1

Geometrical Optics

1.1 Geometrical Optics – Ray and Wave Optics

In describing *optical* systems, in the narrow definition of the term, we might only consider systems that manipulate visible light. However, for the optical engineer, the application of the science of optics extends well beyond the narrow boundaries of human vision. This is particularly true for modern instruments, where reliance on the human eye as the final detector is much diminished. In practice, the term optical might also be applied to radiation that is manipulated in the same way as visible light, using components such as lenses, mirrors, and prisms. Therefore, the word 'optical', in this context might describe electromagnetic radiation extending from the vacuum ultraviolet to the mid-infrared (*wavelengths from ~120 to ~10 000 nm*) and perhaps beyond these limits. It certainly need not be constrained to the narrow band of visible light between about 430 and 680 nm. Figure 1.1 illustrates the electromagnetic spectrum.

Geometrical optics is a framework for understanding the behaviour of light in terms of the propagation of light as highly directional, narrow bundles of energy, or *rays*, with 'arrow like' properties. Although this is an incomplete description from a theoretical perspective, the use of ray optics lies at the heart of much of practical optical design. It forms the basis of optical design software for designing complex optical instruments and geometrical optics and, therefore, underpins much of modern optical engineering.

Geometrical optics models light entirely in terms of infinitesimally narrow beams of light or rays. It would be useful, at this point, to provide a more complete conceptual description of a ray. Excluding, for the purposes of this discussion, quantum effects, light may be satisfactorily described as an electromagnetic wave. These waves propagate through free space (vacuum) or some optical medium such as water and glass and are described by a wave equation, as derived from Maxwell's equations:

$$\frac{\partial^2 E}{\partial x^2} + \frac{\partial^2 E}{\partial y^2} + \frac{\partial^2 E}{\partial z^2} = \frac{n^2}{c^2}\frac{\partial^2 E}{\partial z^2} \tag{1.1}$$

E is a *scalar* representation of the local electric field; **c** is the velocity of light in free space, and **n** is the refractive index of the medium.

Of course, in reality, the local electric field is a ***vector*** quantity and the ***scalar theory*** presented here is a useful initial simplification. Breakdown of this approximation will be considered later when we consider polarisation effects in light propagation. If one imagines waves propagating from a central point, the wave equation offers solutions of the following form:

$$E = \frac{E_0}{r} e^{i(kr-\omega t)} \tag{1.2}$$

Equation (1.2) represents a spherical wave of angular frequency, ω, and spatial frequency, or wavevector, k. The velocity that the wave disturbance propagates with is ω/k or c/n. In free space, light propagates at the speed of light, c, a fundamental and ***defined*** constant in the SI system of units. Thus, the refractive index, n, is the ratio of the speed of light in free space to that in the specified medium. All points lying at the same distance,

Optical Engineering Science, First Edition. Stephen Rolt.

Companion website: www.wiley.com/go/Rolt/opt-eng-sci

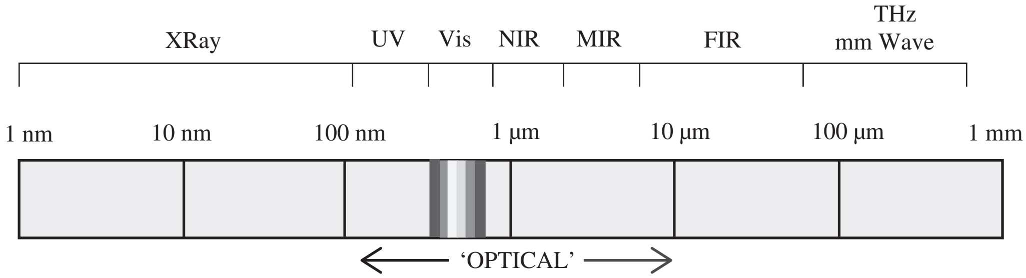

Figure 1.1 The electromagnetic spectrum.

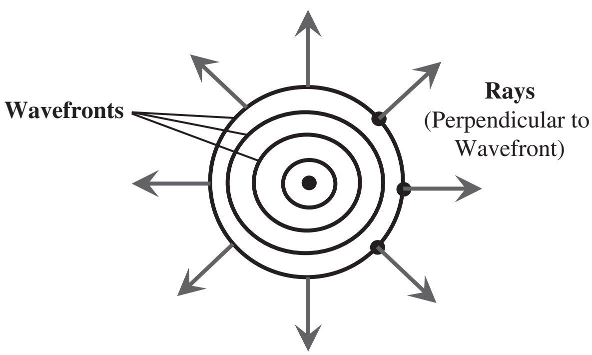

Figure 1.2 Relationship between rays and wavefronts.

r, from the source, will oscillate at an angular frequency, ω, and in the same phase. Successive surfaces, where all points are oscillating entirely in phase are referred to as wavefronts and can be viewed at the crests of ripples emanating from a point disturbance. This is illustrated in Figure 1.2. This picture provides us with a more coherent definition of a ray. *A ray is represented by the vector normal to the wavefront surface in the direction of propagation*. Of course, Figure 1.2 represents a simple spherical wave, with waves spreading from a single point. However, in practice, wavefront surfaces may be much more complex than this. Nevertheless, the precise definition of a ray remains clear:

At any point in space in an optical field, a ray may be defined as the unit vector perpendicular to the surface of constant phase at that point with its sense lying in the same direction as that of the energy propagation.

1.2 Fermat's Principle and the Eikonal Equation

Intuition tells us that light 'travels in straight lines'. That is to say, light propagates between two points in such a way as to minimise the distance travelled. More generally, in fact, all geometric optics is governed by a very simple principle along similar lines. Light always propagates between two points in space in such a way as to minimise the *time* taken. If we consider two points, A and B, and a ray propagating between them within a medium whose refractive index is some arbitrary function, $n(r)$, of position then the time taken is given by:

$$\tau = \frac{1}{c}\int_A^B n(r)ds \tag{1.3}$$

c is the speed of light in vacuo *and ds is an element of path between A and B*

This is illustrated in Figure 1.3.

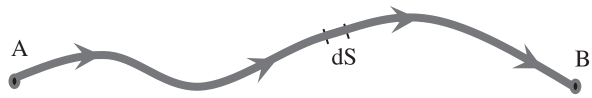

Figure 1.3 Arbitrary ray path between two points.

Fermat's principle may then be stated as follows:

Light will travel between two points A and B such that the path taken represents a local minimum in the total optical path between these points.

Fermat's principle underlies all ray optics. All laws governing refraction and reflection of rays may be derived from Fermat's principle. Most importantly, to demonstrate the theoretical foundation of ray optics and its connection with physical or wave optics, Fermat's principle may be directly derived from the wave equation. This proof demonstrates that the path taken represents, in fact, a ***stationary*** solution with respect to other possible paths. That is to say, technically, the optical path taken ***could*** represent a local ***maximum*** or ***inflexion*** point rather than a minimum. However, for most practical purposes it is correct to say that the path taken represents the minimum possible optical path.

Fermat's principle is more formally set out in the Eikonal equation. Referring to Figure 1.2, if instead of describing the light in terms of rays it is described by the wavefront surfaces themselves. The function $S(r)$ describes the *phase* of the wave at any point and the Eikonal equation, which is derived from the wave equation, is set out thus:

$$\frac{\partial^2 S(r)}{\partial x^2} + \frac{\partial^2 S(r)}{\partial y^2} + \frac{\partial^2 S(r)}{\partial z^2} = n^2 \tag{1.4}$$

The important point about the Eikonal equation is not the equation itself, but the assumptions underlying it. Derivation of the Eikonal equation assumes that the rate of change in phase is small compared to the wavelength of light. That is to say, the radius of curvature of the wavefronts should be significantly larger than the wavelength of light. Outside this regime the assumptions underlying ray optics are not justified. This is where the effects of the wave nature of light (i.e. diffraction) must be considered and we enter the realm of physical optics. But for the time being, in the succeeding chapters we may consider that all optical systems are adequately described by geometrical optics.

So, for the purposes of this discussion, it is one simple principle, Fermat's principle, that provides the foundation for all ray optics. For the time being, we will leave behind specific consideration of the detailed behaviour of individual optical surfaces. In the meantime, we will develop a very generalised description of an idealised optical system that does not attribute specific behaviours to individual components. Later on, this 'black box model' will be used, in conjunction with Gaussian optics to provide a complete first order description of complex optical systems.

