re-plotted in Fig. 1.26(a) as an example. This lens uses Ohara S-FPL53 glass with a small refractive index of 1.44 and a large Abbe number of 94.9. This lens has a 15-mm diameter, a 5-mm central

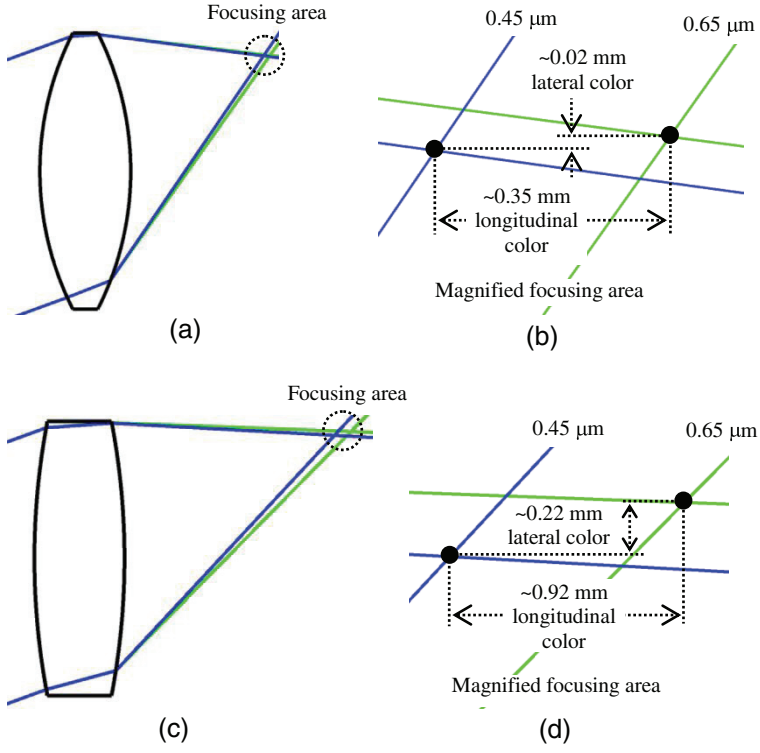

Figure 1.26 Two Zemax-generated raytracing diagrams illustrate color aberrations. Two wavelengths of 0.45 μm and 0.65 μm are used to plot. (a) A 20-mm-focal-length lens made of Ohara S-FPL53 glass focuses a ray bundle with an angle of 20° to the optical axis. The glass has a small refractive index of 1.44 and a relatively large Abbe number of 94.7. (b) The magnified focusing area marked by the dotted-line circle in (a). The longitudinal- and lateral-color-aberration magnitudes are noted. (c) A 20-mm-focal-length lens made of Ohara S-LAH79 glass focuses a ray bundle with an angle of 20° to the optical axis. The glass has a large refractive index of 2.00 and a relatively small Abbe number of 28.3. (d) The magnified focusing area marked by the dotted-line circle in (c). The longitudinal- and lateral-color-aberration magnitudes are much larger than the color aberrations shown in (b) because the smaller Abbe number results in large color aberrations.

thickness, and a 20-mm focal length. The incident rays shown in Fig. 1.26(a) have a 20° angle to the optical axis. In Fig. 1.23(a), only one wavelength of 0.55 μm is used, whereas Fig. 1.26(a) uses a blue-color wavelength of 0.45 μm and a red-color wavelength of 0.65 μm to show the color aberration. Figure 1.26(b) shows the magnified view of the focusing area marked by the dotted-line circle in Fig. 1.26(a).

Figure 1.26(b) shows that the two wavelengths are focused at different locations. The 0.45-μm wavelength is focused at a point closer to the lens than the 0.65-μm wavelength because of the glass dispersion. The horizontal distance between the two focused points is ~0.35 mm; this is the "longitudinal color aberration." Longitudinal color, also called "axial color," is approximately

proportional to the focal length of the lens. The vertical distance between the two focused points is ~0.02 mm; this is the "lateral color aberration," which is approximately proportional to the incident angle of the rays. Both the longitudinal and horizontal color are approximately proportional to the inverse of the Abbe number value. The Abbe number value of S-FPL53 glass is 94.9, which is very large, and thus the glass has very weak color dispersion.

The lens in Fig. 1.23(b) is re-plotted in Fig. 1.26(c). It uses Ohara S-LAH79 glass with a large refractive index of 2.00 and a small Abbe number of 28.3. This lens has a 15-mm diameter, a 5-mm central thickness, and a 20-mm focal length. The incident rays shown in Fig. 1.26(c) have a 20° angle to the optical axis. In Fig. 1.23(b), only one wavelength of 0.55 μm is used, whereas in Fig. 1.26(c) a blue-color wavelength of 0.45 μm and a red-color wavelength of 0.65 μm are used to show the color aberration. Figure 1.26(c) shows that the two wavelengths are focused at different locations. Figure 1.26(d) shows the magnified view of the focusing area marked by the dotted-line circle in Fig. 1.26(c). The color aberrations in Fig. 1.26(d) is larger than those in Fig. 1.26(b) because the Abbe number of glass S-LAH79 is much smaller than the Abbe number of glass S-FPL53. The magnitude of the longitudinal and lateral color aberrations is ~0.92 mm and ~0.22 mm, respectively.

1.12.5 Field curvatures

In Fig. 1.27 a Zemax-generated raytracing diagram for a single lens is used to explain field curvature. This single lens is made of N-BK7 glass with 3-mm central thickness, 11-mm diameter, and 15-mm focal length. The three incident angles of rays are 0°, 10°, and 20°. The wavelength of the rays is 0.55 μm. This lens has severe coma. To avoid unnecessary complexity and to simplify the illustration, only two meridional rays are traced for every field angle. The real chief ray is not traced; instead, a conceived "coma-free" chief ray is manually drawn to help explain the situation. The three focused spots for the three field angles are marked by dark solid dots.

Every lens has a basic curved image surface, called a Petzval surface, as shown in Fig. 1.27. For a simple thin lens, the axial distance between the Petzval surface and the ideal flat image surface is given by $h^2/(2nf)$, where h is the image height, n is the refractive index of the lens material, and f is the lens focal length. The axial position of the focused spot moves toward the lens as the field angle increases. The tangential image surface is the dashed curve that links these three focused spots. Positive lenses introduce inward field curvatures. Negative lenses introduce outward field curvatures. This phenomenon is called field curvature. The image surface of a single lens is more severely curved than the Petzval surface, which is also shown here.

The raytracing and the field curvature shown in Fig. 1.27 are in the tangential plane. In the sagittal plane, a simple positive lens will also introduce

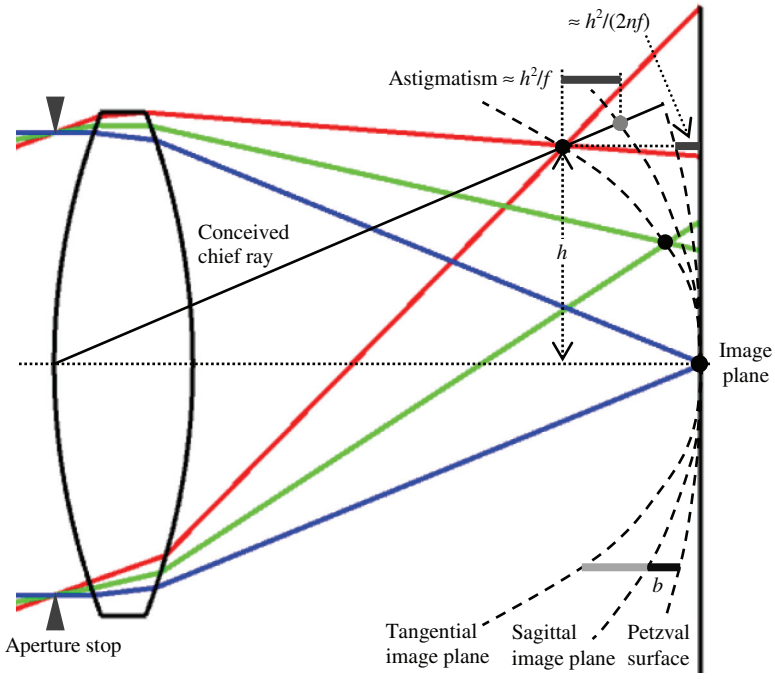

Figure 1.27 Zemax-generated raytracing diagram for a single lens shows a severe, inwardly curved image surface in the tangential plane. The sagittal image surface, which is perpendicular to the book page, is less curved than the tangential image surface but more curved than the Petzval surface.

an inward field curvature with a magnitude smaller than the field curvature magnitude in the tangential plane, as illustrated in Fig. 1.27. The axial distance between the tangential and sagittal image surfaces is about twice the axial distance between the sagittal image surface and the Petzval surface, as indicated by the two thick grey lines in Fig. 1.27.

