Chapter 1
Geometrical Optics

When certain optical phenomena can be explained by geometrical concepts, we are in the realm of what is called geometrical optics. Starting from the idea that light propagates as geometrical rays, and that this propagation is governed by Fermat's principle, we can study the propagation of light in media with a constant or variable refractive index, image formation in instruments comprised of optical elements (e.g., lenses, prisms, and mirrors), optical aberrations that deteriorate such instruments' image quality, the design of optical devices, etc. So, although geometrical optics cannot explain some optical phenomena tied to the wave nature of light, it has a very wide range of applications (Fig. 1.1). Therefore, starting the study of optics from the geometrical point of view is fully justified.

Starting from Fermat's principle, the geometrical properties of optical systems formed by refracting spherical surfaces, spherical mirrors, and lenses

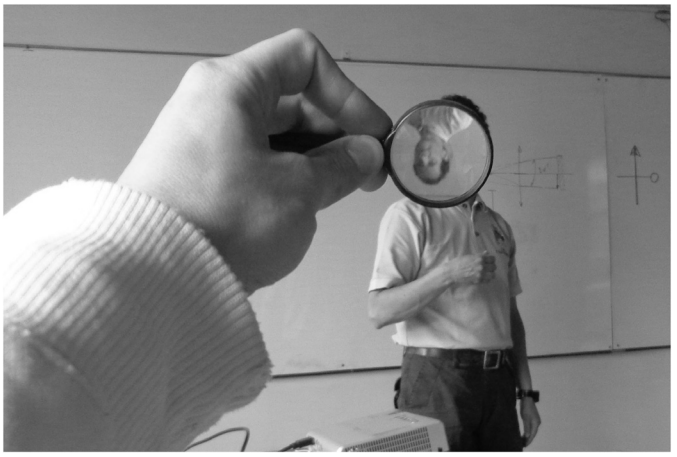


Figure 1.1 Imaging using a positive lens. The image generated by the lens is real and becomes the object for the mobile device's camera that took the photo. We can understand this imaging process from the perspective of geometrical optics. Courtesy of Felix Ernesto Charry Pastrana.

are described in this chapter. While most of the chapter is based on the paraxial approach (small-angle approximation), an introduction to the subject of monochromatic optical aberrations is presented at the end. Chromatic aberration and prisms are briefly discussed in Appendices D and E.

1.1 Rays or Waves

1.1.1 Camera obscura

A camera obscura is basically a closed box with a small circular hole in one of its faces. The function of the hole is to let through light contained in a cone whose vertex is a bright point on the object (object point) and whose base is the hole. The size of the hole determines the angle of the cone.

Suppose that we have a square box of side length L, with a circular hole of radius r_a at the center of one side wall, and that there is an object at distance d from the box (Fig. 1.2). The light that emerges from a point in the object is projected onto the wall opposite the hole as an approximately* circular bright spot of radius $r_a(L+d)/d$. If we define the object as a union of bright points, we will see the superposition of the bright spots on the rear wall inside of the box corresponding to each bright point of the object. If the hole is relatively large, we will essentially see a very blurred (out-of-focus) image of the object on the box's rear wall. But if the size of the hole is gradually reduced, the image will gradually change from out-of-focus to sharp as if the radius of the hole and the angle of the illumination were practically zero. In this limiting case the lighting cone is transformed into what is called a *ray of light*. As a matter of course, theoretically speaking, if the radius of the hole tends to zero, the light beam will be a straight line (carrying no energy). In this limiting (not real) case, we would say that the image is an identical copy of the object,

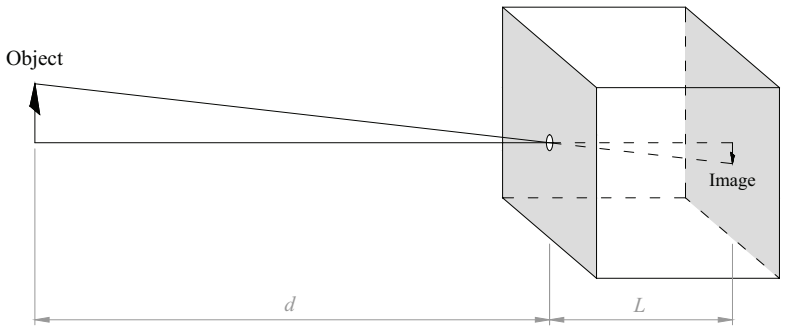


Figure 1.2 Imaging with a camera obscura.

*For an object point on the optical axis, it will be circular; for an object point outside the optical axis, it will be an ellipse.

except for a scale factor. Following the geometry of Fig. 1.2, it is also evident that the image will be inverted (i.e., a right-side-up object results in an upside-down image).

Figure 1.3 shows an example of four images of the same object [shown in (a)]. All the images were taken with a camera obscura [shown in (b)] using the same exposure time, of <1 s, and different aperture sizes (opening diameters) of the iris diaphragm: (c) ≈0.5 mm, (d) ≈1.0 mm, (e) ≈2.0 mm, and

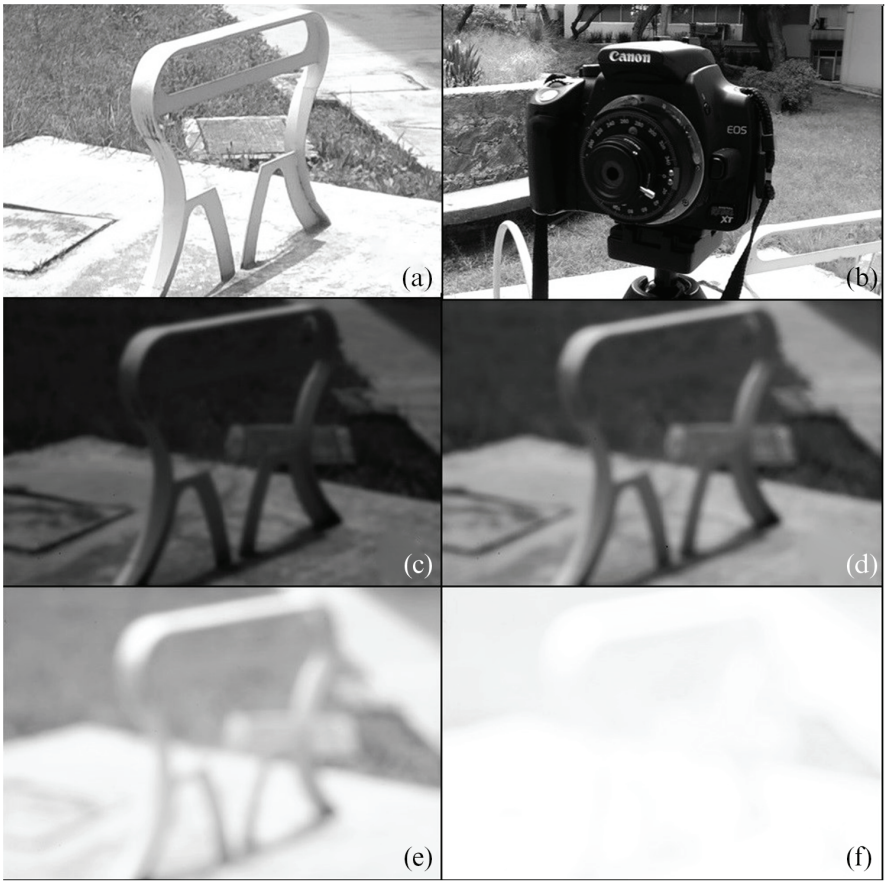


Figure 1.3 Imaging with a camera obscura. (a) Object and (b) camera obscura, created by removing the imaging optical system (camera lens) from a digital camera and replacing it with an iris diaphragm. The images (c–f) were taken with the camera obscura when the diameter of the iris aperture was approximately 0.5, 1.0, 2.0, and 3.0 mm, respectively. In all cases the exposure time was the same (<1 s). With the smaller aperture (c), the image of the object is seen with lower illumination and is slightly out-of-focus. (d) By opening the iris up to 1 mm, the illumination increases but the blur in the image worsens. (e) By further opening the iris diaphragm, up to 2.0 mm, the illumination is noticeably improved but the image is severely out-of-focus. (f) Finally, the iris aperture was increased to 3.0 mm. The resulting image is very bright, but almost all the information about the shape of the object has been lost.

(f) ≈3.0 mm. The camera obscura in this example is a digital camera that has had the imaging optical system (camera lens) removed and replaced with an iris diaphragm (an adjustable opening to allow light to enter and fall upon the image plane).

Alhazen* (965–1040) used a camera obscura (which he himself built) to show that light from objects reaches the human eye traveling in straight lines [1].

In practice, when using a tiny hole, i.e., a very small aperture, in the camera obscura, the wave-like nature of light is made evident by diffraction, which produces a spot in the image that is larger than what would be obtained from geometrical projection. We will deal with the phenomenon of diffraction at the end of the book. For now, we are going to accept that with the use of a camera obscura, it can be observed that light propagates as rays, which follow straight lines in homogeneous media (of constant refractive index).

1.1.2 Newton's corpuscular theory of light

Newton (1642–1727) posits that light travels in straight lines and is made up of small particles that follow the laws of Newtonian mechanics [2]. Newton explains the laws of reflection and refraction of light using this idea. In a homogeneous medium, light particles travel in a straight line and with constant speed. When the light particles hit a reflecting surface, they experience an elastic collision, maintaining the tangential velocity and changing the direction of the velocity orthogonal to the surface. This is outlined in Fig. 1.4, which shows that after the collision, only the y direction component of the velocity changes; the modulus of the velocity vector is

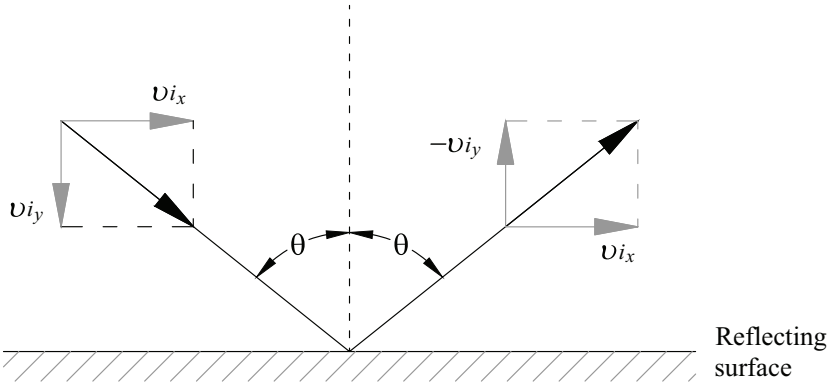


Figure 1.4 The law of reflection according to Newton.

*Alhazen is the Latin form of the first name of Al-Hasan Ibn al-Haytham, who was born in 965 in what is now known as Basra, present-day Iraq. Alhazen lived during the Islamic Golden Age, a period when the Muslim world flourished as a global center of great scientific discovery and economic and cultural advances.

maintained, and the angle with which the light ray is reflected is equal to the angle with which the light ray hits.

On the other hand, in the case of refraction, Newton maintains that the interface that separates two media of different refractive indices (with the refractive index of the incident medium being less than the refractive index of the transmission medium) exerts an attraction on the light particles, thus increasing the normal component of velocity while the tangential component remains unchanged, as shown in Fig. 1.5. This implies that the modulus of the velocity vector increases when light passes from a medium with a lower refractive index to a medium with a higher refractive index, contrary to what actually happens.

1.1.3 Huygens' wave theory

Newton's corpuscular theory is not only wrong about the change in velocity when light is refracted, but it also cannot explain interference phenomena. A different proposition was put forth by Huygens, known as Huygens' principle.

Before stating this principle, let us qualitatively define what a wavefront is by using mechanical waves on the surface of water. When a drop falls on the surface of a water pool that is at rest, we can observe a series of ripples (crests and troughs) on the surface that propagate away from the place where the drop fell. The shape of the crests (and troughs) are circles [Fig 1.6(a)]. Ideally, the points at the peak, or highest point, of one of the crests will be at the same height. We will say that all the points of a peak (and valley) that are at the same height describe a profile that is called a wavefront.

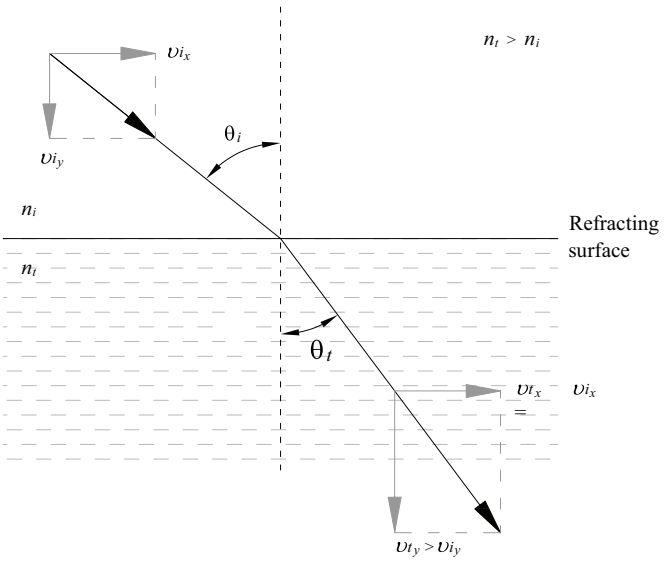


Figure 1.5 Newton's law of refraction.

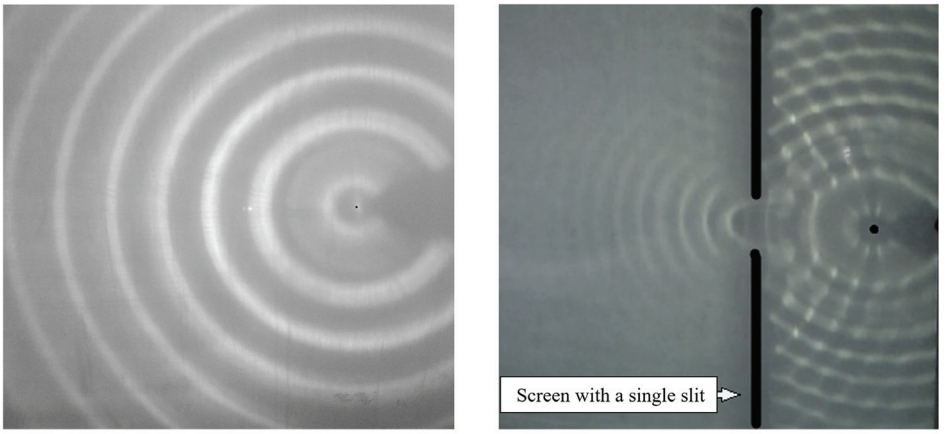


Figure 1.6 Waves in a water bucket. (a) Circular waves and (b) secondary circular waves in an aperture (diffraction).

So, in the example of Fig. 1.6(a), the drop is equivalent to the source that originates the circular waves (spherical in the three-dimensional case). In a homogeneous medium, circular wavefronts propagate with the same speed. As the wavefront moves away from the source, its radius increases, and for large distances, the wavefront in the vicinity of a point will resemble a flat wavefront.

Suppose that a wavefront is obstructed, allowing only a small portion to pass through, as shown in Fig. 1.6(b). What happens is that when crossing the opening, circular wavefronts are generated again that propagate with the same speed of the original wavefront. This fact, which is due to wave diffraction, is the central idea of *Huygens' principle* and can be stated as follows [3]:

> Each point on a wavefront can be thought of as a source of secondary spherical waves that propagate with the same speed as the wavefront. After a while, the propagated wavefront will be the envelope of the secondary spherical waves.

This is illustrated in Fig. 1.7, which recreates the propagation of a wavefront from the secondary Huygens' waves. Huygens' principle is applicable to the propagation of light and allows us to correctly derive the laws of reflection and refraction. In refraction, as shown in Fig. 1.5, the speed of light in the medium with the highest refractive index is less than the speed of light in the medium with the lowest refractive index (contrary to Newton's assumption).

Finally, let us see how the law of refraction is derived from Huygens' principle. A wavefront $S-S'$, incident with an inclination angle θ at an interface that separates two homogeneous media of refractive indices n (incident medium) and n' (refracting medium), is shown in Fig. 1.8(a). When the wavefront reaches point a, secondary waves are generated that propagate

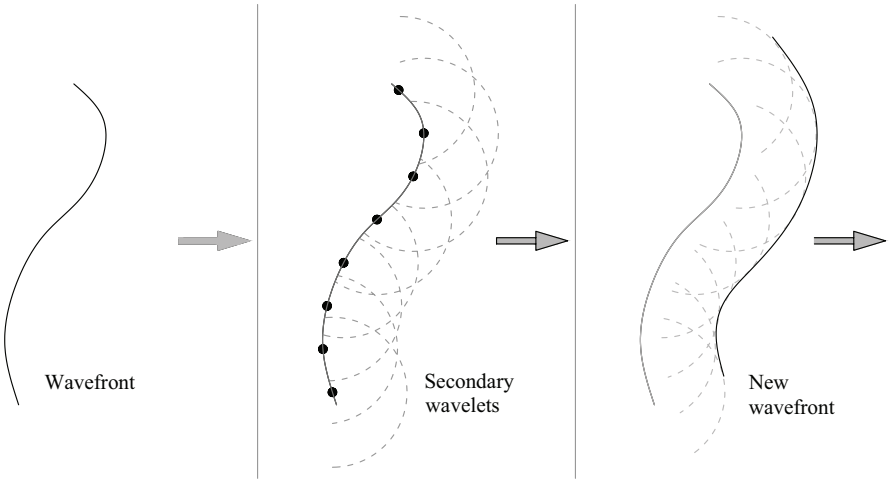


Figure 1.7 Propagating wavefront reconstruction using Huygens' principle.

with speeds υ (in the incident medium) and υ' (refracting medium). At a later time t the wavefront reaches point b, and the two aforementioned secondary waves, emitted at a, will respectively have the radii $\overline{ae} = \upsilon t$ and $\overline{ad} = \upsilon' t$. The speed of light in a medium of refractive index n is defined as $\upsilon = c/n$, where c is the speed of light in vacuum. Therefore, if $n < n'$, then $\upsilon > \upsilon'$, and the case just described will look as presented in Fig. 1.8(a).

According to Huygens' principle, in the time t, the envelope for the reflected wave will be the plane wavefront $\overline{be}$ and the envelope for the refracted wave will be the plane wavefront $\overline{bd}$. For refraction, we have that $\sin\theta' = \overline{ad}/\overline{ab}$ and $\sin\theta = \overline{cb}/\overline{ab}$. Since $\overline{cb} = \overline{ae}$, we get to

$$\frac{\sin\theta'}{\sin\theta} = \frac{\upsilon' t}{\upsilon t}. \tag{1.1}$$

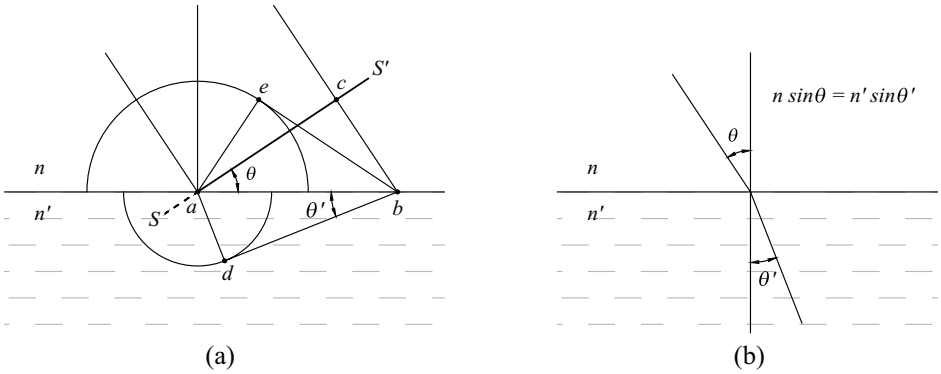


Figure 1.8 Law of refraction according to Huygens' principle.

Finally, using the refractive index definition, we arrive at the *law of refraction*, also known as Snell's law,[*]

$$n \sin \theta = n' \sin \theta', \tag{1.2}$$

which indicates the way in which the inclination of the refracted wavefront changes (angle θ').

For a given medium, the refractive index n is a function of the wavelength of light in vacuum.[**] Therefore, when providing the value of the refractive index, we should also specify the wavelength to which that value corresponds. However, if no wavelength is specified, standard practice in optics is to give a value that corresponds to the refractive index at the spectral Fraunhofer line at 587.56 nm (i.e., a yellow light close to the maximun sensitivity of the eye; see Appendix C).

If instead of observing the wavefront's inclination we look at the wavefront's propagation directions, then Snell's law can also be established, as shown in Fig. 1.8(b). In fact, if we return to the interpretation of light rays, Snell's law can also be stated by saying that if a light ray in the medium with refractive index n hits an interface with an angle θ (measured with respect to the normal of the interface at the point where the light ray hits), the refracted light ray in the medium with refractive index n' will form an angle θ' with respect to the normal. So we have two concepts: the light ray and the wavefront. Geometrically speaking, both concepts are related; the light rays are orthogonal to the wavefronts. Thus, using any of these two representations is equivalent, as long as we consider the propagation of light as a geometrical matter.

In 1678, with knowledge of Snell's works, Huygens (in *Traité de la Lumière*) shows that with his ideas of secondary waves, Snell's law is fulfilled. As we will see later, this law can also be derived from other physical principles.

1.1.4 Graphical ray tracing

Snell's law has a graphical interpretation that enables the tracing of light rays, or simply *rays*, on any refracting surface in a simple (and exact) way.

In Fig. 1.9(a), we have the refraction of an incident ray. Suppose we construct a triangle, as shown in Fig. 1.9(b), as follows. A side of length will be equal to the refractive index n when extending the incident ray from the incidence point A. The point at the end of this side will be B. A side of length equal to the refractive index n' drawn from the incidence point A will be in the direction of the refracted ray. The point at the end of this side will be C. The

[*]Named in honor of Willebrord Snellius (1580–1626), who was a professor of mathematics at the University of Leiden.

[**]The refractive index in dielectric materials is discussed in Appendix B.

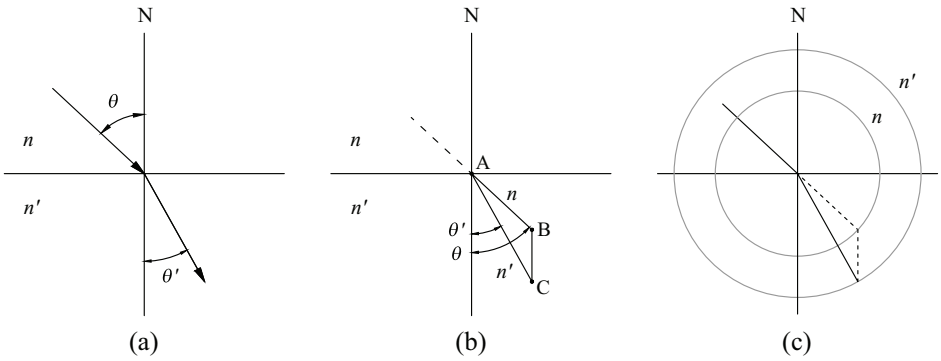


Figure 1.9 Graphical ray tracing according to Snell's law.

third side will be the line that joins the points B and C. This triangle has the property that the side BC is parallel to the normal line N.

Triangle ABC in Fig. 1.9(b) fulfills Snell's law. This results directly from the law of sines for the triangle ABC, as follows:

$$\frac{\overline{AB}}{\sin\theta'} = \frac{\overline{AC}}{\sin(\pi - \theta)} \tag{1.3}$$

or

$$\frac{n}{\sin\theta'} = \frac{n'}{\sin\theta}, \tag{1.4}$$

which is Snell's law.

This method enables the development of a technique for the graphical ray tracing of the refraction phenomena, as illustrated by Fig. 1.9(c), which can be stated as follows:

> Consider a ray that hits a surface that separates two media: an incidence medium with refractive index n and a refracting media with refractive index n'. The refracted ray can be obtained as follows. Determine the normal to the surface at the incidence point, and then draw two circles of radii n and n' with the center at the incidence point. Extend the incident ray to the circle of radius n and use the intersecting point between the extended ray and that circle as a starting point to draw a line segment parallel to the normal. This new line intersects the circle of radius n' at another point that indicates the direction in which the refracted ray should be traced from the point of incidence.

This technique works on flat and curved surfaces, given that Snell's law is applied locally. For example, Fig. 1.10 shows two examples with curved

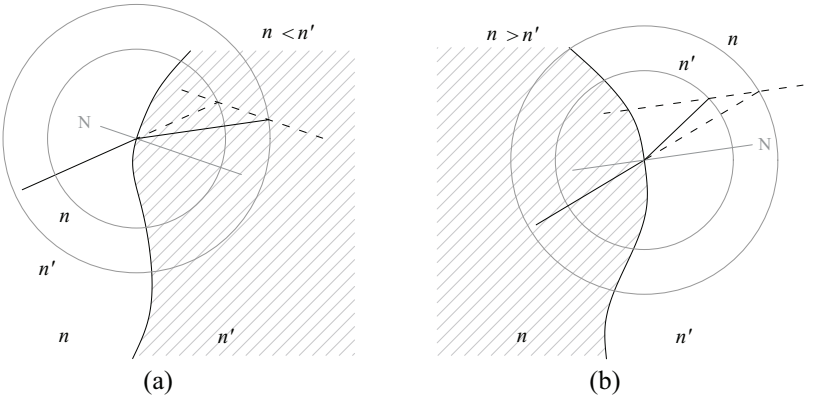


Figure 1.10 Graphical ray tracing on curved surfaces.

surfaces; in (a), $n < n'$, and in (b), $n > n'$. Note the change in the extension of the incident ray in (b), which now goes to the circle with the largest radius. In the latter case, a relevant situation arises when the line parallel to the normal N touches the smaller circle tangentially. The refracted ray forms an angle of $\pi/2$; i.e., the ray is tangent to the interface. In Fig. 1.11, this particular case is illustrated. The angle of the incident ray for which this occurs is denoted by θ_c and is called the *critical angle*.

At $\theta \geq \theta_c$, the phenomenon of total internal reflection occurs, which we will see in detail in Section 2.4.1. From the triangle ABC in Fig. 1.11,

$$\sin \theta_c = \frac{n'}{n}. \tag{1.5}$$

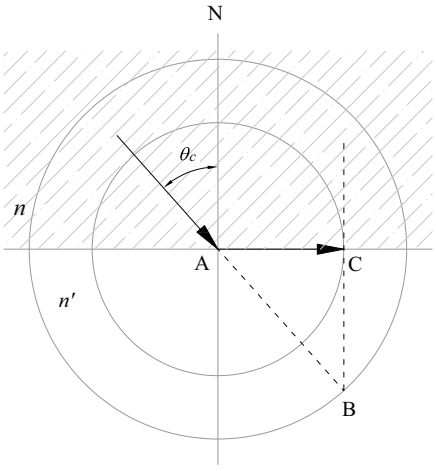


Figure 1.11 When the line parallel to the normal N touches the circle of radius n' tangentially, there is total internal reflection.

1.2 Fermat's Principle

In his treatise on geometry, Euclid (325–265 BCE) postulates that "a straight line segment can be drawn joining any two points." In Euclidean space, the shortest distance between two points is the length of the straight line segment that joins them. Following this postulate, Hero of Alexandria (10–70 BCE) establishes that *for light to go from one point to another, it follows the shortest geometrical path*. This is another way of saying that light propagates as geometrical rays. Given that in a homogeneous medium light propagates with constant speed, it can also be said that *for light to go from one point to another, it follows the path for which the least time is used*. For a homogeneous medium, the two statements are equivalent. However, if the two points are in different media (both homogeneous), the result is that light no longer follows the shortest geometrical path.

In Fig. 1.12(a), three possible light paths are illustrated. A planar interface separates the media with refractive indices n and n'. The shortest geometrical path is the straight line PT_2Q, but it does not correspond to the path that light follows if $n' \neq n$. The path followed by light resembles lines PT_1Q when $n' < n$ and the lines PT_3Q when $n' > n$. So how can the real path followed by light be obtained when $n' \neq n$? The answer to this question is obtained by using the statement above referring to the path for which the least time is used. This statement was formulated by Fermat (1601–1665) and is known as *Fermat's principle*.

To apply Fermat's principle to the problem illustrated in Fig. 1.12(a), consider the geometry of Fig. 1.12(b). The interface is located at $y = 0$ and the point T at x in the horizontal axis. The coordinates of point P are $(0, h)$ and those of point Q are $(a, -b)$. The time for light to go from P to Q through T will be

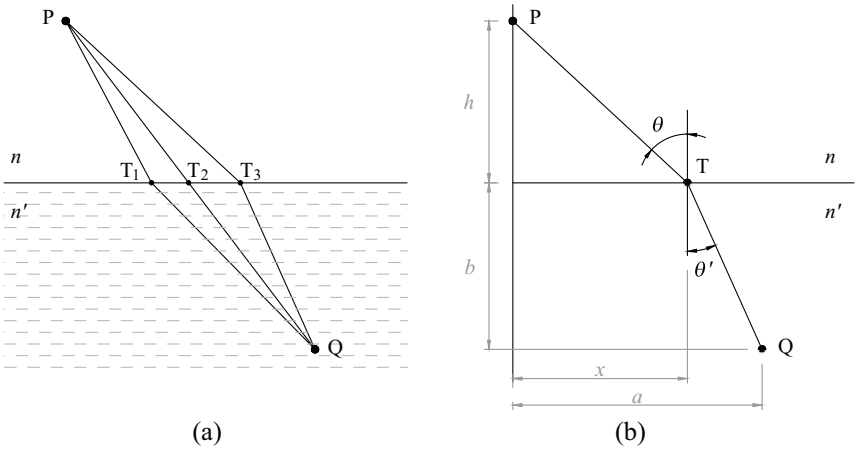


Figure 1.12 Sketch to derive Snell's law from Fermat's principle. (a) Three possible trajectories to go from P to Q. (b) Geometry to calculate the real trajectory.

$$t = \frac{n}{c}\overline{\mathrm{PT}} + \frac{n'}{c}\overline{\mathrm{TQ}}, \tag{1.6}$$

i.e.,

$$t = \frac{n}{c}\sqrt{h^2 + x^2} + \frac{n'}{c}\sqrt{b^2 + (a-x)^2}. \tag{1.7}$$

Fermat's principle states that t must be as small as possible, which is equivalent to finding the value of x for which the right side of Eq. (1.7) is the smallest. In other words, we should differentiate with respect to x and set the first derivative equal to zero, i.e.,

$$\frac{n}{c}\frac{x}{\sqrt{h^2 + x^2}} - \frac{n'}{c}\frac{(a-x)}{\sqrt{b^2 + (a-x)^2}} = 0, \tag{1.8}$$

from which the Snell's law $n' \sin\theta' = n \sin\theta$ [Eq. (1.2)] is derived.

For the reflection case, as in a flat mirror, the geometry of the problem is shown in Fig. 1.13(a). The angle between the incident ray and the normal line at T is θ, and the angle between the reflected ray and the normal line at T is θ'. By applying Fermat's principle to Fig. 1.13(a), we obtain equations equal to those we had in the case of refraction but now with $n' = n$, so Eq. (1.8) remains $\sin\theta' - \sin\theta = 0$. Therefore, the reflected ray has an inclination angle equal to the inclination angle of the incident ray; i.e., $\theta' = \theta$. In Fig. 1.13(b), the virtual image P′ is included. The distance from P′ to the mirror is equal to the distance from P to the mirror. An observer at Q will see the light coming from P′ and not from P. Thus, in effect, the point T is at the intersection of the

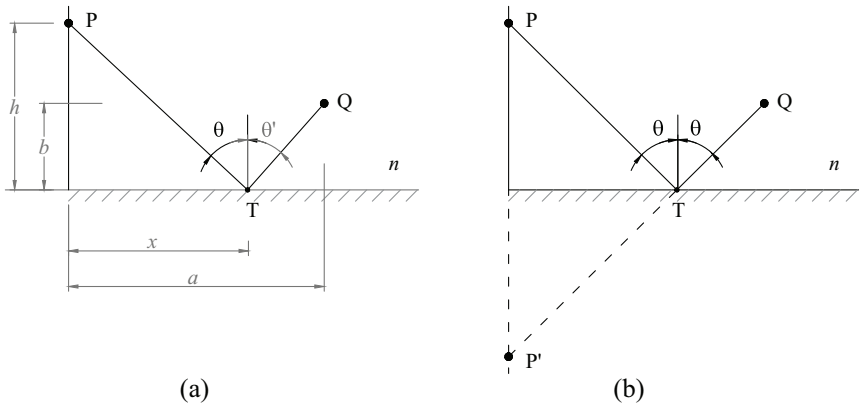


Figure 1.13 Sketch to derive the reflection law from Fermat's principle. (a) Geometry to calculate the path from P to Q through T. (b) T is at the intersection of the line joining P′ with Q. The distance from P′ to the mirror is equal to the distance from P to the mirror.

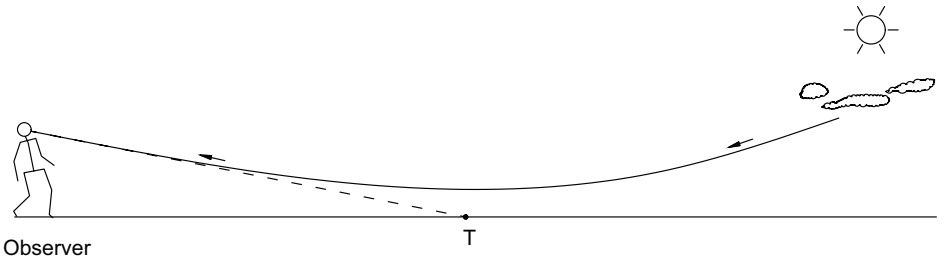


Figure 1.14 A mirage. An observer sees something that is not really at the point T. What the observer sees is light from the sky.

straight line P′Q, so that for the virtual source, Fermat's principle is also satisfied.

An interesting example that is explained by Fermat's principle is the mirage, as illustrated in Fig. 1.14. Suppose we are standing on a straight road on a plain with a clear sky. Due to the heat, the air that is closer to the road's surface will have a lower density than the air that is higher. Consequently, the refractive index varies with height and is lower near the road's surface. If we now direct our gaze toward a point T on the road, the light that arrives in that direction is not the one that comes out of point T; it is light from the sky. Thus, we see it as if the road in front of us is wet or as if there is a mirror. What is happening is that the light is avoiding the areas of higher refractive index, looking for the minimum time trajectory, and some of the rays coming from the sky are reaching our eyes when we look in the indicated direction.