1.3 Sequential Geometrical Optics – A Generalised Description

In applying geometrical optics to a real system, we are attempting to determine the path of a ray(s) through the system. There are a few underlying characteristics that underpin most optical systems and help to simplify analysis. First, *most* optical systems are sequential. An optical system might comprise a number of different elements or surfaces, e.g. lenses, mirrors, or prisms. In a sequential optical system, the order in which light propagates through these components is unique and pre-determined. Second, in most practical systems, light is constrained with respect to a mechanical or optical axis of symmetry, the optical axis, as illustrated in Figure 1.4. In real optical systems, light is constrained by the use of physical apertures or 'stops'; this will be discussed in more detail later.

Of course, in practice, the optical axis need not be a continuous, straight line through an optical system. It may be bent, or folded by mirrors or prisms. Nevertheless, there exists an axis throughout the system with respect to which the rays are constrained.

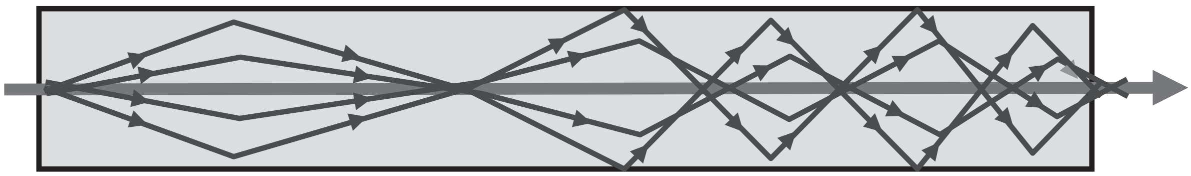

Figure 1.4 Constraint of rays with respect to optical axis.

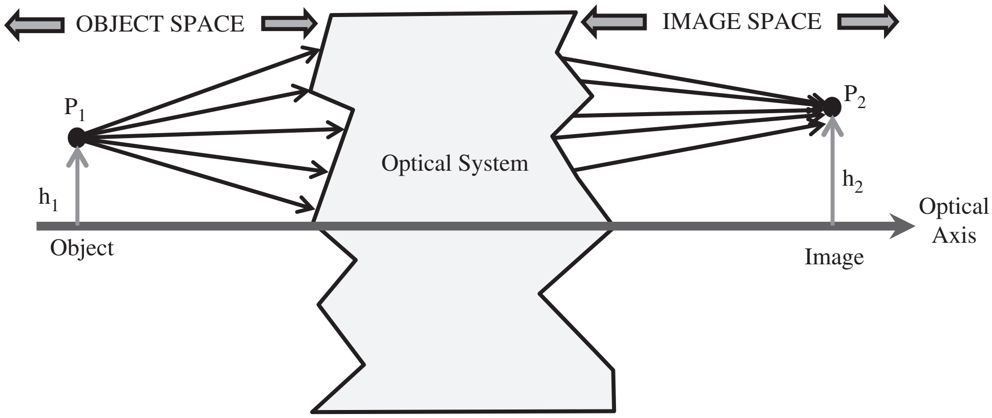

Figure 1.5 Generalised optical system and conjugate points.

1.3.1 Conjugate Points and Perfect Image Formation

We consider an ideal optical system which consists of a point source of light, the ***object***, and an optical system that collects the light and re-directs *all rays* emanating from this point source or object, such that the rays converge onto a single point, the ***image*** point. At this stage, the interior workings of the optical system are undefined; the system behaves as a 'black box'. The object is said to be located in object space and the image in image space and the pair of points are said to be ***conjugate points***. This is illustrated in Figure 1.5.

In Figure 1.5, the two points P_1 and P_2 are conjugate. The optical system can be simple, for example a single lens, or it can be complex, containing many optical elements. The description above is entirely generalised. Where the object point lies on the optical axis, its image or conjugate point also lies on the optical axis. In Figure 1.5, the object point has a height of h_1 with respect to the optical axis and its corresponding image point has a height of h_2 with respect to the same axis. The ratio of these two heights gives the system (transverse) magnification, M:

$$M = h_2/h_1 \tag{1.5}$$

Points occupying a plane perpendicular to the optical axis are conjugate to points lying on another plane perpendicular to the optical axis. These planes are known as conjugate planes.

1.3.2 Infinite Conjugate and Focal Points

Where an image or object is located at infinity, all rays emerging from or travelling to these locations will be parallel with respect to each other. In this instance, the point located at infinity is said to be at an ***infinite conjugate***. The corresponding conjugate point to the infinite conjugate is known as a ***focal point***. There are two focal points. The ***first focal point*** is located in the object space with the corresponding image located at the infinite conjugate. The ***second focal point*** is located in the image space with the object placed at the infinite conjugate. Figure 1.6 depicts the first focal point:

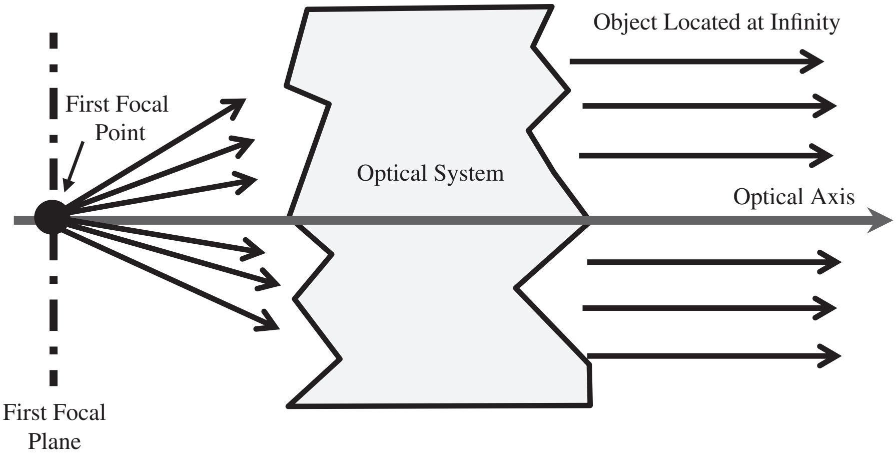

Figure 1.6 Location of first focal point.

As well as focal points, there are two corresponding ***focal planes***. The two focal planes are planes perpendicular to the optical axis that contain the relevant focal point. For all points lying on the relevant focal plane, the conjugate point will lie at the infinite conjugate. In other words, all rays will be parallel with respect to each other. In general, the rays will not be parallel to the optic axis. This would only be the case for a conjugate point lying on the optical axis.

1.3.3 Principal Points and Planes

All points lying on a particular conjugate plane are associated with a specific transverse magnification, M, which is equal to the ratio of the image and object heights. For an ideal system, there exist two conjugate planes where the magnification is unity. These are known as the ***principal planes***. Thus, there are two principal planes and the points where the optical axis intersects the principal planes are known as principal points. The first principal point (plane) is located in object space and the second principal point (plane) is located in image space. The arrangement is illustrated schematically in Figure 1.7.

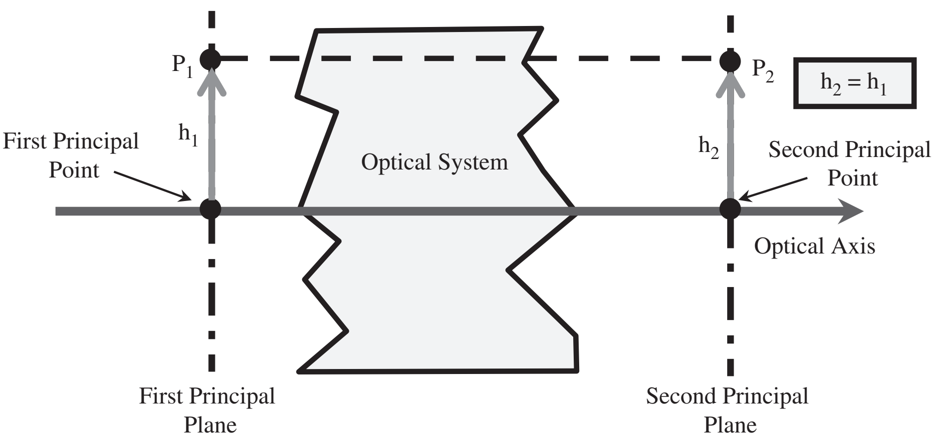

Figure 1.7 Principal points and principal planes.

1.3.4 System Focal Lengths

The reader might be used to ascribing a single focal length to an optical system, such as for a magnifying lens or a camera lens. However, in this general description, the system has two focal lengths. The ***first focal length***, f_1, is the distance from the first focal plane (or point) to the first principal plane (or point) and the ***second focal length***, f_2, is the distance from the second principal plane to the second focal plane. In many cases, f_1 and f_2 are identical. In fact, the ratio f_1/f_2 is equal to n_1/n_2, the ratio of the refractive indices of the media associated with the object and image spaces. However, this need not concern us at this stage, as the treatment presented here is entirely general and independent of the specific attributes of components or media.

