

Geometrical Optics

QuadriviumPress

Abstract

This comprehensive textbook covers the fundamental principles of optics, from geometrical optics and wave phenomena to advanced topics including lasers, fiber optics, and optical instrumentation. Designed for undergraduate students in physics and electrical engineering, it provides a rigorous yet accessible treatment of topics including ray optics, interference, diffraction, polarization, lasers, and fiber optics.

Keywords optics, physics, education, textbook, light, lenses, lasers, fiber optics, polarization, interference, diffraction

What you should know and be able to do after studying this chapter

- Principle of Fermat.
- Understand the approximation made in Gaussian geometrical optics.
- Use the signed object and image coordinates in the spherical-interface imaging equation.
- Obtain the thin-lens imaging equation by applying the interface formula at both surfaces.
- Understand how the image of two and more lenses is derived from that of a single lens by construction and by computing the intermediate images. You do not need to know the imaging equation and the formulae for the focal distances of two thin lenses.
- Understand the matrix method (you do not need to know the matrices by heart).
- Recognize where the thin-lens approximation stops applying.
- Understand the limitations of geometrical optics, in particular when diffraction optics is needed.

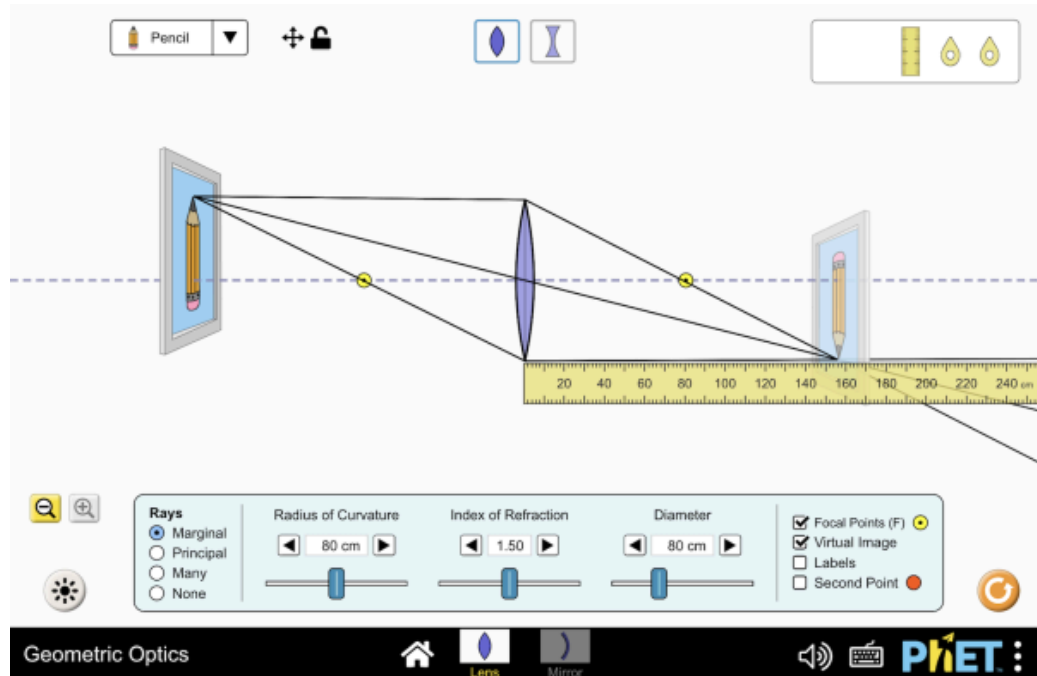


Figure 1: Place a lens or mirror, drag the object, and watch real and virtual images form. This is the interactive companion to the Gaussian imaging constructions later in the chapter.

Interactive simulation: [Geometric Optics](#)

1. INTRODUCTION

Geometrical optics is an old subject, but it is still essential to understand and design optical instruments such as cameras, microscopes, telescopes etc. Geometrical optics started long before light was described as a wave as is done in wave optics, and long before it was discovered that light is an electromagnetic wave and that optics is part of electromagnetism.

In this chapter, we go back in history and treat geometrical optics. That may seem strange now that we have a much more accurate and better theory at our disposal. However, the predictions of geometrical optics are under quite common circumstances very useful and also very accurate. In fact, for many optical systems and practical instruments there is no alternative to geometrical optics because more accurate theories are much too complicated to use.

