

Chapter 1

Geometrical Optics

When we consider optics, the first thing that comes to our minds is probably light. Light has a dual nature: light is particles (called photons) and light is waves. When a particle moves, it processes momentum, p . And when a wave propagates, it oscillates with a wavelength, λ . Indeed, the momentum and the wavelength is given by the *de Broglie relation*

$$\lambda = \frac{h}{p},$$

where $h \approx 6.62 \times 10^{-34}$ Joule-second is Planck's constant. Hence from the relation, we can state that every particle is a wave as well.

Each particle or photon is specified precisely by the frequency ν and has an energy E given by

$$E = h\nu.$$

If the particle is traveling in free space or in vacuum, $\nu = c/\lambda$, where c is a constant approximately given by 3×10^8 m/s. The speed of light in a transparent linear, homogeneous and isotropic material, which we term v , is again a constant but less than c . This constant is a physical characteristic or signature of the material. The ratio c/v is called the *refractive index*, n , of the material.

In *geometrical optics*, we treat light as particles and the trajectory of these particles follows along paths that we call *rays*. We can describe an optical system consisting of elements such as mirrors and lenses by tracing the rays through the system.

Geometrical optics is a special case of *wave* or *physical* optics, which will be mainly our focus through the rest of this Chapter. Indeed, by taking the limit in which the wavelength of light approaches zero in wave optics, we recover geometrical optics. In this limit, diffraction and the wave nature of light is absent.

1.1 Fermat's Principle

Geometrical optics starts from *Fermat's Principle*. In fact, Fermat's Principle is a concise statement that contains all the physical laws, such as the *law of reflection* and *the law of refraction*, in geometrical optics. Fermat's principle states that the path of a light ray follows is an extremum in comparison with the nearby paths. The extremum may be a minimum, a maximum, or stationary with respect to variations in the ray path. However, it is usually a minimum.

We now give a mathematical description of Fermat's principle. Let $n(x, y, z)$ represent a position-dependent refractive index along a path C between end points A and B , as shown in Fig. 1.1. We define the *optical path length (OPL)* as

$$OPL = \int_C n(x, y, z) ds, \quad (1.1-1)$$

where ds represents an infinitesimal arc length. According to Fermat's principle, out the many paths that connect the two end points A and B , the light ray would follow that path for which the OPL between the two points is an extremum, i.e.,

$$\delta(OPL) = \delta \int_C n(x, y, z) ds = 0 \quad (1.1-2)$$

in which δ represents a small variation. In other words, a ray of light will travel along a medium in such a way that the total OPL assumes an extremum. As an extremum means that the rate of change is zero, Eq. (1.1-2) explicitly means that

$$\frac{\partial}{\partial x} \int n ds + \frac{\partial}{\partial y} \int n ds + \frac{\partial}{\partial z} \int n ds = 0. \quad (1.1-3)$$

Now since the ray propagates with the velocity $v = c/n$ along the path,

$$nds = \frac{c}{v} ds = c dt, \quad (1.1-4)$$

where dt is the differential time needed to travel the distance ds along

the path. We substitute Eq. (1.1-4) into Eq. (1.1-2) to get

$$\delta \int_C n ds = c \delta \int_C dt = 0. \quad (1.1-5)$$

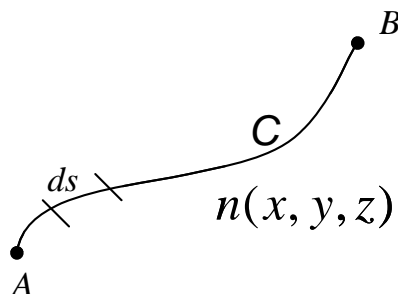


Fig. 1.1 A ray of light traversing a path C between end points A and B .

As mentioned before, the extremum is usually a minimum, we can, therefore, restate Fermat's principle as a *principle of least time*. In a *homogeneous medium*, i.e., in a medium with a constant refractive index, the ray path is a straight line as the shortest *OPL* between the two end points is along a straight line which assumes the shortest time for the ray to travel.

1.2 Reflection and Refraction

When a ray of light is incident on the interface separating two different optical media characterized by n_1 and n_2 , as shown in Fig. 1.2, it is well known that part of the light is reflected back into the first medium, while the rest of the light is refracted as it enters the second medium. The directions taken by these rays are described by the laws of reflection and refraction, which can be derived from Fermat's principle.

In what follows, we demonstrate the use of the principle of least time to derive the law of refraction. Consider a reflecting surface as shown in Fig. 1.3. Light from point A is reflected from the reflecting surface to point B , forming the angle of incidence ϕ_i and the angle of reflection ϕ_r , measured from the normal to the surface. The time required for the ray of light to travel the path $AO + OB$ is given by $t = (AO + OB)/v$, where v is the velocity of light in the medium

containing the points AOB. The medium is considered isotropic and homogeneous. From the geometry, we find

$$t(z) = \frac{1}{v} ([h_1^2 + (d - z)^2]^{1/2} + [h_2^2 + z^2]^{1/2}). \quad (1.2-1)$$

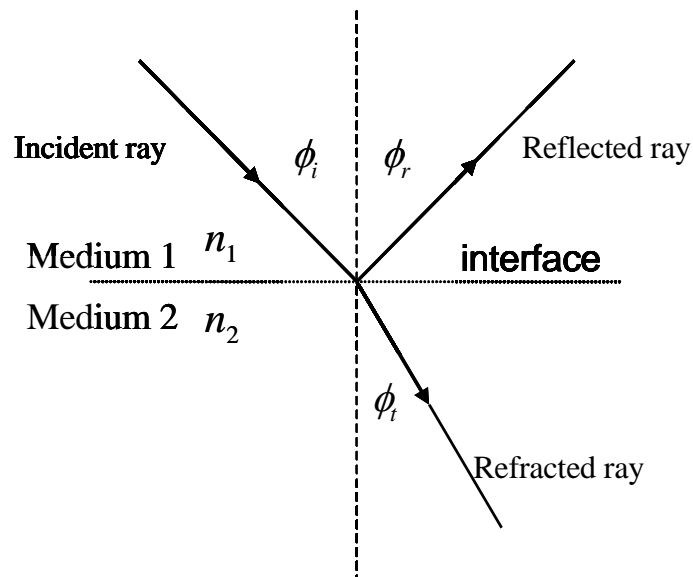


Fig. 1.2 Reflected and refracted rays for light incident at the interface of two media.

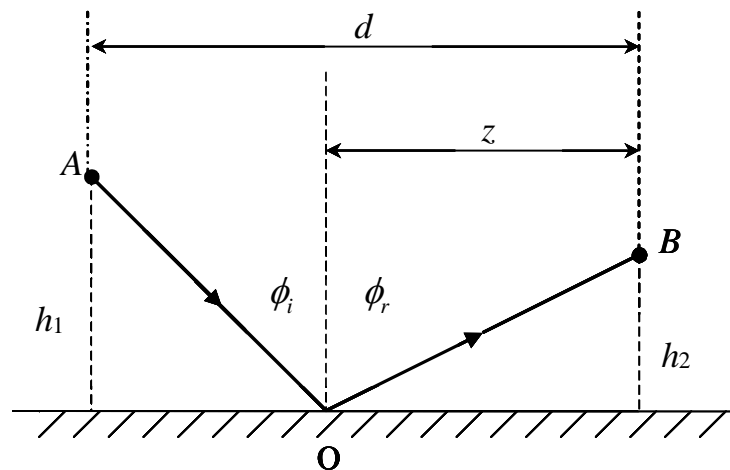


Fig. 1.3 Incident (AO) and reflected (OB) rays.

According to the least time principle, light will find a path that extremizes $t(z)$ with respect to variations in z . We thus set $dt(z)/dz = 0$ to get

$$\frac{d - z}{[h_1^2 + (d - z)^2]^{1/2}} = \frac{z}{[h_2^2 + z^2]^{1/2}} \quad (1.2-2)$$

or

$$\sin \phi_i = \sin \phi_r \quad (1.2-3)$$

so that

$$\phi_i = \phi_r, \quad (1.2-4)$$

which is the law of reflection. We can readily check that the second derivative of $t(z)$ is positive so that the result obtained corresponds to the least time principle. In addition, Fermat's principle also demands that the incident ray, the reflected ray and the normal all be in the same plane, called the *plane of incidence*.

Similarly, we can use the least time principle to derive the law of refraction

$$n_1 \sin \phi_i = n_2 \sin \phi_t, \quad (1.2-5)$$

which is commonly known as *Snell's law of refraction*. In Eq. (1.2-5), ϕ_i is the angle of incidence for the incident ray and ϕ_t is the angle of transmission (or angle of refraction) for the refracted ray. Both angles are measured from the normal to the surface. Again, as in reflection, the incident ray, the refracted ray, and the normal all lie in the same plane of incidence. Snell's law shows that when a light ray passes obliquely from a medium of smaller refractive index n_1 into one that has a larger refractive index n_2 , or an optically denser medium, it is bent toward the normal. Conversely, if the ray of light travels into a medium with a lower refractive index, it is bent away from the normal. For the latter case, it is possible to visualize a situation where the refracted ray is bent away from the normal by exactly 90° . Under this situation, the angle of incidence is called the *critical angle* ϕ_c , given by

$$\sin \phi_c = n_2/n_1. \quad (1.2-6)$$

When the incident angle is greater than the critical angle, the ray

originating in medium 1 is totally reflected back into medium 1. This phenomenon is called *total internal reflection*. The optical fiber uses this principle of total reflection to guide light, and the mirage on a hot summer day is a phenomenon due to the same principle.

1.3 Ray Propagation in an Inhomogeneous Medium: Ray Equation

In the last Section, we have discussed refraction between two media with different refractive indices, possessing a discrete inhomogeneity in the simplest case. For a general inhomogeneous medium, i.e., $n(x, y, z)$, it is instructive to have an equation that can describe the trajectory of a ray. Such an equation is known as the *ray equation*. The ray equation is analogous to the equations of motion for particles and for rigid bodies in classical mechanics. The equations of motion can be derived from *Newtonian mechanics* based on Newton's laws. Alternatively, the equations of motion can be derived directly from *Hamilton's principle of least action*. Indeed Fermat's principle in optics and Hamilton's principle of least action in classical mechanics are analogous. In what follows, we describe Hamilton's principle so as to formulate the so called *Lagrange's equations* in mechanics. We then re-formulate Lagrange's equations for optics to derive the ray equation.

Hamilton's principle states that the trajectory of a particle between times t_1 and t_2 is such that the variation of the line integral for fixed t_1 and t_2 is zero, i.e.,

$$\delta \int_{t_1}^{t_2} L(q_k, \dot{q}_k, t) dt = 0, \quad (1.3-1)$$

where $L = T - V$ is known as the *Lagrangian function* with T being the kinetic energy and V the potential energy of the particle. The q_k 's are called *generalized coordinates* with $k = 1, 2, 3, \dots, n$. Also, $\dot{q}_k = dq_k/dt$.

Generalized coordinates are any collection of independent coordinates q_k (not connected by any equations of constraint) that are sufficient to specify uniquely the motion. The number n of generalized coordinates is the number of *degrees of freedom*. For example, a simple pendulum has one degree of freedom, i.e., $q_k = q_1 = \phi$, where ϕ is the angle the pendulum makes with the vertical. Now if the simple pendulum is complicated such that the string holding the bob is elastic. There will be two generalized coordinates, $q_k = q_1 = \phi$, and $q_k = q_2 = x$, where x is

the length of the string. As another example, let us consider a particle constrained to move along the surface of a sphere with radius R . The coordinates (x, y, z) do not constitute an independent set as they are connected by the equation of constraint $x^2 + y^2 + z^2 = R^2$. The particle has only two degrees of freedom and two independent coordinates are needed to specify its position on the sphere uniquely. These coordinates could be taken as latitude and longitude or we could choose angles θ and ϕ from spherical coordinates as our generalized coordinates.