The axial distance between the two focused points on the tangential and sagittal image surfaces along the same chief ray is the astigmatism, as marked in Fig. 1.27. The magnitude of astigmatism is about h^2/f, a parameter that is special in determining the magnitude of field curvatures.

For a multi-element lens, the field curves in both tangential and sagittal planes can have complex profiles. Since all of the sensors used to sense the image produced by a lens are planar, field curvature will cause defocusing and reduce the image sharpness, particularly at large field angles. Field curvature can often be mostly corrected by combining several positive and negative lens elements with properly selected surface curvatures and glasses or/and the use of some aspheric lenses. A well-designed lens should have a nearly flat image plane in both tangential and sagittal planes.

1.12.6 Wavefront errors and optical path difference

All optical aberrations are some type of wavefront error. For example, when a lens focuses rays to a point, the ideal wavefront is a convergent sphere. The center of the sphere is the focal point. All rays will travel along the radii of the wavefront sphere and be focused at the center. However, because of the presence of various types of aberrations, the real wavefront deviates from the ideal sphere, and the rays will not be focused at the same spot. If a real wavefront is "toroidal," there is astigmatism.

The concept of optical path difference (OPD) quantifies the deviation of a real wavefront from a spherical reference wavefront. Two types of OPD are frequently used: peak-to-valley OPD_{PV} and root-mean-square OPD_{RMS}. Figure 1.28 shows two examples. OPD_{PV} is the largest OPD between the wavefront and a reference sphere over the entire wavefront. OPD_{RMS} is calculated by the following equation using the data of a series of sampling points on the wavefront and the reference sphere:

$$OPD_{RMS} = \sqrt{\frac{\sum_{i=1}^{m} OPD_i^2}{m}}, \tag{1.30}$$

where m is the number of sampling points used to calculate the OPD_{RMS}. There is no rule about how many sampling points should be used. Common sense of mathematics applies here.

OPD_{RMS} contains more complete information about the wavefront and is widely used to evaluate the quality of a wavefront. For example, the wavefront error shown in Fig. 1.28(a) is more severe than the wavefront error shown in Fig. 1.28(b), where the two OPD_{PV} are the same but the OPD_{RMS} shown in Fig. 1.28(a) is larger. Since a real wavefront can have various complex shapes,

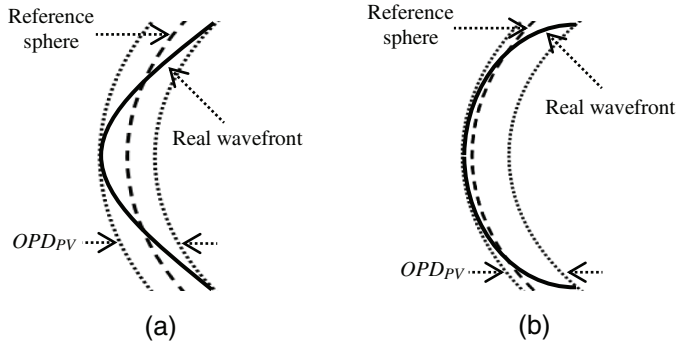

Figure 1.28 The solid curves depict two real wavefronts. The dashed curves are the two reference spheres selected to compare with the real wavefront. The distance between the real wavefront and the reference sphere is the OPD. The dotted curves define the peak-to-valley OPD_{PV}.

there is no simple relation between the OPD_{PV} and OPD_{RMS} because the wavefront can have various shapes. For a mixture of low-order aberrations, $OPD_{PV} \approx 4.5 OPD_{RMS}$. It is usually assumed that $OPD_{PV} \approx 5 OPD_{RMS}$.

1.12.7 How to read an OPD diagram

Optical software can plot the OPD diagram for a lens to evaluate the lens quality. Figure 1.29 shows a Zemax-generated OPD diagram for the double

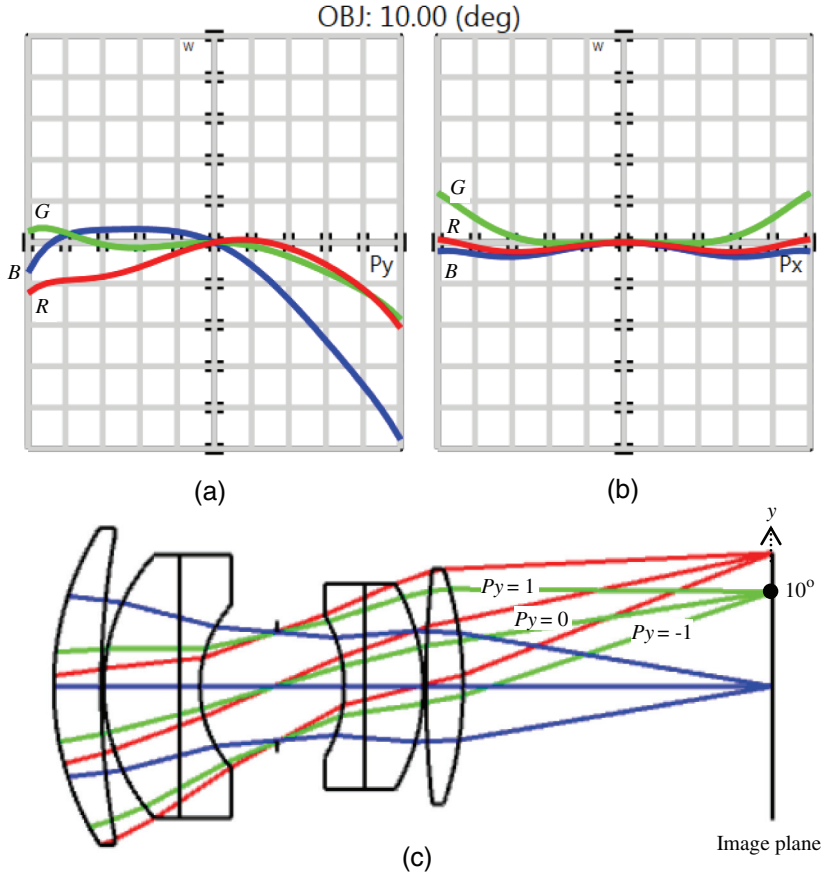

Figure 1.29 Zemax-generated OPD diagram at the image plane for the double Gauss lens shown on Fig. 1.10 for (a) a field angle of 10° in the sagittal direction (perpendicular to the page) and (b) 10° in the tangential direction (page plane), and RGB three colors. In (a) and (b), the vertical axis is the OPD with a unit of "wave" (the wavelength of He-Ne laser). (a) OPD vs. P_y for $P_y = -1$ to $P_y = 1$. P_y is marked in Fig. 1.29(c). The R and G colors have an OPD up to 2 waves. The B color has the largest OPD up to 5 waves. (b) OPD vs. P_x for $P_x = -1$ to $P_x = 1$, which is in the sagittal plane perpendicular to the page and is not marked in Fig. 1.29(c). The OPD is symmetric about $P_x = 0$ and is smaller than the OPD shown in (a). (c) Re-plot of the raytracing diagram for the double Gauss lens to illustrate P_y. The y direction is along the page plane and is the tangential direction. The 10° field and the $P_y = -1$, 0, and 1 are marked.