A situation in which there are infinite trajectories of the same time is found in an ellipsoidal mirror (of rotational ellipsoid surface), which has the property of forming the image of a point source located in one of the foci in the other focus. Consider a cross section of the ellipsoid containing the axis of revolution, as shown in Fig. 1.15. Suppose that inside the mirror we have air ($n = 1$) and that a point source is located at the focus P from which rays

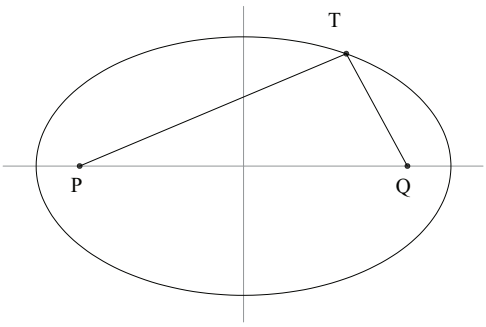


Figure 1.15 Elliptical mirror. The light emitted by a point source at P will arrive after being reflected in the mirror at the point Q following any of the infinite trajectories that satisfy the equation of the ellipse.

diverge in a radial direction. Any ray will follow a straight path until it reaches the mirror at a point T and will be reflected again following a straight path through point Q, which is at the other focus of the ellipse. By changing the position of T, the reflected light ray will pass through Q again. In both cases, the time is the same, and this occurs for any other point T on the surface of the mirror. So none of the trajectories is a minimum. This indicates that the original formulation of Fermat's principle does not cover all cases.

1.2.1 Modern formulation of Fermat's principle

If we multiply Eq. (1.6) by the speed of light, $ct = n\overline{\mathrm{PT}} + n'\overline{\mathrm{TQ}}$, we have an equality of two quantities of different nature. On the left side, we have the distance that light would travel in vacuum during the time t; on the right side, we have a distance that differs from the geometrical distance, i.e., the geometrical distance multiplied by the refractive index. This last distance is called the *optical path length* (*OPL*). Taking this concept into account, Fermat's principle in its modern form states that:

> A ray of light goes from one point to another following a path whose *OPL* is stationary with respect to the variations of such path.

In this statement, "stationary" means that the *OPL* can be a minimum value, a maximum value, or an inflection point with a horizontal slope. In this way, all possible light trajectories are included. For example, in the ellipsoidal mirror of Fig. 1.15, the ray trajectories correspond to the inflection point of Fermat's principle.

So, in a general way, suppose that we have a path C of length s that joins two points, P_1 and P_2, that are in an inhomogeneous medium, as shown in Fig. 1.16. The refractive index will be a function of the spatial coordinates, $n = n(x, y, z)$, and the *OPL* will be given by

$$OPL = \int_C n(s)ds, \tag{1.9}$$

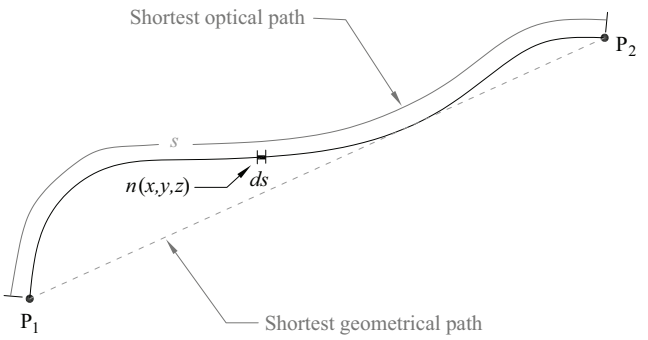


Figure 1.16 *OPL* in an inhomogeneous medium.

where $ds = \sqrt{dx^2 + dy^2 + dz^2}$. If the OPL satisfies

$$\frac{\partial}{\partial s}(OPL) = 0, \tag{1.10}$$

then C is the path that the light will follow. In Fig. 1.16, although the path length of the resulting shortest optical path will be greater than the length of the shortest geometrical path (segment $\overline{P_1P_2}$), the length of the optical path along C is less than the length of the optical path along $\overline{P_1P_2}$.

1.2.2 Rays and wavefronts

Taking into account the above, from the geometrical point of view, a ray of light is the curve that satisfies Fermat's principle. If the medium is homogeneous, $n = n_0$ (constant), then

$$OPL = n_0 \int_C ds = n_0 \overline{P_1P_2},$$

since the shortest path between two points (in Euclidean space) is a straight line segment. Therefore, the rays in a homogeneous medium are straight lines, whereas in an inhomogeneous medium, they will be curved.

For example, in Fig. 1.17(a) a possible configuration for some rays emitted by a point source immersed in a medium of variable refractive index is shown. Now consider a curve orthogonal to the rays. If the distance between the point source and each point intersected by the rays on the orthogonal curve is such that the OPL is the same, the curve is called a *geometrical*

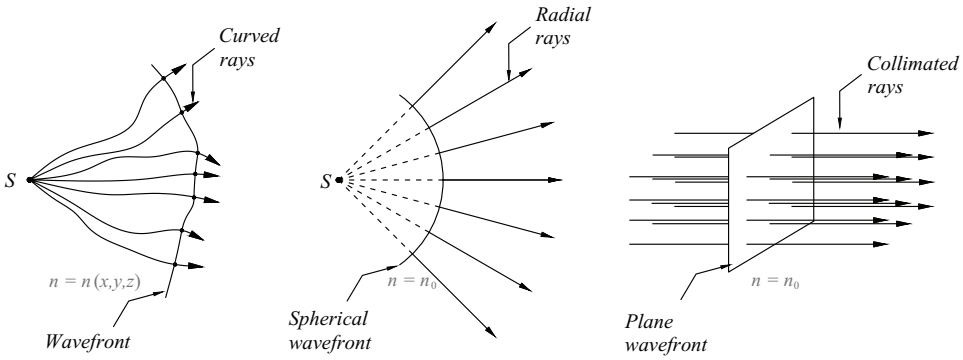


Figure 1.17 Rays and wavefronts are orthogonal to each other. (a) Point source in an inhomogeneous medium; the rays are curved and the wavefront is a distorted surface. (b) Point source in a homogeneous medium; the rays are radial lines and the wavefront is a sphere. (c) At an infinite distance from the point source in a homogeneous medium, the rays are parallel lines (collimated) and the wavefront is a plane.

wavefront. Of course, for another OPL we will have another wavefront. So what we have is a family of rays and a family of wavefronts orthogonal to each other. In the three-dimensional case, the rays will not be limited to one plane and the wavefronts will be surfaces orthogonal to the rays. In the case in which the refractive index is constant, the rays will be radial lines and the wavefronts will be spheres [Fig. 1.17(b)]. If we observe the wavefronts at a far distance from the source that tends to infinity, the wavefronts will be planes and the corresponding rays will be lines parallel to each other [Fig. 1.17(c)]. These rays are called *collimated rays*.

From a mathematical point of view, if $\Phi(x, y, z)$ represents the wavefront, then $\nabla\Phi$ represents the rays and

$$\hat{\mathbf{t}} = \frac{\nabla\Phi}{\|\nabla\Phi\|} \tag{1.11}$$

gives the direction of the ray. Therefore, to refer to the propagation of light we can do it with either of the two representations: rays or wavefronts. Given one of the representations we can move on to the other with Eq. (1.11).

In Section 1.1.3, we qualitatively define the wavefront from the wave motion of the peaks and valleys on the surface of the water. Formally, in waves, the wavefront is defined as the surface where each point has the same phase value. On the other hand, in Fermat's principle, the wavefront is defined as the surface where each point has the same OPL value. The two definitions are equivalent. As we will see later, the phase can be obtained from the OPL.

1.2.3 Image from a point source

Suppose that we have a point source S submerged in a homogeneous medium from which rays diverge, as illustrated in Fig. 1.17(b). Given that the refraction of light at an interface implies that light rays change their direction according to Snell's law, then it is possible to design a surface or a set of refracting surfaces (optical system) with which the rays can rejoin at a point S′ (also in a homogeneous medium). If the OPL is the same for all the rays that leave point S, refracting in the optical system and reaching point S′, then we will say that S′ is the image of S. This implies that the wavefront converging on S′ must also be spherical, as shown in Fig. 1.18.

Figure 1.18 describes the formation of the image of a point object S into a point image S′. But the description can also be made in the reverse sense; i.e., if S′ is the object, then S will be the image. S and S′ are then said to be conjugate points.

If the optical system is such that the rays leaving S do not reach S′ with the same OPL, then we will not have the image of a point but a spot of a certain extent. This is explained by taking into account that a point image implies a

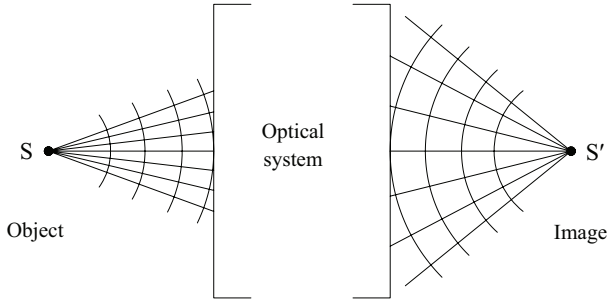


Figure 1.18 Image formation of a point source.

convergent spherical wavefront, and if the rays arriving or approaching the ideal position of S′ do not have the same optical path, the convergent wavefront will not be spherical, so it is not possible to have a point image. In this case, the optical system is said to introduce aberrations that distort the image. We will address this topic in a general way at the end of this chapter. The question that interests us now is to determine the shape that refracting surfaces should have to generate a point image of a point object.

1.3 Refracting Surfaces

In Fig. 1.18, the optical system receives at the input a bundle of divergent rays emerging from a point object (point source) and outputs a bundle of rays converging toward a point image. How should the optical system be set up to perform this task? To answer this question, we can start with the simplest case: image formation by a refracting surface. The shape of the surface can be determined from Fermat's principle.

In Fig. 1.19, the imaging of a point object P is shown using a refracting surface that separates two media of refractive indices n and n' ($n' > n$). The length of any incident ray is denoted by l, and the length of the corresponding refracted ray is denoted by l'. The OPL for any ray should be the same; therefore,

$$OPL = nl + n'l' = \text{constant}. \tag{1.12}$$

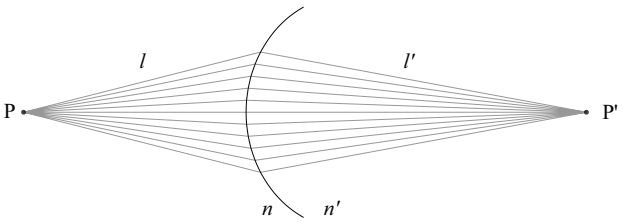


Figure 1.19 The shape of the refracting surface that enables a point image P′ to be obtained from a point object P is called a Cartesian oval.

The surface obtained by solving Eq. (1.12) is a surface of revolution called *Cartesian oval*. The axis of revolution coincides with the straight line $\overline{PP'}$. The intersection of that line with the surface defines the surface vertex. The surface center of curvature at the vertex lies on the axis of revolution.

Therefore, with the Cartesian oval, we form the image of P in P′, as shown in Fig. 1.19. In other cases we may need more than one surface, for example, if we want the image to be in the same medium as the object. In any case, in the rest of this chapter, let us limit ourselves to optical systems whose refracting surfaces have symmetry of revolution. In addition, all the axes of revolution will be coincident; i.e., we will have a single axis of revolution that is called the optical axis of the system. So, the optical axis should contain all the vertices and centers of curvature (at the vertices) of the refracting surfaces.

A simple example of a Cartesian oval is obtained when we want to produce, at a distance b from the refracting surface vertex V, a point image of a point object located at an infinite distance from V. The incident rays (coming from infinity) can be assumed to be collimated and parallel to the optical axis, as shown in Fig. 1.20. Consider a plane orthogonal to the optical axis that passes through the point A. The incident rays at that plane have the same OPL from the object point. Therefore, we can take this plane as a reference to measure the OPL of a ray passing through a point P, going through the point Q on the refracting surface Σ, and arriving at the image point P′. If the point P′ is the image of a point object that is at infinity, all the rays exiting the plane at A and converging to P′ should have the same OPL. In particular, for the ray coming out of P,

$$OPL = n\overline{PQ} + n'\rho = na + n'b, \tag{1.13}$$

where a is the distance from A to V. Note that the right side of Eq. (1.13) is the OPL along the optical axis. Since $\overline{PQ} = a + b - \rho\cos\theta$,

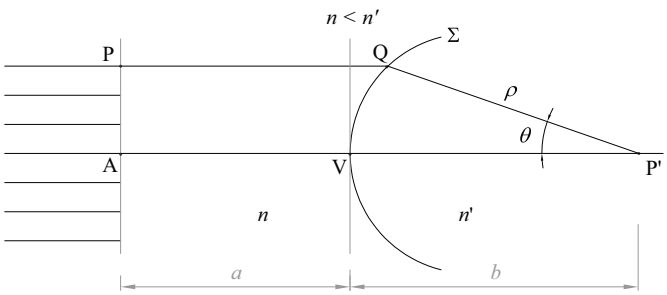


Figure 1.20 The surface shape that generates a point image of a point object that is at infinity is an ellipsoid of revolution when $n < n'$.

$$n'\rho\left(1-\frac{n}{n'}\cos\theta\right)=b(n'-n). \tag{1.14}$$

Defining a new constant $p=b(n'-n)/n'$, we get

$$\rho(1-e\cos\theta)=p, \tag{1.15}$$

which describes a conical surface with eccentricity $e=n/n'$. Since $n'>n$, then $e<1$, corresponding to an ellipse. Therefore, the refracting surface Σ, which focuses the collimated rays parallel to the optical axis at a point image, is an ellipsoid of revolution.

1.3.1 Modeling the cornea of a human eye

An example where the refracting surface resembles an ellipsoid of revolution is the anterior corneal surface of the human eye. The human eye is a biological–optical system that focuses light rays on the concave surface of the retina. Anatomically, the eyeball of a normal adult eye is an approximately spherical oval of about 23 mm in diameter. Light enters the eye first through the cornea, a transparent tissue on the front of the eye (Fig. 1.21). The average refractive index of the cornea is 1.376. In an emmetropic eye, the cornea has an anterior surface with a mean radius of 7.8 mm in the central area (around the vertex), and as it moves away from the center, the radius increases as the cornea overlaps with the sclera (the outer layer covering the eyeball). This makes the shape of the anterior corneal surface resemble an ellipsoid. Various experimental studies [4] have shown that most normal eye corneas can be characterized by an ellipsoidal model with an eccentricity of 0.5 and a vertex offset toward the temporal zone of approximately 0.4 mm (measured with respect to the optical axis).

The posterior surface of the cornea, with a mean radius of 6.5 mm, is separated from the anterior surface by approximately 0.6 mm. Right behind

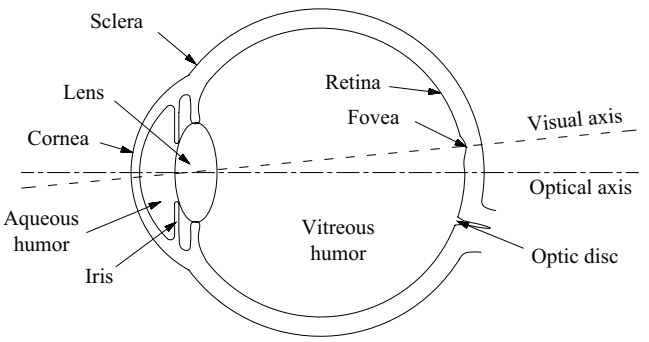


Figure 1.21 Schematic diagram of the human eye. In emmetropic eyes, the anterior surface of the cornea resembles an ellipsoid of revolution.

the cornea is the anterior chamber, which contains aqueous humor with a refractive index of 1.336. Then there is the iris, which controls the amount of light that enters the eye, varying its internal diameter from 2 mm (for high-light objects) to 8 mm (for low-light objects). Behind the iris is the crystalline lens, shaped like a biconvex lens of approximately 9 mm in diameter and 4 mm thickness. The refractive index of the lens varies, from around 1.406 in the inner core to around 1.386 in the outer zones. The crystalline lens can vary its shape to achieve a fine focus, so that the light coming from any external object is focused onto the retinal surface. Behind the crystalline lens there is another chamber with a transparent substance, called vitreous humor, with a refractive index of 1.337. Ultimately, the light is focused onto the retina, a concave surface that contains two classes of photoreceptor cells, the rods and the cones. A detailed description of the anatomy and function of the eye can be found in the works of Helmholtz [5], Le Grand and El Hage [6], Davson [7], and Smith and Atchison [8].

Due to the complexity of the human eye (i.e., the visual axis does not coincide with the optical axis; Fig. 1.21), we usually work in optometry and ophthalmology with simplified models of the human eye that enable us to analyze image formation in the retina. There are several models, some of which include an ellipsoidal model of the cornea [9]. For example, the human eye could be thought of as an optical system with a single refracting surface (the cornea), with an equivalent refractive index n_{eq}, as shown in Fig. 1.22. Rays coming from very distant object points are focused on the same point on the retina, if the shape of the cornea is assumed to be ellipsoidal, according to Eq. (1.15). More simplified models use a sphere, instead of an ellipsoid, to represent the corneal shape. Such simplification is adequate if the bundle of rays that comes from infinity is limited to a region very close to the optical axis (in which we can model a sphere as a second-order surface). If the beam of rays is not limited to a small region, Fermat's principle would not hold for all rays and the image would no longer be a point but a blur spot.

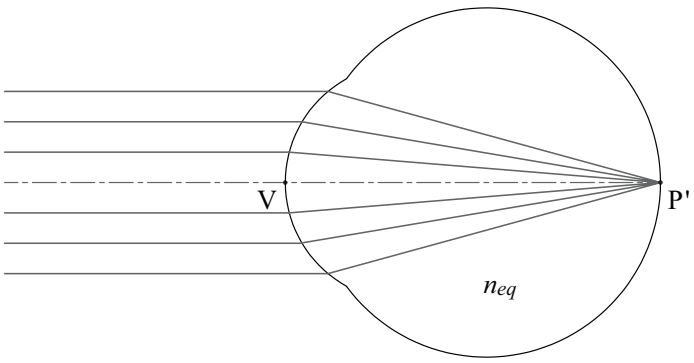


Figure 1.22 Simplified model of the human eye. The eye is modeled as an ellipsoidal refracting surface with an equivalent refractive index n_{eq}.

1.3.2 Refraction at spherical surfaces

Although with Cartesian ovals we have optical systems to form ideal images (point images of point objects), in most image-forming systems, they are impractical because the ideal image is obtained for a fixed position of the object. If the position of the object (along the optical axis) is changed, an ideal image will no longer be formed. On the other hand, spherical surfaces are easy to manufacture and high-quality images can be obtained with them, for different object positions, by combining several spherical refracting surfaces and/or by limiting the transversal extension of a ray bundle to the region close to the optical axis.

To see how the light rays behave on a refractory spherical surface, let us calculate the refraction of the rays according to Fermat's principle based on Fig. 1.23.

In Fig. 1.23, some quantities appear with a negative sign. This is because it uses the Cartesian sign convention, which will be used hereafter. Thus,

- The distances to the left of the vertex (or to the left of the point Q) are negative.
- The distances to the right of the vertex (or to the left of the point Q) are positive.
- The angles with respect to the optical axis (or with respect to the normal N) are negative if they turn clockwise.
- The angles with respect to the optical axis (or with respect to the normal N) are positive if they turn against the clockwise direction.
- The radius of curvature of the refracting surface is negative if the center of curvature is to the left of the vertex.
- The radius of curvature of the refracting surface is positive if the center of curvature is to the right of the vertex.
- The height of an object point outside the optical axis (i.e., the distance from the optical axis) is negative if the point is below the optical axis.

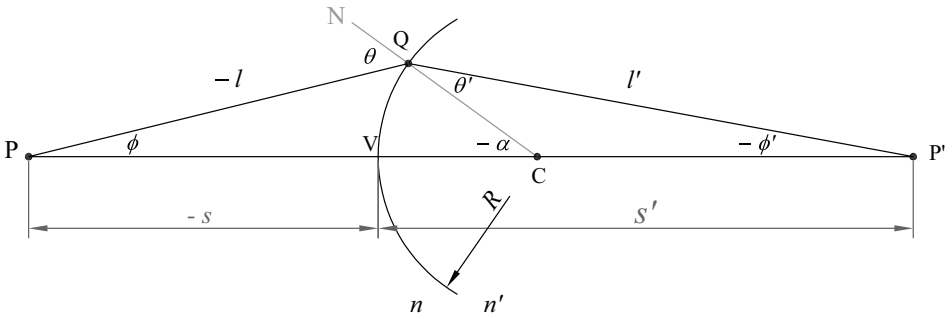


Figure 1.23 Geometry to calculate the refraction of light rays at a spherical surface.

- The height of an object point outside the optical axis (i.e., the distance from the optical axis) is positive if the point is above the optical axis.

Additionally, it is assumed that light travels from left to right and that all rays are contained in the same plane (defined by the points P, Q, and P′).

The sign convention in optical systems varies by author. The convention used in this book is the one usually employed in optical design. In general, the problem with having different sign conventions becomes apparent when some signs in the equations connecting object distances and image distances change.

According to the Cartesian sign convention, in Fig. 1.23 all distances and angles are drawn as positive. For example, P is to the left of V and therefore $s < 0$; in the drawing we put $-s$ to indicate the length separation between P and V. Then, the OPL for the ray PQP′ is

$$OPL = -nl + n'l'. \tag{1.16}$$

Using the triangles PQC and P′QC, we get

$$-l = \sqrt{R^2 + (-s+R)^2 - 2R(-s+R)\cos(-\alpha)} \tag{1.17}$$

and

$$l' = \sqrt{R^2 + (s'-R)^2 - 2R(s'-R)\cos(\pi+\alpha)}. \tag{1.18}$$

Since α is the only variable on the right side of Eqs. (1.17) and (1.18), Fermat's principle for the ray PQP′ implies that $d(OPL)/d\alpha = 0$, leading to

$$0 = n\frac{(-s+R)}{-l} - n'\frac{(s'-R)}{l'}, \tag{1.19}$$

from which

$$\frac{n'}{l'} - \frac{n}{l} = \frac{1}{R}\left[\frac{n's'}{l'} - \frac{ns}{l}\right]. \tag{1.20}$$

This equation tells us how far P′ is from V for the ray PQP′. This is an equation with two unknowns, s' and l'. Therefore, to find the position of P′, we can choose some numerical method. However, to see how a ray bundle is refracted out of the point P, instead of using Eq. (1.20), let us use the graphical ray tracing method described in Section 1.1.4. For example, Fig. 1.24 shows a five ray trace on a spherical surface Σ of radius $R = 41$ mm, with $n = 1.0$ and $n' = 1.8$. The rays emerge from a point P located at $s = -149$ mm and reach the surface Σ with equidistant height increments.

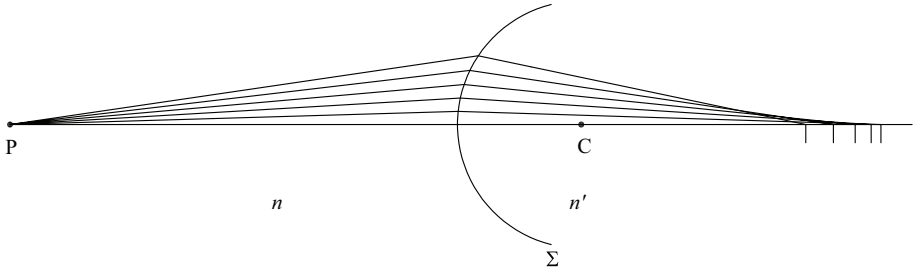


Figure 1.24 Graphical ray tracing on a spherical surface. The refracted rays do not converge at the same point.

From Fig. 1.24, it can be seen that the refracted rays do not converge at a single point on the optical axis. Since this defect appears when using a spherical surface instead of a Cartesian oval, the image is said to be affected by *spherical aberration*. It is also observed that, upon arrival to the optical axis in the image side, the rays that travel closer to the optical axis are less separated from each other than the rays that travel farther away from the optical axis. This suggests that, if the ray bundle reaching the spherical surface is limited to those hitting only a small circular area that is very close to the optical axis, acceptable images can be obtained on the image side. Thus, for an object point on the optical axis, the incident rays that we should consider are those that travel almost parallel to the optical axis, as shown in Fig. 1.25. This is called the *paraxial approximation*.

Using the approximations $l \approx s$ and $l' \approx s'$ in Eq. (1.20), we get

$$\frac{n'}{s'} - \frac{n}{s} = \frac{(n' - n)}{R}, \tag{1.21}$$

which describes image formation by a spherical surface in the paraxial approximation. Equation (1.21) is called the *Gauss equation*.

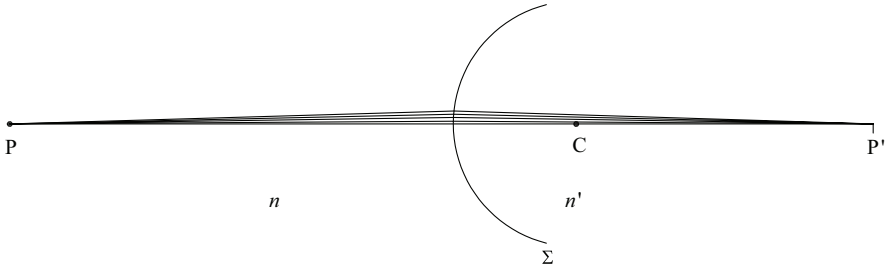


Figure 1.25 In the paraxial approximation, the incident rays emerging from P are very close to the optical axis. Refracted rays converge to the point P′.

1.3.3 Focal lengths and focal points

A refracting spherical surface is called convex if the radius of curvature is positive; it is called concave if the radius of curvature is negative. For both surfaces, we are going to see what happens for an incident or refracted parallel bundle of rays when $n' > n$. Suppose that paraxial rays parallel to the optical axis hit a convex spherical surface, i.e., the point object is located at $s = -\infty$. Therefore, in Eq. (1.21), distance s' is a constant that depends only on the geometrical and optical parameters of the refracting surface. Such a distance is called the secondary focal length f' and is determined by the formula

$$\frac{1}{f'} = \frac{(n' - n)}{n'} \frac{1}{R}. \tag{1.22}$$

Thus, $s' = f'$ is the distance (to the right of the vertex) at which all the rays converge. The point where the rays converge is called the *secondary focal point* F′. This is illustrated in Fig. 1.26(a).

Now, where should a point object be placed so that the refracted rays emerge parallel? This implies that $s' = \infty$ and therefore, in Eq. (1.21), the variable s becomes a constant that depends only on the specifications of the refracting surface. Such a constant is called the primary focal length f and is given by

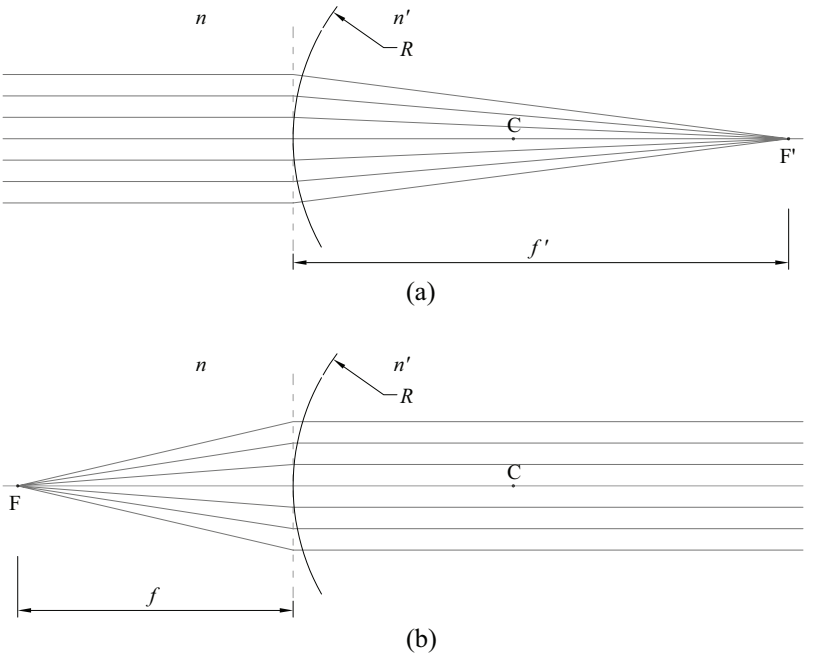


Figure 1.26 Focal lengths and focal points in a convex spherical refracting surface for which $n' > n$. (a) Secondary focal point F′ and (b) primary focal point F.

$$\frac{1}{f} = \frac{(n' - n)}{n} \frac{1}{R}. \tag{1.23}$$

Thus, $s = -f$ is the distance (to the left of the vertex) at which a point object should be placed for its image to be located at infinity. The position where the point object is placed is called the *primary focal point* F. This is illustrated in Fig. 1.26(b).

In Figs. 1.26(a) and 1.26(b), an orthogonal (dashed) line is included that passes through the vertex of the refracting surface. Note that the incident rays are drawn as refracted at the dashed line and not at the actual refracting surface (as it should be). This is done on purpose to emphasize the paraxial approximation and reminds us that in practice the rays travel very close to the optical axis. However, rays are deliberately drawn away from the optical axis to clearly illustrate the ray tracing used for locating the image.

The case of a concave surface is shown in Fig. 1.27. The focal lengths are given by Eqs. (1.22) and (1.23) but with $R < 0$, so $f' < 0$ and $f < 0$. When $s = -\infty$, the refracted rays diverge as if they were coming from a point F′ located at $s' = -|f'|$, as shown in Fig. 1.27(a). In such a case, we would say

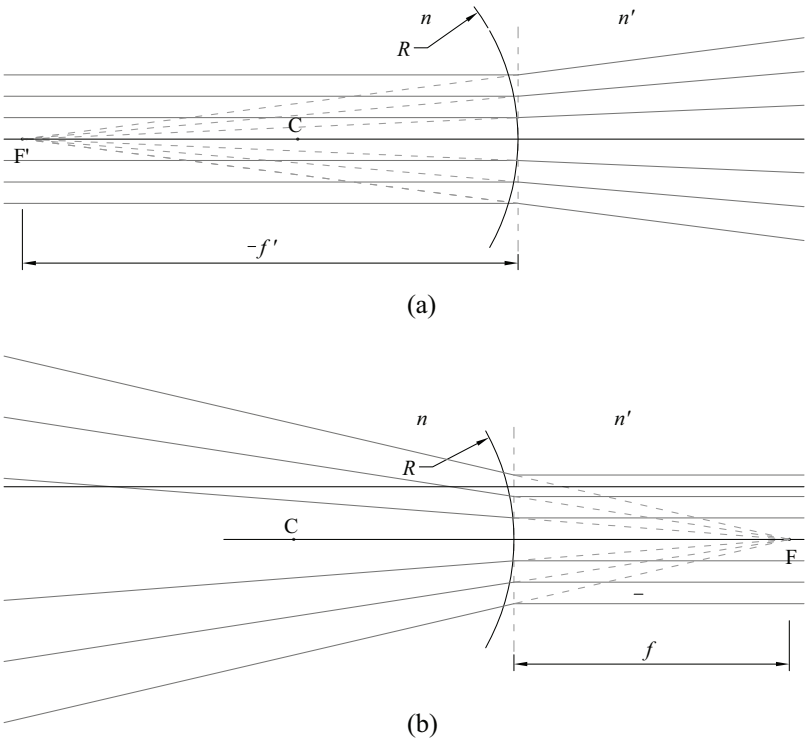


Figure 1.27 Focal lengths and focal points in a concave refracting spherical surface for which $n' > n$. (a) Secondary focal point F′ and (b) primary focal point F.

that there is a *virtual point image* at F′ (secondary focal point). When $s' = \infty$, the incident rays converge to the point F located at $s = |f|$, as shown in Fig. 1.27(b). In such a case, we would say that there is a *virtual point object* at F (primary focal point).

1.3.4 Focal planes

In Fig. 1.26(a), we have rays parallel to the optical axis, limited to the paraxial region, that focus to a secondary focal point. These rays come from an object point (at infinity) that is on the optical axis. If we now consider another object point outside the optical axis (also at infinity), the bundle of parallel rays that hits the refracting surface is oblique. These rays converge at a point that is at the focal length f' from the refracting surface, measured along the auxiliary optical axis drawn from the off-axis object point to the center of curvature C. This is shown in Fig.1.28. With respect to this auxiliary axis, the situation is equivalent to the object on the original optical axis.

As a result, considering the refraction of the oblique bundles of parallel rays, the rays will be focused onto a curved surface at the image side. This surface is called the *focal surface*. The paraxial approximation implies that the inclinations of the rays are limited to small angles. Therefore, in this case, all the oblique bundles of parallel rays will focus onto a small region of the focal surface that can be approximated by a plane. This is called the *focal plane* (Fig. 1.28).

Accordingly, we would say that in the paraxial approximation, the focal planes are defined as planes perpendicular to the optical axis containing the focal points. Hence, primary and secondary focal planes can be established for each refracting surface.

In the paraxial approximation, the following generalizations can be made for $n < n'$:

- An oblique bundle of parallel rays that hit at a convex spherical refracting surface converges to a point on the secondary focal plane;

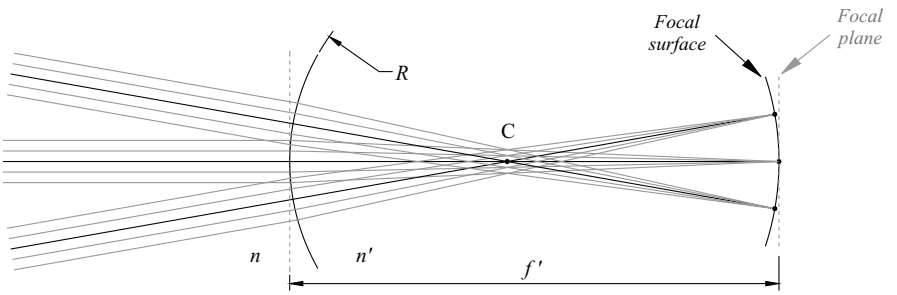


Figure 1.28 Focal surface. In the paraxial approximation, the focal surface can be described as a plane where oblique bundles of parallel rays converge.

and, for a point object located anywhere on the primary focal plane, the rays are refracted as an oblique bundle of parallel rays.