Now, if the force field \mathbf{F} is conservative, i.e., $\nabla \times \mathbf{F} = 0$, the total energy $E = T + V$ is a constant during the motion, and Hamilton's principle leads to the following equations of motion of the particle called *Lagrange's equations*:

$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_k} \right) = \frac{\partial L}{\partial q_k}. \quad (1.3-2)$$

As a simple example illustrating the use of Lagrange's equations, let us consider a particle with mass m having kinetic energy $T = \frac{1}{2}m|\dot{\mathbf{r}}|^2$ under potential energy $V(x, y, z)$, where

$$\mathbf{r}(x, y, z) = x(t)\mathbf{a}_x + y(t)\mathbf{a}_y + z(t)\mathbf{a}_z$$

is the position vector with \mathbf{a}_x , \mathbf{a}_y , and \mathbf{a}_z being the unit vector along the x , y , and z direction, respectively. According to Newton's second law,

$$\mathbf{F} = m\ddot{\mathbf{r}}, \quad (1.3-3)$$

where $\ddot{\mathbf{r}}$ is the second derivative of \mathbf{r} with respect to t . As usual the force is given by the negative gradient of the potential, i.e., $\mathbf{F} = -\nabla V$. Hence, we have the vector equation of motion for the particle

$$m\ddot{\mathbf{r}} = -\nabla V \quad (1.3-4)$$

according to Newtonian mechanics. Now from the Lagrange's equations, we identify

$$L = T - V = \frac{1}{2}m|\dot{\mathbf{r}}|^2 - V.$$

Considering $q_1 = x$, we have

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = m\ddot{x} \quad \text{and} \quad \frac{\partial L}{\partial x} = -\frac{\partial V}{\partial x}. \quad (1.3-5)$$

Now, from Eq. (1.3-2) and using the above results, we have

$$m\ddot{x} = -\frac{\partial V}{\partial x}, \quad (1.3-6)$$

and similarly for the y and z components as $q_2 = y$ and $q_3 = z$. Therefore, we come up with Eq. (1.3-4), which is directly from Newtonian mechanics. Hence, we see that Newton's equations can be derived from Lagrange's equations and in fact, the two sets of equations are equally fundamental. However, the Lagrangian formalism has certain advantages over the conventional Newtonian laws in that the physics problem has been transformed into a purely mathematical problem. We just need to find T and V for the system and the rest is just mathematical consideration through the use of Lagrange's equations. In addition, there is no need to consider any vector equations as in Newtonian mechanics as Lagrange's equations are scalar quantities. As it turns out, Lagrange's equations are much better adapted for treating complex systems such as in the areas of quantum mechanics and general relativity.

After having some understanding of Hamilton's principle, and the use of Lagrange's equations to obtain the equations of motion of a particle, we now formulate Lagrange's equations in optics. Again, the particles of concern in optics are photons. Starting from Fermat's principle as given by Eq. (1.1-2),

$$\delta \int_C n(x, y, z) ds = 0. \quad (1.3-7)$$

We write the arc length ds along the path of the ray as

$$ds^2 = dx^2 + dy^2 + dz^2 \quad (1.3-8)$$

with reference to Fig. 1.4, where for brevity, we have only shown the 2-

D (i.e., $x - z$) version of the configuration. Defining $x' = dx/dz$ and $y' = dy/dz$, we can write Eq. (1.3-8) as

$$ds = dz \sqrt{1 + (x')^2 + (y')^2}. \quad (1.3-9)$$

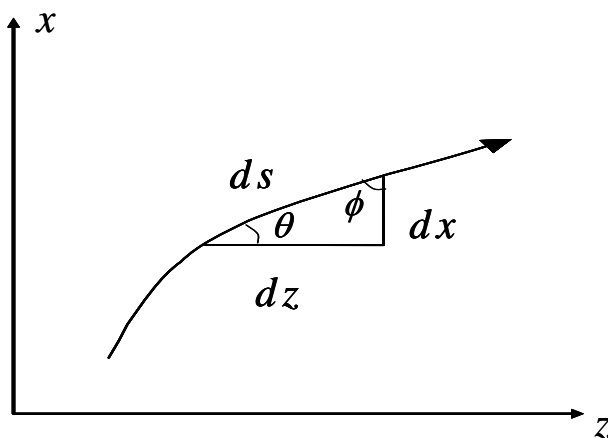


Fig. 1.4 The path of a ray in a continuous Inhomogeneous medium.

Substituting Eq. (1.3-9) into Eq. (1.3-7), we have

$$\delta \int_C n(x, y, z) \sqrt{1 + (x')^2 + (y')^2} dz = 0. \quad (1.3-10)$$

By comparing this equation with Eq. (1.3-1), we can define the so-called *optical Lagrangian* as

$$L(x, y, x', y', z) = n(x, y, z) \sqrt{1 + (x')^2 + (y')^2}. \quad (1.3-11)$$

We can see that Hamilton's principle is based on minimizing functions of time, whereas Fermat's principle minimizes a function of length, z , as we have assumed z to play the same role as t in Lagrangian mechanics, where we have chosen the z -direction as the direction along which the rays are propagating. Now that we have established the optical Lagrangian, we can immediately write down the following Lagrange's equations in optics by referring to Eq. (1.3-2):

$$\frac{d}{dz} \left(\frac{\partial L}{\partial x'} \right) = \frac{\partial L}{\partial x} \quad \text{and} \quad \frac{d}{dz} \left(\frac{\partial L}{\partial y'} \right) = \frac{\partial L}{\partial y}. \quad (1.3-12)$$

From these two equations, we derive the so-called *ray equation*, which tracks the position of the ray (or photon); just like in Lagrangian mechanics, from Lagrange's equations, we derive the equations of motion for a particle.

Using Eqs. (1.3-9) and (1.3-11), Eq. (1.3-12) becomes, after some manipulations,

$$\frac{d}{ds}\left(n\frac{dx}{ds}\right) = \frac{\partial n}{\partial x} \quad \text{and} \quad \frac{d}{ds}\left(n\frac{dy}{ds}\right) = \frac{\partial n}{\partial y}. \quad (1.3-13)$$

The objective is of course, for a given n , we find $x(s)$ and $y(s)$ by solving the above equations. It is important to point out that the two equations above are sufficient to determine the ray trajectory. This indicates that the z -component of the ray equation is really redundant. Indeed the corresponding equation for z , given below, can be derived from the equations for x and y .

$$\frac{d}{ds}\left(n\frac{dz}{ds}\right) = \frac{\partial n}{\partial z}. \quad (1.3-14)$$

Now the desired *ray equation in vectorial form* is obtained by combining Eqs. (1.3-13) and (1.3-14):

$$\frac{d}{ds}\left(n\frac{d\mathbf{r}}{ds}\right) = \nabla n, \quad (1.3-15)$$

where once again $\mathbf{r}(s)$ is a position vector which represents the position of any point on the ray.

Example 1.1 Homogeneous Medium

For $n(x, y, z) = \text{constant}$, Eq. (1.3-15) becomes

$$\frac{d^2\mathbf{r}}{ds^2} = 0, \quad (1.3-16)$$

which has solutions

$$\mathbf{r} = \mathbf{a}s + \mathbf{b} \quad (1.3-17)$$

where \mathbf{a} and \mathbf{b} are some constant vectors determined from the initial conditions, and Eq. (1.3-17) is clearly a straight line equation for the ray path in a homogeneous medium. The situation is shown in Fig. 1.5.

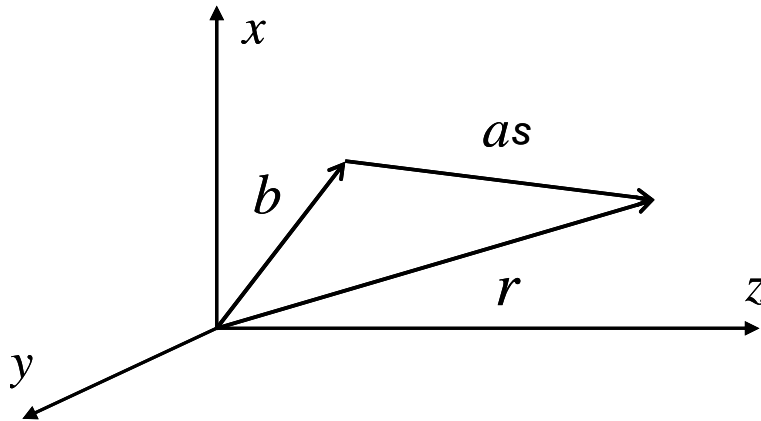


Fig. 1.5 Ray propagating along a straight line in homogeneous medium.

Example 1.2 Law of refraction derived from the ray equation

We consider a 2-D situation involving x and z coordinates where n is a function of x only. The medium consists of a set of thin slices of media of different refractive indices as shown in Fig. 1.6. Since we are interested in how the ray travels along z , we can use Eq. (1.3-14), which becomes

$$\frac{d}{ds} \left(n \frac{dz}{ds} \right) = 0,$$

or

$$n \frac{dz}{ds} = \text{constant}.$$

Since $dz/ds = \cos\theta = \sin\phi$ (see Fig. 1.4), the above equation can be written as

$$n \sin\phi = \text{constant},$$

which holds true throughout the ray trajectory. Hence we have derived the law of refraction, or Snell's law, i.e.,

$$n_1 \sin\phi_1 = n_2 \sin\phi_2 = n_3 \sin\phi_3.$$

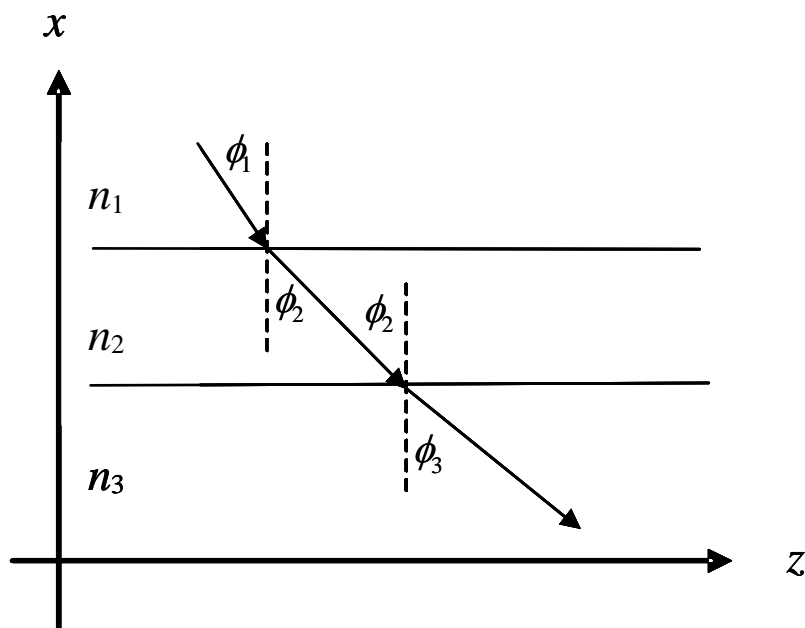


Fig. 1.6 Ray refracted along layers of discrete medium.