When a material is illuminated, its molecules start to radiate spherical waves (more precisely, they radiate like tiny electric dipoles) and the total wave scattered by the material is the sum of all these spherical waves. A time-harmonic wave has at every point in space and at every instant of time a well defined phase. A **wavefront** is a surface of equal phase at a fixed time. In an isotropic medium it moves along its local normal at the phase velocity.

For a plane wave in an isotropic medium, constant-phase surfaces are planes; their normal is the direction of the wave vector and energy flow. A curved wavefront behaves locally like a plane wave when its radius of curvature and the optical features are large compared with the wavelength. Geometrical optics then describes its propagation by rays.

Geometrical optics is based on the intuitive idea that light consists of a bundle of rays. But what is a ray?

Note

In the isotropic media considered here, a ray follows the local normal to the wavefront in the direction of energy flow.

Consider a point source at some distance before an opaque screen with an aperture. According to the ray picture, the light distribution on a second screen further away from the source and parallel to the first screen is simply an enlarged copy of the aperture (see [Figure 2](#)). The copy is enlarged due to the fanning out of the rays. However, this description is only accurate when the wavelength of the light is very small compared to the diameter of the aperture. If the aperture is only ten times the wavelength, the pattern is much broader due to the bending of the rays around the edge of the aperture. This phenomenon is called **diffraction**. Diffraction cannot be explained by geometrical optics and will be studied in [Scalar Diffraction Optics](#).

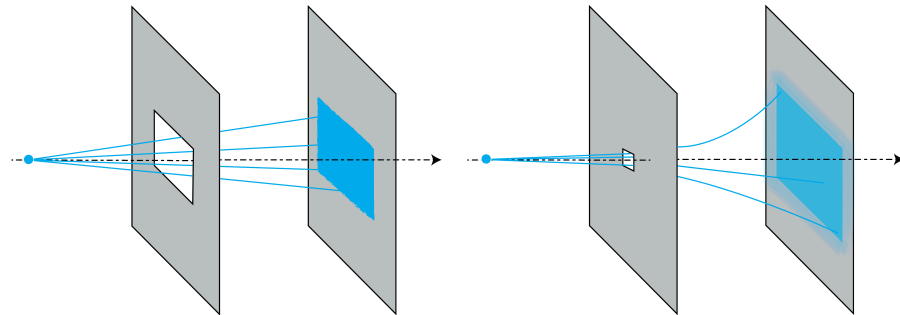


Figure 2: Light distribution on a screen due to a rectangular aperture. Left: for a large aperture, we get an enlarged copy of the aperture. Right: for an aperture that is of the order of the wavelength there is strong bending (diffraction) of the light.

Geometrical optics is accurate when the sizes of the objects in the system are large compared to the wavelength. It is possible to derive geometrical optics from Maxwell's equations by formally expanding the electromagnetic field in a power series in the wavelength and retaining only the first term of this expansion¹. However, this derivation is not rigorous because the power series generally does not converge (it is a so-called asymptotic series).

¹See e.g. Born and Wolf, *Principles of Optics*, Chapter 1.

Although it is possible to incorporate polarization into geometrical optics, this is not standard theory and we will not consider polarization effects in this chapter.

2. PRINCIPLE OF FERMAT

The starting point of the treatment of geometrical optics is the

Important

Fermat's principle. At a fixed frequency, a light ray follows a path whose optical path length is stationary with respect to nearby paths. In the elementary examples below the stationary path is a local minimum.

The phase velocity in a material of refractive index n is c/n , where $c \approx 3 \times 10^8$ m/s is the vacuum speed. The optical path integral below measures the phase delay at one frequency. In a nondispersive medium this delay equals the ray's travel time; a pulse in a dispersive medium generally propagates at a different group velocity. Rømer's 1676 observations of Io's eclipses established a finite light-travel time.

Let $\mathbf{r}(s)$ be a ray with s the length parameter. The ray links two points S and P . Suppose that the refractive index varies with position: $n(\mathbf{r})$. Over the infinitesimal distance from s to $s + ds$, the speed of the light is

$$\frac{c}{n(\mathbf{r}(s))}. \quad (1)$$

The corresponding differential phase delay, expressed in seconds, is:

$$dt = \frac{n(\mathbf{r}(s))}{c} ds, \quad (2)$$

The integrated phase delay from S to P is:

$$t_{S \rightarrow P} = \int_0^{s_P} \frac{n(\mathbf{r}(s))}{c} ds, \quad (3)$$

where s_P is the distance along the ray from S to P . The **optical path length** [m] of the ray between S and P is defined by:

$$\text{OPL} = \int_0^{s_P} n(\mathbf{r}(s)) ds, \quad (4)$$

So the OPL is the distance weighted by the refractive index.