- An oblique bundle of parallel rays that hit at a concave spherical refracting surface diverges from a point on the secondary focal plane; and, for a virtual point object located anywhere on the primary focal plane, the rays are refracted as an oblique bundle of parallel rays.

In both cases, the inclination of the bundle of parallel rays is given by the undeviated ray passing through the center of curvature C (Fig. 1.28).

Example: Refraction in a sphere

As an application of an optical system that contains two spherical surfaces, convex and concave, consider the image for a point object located at infinity that is generated by a sphere of radius R_0 and refractive index $n' = 3/2$. Refraction occurs at two surfaces, labeled as 1 and 2 in Fig. 1.29. The first surface has a radius $R = R_0$. The second has a radius $R = -R_0$ and is separated from the first by a distance $2R_0$. The object distance measured from the first surface is $s_1 = -\infty$. Therefore, the image generated by the first surface is at $s'_1 = f'_1 = 3R_0$. This first image is then seen from the second surface as a virtual object located at $s_2 = 3R_0 - 2R_0 = R_0$. The image for this virtual object, generated by the second surface, would be located at $s'_2 = R_0/2$.

It is worth noting that for the first image to be located right at the vertex of the second surface, i.e., at $s'_1 = 2R_0$ when $s_1 = -\infty$, per Eq. (1.21) the refractive index of the sphere would have to be $n' = 2$. In a way, this explains why the human eye is not completely spherical. If it were with an equivalent refractive index $n_{eq} = 4/3$, all of us would be hypermetropic (farsighted). Instead, the anterior corneal surface of the human eye has a smaller radius than the eyeball so that the refracted light hits the retina (at the vertex

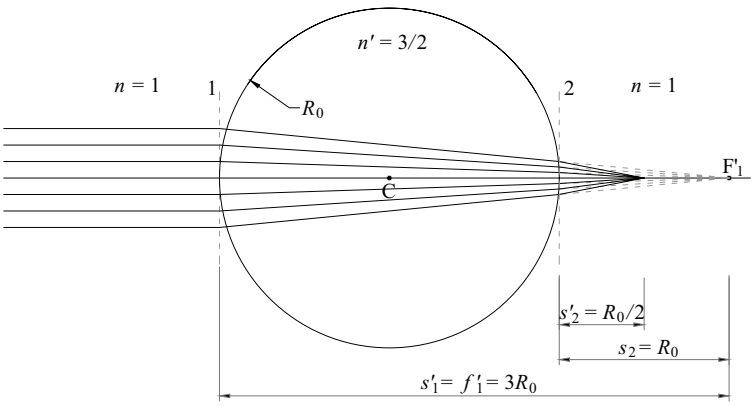


Figure 1.29 The refraction in a sphere of refractive index $n' = 3/2$ generates the image of an object at $s = -\infty$ at the distance $s'_2 = R_0/2$ from the sphere's back face (dashed line labeled as 2).

of the second surface), as illustrated in the simplified model of the eye shown in Fig. 1.22.

1.3.5 Paraxial imaging of extended objects

So far we have only considered point objects. If the object is not a point, we would say it is an extended object and its size can be described by an arrow of a certain height, measured from the optical axis. All rays that leave the object will be contained in a plane that includes the optical axis. This plane is called the *meridional plane*.

With the help of the parallel rays tracing method described in Section 1.3.3, we can graphically find the position and size of the image. Consider Fig. 1.30, which shows two refraction cases: (a) a convex surface and (b) a concave surface. In both cases, let us place an object of height h at a distance $-s$ from the vertex.

In the first case, let us trace a ray that leaves the object tip and travels parallel to the optical axis. This ray is refracted toward the secondary focal point. Then, let us trace another ray, also coming out of the object tip, that passes through the primary focal point and is refracted traveling parallel to the optical axis. The two refracted rays intersect at a point that defines the tip of the image. Such an intersection point reveals the size of the image and its location, i.e., $-h'$ and s', respectively. A third ray that one can draw, exiting the object tip, is the ray that passes through the center of curvature of the surface. Since the angle of incidence at the refracting surface is zero, this ray is

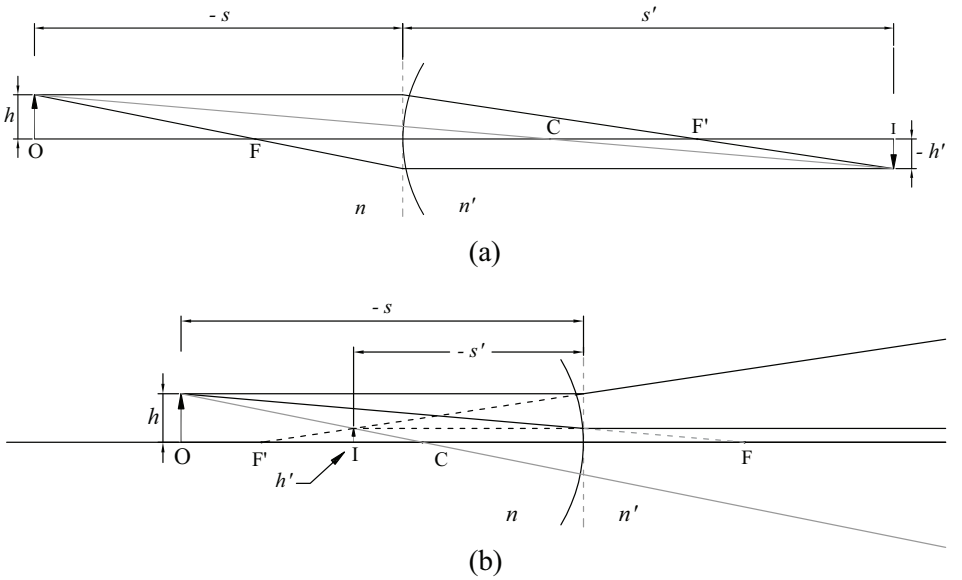


Figure 1.30 Graphical ray tracing to determine location and size of the image generated by spherical refracting surfaces with $n' > n$. (a) Convex, $R > 0$; (b) concave, $R < 0$.

not deflected as it passes through. Of course, this ray will also reach the end point where the first two rays intersect.

In the second case, let us again trace a ray that leaves the object tip and travels parallel to the optical axis. This ray is refracted diverging from the secondary focal point. Next, let us trace a second ray, also coming out of the object tip and passing through the center of curvature of the refracting surface. As in the previous case, this ray is not deflected when passing through the refracting surface because its angle of incidence is zero. The two refracted rays are divergent, so they do not intersect (there is no real image). However, the two refracted rays appear to emerge from a common point behind the refracting surface. Such a point is the tip of a virtual image; its projection on the optical axis locates its position. Again, we have determined the size and position of the image, i.e., h' and $-s'$, respectively. A third ray can be drawn exiting the object tip and directed toward the primary focal point (which is to the right of the surface). When this ray reaches the refracting surface, it is refracted in a direction parallel to the optical axis. This ray also appears to emerge from the tip of the virtual image.

The *image magnification* is the ratio between the image size and the object size, i.e., $m_t = h'/h$. If m_t is negative, the image is inverted, as in Fig. 1.30(a); if m_t is positive, the image is upright, as in Fig. 1.30(b). On the other hand, if $|m_t| > 1$, the image will be larger, and if $|m_t| < 1$, the image will be smaller, when compared to the size of the object. From the geometry of Fig. 1.30(a), looking at similar triangles, we get

$$m_t = \frac{R - s'}{R - s}. \tag{1.24}$$

This relationship is also valid for the case of Fig. 1.30(b), taking into account that $R < 0$.

1.3.6 Optical power and vergence

Another useful way to interpret Eq. (1.21) is by analyzing the change in curvature of the refracted wavefront, with respect to the incident wavefront, by using the *refractive power* of the spherical surface, which is defined as

$$P = \frac{(n' - n)}{R}. \tag{1.25}$$

The unit of optical power is called the diopter (D) and is defined as $1\ \mathrm{D} = 1\ \mathrm{m}^{-1}$. Since we are in the paraxial approximation, the wavefronts emerging from a point object and converging to a point image are spherical. In particular, the curvature of the incident wavefront, hitting the vertex of the surface, would be $1/s$; the curvature of the refracted wavefront leaving the vertex, propagating on the the right side of the surface, would be $1/s'$. These

curvatures multiplied by the corresponding refractive indices are called *vergences*, U and V, respectively. The unit of vergence is also the diopter. In terms of these quantities, Eq. (1.21) becomes

$$V - U = P. \tag{1.26}$$

This relationship between vergences and power is very useful in optometry and is often used as

$$U + P = V. \tag{1.27}$$

Thus, the curvature U/n of the object wavefront is modified by the power P of the refracting surface, resulting in the curvature V/n' of the image wavefront. Taking into account the Cartesian sign convention, a diverging wavefront has a negative curvature and a converging wavefront has a positive curvature. Also, whereas a divergent spherical wavefront corresponds to a real object or a virtual image, a convergent spherical wavefront would correspond to a real image or a virtual object.

In a system of several refracting surfaces, the power of the jth surface is

$$P_j = \frac{n'_j - n_j}{R_j}, \tag{1.28}$$

where n'_j and n_j are the refractive indices to the right and to the left of the jth surface, respectively, and R_j is the jth radius of curvature.

1.4 Reflecting Surfaces

In addition to refracting light, an interface that separates two media can also reflect light. If the surface is smooth (like a mirror), the reflection will be specular; but if it is rough, the reflection will be diffuse. In this section, we will deal with specular reflection of light, which can also be described by light rays and, of course, complies with Fermat's principle (as mentioned in Section 1.2). Using Fig. 1.13, we can describe the reflection phenomena in a plane mirror. By applying Fermat's principle to Fig 1.13(a), an equation similar to Eq. (1.8) is obtained but with $n' = n$. The reflection of light on a plane, and even on spherical mirrors, can be described from the equations obtained in Sections 1.2 and 1.3 (for refraction), if we make $n' = -n$ and follow the sign convention established in Section 1.3.2. Moreover, if we limit this treatment to mirrors immersed in air, then $n = 1$ and, therefore, $n' = -1$. With this, Eq. (1.2) becomes

$$\theta' = -\theta. \tag{1.29}$$

Figure 1.31 illustrates two cases of light reflection: (a) on a flat surface and (b) on a curved surface. In both cases, the angle of incidence of a light ray is

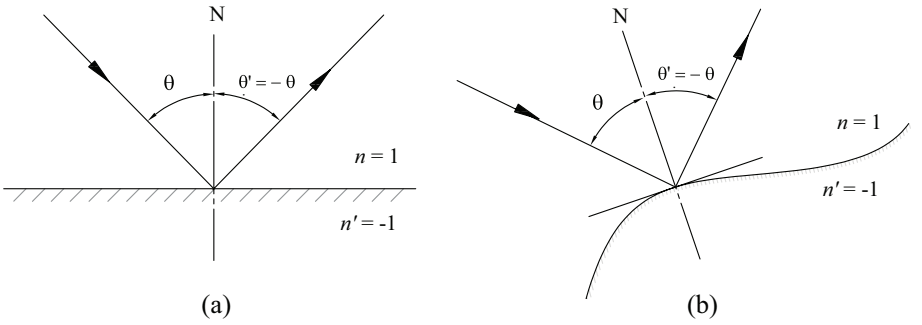


Figure 1.31 Reflection on a smooth surface. (a) A flat mirror and (b) a curved surface. The angle of reflection is equal to the angle of incidence but with a negative sign.

measured with respect to the normal line N at the point on the surface where the ray is incident. The angle of reflection is also measured with respect to the normal N. The incident ray, the reflected ray, and the normal at the point of incidence all lie in the same plane, known as the *plane of incidence*. Thus, the *law of reflection* can be stated as follows:

> The angle formed by the reflected ray and the normal equals the angle formed by the incident ray and the normal. The two rays and the normal are contained in the same plane.

On the other hand, Eq. (1.21) becomes

$$\frac{1}{s'}+\frac{1}{s}=\frac{2}{R} \tag{1.30}$$

for spherical mirrors. According to the sign convention described in Section 1.3.2, $R > 0$ if the center of curvature is to the right of the mirror vertex and $R < 0$ if the center of curvature is to the left of the mirror vertex. In the former case, we say that the mirror is convex; in the latter case, we say that it is concave. The distances s and s' are measured with respect to the mirror vertex.

1.4.1 Ray tracing for spherical mirrors

In spherical mirrors, two incident rays are usually traced; a ray parallel to the optical axis and a ray directed to the mirror vertex, as shown in Fig. 1.32.

For the first ray $s = -\infty$, so $s' = R/2$. Therefore, if $R < 0$, incident paraxial rays parallel to the optical axis will be reflected converging at a point located to the left of vertex V at a distance $s' = f = -|R|/2$. This point is called the focal point F of the mirror and will be the image of a point object located at infinity. If $R > 0$, incident paraxial rays parallel to the optical axis will be

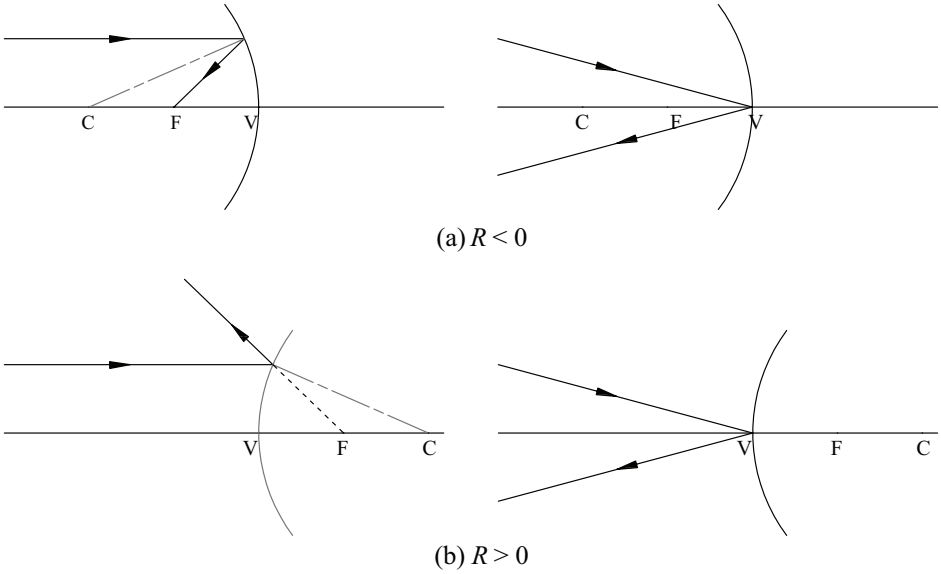


Figure 1.32 Rays in spherical mirrors based in the paraxial approximation. (a) A concave mirror and (b) a convex mirror.

reflected diverging from a point located to the right of V at the distance $s' = f = R/2$. Now the focal point F is virtual, and the image of a point object located at infinity will be a virtual point image at F. On spherical surfaces, the normal at the point of incidence is a radial line with origin at the center of curvature C.

For the second ray, the situation is very simple. At the vertex the normal of the surface coincides with the optical axis, so for both (concave and convex) mirrors the ray is reflected to the left and downward. The trace of these two rays is shown in Fig. 1.32(a) for the concave mirror and in Fig. 1.32(b) for the convex mirror.

With these two rays, we can locate the image generated by the mirrors. In Fig. 1.33, the ray tracing for four different scenarios in a concave mirror is shown. In Figs. 1.33(a), (b), and (c), the image is real and inverted, going from smaller to larger than the object. In particular, in (b) $s = R$ and $s' = R$, and the image is the same size as the object. In (d) the image is virtual, upright, and larger than the object.

The image magnification $m_t = h'/h$ can be obtained by considering the ray that goes to the mirror vertex. According to the trace of that ray, we have two similar triangles that satisfy the relation $h/(-s) = (-h')/(-s')$. Therefore, the image magnification will be

$$m_t = -\frac{s'}{s}. \tag{1.31}$$

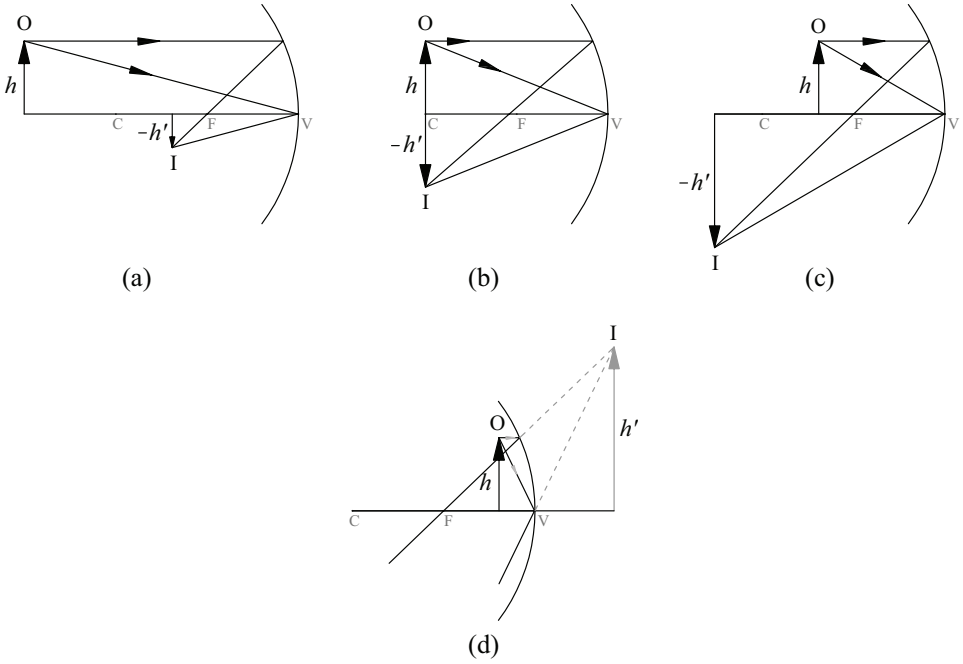


Figure 1.33 Image formation in concave spherical mirrors. (a) Real image, inverted, and smaller than the object. (b) Real image, inverted, and of the same size as the object. (c) Real image, inverted, and larger than the object. (d) Virtual image, upright, and equal in size to the object.

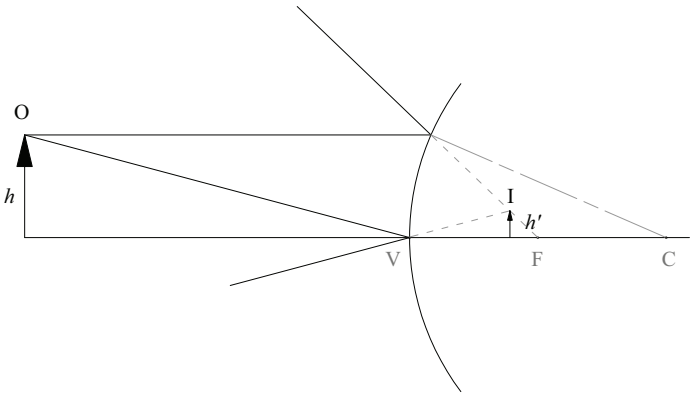


Figure 1.34 Image formation in a convex mirror. The image is always virtual, upright, and smaller than the object, and its apparent position is between the vertex V and the focal point F.

Thus, in the case shown in Fig. 1.33(b) the magnification turns out to be $m_t = -1$. The minus sign $(-)$ indicates that the image is inverted.

In Fig. 1.34, the ray tracing in a convex mirror is shown. For any position of the object, the image is virtual, upright, and smaller than the object. The

image seems to be behind the mirror, and the image magnification can still be computed from Eq. (1.31).

1.4.2 The parabolic mirror

The ray tracing presented in Section 1.4.1 for spherical mirrors is limited to the paraxial approximation. For the general case, not paraxial, consider the imaging process in a concave spherical mirror for a point object located at infinity but on the optical axis. This implies that the light reaching the mirror can be considered a bundle of parallel rays. In the paraxial approximation, all the rays in the bundle converge to the focal point [Fig. 1.32(a)]. However, by ignoring the paraxial approximation, if we apply Snell's law for each of the rays in the bundle, the reflection of rays occurs as shown in Fig. 1.35. As with spherical refracting surfaces, this is called spherical aberration.

This can be easily verified with the help of Fig. 1.36. A ray parallel to the optical axis hits Q, and then is reflected through point X on the optical axis. The triangle CXQ is isosceles, so the following relationship can be established:

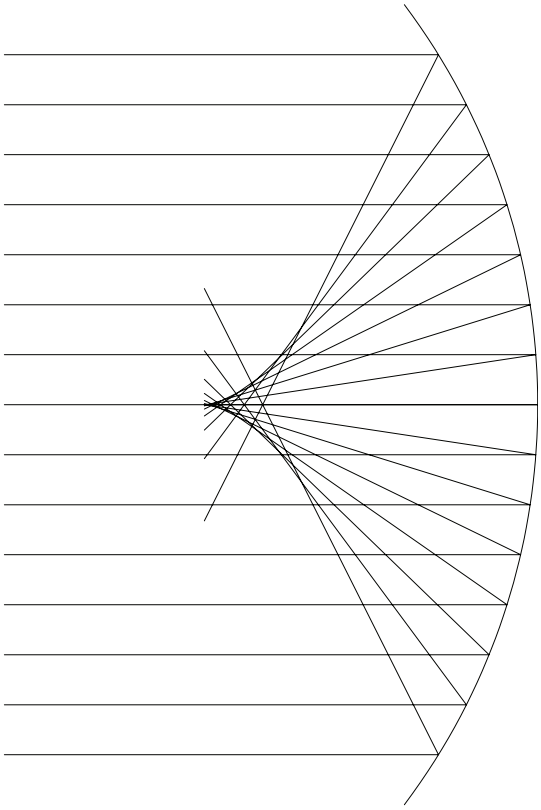


Figure 1.35 Parallel rays in a concave spherical mirror do not converge to the same point.

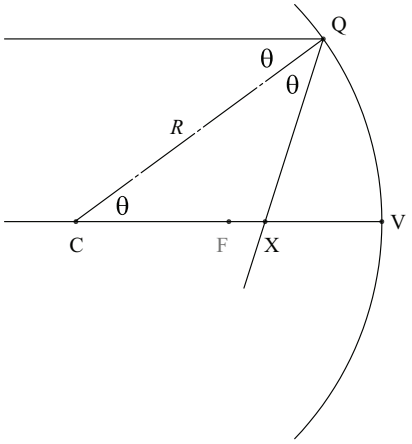


Figure 1.36 Nonparaxial rays in a concave spherical mirror, traveling parallel to the optical axis, reflect crossing the optical axis at a point other than the focal point. As the angle θ increases, the reflected ray crosses the optical axis at a point closer to the vertex.

$$\overline{\mathrm{CX}} + \overline{\mathrm{QX}} > R, \tag{1.32}$$

as long as Q does not coincide with V. Also, $\overline{\mathrm{CX}} = \overline{\mathrm{QX}}$; therefore,

$$\overline{\mathrm{CX}} > \frac{R}{2}, \tag{1.33}$$

and the focal point F is at a distance $R/2$ from C. This implies that the nonparaxial rays that hit parallel to the optical axis do not converge to F. As the incidence angle θ increases, the reflected ray crosses the optical axis at a position that is closer to the mirror vertex.

So, what shape should the concave mirror have to focus the rays parallel to the optical axis at a single point? The answer is, again, given by Fermat's principle. Based on the geometry shown in Fig. 1.37, applying Fermat's principle,

$$OPL = a + \rho\cos\theta + \rho. \tag{1.34}$$

All rays from the orthogonal line to the optical axis passing through P must have the same OPL to reach F. Defining a new constant $p = OPL - a$, then

$$\rho(1 + \cos\theta) = p. \tag{1.35}$$

This is the equation of a parabola. Thus, the shape of the mirror we are looking for is a paraboloid of revolution. Note that Eq. (1.35) can also be obtained if we change $n' = -n = -1$ in Eq. (1.15).

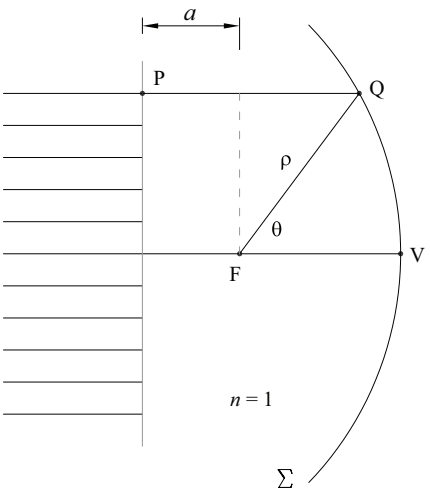


Figure 1.37 A parabolic mirror focuses the parallel rays coming from infinity at the focal point.

In Cartesian coordinates, the profile of the paraboloid can be written as

$$z = \frac{y^2}{2R_v}, \tag{1.36}$$

where z is the coordinate along the optical axis, y is the meridional coordinate, and R_v is the radius of curvature of the parabola at the vertex. Then, the focal point of the mirror is located at the distance $R_v/2$ from the vertex V (Fig. 1.37).

1.5 Lenses: Thin Lens Approximation

If we want to obtain images in a medium other than the spherical refracting surface discussed in Section 1.3.2, another refracting surface should be included. In this section, we will deal with refracting elements limited by two spherical surfaces with a common optical axis. These types of elements are called lenses.

To obtain the position of the image of a point object, we should use Eq. (1.21). The Gaussian equation for surface 1 is

$$\frac{n_1'}{s_1'} - \frac{n_1}{s_1} = \frac{(n_1' - n_1)}{R_1}; \tag{1.37}$$

for surface 2, it is

$$\frac{n_2'}{s_2'} - \frac{n_2}{s_2} = \frac{(n_2' - n_2)}{R_2}. \tag{1.38}$$

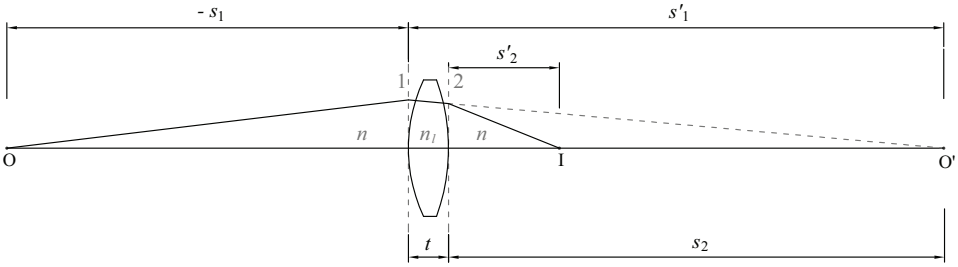


Figure 1.38 Locating the image of an object generated by a positive lens.

Limiting ourselves to a lens of refractive index n_l immersed in a single medium of refractive index $n < n_l$, as shown in Fig. 1.38, then $n_1 = n'_2 = n$ and $n'_1 = n_2 = n_l$. By adding Eqs. (1.37) and (1.38), we get

$$\frac{1}{s'_2} - \frac{1}{s_1} = \frac{(n_l - n)}{n}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{n_l}{s_2} - \frac{n_l}{s'_1}. \tag{1.39}$$

The lens thickness t is the distance between the vertices of surfaces 1 and 2, related to the distances s'_1 and s_2 through the relationship $t = s'_1 - s_2$. If $t \ll s'_1$ and $t \ll s_2$, such that $s'_1 \approx s_2$, then the last two terms on the right side of Eq. (1.39) would cancel each other. This approximation implies that the lens thickness is negligible compared with the radii of curvature of the two surfaces and is known as the *thin lens*. Defining the object distance as $s_o = s_1$ and the image distance as $s_i = s'_2$, Eq. (1.39) can be rewritten as

$$\frac{1}{s_i} - \frac{1}{s_o} = \frac{(n_l - n)}{n}\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{1.40}$$

which is known as the *thin lens equation*.

The right side of Eq. (1.40) is a constant that only depends on the lens parameters. The inverse of this is called the lens *focal length* f and is given by

$$\frac{1}{f} = \frac{(n_l - n)}{n}\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{1.41}$$

Using the definition of the focal length, we arrive at the *Gaussian equation for lenses*:

$$\frac{1}{s_i} - \frac{1}{s_o} = \frac{1}{f}. \tag{1.42}$$

The focal length is a parameter that identifies the lens. The sign of the focal length depends on the values of the refractive indices n and n_l and the radii of curvature R_1 and R_2. Since in practice most lenses are used in

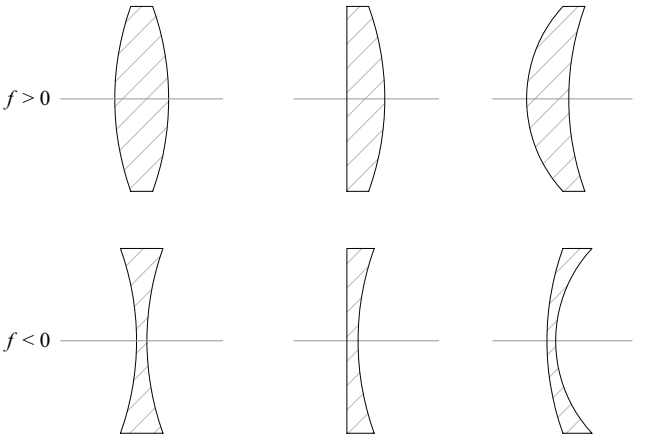


Figure 1.39 Positive and negative lenses immersed in air ($n = 1$).

air, in what follows we will use $n = 1$, and the thin lens focal length will be defined by

$$\frac{1}{f} = (n_l - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{1.43}$$

Taking this into account, Fig. 1.39 shows the general shapes of the positive and negative lenses. As a rule of thumb, we can say that a lens is positive if the thickness at the edges is less than the thickness at the center region. Similarly, it can be said that a lens is negative if the the thickness at the edges is greater than the thickness at the center.

1.5.1 Ray tracing for thin lenses

In Section 1.3.3, we defined the focal points for refracting spherical surfaces as the points on the optical axes located at distances f' and f from the surfaces' vertices. For thin lenses, following Fig. 1.38, the extension of the spherical surfaces 1 and 2 meet in a plane orthogonal to their optical axis. Such a plane is used as a reference plane to measure distances. Thus, for a positive lens, a bundle of paraxial parallel rays coming from a point object located on the optical axis at infinity, $s_o = -\infty$, is refracted by the lens converging to a point image on the optical axis located at a distance $s_i = f$. This point is called the *secondary focal point*. Whereas, if we locate a point object at a distance $s_o = -f$, on the optical axis, the paraxial rays refracted by the lens travel parallel to the optical axis. Such positional location is called the *primary focal point*. For a negative lens, the situation is analogous; a bundle of paraxial parallel rays coming from a point object located on the optical axis at infinity, $s_o = -\infty$, is refracted by the lens as if the rays diverge out from a point on the optical axis at $s_i = -|f|$ (secondary focus). For a virtual object located on the

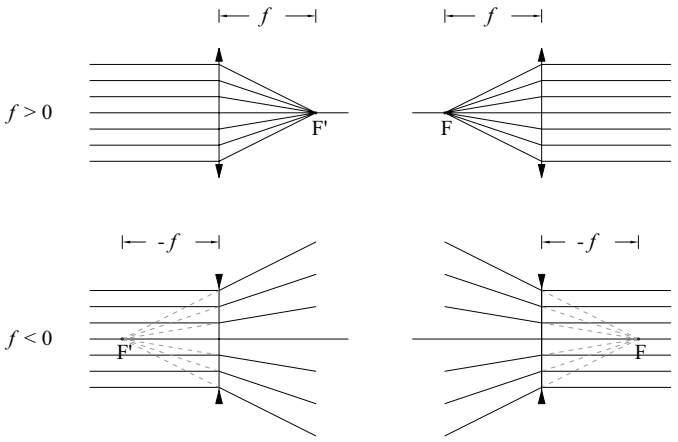


Figure 1.40 Focal lengths and focal points of thin lenses.

optical axis at $s_o = |f|$ (primary focus), the paraxial rays refracted by the lens travel parallel to the optical axis. The focal lengths and focal points of thin lenses are illustrated in Fig. 1.40. Positive lenses are represented by vertical line segments with outward arrows at the ends, and negative lenses are represented by vertical line segments with inward arrows at the ends.

From the ray tracing shown in Fig. 1.40, we can obtain the position and size of the image generated by a thin lens for an extended object. Consider an object of height h_o located at distance $-s_o$ from the lens, as shown in Fig. 1.41. To find the image we can trace two rays: one ray that leaves the object tip and travels parallel to the optical axis, and a second ray, also coming out of the object tip but passing through the primary focal point. After being refracted by the thin lens, the first ray passes through the secondary focal point and the second ray travels parallel to the optical axis. The intersection of these two refracted rays determines the tip of the image and, therefore, the image size $-h_i$ and the distance location s_i.

In general, any ray exiting the tip of an object will be refracted after passing through the lens in such a way that it reaches the tip of the image. In

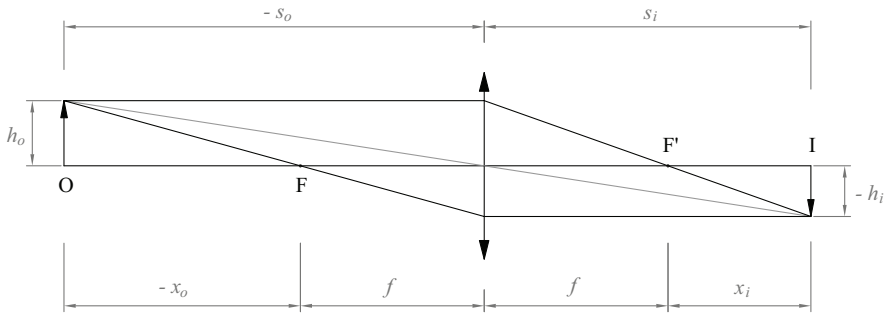


Figure 1.41 Ray tracing to locate the image generated by a positive thin lens.