Example 1.3 Square-Law Medium: $n^2(x, y) = n_0^2 - n_2(x^2 + y^2)$

In this example, we first consider that an optical waveguide with a z -independent refractive index, i.e., $n = n(x, y)$ in general, and then find a solution to a special case of a *square-law medium* where $n^2(x, y) = n_0^2 - n_2(x^2 + y^2)$. Note that n_2 is considered small enough such that $n_0^2 \gg n_2(x^2 + y^2)$ for all practical values of x and y . For the case that $n(x, y)$ is not a function of z , we can inspect Eq. (1.3-14) to get some insight into the problem:

$$\frac{d}{ds} \left(n \frac{dz}{ds} \right) = \frac{\partial n}{\partial z} = 0,$$

which means that $n \frac{dz}{ds}$ is not a function of s , i.e., it is constant along the ray path. In fact, it is not a function of any coordinates x, y and z as $s(x, y, z)$. Hence $n \frac{dz}{ds}$ is strictly a constant. We let $n \frac{dz}{ds} = \tilde{\beta}$ and refer to

Fig. 1.4 to use the fact that $\frac{dz}{ds} = \cos\theta(x, y)$, and by taking into account the y -dimension for a general situation, we arrive at an equation

$$n(x, y)\cos\theta(x, y) = \tilde{\beta}. \quad (1.3-18)$$

The above equation is *generalized Snell's law* and it means physically that as the ray travels along a trajectory inside the waveguide, the ray would bend in such a way that the product $n(x, y)\cos\theta(x, y)$ or $n(x, y)\sin\phi(x, y)$ remains the same. Now let us find the equations so that we can solve for $x(z)$ and $y(z)$. To find $x(z)$, we can use Lagrange's equation involving x [see Eq. (1.3-12)], i.e.,

$$\frac{d}{dz}\left(\frac{\partial L}{\partial x'}\right) = \frac{\partial L}{\partial x} \quad (1.3-19)$$

with $L = n(x, y)\sqrt{1 + (x')^2 + (y')^2}$ for our current example. We can show that Eq. (1.3-19) becomes

$$\frac{d^2x}{dz^2} = \frac{n}{\tilde{\beta}^2} \frac{\partial n}{\partial x} = \frac{1}{2\tilde{\beta}} \frac{\partial n^2}{\partial x}. \quad (1.3-20)$$

Similarly, we can derive the ray equation for $y(z)$ by using the y -component of Eq. (1.3-12):

$$\frac{d^2y}{dz^2} = \frac{n}{\tilde{\beta}^2} \frac{\partial n}{\partial y} = \frac{1}{2\tilde{\beta}} \frac{\partial n^2}{\partial y}. \quad (1.3-21)$$

The above two equations are rigorous equations for media with the index of refraction independent of z .

We now consider a simple example in the square-law medium and find the ray path for propagation in x - z plane when we launch a ray from $x = x_0$ with a launching angle α with respect to the z -axis. We use Eq. (1.3-20), which becomes

$$\frac{d^2x}{dz^2} = -\frac{n_2}{n^2(x_0)\cos^2\alpha}x(z), \quad (1.3-22)$$

where we have used the definition that

$$\tilde{\beta} = n(x_0)\cos\theta(x_0) = n(x_0)\cos\alpha.$$

Equation (1.3-22) has a general solution of the form given

$$x(z) = A \sin \left(\frac{\sqrt{n_2}}{n(x_0) \cos \alpha} z + \phi_0 \right), \quad (1.3-23)$$

where the constants A and ϕ_0 can be determined from the initial position and slope of the ray. Note that rays with smaller launching angles α have a larger period; however, in the paraxial approximation (i.e., for small launching angles), all the ray paths have approximately the same period. These rays, which lie in the plane containing the so-called *optical axis* (z -axis), are called *meridional rays* and all other rays are called *skew rays*.

Let us now discuss a case in that the ray is launched on the y - z plane at $x = x_0$, $y = 0$ and $z = 0$ with a launching angle α with respect to the z -axis. Under these considerations, Eqs. (1.3-20) and (1.3-21) become

$$\frac{d^2x}{dz^2} = - \frac{n_2}{\tilde{\beta}^2} x(z) \quad (1.3-24a)$$

and

$$\frac{d^2y}{dz^2} = - \frac{n_2}{\tilde{\beta}^2} y(z), \quad (1.3-24b)$$

respectively, where $\tilde{\beta} = n(x, y) \cos \theta(x, y) = n(x_0, 0) \cos \alpha$.

The corresponding boundary conditions for Eqs. (1.3-24a) and Eq. (1.3-24b) are

$$x(0) = x_0, \quad \frac{dx(0)}{dz} = 0 \quad (1.3-25a)$$

and

$$y(0) = 0, \quad \frac{dy(0)}{dz} = \tan \alpha. \quad (1.3-25b)$$

The solutions of Eqs. (1.3-24a) and (1.3-24b) are

$$x(z) = x_0 \cos\left(\frac{\sqrt{n_2}}{\tilde{\beta}} z\right) \quad (1.3-26a)$$

and

$$y(z) = \tilde{\beta} \frac{\tan\alpha}{\sqrt{n_2}} \sin\left(\frac{\sqrt{n_2}}{\tilde{\beta}} z\right), \quad (1.3-26b)$$

respectively. In general, the two equations are used to describe skew rays. As a simple example, if $x_0 = \tilde{\beta} \tan\alpha / \sqrt{n_2}$ and from Eq. (1.3-26), we have

$$x^2(z) + y^2(z) = x_0^2. \quad (1.3-27)$$

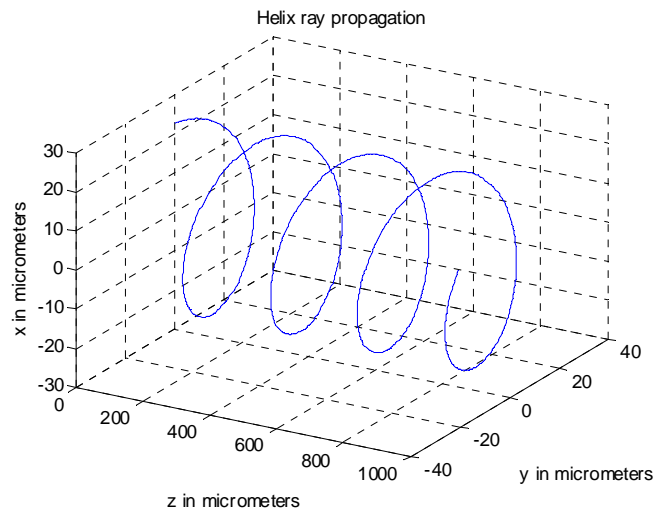


Fig. 1.7 Helix ray propagation.

The ray spirals around the z -axis as a helix. Figure 1.7 shows a MATLAB output for the m-file presented in Table 2.1. For $n(x_0, 0) = 1.5$, $n_2 = 0.001$, and a launching angle α of 0.5 radian, we have $x_0 = 22.74 \mu\text{m}$.

Table 1.1 Helix.m: m-file for plotting helix ray propagation, and its corresponding output for the input parameters used.

```

-----
%Helix.m
%Plotting Eq. (1.3-27)
clear
nxo = input('n(xo) = ');
n2 = input('n2 = ');
alpha = input('alpha [radian] = ');
zin = input('start point of z in micrometers = ');
zfi = input('end point of z in micrometers = ');
Beta = nxo*cos(alpha);
z=zin:(zfi-zin)/1000:zfi;
xo=Beta*tan(alpha)/(n2^0.5);
x=xo*cos((n2^0.5)*z/Beta);
y=xo*sin((n2^0.5)*z/Beta);
plot3(z,y,x)
title('Helix ray propagation')
xlabel('z in micrometers')
ylabel('y in micrometers')
zlabel('x in micrometers')
grid on
sprintf('%f [micrometers]', xo)
view(-37.5+68, 30)
-----
n(xo) = 1.5
n2 = 0.001
alpha [radian] = 0.5
start point of z in micrometers = 0
end point of z in micrometers = 1000

ans =

22.741150 [micrometers]
-----

```

1.4 Matrix Methods in Paraxial Optics

Matrices may be used to describe ray propagation through optical systems comprising, for instance, a succession of spherical refracting and/or reflecting surfaces all centered on the same axis - the *optical axis*. We take the optical axis to be along the z -axis, which is also the general direction in which the rays travel. We will not consider skew rays and our discussion is only confined to those rays that lie in the x - z plane and that are close to the z -axis (called *paraxial rays*). Paraxial rays are close to the optical axis such that their angular deviation from it is small; hence, the sine and tangent of the angles may be approximated by the

angles themselves. The reason for this paraxial approximation is that all paraxial rays starting from a given object point intersect at another point after passage through the optical system. We call this point the image point. Nonparaxial rays may not give rise to a single image point. This phenomenon, which is called *aberration*, is outside the scope of this book. Paraxial optical imaging is also sometimes called *Gaussian optics* as it was Karl Friedrich Gauss (1777-1855) who laid the foundations of the subject.

A ray at a certain point along the x -axis can be specified by its "coordinates," which contains the information of the position of the ray and its direction. Given this information, we want to find the coordinates of the ray at another location further down the optical axis, by means of successive operators acting on the initial ray coordinates, with each operator characteristic of the optical element through which the ray travels along the optical axis. We can represent these operators by matrices. The advantage of this matrix formalism is that any ray can be tracked during its propagation through the optical system by successive matrix multiplications, which can be easily done on a computer. This representation of geometrical optics is widely used in optical element designs.

In what follows, we will first develop the matrix formalism for paraxial ray propagation and examine some of the properties of ray transfer matrices. We then consider some illustrative examples.

1.4.1 The Ray Transfer Matrix

Consider the propagation of a paraxial ray through an optical system as shown in Fig. 1.8. Our discussion is confined to those rays that lie in the xz -plane and are close to the z -axis (the optical axis). A ray at a given cross-section or plane may be specified by its height x from the optical axis and by its angle θ or slope which it makes with the z -axis. The convention for the angle is that θ is measured in radians and is anti-clockwise positive measured from the z -axis. The quantities (x, θ) represent the coordinates of the ray for a given z -constant plane. However, instead of specifying the angle the ray makes with the z -axis, it is customary to replace the corresponding angle θ by $v = n\theta$, where n is the refractive index at the z -constant plane.

In Fig. 1.8, the ray passes through the input plane with *input ray coordinates* $(x_1, v_1 = n_1\theta_1)$, then through the optical system, and finally

through the output plane with *output ray coordinates* $(x_2, v_2 = n_2\theta_2)$. In the paraxial approximation, the corresponding output quantities are linearly dependent on the input quantities. We can, therefore, represent the transformation from the input to the output in matrix form as

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}. \quad (1.4-1)$$

The above $ABCD$ matrix is called the *ray transfer matrix*, and it can be made up of many matrices to account for the effects of a ray passing through various optical elements. We can consider these matrices as operators successively acting on the input ray coordinates. We state that the determinant of the ray transfer matrix equals unity, i.e., $AD - BC = 1$. This will become clear after we derive the translation, refraction and reflection matrices.

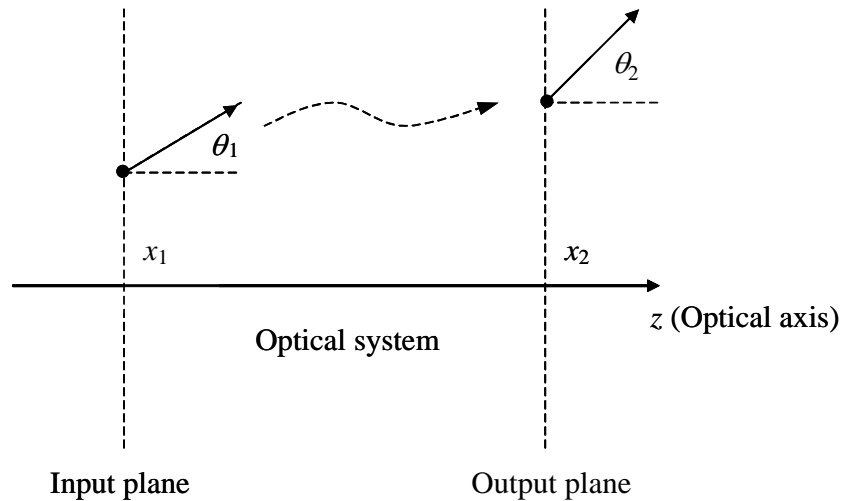


Fig. 1.8 Reference planes in an optical system.