Gauss lens shown in Fig. 1.10, which is re-plotted here in Fig. 1.29(c) for convenience.

For any given field angle, $P_y = 0$ is the chief ray, $P_y = -1$ is the low ray, and $P_y = 1$ is the high ray, as marked in Fig. 1.29(c) for the field angle of 10°. All of the rays for any given field angle in the tangential plane are in the range of $-1 \le P_y \le 1$.

The OPD in the tangential direction is shown in Fig. 1.29(a) and is larger than the OPD in the sagittal direction shown in Fig. 1.29(b). This phenomenon is commonly seen and can also be seen in the CTF diagram in Fig. 1.33.

OPD is closely related to image quality. The Rayleigh criterion states that the $OPD_{PV} < 0.25$ wave or $OPD_{RMS} < 0.07$ wave can be translated to diffraction-limited quality.

1.13 Evaluation of Image Quality

Many lenses are used to generate an image of an object on a sensor for display. The image quality evaluation is an important subject. Image resolution and image distortion are the two most important parameters used to describe the image quality.

Several criteria are used to describe image resolution. Among these, the contrast transfer function is the most commonly used. The Rayleigh criterion (another criterion from the same Rayleigh mentioned earlier) is also frequently used. These two criteria are related.

1.13.1 Image resolution: Rayleigh criterion

Image resolution involves the diffraction of optical waves, which will be discussed later in Chapter 2. The results are presented here in advance.

The image of an infinitely small object point formed by a perfect lens has a certain size because of diffraction. The intensity profile of the image spot contains a central lobe and several side lobes with gradually decreasing intensity, as shown in Fig. 2.12(a). The spatial resolution limit of this lens is defined by Rayleigh as when the maximum of one image spot falls on the first minimum of the next image spot, as shown by the solid curves in Fig. 1.30(a). In such a case, the intensity profile of the two-image-spot combination has a central dip of ~74% of the maximum intensity of the two-image-spot combination, as shown by the dashed curve in Fig. 1.30(a), and the two image spots are said to be just resolvable.

The distance R between the two spots is the resolution limit of the lens. Any two image spots with a distance between them smaller than R is not resolvable (see Fig. 1.30(b)). Any two image spots with a distance between them larger than R is resolvable (see Fig. 1.30(c)).

This criterion is the Rayleigh criterion and is widely used to define the resolution limit of lenses. The image spots can either be diffraction limited

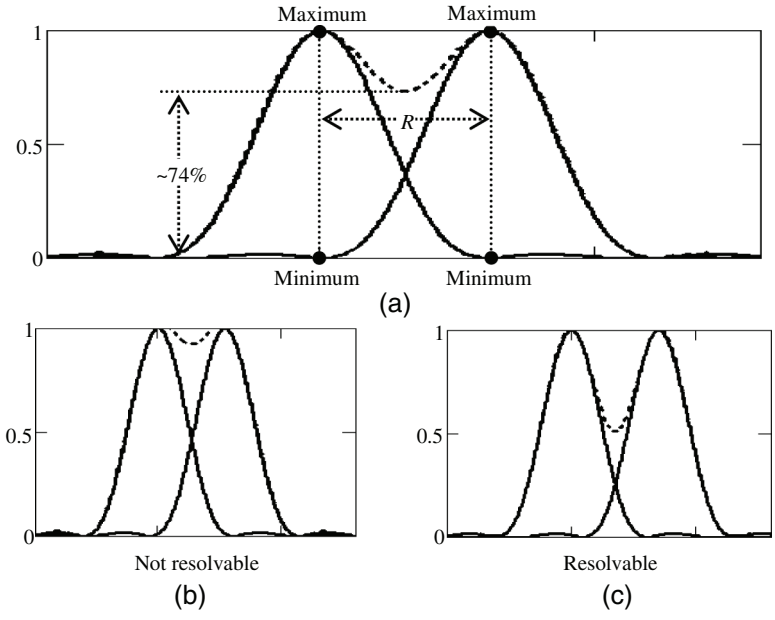

Figure 1.30 Rayleigh criterion of resolvable images. The solid curves are the intensity profiles of image spots. The dashed curves are the intensity profile of the two-image-spot combination. (a) The two image spots are just resolvable according to the widely used Rayleigh criterion. Distance R between the centers of the two spots is the image resolution of the lens used to generate these image spots. (b) The two image spots are too close to be resolvable. (c) The two image spots are clearly resolvable.

(aberration free) for a very-high-quality lens or have severe aberration for a low-quality lens. A lens that can form smaller focused spots has higher image resolution.

1.13.2 US Air Force resolution test chart

The US Air Force (USAF) resolution test chart, as shown in Fig. 1.31, is widely used as an object to measure the resolution of an image lens. The lens under test generates an image of the chart. The resolution of the lens can be found by comparing the chart and the image of the chart.

A resolution test chart contains many black-and-white bars. All of these bars are divided into several groups. Each group contains several bar elements with different widths or spatial frequencies. The chart is designed so that the spatial frequency R of a bar can be calculated by

$$R = 2^{group+\frac{element-1}{6}}, \tag{1.31}$$

where *group* and *element* are the group number and element number, respectively, as marked in Fig. 1.31.

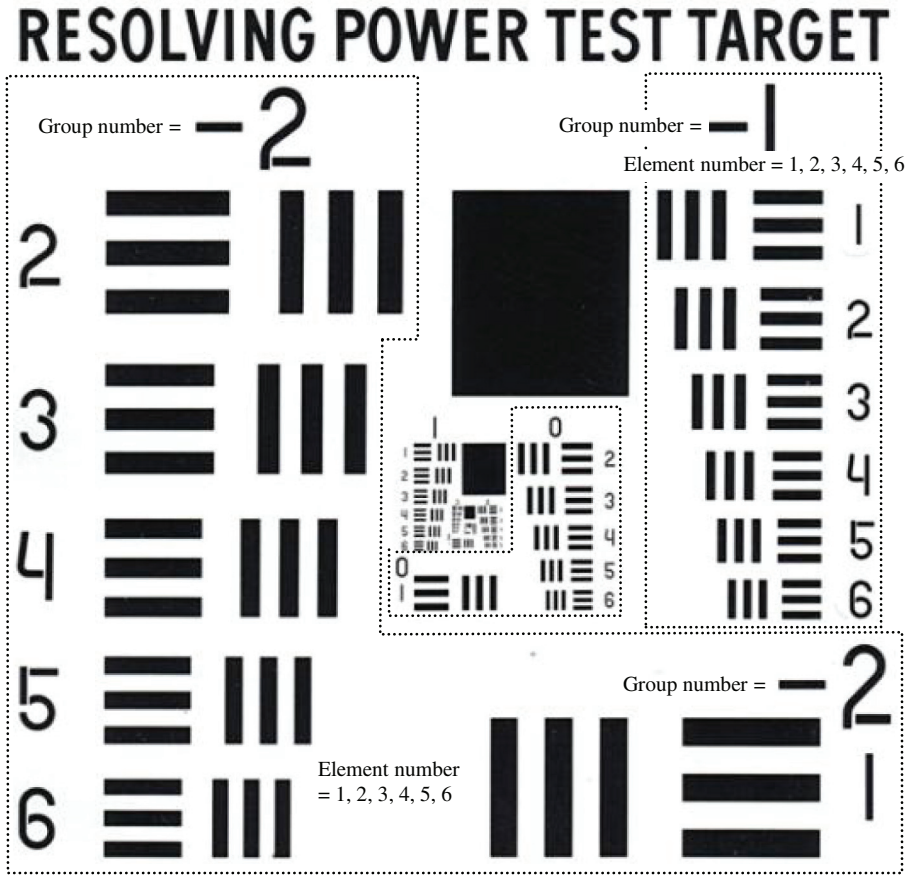

Figure 1.31 A US Air Force resolution test chart consists of many groups of black-and-white bars with different widths or spatial frequencies. The three dotted-line frames have been added by the author to mark the three groups of −2, −1, and 0, respectively, and the notes "Group number = −2", "Element number = 1, 2, 3, 4, 5, 6", and "Group number = −1" have been added for clarification.