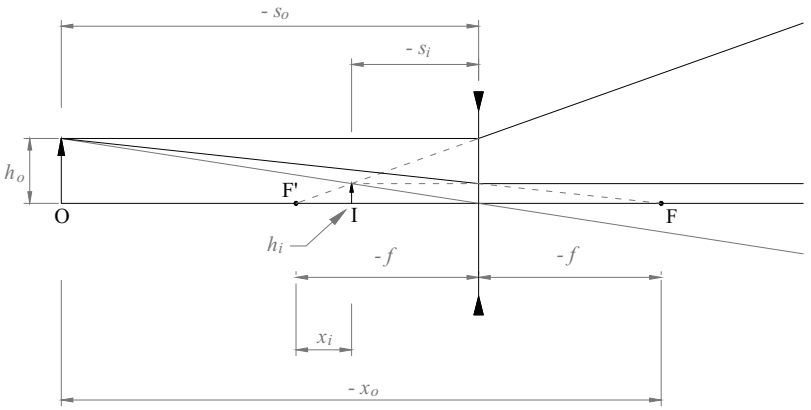


Figure 1.42 Ray tracing to locate the image generated by a negative thin lens.

particular, in Fig. 1.41, a third ray can be traced that goes from the tip of the object to the tip of the image passing through the center of the lens without deviating. Using the geometry of the similar triangles generated by this ray with the optical axis and the object and image, the image magnification is given by

$$m_t = \frac{h_i}{h_o} = \frac{s_i}{s_o}. \tag{1.44}$$

For a thin negative lens, the ray tracing is shown in Fig. 1.42. In such a case, the image magnification is also given by Eq. (1.44).

1.5.2 Newton's lens equation

In Figs. 1.41 and 1.42, two new parameters have been included, x_o and x_i, which are defined as follows: x_o is the distance measured from the primary focal point to the object and x_i is the distance measured from the secondary focal point to the image. For these measures, the sign convention established in Section 1.3.2 also applies. Thus, $x_o < 0$ if the object is to the left of F and $x_o > 0$ if the object is to the right of F; $x_i < 0$ if the image is to the left of F′ and $x_i > 0$ if the image is at the right of F′.

Following the geometry of Fig. 1.41,

$$\frac{h_o}{-x_o} = \frac{-h_i}{f} \tag{1.45}$$

and

$$\frac{-h_i}{x_i} = \frac{h_o}{f}. \tag{1.46}$$

Substituting h_i/h_o from one equation into the other, we obtain

$$x_o x_i = -f^2. \tag{1.47}$$

This is Newton's equation for both positive and negative thin lenses. On the other hand, the h_i/h_o ratio is the image magnification. Thus, we have two other expressions for the magnification given by

$$m_t = \frac{f}{x_o} \tag{1.48}$$

and

$$m_t = -\frac{x_i}{f}. \tag{1.49}$$

Distances in Gauss' formulation are related to distances in Newton's formulation through $s_o = x_o - f$ and $s_i = x_i + f$. In some cases, compared with Gauss' equation, Newton's equation is advantageous. For example, suppose we are faced with the following situation: given a positive lens of focal length f, where should an object be placed so that its image magnification is $m_t = -2$? According to Gauss' formulation, the magnification is $m_t = s_i/s_o$. But none of the distances are known, so we would also have to use Eq. (1.42); i.e., this is a problem of two equations with two unknowns. Whereas, with Newton's formulation, the magnification is $m_t = f/x_o$, from which the position of the object, $x_o = -f/2$ ($s_o = -3f/2$), is directly obtained.

1.5.3 Real and virtual images domain

Let us explore object positional ranges for which positive and negative lenses generate real and virtual images.

For the positive lens illustrated in Fig. 1.43, five positions have been selected for an object, labeled as a, b, c, d, and e:

- Position a corresponds to an object at infinity ($x_o = -\infty$), so its image a' will be at the secondary focal point ($x_i = 0$) with the magnification $m_t = 0$.
- Position c corresponds to $x_o = -f$, so its image c' will be at $x_i = f$ with magnification $m_t = -1$. This is the setting where the image is the same size as the object.
- For the intermediate position b, which is between a and c, if $x_o = -3f/2$, its image b' will be at $x_i = 2f/3$ with magnification $m_t = -2/3$. In this case, the image is smaller than the object.
- The object at d is at position $x_o = -f/2$, so its image d' will be at $x_i = 2f$ with magnification $m_t = -2$.

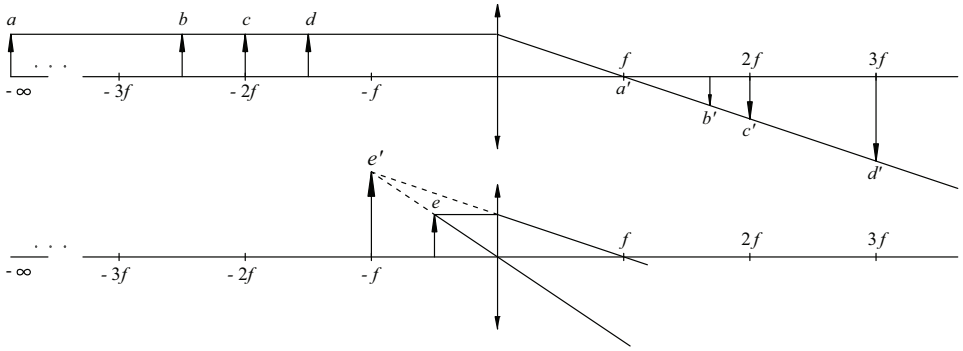


Figure 1.43 Object positional ranges for which real and virtual images are obtained with a positive lens. For an object located in the range $(-\infty, -2f]$ to the left of the lens, the image is real, inverted, smaller than the object, and located in the range $[f, 2f]$ to the right of the lens. For an object in the range $[-2f, -f]$, the lens forms an image that is real, inverted, larger than the object, and located in the range $[2f, \infty)$. For an object in the range $[-f, 0)$, the image is virtual, upright, larger than the object, and located in the range $(-\infty, 0)$. In the latter case, the lens works as a magnifying glass.

- Moving the object closer to the primary focal point moves the image farther away, increasing its size. When the object is located at $x_o = 0$ (primary focal point), the image will be at $x_i = \infty$ with magnification $m_t = -\infty$.
- Finally, at e, the object is in an intermediate position between the primary focal point and the lens, $x_o = f/2$. The refracted rays diverge, so the image e' located at $x_i = -2f$ ($s_i = -f$) will be virtual with magnification $m_t = 2$. The virtual image will be upright, twice the size of the object, and located behind the object.

In summary, for objects located in the range $-\infty < s_o \leq -2f$, the image will be real, inverted, smaller than the object, and located in the range $f \leq s_i \leq 2f$. For objects in the range $-2f \leq s_o \leq -f$, the image will be real, inverted, larger than the object, and located in the range $2f \leq s_i < \infty$. For objects in the range $-f \leq s_o < 0$, the image will be virtual, upright, larger than the object, and located in the range $-\infty < s_i < 0$. The latter case corresponds to the use of the lens as a magnifying glass.

In the case of a negative lens ($f < 0$), there is only one configuration for the image assuming a real object. A real object will be within the range $-\infty < x_o < -|f|$ or, equivalently, $-\infty < s_o < 0$, since x_o is measured from the primary focal point (which in a negative lens is located to the right of the lens). Thus, for any distance, $x_o = -\alpha|f|$, where $\alpha > 1$, the image will always be to the right of the secondary focal point (at $x_i = |f|/\alpha$) and to the left of the lens; i.e., the positional range of the image is $-|f| \leq s_i < 0$. The image magnification will be $m_t = 1/\alpha$. In Fig. 1.44, the positional ranges for an object and the corresponding positional ranges for its image through a negative lens are displayed. The image is always smaller than the object.

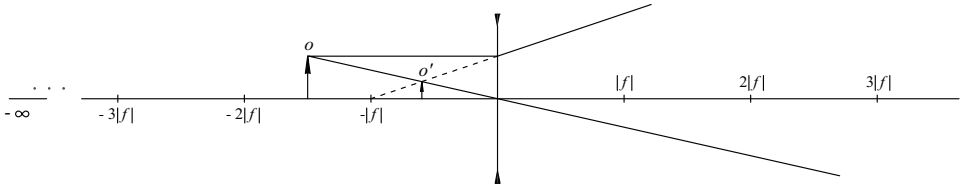


Figure 1.44 A negative lens forms virtual images that are upright and smaller than the object, located between the object and the lens.

1.5.4 Focal planes in thin lenses

Similar to the definition of focal planes given in Section 1.3.4, thin lens focal planes are planes perpendicular to the optical axis containing the focal points. Within the paraxial approximation, oblique bundles of parallel rays hitting a lens can be refracted in two ways: (1) by converging to a point located in the secondary focal plane if the lens is positive, and (2) by diverging from a point located in the secondary focal plane if the lens is negative. On the other hand, refracted rays come out of the lens as oblique bundles of parallel rays when: (1) a point object is placed in the primary focal plane of a positive lens, and (2) a virtual object is located on the primary focal plane of a negative lens. These cases are illustrated in Fig. 1.45. In any case, a ray passing through the center of the lens is not deflected when refracted.

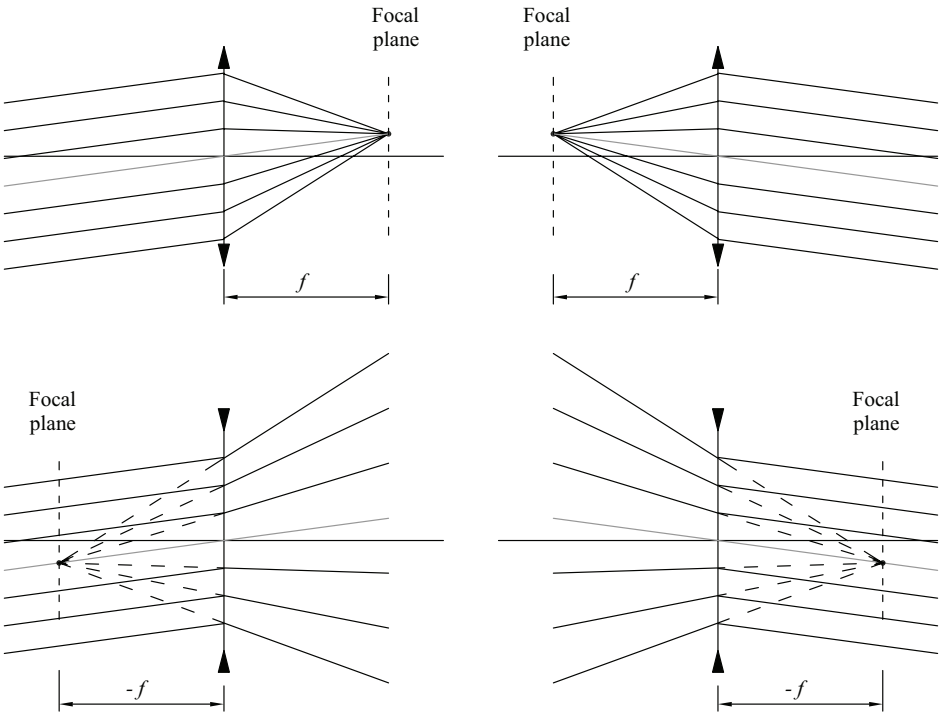


Figure 1.45 Focal planes in positive and negative lenses.

1.5.5 Ray tracing for oblique rays

Although they have been mentioned in Sections 1.3.4 and 1.5.4, oblique rays can be specifically defined as those rays that leave the tip of an object and do not travel in a direction parallel to the optical axis. In particular, we are going to deal with oblique rays that are kept in the meridional plane. We have already dealt with some of them in Section 1.5.4, e.g., the ray passing through the center of a lens and the ray directed toward a primary focal point. In this section, we are interested in oblique rays that are directed in any other direction, which can occur in ray tracings that involve a combination of two or more lenses.

For example, in Fig. 1.46, a ray refracted by the lens L_1 is directed to the secondary focal point of L_1. As it hits the lens L_2, how is the ray refracted? With the ray tracing technique illustrated in Fig. 1.40, we do not have a solution. However, the ray tracing shown in Fig. 1.45 gives us a hint as to how to graphically determine the refraction generated by L_2. If we assume that the oblique ray reaching L_2 is part of a bundle of parallel rays, as illustrated in Fig. 1.47(a), such a ray will be refracted (according to Fig. 1.45), diverging

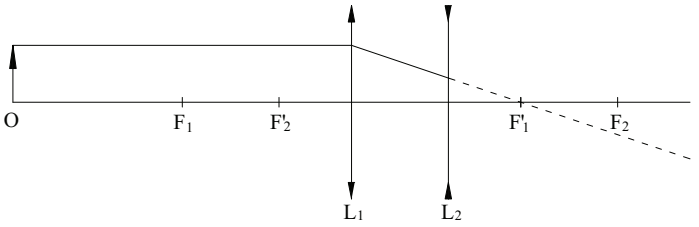


Figure 1.46 Ray tracing for a ray parallel to the optical axis coming out of the object tip. The ray refracted by lens 1 becomes an oblique ray for lens 2.

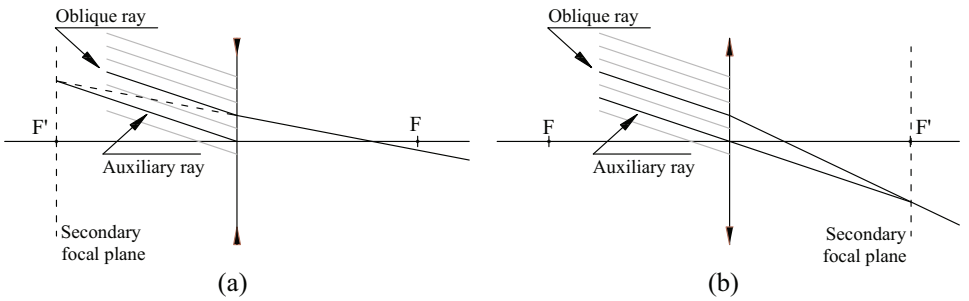


Figure 1.47 Oblique rays incident on thin lenses. (a) A negative lens and (b) a positive lens. To determine the refraction of an oblique ray hitting a lens, an auxiliary ray passing through the center of the lens and parallel to the oblique ray is drawn. The oblique ray is refracted traveling toward, or coming from, the point where the auxiliary ray intersects the secondary focal plane of the lens.

from a point located at the secondary focal plane of the lens L_2. Such a point is determined by the intersection of a ray that goes through the center of the lens (auxiliary ray) and the secondary focal plane of L_2. With this in mind, we can determine the refraction direction of the oblique ray reaching the lens L_2, as illustrated in Fig. 1.47(a). The backward extension of the refracted ray should join the auxiliary ray, which passes through the center of the lens at the focal plane. A similar procedure can be used for a positive lens, as illustrated in Fig. 1.47(b).

In summary, to determine the refraction direction of an oblique ray incident on a lens, an auxiliary ray can be drawn parallel to the oblique ray so that it passes through the center of the lens. The oblique ray is refracted so that the ray, or its backward extension, passes through the point where the auxiliary ray intersects the secondary focal plane of the lens.

Example: two positive lenses

Consider an optical system with two thin lenses of focal lengths $f_1 = 50$ mm and $f_2 = 35$ mm separated by 20 mm. Let us find the position and size of the image generated by this optical system when an object is located at $s_{o1} = -70$ mm and its size is $h_o = 11.80$ mm. Graphical ray tracing should be shown as well as analytical verification of the results.

The graphical solution is shown in Fig. 1.48, in which two rays have been drawn exiting the object tip: one ray parallel to the optical axis and a second ray directed to the center of the first lens. The two rays refracted by lens 1 reach lens 2 as oblique rays. Using the oblique ray tracing method, the final refraction of the two rays is obtained. At the point where these two rays intersect, we find the tip of the image.

Analytical verification can be performed with Newton's equation. For the first lens, $x_{o1} = -20$; hence, $x_{i1} = (50^2/20) = 125$. For the second lens, $s_{o2} = (125 + 50 - 20) = 155$, from where $x_{o2} = (155 + 35) = 190$ and thus

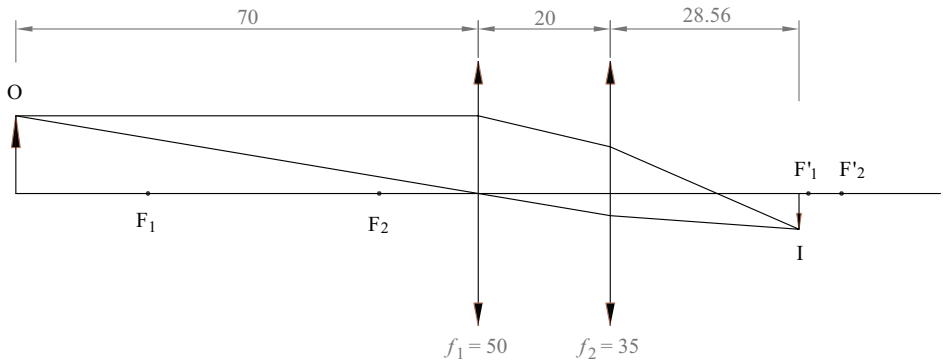


Figure 1.48 Ray tracing in a system of two positive lenses.

$x_{i2} = (-35^2/190) = -6.44$. In this way, we get $s_{i2} = (35 - 6.44) = 28.56$, as in Fig. 1.48. All distances are given in millimeters.

The image magnification of an optical system is the multiplicative product of the magnifications generated by each optical element in the system. The first lens magnification is $m_{t1} = -125/50 = -2.5$, and the second lens magnification is $m_{t2} = 6.44/35 = 0.18$; thus, the magnification of the optical system in this example is $m_t = m_{t1}m_{t2} = -0.46$. Lastly, the image size is $h_i = m_t h_o = -5.43$ mm.

1.6 Lenses: Principal Planes

In Section 1.5, we discussed lens imaging in the paraxial range using the thin lens approximation. A thin lens can be thought of as a simplification of a real lens in which two refracting surfaces with a common optical axis meet in a plane, orthogonal to the optical axis, from where the object distance and image distance are measured. The Gaussian equation for lenses, i.e., Eq. (1.42), connects these two distances by using a lens parameter, the focal length, which depends on the radii of curvature of the two refracting surfaces, the refractive index of the lens, and the refractive index of the medium surrounding the lens. Once the focal length is specified, the primary and secondary focal points can be identified. In relation to these focal points, Newton's lens equation is established, from which the position and size of the image can be determined.

With the thin lens approximation, we can design and analyze an optical system quite well; i.e., we can establish the most fundamental parameters of the optical system, such as the number of lenses to be used and their focal lengths, the spacing between lenses, the angular size of the image, the amount of energy the system can collect, etc. However, to analyze image quality, i.e., to determine how similar the image is in relation to the object, it is necessary to consider optical aberrations of the system and analyze light diffraction. Because this requires precise knowledge of the actual geometry of the lenses, the thickness of the lenses must be taken into account.

The first parameter with which we identify a lens is its focal length. The focal length can be determined by using a bundle of rays that travels toward the lens in a direction parallel to the optical axis. The point where the refracted rays converge, if the lens is positive, or the point from where the rays diverge, if the lens is negative, must be located. That point of convergence or divergence is the secondary focal point F′. In the thin lens approximation, the focal length corresponds to the distance between the plane representing the thin lens and the secondary focal point. In the case of a real lens (commonly referred to as a "thick lens"), we can easily define the secondary focal point as

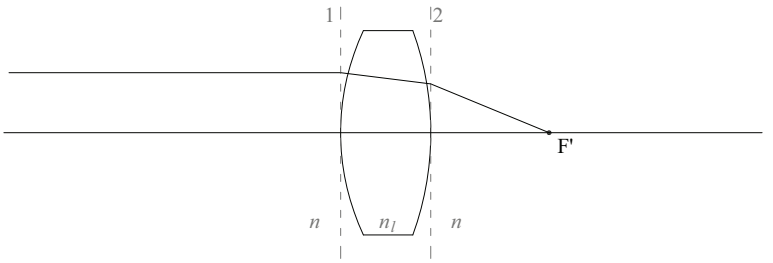


Figure 1.49 Secondary focal point in a real (thick) lens.

shown in Fig. 1.49. But what is the focal length? And what is the point or plane from where such length is measured?

In a thin lens, the forward projection of a ray incident parallel to the optical axis intersects the backward projection of its corresponding refracted ray in a single plane, i.e., a plane that serves as the (principal) plane representing the thin lens. Now, if in Fig. 1.49 we extend forward the ray incident parallel to the optical axis and extend backward the corresponding refracted ray after the lens, to the point where the two rays intersect, we can identify a plane orthogonal to the optical axis that contains the point of intersection. A thin lens can be placed in that plane, omitting the actual lens, with a focal length that is equal to the distance between that plane and the secondary focal point. With this construction, the refraction of an incident ray parallel to the optical axis in both the real lens and the thin lens would be equivalent; the ray would converge toward the secondary focal point. The plane in which the equivalent thin lens can be placed is called the secondary principal plane and is denoted by H′, as in Fig. 1.50, and the focal length of such a thin lens will be the focal length f of the real lens, which is given by*

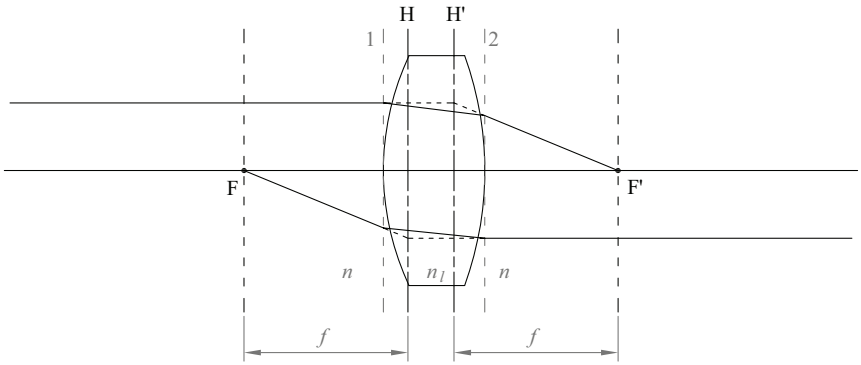


Figure 1.50 Establishing focal points and focal lengths in a real lens. The focal lengths are measured from the principal planes H and H′.

*In Appendix A, this expression for the focal length of a lens immersed in air is derived. The derivation uses paraxial ray tracing equations.

$$\frac{1}{f} = (n_l - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{(n_l - 1)^2}{R_1 R_2}\frac{t}{n_l}, \tag{1.50}$$

where t is the thickness of the lens. The primary focal point F can be obtained with a similar strategy by sending a ray (parallel to the optical axis) that travels from right to left toward the lens. The focal length obtained in such a case is equal to the one computed from Eq. (1.50) but measured from the primary principal plane, denoted by H, as shown in Fig. 1.50.

A real lens can be represented using the principal planes. From these, we can measure the object distance and the image distance, as illustrated by Fig. 1.51. By comparing the geometry of Figs. 1.41 and 1.51, we can establish that the Gaussian and Newtonian equations for lenses, Eqs. 1.42 and 1.47, are also valid in real lenses, although the object and image distances are to be measured from the principal planes and the lens focal length is given by Eq. (1.50).

In a thin lens, a ray directed to the center of the lens does not deviate. In the real lens, there is an equivalent ray but not with a straight path line. There is a ray that leaves the object tip and is directed to the point where the optical axis intersects the primary principal plane. The point of intersection is denoted by P and is called the primary principal point. Such a ray is refracted, after the lens, traveling with the same inclination of the incident ray as if it were leaving the point where the optical axis intersects the secondary principal plane. That point of intersection is denoted by P′ and is called the secondary principal point (see Fig. 1.51). The actual trajectory of the ray within the lens will, of course, be a straight line going from the point where the incident ray intersects the first surface to the point where the refracted ray intersects the second surface.

When the lens is immersed in two media of different refractive indices, i.e., the object is in one medium and its image is in another (both with a

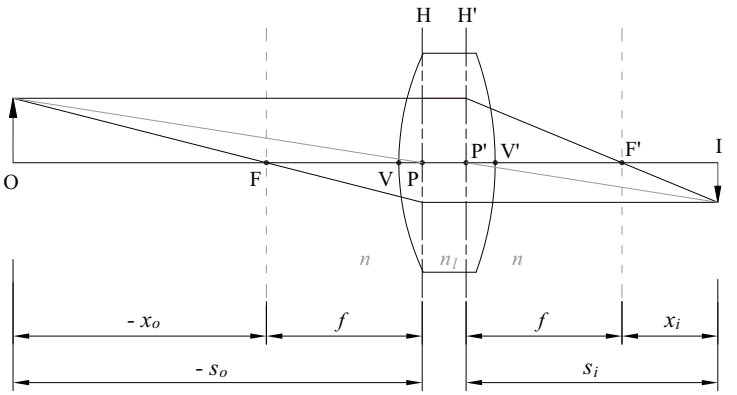


Figure 1.51 Object and image distances in a real lens with respect to its principal planes.

different refractive index than the lens), the oblique ray that does not change the inclination when refracted no longer passes through the principal points. However, another pair of points can be defined on the optical axis that allow us to know the direction of this oblique ray. These points are called nodal points, and the ray is called the nodal ray. An example of lenses immersed in different media is found in underwater cameras (water–glass–air). In cases like these, Eqs. (1.42) and (1.47) are no longer valid. We will not consider these types of cases in this book. We will deal with lenses immersed in the same medium, so that the nodal points and the principal points coincide.

The position of the principal planes is usually given by the distance between the vertex of the refracting surface and the principal plane (Fig. 1.51; see also Appendix A). For the plane H,

$$\overline{\mathrm{VP}} = -f\frac{(n_l - 1)t}{R_2 n_l}, \tag{1.51}$$

and for H′,

$$\overline{\mathrm{V'P'}} = -f\frac{(n_l - 1)t}{R_1 n_l}. \tag{1.52}$$

In Fig. 1.52, the location of the principal planes is shown for some positive and negative lens configurations. In symmetrical lenses, the principal planes are also symmetrical; in lenses with a flat surface, one of the principal planes is

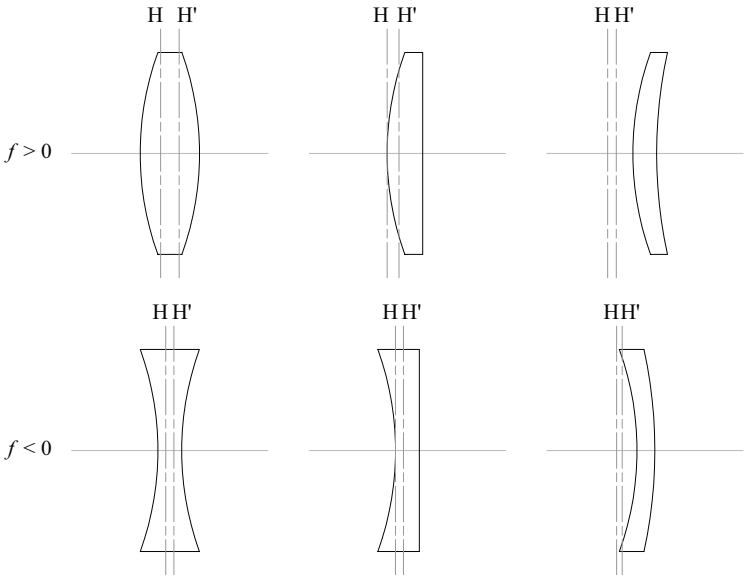


Figure 1.52 Principal planes in some positive and negative lenses.

located at the vertex of the other surface; and in meniscus-type lenses, the principal planes may be outside the lens.

Example: from a thin lens to a real lens

Suppose we have a complementary metal-oxide semiconductor (CMOS) video camera with a sensor that has an image registration region of 4.5×2.8 mm (1/3″ format, 752×480 pixels). We want to use this camera to record images of objects whose size $2h_o$ is around 11 mm. The working distance between the camera sensor and the object should be close to 200 mm, but not more. What single lens can be used so that the image is recorded in the largest possible region of the sensor?

One solution to the problem may be to start with a thin lens design and then replace it with a real lens that is commercially available.

The system magnification, which is estimated from the ratio between the smallest dimension of the sensor and the largest size of the object, i.e., $2.8/11 = 0.254$, must be established first. Based on this result, let us use magnification $m_t = -1/4$. Then we can determine the distances from the object and the image to the thin lens. The object distance should be $x_o = f/m_t = -4f$, and the image distance should be $x_i = f/4$. Thus, $s_o = -5f$ and $s_i = 5f/4$. The sum of these two distances must be such that $s_i - s_o = 6.25f \leq 200$ mm. Using the equality limit, the focal length should be 32 mm.

In the catalog of a lens manufacturing company, we find that the closest single lens to the one we are looking for is one with a 30 mm focal length. With $f = 30$ mm, the working distance would be 187.5 mm, which is suitable. Therefore, using the thin lens approximation, a solution would be as follows: use a lens of focal length $f = 30$ mm, with the object at distance $s_o = -150$ mm and the sensor at distance $s_i = 37.5$ mm. With this, magnification $m_t = s_i/s_o = -1/4$ is obtained and an object of size of 11 mm would be registered in the CMOS camera with a size of 2.75 mm. This solution, using the thin lens approximation, is shown in Fig. 1.53(a).

For a practical solution, we should replace the thin lens with a real lens of 30 mm focal length. The lens selected from a company's catalog is a convex-flat lens (30 mm in diameter) of thickness $t = 6.90$ mm. The radius of the convex surface is $R_1 = 23.54$ mm. The lens material is N-SF11 glass from Schott AG. For the wavelength $\lambda = 587.56$ nm,* the refractive index is 1.7847. The principal planes lie at $\overline{VP} = 0$ and $\overline{V'P'} = -3.8662$ mm. Therefore, if the principal plane H is made to coincide with the position of the thin lens, the distance between the object and the vertex of the first surface is 150 mm. Since the principal plane H′ is separated by 3.03 mm from H, the plane in which the image will be at (where the sensor should be placed) needs to be moved to the

*This wavelength corresponds to spectral Fraunhofer line *d*. See Appendix C.

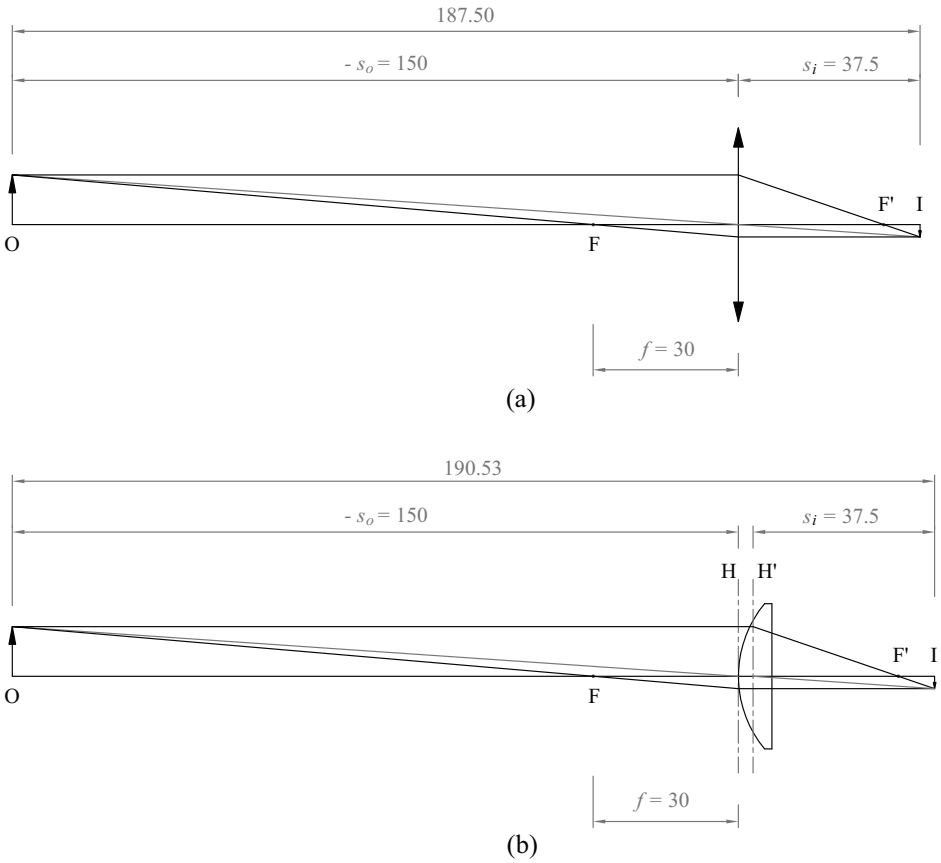


Figure 1.53 Imaging with a single lens that generates magnification $m_t = -1/4$. (a) A simple solution using the thin lens approximation. (b) The thin lens is exchanged for a real (commercial) convex-planar lens with a refractive index of 1.7847, radius of the convex surface $R_1 = 23.54$ mm, and thickness $t = 6.90$ mm.

right by the same distance, as shown in Fig. 1.53(b); i.e., the total working distance between the object and the sensor increases by 3.03 mm (it is worth noting that this distance increment is not equal to the thickness of the lens). At last, a practical solution has been obtained, achieving a $-1/4$ magnification in a working distance (object-sensor) of 190.53 mm with a lens of 30 mm focal length.

1.6.1 A lens system

In the previous section, the principal planes of a simple lens are defined. The great advantage of introducing the principal planes is that to determine the position and size of an image, we can use the Gaussian and Newtonian equations for a thin lens. Moreover, the concept of principal planes can be generalized to a lens system and is defined in the same way as for simple

lenses. The Gaussian and Newtonian equations are still valid in a multi-lens system, but now the focal length will be an effective focal length for all the lenses in the system. The positions of the principal planes are no longer calculated with Eqs. (1.51) and (1.52).

Example: a three-lens system

To illustrate the construction of the principal planes in a multi-lens system, consider the example given in Table 1.1, which lists a series of surfaces that include the object plane and the image plane. The radius of curvature is indicated for each surface element. The thickness or separation between one surface and the other is also provided, as well as the refractive index of the medium between the surfaces for the wavelength $\lambda = 587.56$ nm. Radius and thickness parameters are given in millimeters as is customary in optical design. The distance between the object plane and the first refractive surface, to the right of the object, is denoted by l_o; the distance between the last refractive surface and the image plane is denoted by l_i. Note that we have not used the notations s_o and s_i because these would represent the distances to the principal planes; the distances shown in Table 1.1 are the distances between surfaces.

As shown in Table 1.1, the first lens (surfaces 1 and 2) is convex-planar with focal length $f_1 = 30$ mm and thickness $t = 4.5$ mm. The second lens (surfaces 3 and 4) is biconcave with focal length $f_2 = -20$ mm and thickness $t = 2.0$ mm, and the third lens (surfaces 5 and 6) is biconvex with focal length $f_3 = 25$ mm and thickness $t = 8$ mm. The focal length of this lens system is $f = 40.17$ mm.

Table 1.1 Three-lens optical system. Units for radius and thickness are given in millimeters.