Let us now attempt to understand better the significance of A , B , C , and D by considering what happens if one of them vanishes within the ray transfer matrix.

a) If $D = 0$, we have from Eq. (1.4-1) that $v_2 = Cx_1$. This means that all rays crossing the input plane at the same point x_1 , emerge at the

output plane making the same angle with the axis, no matter at what angle they entered the system. The input plane is called the *front focal plane* of the optical system [see Fig. 1.9(a)].

b) If $B = 0$, $x_2 = Ax_1$ [from Eq. (1.4-1)]. This means that all rays passing through the input plane at the same point x_1 will pass through the same point x_2 in the output plane [see Figure 1.9(b)]. The input and output planes are called the *object* and *image planes*, respectively. In addition, $A = x_2/x_1$ gives the *magnification* produced by the system.

Furthermore, the two planes containing x_1 and x_2 are called *conjugate planes*. If $A = 1$, i.e., the magnification between the two conjugate planes is unity, these planes are called the *unit* or *principal planes*. The points of intersection of the unit planes with the optical axis are the *unit* or *principal points*. The principal points constitute one set of *cardinal points*.

c) If $C = 0$, $v_2 = Dv_1$. This means that all the rays entering the system parallel to one another will also emerge parallel, albeit in a new direction [see Figure 1.9(c)]. In addition, $D(n_1/n_2) = \theta_2/\theta_1$ gives the *angular magnification* produced by the system.

If $D = n_2/n_1$, we have unity angular magnification, i.e., $\theta_2/\theta_1 = 1$. In this case, the input and output planes are referred to as the *nodal planes*. The intersections of the nodal planes with the optical axis are called the *nodal points* [see Figure 1.9(d)]. The nodal points constitute a second set of cardinal points.

d) If $A = 0$, $x_2 = Bv_1$. This means that all rays entering the system at the same angle will pass through the same point at the output plane. The output plane is the *back focal plane* of the system [see Figure 1.9(e)]. Note that the intersection of the front focal and back focal planes with the optical axis are called the *front and back focal points*. The focal points constitute the last set of cardinal points.

Translation Matrix

Figure 1.10 shows a ray traveling a distance d in a homogeneous medium of refractive index n . Since the medium is homogeneous, the ray travels in a straight line [see Eq. (1.3-17)]. The set of equations of translation by a distance d is

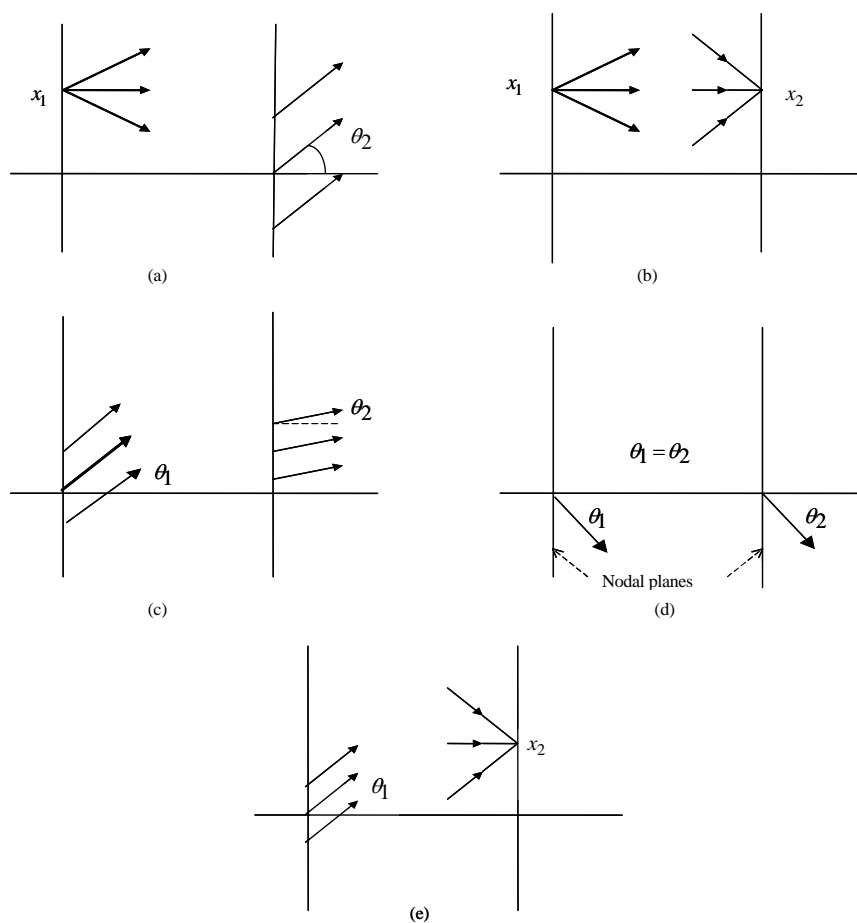


Fig. 1.9 Rays at input and output planes for (a) $D = 0$, (b) $B=0$, (c) $C = 0$, (d) the case when the planes are nodal planes, and (e) $A = 0$.

$$x_2 = x_1 + d \tan\theta_1, \quad (1.4-2a)$$

and

$$n\theta_2 = n\theta_1 \text{ or } v_2 = v_1. \quad (1.4-2b)$$

From the above equations, we can relate the output coordinates of the ray with its input coordinates. We can express this transformation in a matrix form as

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}. \quad (1.4-3)$$

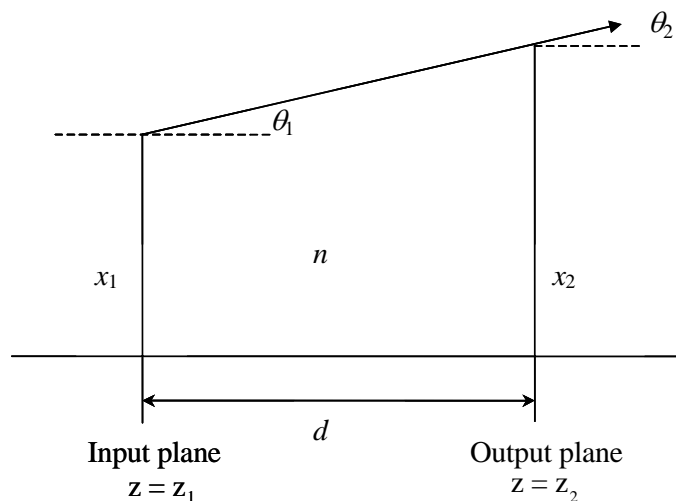


Fig. 1.10 A ray in a homogeneous medium of refractive index n .

The 2×2 ray transfer matrix, for a translation distance of d in a homogeneous medium of refractive index n , is called the *translation matrix* \mathcal{T}_d :

$$\mathcal{T}_d = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix}. \quad (1.4-4)$$

Note that the determinant of the above equation is unity.

Refraction Matrix

We now consider the effect of a spherical surface separating two regions of refractive indices n_1 and n_2 as shown in Fig. 1.11. The center of the curved surface is at C and its radius of curvature is R . The ray strikes the surface at the point A and gets refracted. ϕ_i is the angle of incidence and ϕ_t is the angle of refraction. Note that the radius of curvature of the surface will be taken as positive (negative) if the center C of curvature

lies to the right (left) of the surface. Let x be the height from A to the optical axis. Then the angle ϕ subtended at the center C becomes

$$\sin \phi \approx x/R \approx \phi. \quad (1.4-5)$$

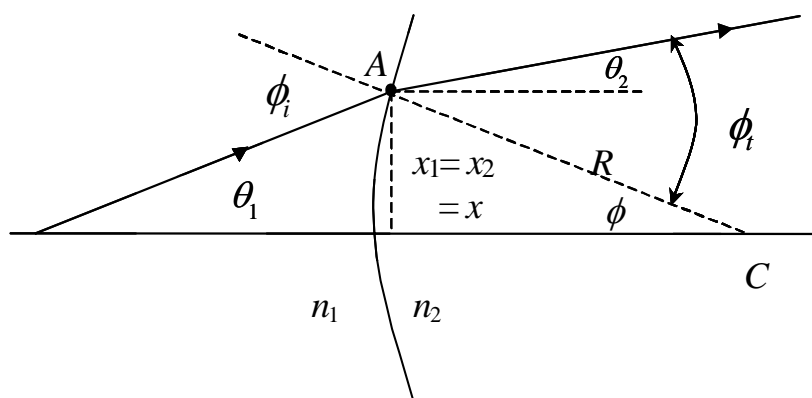


Fig. 1.11 Ray trajectory during refraction at a spherical surface.

We see that in this case, the height of the ray at A, before and after the refraction, is the same, i.e., $x_2 = x_1$. We therefore need to obtain the relationship for v_2 in terms of x_1 and v_1 . Applying Snell's law [see Eq. (1.2-5)] and using the paraxial approximation, we have

$$n_1 \phi_i = n_2 \phi_t. \quad (1.4-6)$$

From geometry, we know from Fig. 1.11 that $\phi_i = \theta_1 + \phi$, and $\phi_t = \theta_2 + \phi$. Hence,

$$n_1 \phi_i = v_1 + n_1 x_1 / R, \quad (1.4-7a)$$

$$n_2 \phi_t = v_2 + n_2 x_2 / R. \quad (1.4-7b)$$

Using Eqs. (1.4-6), (1.4-7) and the fact that $x_1 = x_2$, we obtain

$$v_2 = \frac{n_1 - n_2}{R} x_1 + v_1. \quad (1.4-8)$$

The matrix relating the coordinates of the ray after refraction to those before refraction becomes

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -p & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}, \quad (1.4-9a)$$

where the quantity p given by

$$p = \frac{n_2 - n_1}{R} \quad (1.4-9b)$$

is called the *refracting power* of the spherical surface. When R is measured in meters, the unit of p is called *diopters*. If an incident ray is made to converge (diverge) by a surface, the power will be assumed to be positive (negative) in sign. The (2×2) transfer matrix is called the *refraction matrix* \mathcal{R} and it describes refraction for the spherical surface:

$$\mathcal{R} = \begin{pmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R} & 1 \end{pmatrix}. \quad (1.4-10)$$

Note that the determinant of \mathcal{R} is also unity.

Thin-Lens Matrix

Consider a thick lens as shown in Fig. 1.12. We can show that the input ray coordinates (x_1, v_1) and the output ray coordinates (x_2, v_2) are connected by three matrices (a refraction matrix followed by a translation matrix and then by another refraction matrix):

$$\begin{pmatrix} x_2 \\ v_2 \end{pmatrix} = \mathcal{S} \begin{pmatrix} x_1 \\ v_1 \end{pmatrix}, \quad (1.4-11)$$

where \mathcal{S} is called the *system matrix* and given by, using Eqs. (1.4-4) with $n = n_2$ and Eq. (1.4-10),

$$\begin{aligned} \mathcal{S} &= \mathcal{R}_2 \mathcal{T}_d \mathcal{R}_1 \\ &= \begin{pmatrix} 1 & 0 \\ \frac{n_2 - n_1}{R_2} & 1 \end{pmatrix} \begin{pmatrix} 1 & d/n_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R_1} & 1 \end{pmatrix}. \\ &\quad \text{refraction at} \quad \text{translation} \quad \text{refraction at} \\ &\quad \text{surface 2} \quad \quad \quad \text{surface 1} \end{aligned}$$

Note that in \mathcal{R}_2 , we have interchanged n_1 and n_2 to take into the account that the ray is traveling from n_2 to n_1 .