In classical geometrical optics, the object location is denoted by the ***object distance***, u, and the image location by the ***image distance***, v. In the context of this general description, the object distance is simply the distance from the object to the first principal plane. Correspondingly, the image distance, v, is the distance from the second principal plane to the image. In addition, the object location can be described by the distance, x_1, separating the object from the corresponding focal plane. Similarly, x_2 represents the distance from the image to the second focal plane. This is illustrated in Figure 1.8.

1.3.5 Generalised Ray Tracing

This general description of an optical system is very economical in that the definition of conjugate points, focal planes, and principal planes provides sufficient information to determine the path of a ray in the image space, given the path of the ray in the object space. No assumptions are made about the internal workings of the optical system; it is merely a ‘black box’.

We see how input rays originating in the object space are mapped onto the image space for specific scenarios where the object is located at the input focal plan, the infinite conjugate, or the first principal plane. How can this be extended to determine the output path of any input ray? The general principle is set out in Figure 1.9. First, the input ray is traced from point P_1 as far as its intersection with the (first) principal plane at A_1. We know that this point, A_1, is conjugated with point A_2, lying at the same height at the second principal plane. This follows directly from the definition of principal planes. Second, we draw a dummy ray originating from the first focal point, f_1, but parallel to the input ray and trace it to where it intersects the first principal plane at B_1. We know that B_1 is conjugated with point B_2, lying at the same height on the second principal plane.

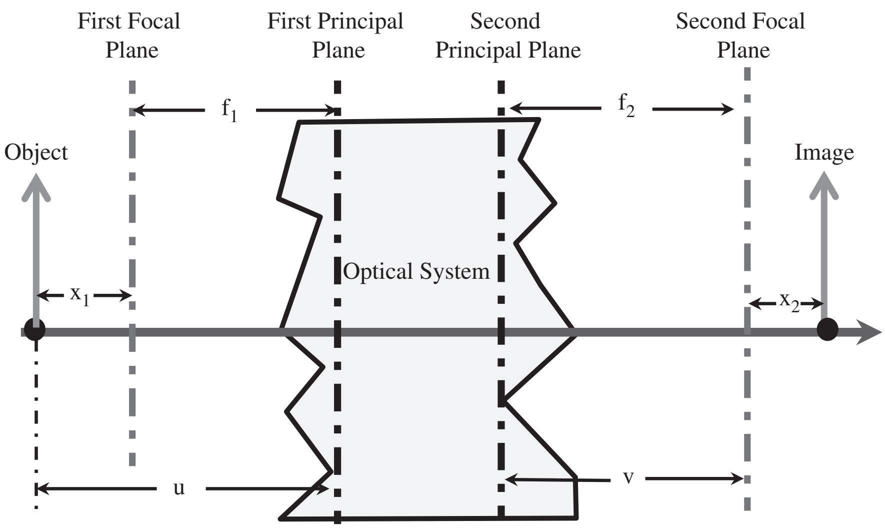

Figure 1.8 System focal lengths.

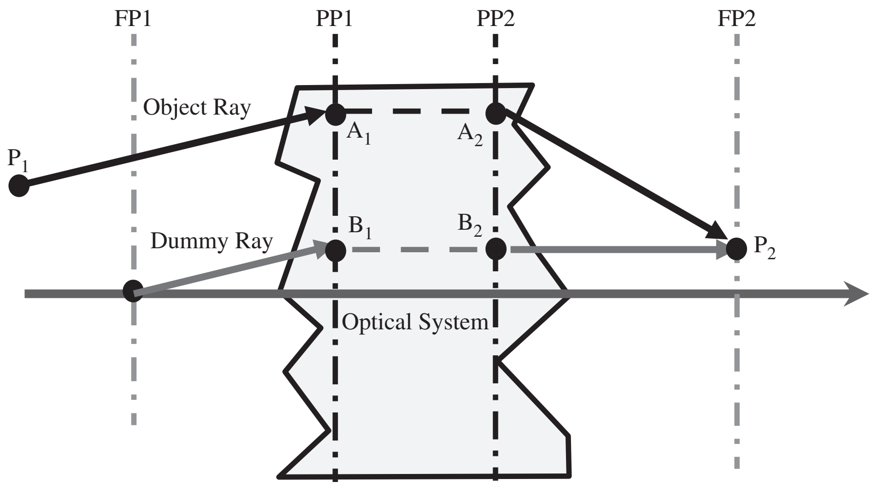

Figure 1.9 Tracing of arbitrary ray.

Since this ray originated from the first focal point, its path must be parallel to the optical axis in image space and thus we can trace it as far as the second focal plane at P_2. Finally, since the object ray and dummy rays are parallel in object space, they must meet at the second focal plane in the image space. Therefore, we can trace the image ray to point P_2, providing a complete definition of the path of the ray in image space.

1.3.6 Angular Magnification and Nodal Points

The ***angular magnification*** of an optical system is the ratio of the angle (with respect to the optical axis) of a ray in image space and that of its conjugate in object space. There exists a pair of conjugate points lying on the optical axis where, for all possible rays, the angular magnification is unity. These are the ***nodal points***. The ***first nodal point*** is located in object space and the ***second nodal point*** is located in image space. This is set out in Figure 1.10, where for a general conjugate pair, the angular magnification, α, is equal to θ_2/θ_1. For

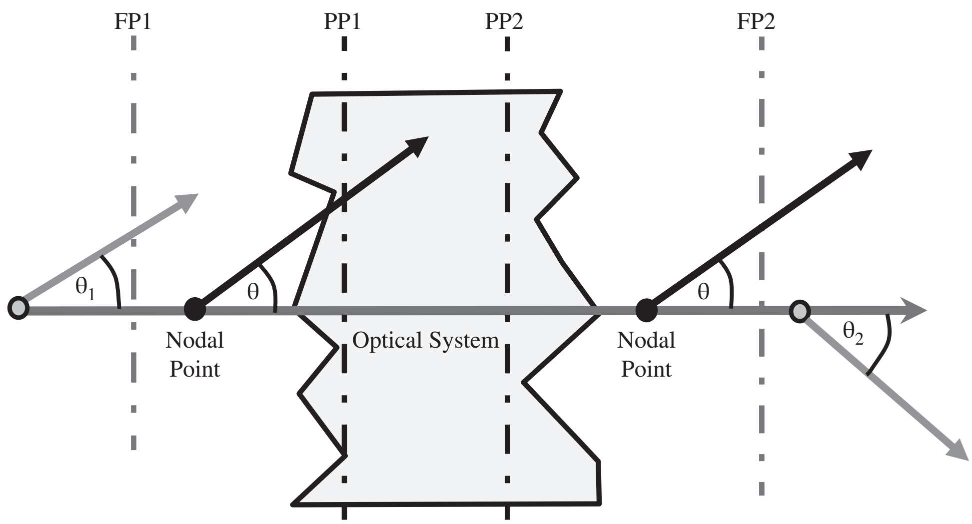

Figure 1.10 Angular magnification and nodal points.

the nodal points, $\theta_2 = \theta_1$; that is to say, the angular magnification is unity. Where the two focal lengths are identical, or the object and image spaces are within media of the same refractive index, the nodal points are co-located with the principal points.

1.3.7 Cardinal Points

This brief description has provided a complete definition of an ideal optical system. No matter how complex (or simple) the optical system, this analysis defines the complete end-to-end functionality of an ideal system. On this basis, an optical designer will specify the six ***cardinal points*** of a system to describe the ideal behaviour of a design. These six cardinal points are:

First Focal Point
Second Focal Point
First Principal Point
Second Principal Point
First Nodal Point
Second Nodal Point

The principal and nodal points are co-located if the two system focal lengths are identical.

1.3.8 Object and Image Locations - Newton's Equation

The location of the cardinal points has given us a complete description of a generalised optical system. Given that the function of an optical system might be to produce an image of an object located at a specific point, we might want to know the location of that image. Figure 1.11 shows the relationship between a generalised object and image.