The modulation depth of the chart is defined by $(w-b)/(w+b)=1$, where $w=1$ and $b=0$ are the normalized optical intensities of the black-and-white bars, respectively. Detailed information about the USAF resolution test chart can be found in the literature.[8]

1.13.3 Image resolution: contrast transfer function

The spatial resolution of an imaging lens can be quantized by the contrast transfer function (CTF) or modulation transfer function (MTF). These two quantities are related, and the CTF is the more commonly used of the two.

The CTF of an imaging lens can be measured using a USAF resolution test chart or any similar chart, such as a Siemens star chart. The value of the

CTF is defined as the modulation depth of the image of a target with a modulation depth of 1. The mathematical form of the CTF is

$$CTF = \frac{I_w - I_b}{I_w + I_b}, \tag{1.32}$$

where I_b and I_w are the intensities of the images of the black/white bar, respectively, and the intensities of the target are normalized to $b = 0$ and $w = 1$.

Because of the diffraction, the image of a black/white vertical edge formed by a perfect lens is a gradient grey area, which is called the "line spread function." For an imperfect lens, the presence of various types of aberrations increases the width of this gradient grey area, as shown in Fig. 1.32(b), where the grey area width $2s$ is exaggerated. If the black-and-white bar width P is larger than s, the portions of the image outside the grey area still have

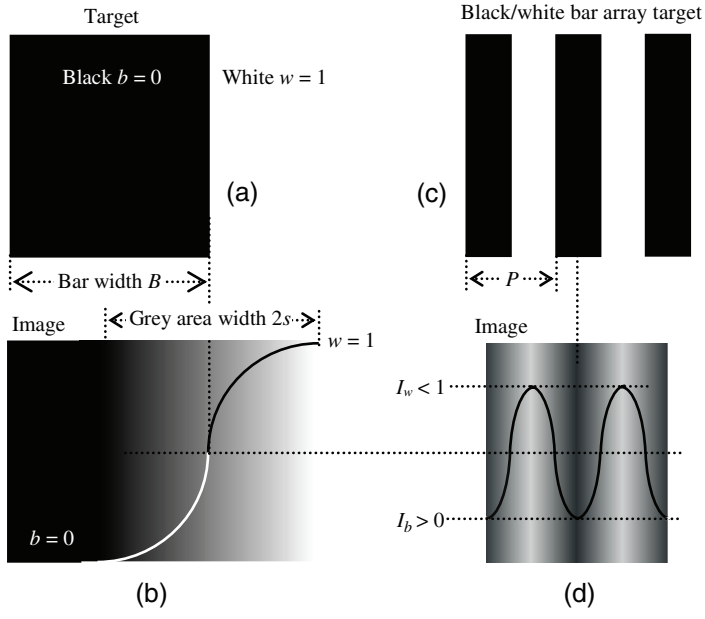

Figure 1.32 Illustration the relation between black-and-white targets and their images. (a) A black-and-white edge target with normalized intensity of $b = 0$ and $w = 1$. (b) Image of the black-and-white edge target formed by a lens is a gradient grey area with width $2s$ determined by the diffraction and aberrations of the lens. The curve is the intensity level of the grey area. Inside the grey area, the modulation depth of the image is $CTF = (I_w - I_b)/(I_w + I_b) < 1$. Outside the grey area, the modulation depth of the image is $CTF = (w - b)/(w + b) = 1$. (c) A black-and-white bar array target with spatial period P and normalized intensities of $b = 0$ and $w = 1$. (d) Image of the black-and-white bar array. When the bar width is $< s$, the modulation depth of the image is always < 1. As the bar width decreases, the image modulation depth decreases. When the bar width is much smaller than s, the modulation depth of the image approaches zero.

normalized intensities of 1 and 0, respectively, as shown in Fig. 1.32(b), and the image *CTF* is still 1.

If the black-and-white bar width *P*/2 is smaller than *s*, the image only consists of periodic grey areas, as shown in Fig. 1.32(d). The dark portion of the image is no longer black and has a normalized intensity $I_b < 1$. The light portion of the image is no longer white and has a normalized intensity $I_w > 0$. The image *CTF* is < 1, and the *CTF* value decreases as the ratio *P*/*s* decreases. The lens spatial resolution limit is widely defined as $CTF = 0.3$. The test bar width of the corresponding image bar width for $CTF = 0.3$ is the resolution limit of the lens under test. For $CTF = 0.3$, higher-quality lenses have a smaller *s* value and can resolve a smaller *P* (narrower test bar). The CTF will eventually decrease to zero for any lenses with a very small *P*/*s*.

When a test chart is used to measure the resolution limit of an imaging lens, the distance between the test chart and the lens must be specified because the same bars will look smaller if the distance increases. More frequently, people use the image bar width instead of the test chart bar width to specify the resolution limit of a lens. There is a one-to-one relation between the widths of the test bar and the image for a given lens focal length and the distance between the test chart and the lens.

The widths of the test chart bar and image bar are often expressed in terms of cycle/mm or line pair/mm. One cycle or one line pair means one black-and-white bar pair.

1.13.4 How to read a CTF diagram

Figure 1.33 shows a CTF diagram produced by Zemax for the double Gauss lens shown in Figs. 1.10 and 1.29(c):

1. The vertical axis "Square Wave MTF" means the CTF.
2. The horizontal axis "Spatial frequency in cycles per mm" is the image bar width. For example, 100 cycles/mm means that there are 100 black-and-white bar pairs in one millimeter in the image, and each bar has a 5-μm width.
3. For a given lens, the CTF varies as the angle of incident ray varies. A larger incident ray angle usually results in a smaller CTF. The CTFs for 0°, 10°, and 14° are plotted in Fig. 1.33.
4. For any angles of ray incidence, the CTF in the tangential and sagittal directions are represented by the dotted and solid curves in Fig. 1.33, respectively, and are often different, which means the lens has astigmatism. The black-and-white bars shown at the left of Fig. 1.33 illustrate this phenomenon. For a 0° incident angle, the rays are symmetric about the optical axis. The concepts of tangential and sagittal are invalid.
5. The top black curve in the diagram is the "diffraction-limited" CTF curve that is the theoretical best possible CTF curve that can be

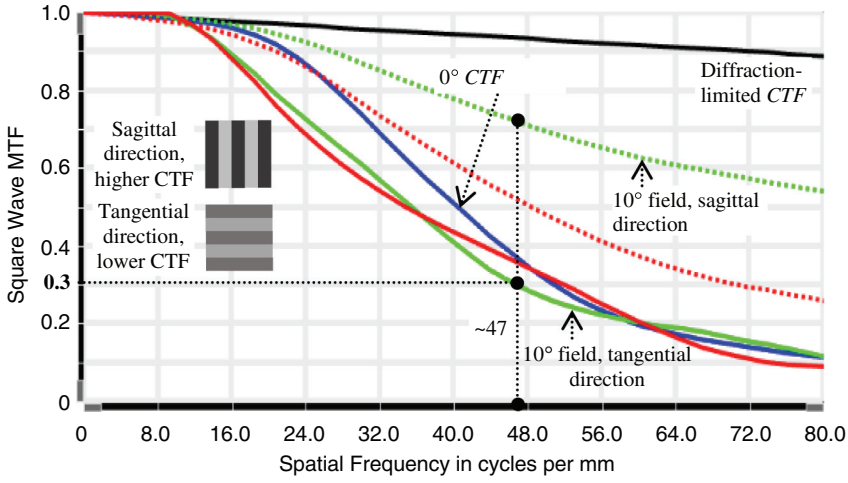

Figure 1.33 A CTF diagram generated by Zemax for the double Gauss lens shown in Figs. 1.10 and 1.29. The CTF value is a function of bar spatial frequency. The CTF curves for three field angles of 0°, 10°, and 14° are plotted. The CTF in the tangential and sagittal directions are plotted by dotted and solid curves, respectively; these two curves are usually different. At field center (0°), the CTFs in the tangential and sagittal directions coincide. The spatial resolution limit ($CTF \geq 0.3$) is ~47 cycle/mm in the tangential direction for a 10° field, as marked.

obtained only by a perfect lens of the same *F*-number. Different lenses have different diffraction-limited CTF curves, which are always smaller than 1.