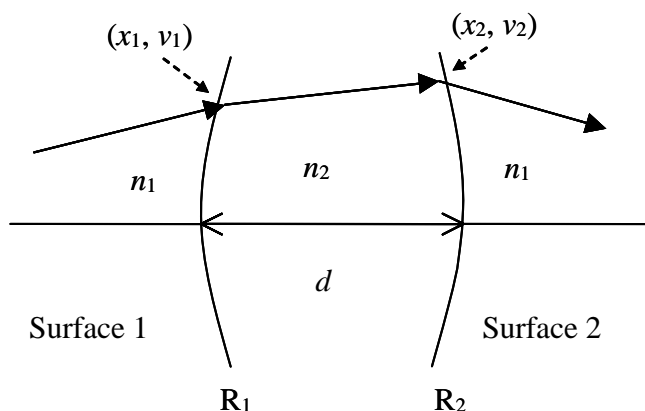


Fig. 1.12 A thick lens: The radii of curvatures of surfaces 1 and 2 are R_1 and R_2 , respectively.

For an ideal thin lens in air, $d \rightarrow 0$ and $n_1 = 1$. Writing $n_2 = n$ for notational convenience, Eq. (1.4-11) becomes

$$\mathcal{S} = \begin{pmatrix} 1 & 0 \\ -p_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -p_1 & 1 \end{pmatrix}, \quad (1.4-12)$$

where $p_1 = (n - 1)/R_1$ and $p_2 = (1 - n)/R_2$ are the *refracting powers* of surfaces 1 and 2, respectively. Note that the translation matrix degenerates into a unit matrix. Equation (1.4-12) can be rewritten as

$$\mathcal{S} = \begin{pmatrix} 1 & 0 \\ -p_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -p_1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} = \mathcal{S}_f, \quad (1.4-13)$$

where \mathcal{S}_f is called the *thin-lens matrix* and f is the *focal length* given by

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right). \quad (1.4-14)$$

For $R_1 > (<) 0$ and $R_2 < (>) 0$, we have $f > (<) 0$. If a ray of light is incident on the left surface of the lens parallel to the axis and for $f > (<) 0$, the ray bends towards (away from) the axis upon refraction through the lens. In the first case, the lens is called a *converging (convex)* lens, while in the second case, we have a *diverging (concave)* lens.

1.4.2 Illustrative examples

Example 1.4 Ray tracing through a single thin lens

(a) *Ray traveling parallel to the optical axis:*

The input ray coordinates are $(x_1, 0)$, and hence the output ray coordinates are given, using Eqs. (1.4-1) and (1.4-13), as $(x_1, -x_1/f)$. This ray now travels in a straight line at an angle $-1/f$ with the axis, which means that if x_1 is positive (or negative), the ray after refraction through the lens intersects the optical axis at a point a distance f behind the lens if the lens is converging ($f > 0$). This justifies why f is called the focal length of the lens. All rays parallel to the optical axis in front of the lens converge behind the lens to a point called the *back focus* [see Fig. 1.13(a)]. In the case of a diverging lens ($f < 0$), the ray after refraction diverges away from the axis as if it were coming from a point on the axis a distance f in front of the lens. This point is called the *front focus*. This is also shown in Fig. 1.13(a).

(b) *Ray traveling through the center of the lens:*

The input ray coordinates are $(0, v_1)$, and hence the output ray coordinates are given, using Eqs. (1.4-6) and (1.4-13), as $(0, v_1)$, which means that a ray traveling through the center of the lens will pass undeviated as shown in Fig. 1.13(b).

(c) *Ray passing through the front focus of a converging lens:*

The input ray coordinates are given by $(x_1, x_1/f)$, so that the output ray coordinates are $(x_1, 0)$. This means that the output ray will be parallel to the axis, as shown in Fig. 1.13(c).

In a similar way, we can also show that for an input ray on a diverging lens appearing to travel toward its *back focus*, the output ray will be parallel to the axis.

Example 1.5 Imaging by a single thin lens

Consider an object OO' located a distance d_o in front of a thin lens of focal length f , as shown in Fig. 1.14. Assume that (x_o, v_o) represents the input ray coordinates originally from point O' , and traveling towards the lens for a distance of d_o . Then the output ray coordinates (x_i, v_i) at a distance d_i behind the lens can be written in terms of the input ray

coordinates, two translation matrices for air ($n = 1$) [see Eq.(1.4-4)] and the thin-lens matrix [see Eq. (1.4-13)] as:

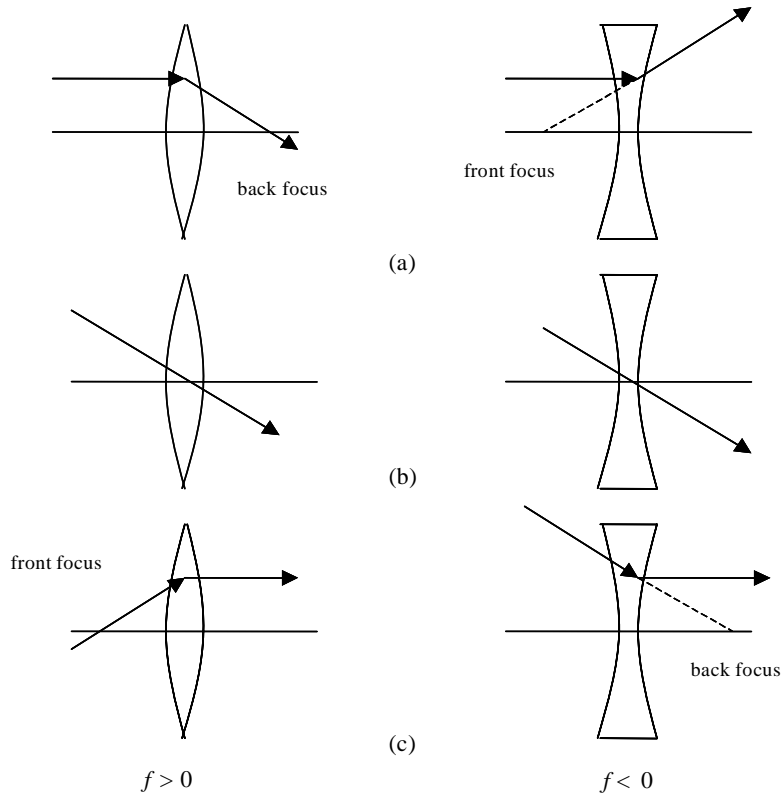


Fig. 1.13 Ray tracing through thin converging and diverging lenses.

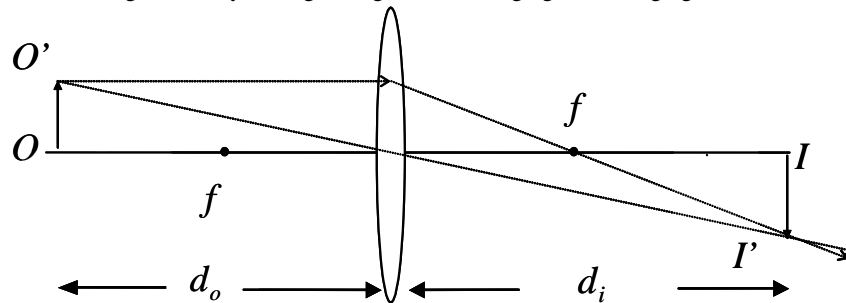


Fig. 1.14 Imaging by a single lens.

$$\begin{aligned}
\begin{pmatrix} x_i \\ v_i \end{pmatrix} &= \mathcal{T}_{d_i} \mathcal{S}_f \mathcal{T}_{d_o} \begin{pmatrix} x_o \\ v_o \end{pmatrix} \\
&= \begin{pmatrix} 1 & d_i \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix} \begin{pmatrix} 1 & d_o \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_o \\ v_o \end{pmatrix} \\
&= \begin{pmatrix} 1 - d_i/f & d_o + d_i - d_o d_i/f \\ -1/f & 1 - d_o/f \end{pmatrix} \begin{pmatrix} x_o \\ v_o \end{pmatrix} \\
&= \mathcal{S} \begin{pmatrix} x_o \\ v_o \end{pmatrix}.
\end{aligned} \tag{1.4-15}$$

We see that \mathcal{S} is the system matrix in our case and by setting $B = 0$ [see Eq. (1.4-10)] in the matrix we have the following celebrated *thin-lens formula* for the imaging lens:

$$\frac{1}{d_o} + \frac{1}{d_i} = \frac{1}{f}. \tag{1.4-16}$$

The *sign convention* for d_o and d_i is as follows. d_o is positive (negative) if the object is to the left (right) of the lens. If d_i is positive (negative), the image is to the right (left) of the lens and it is real (virtual).

Now, returning to Eq. (1.4-15) with Eq. (1.4-16), we have, corresponding to the image plane, the relation

$$\begin{pmatrix} x_i \\ v_i \end{pmatrix} = \begin{pmatrix} 1 - d_i/f & 0 \\ -1/f & 1 - d_o/f \end{pmatrix} \begin{pmatrix} x_o \\ v_o \end{pmatrix}. \tag{1.4-17}$$

For $x_o \neq 0$, we obtain

$$\frac{x_i}{x_o} = M = 1 - \frac{d_i}{f} = \frac{f - d_i}{f} = \frac{f}{f - d_o} = -\frac{d_i}{d_o} \tag{1.4-18}$$

using Eq. (1.4-16), where M is called the *lateral magnification* of the system. If $M > 0$ (< 0), the image is erect (inverted).

1.4.3 Cardinal points of an optical system

We have briefly mentioned cardinal points in Section 1.4.1 and pointed out that there are six *cardinal points* on the optical axis that characterize an optical system. They are the first and second principal (unit) points (H_1, H_2), first and second nodal points (N_1, N_2) and the front and back

focal points (F_1, F_2). The transverse planes normal to the optical axis at these points are called the *cardinal planes*: principal planes, nodal planes and the back and front focal planes. We shall learn how to find their locations in a given optical system. In fact, there is a relationship between the A, B, C , and D system matrix elements and the location of the cardinal planes.

Locating the Principal Planes

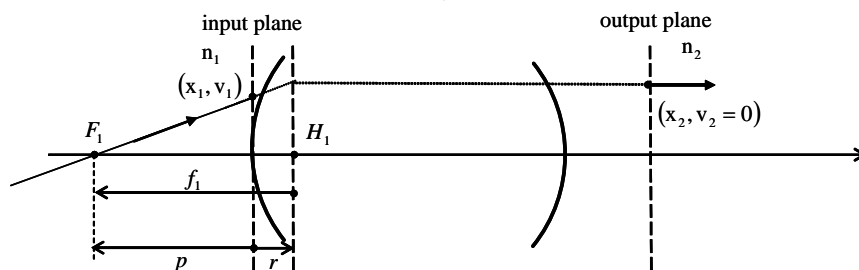
For a given optical system, shown in Fig. 1.15, we first choose the input plane and the output plane. We then assume that we know the $ABCD$ system matrix linking the two chosen planes. Now for the sake of generality, we take n_1 and n_2 to be the refractive indices to the left and to the right of the two planes, respectively.