Referring to Figure 1.11 and by using similar triangles it is possible to derive two separate relations for the magnification h_2/h_1:

$$M = \frac{h_2}{h_1} = -\frac{f_1}{x_1} = -\frac{x_2}{f_2}$$

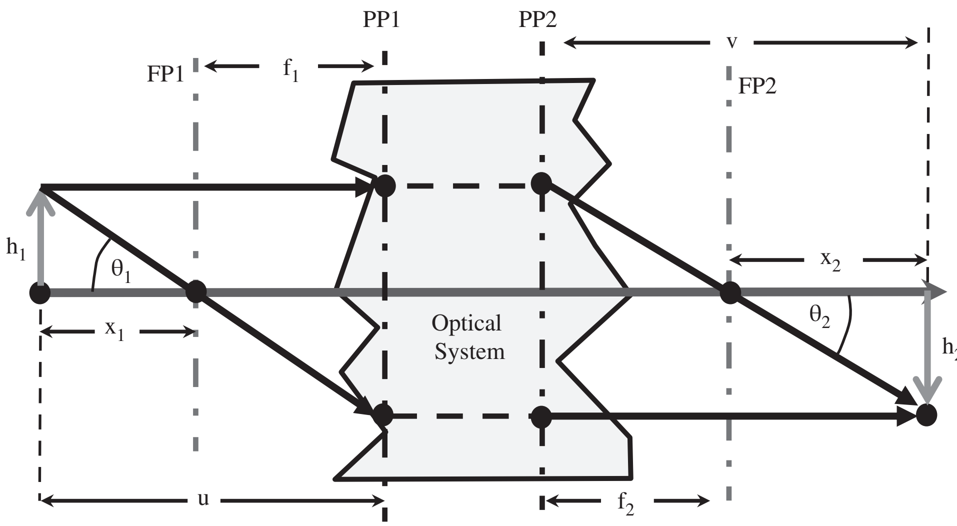

Figure 1.11 Generalised object and image.

And:

$$\text{Newton's Equation}: \quad x_1x_2 = f_1f_2 \tag{1.6}$$

The above equation is ***Newton's Equation*** and may be re-cast into a more familiar form using the definitions of object and image distances, u and v, as previously set out.

$$\frac{1}{u} + \left(\frac{f_2}{f_1}\right)\frac{1}{v} = \frac{1}{f_1} \tag{1.7}$$

If $f_1 = f_2 = f$, we are left with the more familiar lens equation. However, Eq. (1.7) is generally applicable to all optical systems. Most importantly, Eq. (1.7) will give the locations of the object and image in systems of arbitrary complexity. Many readers might have encountered Eq. (1.7) in the context of a simple lens where object and image distances are obvious and easy to determine. For a more complex system, one has to know the location of the principal planes as well in order to determine the object and image distances.

1.3.9 Conditions for Perfect Image Formation – Helmholtz Equation

Thus far, we have presented a description of an idealised optical system. Is there a simple condition that needs to be fulfilled in order to generate such an ideal image? It is easy to see from Figure 1.11 that the following relations apply:

$$f_1 \tan\theta_1 = h_2 \, and f_2 \tan\theta_2 = h_1$$

Therefore:

$$h_1f_1 \tan\theta_1 = h_2f_2 \tan\theta_2$$

As we will be able to show later, the ratio f_2/f_1 is equal to the ratio of the refractive indices, n_2/n_1, in the two media (object and image space). Therefore it is possible to cast the above equation in its more usual form, the ***Helmholtz equation***:

$$\text{Helmholtz equation}: \quad h_1n_1 \tan\theta_1 = h_2n_2 \tan\theta_2. \tag{1.8}$$

One important consequence of the Helmholtz equation is that there is a clear, inextricable linkage between transverse and angular magnification. Angular magnification is inversely proportional to transverse magnification. For small θ, tan θ and θ are approximately equal. So in the small signal approximation, the angular magnification, α is given by:

$$\alpha = \frac{\theta_2}{\theta_1} \approx \frac{h_1n_1}{h_2n_2}$$

Hence:

$$\alpha \approx \left(\frac{n_1}{n_2}\right)\frac{1}{M} \tag{1.9}$$

We have, thus far, introduced two different types of optical magnification – transverse and angular. There is a third type of magnification that we need to consider, ***longitudinal magnification***. Longitudinal magnitude, L, is defined as the shift in the axial image position for a unit shift in the object position, i.e.:

$$L = \frac{dx_2}{dx_1} \tag{1.10}$$

From Newton's Eq. (1.6):

$$\frac{dx_2}{dx_1} = \frac{f_1f_2}{x_1^2} = -\left(\frac{f_2}{f_1}\right)M^2$$

And:

$$L = -\left(\frac{f_2}{f_1}\right)M^2 \tag{1.11}$$

Thus, the longitudinal magnification is proportional to the square of the transverse magnification.

1.4 Behaviour of Simple Optical Components and Surfaces

1.4.1 General

The analysis presented thus far is entirely independent of the optical components that might populate the idealised optical system. In this section we will begin to consider, from the perspective of ray optics, the behaviour of real elements that make up this generalised system. At a basic level, only a few behaviours need to be considered in order to understand the propagation of rays through a real optical system. These are:

Propagation through a homogeneous medium
Refraction at a planar surface
Refraction at a curved (spherical) surface
Refraction through lenses
Reflection at a planar surface
Reflection at a curved (spherical) surface

As previously set out, the path of rays through a system is governed entirely by Fermat's principle. From this point, we will apply the simplest definition of Fermat's principle and assume that the time or optical path of rays is ***minimised***. As far as ***propagation through a homogeneous medium*** is concerned, this leads to a perhaps obvious and trivial conclusion that ***light travels in straight lines***. In fact, this describes a specific application of Fermat's principal, known as ***Hero's principle***, namely that light follows the path of minimum distance between two points within a homogeneous medium.

1.4.2 Refraction at a Plane Surface and Snell's Law

The law governing refraction at a planar surface is universally attributed to Willebrord Snellius and referred to as ***Snell's law***. This states that both incident and refracted rays lie in the same plane and their angles of incidence and refraction (*with respect to surface normal*) are given by:

$$\text{n}_1 \text{ and n}_2 \text{ are the refractive indices of the two media}: \quad n_1 \sin\theta_1 = n_2 \sin\theta_2. \tag{1.12}$$

This is illustrated in Figure 1.12.

The refractive indices of some optical materials (at 550 nm) are listed below:

Glass (BK7): 1.52
Plastic (Acrylic): 1.48
Water: 1.33
Air: 1.00027

Snell's law is, in fact, a direct consequence of Fermat's principle. The reader may wish to derive this through the application of differential calculus. In finding the optimum path from a point in one medium to a point in another medium, the ray will attempt, as far as possible, to minimise its path through the higher index medium. Snell's law thus represents the minimum optical path condition in this instance. Where the ray passes from a high index material to a low index material, there exists an angle of incidence where the angle of refraction

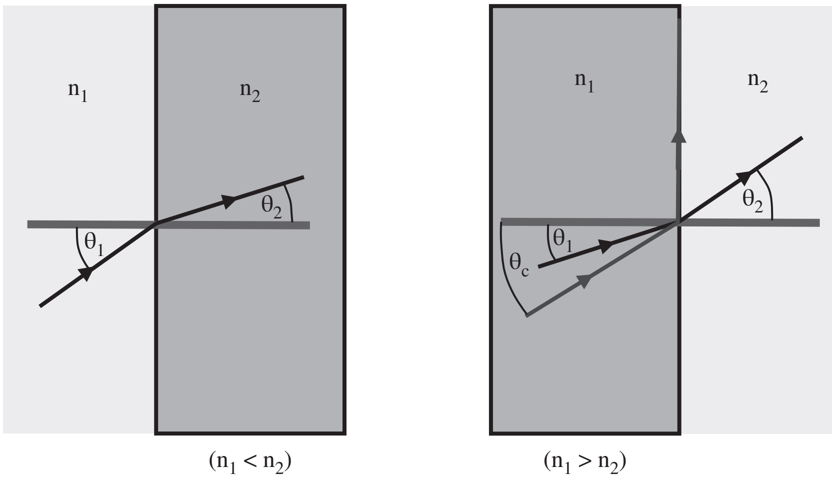

Figure 1.12 Refraction at a plane surface.

is 90°. This angle is known as the ***critical angle*** and, for angles of incidence beyond this, the ray is ***totally internally reflected***. The critical angle, θ_c, is given by:

$$\mathrm{n}_2 < \mathrm{n}_1 \quad n_1 \sin\theta_c = \frac{n_2}{n_1} \tag{1.13}$$

A single refractive surface is an example of an ***afocal system***, where both focal lengths are infinite. Although it does not bring a parallel beam of light to a focus, it does form an image that is a geometrically true representation of the object.