6. The values of all CTF curves start as 1 at 0 cycles/mm, which means that the black-and-white bar width is infinitely large and decreases as the spatial frequency increases.
7. All CTF curves at a given spatial frequency are much smaller than the diffraction-limited CTF curve. This phenomenon indicates that this lens has large aberrations.
8. Spatial frequency of ~47 cycle/mm is about the resolution limit of this lens. Below this spatial frequency, all CTF curves >0.3. 47 cycle/mm translates to a ~10-μm bar width in the image.
9. The CTF for a 10° field in the tangential direction is much smaller than the CTF in the sagittal direction. This phenomenon can also be seen in the OPD diagram in Fig. 1.29.

1.13.5 Image resolution: modulation transfer function

Traditionally, a test chart with a sinusoidal periodic intensity pattern is also used. The normalized intensity of such a chart gradually varies between 0 and 1. The image modulation depth of such a test chart is the MTF. Because the sinusoidal intensity pattern does not have a sharp black-and-white edge,

the value of the MTF is always smaller than the value of the CTF for the same spatial frequency and the same lens. The equations linking the CTF to MTF are[9]

$$MTF(\nu) = \frac{\pi}{4}\left[CTF(\nu) + \frac{CTF(3\nu)}{3} - \frac{CTF(5\nu)}{5} + \frac{CTF(7\nu)}{7} \ldots\right] \quad (1.33)$$

or

$$CTF(\nu) = \frac{\pi}{4}\left[MTF(\nu) - \frac{MTF(3\nu)}{3} + \frac{MTF(5\nu)}{5} - \frac{MTF(7\nu)}{7} \ldots\right], \quad (1.34)$$

where ν is the spatial frequency. Nowadays, the USAF resolution test chart is the most widely used test chart, and therefore the CTF is more widely used than the MTF.

All of the optical aberrations described in Section 1.12 can reduce the CTF or MTF of a lens. Increasing either value requires an overall improvement of the lens and can be achieved by carefully designing the lens using more appropriate optical elements.

1.13.6 Effect of sensor pixel size: Nyquist sampling theorem

The image formed by a lens is often detected by a 2D sensor array and sent to a display device. The sensor pixel size will have an impact on the resolution of the image sent to the display device.

The Nyquist sampling theorem[10] states that when sampling a periodic signal, at least two samplings per period are needed to recover the signal. In the case here, one period of signal consists of one black-and-white bar pair, and one sampling means one pixel. The resolution limit of the lens determines the smallest image of the bars that the lens can produce. When the pixel size of the sensor is larger than the resolution limit of the lens, the detector cannot fully resolve the image generated by the lens, so the quality of the lens is partially wasted, and vice versa. For example, to fully utilize the resolution of the image in Fig. 1.33, the sensor pixel must be smaller than 10 μm.

In the visible range, CCD arrays are widely used as the sensor. The pixel size of CCD arrays can be as small as a couple of microns now. Most lenses in the visible range have an image resolution lower than the sensor resolution and limit the resolution of the lens/sensor combination.

1.13.7 Image distortion

Any optical image is more or less distorted. The image distortion of a lens can be analyzed by imaging a grid target. There are two typical types of image distortion: barrel distortion and pincushion distortion, as shown in Fig. 1.34.

Barrel distortion shrinks the image, whereas pincushion distortion stretches the image. Some optical lenses produce images with a mixture of

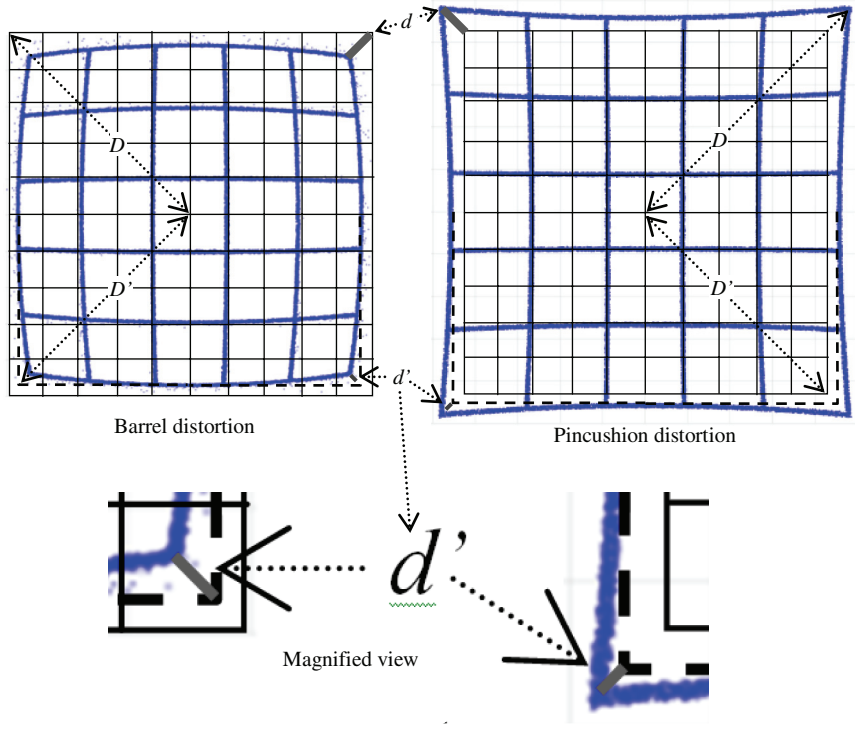

Figure 1.34 Two typical types of image distortion: barrel and pincushion. The thin line grid is the image of a grid target without distortion plotted here for comparison. The thick curved grids are Zemax-generated distorted images of the grid target generated by a lens under test. The very thick and short grey lines at the top right and left corners mark the absolute distortion value d. The relative distortion d/D is what is really sought, where D is the half-diagonal of the grid. Another way to define the relative distortion is d'/D', where d' is shown in the magnified view. Distortion d'/D' ignores the image size change caused by distortion.

these two types of distortion. Figure 1.34 depicts pure barrel and pincushion distortions. The magnitude of the relative image distortion at a certain location in the image is the distorted amount divided by the radial position of this location. The maximum magnitude of the relative image distortion often appears at the image corner and is marked by the two thick, short grey lines at the top corner in Fig. 1.34. The maximum relative distortion is the length of these thick, short grey lines divided by the half-diagonal of the grid image. The maximum relative distortions in both images in Fig. 1.34 are ~8%.

The image size compression or stretching caused by distortion can often be compensated by intentionally designing a lens with a "too large" or "too small" image, respectively. Lens users are annoyed by the curved images of straight lines. Based on this argument, the image compression or stretching can be excluded from the distortion, and the maximum relative distortion can be defined by d'/D', as shown in Fig. 1.34.

Image distortion can be minimized by carefully designing the lens using more appropriate optical elements, which is an issue when balancing the performance and cost. For a camera lens, image distortion with a couple of percent magnitude is acceptable. For an inspection lens, the required distortion can be below 0.1%.

1.14 Illumination Optics versus Imaging Optics

Thus far, only imaging optics has been discussed, i.e., where an image of an object is formed. Every image point has an exclusive one-to-one relation with an object point, as illustrated in Fig. 1.35(a).

Illumination optics is different from imaging optics and is also a part of geometrical optics. In illumination applications, certain illumination intensity patterns (usually a uniform intensity pattern or a flat-top pattern) are desired. Many light sources, such as a tungsten bulb, have a certain source structure. If imaging optics is used to handle the illumination light, the light source structure will be imaged in the illumination pattern. Such a result is not desired. Therefore, illumination optics intentionally avoids forming any images, as illustrated in Fig. 1.35(b). Light from many source points can reach the same point on the working plane, or light from one source point can reach many points on the working plane. That is why illumination must be discussed in a separate section. Generally speaking, illumination optics is simpler than imaging optics, in terms of the number of variables involved and the complexity of the merit function, because the requirements for an illumination pattern are often not specified to a high accuracy. Illumination optics is also less widely used than imaging optics.