Note

Fermat's principle states that the OPL is stationary under small changes to the ray path. It is often, but not always, the shortest OPL.

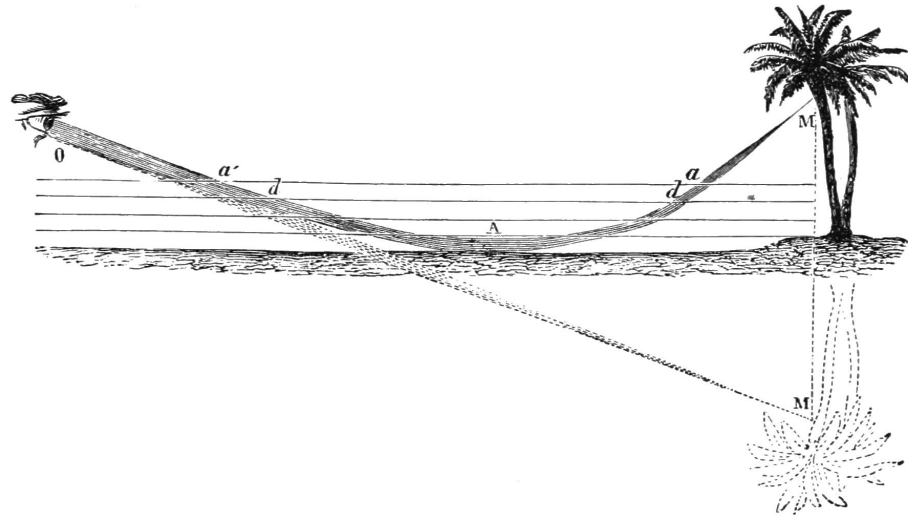


Figure 3: Because the temperature close to the ground is higher, the refractive index is lower there. Therefore the rays bend upwards, creating a mirror image of the tree below the ground. (From Popular Science Monthly Volume 5, Public Domain, [link](#)).

Remark. More than one stationary path can connect two points. For example, a direct ray and a reflected ray can have different travel times. The reflection calculation below minimizes travel time among paths constrained to touch the mirror; it does not compare that path with the direct ray.

3. SOME CONSEQUENCES OF FERMAT'S PRINCIPLE

- **Homogeneous matter**

In homogeneous matter, the refractive index is constant and therefore paths of shortest OPL are straight lines. Hence in homogeneous matter rays are straight lines.

- **Inhomogeneous matter**

When the refractive index is a function of position, as in air with a temperature gradient, the rays bend towards regions of higher refractive index. In the case of [Figure 3](#) for example, on a warm day, the ray from the top of the tree to the eye of the observer passes close to the ground because there the temperature is higher and hence the refractive index is smaller. Although the curved path is longer than the straight path, the total travel time of the light is less because near the ground the light speed is higher (since the refractive index is smaller). The observer gets the impression that the tree is upside down under the ground.

- **Law of reflection**

Consider the mirror shown in [Figure 4](#). In the homogeneous medium above it, a direct ray and a ray reflected by the mirror can connect P and Q . We now minimize travel time over the paths that touch the mirror. Let the x -axis be the intersection of the mirror and the plane through the points P and Q and perpendicular to the

mirror. Let the y -axis be normal to the mirror. Let (x_P, y_P) and (x_Q, y_Q) be the coordinates of P and Q , respectively. If $(x, 0)$ is the point where a ray from P to Q hits the mirror, the travel time of that ray is

$$\frac{n}{c}d_1(x) + \frac{n}{c}d_2(x) = \frac{n}{c}\sqrt{(x - x_P)^2 + y_P^2} + \frac{n}{c}\sqrt{(x_Q - x)^2 + y_Q^2}, \quad (5)$$

where n is the refractive index in $y > 0$. Among paths constrained to reflect at the mirror, the stationary point $(x, 0)$ satisfies

$$\frac{d}{dx}[d_1(x) + d_2(x)] = \frac{(x - x_P)}{d_1(x)} - \frac{(x_Q - x)}{d_2(x)} = 0. \quad (6)$$

Hence

$$\sin \theta_i = \sin \theta_r, \quad (7)$$

or

$$\theta_r = \theta_i. \quad (8)$$

where θ_i and θ_r are the angles of incidence and reflection as shown in [Figure 4](#).