Consider first, as in Fig. 1.15a), a ray crossing F_1 , by definition, is bent parallel to the optical axis at the first principal plane. The focal point is located at a distance f_1 from the principal plane and at a distance p from the input plane. Furthermore, the distance r locates the principal point from the input plane. The convention for distances are that distances measured to the right of their planes are considered positive and to the left, negative. Since the input ray coordinates and the output ray coordinates are related by the $ABCD$ matrix given by Eq. (1.4-1), we can write

$$v_2 = Cx_1 + Dn_1\theta_1 = 0, \quad (v_1 = n_1\theta_1). \quad (1.4-19)$$

Now, $p = -x_1/\theta_1$ and the negative sign is included because p is to the left side of the input plane according to the convention. Incorporating Eq. (1.4-19), we have

$$p = -Dn_1/C. \quad (1.4-20)$$



(a)

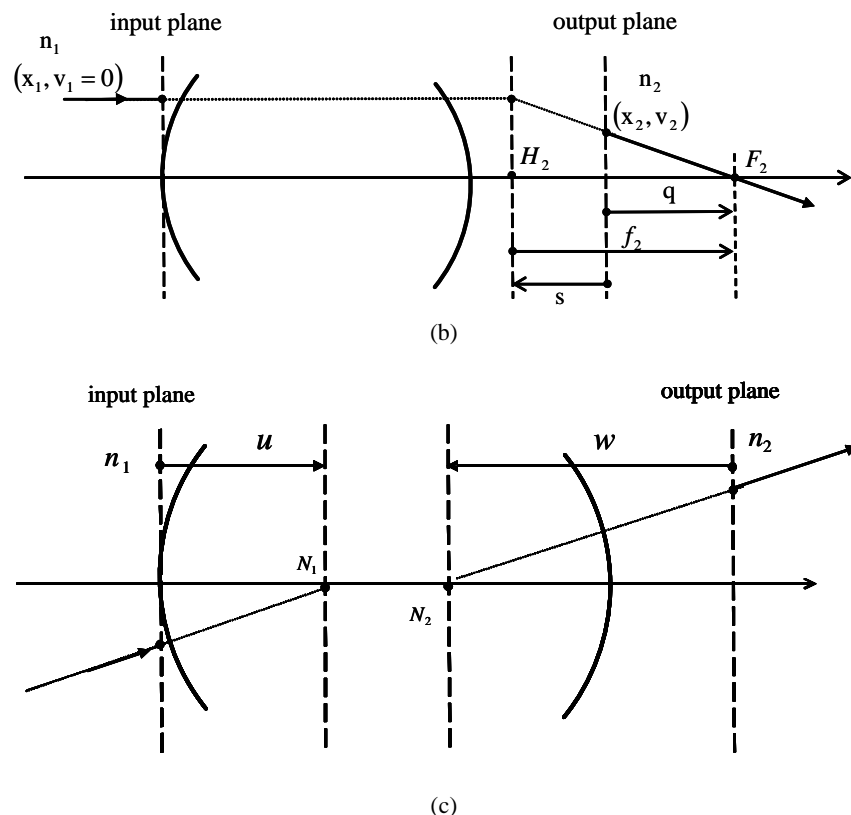


Fig. 1.15 (a) Ray crossing first focal point F_1 is bent parallel to the optical axis at the first principal plane, (b) Ray entering the system parallel to the optical axis is bent at the second principal plane in such a way that it passes through the second focal point F_2 , and (c) Ray entering the system directed towards N_1 is emerged as a ray coming from N_2 .

Now $f_1 = -x_2/\theta_1$, where $x_2 = Ax_1 + Bn_1\theta_1$ from Eq. (1.4-1). We can derive, using the fact that $AD - BC = 1$,

$$f_1 = n_1/C. \tag{1.4-21}$$

Finally to find the location of the first principal point, we notice that $r = -(f_1 - p)$. By incorporating Eqs. (1.4-20) and (1.4-21), we have

$$r = n_1(D - 1)/C. \tag{1.4-22}$$

Similarly, with reference to Fig. 1.15b) we can find the location of the second principal plane. By definition, a ray which enters the system parallel to the optical axis at the height x_1 arrives the same height at the second principal plane. The ray is then bent at the second principal plane in such a way that it passes through the second focal point F_2 . Again, the convention for distances q , s , and f_2 are that distances measured to the right of their planes (output plane and the second principal plane) are considered positive and to the left, negative. We, therefore, write that

$$q = -x_2/\theta_2, \quad (v_2 = n_2\theta_2), \quad (1.4-23)$$

where the negative sign is included in the above equation as $\theta_2 < 0$. Now, from Eq. (1.4-1), we have $v_2 = Cx_1$ and $x_2 = Ax_1$, and we can re-write Eq. (1.4-23) in terms of the elements of the $ABCD$ matrix:

$$q = -An_2/C. \quad (1.4-24)$$

To find the second focal length, write $f_2 = -x_1/\theta_2$. Using $v_2 = Cx_1$, the second focal length is

$$f_2 = -n_2/C. \quad (1.4-25)$$

Finally, to find s , we refer to Fig. 1.15b) and write $s = q - f_2$. Using Eqs. (1.4-24) and (1.4-25), we have

$$s = n_2(1 - A)/C. \quad (1.4-26)$$

Locating the Nodal Planes

Similarly, we can find the location of the Nodal planes with reference to Fig. 1.15c). Again, the convention for distances u and w are that distances measured to the right of their planes (output plane and input planes) are considered positive and to the left, negative. We state the results as follows:

$$u = (Dn_1 - n_2)/C, \quad (1.4-27)$$

and

$$w = (n_1 - An_2)/C. \quad (1.4-28)$$

Example 1.6 Ray tracing using principal planes and nodal planes

An object (an erected arrow denoted by O) is located 20cm from the ideal positive lens of focal length $f_p = 10\text{cm}$. The distance between the positive lens and the ideal negative lens of focal length $f_n = -10\text{cm}$ is 5cm , as shown in Fig. 1.16. We shall draw a ray diagram for the two-lens optical system for the image (I).

We choose the input and output planes to be the location where the positive lens and the negative lens are situated, respectively. The system matrix S linking the two planes is

$$S = \begin{pmatrix} 1 & 0 \\ 10 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0.05 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -10 & 1 \end{pmatrix} = \begin{pmatrix} 0.5 & 0.05 \\ -5 & 1.5 \end{pmatrix}, \quad (1.4-29)$$

where we have converted distances to meters and focal lengths to diopters.

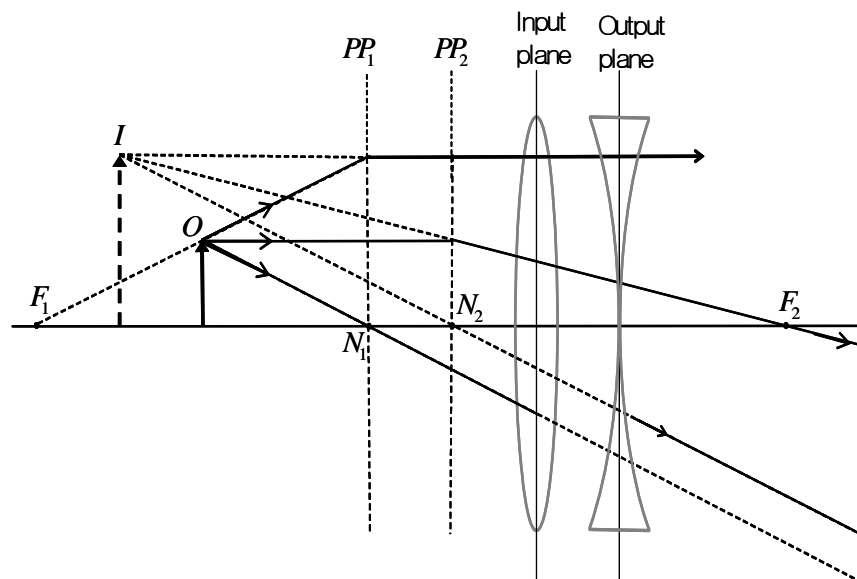


Fig. 1.16 Ray tracing using principal planes and nodal planes.

Since we have found the $ABCD$ matrix of the system, we can now find all the cardinal points and planes, which will help us to draw a ray

diagram for the optical system. Assuming $n_1 = n_2 = 1$, i.e., the optical system is immersed in air, we tabulate the results as follows:

Front focal point F_1 :

$$\begin{aligned} p &= D/C = 1.5/(-5) = -0.3m \\ &= -30cm < 0 \text{ (30cm left of the input plane).} \end{aligned}$$

Back focal point F_2 :

$$\begin{aligned} q &= -A/C = -0.5/(-5) = 0.1m \\ &= 10cm > 0 \text{ (10cm right of the output plane).} \end{aligned}$$

First Principal point H_1 :

$$\begin{aligned} r &= (D-1)/C = (1.5-1)/(-5) = -0.1m \\ &= -10cm < 0 \end{aligned}$$

(10cm left of the input plane, PP_1 denotes the first principal plane).

Second Principal point H_2 :

$$\begin{aligned} s &= (1-A)/C = (1-0.5)/(-5) = -0.1m \\ &= -10cm < 0 \end{aligned}$$

(10cm left of the output plane, PP_2 denotes the second principal plane).

First Nodal point N_1 :

$$\begin{aligned} u &= (D-1)/C \\ &= -10cm < 0 \text{ (10cm left of the input plane).} \end{aligned}$$

Second Nodal point N_2 :

$$\begin{aligned} w &= (1-A)/C \\ &= -10cm < 0 \text{ (10cm left of the output plane).} \end{aligned}$$

Notice that the equivalent focal length of the optical system is f_2
 $= -1/C = 20cm$.

1.5 Reflection Matrix and Optical Resonators

There is a rule that will enable us to use the translation matrix \mathcal{T}_d and refraction matrix \mathcal{R} even for reflecting surfaces such as mirrors. When a light ray is traveling in the $-z$ direction, the refractive index of the medium through which the ray is transversing is taken as negative.

According to the rule, from the refraction matrix [see Eq. (1.4-10)], we can modify it to become the *reflection matrix* $\tilde{\mathcal{R}}$:

$$\tilde{\mathcal{R}} = \begin{pmatrix} 1 & 0 \\ -p & 1 \end{pmatrix},$$

where $p = \frac{n_2 - n_1}{R} = \frac{(-n_1) - n_1}{R} = \frac{-2n}{R}$ and n is the refractive index for the medium in which the mirror is immersed. The situation is shown in Fig. 1.17. Hence we can write the reflection matrix as

$$\tilde{\mathcal{R}} = \begin{pmatrix} 1 & 0 \\ \frac{2n}{R} & 1 \end{pmatrix}. \quad (1.5-1)$$

We see that if the rule is used on the equation for the power of a surface, we find that a concave mirror (R being negative) will give a positive power p , which is in agreement with the common knowledge that a concave mirror will focus rays, as illustrated in Fig. 1.17. The focal length of the spherical mirror is $f = -R/2n$.

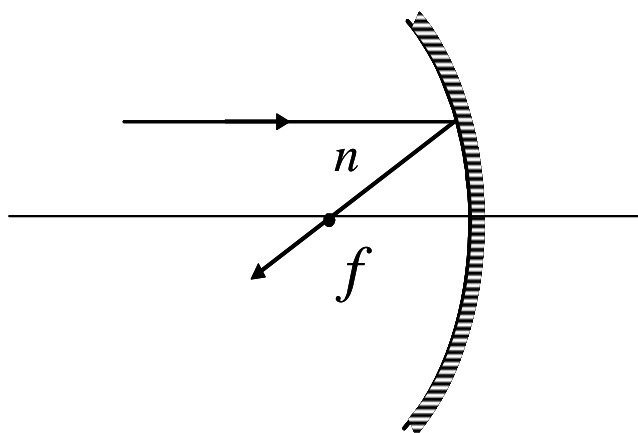


Fig. 1.17 Spherical mirror.