Surface	Radius	Thickness	Refractive index
0 (Obj)	∞		
		l_o	1.0000
1	20.10		
		4.5	1.6700
2	∞		
		12.0	1.0000
3	− 31.82		
		2.0	1.7847
4	31.82		
		7.0	1.0000
5	31.94		
		8.0	1.6727
6	− 31.94		
		l_i	1.0000
7 (Imag)	∞		

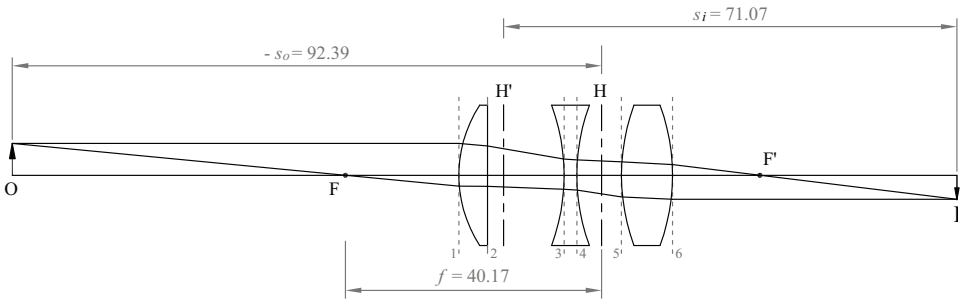


Figure 1.54 Location of the principal planes in a three-lens system.

In Fig. 1.54, the optical system is shown when imaging an object that is at the distance $l_o = -70$ mm. In this case, the image will be at $l_i = 44.60$ mm. Note that in this example the principal planes have a different order than in simple lenses (Fig. 1.52). Taking this into account, the object and image distances for the Gaussian equation are $s_o = -92.39$ mm and $s_i = 71.07$ mm.

From this example it is clear that, in general, an optical system can be represented by the object and image planes, the primary and secondary focal planes, and the primary and secondary principal planes.

The focal length in a lens system is determined in the same way as for a single lens. By using an incident beam of rays coming from infinity that is parallel to the optical axis, we should be able to determine a point where the refracted rays would converge after leaving the system, if the system acts as a positive lens, or the point from which the rays diverge, if the system acts as a negative lens. The point of convergence or divergence is the secondary focal point F′. The distance between the secondary principal plane H′ and F′ is the focal length of the system. In particular, for a system of two thin lenses separated by a distance d,*

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{1}{f_1}\frac{1}{f_2}d. \tag{1.53}$$

The power of the jth thin lens is defined as

$$\tilde{P}_k = \frac{1}{f_k}. \tag{1.54}$$

With this definition, the power of the lens combination given by Eq. (1.53) can be written as

$$P = \tilde{P}_1 + \tilde{P}_2 - \tilde{P}_1\tilde{P}_2 d. \tag{1.55}$$

*The expression for the focal length of a system of two thin lenses is derived by using paraxial ray tracing equations (see Appendix A).

This equation is analogous to the power of a real lens of thickness t and refractive index n_l, also defined as $1/f$. Using the definition of power given by Eq. (1.28) for a refractive surface, Eq. (1.50) can be rewritten as

$$P = P_1 + P_2 - P_1 P_2 (t/n_l). \tag{1.56}$$

Indeed, Eqs. (1.55) and (1.56) are analogous and very useful in practice. It is straightforward to use Eq. (1.56) for the case of two thin lenses, where $t = d$ and $n_l = 1$.

1.7 Stops and Pupils

So far, we have dealt only with the geometrical aspect of optical imaging in which the ideal image of an extended object would be a faithful copy of the object, except for a scale factor. However, for a complete description of image formation, the energy distribution in the image should also be analyzed in terms of the energy distribution of the object. Since an extended object can be modeled as a set of point objects, we are going to limit the energy flux analysis within an optical imaging system to point objects and point images, located on the optical axis, using the paraxial approximation. A point object, by definition, is an isotropic energy radiator that emits energy radially and at the same flux rate in any direction. Thus, from the point of view of light energy, the image of a point object would be a faithful copy of the object if it were possible to collect all the energy emitted by the object in the image. For example, a point object located in one of the foci of an ellipsoidal mirror will have its image (also a point) in the other focus of the ellipsoid (Fig. 1.15). In such a case, omitting the fact that there would be a small amount of energy absorption in the mirror, we would say that the mirror makes a perfect image of the object. But when imaging with a lens, the edge of the lens will determine the amount of energy that can be collected to form the image. In this sense, we cannot have a faithful copy of the object.

The *intensity* for a point source is defined as flux of radiant energy dF [in watts (W)] in a given direction contained in a certain solid angle $d\Omega$ [in steradians (sr)], i.e.,

$$\tilde{I} = \frac{dF}{d\Omega}. \tag{1.57}$$

For a point source that emits a constant flux, $F = 4\pi\tilde{I}$, so $\tilde{I}$ (in W/sr) is also constant.

Figure 1.55 shows the imaging of a point object through a single lens. The energy flux that leaves the object is contained in a cone whose vertex is the object and whose base is the lens; the energy flux that reaches the image is contained in the cone whose vertex is the image and whose base is also the

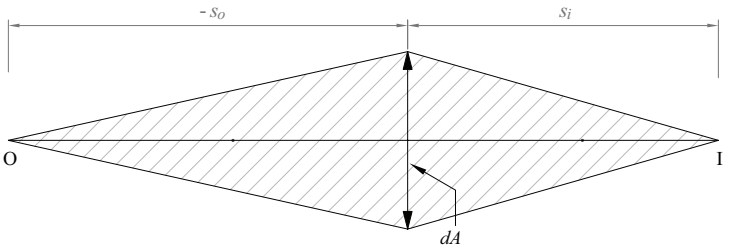


Figure 1.55 The amount of light energy collected by a lens is determined by the edge of the lens.

lens. If the area of the lens is dA, the solid angle that the lens subtends with respect to the object will be $d\Omega_o = dA/s_o^2$, while the solid angle that the lens defines with respect to the image will be $d\Omega_i = dA/s_i^2$. Assuming a radiant intensity at the point object of $\tilde{I}_o$, the energy flux that the lens can collect will be

$$dF = \tilde{I}_o \frac{dA}{s_o^2}. \tag{1.58}$$

In absence of energy losses by absorption in the lens, the same energy flux dF will reach the point image. If $\tilde{I}_i$ is the intensity of the image point, then $dF = \tilde{I}_i dA/s_i^2$. Thus, the radiant intensity at the point image will be [10]

$$\tilde{I}_i = \tilde{I}_o \frac{s_i^2}{s_o^2}. \tag{1.59}$$

1.7.1 Aperture stop

If we want to reduce the energy flow that reaches the image, we need to reduce the area of the lens through which the energy flows. This can be achieved by inserting into the optical system an opaque screen with a circular diaphragm, centered on the optical axis, in front of or behind the lens, as shown in Fig.1.56.

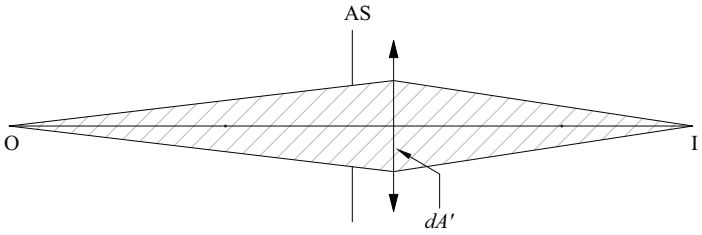


Figure 1.56 Control of the flow of energy in the image by means of an aperture stop (AS) in the optical system.

In this case, the energy flow that reaches the image will be

$$dF' = \tilde{I}_i \frac{dA'}{s_i^2}, \tag{1.60}$$

where dA' is the area that the aperture (seen from the object) projects onto the lens. This type of diaphragm is called an *aperture stop* (AS). If the radius of the stop is variable, e.g., in an iris diaphragm, the amount of energy entering the imaging system can be controlled. If the stop radius is fixed, the amount of energy entering the imaging system will also be fixed.

Suppose the circular region defined by dA' in the lens has a diameter D, so that $dA' = \pi D^2/4$ and the energy flow to the image is

$$dF' = \frac{\pi \tilde{I}_i}{4} \left(\frac{D}{s_i}\right)^2. \tag{1.61}$$

Since $\tilde{I}_i$ is fixed, the energy flux is proportional to the square of D/s_i. A particularly interesting situation arises when the point object is at infinity, as illustrated in Fig. 1.57. In such a case, the energy flux will be proportional to the square of D/f, where f is the focal length of the lens. The quantity (f/D) is called the *f-number* and is denoted by

$$f/\# = \frac{f}{D}. \tag{1.62}$$

This quantity is commonly used in astronomy to characterize the objective lenses (or mirrors) of telescopes. In photography, the f-number is an indicator to establish exposure time. In terms of $f/\#$, the energy flux is

$$dF' \propto \frac{1}{(f/\#)^2}. \tag{1.63}$$

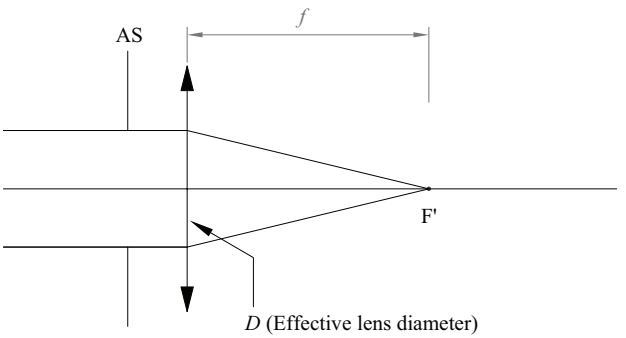


Figure 1.57 The quantity f/D is a parameter that is often used to characterize lenses (or mirrors) in relation to the amount of energy that an optical system can capture.

The lenses of manual cameras usually have a ring with markings denoting the *f*/# as *f*/1.4, *f*/2, *f*/2.8, *f*/4, *f*/5.6, *f*/8, etc. These numbers correspond to the square root of 2, 4, 8, 16, 32, and 64. For a fixed focal length, by consecutively switching from one *f*/# to another (by turning the ring), one can increase or decrease the effective area on the lens by a factor of 2; e.g., by switching from *f*/1.4 to *f*/2, the power flow at the image is cut in half. Because the energy flow is measured as energy per unit of time, the same amount of light energy will reach the image if the image is recorded, for example, with *f*/1.4 at 1/500th of a second, or with *f*/2 at 1/250th of a second. In Fig. 1.58, a camera lens from Nikon (AF Nikkor 28 mm f/2.8D) is shown with a focal length $f = 28$ mm and an *f*/# ranging from *f*/2.8 to *f*/22. For example, *f*/4 means that $f/4 = 28/D = 4$, and therefore $D = 7$ mm.

When the point object is off-axis, e.g., the end point of an extended object, the energy cone is not symmetrical and the cross-sectional area (orthogonal to the axis of the cone) projected onto the lens (dA'' in Fig. 1.59) will be smaller than the area dA' when the point object is on the axis (Fig. 1.56, assuming the same AS). Therefore, if all the points of an extended object have the same radiant intensity, the radiant intensity at the image points will not be the same; the intensity decreases as we move away from the optical axis at the image plane. This phenomenon is called *vignetting*.

In the paraxial approximation, we assume that the rays leaving an object toward a lens travel with very small inclinations, i.e., practically parallel to the lens optical axis. This implies that the size extent of an object must be much less than the object distance s_o. With this in mind, we can approximate the energy cones of the off-axis point objects (which make up the extended object) from the energy cone of the point object on the axis.

Figure 1.58 Objective of a 28 mm focal length photographic camera from Nikon. The *f*/# ranges from *f*/2.8 to *f*/22. Image reprinted from https://imaging.nikon.com/lineup/lens/f-mount/singlefocal/wide/af_28mmf_28d/ with permission from Nikon.

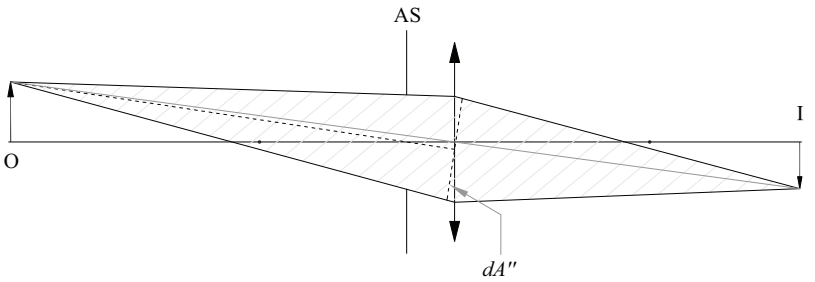


Figure 1.59 For an off-axis point object, the energy cone is not symmetrical. However, using the paraxial approximation, the energy cone of any point on the object (in an extended object) can be taken as approximately equal to the energy cone of the point object that is located on the axis. Within this approximation, the dotted ray in the figure would pass through the center of the AS.

The quantity

$$I = \frac{dF}{dA} = \frac{\tilde{I}}{s^2} \tag{1.64}$$

(in watts per square meter) is called *irradiance* and is a measure of the energy distribution in a flat region of area dA. Thus, the irradiance can characterize the image of extended objects. This equation also describes the inverse square distance law for irradiance, where s is the distance from the point source.

1.7.2 Pupils

In Eq. (1.61), we expressed the flow of energy that reaches the image in terms of the effective diameter of a lens, which results from the projection of an AS. Such an effective diameter is the diameter of the same circular base for both the energy cone of the object and the energy cone of the image. Another more convenient way to define the energy cones is by using the images of the AS generated by the optical system, i.e., the images seen through the lenses located before the AS and through the lenses located after the AS. These images are called entrance and exit pupils, respectively. The *entrance pupil* defines the base of the energy cone for the object, and the *exit pupil* defines the base of the energy cone for the image.

Figure 1.60 illustrates the concept of entrance (P_{en}) and exit (P_{ex}) pupils in an optical system with two positive lenses and an AS in the middle of the lenses. The entrance pupil is the image of the AS made by lens 1. The exit pupil is the image of the AS made by lens 2. Note that in both cases, these images are virtual because the AS is between a lens and its focal point. Denoting the diameters of the entrance and exit pupils by D_{en} and D_{ex}, respectively, and denoting the distances between the entrance pupil and the object and between the exit pupil and the image by $-l_{en}$ and l_{ex}, respectively, the energy flux will be given by

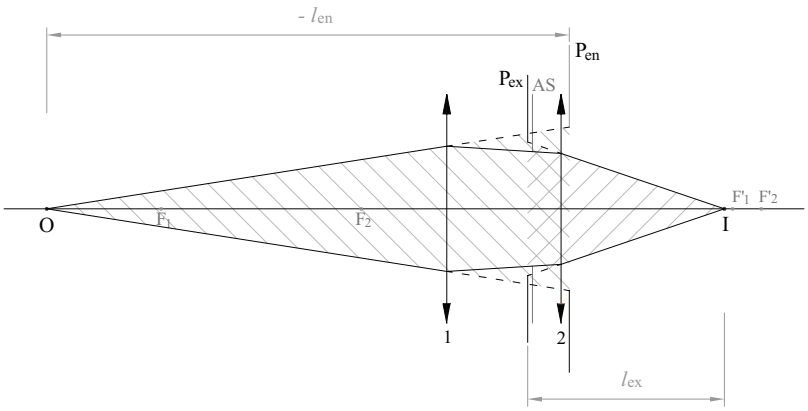


Figure 1.60 The entrance pupil P_{en} and exit pupil P_{ex} determine the base of the effective cones of illumination for an object and its image in an optical system.

$$dF' = \frac{\pi \tilde{I}_o}{4}\left(\frac{D_{en}}{l_{en}}\right)^2 = \frac{\pi \tilde{I}_i}{4}\left(\frac{D_{ex}}{l_{ex}}\right)^2. \tag{1.65}$$

And the effective *f*/# of the system is defined as

$$f/\# = \frac{l_{ex}}{D_{ex}}. \tag{1.66}$$

While the size of the AS determines the amount of energy that enters the optical system, the position of the stop (along with its size) plays a very important role in controlling the optical aberrations of the system. Therefore, the location of the AS must meet the criteria that optimize image quality.

Example: identifying the AS in an optical system
In Table 1.1 (Section 1.6.1), which describes the optical system of Fig. 1.54, the AS is not evident. However, even if it is not explicitly stated, the AS will be the edge of one of the lenses. To determine that edge, each lens edge should be imaged through the optical elements preceding it. The image that subtends the smallest solid angle with respect to the point object defines the edge that corresponds to the AS. In the case of Table 1.1, the edge of surface 1 (first lens) is the AS, as shown in Fig. 1.61. But using the edge of the first lens as the AS is not practical because the amount of illumination reaching the image cannot be controlled and, moreover, such a stop would have a relatively large aperture, which can affect image quality with optical aberrations.

An improvement to the optical system shown in Fig. 1.61 is obtained by inserting an AS in the middle of the lenses, as shown in Fig. 1.62.

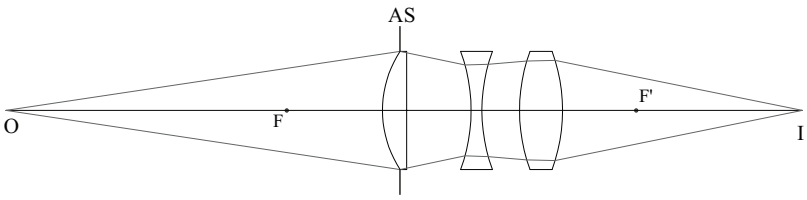


Figure 1.61 The edge of the optical element whose image subtends the smallest solid angle with respect to the on-axis point object is the AS.

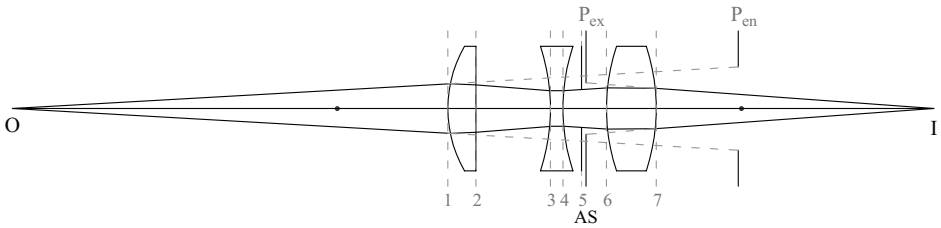


Figure 1.62 An AS outside the lens edges is counted as another surface of the optical system.

By introducing the AS as a separate element outside the lens edges, the stop is counted as an additional surface of the optical system. Thus, Table 1.1 becomes Table 1.2, and surface 5 is the AS at 3 mm from the fourth surface. Radius and thickness parameters are expressed in millimeters.

1.7.3 Marginal and chief rays

Suppose we have an extended object to image through the optical system of Fig. 1.60. This situation is illustrated in Fig. 1.63. Two new rays are defined from the entrance and exit pupils: the *marginal ray*, which passes through the edge of the AS and allows the image to be located, and the *chief ray*, which passes through the center of the AS and determines the image size. To trace these two rays, we follow the following procedure:

- Marginal ray: It exits the base of the object directed toward the edge of the entrance pupil, is refracted in the optical system, passes through the edge of the stop, and exits the optical system as if it were coming from the edge of the exit pupil. The intersection of the marginal refracted ray with the optical axis locates the base of the image.
- Chief ray: It leaves the object tip directed toward the center of the entrance pupil, is refracted in the optical system, passes through the center of the stop, and leaves the optical system as if it were coming from the center of the exit pupil. The intersection of the refracted chief

Table 1.2 Three-lens optical system with an AS. Radius and thickness parameters are in millimeters.

Surface	Radius	Thickness	Refractive index
0 (Obj)	∞		
		l_o	1.0000
1	20.10		
		4.5	1.6700
2	∞		
		12.0	1.0000
3	− 31.82		
		2.0	1.7847
4	31.82		
		3.0	1.0000
5 (AS)	∞		
		4.0	1.0000
6	31.94		
		8.0	1.6727
7	− 31.94		
		l_i	1.0000
8 (Imag)	∞		

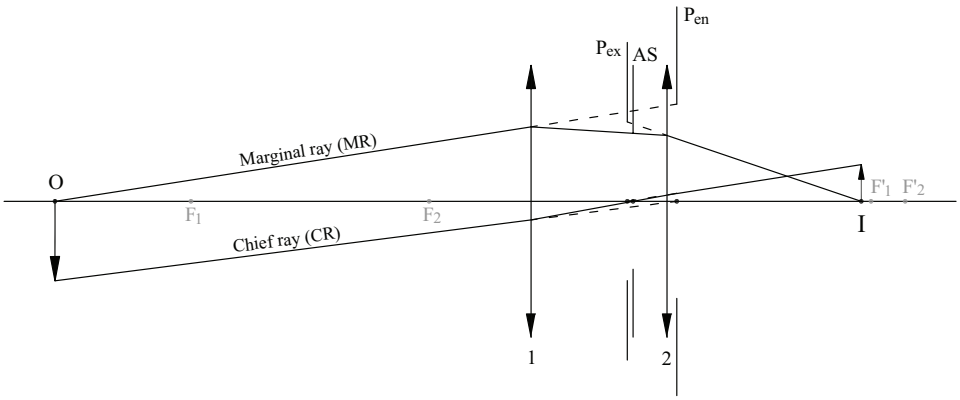


Figure 1.63 Defining chief and marginal rays.

ray with the line normal to the optical axis, at the base of the image, gives the height of the image.

These two rays can also be traced for a point object. For example, for the point object shown in Fig. 1.60, the chief ray coincides with the optical axis, and for the point at the object tip shown in Fig. 1.59, we have a marginal ray going to the upper edge of the AS and a marginal ray going to the lower edge

of the AS. Thus, for a point object, whereas the marginal ray defines the extent of the illumination cone entering the optical system, the chief ray determines the main axis of the illumination cone.

1.7.4 Field stop, field of view, and angle size

In optical systems there is another type of stop that determines the region within which the image is observed. This type of stop is called a *field stop* (FS). This stop can be in the object plane (not very common), in the image plane (as in a photographic camera), in an intermediate plane where there may be a first image (as in a microscope or a telescope), or it can even be the edge of a lens. The images seen through the lenses located before the FS and through the lenses located after the FS are called *entrance* and *exit windows*, respectively,

Consider an optical system such as in a photographic camera with a fixed focal length. A simplified schematic of the camera's optics and sensor is shown in Fig. 1.64. Since the camera is designed to take pictures of objects at distances much greater than the lens focal length, the image plane coincides with the secondary focal plane. The image is recorded with an irradiance sensor [photographic film, CMOS solid state device, or charge-coupled device (CCD)] within the region defined by the edge of the sensor, usually rectangular in geometry. This edge is the top of the FS of the optical system and sets the angular size of the image α', which is defined as the angle between the chief refracted ray and the optical axis, i.e.,

$$\tan\alpha' = h'_{\max}/l_{ex}, \tag{1.67}$$

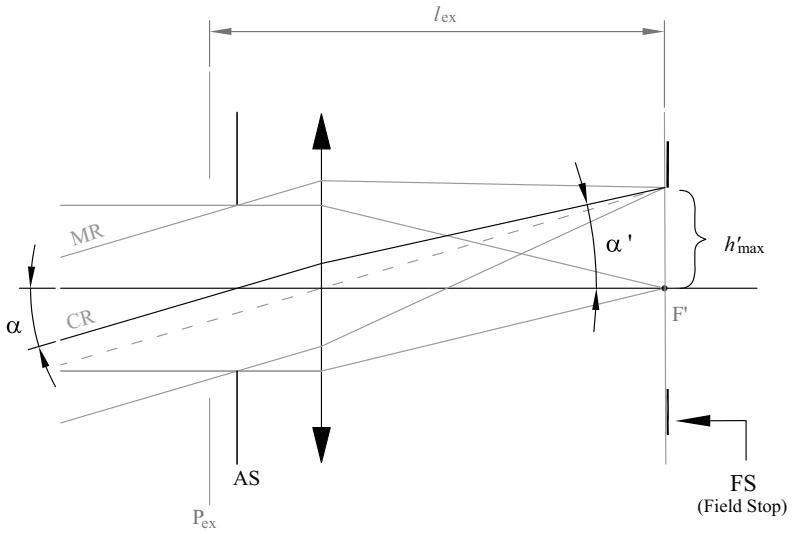


Figure 1.64 The FS determines the region within which the image can be viewed.

where h'_{max} is the maximum height that the image can have (measured from the optical axis). Since α' is the maximum angle that subtends the image (with respect to the center of the exit pupil), the angle $2|\alpha'|$ is called the *field of view*, usually denoted in optical instruments by the acronym FOV and given in degrees.

On the side of the object, the *object field* is also defined as the angle $2|\alpha|$, where α is the angle formed by the incident chief ray and the optical axis; i.e., α is the angle that the object subtends with respect to the center of the entrance pupil.

We have seen that the AS determines (through the pupils) the illumination cones and the marginal ray delimits the surface of these cones. Similarly, the FS determines the field of view cones and the chief ray delimits the boundary of these cones. If we define the *object field cone* as the region corresponding to the object field and the *image field cone* as the region corresponding to the field of view, then the object field cone has its vertex at the center of the entrance pupil and its base at the entrance window, and the field of view cone has its vertex at the center of the exit pupil and its base at the exit window. In Fig. 1.64, the exit window coincides with the FS. The entrance window is at infinity in the object space.

Example: energy and field of view cones

Consider a photographic camera objective consisting of a negative lens with focal length $f_1 = -40$ mm and a positive lens with focal length $f_2 = 40$ mm, separated by 40 mm. Both lenses have a diameter of 40 mm. The image of the optical system is recorded on a 35 mm (horizontal) $\times$ 24 mm (vertical) film. The maximum image height will be half the vertical dimension of the film, i.e., 12 mm, as illustrated in Fig. 1.65.

Assuming that the objects to be registered in the photographic film are at distances much greater than the effective focal length of the two-lens system, i.e., $s_o \rightarrow -\infty$, the image plane must be located at the secondary focal point. To determine the secondary focal point and the focal length, let us examine the ray tracing of an incident ray, coming from left to right, that is parallel to

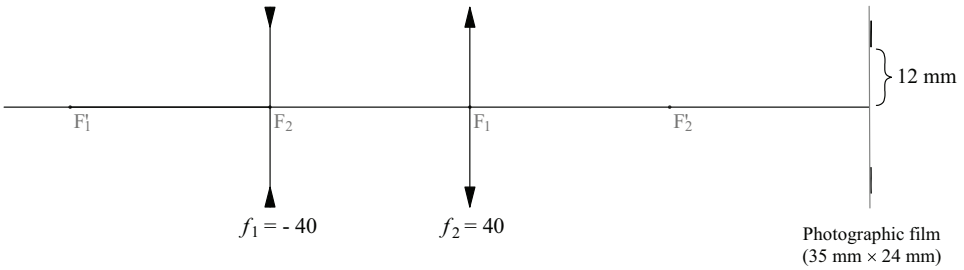


Figure 1.65 Two-lens optical system for capturing images on a 35 mm $\times$ 24 mm photographic film.

the optical axis. After refracting through the lenses, this ray intersects the optical axis of the system at the point that determines the secondary focal point F′. On the other hand, the intersection of the forward extension of the incident ray and the backward extension of the refracted ray (in the image space) will locate the principal plane H′. The distance from H′ to the focal point turns out to be $f = 40$ mm. The focal point F and the principal plane H can be located in a similar way, by examining the refraction of another incident ray, from right to left, as shown in Fig. 1.66.

The FS is the edge of the photographic film, and the image height is $h'_{\max} = 12$ mm. What is the AS? Since the object is at infinity, if we first trace the marginal ray to the edge of the first lens, this ray would be refracted, diverging, and would not pass through the second lens. Therefore, the second lens is the one that limits the energy cone. The surface boundary of that cone can be outlined by using a ray parallel to the optical axis, coming from left to right, as shown in Fig. 1.66. Hence, the edge of the second lens is the AS. The image of this stop seen from the object side is the entrance pupil, which is virtual and smaller (half in size) than the AS. Since there are no more lenses between the second lens and the image plane, the edge of the second lens is also the exit pupil, i.e., a real image of the same size as the AS, as shown in Fig. 1.67(a). The exit pupil distance would then be $l_{ex} = 80$ mm.

On the other hand, the chief ray determines the height of the image, which would be $h'_{\max}$ in this case. Therefore, we can determine the field of view as $2|\alpha'| = 2\tan^{-1}(12/80) = 17.06^{\circ}$. Because the size of the image is already known, the trace of the chief ray in the optical system can be performed from right to left. The result is shown in Fig. 1.67(a). The object field turns out to be $2|\alpha| = 33.3^{\circ}$.

Finally, in Fig. 1.67(b) the illumination and field of view cones are shown. The energy that is inside the illumination cone will reach the image. Objects that are within the object field cone will have their image within the image field cone. Changing the AS size would modify the energy, but there would be no effect on the image extension because the chief ray passes through the

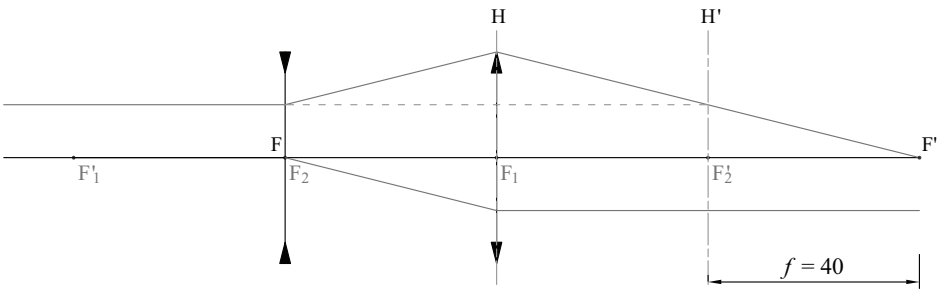


Figure 1.66 Principal planes and focal length of the two-lens optical system from Fig. 1.65.

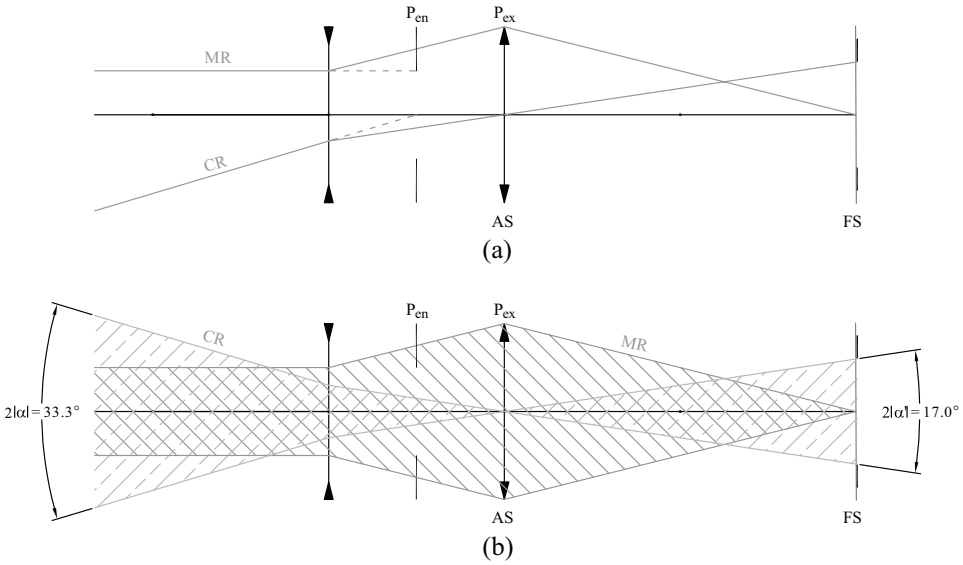


Figure 1.67 Energy and field of view cones for the two-lens optical system from Fig. 1.65. (a) Chief and marginal ray tracing. (b) Whereas the surface boundary of the energy cone is defined by the marginal ray, the surface of the field of view cone is determined by the chief ray.

center of the AS. By resizing the FS, the image extension would be modified, but there would be no effect on the amount of energy reaching each point in the image because the marginal ray passes through the center of the FS.

1.8 Some Optical Instruments

In this section, we will look at the basic setup of some optical instruments and describe their operation in terms of topics covered in the previous sections. First, we will view the human eye as a system of spherical surfaces with symmetry of revolution and simplify it to a thin lens and image plane. Next, we will see the magnifying glass coupled with the thin lens model of the eye. Finally, we will study the telescope and the microscope as extensions of the magnifying glass, i.e., when another lens is added to form the image of very distant objects or very close but small objects.

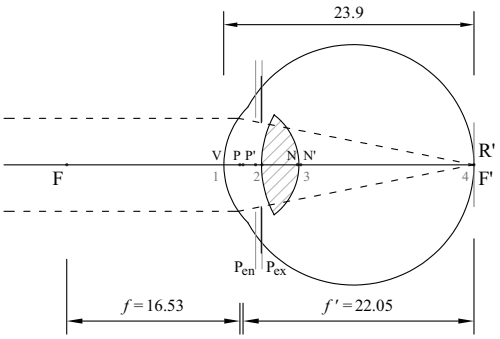
1.8.1 The human eye (schematic representation)

In Section 1.3.1, an overview of the eye as an optical system with various refracting surfaces is given. Due to the existing variation of the characteristics of the eye from one person to another, to study the formation of an image in the eye, models with ocular parameters that represent the mean values of the population are used. This can be done with varying degrees of precision. If the refracting surfaces of the eye are assumed to be spherical and centered on

the same axis (optical axis), a family of models is obtained, the so-called schematic eyes [9]. In particular, they are used in the paraxial region to obtain information such as refractive power, magnification, illumination on the retina, Purkinje images (reflections on the refracting surfaces of the eye), location of the pupils, location of the focal points (F, F′), location of the principal points (P, P′) and the nodal points (N, N′), and the effects of refractive errors (myopia and hyperopia). When image information is required beyond the paraxial region, more refined models called wide-field schematic eyes are used, including nonspherical refracting surfaces, off-center refracting surfaces, and *Gradient Index* (Grin) models for the refractive index of the crystalline lens [9].

