CHAPTER 1
GEOMETRICAL OPTICS

In this chapter we will introduce most of the basic concepts of geometrical optics, although it is likely that most readers will be familiar with these concepts from a study of more elementary texts.

Although all of the basic principles of geometrical optics can be derived from a knowledge of the wave nature of light, we will not follow this approach here. Except in special cases, an understanding of the rigorous derivations of these principles is not helpful in lens design.

However, we will discuss the limitations of geometrical optics when they are relevant to lens design, because there are situations in which an understanding of physical optics is absolutely essential. In lens design, as in other branches of applied science, it is most helpful to use the simplest approximation that can be used for any given task.

Geometrical optics can be considered to describe, with a high degree of accuracy, the properties of lenses as the wavelength of the radiation, \( \lambda \), approaches zero. In this situation, diffraction effects disappear. So geometrical optics will be quite accurate for the design of short-wavelength x-ray imaging systems (if the wavelength is short enough). On the other hand, geometrical optics is rarely completely adequate for the design of thermal imaging systems operating at wavelengths around 10 \( \mu \)m. In the visible waveband, some lenses can be designed and evaluated completely by using geometrical optics, while the evaluation of other lenses must use physical optics. However, in almost all cases, lenses are actually designed using the results of geometrical optics.

1.1 Coordinate system and notation

In this book we discuss primarily the design of centered optical systems. We define the optical axis of a lens to be the \( z \)-axis, with the \( y \)-axis in the plane of the diagram, in Fig. 1.1. The \( x \)-axis is orthogonal to the \( y \)- and \( z \)-axes; in a right-handed coordinate system the \( x \)-axis is positive into the diagram. In the case of lenses that are centered, the \( z \)-axis represents the common optical axis of the refracting and reflecting surfaces.

In many equations in geometrical optics, we are concerned with quantities that are affected by refraction or reflection at a surface or at a lens. In these cases, we represent quantities after refraction or reflection as primed quantities; for example, we shall see that we write \( n' \) for the refractive index after a surface in Eq. (1.1) below.
1.2 The rectilinear propagation of light

One of the most obvious properties of light, which is seen very clearly by observation of the path of a laser beam, is that it propagates in straight lines. In reality this is an approximation. As physical optics predicts and experiment confirms, any beam of light diverges to an extent determined by the beam width and the wavelength.

Furthermore, the rectilinear propagation of light is dependent on the uniformity of the medium through which it is passing. The classical example of a situation in which the medium is not uniform is the atmosphere; the existence of mirages and the compression of the image of the sun, when it is very near the horizon, are both due to nonuniformity of the atmosphere. More recent examples of nonuniform materials are gradient-index lenses.

Despite these reservations, the geometrical optics presented here will assume rectilinear propagation of light, embodied in the concept of the light ray. Many results that are discussed in Chapters 1 through 7 will demonstrate the usefulness of this approximation.

1.3 Snell’s law

The inception of optical design, in my opinion, occurred in 1621. In that year, Snell formulated the law of refraction, which states that if the angle between an incident ray and the surface normal at the point of incidence, called the angle of incidence, is \( I \); and if the angle of refraction, the angle between the refracted ray and the normal, is \( I’ \); then the angles are related by the equation

\[
n \sin I = n’ \sin I’. \quad (1.1)
\]

In addition, Snell’s law states that the incident ray, the refracted ray and the normal to the surface at the point of incidence are all in the same plane (Fig. 1.2). The quantities \( n \) and \( n’ \) are the refractive indices of the two materials. While
Eq. (1.1) can be taken as the definition of refractive index, it should also be noted that the refractive index of a material is more fundamentally defined as

\[
n = \frac{c}{v},
\]

(1.2)

where \( c \) is the velocity of light in a vacuum and \( v \) is the velocity of light in the material.

\[\text{(n)} \quad \text{(n')} \]

Figure 1.2. Snell’s law (the law of refraction).

Since, at any surface, the ratio of the refractive indices determines the refracted ray angle, it is convenient to write

\[
q = \frac{n}{n'},
\]

(1.3)

so that Snell’s law is simplified to

\[
\sin I' = q \sin I.
\]

(1.4)

In the case of reflection (Fig. 1.3), the angle of the reflected ray is equal to the angle of incidence. Because of the sign convention for angles in ray tracing, the two angles have opposite sign. Therefore, the law of reflection is given as

\[
I' = -I.
\]

(1.5)

In lens design, it is quite usual to treat reflection as a special case of refraction, by assuming that

\[
n' = -n \text{ or } q = -1.
\]

(1.6)

This device is very useful in designing centered systems with reflecting surfaces, as the formulae for refraction can be applied unchanged to reflective surfaces, provided we adopt the convention that the refractive index changes sign after each reflection. After an even number of reflections, when the rays are
traveling in the same sense as their initial direction, the refractive index will be positive; after an odd number of reflections, the refractive index will be assumed to be negative.

![Figure 1.3. The law of reflection.](image)

In the case of complex decentered systems, such as systems with several folding mirrors, this convention can become very confusing, and it is probably then more convenient to treat reflection as a separate case, keeping all refractive indices positive.

### 1.4 Fermat’s principle

Fermat’s principle is one of the most important theorems of geometrical optics. While it is not used directly in practical lens design (unlike Snell’s law) it is used to derive results that would be impossible, or more difficult, to derive in other ways.

It may be stated as follows.

Figure 1.4 shows a physically possible path for a ray from A to D, and let the lengths of the segments along the ray be $d_1, d_2, d_3$.

![Figure 1.4. Optical path length.](image)

We then define the optical path length in any medium to be the product of the distance traveled and the refractive index:
Optical path length $= [ABCD] = \sum n_i d_i$ \hspace{1cm} (1.7)

where the square brackets are used to distinguish the optical path length from a geometrical distance.