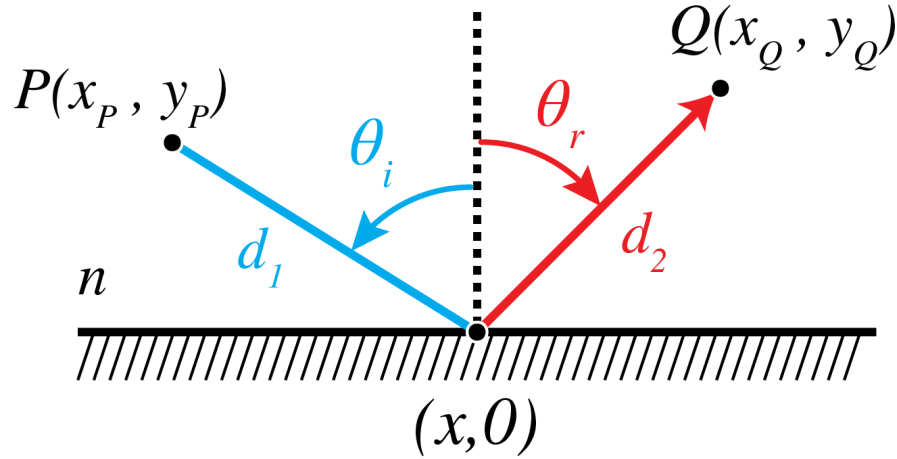


Figure 4: Ray from P to Q via the mirror.

- **Snell's law of refraction**

Next, we consider refraction at an interface. Let $y = 0$ be the interface between a medium with refractive index n_i in $y > 0$ and a medium with refractive index n_t in $y < 0$. We use the same coordinate system as in the case of reflection above. Let (x_P, y_P) and (x_Q, y_Q) with $y_P > 0$ and $y_Q < 0$ be the coordinates of two points P and Q as shown in [Figure 5](#). What path will a ray follow that goes from P to Q ? Since the refractive index is constant in both half spaces, the ray is a straight line in both media. Let $(x, 0)$ be the coordinate of the intersection point of the ray with the interface. Then the travel time is

$$\frac{n_i}{c}d_1(x) + \frac{n_t}{c}d_2(x) = \frac{n_i}{c}\sqrt{(x-x_P)^2 + y_P^2} + \frac{n_t}{c}\sqrt{(x_Q-x)^2 + y_Q^2}. \quad (9)$$

The OPL is stationary with respect to the crossing point, so

$$\frac{d}{dx}[n_i d_1(x) + n_t d_2(x)] = n_i \frac{(x-x_P)}{d_1(x)} - n_t \frac{(x_Q-x)}{d_2(x)} = 0. \quad (10)$$

where the phase delay has been multiplied by the vacuum light speed. Eq. (10) implies

$$n_i \sin \theta_i = n_t \sin \theta_t, \quad (11)$$

where θ_i and θ_t are the angles between the ray and the normal to the surface in the upper half space and the lower half space, respectively (Figure 5).

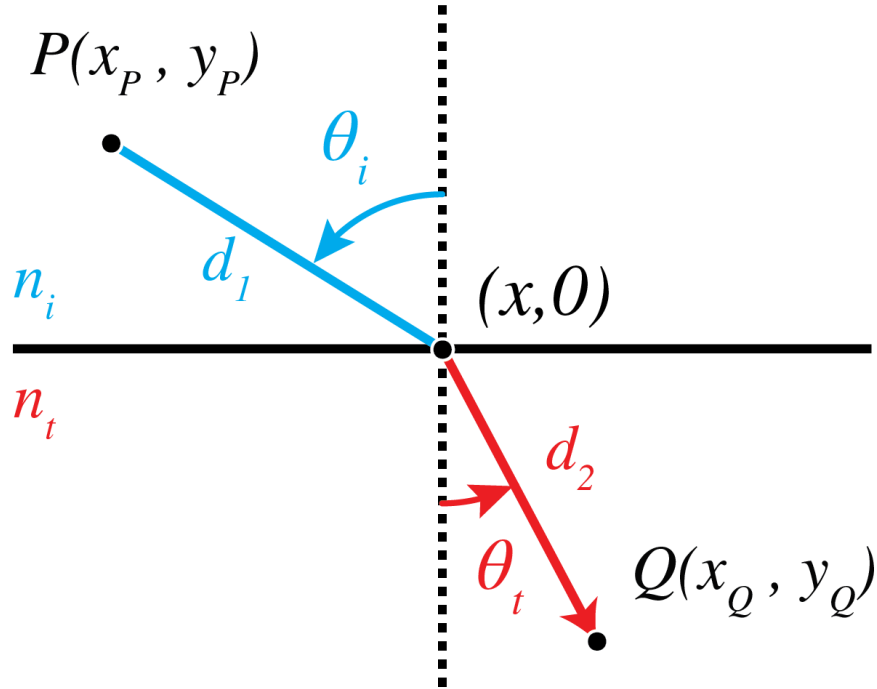


Figure 5: Ray from P to Q refracted by an interface.