In this section, we will consider the Gullstrand–Emsley schematic eye in the relaxed and fully accommodated eye configurations, assuming the eye has no refractive errors (emmetropic eye). A relaxed eye is understood as the configuration that the eye has when observing an object located at infinity. In this situation, the crystalline lens has the longest focal length it can have. As the object approaches the eye, the crystalline lens changes its geometry to decrease the focal length so that the image remains focused on the retina. This process is called *accommodation*. Accommodation has a limit; thus, there is a distance limit from which, for distances less than this limit, the image on the retina can no longer remain in focus. The distance where the accommodation limit is located is called the *near point distance* (*npd*). The farthest distance, where the eye can see without accommodation, is called the *far point distance* (*fpd*). In an emmetropic eye, we have $fpd = \infty$. The range in diopters between *npd* and *fpd* is called the *accommodation range*. The *npd* varies with age and is lower in children. For example, a 10-year-old child has an *npd* of $\simeq 70$ mm and his accommodation range will be $[1000/(70\text{ mm}) - 1000/\infty] = 14.29$ D, whereas a 50-year-old adult has an *npd* of $\simeq 400$ mm and her accommodation range will be $[1000/(400\text{ mm}) - 1000/\infty] = 2.5$ D.

In Fig. 1.68, the Gullstrand–Emsley schematic eye is shown in the relaxed eye configuration [9]. The parameters R and d are expressed in millimeters, and the optical power is expressed in diopters. The optical system is represented by four surfaces: (1) cornea, (2) iris (AS) and anterior surface of the lens, (3) posterior surface of the lens, and (4) retina. The distances of the cardinal points (focal points F-F′, principal points P-P′, and nodal points N-N′) with respect to surface 1 are: $VF = -14.983$, $VF' = 23.896$, $VP = 1.550$, $VP' = 1.851$, $VN = 7.062$, and $VN' = 7.363$. The secondary focal point F′ coincides with the retina R′. The focal lengths are $f = 16.53$ and $f' = 22.05$. On the other hand, with respect to surface 1, the pupils are located at $VP_{en} = 3.052$ and $VP_{ex} = 3.687$. When examining the refractive power of each element $[P = (n' - n)/R]$, the cornea contributes approximately two-thirds of the total refraction of the eye. All distances are in millimeters.



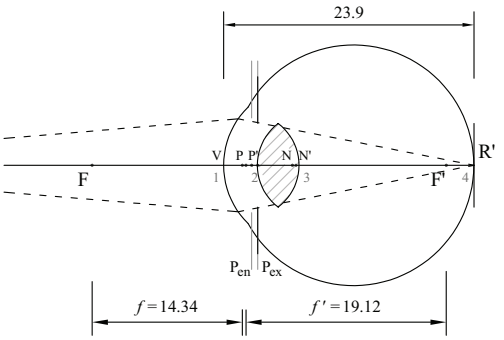
#	R	d	n	Equivalent power: Surface	Equivalent power: Component	Equivalent power: Whole eye
0 (Obj)	∞					
		∞	1.000			
1 Cornea	7.8			42.74	42.74	
		3.6	4/3			
2 Iris (AS)	∞			0		
		0	4/3			60.48
2 ----------	10.0			8.27		
Lens		3.6	1.416		21.75	
3 ----------	-6.0			13.78		
		16.69	4/3			
4 Retina						

Figure 1.68 Gullstrand–Emsley relaxed schematic eye (*fpd* = ∞) [9]. Parameters *R* and *d* are expressed in millimeters, and the optical power is expressed in diopters.

In Fig. 1.69, the Gullstrand—Emsley schematic eye is shown in the full accommodation configuration for the distance $npd = 116.3$ mm [9]. Although the secondary focal point F′ does not fall on the retina R′, the size of the eye remains the same as in the relaxed eye. It is worth noting the change in crystalline lens geometry, for which its radii are 5 and -5 mm; its thickness increases to 4 mm, increasing the lens optical power. The total power of the fully accommodated eye turns out to be 69.72 D, i.e., a change of 8.59 D from the relaxed eye. The distances of the cardinal points with respect to surface 1 are: $VF = -12.561$, $VF' = 21.252$, $VP = 1.782$, $VP' = 2.128$, $VN = 6.562$, and $VN' = 6.909$. The distance between F′ and R′ is $F'R' = 2.644$. The focal lengths are $f = 14.343$ and $f' = 19.124$. On the other hand, with respect to surface 1, the pupils are located at $VP_{en} = 2.674$ and $VP_{ex} = 3.249$. All distances are in millimeters.

Example: size of the moon's image on the retina

As an example of the use of the schematic eye, we are going to calculate the size of an object's image on the retina using the nodal ray. Suppose that a very distant object subtends an angle θ with respect to the nodal point N, as illustrated in Fig. 1.70. Considering the nodal point N′, $h' = \theta N'F'$. For example, the full moon subtends an angle of 0.52° (0.0091 rad) with respect to an observer on earth, so the size of the moon's image on the retina of the



#	R	d	n	Equivalent power: Surface	Equivalent power: Component	Equivalent power: Whole eye
0 (Obj)	∞					
		116.3	1.000			
1 Cornea	7.8			42.74	42.74	
		3.2	4/3			
2 Iris (AS)	∞			0		
		0	4/3			69.72
2 ----------	5.0			16.53		
Lens		4.0	1.416		32.30	
3 ----------	-5.0			16.53		
		16.69	4/3			
4 Retina						

Figure 1.69 Gullstrand–Emsley schematic eye in full accommodation for *npd* = 116.3 mm [9]. Parameters *R* and *d* are expressed in millimeters, and the optical power is expressed in diopters.

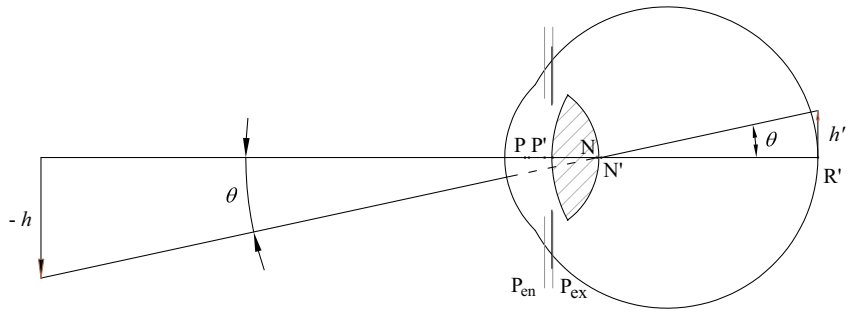


Figure 1.70 From the nodal ray, the size of an object's image on the retina of the schematic eye can be measured. For this, it is necessary to know the distance N′R′.

relaxed eye would be $0.0091 \times (\mathrm{VR}' - \mathrm{VN}') = 0.155\,\mathrm{mm}$. If we consider the eye in complete accommodation, the same angle is obtained with an object of height $-h = 1.12\,\mathrm{mm}$ at the distance of $npd + \mathrm{VN} = (116.3 + 6.562)\,\mathrm{mm} = 122.86\,\mathrm{mm}$. Then, to determine the size of the moon's image on the retina, the distance from the secondary nodal point to the retina and the angle that the object subtends with respect to the primary nodal point must be known. Note that because we are working in the paraxial approximation, the

image is not drawn on the curved surface of the retina, but rather on the orthogonal plane that passes through the vertex of the retina (R′).

The eye as a thin lens

To simplify the analysis below, we will use a simplified model of the eye: an optical system composed of a thin positive lens immersed in air, the edge of which is the AS (iris) and an image plane (retina) at a distance l_{eye} from the lens. When the eye is relaxed, $f_{eye} = l_{eye}$, and when the eye is in accommodation, $f_{eye} < l_{eye}$. For the design of ophthalmic lenses and visual instruments, there is a standard that establishes that in an emmetropic eye, $npd = 250$ mm (10 in). This is illustrated in Fig. 1.71: the situation of the relaxed eye (top) and the eye with maximum accommodation (bottom), according to the standard. In this case, the accommodation range would be 4D.

Myopia (nearsightedness). When $fdp < \infty$ in an eye, the eye is said to be myopic or nearsighted. The secondary focal point is located before the retina, and the image of an object at infinity on the retina is blurred. This also implies that the accommodation range decreases. The myopic eye can then focus retinal images of objects that are a finite distance between the *fdp* and *npd*. Since the myopic eye in a relaxed state focuses parallel rays before the retina, we can say that the eye has an optical power greater than that required when observing objects at infinity. Therefore, we can decrease the optical power with the help of a negative lens.

Hyperopia (farsightedness). An eye is said to be hyperopic or farsighted if the secondary focal point is behind the retina. Again, an object at infinity will

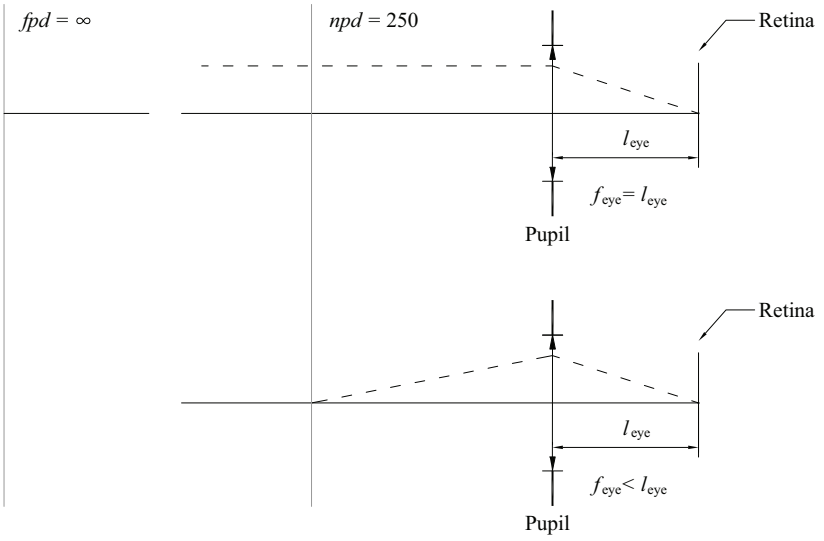


Figure 1.71 A simple representation of the eye consists of a thin positive lens immersed in air. The distance between the lens and the image plane is fixed and is denoted by l_{eye}.

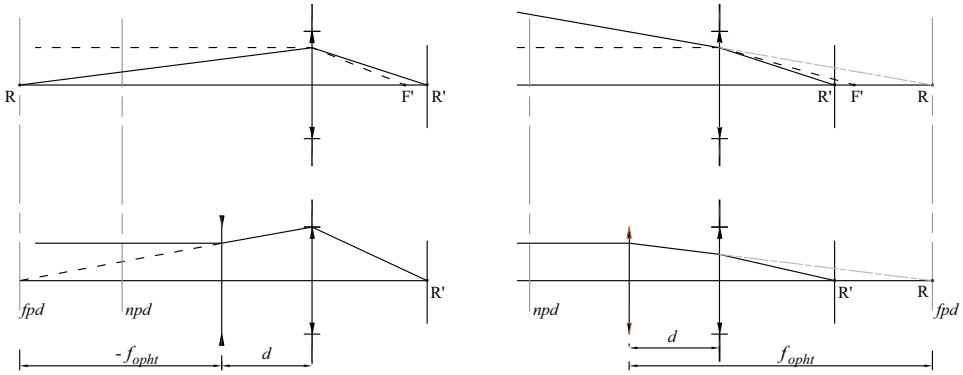


Figure 1.72 Refractive defects. (a) Myopia: the far point R is at a distance *fdp* < ∞. With a negative lens, this nearsightedness defect can be corrected. (b) Hyperopia: the far point R is behind the eye. With a positive lens, this farsightedness defect can be corrected.

be blurred on the retina. Since the hyperopic eye in a relaxed state focuses parallel rays after the retina, we can say that the eye has a lower optical power than required when observing objects at infinity. Therefore, we can increase the optical power with the help of a positive lens.

Figure 1.72 shows the myopic eye and its correction in (a) and the hyperopic eye and its correction in (b). The focal length of the negative or positive ophthalmic lens is indicated by f_{opht}. From the geometry of Fig. 1.72, f_{opht} can be determined if the distance d between the ophthalmic lens and the eye is known. This distance should ensure that the power of the eye does not change when the ophthalmic lens is added, since the image of an object seen with both eyes must be the same size even though the refractive error in each eye is different. Otherwise, we would have a complex situation for the brain, which would have to process two images of different sizes of the same object.

To determine the distance d, consider Eq. (1.53) for the focal length of two thin lenses: $1/f = 1/f_1 + 1/f_2 - d/(f_1 f_2)$. Let $f_1 = f_{opht}$ and $f_2 = f_{eye}$. If $d = f_{eye}$, then the focal length of the eye-ophthalmic lens combination will be $f = f_{eye}$. This is a remarkable result, which in a way is guaranteed by the physiognomy of our face; the nose not only supports eyeglasses, but it also helps to ensure that the distance between the eyeglasses and the eye is close to the focal length of the eye (≈20 mm).

In practice, a person with $fdp < 0.167$ m (6 D) is considered very myopic. In the range 0–6.0 D, the distance d can be omitted in the calculation of the ophthalmic lens power, $P = 1/f_{opht}$, so that the focal length of the ophthalmic lens can be equaled by $-fdp$. For example, a person with a myopia of -2 D will have a far point located at $fdp = 0.5$ m, so the ophthalmic lens needed to correct or improve the reduced vision of this person will be a negative lens of

focal length $f_{opht} = -500$ mm (since $500\,\text{mm} \gg f_{eye} = 22\,\text{mm}$). In the case of hyperopia, there is an analogous situation.

1.8.2 Magnifiers

In Section 1.5.3, we discuss that for a positive lens, the image of an object placed between the primary focal point and the lens is virtual, upright, larger than the object, and located to the left of the object. This is the layout of a magnifying glass (or simple microscope).

Before discussing the operation of a magnifying glass in detail, let us review the field of view and the object field for a lens (L) placed in front of the eye. The lens–eye system is shown in Fig. 1.73. Let us assume that the eye is emmetropic and we model it as shown in Fig. 1.71. If the diameter of the lens is much larger than the iris aperture, the iris will be the AS and, therefore, the edge of the lens will be the FS because it determines the region that the eye can see.

In this case, the entrance pupil is a real image, inverted and larger than the iris aperture and located to the left of the primary focus of L. On the other hand, the iris is the exit pupil. Although we have not included an object in Fig. 1.73, the system's chief ray can be traced from the pupils and the FS, as in Section 1.7.4. This determines the field of view cone and the object field cone. The angle α' determines the maximum height of the image that can be observed by the eye.

An object placed within the object field will have its conjugate within the field of view. With this in mind, consider four positions for an object of height h as shown in Fig. 1.74. For position a, the image of the object will be completely contained in the field of view cone. For position b, only part of the image will be visible; the top part will be cut off. For position c, the object is completely outside the object field cone, so there is no image anywhere in the field of view cone. For position d, again, the full image of the object can be seen, but it will now be inverted compared to position a. This can be easily verified with a simple lens. By setting distances between the lens and the eye greater than the focal

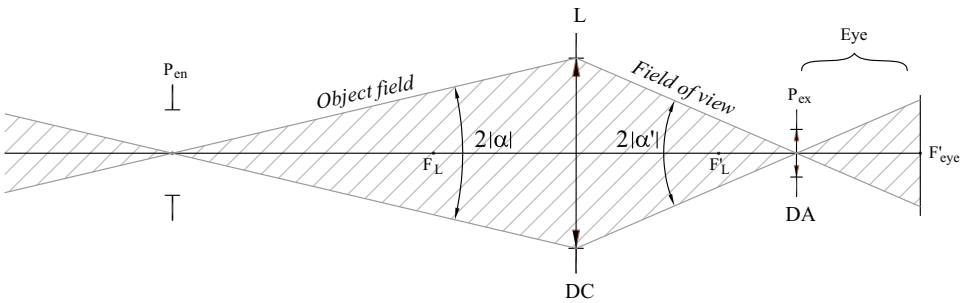


Figure 1.73 Fields of view and object field of a positive lens coupled with an eye.

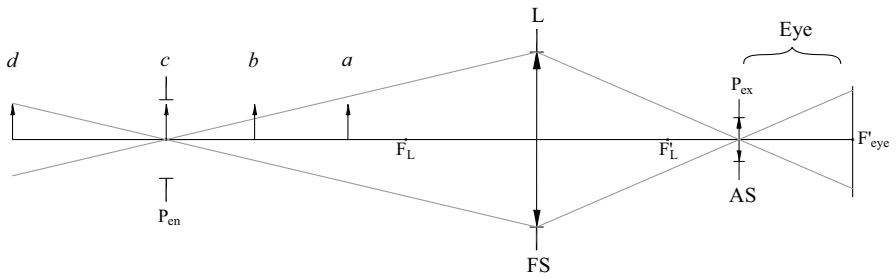


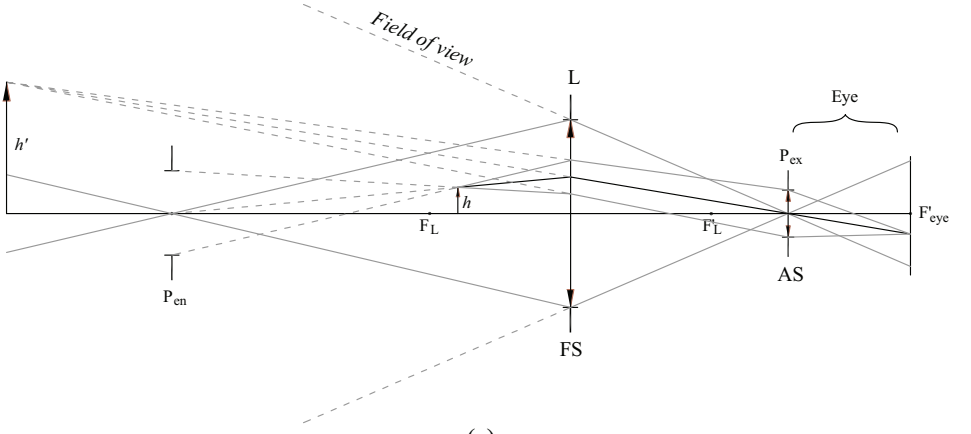
Figure 1.74 The extent of an object that can be seen in the image plane of a positive lens depends on the location of the object within the object field.

length of the lens, we will be zooming in and out of the object. We can see that in one position the object's image disappears, and as we move forward the orientation of that image is reversed.

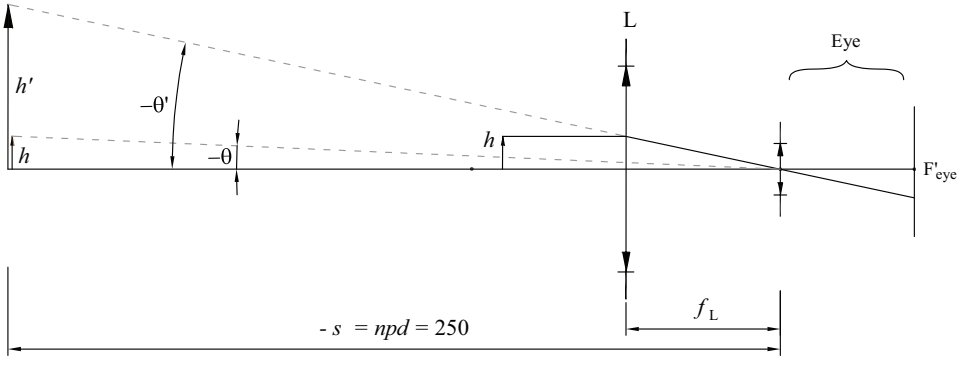
When the object is located between the primary focal point and the lens, a larger image is seen in the eye with the same orientation as the object (image processing in the brain involves reversing the orientation of the physical image on the retina). In this configuration, the lens works like a magnifying glass. In Fig. 1.75(a), the image is located using the chief ray and two marginal rays emerging from the object tip. Note that the image is still within the field of view cone, which is projected into the object's space. The image will be focused on the retina if the distance from the virtual image to the eye is greater than or equal to the *npd* of the eye.

To calculate the nominal magnification of the magnifying glass, we should take $npd = 250$ mm, according to the standard. Thus, the shortest distance that we can bring an object closer to be able to see it with an eye (without the help of a magnifying glass) using the largest possible angular size is the *npd*. With the magnifying glass, we can bring the object even closer; now the image we would see is the virtual image generated by the magnifying glass, and again we could see it in the largest possible angular size if this image were located in the *npd*. The ratio between these two angular sizes, i.e., from the angles with and without the magnifying glass, is the nominal magnification of the magnifying glass.

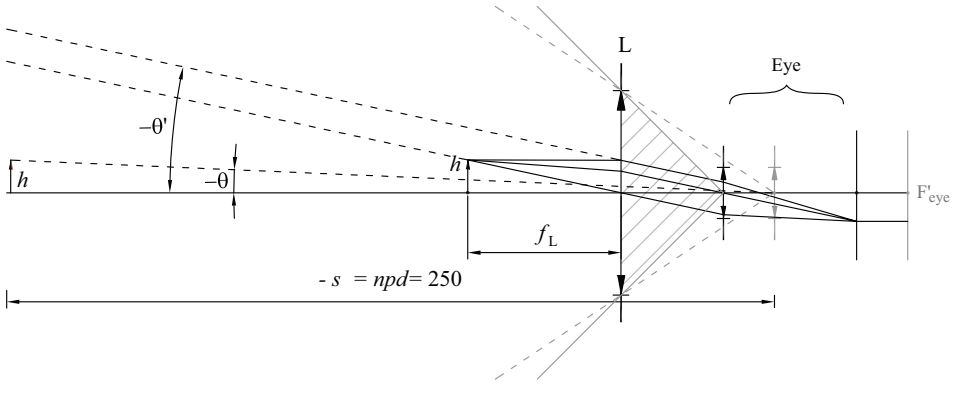
One way to determine the angular magnification of the magnifying glass is illustrated in Fig. 1.75(b). Let us assume that we are in the paraxial range. Without the magnifying glass, when we locate the object of height h at the *npd* of the eye, the object subtends the angle $\theta = -\arctan(h/npd)$. With the magnifying glass the object is located in such a way that the virtual image of height h' is at the *npd* of the eye. The angle that the virtual image subtends is $\theta' = -\arctan(h'/npd)$. If the eye's crystalline lens is located at the secondary focal point of the magnifying glass, the chief ray is parallel to the optical axis. With this in mind, $h'/npd = h/f_L$. Therefore, the nominal angular magnification $m_\theta = \tan\theta'/\tan\theta$ will be given by



(a)



(b)



(c)

Figure 1.75 The magnifying glass. (a) Trace of the chief ray and two marginal rays emerging from the object tip and forming a virtual image larger than the object. (b) The magnification of the magnifying glass is the ratio between the angle that subtends the virtual image and the angle that subtends the object, with respect to the center of the iris, as if both the object (seen without the magnifying glass) and the virtual image seen with the magnifying glass were at a plane located at the near point. (c) Bringing the eye closer to the magnifying glass increases the field of view of the system without changing the size of the image on the retina.

$$m_\theta = \frac{npd}{f_L} = \frac{250\,\text{mm}}{f_L}, \tag{1.68}$$

with f_L in millimeters. Note that by placing the eye's crystalline lens at the secondary focal point of the magnifying glass, the object can be moved between the magnifying glass primary focal point and the magnifying glass location without changing the angular size of the object. Since the eye is physiologically relaxed when observing an object at infinity, it is convenient to place the object at the primary focal point of the magnifying glass to observe its virtual image at infinity without changing the angular magnification.

What happens if we bring our eye closer to the magnifying glass while keeping the object at the primary focal point of the magnifying glass? This situation is illustrated in Fig. 1.75(c). Because the rays emerging from any point of the object, particularly from the object tip, are refracted as oblique bundles of parallel rays with angular inclination θ', the ray that crosses the center of the eye's crystalline lens will reach the retina at the same point as shown in Fig. 1.75(b). Therefore, the size of the image on the retina does not change. However, the field of view increases, since the eye's iris is the exit pupil; thus, when the eye is closer to the magnifying glass, the angle $2|\alpha'|$ that the magnifying glass subtends is greater. So if we put an eye very close to the magnifying glass (as watchmakers do), we will have a magnified image within the widest field of view that can be achieved.

The nominal magnification of a magnifying glass (also referred to as a "magnifier") is indicated by m×, e.g., a 10× magnifier is a 25 mm focal length magnifying glass [Eq. (1.68)]. Commercial magnifiers vary from single-digit magnifications to about 30×. Two magnifiers are shown in Fig. 1.76. In

(a) (b)

Figure 1.76 Optical magnifiers. (a) A Carson LinenTest™ 5× loupe magnifier with a base on the focal plane of the lens. Image reprinted from https://carson.com/ with permission from Carson. (b) An Iwamoto wide-field 20× magnifier with three lenses in place to correct chromatic aberration. Image reprinted from https://www.ascscientific.com/ with permission from ASC Scientific.

(a), a 5× magnifier is made up of a single lens and a base at the focal length of the lens, ensuring that the object is always at the proper distance. In (b), a 20× magnifier is shown. The higher the magnification, the more complex the magnifier will be, as aberrations that are most evident in high-magnification magnifiers must be corrected to ensure a good-quality image. In (b), the magnifier is made up of three lenses with a diameter of 15 mm. All three lenses are linked. This magnifier is labeled Achromatic 20×. The word "Achromatic" indicates that the system is designed to correct for chromatic aberration, and 20× indicates that the focal length of the system is 12.5 mm [Eq. (1.68)].

1.8.3 The telescope

Keplerian refracting telescope

A positive lens forms the image of an object that is at infinity in its secondary focal plane (Section 1.5.3). The image is real, inverted, and small. If this image is placed in the primary focal plane of a magnifier, we will have a magnified image of an object that is at infinity. This combination of a positive imaging lens (L_{ob}) and a magnifier ocular lens (L_{oc}) defines the basic structure of a Keplerian refracting telescope, as shown in Fig. 1.77.

The L_{ob} lens is called *objective* because it is on the side of the object, and the L_{oc} lens is called *ocular* or eyepiece because it is on the side of the eye. The secondary focal point of the objective lens and primary focal point of the eyepiece coincide. The marginal ray passes through the edge of the objective lens, which would be the AS. Since the edge of the ocular lens will be the FS, the chief ray must pass through this edge and through the center of the entrance and exit pupils, defining the angle $-\alpha$ of the object field ($2|\alpha|$) and the angle α' of the field of view ($2|\alpha'|$) of the telescope, as illustrated in Fig. 1.77. The entrance pupil is the edge of the objective lens, and the exit pupil is a real image that the eyepiece forms for the edge of the objective.

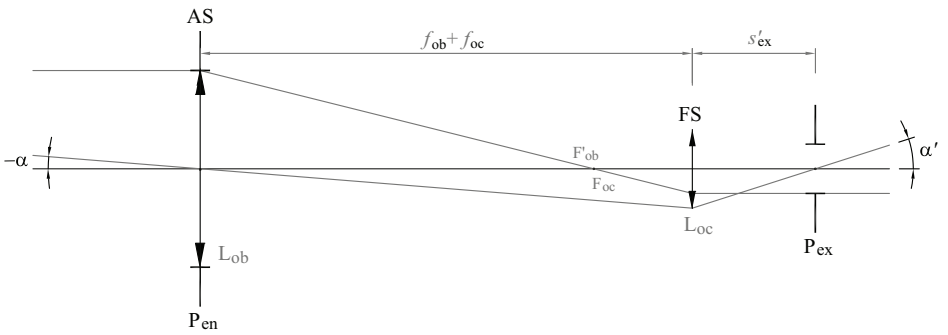


Figure 1.77 A schematic setup of the Keplerian refracting telescope. The secondary focal point of the objective lens coincides with the primary focal point of the ocular lens or eyepiece.

From the geometry of Fig. 1.77 and Eq. (1.42), the exit pupil is located at the distance

$$s'_{ex} = f_{oc}\left(\frac{f_{ob} + f_{oc}}{f_{ob}}\right) \tag{1.69}$$

from the ocular lens.

A telescope as an optical system is afocal; i.e., it does not have focal points. In the Keplerian telescope, this becomes clear by examining the marginal ray shown in Fig. 1.77. The marginal ray strikes parallel to the optical axis and is refracted parallel to the optical axis. This also implies that the object is at $-\infty$ and the image is at $+\infty$. Therefore, in a telescope, the magnification should also be defined in angular form, i.e., $m_\alpha = \tan\alpha'/\tan\alpha$. From the geometry of Fig. 1.77, the object field and the field of view angles are given by $\tan\alpha = -r_{oc}/(f_{ob} + f_{oc})$ and $\tan\alpha' = r_{oc}/s'_{ex}$, respectively, where r_{oc} is the radius of the ocular lens. Therefore, the telescope magnification will be

$$m_\alpha = -\frac{f_{ob}}{f_{oc}}. \tag{1.70}$$

As a last step to observe an image with the telescope, we place an eye behind the eyepiece. Let us assume that the eye's pupil diameter is greater than the telescope exit pupil diameter. Thus, the image formation process is as follows. The telescope objective forms an intermediate image of the object at infinity in the secondary focal plane (F'_{ob}). The maximum size of this image will be $h' = f_{ob}\tan\alpha$. Then, the ocular lens, which works like a magnifying glass, sends bundles of parallel rays to the eye, and the crystalline lens of the eye focuses them on the retina, forming a final image of size $h'' = l_{eye}\tan\alpha'$. As we see in Fig. 1.75(c), the eye can move closer to or farther from the eyepiece magnifying glass without changing the magnification of the image. However, there is an optimal position for the eye. Because the telescope exit pupil is fixed, the field of view will also be fixed. Therefore, if the eye's pupil is to the left or to the right of the telescope exit pupil, the eye's pupil may decrease the field of view for the eye. If the eye's pupil is placed in the same position as the telescope exit pupil, the field of view will not be altered. This latter configuration of the telescope-eye system is ideal, and the two pupils are said to be coupled. In other words, the eye should be positioned at the distance s'_{ex} from the ocular lens. The imaging formation on the retina and the pupils coupling are illustrated in Fig. 1.78. The energy cone (delimited by the marginal ray) and field of view cone (delimited by the chief ray) are also shown.

In Fig. 1.78, an aperture (of circular shape) has been inserted into the secondary focal plane of the objective lens to be the FS. If this aperture is of variable diameter, then the extent of the image to be observed can be adjusted without affecting the energy flow. More important is the fact that in this

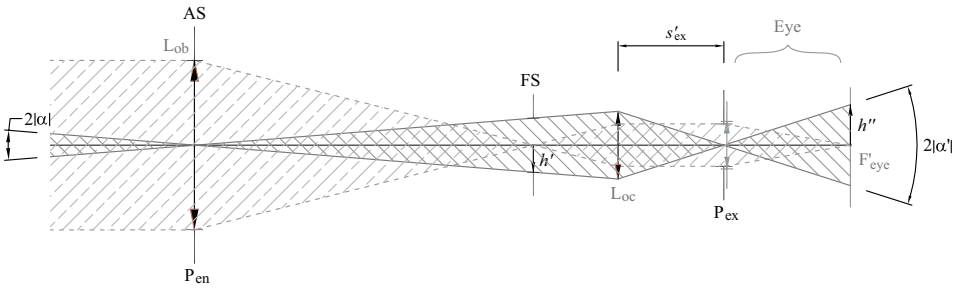


Figure 1.78 Object field cone and field of view cone, which are defined through the tracing of the chief ray, in a Keplerian refracting telescope. The eye's pupil is coupled to the telescope exit pupil. The optimal position of the FS is in the secondary focal plane of the objective lens. The energy cone is determined by the marginal ray.

plane, we can introduce a transparency film with additional information that we would like to show in the image, e.g., a metric scale. Thus, as a general rule of thumb, we can say that the proper place to place an FS in an optical system should be where the marginal ray intersects the optical axis. In practice, eyepieces are made up of several lenses and an FS separated from the lenses. For example, the following types of eyepieces can be consulted: Ramsden, Kellner, orthoscopic, symmetrical, and Erfle [3]. The optical elements that constitute a Kellner-type eyepiece are shown in Fig. 1.79.

Example: a 36× telescope

Consider a commercial refracting telescope that has an objective lens of focal length $f_{ob} = 900$ mm and diameter $2r_{ob} = 80$ mm, and a 10× eyepiece with diameter $2r_{oc} = 10$ mm.

From these parameters, the ocular lens system has a focal length $f_{oc} = 25$ mm [Eq. (1.68)]; thus, the telescope magnification is $m_\alpha = -36$. The object field turns out to be $2|\alpha| = 10/925\,\text{rad} = 0.62°$ and the field of view

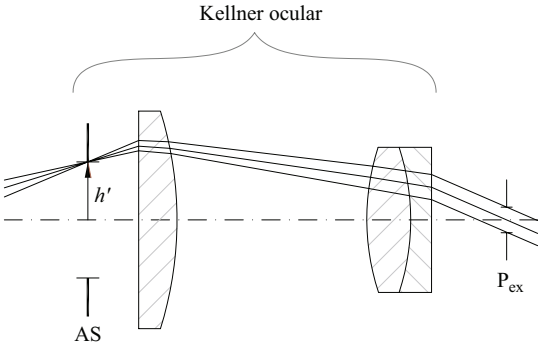


Figure 1.79 Diagram of the optical elements that form a Kellner-type eyepiece.

would be $2|\alpha'| = 10/s'_{ex} = 0.3892\,\text{rad} = 22.30°$ (the correct value is 22.02°). In this example, we can perceive how small the object field can be (0.62°). For this reason, some astronomical telescopes usually have a small telescope attached (called a finder scope) with a lower magnification and a larger object field that allows us to quickly locate what we want to observe (an object in the sky). The finder scope must be aligned with the telescope. Using the prameters provided above, the diameter of the telescope exit pupil should be $D_{ex} = 2r_{ob}(f_{oc}/f_{ob}) = 2.5$ mm and the f-number would be $f/\# = 10$.

A commercial telescope with these parameters is shown in Fig. 1.80. Note that the eyepiece is not located along the optical axis that passes through the objective, but rather at 90°, this is to make the observation more comfortable. A right-angle prism is used to change the direction of the eyepiece.

Galilean refracting telescope

Two of the telescopes (and an objective lens) built around 1610 by Galileo are exhibited in the Museo Galileo, the former Istituto e Museo di Storia della Scienza (Institute and Museum of the History of Science), located in Florence, Italy. Unlike the Keplerian refracting telescope described above, which uses a positive lens as an ocular lens, Galilean telescopes use a negative lens, although they still maintain an afocal configuration; i.e., the secondary focal point of the objective lens coincides with the primary focal point of the ocular lens. At the primary focal plane of the ocular lens there will be a virtual object (corresponding to the smallest inverted real image that the objective lens forms in the absence of the ocular lens), and its image will be infinite. An analysis of the geometry of this telescope returns us to Eqs. (1.69) and (1.70), and now $f_{oc} < 0$. According to Fig. 1.81(a), $\tan\alpha = r_{oc}/(f_{ob} - |f_{oc}|)$ and $\tan\alpha' = r_{oc}/s'_{ex}$. The exit pupil diameter would be $D_{ex} = 2r_{ob}(|f_{oc}|/f_{ob})$.