Since illumination optics does not focus the light from an object to form an image, all of the aberrations discussed in Section 1.12 are no longer relevant. Aberrations are sometimes utilized to generate the desired illumination pattern. The source wavelength can still be an issue. Besides the capability of generating

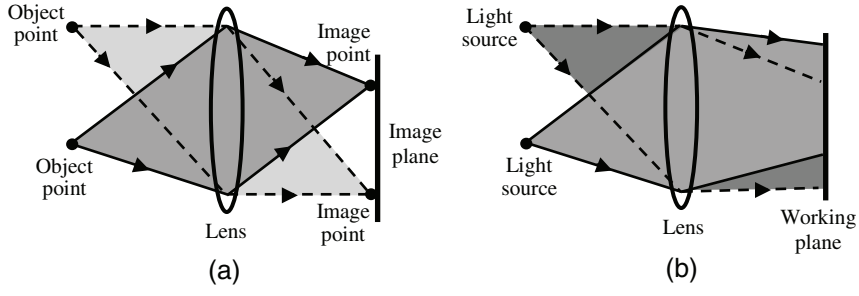

Figure 1.35 (a) An imaging lens forms an image of an object on an image plane. (b) An illumination lens forms the light from light source points to illumination patterns on a working plane. The combination of these illumination patterns is the desired illumination pattern. There is no image formed.

a certain illumination pattern, the light-collecting power is another important issue for image optics.

Optical design software for imaging optics are often capable of designing illumination optics as well, but their capability to design illumination optics is inferior. The designer sets the desired illumination pattern as the target in the merit function in the software, and the software will try to find a certain optical structure to meet the target. The raytracing technique used to design illumination optics is called "non-sequential raytracing," whereas the raytracing technique used when designing imaging optics is called "sequential raytracing."

Illumination optics deal not only with optical components to handle the light but also the light sources, including various types and shapes of light bulbs, lasers, and light-emitting diodes (LEDs). These light sources can have dramatically different characteristics and illuminations. Lenses used to work with different light sources can also be very different.

Section 12.4 will discuss the modeling of illumination lenses and light sources.

1.15 Radiometry

Radiometry is the science about handling light intensity. Among the units and terminologies used, some are unique and may be confusing. In this book, only the International System (SI) of units is used to avoid unnecessary complexity. The two widely used terminologies in radiometry are "radiance" and "irradiance." The former means the power per unit solid angle per unit area, and the latter means the power per unit area.

Many optical engineers and scientists are more experienced handling imaging optics than handing radiometry, although radiometry contains less substance than imaging optics. See Palmer and Grant[11] and McCluney[12] for more about radiometry.

1.15.1 Lambert's cosine law

One basic law in radiometry is Lambert's cosine law. The radiance characteristics of most diffusive surfaces can be approximately described by[13]

$$J(\theta) = J_0 \cos(\theta), \tag{1.35}$$

where J_0 is the maximum radiance that appears in the normal direction of the radiation source surface, θ is the angle between the direction of interest and the radiation normal, and $J(\theta)$ is the radiance in the θ direction.

One interesting characteristic of Lambertian sources is that the apparent radiance viewed in any direction is the same. In a direction θ off the source normal, the radiance falls from J_0 to $J_0\cos(\theta)$, and at the same time the source area being viewed is increased by a factor of $1/\cos(\theta)$, as illustrated in Fig. 1.36; these two factors compensate each other.

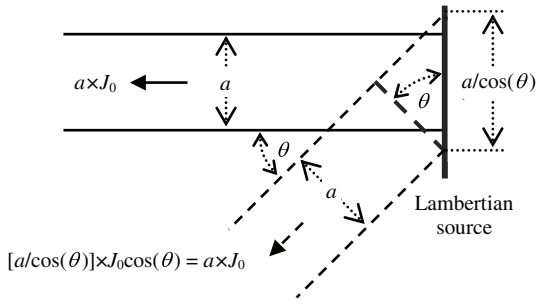

Figure 1.36 A Lambertian source has the same radiance in any direction.

1.15.2 Light-collecting power of a lens

One key specification of an illumination lens is its light-collecting power. Consider the case shown in Fig. 1.37, where the irradiance power P collected by this lens is given by

$$\begin{aligned} P &= \int AJ_0 \cos(\theta) d\Omega \\ &= \int_0^R AJ_0 \cos(\theta) \frac{2\pi r \cos(\theta) dr}{\left[\frac{o}{\cos(\theta)}\right]^2} \\ &= \frac{2\pi AJ_0}{o^2} \int_0^R r\cos^4(\theta) dr \\ &= 2\pi AJ_0 \int_0^{\theta_M} \sin(\theta)\cos(\theta) d\theta \\ &= \pi AJ_0 \frac{R^2}{o^2 + R^2}, \end{aligned} \tag{1.36}$$

where A is the area size of the light source; o is the distance between the light source and the lens; i is the distance between the lens and the illuminated area with size A'; J_0 and J_0' are the peak radiance and irradiance of the light source and illuminated area, respectively; R is the radius of the lens aperture; θ and θ′ are the angles between the point of interest on the lens and the optical axis, respectively; the largest value of θ is $\theta_M = \sin^{-1}[R/(o^2 + R^2)^{0.5}]$; $r = o \times \tan(\theta)$ is the radial variable on the lens, $dr = o \times d\theta/\cos(\theta)^2$; $d\Omega = 2\pi\cos(\theta) r dr/[o/\cos(\theta)]^2$ is the incremental solid angle marked by the grey color ring on the lens imposes on the light source; and $d\Omega$ is integrated over the entire lens aperture.

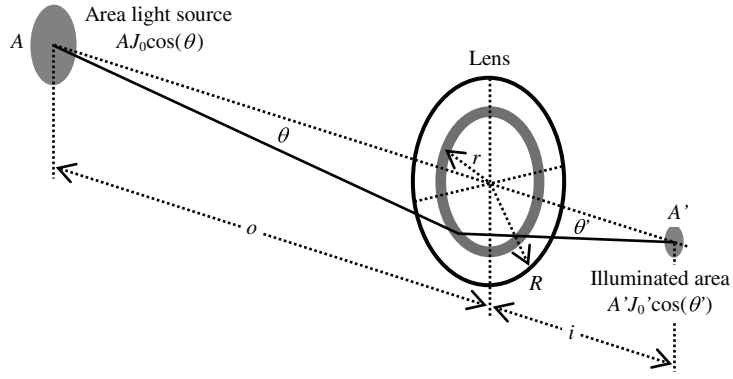

Figure 1.37 A lens with aperture radius R collects light power emitted by an area light source.

When $o \ll R$, Eq. (1.36) leads to $P = \pi A J_0$, which is the total power of a Lambertian source emitted to a half-sphere.

1.15.3 Inverse square law

For $o \gg R$, Eq. (1.36) reduces to $P = \pi A J_0 R^2/o^2 \sim 1/o^2$. This means that the light power collected by a lens is inversely proportional to the square of the distance between the lens and the light source, which is the inverse square law. When $R/o \to 0$, Eq. (1.36) leads to $P \to 0$, and the lens is too small or placed too far from the source, or both.