1.4.3 Refraction at a Curved (Spherical) Surface

Most, if not all, curved optical surfaces are at least approximately spherical and are widely employed in the fabrication of lens components. Figure 1.13 illustrates refraction at a spherical surface.

As before, the special case of refraction at a spherical surface may be described by Snell's law:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

If we now wish to calculate the angle ϕ in terms of θ, this process is, in principle, straightforward. We need also to take into account the angle the surface normal makes with the optical axis, Δ, and the radius

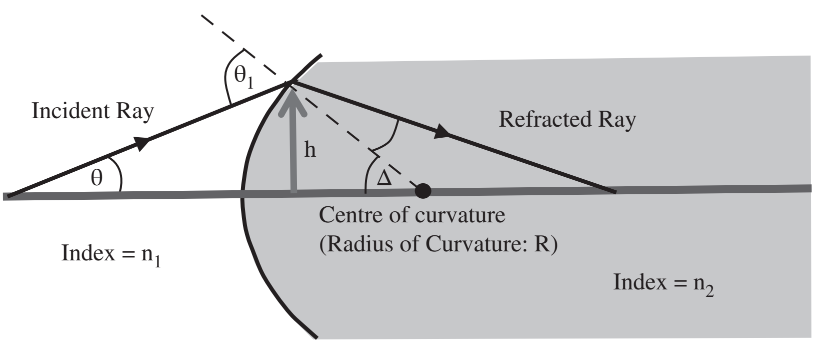

Figure 1.13 Refraction at a spherical surface.

of curvature, R, of the spherical surface. However, calculation is a little unwieldy, so therefore we make the simplifying assumption that all angles are small and:

$$\sin\theta \approx \theta$$

Hence:

$$\theta_2 \approx \frac{n_1}{n_2}\theta_1; \quad \Delta \approx \frac{h}{R} \text{ and } \theta_1 = \theta + \Delta; \quad \varphi = \theta_2 - \Delta$$

We can finally calculate ϕ in terms of θ:

$$\phi \approx \frac{n_1}{n_2}\theta - \left(\frac{n_2 - n_1}{n_2}\right)\frac{h}{R} \tag{1.14}$$

There are two terms on the RHS of Eq. (1.14). The first term, depending on the input angle θ is of the same form as Snell's law (for small angles) for a plane surface. The second term, which gives an angular deflection proportional to the height, h, and inversely proportional to the radius of curvature R, provides a focusing effect. That is to say, rays further from the optic axis are bent inward to a greater extent and have a tendency to converge on a common point. The sign convention used here assumes that positive height is vertically upward, as displayed in Figure 1.13 and a positive spherical radius corresponds to a scenario in which the centre of the sphere lies to the right of the point where the surface intersects the optical axis. Finally, a positive angle is consistent with an increase in ray height as it propagates from left to right in (1.13).

Equation (1.14) can be used to trace any ray that is incident upon a spherical refractive surface. If this surface is deemed to comprise 'the optical system' in its entirety, then one can use Eq. (1.14) to calculate the location of all Cardinal Points, expressed as a displacement, z along the optical axis. Positive z is to the right and the origin lies at the intersection of the optical axis and the surface. The Cardinal points are listed below.

Cardinal points for a spherical refractive surface

First Focal Point: $z = -\left(\frac{n_1}{n_2 - n_1}\right)R$ *First Focal Length*: $\left(\frac{n_1}{n_2 - n_1}\right)R$

Second Focal Point: $z = \left(\frac{n_2}{n_2 - n_1}\right)R$ *Second Focal Length*: $\left(\frac{n_2}{n_2 - n_1}\right)R$

Both Principal Points: $z = 0$

Both Nodal Points: $z = R$

In this instance, the two focal lengths, f_1 and f_2 are different since the object and image spaces are in different media. If we take the first focal length as the distance from the first focal point to the first principal point, then the first focal length is positive. Similarly, the second focal length, the distance from the second principal point to the second focal point, is also positive. The principal points are both located at the surface vertex and the nodal points at the centre of curvature of the sphere. It is important to note that, in this instance, the principal and nodal points do not coincide. Again, this is because the refractive indices of object and image space differ.

1.4.4 Refraction at Two Spherical Surfaces (Lenses)

Figure 1.14 shows a lens made up of two spherical surfaces, of radius, R_1 and R_2. Once again, the convention is that the spherical radius is positive if the centre of curvature lies to the right of the relevant vertex.

So, in the biconvex lens illustrated in Figure 1.14, the first surface has a positive radius of curvature and the second surface has a negative radius of curvature. The lens is made from a material of refractive index n_2 and is bounded by two surfaces with radius of curvature R_1 and R_2 respectively. It is immersed totally in a medium of refractive index, n_1 (e.g. air). In addition, it is assumed that the lens has negligible thickness (***the thin lens approximation***). Of course, as for the treatment of the single curved surface, we assume all angles are small

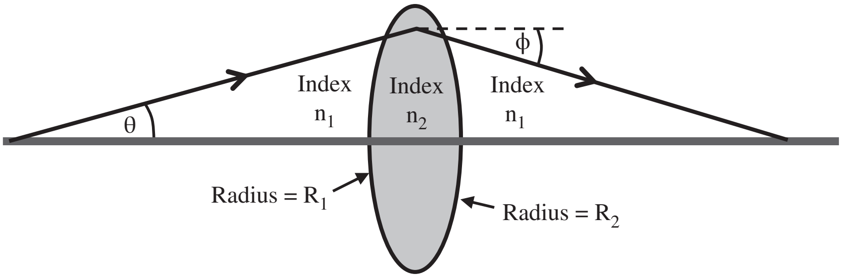

Figure 1.14 Refraction by two spherical surfaces (lens).

and $\theta \sim \sin\theta$. First, we might calculate the angle of refraction, ϕ_1, produced by the first curved surface, R_1. This can be calculated using Eq. (1.14):

$$\phi_1 \approx \frac{n_1}{n_2}\theta - \left(\frac{n_2 - n_1}{n_2}\right)\frac{h}{R_1}$$

Of course, the final angle, ϕ, can be calculated from ϕ_1 by another application of Eq. (1.14):

$$\phi \approx \frac{n_2}{n_1}\phi_1 - \left(\frac{n_1 - n_2}{n_1}\right)\frac{h}{R_2}$$

Substituting for ϕ_1 we get:

$$\phi \approx \theta - \left(\frac{n_2 - n_1}{n_1}\right)\left[\frac{h}{R_1} - \frac{h}{R_2}\right] \tag{1.15}$$

As for Eq. (1.14) there are two parts to Eq. (1.15). First, there is an angular term that is equal to the incident angle. Second, there is a focusing contribution that produces a deflection proportional to ray height. Equation (1.15) allows the tracing of all rays in a system containing the single lens and it is straightforward to calculate the Cardinal points of the thin lens:

Cardinal points for a thin lens

First Focal Point:	$z = -\left(\frac{n_1}{n_2 - n_1}\right)\frac{R_1R_2}{R_1 - R_2}$	*First Focal Length*:	$\left(\frac{n_1}{n_2 - n_1}\right)\frac{R_1R_2}{R_1 - R_2}$
Second Focal Point:	$z = \left(\frac{n_1}{n_2 - n_1}\right)\frac{R_1R_2}{R_1 - R_2}$	*Second Focal Length*:	$\left(\frac{n_1}{n_2 - n_1}\right)\frac{R_1R_2}{R_1 - R_2}$

Both Principal Points: At centre of lens

Both Nodal Points: At centre of lens

Since both object and image spaces are in the same media, then both focal lengths are equal and the principal and nodal points are co-located. One can take the above expressions for focal length and cast it in a more conventional form as a ***single*** focal length, f. This gives the so-called ***Lensmaker's Equation***, where it is assumed that the surrounding medium (air) has a refractive index of one (i.e. $n_1 = 1$) and we substitute n for n_2.

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2}\right] \tag{1.16}$$

1.4.5 Reflection by a Plane Surface

Figure 1.15 shows the process of reflection at a plane surface. As in the previous case of refraction, the reflected ray lies in the same plane as the incident ray and the angle of reflection is equal and opposite to the angle of incidence.

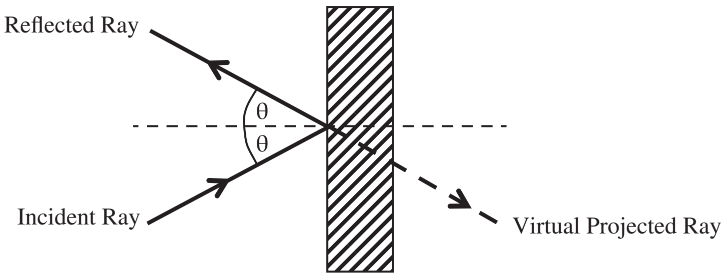

Figure 1.15 Reflection at a plane surface.