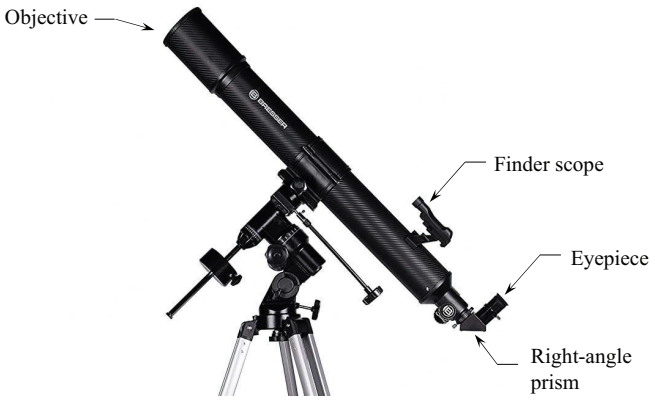


Figure 1.80 A Bresser Quasar 80/900 EQ refracting telescope. It includes an optical finder scope (small telescope with a lower magnification and greater object field) to easily locate objects. Image reprinted from https://www.bresser.de/ with permission from Bresser GmbH.

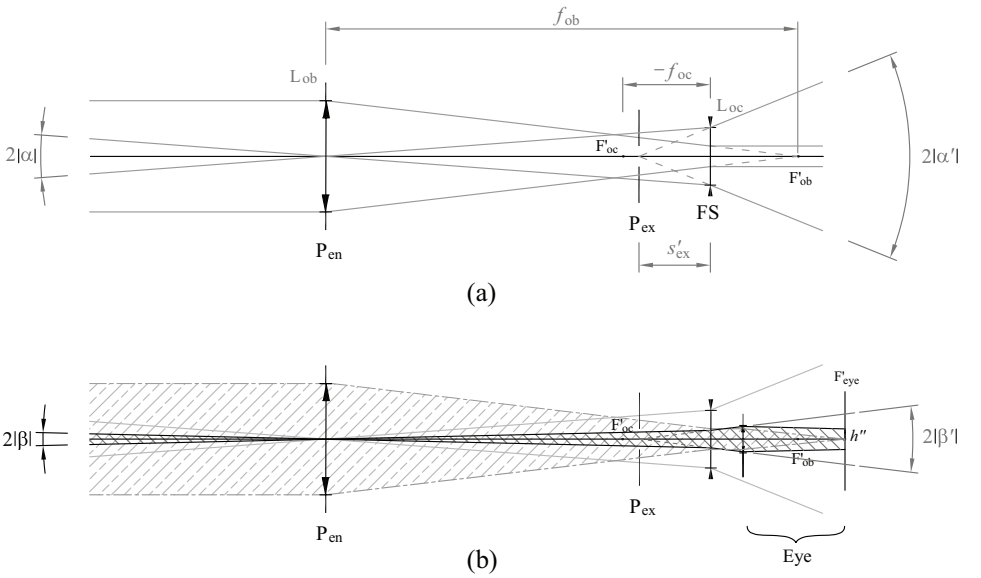


Figure 1.81 Galilean telescope. The telescope exit pupil cannot be coupled with a human eye's pupil. The effective field of view is determined by the eye's pupil.

When using a negative lens as the eyepiece or ocular lens, the exit pupil, which is the image of the AS, is located between the objective lens and the eyepiece as a virtual image, upright and smaller than the AS. Consequently, it is not possible to couple the eye's pupil with the telescope exit pupil. So where should the eye be positioned for optimal vision? In Fig. 1.81(b), the eye is shown close to the eyepiece. Assuming that the eye's pupil diameter is greater than the telescope exit pupil diameter, the eye's pupil cuts the field of view to an angle of $2|\beta'| < 2|\alpha'|$. The farther the eye is, the smaller the telescope's effective field of view (and the object field) will be. Because the chief ray no longer passes through the center of the eye's crystalline lens, the maximum size that the image can be on the retina depends on the position of the eye. In other words, in the Galileo eye-telescope combination, the FS will be the eye's pupil. Therefore, the best way to use a Galilean telescope is by placing the eye as close to the eyepiece as possible.

One of Galileo's telescopes on display in the Museo Galileo consists of a wooden tube covered in leather. The objective and ocular lenses are housed at the two ends of the tube. The objective lens has a focal length $f_{ob} = 980$ mm and diameter $2r_{ob} = 37$ mm, and the ocular lens has a focal length $f_{oc} = -47.5$ mm and diameter $2r_{oc} = 22$ mm [11]. The magnification of the telescope would be $m_\alpha = 20.63$, and the object field would be $2|\alpha| = 0.0236$ rad $= 1.35°$. However, the effective object field will be less when positioning the eye behind the telescope's eyepiece. Assuming that the eye is positioned very close to the eyepiece and that the eye's pupil diameter is

4 mm (mean aperture diameter of a human eye's iris), the field object would be $2|\beta| = 4/(980 - 47.5)\,\text{rad} = 0.0043\,\text{rad} = 0.25°$ (15 arc min).

Newtonian reflecting telescope

In refracting telescopes, the objective lens is positive and forms the image of an object (which is at infinity) in the secondary focal plane; the magnified image is then observed with the eyepiece (magnifying glass). The image of distant objects can also be brought to the primary focal plane of the eyepiece by using a concave mirror instead of a positive lens. In Section 1.4.2, we see that a concave mirror, whose reflecting surface is a paraboloid of revolution, focuses rays parallel to the optical axis at the focal point. This point is located $R_v/2$ from the mirror vertex, where R_v is the radius of curvature of the paraboloid at the vertex (Fig. 1.37).

In the refracting telescope, light always travels in the same direction until it reaches the eye. But by changing the main light-gathering element (the objective) from a positive lens to a concave mirror, the light changes direction as it reflects off the mirror. So another optical element is required to change the direction of light propagation again. In 1668, Newton proposed a configuration that solves this problem. The solution consisted of placing a flat mirror (secondary mirror) near the focal point of the concave mirror (primary mirror) with respect to the optical axis. The secondary mirror is several times smaller than the primary mirror and must receive all the rays reflected by the primary mirror. The basic configuration of the Newtonian telescope is illustrated in Fig. 1.82. The nominal magnification of the telescope is given by Eq. (1.70), where $f_{ob} = R_v/2$.

There are other reflecting telescope configurations that, instead of using a flat secondary mirror, use a hyperbolic mirror (Cassegrain) or an elliptical mirror (Gregorian), keeping a parabolic mirror as the primary mirror.*

Unlike refracting telescopes that depend on the refractive property of lenses and the wavelength of light, reflecting telescopes prevent chromatic aberration of the telescope objective. Although achromatic refracting systems, which correct for chromatic aberration, can be built by combining lenses, there is a practical limitation when we want to build large diameter (few meters) objectives. In fact, the largest refracting telescope ever built is at the Yerkes Observatory, an astronomical observatory in Williams Bay, Wisconsin, US. The telescope objective consists of two 1016 m (40 in) diameter lenses 218 mm apart; the focal length turns out to be 18897 m (744 in).** From these

*For the mathematical and optical details of the most common types of reflecting telescopes and their variants, see "Chapter 6: Calculation of primary aberrations: reflecting and catadioptric systems" in *Optical Imaging and Aberrations: Part I. Ray Geometrical Optics* by Virendra N. Mahajan (SPIE Press, 1998).

**The telescope data was obtained from the Yerkes Observatory page (https://www.yerkesobservatory.org/).

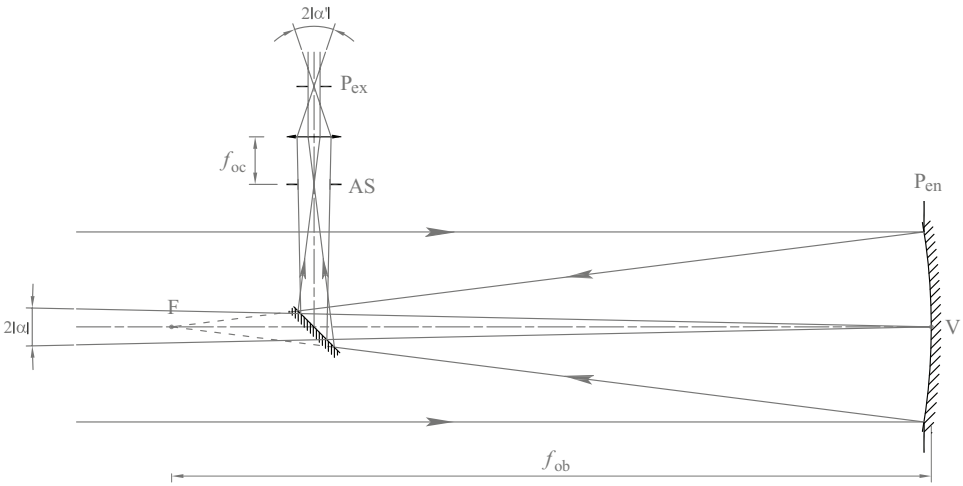


Figure 1.82 The Newtonian telescope uses a concave mirror in the shape of a paraboloid of revolution (primary mirror) and a diagonal plane mirror (secondary mirror) to focus light from infinity to the primary plane of the ocular lens.

data, the objective telescope has an $fl\# = 18.6$. The challenge of building large refracting telescopes is to obtain glass blocks without imperfections and with a homogeneous refractive index, and to achieve the desired curvature on both faces of the lens. This is not a limitation for the primary mirrors of reflecting telescopes, where care must be taken instead with the quality of the reflecting surface. Thus, several astronomical observatories have large reflecting telescopes with primary mirrors that measure up to 10 m in diameter.

The size of the primary mirror plays a fundamental role in the *resolution of the telescope*. This can be given as the minimum angular distance between two point objects with which the two points of the image can be distinguished (limit of angular resolution), or as the minimum transverse distance between the images of the two points that allows the points to be identified (resolution limit). According to the Airy criterion [3], the resolution limit is given by*

$$(\Delta l)_{\min} = 1.22\lambda(f/\#). \tag{1.71}$$

1.8.4 The microscope

At present, there are many types of microscopes that can be consulted in the specialized literature. In this book, we will deal only with the traditional light microscope setup that focuses the image on the retina of the eye. Basically, a light microscope adds a positive lens (objective, L_{ob}) to a magnifying glass (eyepiece or ocular lens, L_{oc}) to achieve magnified images of close and very small objects, as shown in Fig. 1.83. The distance between the secondary focal

*This criterion will be discussed in Chapter 4.

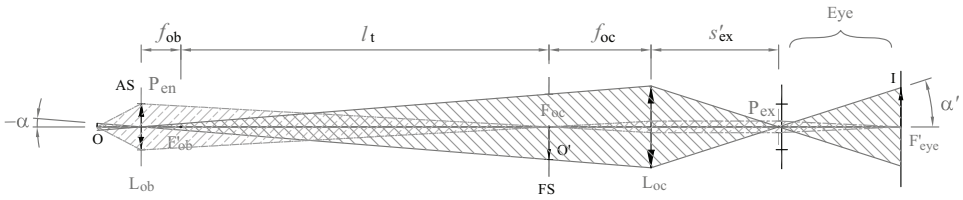


Figure 1.83 The microscope. The length of the tube l_t is set at 160 mm. With this, the distance between the objective lens and the ocular lens is kept low, and as the optical power of the objective lens changes, the object or specimen should be moved in or out until it is in focus.

point of the objective and the primary focal point of the eyepiece, denoted by l_t, is called the length of the tube. This length is set by microscope manufacturers so that when the refractive power of the objective is changed, the distances between the lenses do not change; thus, it is the object that must move until it is in focus. Common distances for tube length are 160 mm (DIN standard) or 170 mm (JIS).*

Microscope magnification is the product of the magnification produced by the objective lens and the magnification produced by the ocular lens. The objective lens forms the image of the object O at O′ at the distance l_t from the secondary focal point of the objective lens; thus, the magnification of this image will be $m_{ob} = -l_t/f_{ob}$. On the other hand, the magnification of the ocular lens is $m_{oc} = 250/f_{oc}$. Therefore, the magnification of the microscope is

$$m_t = -\frac{l_t}{f_{ob}}\frac{250}{f_{oc}}, \tag{1.72}$$

with the focal lengths given in millimeters. Microscope objectives when used in air range from about 4× to about 40×. Higher magnification objectives are also found, but they are used by immersing the object in oil for light microscopy. These objectives are called immersion objectives.

The AS is defined by the edge of the objective lens, so the entrance pupil will be there. The exit pupil will be to the right of the ocular lens and is a real image and smaller than the AS. The exit pupil is usually smaller in diameter than the eye's pupil. As in the refracting telescope, pupillary coupling is achieved by placing the eye's pupil at the microscope exit pupil.

In telescope objectives, the *f*/# [Eq. (1.62)] is the parameter used to establish the amount of energy collected by the objective and measure the resolution limit. In microscopes, a slightly different quantity is used called the *numerical aperture NA*, defined as

*The Deutsches Institut für Normung (DIN) standard is established in Germany. The Japanese Industrial Standard (JIS) is established in Japan.

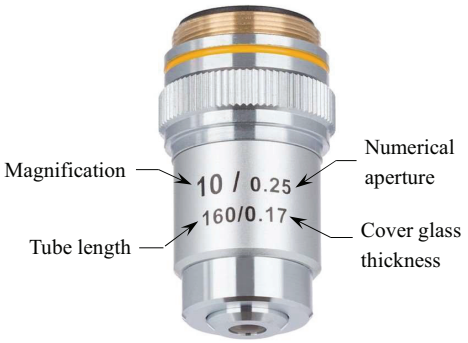


Figure 1.84 An AmScope 10× achromatic objective designed to work with DIN-standard microscopes using a 160 mm tube length and a 45 mm distance between the objective lens mounting plane and the object or specimen. Image reprinted from https://amscope.com/ with permission from AmScope.

$$NA = n \sin \varphi, \tag{1.73}$$

with n the refractive index of the medium surrounding the object and $2|\varphi|$ the angle that the entrance pupil subtends (of diameter $2r_{Pen}$) with respect to the axial point of the object; i.e., $\tan \varphi = r_{Pen}/l_{en}$, where l_{en} is the distance from the object to the entrance pupil. In Fig. 1.84, a commercial microscope objective is shown, identified with the magnification, the length of the tube, the numerical aperture (in air), and the thickness of the coverslip to be used. In this example, the focal length of the objective lens is $f_{ob} = (160/10)$ mm = 16 mm.

The resolution of the microscope is given by an amount analogous to that of the telescope. This quantity is called the resolving power (resolution limit) and is given by

$$(\Delta l)_{\min} = 1.22\lambda\left(\frac{1}{2NA}\right). \tag{1.74}$$

The object field in a microscope is greater than that in a telescope. The angle $-\alpha$ in Fig. 1.83 can be calculated as $-\alpha = r_{DC}/(l_t + f_{ob})$, where r_{DC} is the radius of the FS. Thus, $2|\alpha|$ is the estimated region of the object that can be seen under the microscope.

The basic parts of a microscope are shown in Fig. 1.85. This microscope illuminates the sample from below, so that what is observed is the light transmitted by the object (transmission microscope). The object, a sample or specimen, is positioned on the stage, which can be moved vertically using the coarse (large shift) and fine (small shift) drive screws to achieve focus on the sample. The revolver accommodates multiple targets of different magnifications, and turning it selects the desired target. At the end of the tube(s), we can

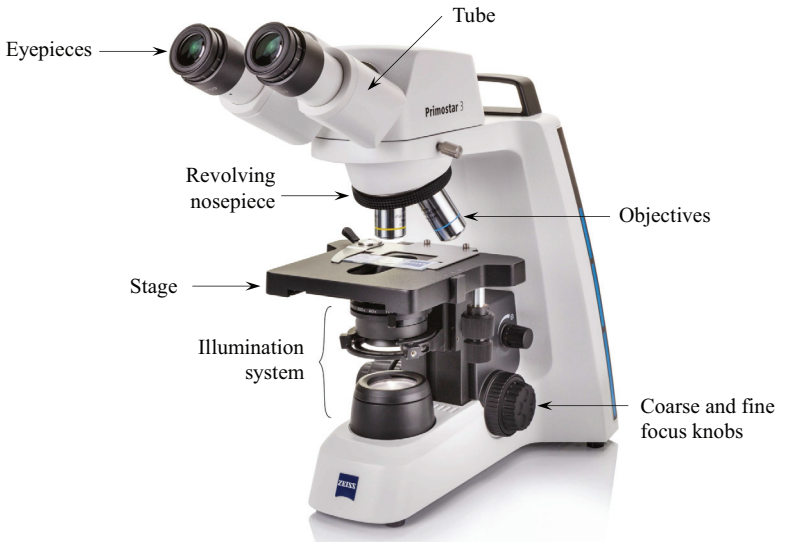


Figure 1.85 Components of a ZEISS Primostar transmission microscope. Image reprinted with permission from ©ZEISS Microscopy (https://www.zeiss.de/).

find the eyepiece(s), in which the eye(s) can be placed to observe the final image. The object can also be illuminated from above (for example, with a concave mirror), in which case what will be seen is the light reflected by the object (reflection microscope).

1.9 Monochromatic Optical Aberrations

In Section 1.3, the Cartesian oval is defined as the surface of revolution that forms a point image of a point object, both on the optical axis. The shape of this surface is derived from Fermat's principle and depends on the object distance and image distance, with respect to the surface vertex, and on the relation between the media (surface and surroundings) refractive indices. Therefore, for different object distances, we will have different Cartesian ovals. This imposes a practical limitation when designing optical imaging systems; the image would be well in focus for only one object position. A practical solution is to use reflecting and/or refracting spherical surfaces and limit the optical region through which the energy flow will pass to a very small region around the optical axis, which can be done with a diaphragm (AS) of small diameter. This condition is known as the paraxial approximation and gives rise to the Gaussian equation [Eq. (1.21)], which implies that the image of a point object will be a point image, so the image of an extended object will be a copy of the object except for a scale and orientation factor.

When the optical system is not limited to the paraxial region, the image of an extended object will no longer be a faithful replica of the object. The

differences between the real image and the paraxial (ideal) image are called optical aberrations. A detailed analysis of optical aberrations is beyond the scope of this book. The reader interested in expanding on this topic can consult Malacara and Malacara's book [12], among others. In this section we will make a very general qualitative description that allows us to conceptually understand monochromatic primary optical aberrations, which do not depend on wavelength. The treatment presented in this section is made for a refracting spherical surface, but the results can be extended to lens systems. The main aberrations due to the color of light (chromatic aberrations) are presented in Appendix D.

1.9.1 Field curvature

In the paraxial approximation, the rays leaving a point object must be almost parallel to the optical axis of an optical system and travel very close to that axis. This means that the extent of the object (and its image through the optical system) must also be small compared with the object distance (and its corresponding image distance). What if the extension of the object is comparable to the object distance even though the AS size is kept small? Figure 1.86 illustrates an optical system composed of a refracting spherical surface and an AS located at the center of curvature of the surface. An object of height $h = -\overline{OP}$, according to paraxial optics, will have its image in a plane (Gaussian image plane) located at O′ and would have a height $h' = \overline{O'P''}$. However, due to the extension of the object, the image of P is not in P″ but rather a bit in front of P″. To see this, let us consider two bundles of rays: one coming out of O and the other coming out of P. The chief ray coming out of O will be the optical axis OVO′, and the chief ray coming out of P will be the auxiliary axis PV′P″. Both the rays leaving O and those leaving P remain almost parallel to the corresponding chief ray, but the chief ray coming from point P is no longer paraxial. The object distance $\overline{PV'}$ is greater than the object distance $\overline{OV}$ and, from Eq. (1.21), it turns out that the image distance $\overline{V'P'}$ is

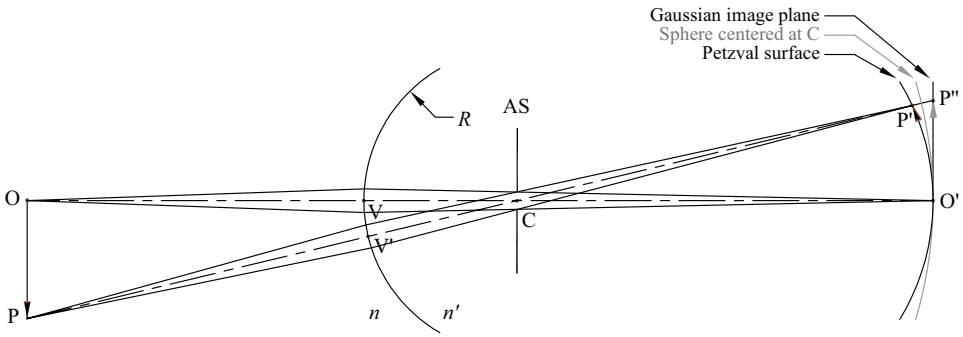


Figure 1.86 Field curvature. The image of a plane is the so-called Petzval surface.

less than the image distance $\overline{VO'}$. The image P′ will be on the auxiliary axis, but to the left of the sphere of radius $\overline{CO'}$ centered at C. In conclusion, the image of a plane located at O is a surface of revolution with vertex at O′. This surface is called the *Petzval surface*. As a first approximation, the Petzval surface (*Ptz*) is a paraboloid of revolution with vertex at O′; i.e.,

$$Ptz = \frac{h'^2}{2r_{Ptz}}, \tag{1.75}$$

where r_{Ptz} is the radius of curvature at the paraboloid vertex. In an optical system with several refracting surfaces, the inverse of the radius of curvature depends on the sum of the refractive power of the surfaces. In the thin lens approximation, the Petzval radius for N lenses is given by

$$\frac{1}{r_{Ptz}} = -\sum_{k=1}^{N} \frac{\tilde{P}_k}{n_k}, \tag{1.76}$$

with $\tilde{P}_k$ given by Eq. (1.54) and n_k the refractive index of lens k. Thus, by combining positive and negative lenses, it is possible to increase the radius of curvature, i.e., to flatten the image surface.

An image is usually recorded on a flat sensor (photographic film, CCD, or CMOS camera). Let us assume that the sensor is located in the Gaussian image plane. Given that the image of a point is also a point on the Petzval surface, the image of an extended object in the Gaussian image plane is focused on the optical axis; and as we move away from the axis, it will become more and more blurred (out-of-focus), as shown in Fig. 1.87. This defect is known as *field curvature aberration*. Reducing the diameter of the AS also

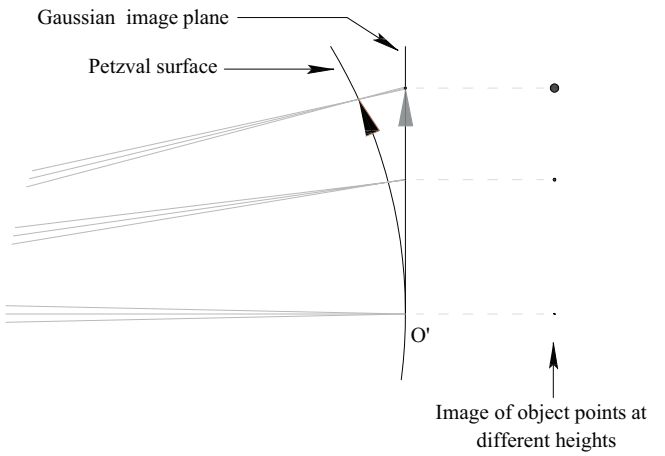


Figure 1.87 Image of object points at different heights affected by the field curvature aberration.

reduces this defect. At the limit, an AS with diameter tending to zero would only allow the chief ray to pass through, and the image of any point on the object would look like a point image. In other words, the effect of field curvature on the image can also be reduced by reducing the size of the AS (at the expense of letting less light energy pass through).

1.9.2 Spherical aberration

This aberration is mentioned in Section 1.3 for the refraction on a spherical surface and in Section 1.4 for the reflection on a spherical surface. The refraction and reflection, respectively, of ray bundles limited by a large effective AS diameter are shown in Figs. 1.24 and 1.35. The marginal rays when they are refracted, in the first case, and when they are reflected, in the second case, do not converge to the point where the corresponding paraxial rays would converge. This defect is called spherical aberration, and its effect can be diminished by reducing the size of the AS.

The refraction on a spherical surface of the marginal ray and of a paraxial ray is shown in Fig. 1.88. The AS has been positioned to the left of the surface vertex. Quantitatively, spherical aberration can be measured as

$$SphL = S' - s', \tag{1.77}$$

i.e., the longitudinal difference between the point of intersection of the marginal ray and the point of intersection of the paraxial ray on the optical axis. In a first approximation, this aberration depends quadratically on the height of the marginal ray, so that

$$SphL = a_s y^2, \tag{1.78}$$

where the coefficient a_s depends on the parameters of the optical system (object distance, surfaces radii of curvature, and media refractive indices), but not on the height of the object (image). The spherical aberration is positive if $S' > s'$ ($a_s > 0$) and negative if $S' < s'$ ($a_s < 0$).

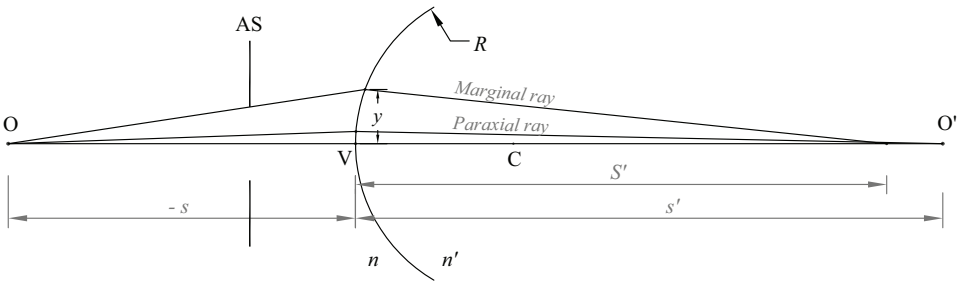


Figure 1.88 Spherical aberration. Nonparaxial marginal rays do not converge at the same point on the optical axis.

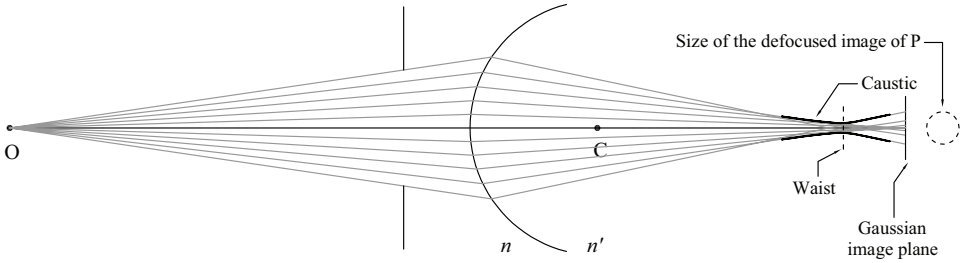


Figure 1.89 A caustic of rays affected by spherical aberration. The best image for P is achieved in the caustic waist.

Figure 1.89 shows the refraction of a ray bundle, diverging from the point O, that is affected by spherical aberration. In the Gaussian image plane, the intersection of the rays determines the size of the point image, as shown in the figure. The curve tangent to the intersection of the refracted rays is called a *caustic* and is a maximum energy curve. The caustic is seen to have a minimal cross section called the *waist*. In this section, the refracted rays have the lowest radial scatter. There you will have the best image of the point O, i.e., the least possible blur for the given AS size. Spherical aberration affects off-axis points of the object in the same way.

1.9.3 Distortion

Let us assume that the AS in Fig. 1.89 is closed enough so that the effect of spherical aberration can be ignored for rays exiting the axial point O; again, the image of O is a point in O′, as shown in Fig. 1.90. On the other hand, from the point P, which is off-axis, a ray bundle also comes out with the rays traveling very close to the chief ray and hitting point Q on the refracting surface. Therefore, for the rays leaving P, the spherical aberration effect with respect to the chief ray can also be ignored. However, the distance between the point Q and the auxiliary axis PV′P″ is greater than the size of the AS; thus,

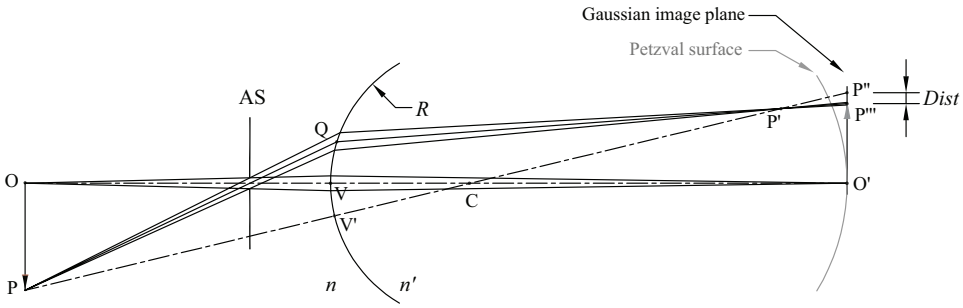


Figure 1.90 Distortion aberration warps the image but maintains the focus of each point. This aberration depends on the position of the AS.

the chief ray leaving P is affected by spherical aberration, and the image P′ of P will be at a distance $SphL \approx a_s \overline{QV}'^2$ from the Petzval surface.

Because the image is observed in the Gaussian image plane, what we will see is the projection of rays that diverge from P′. Since the AS is small and in practice $\overline{VO'} \gg SphL$, the intersection of these rays in the plane of the Gaussian image will look roughly like a point, denoted by P‴ in Fig. 1.90. Therefore, P‴ will be the image of P in the Gaussian image plane. The difference between the height of P‴, $\bar{h}'$, and the Gaussian image P″, h', is called distortion aberration or simply *distortion* (*Dist*).

In Fig. 1.86, the AS is located at the center of curvature of the surface. Note that the projection of the rays diverging from P′ reaches P″, so there is no distortion. The distortion depends on the position of the AS. In Fig. 1.90, the AS is to the left of C and the height of the image P‴ is less than the height of the Gaussian image P″; in this case, the image is said to have barrel distortion. If the AS is placed to the right of C, we will have the opposite: the height of the image P‴ will be greater than the height of the Gaussian image P″. In this case, the image is said to have a cushion deformation.

In a first approximation, the distortion also depends on the cube of h',

$$Dist = a_t h'^3 = \bar{h}' - h', \tag{1.79}$$

where a_t is the distortion coefficient that depends on the system parameters and the position of the AS. The possible configurations for an image affected by distortion are shown in Fig. 1.91 for an optical system of magnification $m_t = -1$ formed by a thin biconvex lens. When the AS is right next to the lens, the distortion is roughly zero. Distortion is strictly zero when the AS is at the center (inside) of the lens. In a multiple lens system, the AS will be located

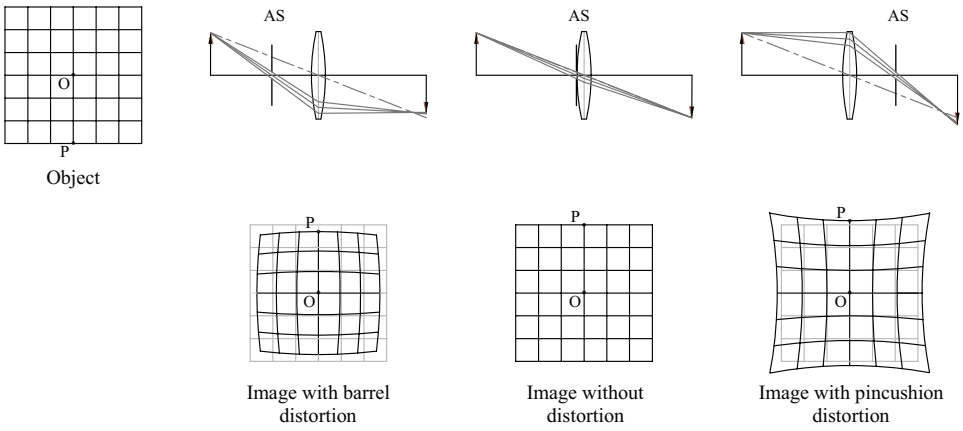


Figure 1.91 Examples of distortion aberration. If the AS is to the left of the lens, the image has barrel distortion. If the AS is right next to the lens, the image is (practically) without distortion. If the AS is to the right of the lens, the image has cushion distortion.

between the lenses seeking the greatest possible symmetry to have the least distortion.

1.9.4 Astigmatism and coma

To see these two aberrations, we will first define the tangential plane and the sagittal plane. In Section 1.3.5, the meridional plane is defined as the plane containing the optical axis. In that section, the extended objects are represented as arrows of a certain height contained in the meridional plane, and the rays emerging from any point on the object are also limited to the meridional plane. In an optical system, the plane that contains the optical axis and the chief ray is called the *tangential plane*, and the plane that contains the chief ray and is orthogonal to the tangential plane is called the *sagittal plane*. The tangential plane is a meridional plane, but the sagittal plane is not. The tangential plane maintains its spatial orientation throughout the optical system, whereas the sagittal plane changes its inclination in the same way as the chief ray. In Fig. 1.92, these planes are illustrated in a system composed of an AS and a convex refracting surface. The height of the object is $h = -\overline{OP}$, and the height of the image is $h' = \overline{O'P'}$. In Fig. 1.92, the rays that leave P and are in the tangential plane are called tangential rays (blue); the rays leaving P and contained in the sagittal plane are called sagittal rays (red). The projection of the AS on the refracting surface is shown in the figure and four points are marked: T_1 and T_2 correspond to the intersections of the tangential marginal rays on the refracting surface, and S_1 and S_2 correspond to the intersections of the marginal sagittal rays on the refracting surface.