1.15.4 A point illuminated by a circular Lambertian source

Consider a point that is illuminated by a circular Lambertian source. The point is at the optical axis of the light source and a distance o away, as shown in Fig. 1.38. The source has a radius R and radiance $J_0\cos(\theta)$; the irradiance power the point receives can be calculated by

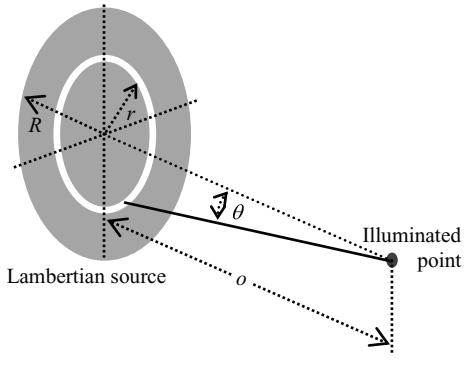

Figure 1.38 A circular Lambertian light source with radius R illuminates a point at the optical axis of the source and a distance o away.

$$
\begin{aligned}
P &= \int_0^R J_0 \cos(\theta) \frac{2\pi r \cos(\theta) dr}{\left[\frac{o}{\cos(\theta)}\right]^2} \\
&= 2\pi J_0 \int_0^{\theta_M} \sin(\theta)\cos(\theta) d\theta \\
&= \pi J_0 \sin^2(\theta_M) \\
&= \pi J_0 \frac{R^2}{o^2 + R^2},
\end{aligned}
\tag{1.37}
$$

where $2\pi\cos(\theta)rdr/[o/\cos(\theta)]^2$ is the incremental solid angle marked by the white ring on the source, $r = o \times \tan(\theta)$ is the radial variable on the source, $dr = o \times d\theta/\cos(\theta)^2$, and the largest value of θ is $\theta_M = \sin^{-1}[R/(o^2 + R^2)^{0.5}]$.

Note that the derivation of Eq. (1.37) is similar to the derivation of Eq. (1.36), and the results of these two equations are similar. Again, for $o \gg R$, Eq. (1.37) reduces to $P = \pi J_0 R^2/o^2 \sim 1/o^2$, and for $o \ll R$, Eq. (1.37) reduces to $P = \pi J_0$.

1.15.5 A working plane illuminated by a point source: the $\cos^4(\theta)$ off-axis relation

Consider a Lambertian point light source that illuminates a working plane perpendicular to the normal of the source radiance and has a distance o to the light source, as shown in Fig. 1.39. This radiance normal direction is the optical axis. A unit size area a on the optical axis at the working plane obviously receives the peak irradiance. For an off-axis unit size area b on the

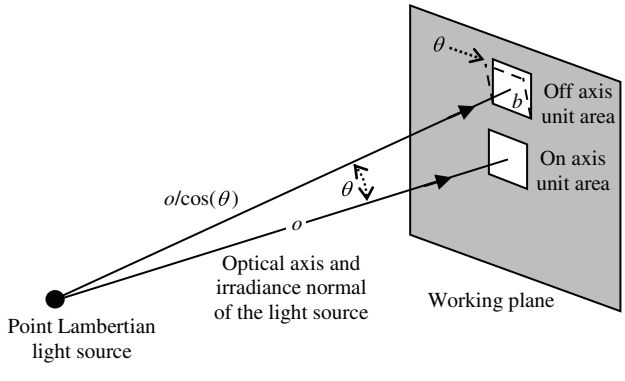

Figure 1.39 The radiance received by an off-axis unit area from a point Lambertian light source is $\cos^4(\theta)$ of the radiance received by the on-axis unit area, where θ is the off-axis angle.

working plane with an angle θ to the optical axis, the irradiance received is less than the irradiance received by *a* because of three factors:

1. The area projected by *b* in the direction of the light source is smaller than *b* by a factor of cos(θ), so the irradiance received by *b* is lowered by a factor of cos(θ).
2. The distance from the light source to area *b* is *o*/cos(θ). Thus, the irradiance received by *b* is lowered by a factor of $\cos^2(\theta)$, according to the inverse square law.
3. Finally, if the light source is Lambertian, the radiance in the direction of *b* is lower than the peak radiance in the normal direction by a factor of cos(θ).

The irradiance received by unit area *b* can be found by multiplying all three factors, i.e., $\cos^4(\theta)$ of the peak irradiance received by unit area *a*.

1.15.6 Etendue and radiance conservation

There are two conservation quantities in illumination optics: etendue and radiance. These two quantities are related, and familiarity with them can help one understand the performance of existing illumination lenses and design new illumination lenses.

Consider a lens with radius *R* that takes light from a light source a distance *o* away and transforms the light to illuminate an area a distance *i* away, as shown in Fig. 1.40, where *h* and *h*′ are the half height of the light source and the illuminated area, respectively, and $\Omega = \pi R^2/o^2$ and $\Omega' = \pi R^2/i^2$ are the solid angle the lens imposes on the light source and the illuminated area, respectively. Note that the *h*′ shown here is an ideal case. For a real lens, there are always some defects or aberrations; the real *h*′ will be somewhat larger than the *h'* shown here. Thus, the relation $h/o \leq h'/i$ in the plane of the page can be established. A similar relation in the plane perpendicular to the page plane also holds. The relation

$$\frac{A}{o^2} \leq \frac{A'}{i^2} \tag{1.38}$$

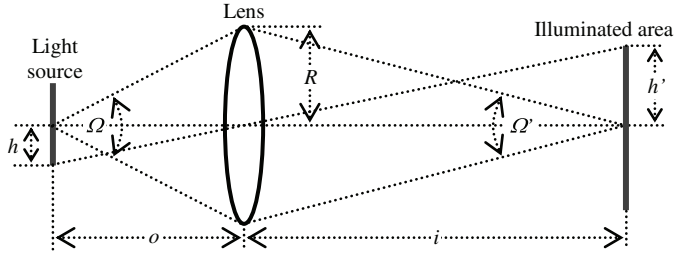

Figure 1.40 A lens with a radius *R* takes light from a light source a distance *o* away and illuminates an area a distance *i* away.

can then be established, where $A = h^2$ and $A' = h'^2$ are the size of the light source area and the illuminated area, respectively. Multiply both sides of Eq. (1.38) by πR^2 to produce

$$A\Omega \leq A'\Omega'. \tag{1.39}$$

$A\Omega$ is the etendue or throughput of the lens. Equation (1.39) is the etendue conservation law, which says that for a perfect lens the etendue remains a constant as the light propagates through the lens. If the lens is not perfect, etendue increases as the light propagates through it.

Etendue has applications in the design of illumination lens. For example, consider the use of a lens to couple all of the light power from a light source 100 mm away into a light guide 100 mm away. The light source size $A = 2$ cm^2, and the light-guide area size is $A' = 1$ cm^2. Then Eq. (1.39) is violated: only a portion of the light power can be coupled into the light guide, and the intended outcome will certainly fail.

When a light source and an illumination lens illuminate an area, the power that the illuminated area receives can be written as $T \times P$, where $T \leq 1$ is the transmission of the illumination lens, and P is the power collected by the lens from the light source. By taking an inverse of Eq. (1.39) and multiplying both sides by P, the result is

$$\frac{P}{A\Omega} \geq \frac{P}{A'\Omega'} \geq \frac{TP}{A'\Omega'} \geq \frac{RTP}{A'\Omega'}, \tag{1.40}$$

where $R \leq 1$ is the reflectivity of the area being illuminated. By definition, $P/(A\Omega) = J_0$ is the radiance of the light source, and $RTP/(A'\ \Omega') = J'_0$ can be considered as the radiance of the illuminated area. Then the radiance conservation law is

$$J_0 \geq J'_0. \tag{1.41}$$

Equation (1.41) states that the radiance of an area can never exceed the radiance of the light source that illuminates this area.

Etendue exists beyond geometrical optics; a similar relation also exists in wave optics and Gaussian beam optics, which will be discussed in Chapters 2 and 3.

1.15.7 Radiometry and photometry

Radiometry deals with the detection and measurement of electromagnetic radiation across the entire spectrum. The measured radiation is an absolute power. Photometry is a subfield of radiometry that scales the measured,

absolute radiometric power by the spectral response (photopic curve) of human eyes. (The spectral response of the human eye is plotted later in Fig. 7.22.)

Radiometric quantities, particularly photometric quantities, may appear to be confusing or even odd. Table 1.1 summarizes the correspondence among the most commonly used radiometric quantities for SI units and photometric units.