In dealing with the translation matrix upon ray reflection, we adopt the convention in that when light rays travel between planes $z = z_1$ and $z = z_2 > z_1$, $z_1 - z_2$ is taken to be positive (negative) for a ray traveling in the $+z$ ($-z$) direction. Again, the refractive index of the

medium is taken to as negative. By taking the value of the refractive index to be negative when a ray is traveling in the $-z$ direction, we can use the same translation matrix throughout the analysis when reflecting surfaces are included in the optical system. With reference to Fig. 1.18, the translation matrices between various planes are given as follows:

$$\mathcal{T}_{21} = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix} \quad \text{between planes 1 and 2;}$$

$$\mathcal{T}_{32} = \begin{pmatrix} 1 & -d/-n \\ 0 & 1 \end{pmatrix} \quad \text{between planes 2 and 3, and}$$

$$\mathcal{T}_{31} = \mathcal{T}_{32}\mathcal{T}_{21} = \begin{pmatrix} 1 & 2d/n \\ 0 & 1 \end{pmatrix} \quad \text{between planes 1 and 3.}$$

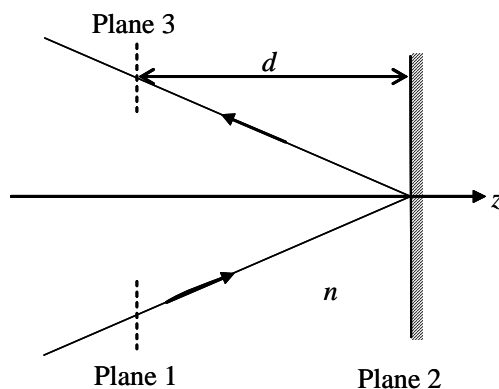


Fig. 1.18 Rays reflected from a plane mirror.

Optical Resonators

An *optical resonator* is an optical system consisting of two mirrors of radii of curvature R_1 and R_2 , separated by a distance d , as shown in Fig. 1.19. The resonator forms an important part of a laser system. Indeed, for sustained oscillations, implying a constant laser output, the resonator must be stable. We shall now obtain the condition for a stable resonator. In stable resonators, a light ray must keep bouncing back and forth and remain trapped inside in order that oscillations are sustained.

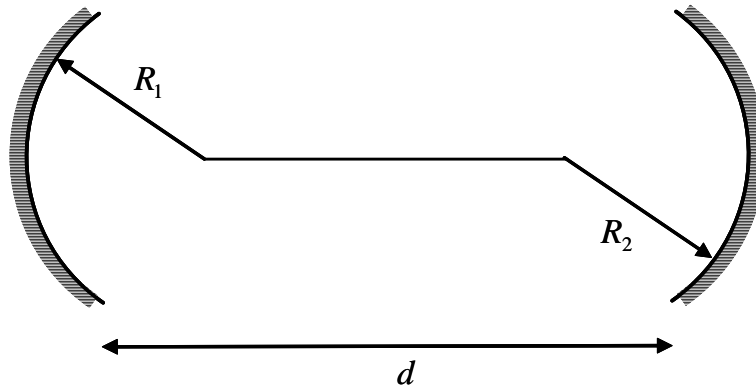


Fig. 1.19 Resonator consisting of two spherical mirrors.

To follow a light ray through a resonator, we start the ray at the left mirror traveling toward to the right mirror, and then reflecting back to the left mirror. The system matrix describing a round trip through the resonator is

$$\begin{aligned} \mathcal{S} &= \tilde{\mathcal{R}}_1 \mathcal{T}_d \tilde{\mathcal{R}}_2 \mathcal{T}_d \\ &= \begin{pmatrix} 1 & 0 \\ \frac{2}{R_1} & 1 \end{pmatrix} \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{2}{R_2} & 1 \end{pmatrix} \begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \end{aligned} \quad (1.5-2a)$$

where

$$\begin{aligned} A &= 1 + 2d/R_2, & B &= 2d(1 + d/R_2), \\ C &= 2\left[\frac{1}{R_1} + \frac{1}{R_2}\left(1 + \frac{2d}{R_1}\right)\right], & (1.5-2b) \\ D &= \frac{2d}{R_1} + \left(1 + \frac{2d}{R_1}\right)\left(1 + \frac{2d}{R_2}\right). \end{aligned}$$

Hence, we can write

$$\begin{pmatrix} x_1 \\ v_1 \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x_0 \\ v_0 \end{pmatrix},$$

where (x_1, v_1) is the ray coordinates after one round trip and (x_0, v_0) is the ray coordinates when it started from the left mirror. Now, the

coordinates of the ray (x_m, v_m) after m complete round trips (oscillations) would be

$$\begin{pmatrix} x_m \\ v_m \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^m \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \quad (1.5-3)$$

We can show that

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^m = \frac{1}{\sin\theta} \begin{pmatrix} A\sin m\theta - \sin(m-1)\theta & B\sin m\theta \\ C\sin m\theta & D\sin(m-1)\theta \end{pmatrix} \quad (1.5-4)$$

where the angle θ has been defined as

$$\cos\theta = \frac{1}{2}(A + D). \quad (1.5-5)$$

In order to achieve stability, the coordinates of the ray after m trips should not diverge as $m \rightarrow \infty$. This happens if the magnitude of $\cos\theta$ is less than 1. In other words, if θ is a complex number, the terms $\sin m\theta$ and $\sin(m-1)\theta$ in Eq. (1.5.4) diverges. Hence the *stability criterion* is

$$-1 \leq \cos\theta \leq 1 \quad (1.5-6)$$

or, when using Eqs. (1.5-5) and (1.5-2b),

$$0 \leq \left(1 + \frac{d}{R_1}\right)\left(1 + \frac{d}{R_2}\right) \leq 1. \quad (1.5-7)$$

The stability criterion is often written using the so-called *g parameters* of the resonator as

$$0 \leq g_1 g_2 \leq 1, \quad (1.4-26)$$

where $g_1 = \left(1 + \frac{d}{R_1}\right)$ and $g_2 = \left(1 + \frac{d}{R_2}\right)$.

Figure 1.20 shows the *stability diagram* for optical resonators. Only those resonator configurations that lie in the shaded region correspond to a stable configuration. The point marked O corresponds to the so-called *confocal configuration*, where $R_1 = R_2 = -d$ or $g_1 g_2 = 0$. Figure 1.21 shows ray propagation inside such a resonator.

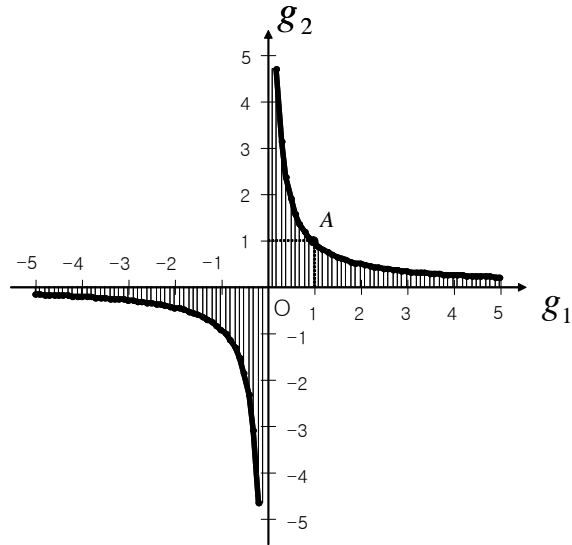


Fig. 1.20 Stability diagram for optical resonators.

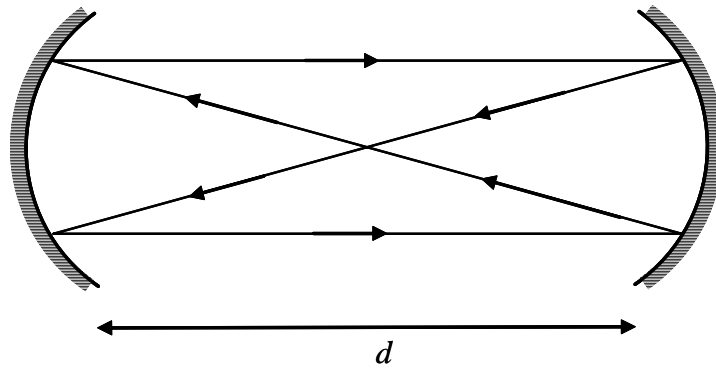


Fig. 1.21 Ray propagation inside a stable resonator.

1.6 Ray Optics using MATLAB

Example 1 Obtaining output ray coordinates of a single lens system

We shall find the ray coordinates $\mathbf{r}_i = (x_i, v_i)$ an arbitrary distance z behind a lens of focal lens f when the input ray coordinates

$\mathbf{r}_o = (x_0, v_0)$ for a ray starting from an object located a distance d_o in front of the lens is specified. \mathcal{T}_{d_o} and \mathcal{T}_z denote the translation matrices for the ray in air before and after the lens (corresponding to object and image distances, respectively), while \mathcal{S}_f is the lens matrix. The product of the three $\mathcal{S} = \mathcal{T}_z \mathcal{S}_f \mathcal{T}_{d_o}$ gives the overall system matrix for the optical system. The program gives the output ray coordinates \mathbf{r}_i . All distance dimensions have been written in centimeter.

As an example, we create the MATLAB function Ray_s. In MATLAB, after the prompt \gg , we type `[detS,ri]=Ray_s([0;1],15,10,30)` to denote input conditions: $\mathbf{r}_o = (0, 1)$, $d_o = 15\text{cm}$, $f = 10\text{cm}$, and $z = 30\text{cm}$. We obtain $\mathbf{r}_i = (0, -0.5)$ as an output.

Table 1.2 MATLAB code for ray traveling through a single lens, and the corresponding MATLAB output.

```
-----
function [detS, ri]=Ray_s(ro, do, f,z);
%This function is for output ray vector of
%a single lens system
To=[1, do;0,1];
Sf=[1,0;-(1/f),1];
Ti=[1,z;0,1];
S=Ti*Sf*To;
%Checking determinant for overall matrix
detS=det(S);
%"image" ray coordinate is ri
ri=S*ro;
-----
```

```
Type in Matlab prompt
>>[detS, ri]=Ray_s([0;1], 15, 10, 30)
```

```
-----
Output from Matlab
```

```
detS =
      1
ri =
      0
     -0.5000
-----
```

We interpret the program as follows: if the input ray starts from the optical axis at a distance of 15cm from the lens with $v = 1$ rad, the output ray meets the axis a distance of 30cm behind the single lens with $v = -0.5$ rad. In other words, for an object distance d_o of 15cm, the image distance $z = d_i$ is 30 cm. Finding the determinant of the overall

system matrix \mathcal{S} being unity is a check of the computations. Note that the ray coordinates at any plane z behind the lens can be found by substituting a number for the value of z in the program.

To find the lateral magnification of the imaging system, we can enter in input ray coordinates of, say, (1,1). Using the same program as above, the output ray coordinates at $z = 30\text{cm}$ (the image plane) works out to be (-2.0, -0.6). This means that the magnification of the system equals -2, which corresponds to an inverted real image of twice the size as the object, as expected.

Example 2 Obtaining output ray coordinates of a single lens system

We shall find the ray coordinates $\mathbf{r}_i = (x_i, v_i)$ at an arbitrary distance z behind a two-lens combination of focal lengths $f_{1,2}$ and separated by a distance d when the input ray coordinates \mathbf{r}_o for rays starting from an object located a distance d_o in front of the lens are specified. \mathcal{T}_{d_o} and \mathcal{T}_{d_i} denote the translation matrices for the ray in air before and after the lens (corresponding to object and image distances, respectively), \mathcal{T}_d denotes the translation matrix for a ray traveling between the two lenses, while $\mathcal{S}_{f_{1,2}}$ are the lens matrices, respectively. The product of $\mathcal{S} = \mathcal{T}_{d_i}\mathcal{S}_{f_2}\mathcal{T}_d\mathcal{S}_{f_1}\mathcal{T}_{d_o}$ gives the overall system matrix for the optical system. The program gives the output ray coordinates r_i . All dimensions have been written in centimeter.