The meridional plane containing the marginal tangential rays passing through T_1 and T_2 is shown in Fig. 1.93. Sagittal points S_1 and S_2 project into this meridional plane at a single point. Given the effect of spherical aberration

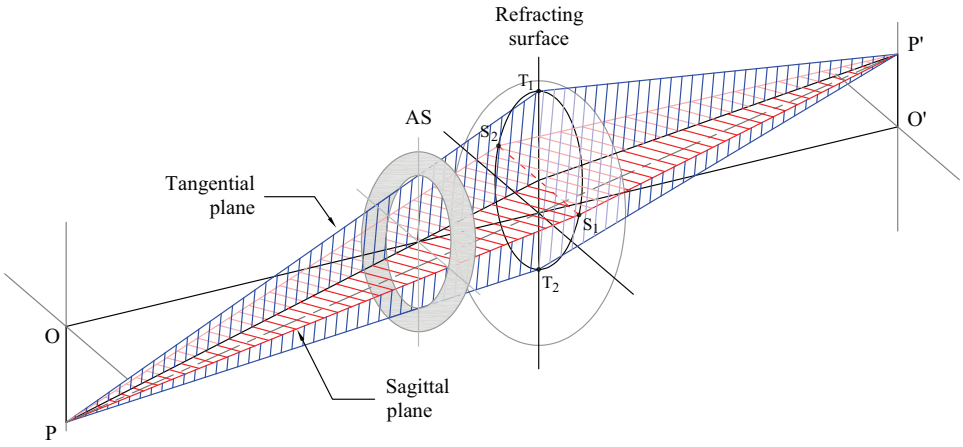


Figure 1.92 Defining the tangential (blue) and sagittal (red) planes.

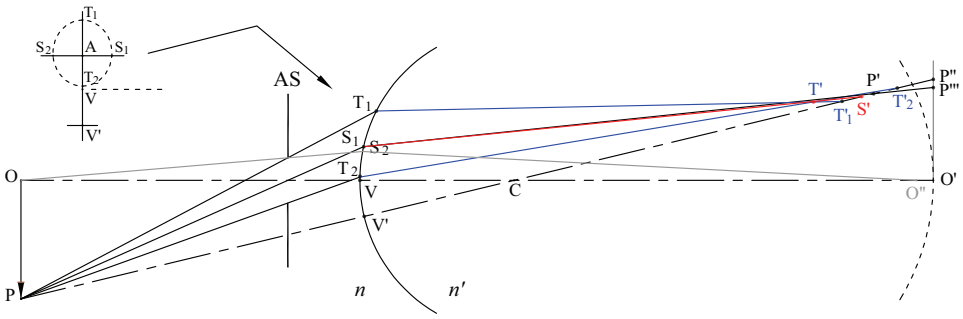


Figure 1.93 Defining astigmatism and coma aberrations for marginal tangential rays (blue) and sagittal rays (red). The distance $\overline{S'T'}$ along the chief ray is the (longitudinal) astigmatism. The distance between T′ and the chief ray is the tangential (transversal) coma, and the distance between S′ and the chief ray is the sagittal (transversal) coma. The chief ray cuts the surface at A and the auxiliary axis at P′.

on the marginal rays, it follows that, due to the symmetry of S_1 and S_2 with respect to vertex V′, the marginal sagittal rays (Fig. 1.93, rays in red) intersect in the auxiliary optical axis after refracting at points S_1 and S_2. On the other hand, because T_1 and T_2 are not at the same distance from the vertex V′ (top left part of Fig. 1.93) the marginal tangential rays (rays in blue) do not intersect on the auxiliary optical axis after refracting at points T_1 and T_2. This asymmetry gives rise to astigmatism and coma aberrations, which are defined from the following intersections: T'_1, intersection of the superior marginal tangential ray with the auxiliary axis; T'_2, intersection of the lower marginal tangential ray with the auxiliary axis; T′, intersection between the upper and lower marginal tangential rays; and S′, intersection between the marginal sagittal rays that is located on the auxiliary axis. The longitudinal distance between the transversal projections of S′ and T′ on the chief ray defines the *longitudinal astigmatism aberration* (*AstL*). The transversal distance between T′ and the chief ray defines the *tangential coma aberration* ($Coma_T$). The transversal distance between S′ and the chief ray defines the *sagittal coma aberration* ($Coma_S$).

Astigmatism

As the height of an object decreases, the spherical aberration experienced by the marginal rays (measured from the auxiliary axis) also decreases; thus, the intersections T′ and S′ approach the Petzval surface. On the optical axis, T′ and S′ coincide (astigmatism is equal to zero) at the point O″ that goes from O′ to the distance $SphL \approx a_s\overline{T_1V}^2 = a_s\overline{T_2V}^2$ (Fig. 1.93). In a first approximation, for a plane object, T′ and S′ describe two paraboloids of revolution with vertex at O″, called *astigmatic surfaces*. As the AS closes, O″ approaches O′ and the astigmatic surfaces also move toward O′. For a fixed height of the object, the *AstL* does not depend on the size of the AS.

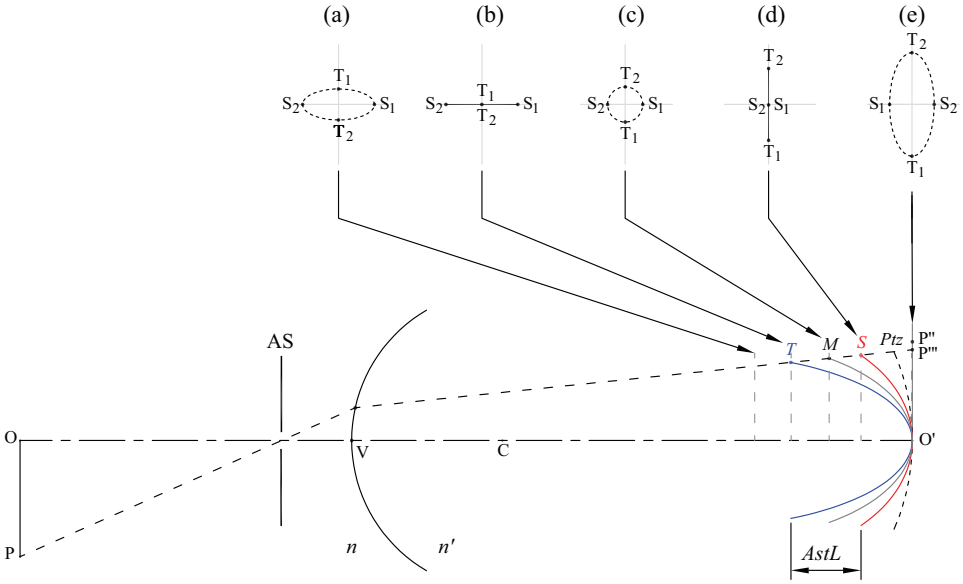


Figure 1.94 Astigmatic curves *T* and *S* and the intersection of the marginal rays in different planes in the absence of coma aberration. The projection of the marginal rays in the Gaussian image plane is an ellipse. In *M*, the circle of least confusion is obtained.

Figure 1.94 shows the sections of the astigmatic surfaces in the tangential plane, denoted by *T* and *S*, assuming that the diameter of AS tends to zero* (O″ coincides with O′). At *T*, the tangential marginal rays are in focus while the sagittal marginal rays are not yet in intersection. In a transversal plane passing through *T*, the marginal rays describe a (horizontal) line called the *tangential focal line*, shown in (b). At *S*, the sagittal marginal rays are in focus, while the tangential marginal rays diverge from T′. In a transversal plane passing through *S*, the marginal rays describe a (vertical) line called the *sagittal focal line*, shown in (d). The length of these focal lines is equal and depends on the size of the AS. In the middle of the astigmatic surfaces, there is a surface *M* where the marginal rays describe a circumference whose diameter is half the length of the focal lines, as shown in (c). In the vertical direction, we have the intersections of the marginal tangential rays T_1 and T_2, and in the horizontal direction, we have the intersections of the marginal sagittal rays S_1 and S_2. In particular, the marginal tangential rays have reversed their positions. The circle in *M* is called the circle of least confusion. For a plane located before *T*, the marginal rays describe an ellipse with the major axis defined by the sagittal rays, as shown in (a). Finally, in the Gaussian image plane, the marginal rays also describe an ellipse, but now the major axis is defined by the tangential rays. The Petzval

*This condition implies that the coma also tends to zero.

surface, in which the image would be formed in the absence of astigmatism, is also shown in Fig. 1.94.

According to the above, the image of an off-axis point object will change its geometry depending on the plane that is used to observe the image. The marginal rays describe the edge of the image corresponding to that off-axis point. The other rays, which go inside the AS, will fill the shape described by the marginal rays. Reducing the size of the AS decreases the length of the focal lines; i.e., in general, the ellipse size decreases, thus optimizing the image of the point. On surface M, we will have the best image for the point (a symmetrical point). However, astigmatic curves do not change shape with variation in the size of the AS.

The distances from T and S to Ptz satisfy the relationship $\overline{PtzT} = 3\overline{PtzS}$. Like the Petzval surface, astigmatic surfaces can be characterized by the radii of curvature at the vertex O′ that depend on the parameters of the optical system and the position of the AS. Therefore,

$$T = \frac{h'^2}{2r_T} \tag{1.80}$$

and

$$S = \frac{h'^2}{2r_S}. \tag{1.81}$$

Coma aberration

Now suppose we have coma aberration and there is no astigmatism. Thus, the points T′ and S′ will be in the same plane orthogonal to the optical axis, and we have the relation $Coma_T = 3Coma_S$. The $Coma_S$ depends linearly on the height h'; thus,

$$Coma_S = a_c h', \tag{1.82}$$

where the coma coefficient depends on the parameters of the optical system and the position and size of the AS. The shape described by the marginal rays is a circle (comatic circle) of radius equal to $Coma_S$ with its center at the distance $2Coma_S$ from the chief ray. Decreasing the AS radius decreases the $Coma_S$ value, thus decreasing the radius of the comatic circle and moving closer to the chief ray, as shown in Fig. 1.95. The shape we have looks like a comet, hence the name of this aberration. Going from T_1 to S_1 in Fig. 1.92, following the projection of the AS on the refracting surface, implies going from T′ to S′ in the corresponding circle of the the comatic shape; i.e., we would go around the corresponding comatic circle twice as we go around the edge of the AS once.

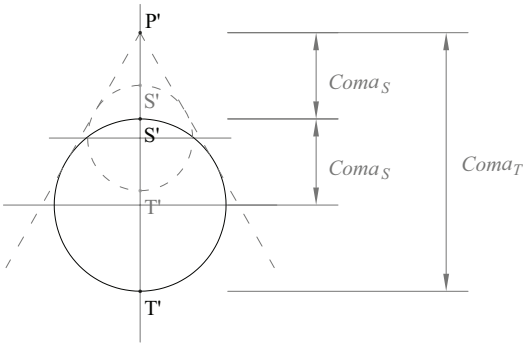


Figure 1.95 Coma aberration geometry. Each circle corresponds to a diameter of the AS.

Example: the geometrical point spread function

So far, we have made a qualitative description of the primary optical aberrations separately in order to understand the main characteristics of each of them. However, in practice, an image has a mixture of aberrations. In the Gaussian image plane, the image of a point outside the optical axis will be, in general, an asymmetrical point called a geometrical *point spread function* (PSF).

To see this, consider the optical system shown in Fig. 1.96, whose data are given in Table 1.3. In particular, we are going to consider the formation of the image of two points in the plane of the object: the point on the axis with $h = 0$ and the point off-axis with $h = -10$ mm. The focal length of this lens is $f = 49.55$ mm, the height of the Gaussian image is $h' = 9.3844$ mm, and the height of the chief ray in the Gaussian image plane is $\bar{h}' = 9.2817$ mm. Thus, the distortion of the chief ray is $Dist(h = -10) = -0.1027$ mm.

Color convention (Figs. 1.97–1.101). The PSF will be drawn as a diagram of points (small circles) corresponding to the intersections of the rays emerging from a point object with the image plane. This type of diagram is called a *spot diagram*. We will present the spot diagram in the image plane in two ways: first, for tangential and sagittal rays uniformly distributed at the AS (both in black, along the x' and y' directions); second, for marginal rays uniformly

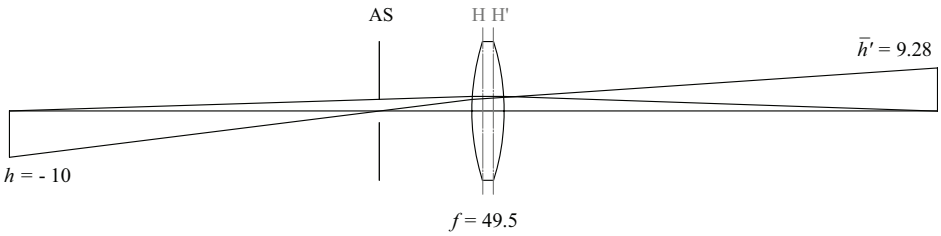


Figure 1.96 Optical system to view the point spread function in different image planes.

Table 1.3 A biconvex lens optical system with an AS. Radii and thicknesses are in millimeters.

Surface	Radius	Thickness	Refractive index
0 (Obj)	∞		
		80	1.0000
1 (AS)	∞		
		20	1.0000
2	50		
		7	1.5168
3	− 50		
		93.7	1.0000
4 (Imag)	∞		

distributed on the perimeter of three concentric circles with the AS, with diameters $D_1 = D_{DA}$ (black), $D_2 = 2D_{DA}/3$ (green), and $D_3 = D_{DA}/3$ (magenta). Points T_1 and T_2 of the marginal tangential rays are blue, and points S_1 and S_2 of the marginal sagittal rays are red.

The calculations to obtain the spot diagrams are performed with the equations for the exact ray tracing presented by Kingslake.* In all cases, the origin of the coordinates corresponds to the intersection of the chief ray with the image plane.

In Fig. 1.97, spot diagrams for the point object are shown on the axis when the AS has a diameter of 5 mm. In (a), the image is observed in the Gaussian image plane. The size of the image turns out to be a circular spot approximately 55 μm in diameter. Note that although the marginal rays are distributed in the AS in three circles whose radii vary uniformly, the same is not the case with the radii of the circles in the image. Due to the quadratic dependence of longitudinal spherical aberration on the size of the AS, the size of the circles in the image varies with the cube of the AS diameter. Thus, while for an aperture of 5 mm in diameter a spot of approximately 55 μm in diameter is obtained, for an aperture of 1.67 mm in diameter a spot of approximately 2 μm in diameter is obtained, which represents a great improvement in the image. In (b), the image size is reduced when the image plane is displaced $\Delta s' = -0.63$ mm from the Gaussian image plane. This situation corresponds to the caustic waist (Fig. 1.89), in which the circular spot of smaller diameter is obtained. In this case, the image is now about 12 μm in diameter.

Other aberrations appear in the image when the point object is displaced from the optical axis. To see astigmatism aberration consider the off-axis

*In the book *Lens Design Fundamentals* by Rudolf Kingslake (Academic Press, Elsevier, 1978), there is a version of the equations for exact ray tracing in optical systems with symmetry of revolution. The calculations shown in this example were performed with a program that I developed.

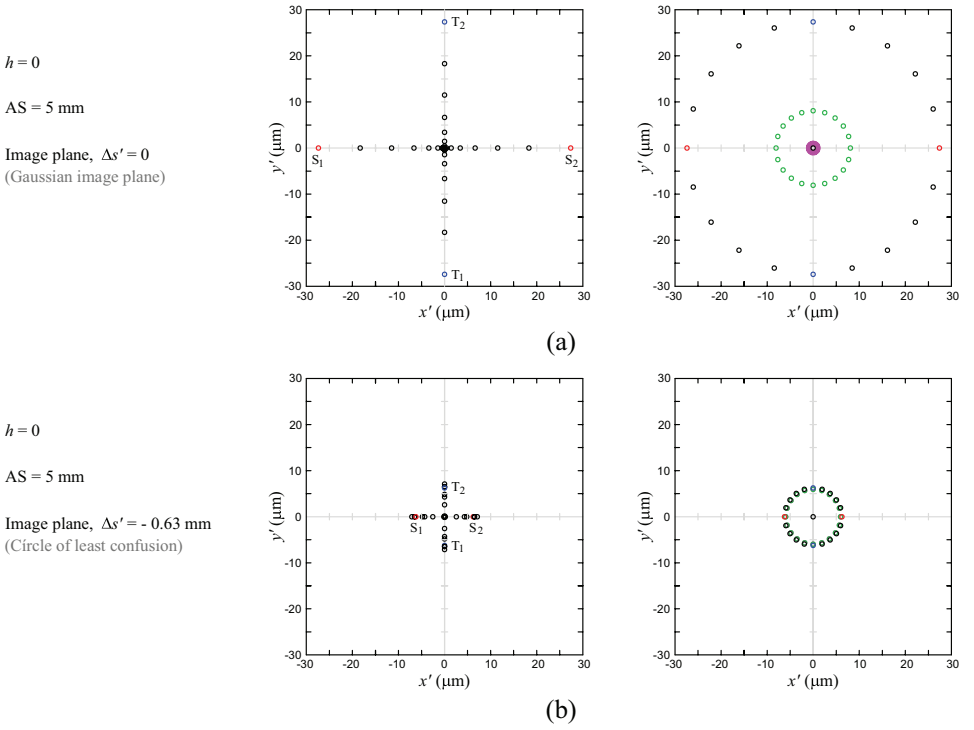


Figure 1.97 Spherical aberration with a 5 mm diameter AS. (a) In the Gaussian image plane, the image is a circular spot of approximately 60 μm. (b) Moving the image plane a distance of -0.63 mm from the Gaussian image plane, the smallest circular spot (circle of least confusion) can be seen with a size of approximately 12 μm. See the color convention described on page 94.

point object ($h = -10$ mm), with an AS of 1 mm in diameter. The corresponding spot diagram, which resembles an ellipse, is shown in Fig. 1.98(a). The asymmetry of the ellipse is due to the presence of a coma. In the direction of the tangential plane, the size is approximately 70 μm. Note that for the 1 mm aperture, the diameter of the circular spot, due to spherical aberration, will be about 60/125 mm, much smaller than the size of the major axis of the astigmatic ellipse. Therefore, the increase in PSF is mainly due to astigmatism. In Fig. 1.98(b), the image plane was shifted $\Delta s' = -2.0$ mm with respect to the Gaussian image plane to find the sagittal focal line. In fact, points S_1 and S_2 coincide. In Fig. 1.98(c), the image plane was shifted $\Delta s' = -4.5$ mm with respect to the Gaussian image plane to find the tangential focal line. Now the points T_1 and T_2 coincide and the points S_1 and S_2 separate, exchanging their positions with respect to the positions they have in Fig. 1.98(a). This tells us that $AstL \approx 2.5$ mm.*

*In the ray tracing program, for astigmatism and coma, the image plane is shifted in 0.1 mm steps. Therefore, the value of the $AstL$ is an approximate number.

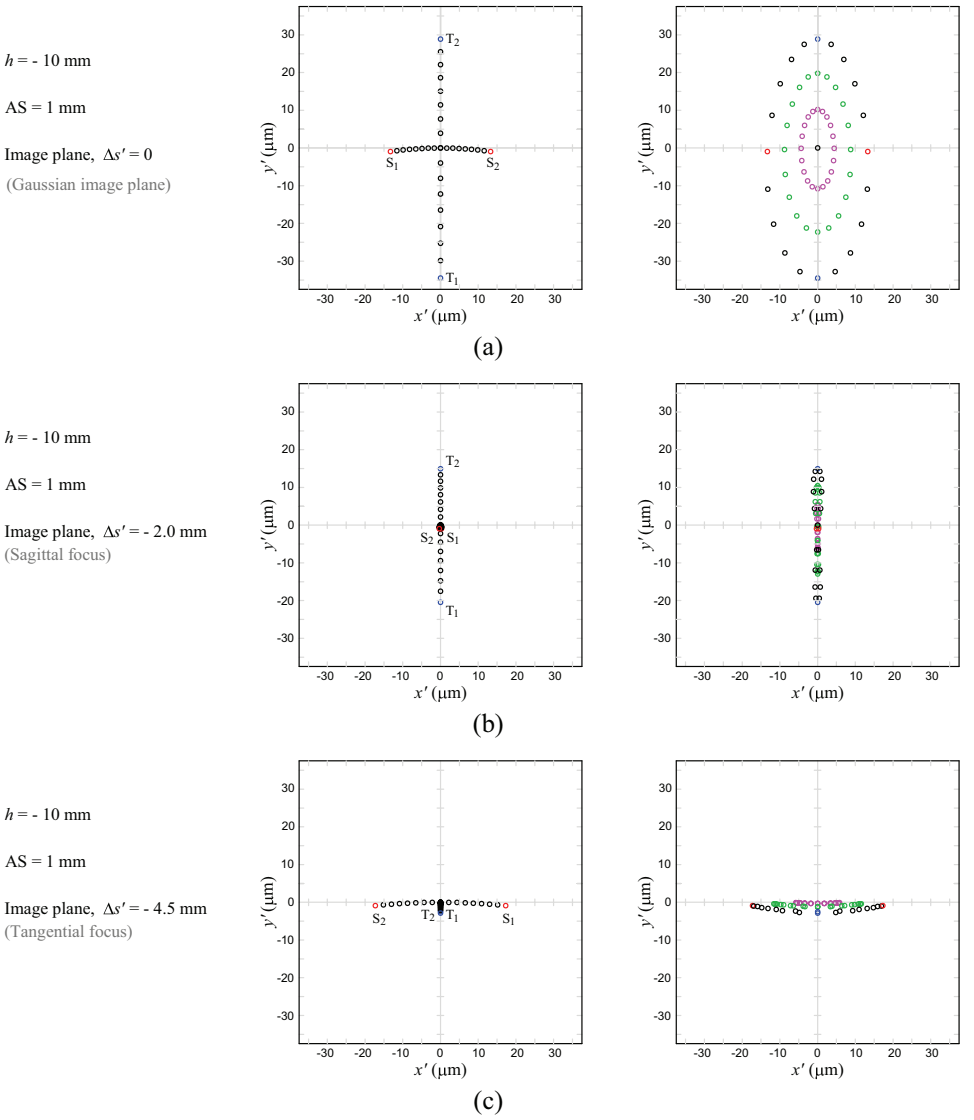


Figure 1.98 Astigmatism aberration. (a) In the Gaussian image plane, the spot diagram resembles an ellipse, as shown in Fig. 1.94(e). Here the ellipse is slightly asymmetrical due to the presence of a coma. In (b) and (c), the sagittal and tangential focal lines are shown, respectively. The effect of coma aberration is small, and therefore the sagittal and tangential focal lines show up nicely in the figure. See the color convention on page 94.

By changing the size of the AS to 5 mm in diameter, the effect of coma aberration becomes noticeable. In Fig. 1.99(a), the spot diagram is more asymmetrical and now looks like a comatic shape (Fig. 1.95). The image size also increased significantly, measuring 370 μm along the tangential plane, compared with the image size affected only by spherical aberration (60 μm),

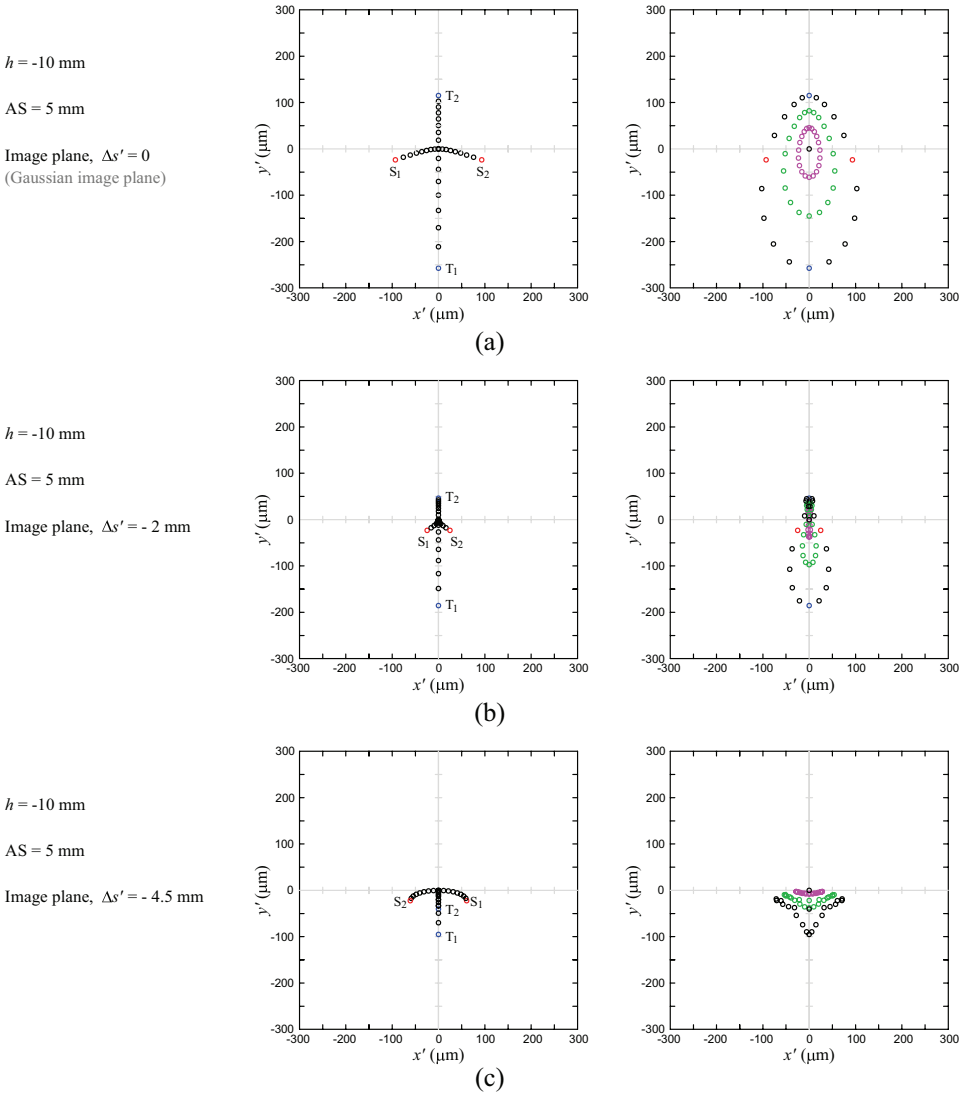


Figure 1.99 The coma aberration becomes more relevant as the diameter of the AS increases. (a) On the Gaussian image plane, the cometic form becomes apparent. (b and c) The image plane is shifted to find the astigmatic lines, which are heavily affected by the coma. See the color convention on page 94.

as shown in Fig. 1.97(a). When moving the image plane to the positions where the focal lines shown in Fig. 1.98 were found, the spot diagram now differs greatly from a couple of lines (vertical and horizontal). This is a result of the increase in coma as the diameter of the AS increases. Note that as the aperture diameter increases, the sagittal marginal radii S_1 and S_2 and the tangential marginal radii T_1 and T_2 also increases; therefore, these rays will be focused at

different positions than when the AS is 1 mm diameter. This explains why in Figs. 1.99(b) and (c) the marginal sagittal radii and the marginal tangential radii do not coincide as they do in Figs. 1.98(b) and (c).

For the marginal sagittal and marginal tangential rays to coincide, the image plane must be shifted further. In Fig. 1.100(a), the image plane has shifted $\Delta s' = -2.7$ mm, where S_1 and S_2 coincide, i.e., where the sagittal focal line would be. In Fig. 1.100(b), the image plane has shifted $\Delta s' = -5.3$ mm, where T_1 and T_2 coincide, i.e., where the tangential focal line would be. Although the spot diagrams in (a) and (b) are far from being a vertical line [in (a)] and a horizontal line [in (b)], the overall size is smaller by an approximate factor of 2. Note that the difference between the two positions of the image planes is $AstL \approx 2.6$ mm. This shows that the $AstL$ does not in fact depend on the size of the AS.

Finally, consider again Fig. 1.98. Here we have three positions for the image plane. Which one best reproduces the point object? Ideally, the image of a point object is a point, so we would expect the image to resemble a circular spot. In Fig. 1.94, in the middle of the astigmatic surfaces is the circle of least confusion. This would be the best image we can have in the presence

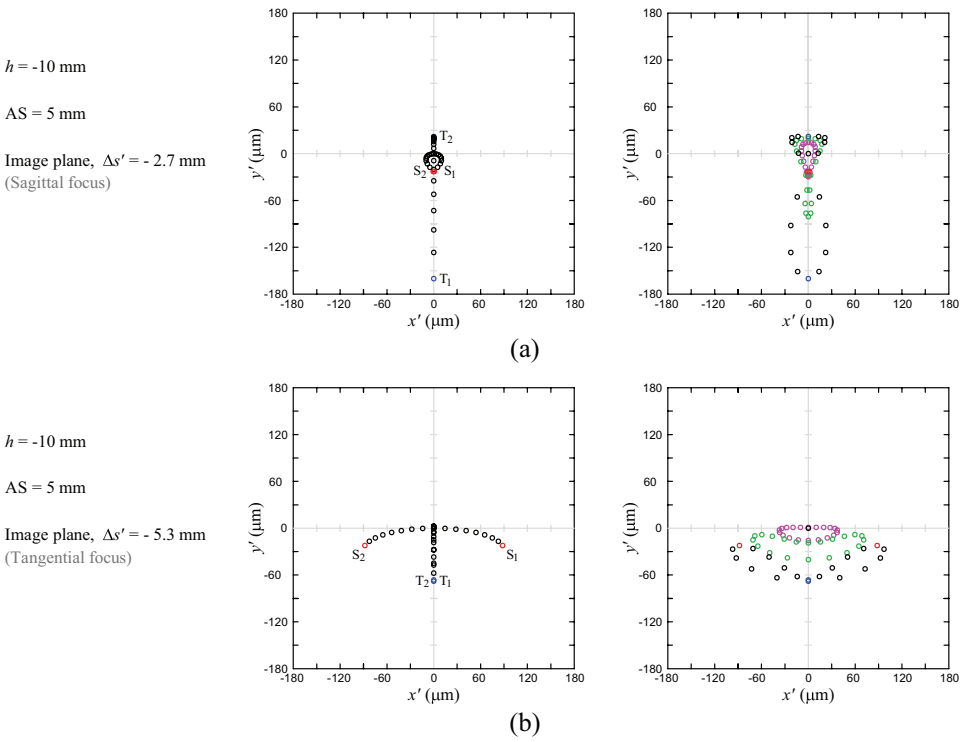


Figure 1.100 Spot diagram in which the sagittal marginal rays (a) and the tangential marginal rays (b) coincide. This corresponds to the focal lines when the AS has a diameter of 5 mm. See the color convention on page 94.

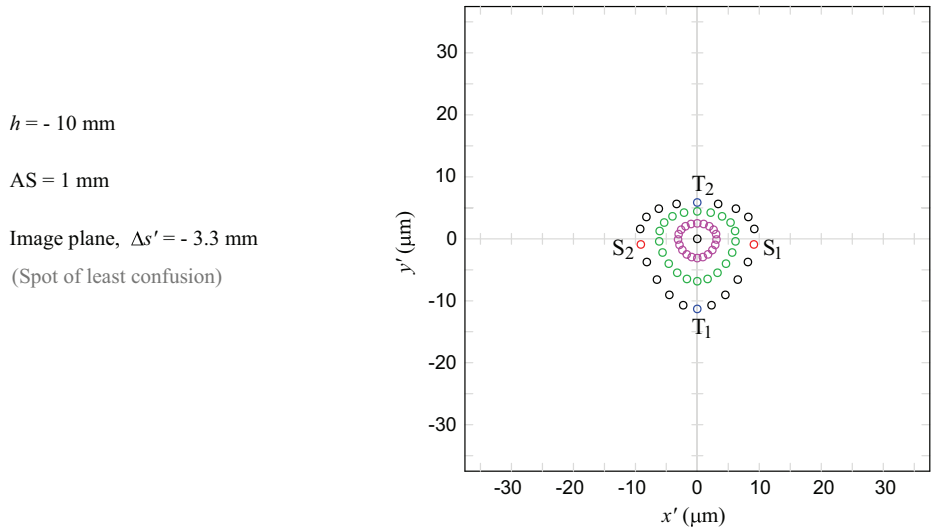


Figure 1.101 Spot diagram corresponding to the circle of least confusion for astigmatic surfaces when the diameter of the AS is 1 mm. The asymmetry is due to the presence of a coma aberration. See the color convention on page 94.

of astigmatism. For our example, with the 1 mm diameter AS, the image plane should be shifted $\Delta s' = -3.3$ mm from the image plane, which turns out to be $AstL/2 = 1.3$ mm toward the left surface S. The corresponding spot diagram is shown in Fig. 1.101. Although the diaphragm is small, the effect of the coma on the asymmetry of the PSF is noticeable. This is the best image for the point object $h = -10$ mm. The center point of the image is at $\bar{h}' = 9.28$ mm. If the height of the point object is changed, the circle of least confusion will be in another position for the image plane. Therefore, on an extended object, the image will be well in focus in some regions and out of focus in others. Lenses are combined in an optical system to reduce optical aberrations, so that the overall focus of the image is optimal.

When the illumination of the object is polychromatic, another type of aberrations occurs: axial chromatic aberration and transverse chromatic aberration. These aberrations are defined and briefly described in Appendix D.

References

[1] S. R.Wilk, "Ibn al-Haytham: 1,000 years after the Kitāb al-Manāzir," *Optics & Photonics News*, 26(10), 42–48 (2015).

[2] I. Newton, *Opticks: or, a Treatise of the Reflections, Refractions, Inflections and Colours of Light*, Sam. Smith and Benj. Walford, London (1704).

[3] E. Hecht, *Optics*, Global Edition, 5th ed., Pearson, Harlow, England (2017).

[4] W. E. Humphrey, "Method and apparatus for analysis of corneal shape," US Patent 4420228 (1983).
[5] H. von Helmholtz, *Helmholtz's Treatise on Physiological Optics*, Vol. 1, Optical Society of America, Washington, D.C. (1924).
[6] Y. LeGrand and S. G. El Hage, *Physiological Optics*, Vol. 13, Springer, New York (2013).
[7] H. Davson, *Physiology of the Eye*, Macmillan International Higher Education, New York (1990).
[8] G. Smith and D. A. Atchison, *The Eye and Visual Optical Instruments*, Cambridge University Press, Cambridge, England (1997).
[9] D. A. Atchison and G. Smith, *Optics of the Human Eye*, Vol. 35, Butterworth-Heinemann, Oxford, England (2000).
[10] V. N. Mahajan, *Optical Imaging and Aberrations: Ray Geometrical Optics*, SPIE Press, Bellingham, Washington (1998).
[11] S. Dupr, "Galileo, the Telescope and the Science of Optics in the Sixteenth Century," Ph.D. thesis, Ghent University, Ghent, Belgium (2002).
[12] D. Malacara and Z. Malacara, *Handbook of Optical Design*, CRC Press, Boca Raton, Florida (2016).