Fermat’s principle states that the optical path length along a physically possible ray is stationary. For example, take the simple case of a plane refracting surface, as shown in Fig. 1.5.

Here we have a ray passing through two points A and C; it is assumed to intersect the plane refracting surface at B. Fermat’s principle states that if we write an expression for the optical path length as a function of $y$, and then differentiate with respect to $y$, the point where the differential is zero will represent the point B.

**1.5 Rays and wavefronts—the theorem of Malus and Dupin**

We have encountered the concepts of rays and optical path length, and we must now define the wavefront. In geometrical optics we define a wavefront as follows:

A wavefront is a surface of constant optical path length, from a point in the object.

In other words, if we trace several rays from a source at a point A, as shown in Fig. 1.6, then the points $B_1$, $B_2$, $B_3$, etc., all represent points that have the same optical path length from A. In the case of these points it is clear that since they are in the same medium as A, the locus of the points $B_1$, $B_2$, $B_3$, etc., is a surface of constant optical path length and therefore a wavefront. It is a sphere centered on A. If, however, we take the points $C_1$, $C_2$, $C_3$, etc., which are not in the same medium as A, the wavefront that is described by these points is not, in general, a sphere.
Note that this definition of a wavefront depends purely on geometrical optics, and we have not considered physical optics at all. In fact, in most cases, this geometrical wavefront does correspond to a surface of constant phase, as determined by physical optics; the major exception to this correspondence is the case of a Gaussian laser beam, with a very low convergence angle.

The theorem of Malus and Dupin states that these geometrical wavefronts are orthogonal to the rays from the point A; in other words, the rays are always perpendicular to the geometrical wavefront.\(^1\)

There is an exception to this. In the case of a nonisotropic material, such as a birefringent crystal, the refractive index depends on the direction of propagation of the ray, and in this case the rays are not, in general, normals to the wavefront.

The major use of this theorem in practical lens design is to assist in understanding the relationship between transverse ray aberrations and wavefront aberrations, which we will discuss later.

### 1.6 Stops and pupils

The diagram below (Fig. 1.7) shows a simple lens system, with a stop between two lenses. If this stop limits the size of the beam from an axial point, it is known as the aperture stop.

Since the diameter of the beam passing through an optical system is always limited by something, every optical system has an aperture stop of some sort. In some cases, the aperture stop is a separate entity, as in the diagram. In other cases the aperture stop may simply be part of the lens mount.
1.6.1 Marginal and chief rays

In Fig. 1.7, the ray that passes from the center of the object, at the maximum aperture of the lens, is normally known as the marginal ray. It therefore passes through the edge of the aperture stop. Conventionally, this ray is in the $y$-$z$ plane, usually called the meridian plane.

The chief ray is defined to be the ray from an off-axis point in the object passing through the center of the aperture stop; although there can be an infinite number of such rays, we can usually assume, at least for centered systems, that the chief ray is also restricted to the meridian plane.

1.6.2 Entrance and exit pupils

In the case of a lens with an internal aperture stop, as in Fig. 1.7, the concepts of entrance pupil and exit pupil are particularly important. If we view the lens in Fig. 1.7 from the object space, we see an image of the aperture stop, known as the entrance pupil. Its position is found by the construction shown in Fig. 1.8.
In the image space, the image of the aperture stop is known as the exit pupil, and its position is found using a similar construction.

### 1.6.3 Field stops

In addition to the aperture stop, which limits the size of the image-forming beam, all optical systems also have a field stop, which limits the size of the image or the object. This field stop is normally at, or close to, the object or image surface.

In some cases, the function of the field stop is very clear. For example, in the case of eyepieces used in microscopes or binoculars, the size of the field of view is limited by a physical stop at the plane of the intermediate image. In the case of a camera, the field stop is effectively the “gate” that defines the area of the film that is exposed.

### 1.7 Surfaces

The interface between media of different refractive indices may be described by any number of geometrical shapes. Here we will consider only some of the more common forms.

#### 1.7.1 Spheres

Since many lenses are constructed from surfaces that are nominally spherical in form, it is convenient at this point to discuss several forms of the equations that can be used to represent a spherical surface.

The equation of a sphere, of radius $R$, and passing through the origin is given by

$$R^2 = x^2 + y^2 + (z - R)^2. \quad (1.8)$$

Writing $r^2 = x^2 + y^2$, for convenience

$$R^2 = r^2 + (z - R)^2, \quad (1.9)$$

or

$$z^2 - 2zR + r^2 = 0, \quad (1.10)$$

giving

$$z = R \pm \sqrt{R^2 - r^2}. \quad (1.11)$$
For the first intersection point (as shown in Fig. 1.9) we need the negative square root, giving

\[ z = R - \sqrt{R^2 - r^2}. \]  \hspace{1cm} (1.12)

**Figure 1.9.** The sphere.

Since \( R \) is often large compared with \( r \), we would be determining \( z \) by the difference between two large numbers.

To overcome this, we first replace Eq. (1.11) by

\[ z = R \left[ 1 - \sqrt{1 - \left(\frac{r}{R}\right)^2} \right], \]  \hspace{1cm} (1.13)

which always gives us the solution nearest to the vertex of the surface, and we also replace the radius of curvature of the surface, \( R \), with \( 1/c \), where \( c \) is the curvature of the surface. Then we have

\[ z = \frac{1 - \sqrt{1 - (cr)^2}}{c}. \]  \hspace{1cm} (1.14)

By simple algebra we can also write this equation as

\[ z = \frac{cr^2}{1 + \sqrt{1 - (cr)^2}}. \]  \hspace{1cm} (1.15)