By definition, 1 watt (W) of radiant power with a wavelength of 555 nm, which is the peak wavelength of the photopic curve, equals 683 lumens (lm). Then, from Fig. 7.22, 1 W radiant approximately equals 223 lm at 500 nm and approximately equals 429 lm at 600 nm.

There are other, less frequently used photometric quantities. For example, candela = lm/sr is also used. However, some of these quantities appear to be already or gradually becoming obsolete.

1.15.8 Blackbody

Any object radiates. The spectrum of the radiance is a function of the temperature described by Planck's law:[14]

$$I(\lambda,T) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda KT}-1}, \tag{1.42}$$

where $I(\lambda, T)$ is the spectral radiance with units of W/(Sr·m^2·$d\lambda$) (watt per solid angle per square meter and per meter wavelength), λ is the wavelength, T is the temperature in kelvin, $h = 6.62606957 \times 10^{-34}$ m^{2}kg/s is the Planck constant, $c = 3 \times 10^8$ m/s is the velocity of light in vacuum, and $K = 1.3806488 \times 10^{-23}$ m^{2}kg s^{-2}K^{-1} is the Boltzmann constant.

Equation (1.42) is plotted in Fig. 1.41 for six values of T from 250 K (−23 °C) to 500 K (227 °C) and in Fig. 1.42 for five values of T from 500 K to 4000 K. Note that the unit used in Figs. 1.41 and 1.42 is watt per solid angle per square meter area per nm spectrum.

Table 1.1 Correspondence among the most commonly used radiometric quantities of SI units and photometric quantities

Radiometric		Photometric	
Quantity	**Units**	**Quantity**	**Units**
Radiant power	W	Luminous flux	lm
Radiant intensity	W/sr	Luminous intensity	lm/sr
Irradiance	W/m^2	Illuminance	lm/m^2
Radiance	W/m^2·sr	Luminance	lm/m^2·sr

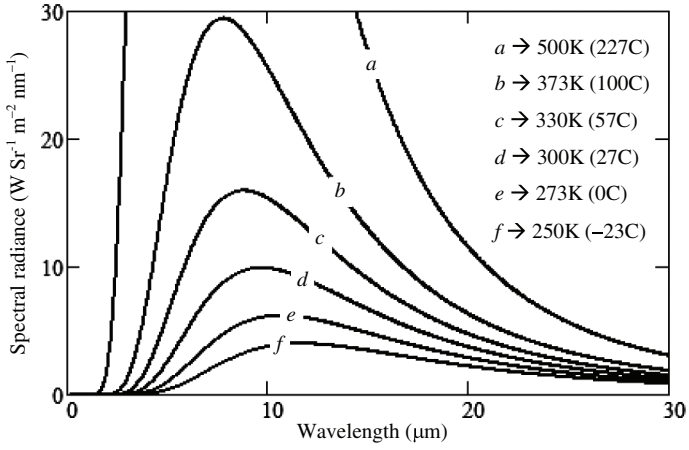

Figure 1.41 Plot of Eq. (1.42), the blackbody radiance spectra for six different temperatures from 250–500 K.

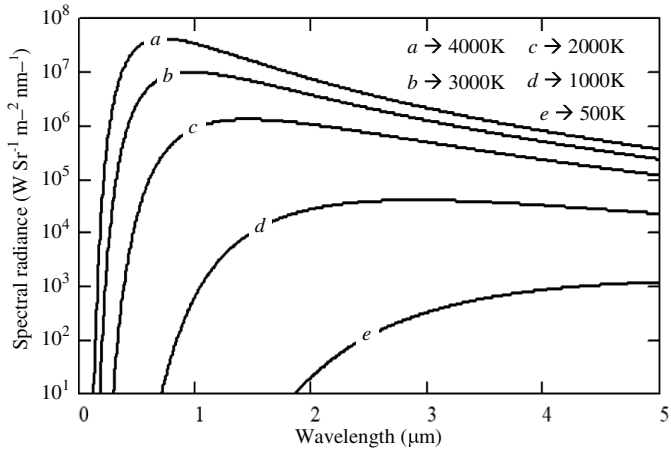

Figure 1.42 Plot of Eq. (1.41), the blackbody radiance spectra for five different temperatures from 500 K to 4000 K.

The fraction of power contained in a certain spectral range can be found by

$$F(T) = \frac{\int_a^b I(\lambda,T)dT}{\int_0^\infty I(\lambda,T)dT}, \tag{1.43}$$

where a and b are the lower and upper limits of the spectral range of interest, respectively. $F(T)$ is plotted in Fig. 1.43(a) for $a = 9$ μm and $b = 15$ μm. This spectral range is best for detecting objects with temperature around $T \approx 320$ K and has the highest fraction power of ~0.36.

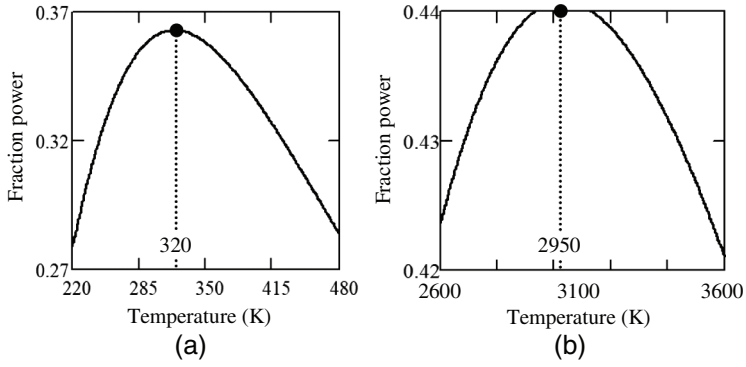

Figure 1.43 (a) The fractional power contained in the 9–15-μm range for temperatures from 220–480 K. (b) The fractional power contained in the 0.9–1.7-μm range for temperatures from 2600–3600 K.

For near-IR hyperspectral imaging applications, the spectral range of interest is about 0.9–1.7 μm. $F(T)$ is plotted in Fig. 1.43(b) for this spectral range. The highest fractional power of ~0.442 appears at $T \approx 2950$ K. Illumination lamps with a color temperature of ~2950 K have the best efficiency.

References

1. W. Smith, *Optical Engineering*, 4th ed., McGraw-Hill Education, New York (2007).
2. E. Hecht, *Optics*, 5th ed., Pearson Education, New York (2016).
3. Wikipedia, "Snell's law," https://en.wikipedia.org/wiki/Snell%27s_law
4. Wikipedia, "Total internal reflection," https://en.wikipedia.org/wiki/Total_internal_reflection
5. Wikipedia, "Paraxial approximation," https://en.wikipedia.org/wiki/Paraxial_approximation
6. Wikipedia, "Cardinal points (optics)," https://en.wikipedia.org/wiki/Cardinal_point_(optics)
7. M. Born and E. Wolf, "V Geometrical Theory of Aberrations," *Principles of Optics*, 7th ed., Cambridge University Press, Cambridge, England, pp. 228–260 (1999).
8. Wikipedia, "1951 USAF Resolution Test Chart," https://en.wikipedia.org/wiki/1951_USAF_resolution_test_chart
9. N. B. Nill, "Conversion Between Sine Wave and Square Wave Spatial Frequency Response of an Imaging System," Eqs. (10) and (11), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.914.2697&rep=rep1&type=pdf (2001)
10. Wikipedia, "Nyquist-Shannon Sampling Theorem," https://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem

11. J. Palmer and B. G. Grant, *The Art of Radiometry*, SPIE Press, Bellingham, WA (2009) [doi: 10.1117/3.798237].
12. R. McCluney, *Introduction to Radiometry and Photometry*, 2nd ed., Artech House Publishers, Norwood, MA (2014).
13. Wikipedia, "Lambert's Cosine's Law," https://en.wikipedia.org/wiki/Lambert%27s_cosine_law
14. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*, 2nd ed., Wiley & Sons, New York, Eqs. (9.2)–(9.18), p. 338 (2007).