The virtual projected ray shown in Figure 1.15 illustrates an important point about reflection. If one considers the process as analogous to refraction, then a mirror behaves as a refractive material with an index of −1. This, in itself has an important consequence. The image produced is inverted in space. As such, there is no combination of positive magnification and pure rotation that will map the image onto the object. That is to say, a right handed object will be converted into a left handed image. More generally, if an optical system contains an ***odd number of reflective elements***, the parity of the image will be reversed. So, for example, if a complex optical system were to contain nine reflective elements in the optical path, then the resultant image could not be generated from the object by rotation alone. Conversely, if the optical system were to contain an even number of reflective surfaces, then the parity between the object and image geometries would be conserved.

Another way in which a plane mirror is different from a plane refractive surface is that a plane mirror is the one (and perhaps only) example of a ***perfect imaging system***. Regardless of any approximation with regard to small angles discussed previously, following reflection at a planar surface, ***all rays*** diverging from a single image point would, when projected as in Figure 1.15, be seen to emerge ***exactly*** from a ***single*** object point.

1.4.6 Reflection from a Curved (Spherical) Surface

Figure 1.16 illustrates the reflection of a ray from a curved surface.

The incident ray is at an angle, θ, with respect to the optical axis and the reflected ray is at an angle, φ to the optical axis. If we designate the incident angle as θ_1 and the reflected angle as θ_2 (with respect to the local surface normal), then the following apply, ***assuming all relevant angles are small***:

$$\theta_1 = \theta + \Delta; \quad \theta_2 = -\theta_1; \quad \varphi = \theta_2 - \Delta \text{ and } \Delta \approx \frac{h}{R}$$

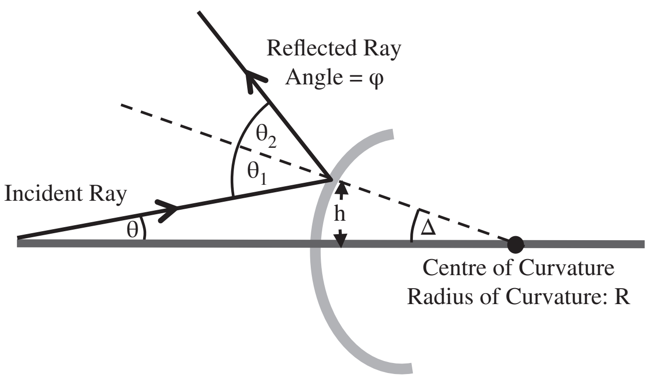

Figure 1.16 Reflection from a curved surface.

We now need to calculate the angle, φ, the refracted ray makes to the optical axis:

$$\phi = -\theta - \frac{2h}{R} \tag{1.17}$$

In form, Eq. (1.17) is similar to Eq. (1.14) with a linear dependence of the reflected ray angle on both incident ray angle and height. The two equations may be made to correspond exactly if we make the substitution, $n_1 = 1$, $n_2 = -1$. This runs in accord with the empirical observation made previously that a reflective surface acts like a medium with a refractive index of −1. Once more, the sign convention observed dictates that positive axial displacement, z, is in the direction from left to right and positive height is vertically ***upwards***. A ray with a positive angle, θ, has a positive gradient in h with respect to z.

As with the curved refractive surface, a curved mirror is image forming. It is therefore possible to set out the Cardinal Points, as before:

Cardinal points for a spherical mirror

First Focal Point: $z = \frac{R}{2}$ *First Focal Length*: $-\frac{R}{2}$

Second Focal Point: $z = \frac{R}{2}$ *Second Focal Length*: $\frac{R}{2}$

Both Principal Points: At vertex

Both Nodal Points: At centre of sphere

The focal length of a curved mirror is half the base radius, with both focal points co-located. In fact, the two focal lengths are of opposite sign. Again, this fits in with the notion that reflective surfaces act as media with a refractive index of −1. Both nodal points are co-located at the centre of curvature and the principal points are also co-located at the surface vertex.

1.5 Paraxial Approximation and Gaussian Optics

Earlier, in order to make our lens and mirror calculations simple and tractable, we introduced the following approximation:

$$\theta << 1 \text{ and } \theta \approx \sin\theta \approx \tan\theta$$

That is to say, all rays make a sufficiently small angle to the optical axis to make the above approximation acceptable in practice. When this approximation is applied more generally to an entire optical system, it is referred to as the ***paraxial approximation*** (i.e. *'almost axial'*). If the same consideration is applied to ray heights as well as angles, the paraxial approximations lead to a series of equations describing the transformation of ray heights and angles that are linear in both ray height and angle. This first order theory is generally referred to as ***Gaussian optics***, named after Carl Friedrich Gauss.

If we now assume that all rays are confined to a single plane containing the optical axis, then we can describe all rays by two parameters: θ – the angle the ray make to the optical axis and h – the height above the optical axis. If, after transformation by an optical surface, these parameters change to θ' and h′, it is possible to write down a series of linear equations describing all transformations. These are set out in Eqs. 1.18–1.21:

Refraction at a plane surface: $$\theta' = \frac{n_1}{n_2}\theta \qquad (h' = h) \tag{1.18}$$

Refraction at a curved surface: $$\theta' = \frac{n_1}{n_2}\theta - \left(\frac{n_2 - n_1}{n_2}\right)\frac{h}{R} \qquad (h' = h) \tag{1.19}$$

Reflection at plane mirror: $$\theta' = -\theta \qquad (h' = h) \tag{1.20}$$

Reflection at curved surface: $$\theta' = -\theta - \frac{2h}{R} \qquad (h' = h) \tag{1.21}$$

Even the most complex optical system may be described as a combination of all the above elements. At first sight, therefore, it would seem that this provides a complete description of the first order behaviour of an optical system. However, there is one important, but seemingly trivial, aspect that is not considered here. This is the case of ray propagation through space. The equations are, of course simple and obvious, but we include them for completeness.

$$\textbf{\textit{Propagation through space}}\text{:} \qquad \theta' = \theta \qquad h' = h + d\theta \qquad (d \text{ is propagation distance}) \tag{1.22}$$

Equation (1.8) introduced the Helmholtz equation, a necessary condition for perfect image formation for an ideal system. It is clear that Gaussian optics represents a mere approximation to the ideal of the Helmholtz equation. The contradiction between the two suggests that there may be imperfections in the ideal treatment of Gaussian optics. This will be considered later when we will look at optical imperfections or aberrations. In the meantime, we will consider a very powerful realisation of Gaussian optics that takes the basic linear equations previously set out and expresses them in terms of matrix algebra. This is the so-called Matrix Ray Tracing technique.

1.6 Matrix Ray Tracing

1.6.1 General

In Section 1.4 we looked at the behaviour of some very simple components, mirrors and lenses, deriving the locations of the Cardinal Points. As discussed previously, the Cardinal Points provide a complete description of the first order properties of an optical system, ***no matter how complex***.

The question then is how do we calculate the properties of a more complex optical system, such as the camera lens depicted in Figure 1.17? It is not immediately obvious where the Cardinal points lie or what the focal length is. However, we can combine the generalised description of an optical system with the treatment of Gaussian optics to produce a model that describes the entire system as a black box acting on rays with a simple linear transformation. The black box may be visualised as below in Figure 1.18.

Following the basic premise of Gaussian optics, we can relate the input and output rays using a set of linear equations:

$$h_{out} = Ah_{in} + B\theta_{in} \tag{1.23}$$

$$\theta_{out} = Ch_{in} + D\theta_{in} \tag{1.24}$$

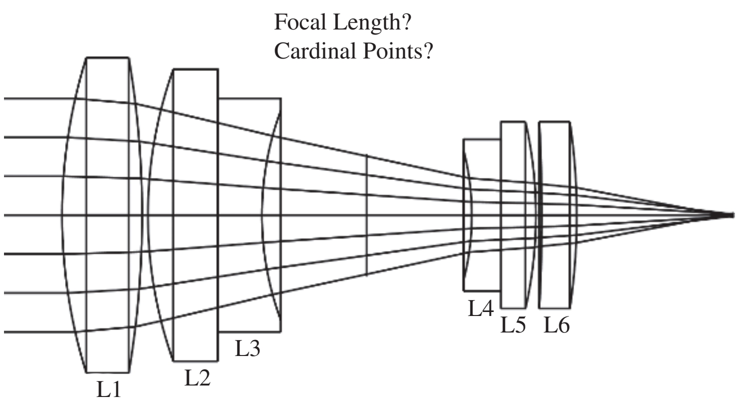

Figure 1.17 Complex optical system.