We create the MATLAB function Ray_d. In MATLAB, after the prompt \gg , we type `[detS,ri]=Ray_d([0;1],10,10,10,10,20)` to denote input conditions of the function: $\mathbf{r}_o = (0, 1)$, $d_o = 10\text{cm}$, $z = 10\text{cm}$, $f_1 = 10\text{cm}$, $f_2 = 10\text{cm}$ and $d = 20\text{cm}$. We obtain output coordinates $\mathbf{r}_i = (0, -1)$ as an output. Note also that when we input $\mathbf{r}_o = (1, 0)$ with other input variables the same, we get $\mathbf{r}_i = (-1, 0)$. This corresponds that when the input ray is parallel to the optical axis, the output ray is also parallel to the axis with an inverted image having a unit magnification.

Table 1.3 MATLAB code for ray traveling through a two lens system, and the corresponding MATLAB output.

```
-----
function [detS, ri]=Ray_d(ro, do, z, f1, f2, d);
%This function is for output ray vector of a double lens
%system
```

```

To=[1, do;0,1];
Sf1=[1,0;-(1/f1),1];
Td=[1,d;0,1];
Sf2=[1,0;-(1/f2),1];
Ti=[1,z;0,1];
S=Ti*Sf2*Td*Sf1*To;

%Checking determinant for overall matrix
detS=det(S);
%"image" ray coordinate is ri
ri=S*ro;
-----
detS =
      1
ri =
      0
     -1
-----

```

Example 3 Finding the image location in a single lens system

The following program is an extension to the first MATLAB example in this section. The MATLAB function Ray_z gives the location of the image plane for a given location of the object in a single imaging lens system. To do this, the object is taken to be an on-axis point, and the ray coordinates monitored behind the lens. If the position of the output ray is sufficiently close to the optical axis behind the lens, the corresponding value of z is the location of the image. As an example, we input $d_o = 15\text{ cm}$, $f = 10\text{ cm}$, $Z_s = 0$, $Z_f = 50\text{ cm}$, and $\Delta z = 0.1\text{ cm}$, where Z_s , Z_f and Δz represent the start and end points of the search range and the resolution of the search, respectively. The program output shows that indeed the image location works out to be a distance of 30 cm behind the lens for an object distance of 15 cm in front of the lens with focal length equal to 10 cm.

Table 1.4 MATLAB code for locating image plane for single lens imaging, and the corresponding MATLAB output.

```

-----
function [z_est, M]=Ray_z(do, f, Zs, Zf, dz);
%This function is for searching image distance of the single
%lens system
To=[1, do;0,1];
Sf=[1,0;-(1/f),1];

ro=[0;1];

```

```

n=0;
for z=Zs:dz:Zf
    n=n+1;
    Z1(n)=z;
    Ti=[1,z;0,1];
    S=Ti*Sf*To;
    %"image" ray coordinate is ri
    ri=S*ro;
    Ri(n)=ri(1,1);
end
[M, N]=min(abs(Ri));
z_est=Z1(N);
-----
>>[z_est, M]=Ray_z(15, 10, 0, 50, 0.1)
z_est =
    30
M =
    0
-----

```

Problems

- 1.1** A laser rocket is accelerated in free space by a photon engine that emits 10 kW of blue light ($\lambda = 450 \text{ nm}$).
- What is the force on the rocket?
 - If the rocket weighs 100 kg, what is its acceleration?
 - How far will it have traveled in one year if it starts from zero velocity?
- [Courtesy of Adrian Korpel, Professor Emeritus, Univ. Iowa]
- 1.2** Derive the laws of reflection and refraction by considering the incident, reflected and refracted light to comprise a stream of photons characterized by a momentum $\mathbf{p} = \hbar\mathbf{k}$, and \mathbf{k} is the wavevector in the direction of ray propagation. Employ the law of conservation of momentum, assuming that the interface, say, $y = \text{constant}$, only affects the y -component of the momentum. This provides an alternative derivation of the laws of reflection and refraction.
- 1.3** Show that the z -component of the ray equation,

$$\frac{d}{ds}\left(n\frac{dz}{ds}\right) = \frac{\partial n}{\partial z},$$

can be derived directly from the equations for x and y [Eq. (1.3-13)]. Hint: make use of $ds^2 = dx^2 + dy^2 + dz^2$.

- 1.4** a) Show that for the square-law medium $n^2(x, y) = n_0^2 - n_2(x^2 + y^2)$,

$$x(z) = \frac{n_0 \sin \alpha}{\sqrt{n_2}} \sin \left(\frac{\sqrt{n_2}}{n_0 \cos \alpha} z \right)$$

when the initial ray position is at $x = 0$ with α being the launching angle.

b) Plot $x(z)$ for $\alpha = 10^\circ$, 20° , and 30° when $n_0 = 1.5$, $n_2 = 0.1 \text{ mm}^{-2}$. Draw some conclusion from the plots.

c) For paraxial rays, i.e., α is small, can you draw a different conclusion from that obtained from part b)?

- 1.5** Show that the ray transfer matrix for the square-law medium $n^2(x, y) = n_0^2 - n_2(x^2 + y^2)$ is

$$\begin{pmatrix} \cos \beta z & \frac{1}{n_0 \beta} \sin \beta z \\ -n_0 \beta \sin \beta z & \cos \beta z \end{pmatrix},$$

where $\beta = \sqrt{n_2}/n_0$.

- 1.6** For medium

$$\begin{aligned} n^2(x, y) &= n_0^2 - \gamma x & x > 0 \\ &= n_0^2 & x \leq 0, \end{aligned}$$

Find $x(z)$ for a ray passing through $x = 0$, $z = 0$ with an angle α with respect to the optical axis, where $\gamma > 0$. Sketch ray path $x(z)$. These ray paths are good examples for radio wave propagation through the ionosphere.

- 1.7** A point object is placed a distance 2m away from a concave mirror of radius of curvature $R = -80\text{cm}$. Find the location of the image and draw the ray diagram of the image formed.

- 1.8** In the imaging system shown in Fig. 1.14, if now the object is displaced axially a small distance δd_o , find an expression for the corresponding distance δd_i of the image. $\delta d_o/\delta d_i$ is called the *longitudinal magnification* M_z . Show that it is the square of the lateral magnification M .
- 1.9** An object is placed 12cm in front of a lens-mirror combination as shown in Fig. P1.8. Using ray transfer matrix concepts, find the position, and magnification of the image. Also, draw the ray diagram of the optical system.

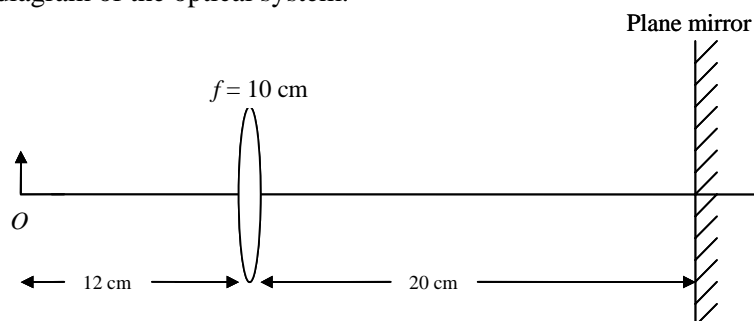


Fig. P1.8

- 1.10** A glass hemisphere, shown in Fig. P1.9, of radius r and refractive index n is used as a lens for paraxial rays. Find the location of the first principal plane, second principal plane and the equivalent focal length of the optical system.

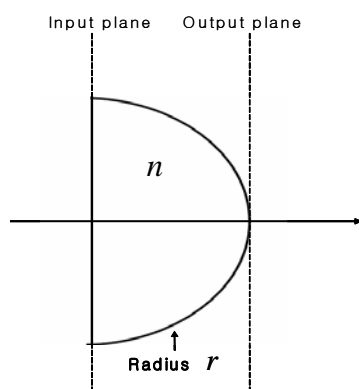


Fig. P. 1.9

- 1.11** Referring to Fig. P.1.10, show that the equivalent focal length f of the two-lens combination can be expressed as

$$\frac{1}{f} = \frac{1}{f_a} + \frac{1}{f_b} - \frac{d}{f_a f_b},$$

assuming $d < f_a + f_b$. Also, locate the second principal plane from which where f is measured.

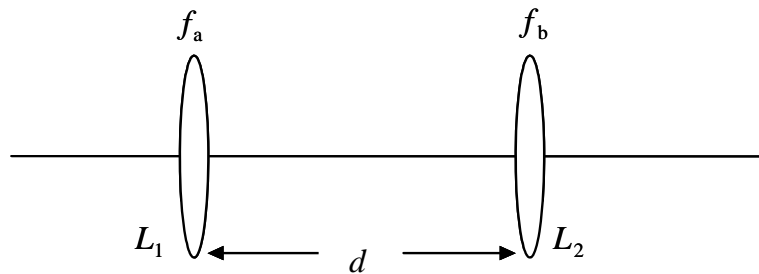


Fig. P. 1.10

- 1.12** Show Eqs. (1.4-27) and (1.4-28).
- 1.13** Show Eq. (1.5-4) by mathematical induction.
- 1.14** Draw the ray propagation diagram similar to that shown in Fig. 1.21 for the following parameters:
- $R_1 = \infty$ and $R_2 = -2d$ (hemispherical resonator);
 - $R_1 = R_2 = -d/2$ (concentric resonator);
 - $R_1 = d$ and $R_2 = \infty$.
 - $R_1 = R_2 = \infty$.

References

- 1.1** Banerjee, P.P. and T.-C. Poon (1991). *Principles of Applied Optics*. Irwin, Illinois.
- 1.2** Feynman, R., R. B. Leighton and M. Sands (1963). *The Feynmann Lectures on Physics*. Addison-Wesley, Reading, Massachusetts.
- 1.3** Fowles, G. R. and G. L. Cassiday (2005). *Analytical Mechanics* (7th ed.). Thomson Brooks/Cole, Belmont, CA.

- 1.4** Gerard, A. and J. M. Burch (1975). *Introduction to Matrix Methods in Optics*. Wiley, New York.
- 1.5** Ghatak, A. K. (1980). *Optics*. Tata McGraw-Hill, New Delhi.
- 1.6** Ghatak, A. K. and Thyagarajan, K. (1998). *An Introduction to Fiber Optics*. Cambridge University Press. Cambridge.
- 1.7** Goldstein, H. (1950). *Classical Mechanics*. Addison-Wesley, Reading, Massachusetts.
- 1.8** Hecht, E. and A. Zajac (1975): *Optics*. Addison-Wesley, Reading, Massachusetts.
- 1.9** Klein, M. V.(1970). *Optics*. Wiley, New York.
- 1.10** Nussbaum, A and R. A. Phillips (1976). *Contemporary Optics for Scientists and Engineers*, Prentice-Hall, New York.
- 1.11** Lakshminarayanan, V., A. K. Ghatak, and K. Thyagarajan (2002). *Lagrangian optics*, Kluwer Academic Publishers, Boston.
- 1.12** Pedrotti F. L. and L. S. Pedrotti (1987). *Introduction to Optics*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
- 1.13** Poon T.-C. and P. P. Banerjee (2001). *Contemporary Optical Image Processing with MATLAB®*. Elsevier, Oxford, UK.