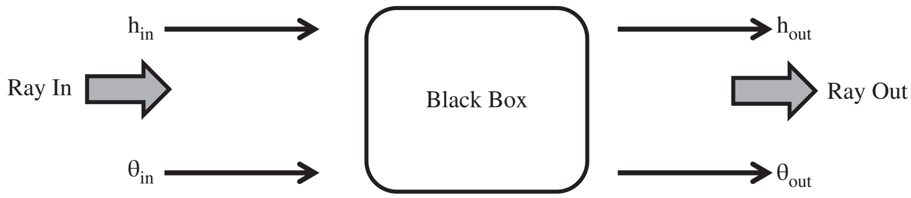

Figure 1.18 Modelling of complex systems.

Equations (1.23) and (1.24) may be combined in a matrix representation:

$$\begin{bmatrix} h_{out} \\ \theta_{out} \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} h_{in} \\ \theta_{in} \end{bmatrix} \tag{1.25}$$

Equation (1.25) sets out the ***Matrix Ray Tracing*** convention used in this book. The reader should be aware that other conventions are used, but this is the most widely used. Equation (1.25) can be used to describe the overall system matrix or that of individual components. The question is how to build up a complex system from a large number of optical elements. The camera lens shown in Figure 1.17 has six lenses and we might represent each lens as a single matrix, i.e. $\mathbf{M_1}, \mathbf{M_2}, \ldots.., \mathbf{M_6}$. Each matrix describes the relationship between rays incident upon the lens and those leaving. The impact of successive optical elements is determined by successive matrix multiplication. So the system matrix for the lens as a whole is given by the matrix product of all elements:

$$\mathbf{M_{system}} = \mathbf{M_6} \times \mathbf{M_5} \times \mathbf{M_4} \times \mathbf{M_3} \times \mathbf{M_2} \times \mathbf{M_1} \tag{1.26}$$

Note the order of the multiplication; this is important. M_1 represents the first optical element seen by rays incident upon the system and the multiplication procedure then works through elements 2–6 successively. For purposes of illustration, each lens has been treated as being represented by a single matrix element. In practice, it is likely that the lens would be reduced to its basic building blocks, namely the two curved surfaces ***plus the propagation (thickness) between the two surfaces.*** We also must not forget the propagation through the air between the lens elements.

Representation of the key optical surfaces can be determined by casting Eqs. (1.18)–(1.22) in matrix format.

$$\textbf{Refraction at a plane surface}: \begin{bmatrix} 1 & 0 \\ 0 & n_1/n_2 \end{bmatrix} \tag{1.27a}$$

$$\textbf{Refraction at a curved surface (Radius R)}: \begin{bmatrix} 1 & 0 \\ \left[\dfrac{n_1 - n_2}{n_2}\right]\dfrac{1}{R} & \frac{n_1}{n_2} \end{bmatrix} \tag{1.27b}$$

$$\textbf{Reflection by plane mirror}: \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \tag{1.27c}$$

$$\textbf{Reflection by curved mirror (Radius R)}: \begin{bmatrix} 1 & 0 \\ -\frac{2}{R} & -1 \end{bmatrix} \tag{1.27d}$$

$$\textbf{Propagation through space (distance d)}: \begin{bmatrix} 1 & d \\ 0 & 1 \end{bmatrix} \tag{1.27e}$$

$$\textbf{Effect of lens (focal length f)}: \begin{bmatrix} 1 & 0 \\ -1/f & 1 \end{bmatrix} \tag{1.27f}$$

n_1 and n_2 represent the refractive index of first and second media respectively.

1.6.2 Determination of Cardinal Points

It is very straightforward to calculate the Cardinal Points of a system from the system matrix:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

The matrix above represents the system matrix after propagating through all optical elements as shown in Figure 1.17. However, the convention adopted here is that an additional transformation is added after the final surface. This additional transformation is free space propagation to the original starting point. It must be emphasised that, this is merely a ***convention***, and that the final step traces a ***dummy ray*** as opposed to a real ray. That is to say, in reality, the light does not propagate backwards to this point. In fact, this step is a virtual back-projection of the real ray which preserves the original ray geometry. The logic of this, as will be seen, is that in any subsequent analysis, the location of all cardinal points is referenced with respect to a common starting point. If this step were dispensed with, then the three first Cardinal Points would be referenced to the start point and the three second Cardinal Points to the end point. With this in mind, the Cardinal Points, as referenced to the common start point are set out below; the reader might wish to confirm this.

$$\text{First FocalPoint}: \quad z = \frac{D}{C} \qquad \textit{First Focal Length}: -\frac{(AD - BC)}{C}$$

$$\textit{Second Focal Point}: \quad z = -\frac{A}{C} \qquad \textit{Second Focal Length}: -\frac{1}{C}$$

$$\textit{First Principal Point}: \quad z = B + D\frac{(1 - A)}{C} \qquad \textit{Second Principal Point}: \quad z = \frac{(1 - A)}{C}$$

$$\textit{First Nodal Point}: \quad z = \frac{(D - 1)}{C} \qquad \textit{Second Nodal Point}: \quad z = \frac{A(D - 1) - BC}{C}$$

The ***determinant*** of the matrix, (**AD−BC**), is a key parameter. The ratio of the two focal lengths of the system is simply given by the determinant. That is to say the ratio of the two focal lengths is given by:

$$\frac{f_1}{f_2} = (AD - BC) = Determinant \tag{1.28}$$

Inspecting all matrix expressions in Eqs. (1.27a–1.27f), the determinant of the matrix is simply n_1/n_2, the ratio of the indices in the two media, ***for all possible scenarios***. Since the determinant of a matrix product is simply the product of the individual determinants, then the determinant of the overall system matrix is simply the ratio of the refractive indices in image and object space. Thus:

$$\frac{f_1}{f_2} = \frac{n_{image}}{n_{object}} \tag{1.29}$$

This relationship was anticipated in the more generalised discussion in 1.3.9. Looking at the relationships for the principal and nodal points, it is clear when the determinant of the system matrix is unity, i.e. object and image space indices are the same, then the principal and nodal points are co-located.

In addition to the principal and nodal points, ***anti-principal points*** and ***anti-nodal points*** are sometimes (rarely) specified. Anti-principal points are conjugate points where the magnification is −1. Similarly, anti-nodal points are conjugate points where the angular magnification is −1.

1.6.3 Worked Examples

We can now use the foregoing analysis to see how matrix ray tracing might be used in practice. Here we focus on a number of useful practical examples.

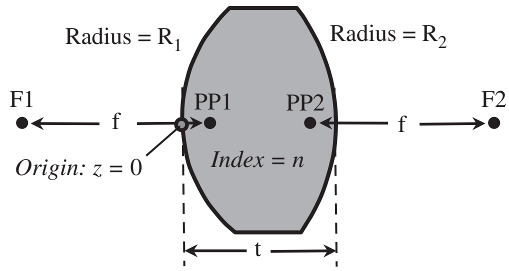

Figure 1.19 Thick lens.

Worked Example 1.1 ***Thick Lens***

The matrix for the system is simply as below – ***note the order***:

$$M = \underbrace{\begin{bmatrix} 1 & -t \\ 0 & 1 \end{bmatrix}}_{\textit{Translate to Origin}} \underbrace{\begin{bmatrix} 1 & 0 \\ \dfrac{(n-1)}{R_2} & 1/n \end{bmatrix}}_{\text{Second Surface}} \underbrace{\begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}}_{\text{Lens Thickness}} \underbrace{\begin{bmatrix} 1 & 0 \\ \dfrac{1-n}{n}\dfrac{1}{R_1} & 1/n \end{bmatrix}}_{\text{First Surface}}$$

We have two translations. The first translation represents the thickness of the lens and the second translation, ***by convention***, traces the refracted rays back to the origin in z. This is so that, in interpreting the formulae for Cardinal points, we can be sure that they are all referenced to a common origin, located as in Figure 1.19. Positive axial displacement (z) is to the right and a positive radius, R, is where the centre of curvature lies to the right of the vertex. The final matrix is as below:

$$M = \begin{bmatrix} \left[1 + t\left((n-1)\left(\dfrac{1}{R_1} - \dfrac{1}{nR_1} - \dfrac{1}{R_2}\right) + \dfrac{t(n-1)^2}{nR_1R_2}\right)\right] & \left[\dfrac{t^2(1-n)}{nR_2}\right] \\ \left[-(n-1)\left(\dfrac{1}{R_1} - \dfrac{1}{R_2}\right) - \dfrac{t(n-1)^2}{nR_1R_2}\right] & \left[1 + \dfrac{t(n-1)}{nR_2}\right] \end{bmatrix}$$

As both object and image space are in the same media, there is a common focal length, f, i.e. $f_1 = f_2 = f$. All relevant parameters are calculated from the above matrix using the formulae tabulated in Section 1.6.2.

The focal length, f, is given by:

$$\frac{1}{f} = \underbrace{(n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right)}_{\text{Lensmaker}} + \underbrace{\frac{t(n-1)^2}{nR_1R_2}}_{\text{'Thickness Term'}}$$

The formula above is similar to the simple, 'Lensmaker' formula for a thin lens. In addition there is another term, linear in thickness, t, which accounts for the lens thickness.

The focal positions are as follows:

$$F_1 = -f\left(1 + \frac{t(n-1)}{nR_2}\right) \qquad F_2 = f + t - \frac{(n-1)}{nR_1}tf$$

The principal points are as follows:

$$P_1 = -f\frac{t(n-1)}{nR_2} \qquad P_2 = t - \frac{(n-1)}{nR_1}tf$$

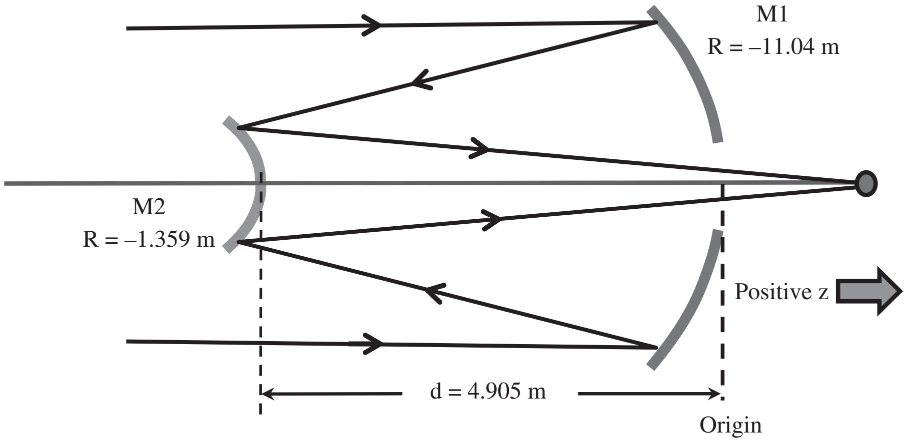

Figure 1.20 Hubble space telescope schematic.

Of course, since the refractive indices of the object and image spaces are identical, the nodal points are located in the same place as the principal points. If we take the example of a biconvex lens where $R_2 = -R_1$, then:

$$P_1 = \left[\frac{t}{2n}\right]\left[\frac{1}{1 + (t/2n(n-1)R_1)}\right]$$

So, for a biconvex lens with a refractive index of (1.5), then the principal points lie about one third of the thickness from their respective vertices.

Worked Example 1.2 ***Hubble Space Telescope***

The telescope part of the Hubble Space Telescope instrument is made up of two mirrors, a primary and a secondary. Characteristics of the telescope are shown in Figure 1.20. *Data is courtesy of the National Aeronautics and Space Administration.*

There are four matrix elements to consider here. First, there is a mirror with a radius of −11.04 m (note sign), followed by a translation of −4.905 m (again note sign). The third matrix element is a mirror (M2) of radius − 1.359 m. Finally, we translate by +4.905 m, so that both the input and output co-ordinates are referenced with respect to the same origin. The matrices are as below:

$$M = \begin{bmatrix} 1 & 4.905 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{-1.359} & -1 \end{bmatrix} \begin{bmatrix} 1 & -4.905 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\dfrac{2}{-11.04} & -1 \end{bmatrix} = \begin{bmatrix} .0271 & 45.217 \\ -.0172 & 8.219 \end{bmatrix}$$

The focal positions are:

F1: −477.93 m **F2:** 1.574 m

The principal points are at:

P1: −419.777 m **P2:** −56.579 m

$f_1 = f_2 = 58.153$ m

Since object and image space are in the same media, then the two focal lengths are the same. In addition, the nodal and principal points are co-located. However, when dealing with mirrors, one must be a little cautious. Each reflection is equivalent to a medium with a refractive index of −1, so that the matrix of a reflective surface will always have a determinant of −1. Therefore, any system having an ***even number*** of reflective surfaces, as in this example, then its matrix will have a determinant of 1. As such, the two focal lengths will be the same and principal and nodal points co-located. However, where there are an ***odd number*** of reflective surfaces, assuming object and image spaces are surrounded by the same media, then $f_2 = -f_1$. In this instance, principal and nodal points are separated by twice the focal length.

Although, in terms of overall length, the telescope is compact, ~5 m primary–secondary separation, the focal length, at 58 m, is long. The focal length of the instrument is fundamental in determining the 'plate scale' the separation of imaged objects (stars, galaxies) at the (second) focal plane as a function of their ***angular*** separation. As such, a long focal length, of the order of 60 m, may have been a ***requirement*** at the outset. At the same time, for practical reasons, a compact design may also have been desired. One may begin to glance, therefore, at the significance, at the very outset of these very basic calculations in the design of complex optical instruments.

1.6.4 Spreadsheet Analysis

For the examples previously set out, matrix multiplication is a quick and convenient method for calculating the first order parameters of an optical system. Nonetheless, it must be recognised that as systems become more complex, with more optical surfaces, these calculations can become quite tedious. However, these matrix calculations are easy to embed with spreadsheet tools enabling the automatic computation of all cardinal points. By way of example, the previous calculation is set out and automated using a simple spreadsheet tool.

Translate		2nd Mirror		Translate		1st Mirror			
d	**4.905**	**R**	**–1.359**	**d**	**–4.905**	**R**	**–11.04**		
1	4.905	1	0	1	–4.905	1	0		
0	1	1.472	–1	0	1	0.181	–1		
0.027	45.217	0.111	4.905	0.111	4.905	1	0	1	0
–0.017	8.219	–0.017	8.219	0.181	–1	0.181	–1	0	1
1st Focal Point		–477.93	m		**2nd Focal Point**	1.574	m		
1st Principal Point		–419.78	m		**2nd Principal Point**	–56.579	m		
1st Nodal Point		–419.78	m		**2nd Nodal Point**	–56.579	m		
Focal Length 1		58.153	m		**Focal Length 2**	58.153	m		

In the exercises that follow, the reader may choose to use this method to simplify calculations.

Further Reading

Born, M. and Wolf, E. (1999). *Principles of Optics*, 7e. Cambridge: Cambridge University Press. ISBN: 0-521-642221.

Haija, A.I., Numan, M.Z., and Freeman, W.L. (2018). *Concise Optics: Concepts, Examples and Problems*. Boca Raton: CRC Press. ISBN: 978-1-1381-0702-1.

Hecht, E. (2017). *Optics*, 5e. Harlow: Pearson Education. ISBN: 978-0-1339-7722-6.

Keating, M.P. (1988). *Geometric, Physical, and Visual Optics*. Boston: Butterworths. ISBN: 978-0-7506-7262-7.

Kidger, M.J. (2001). *Fundamental Optical Design*. Bellingham: SPIE. ISBN: 0-81943915-0.

Kloos, G. (2007). *Matrix Methods for Optical Layout*. Bellingham: SPIE. ISBN: 978-0-8194-6780-5.

Longhurst, R.S. (1973). *Geometrical and Physical Optics*, 3e. London: Longmans. ISBN: 0-582-44099-8.

Riedl, M.J. (2009). *Optical Design: Applying the Fundamentals*. Bellingham: SPIE. ISBN: 978-0-8194-7799-6.
Saleh, B.E.A. and Teich, M.C. (2007). *Fundamentals of Photonics*, 2e. New York: Wiley. ISBN: 978-0-471-35832-9.
Smith, F.G. and Thompson, J.H. (1989). *Optics*, 2e. New York: Wiley. ISBN: 0-471-91538-1.
Smith, W.J. (2007). *Modern Optical Engineering*. Bellingham: SPIE. ISBN: 978-0-8194-7096-6.
Walker, B.H. (2009). *Optical Engineering Fundamentals*, 2e. Bellingham: SPIE. ISBN: 978-0-8194-7540-4.