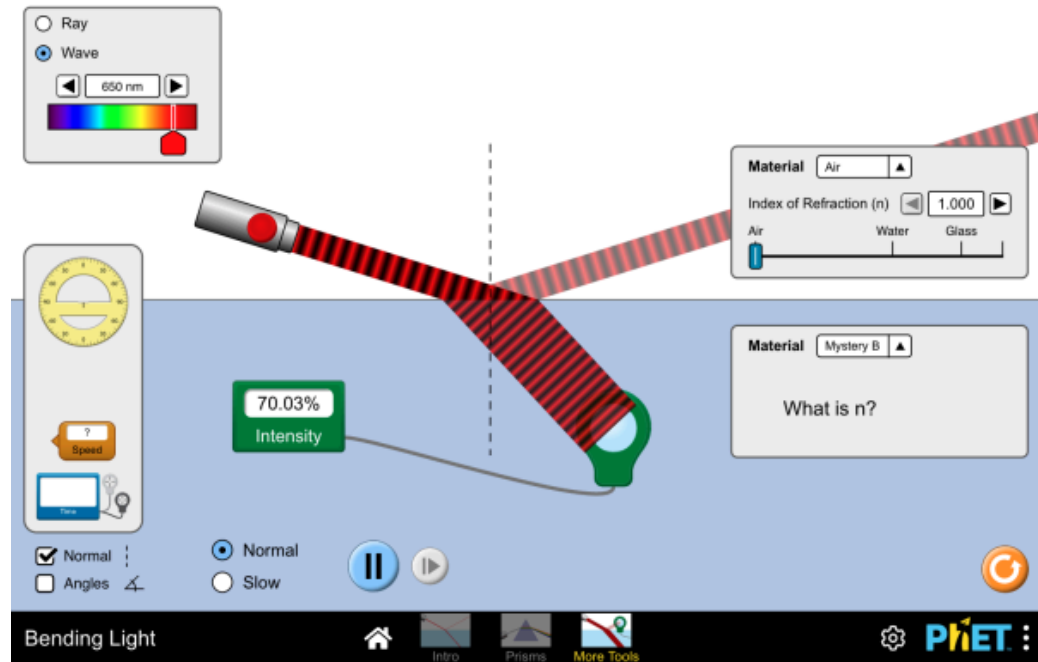


Figure 6: Refraction and reflection at a planar interface with adjustable indices. Measure θ_i and θ_t and check that $n_i \sin \theta_i = n_t \sin \theta_t$ matches the protractor — the same Snell's law just derived from Fermat's principle.

Interactive simulation: [Bending Light](#)

Thus Fermat's principle gives the reflection and refraction laws. Matching the phase of electromagnetic fields along an interface gives an independent derivation of Snell's law.

4. PERFECT IMAGING BY CONIC SECTIONS

In this section, the conic sections ellipse, hyperbola and parabola are important. In [Figure 8](#) their definitions are shown as a quick reminder².

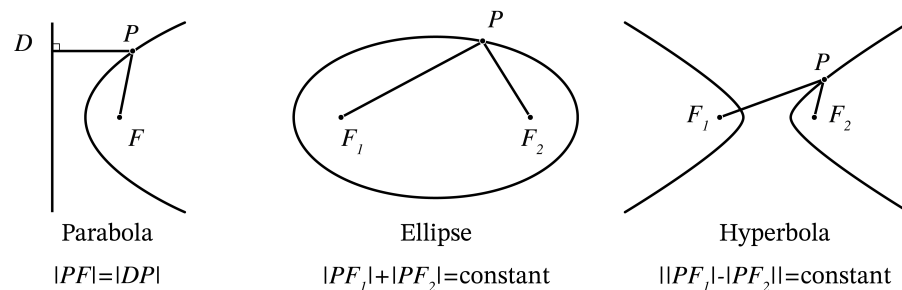


Figure 7: Geometric construction of conic sections (ellipse, hyperbola, and parabola) showing their traditional definitions. The ellipse is the locus of points with constant

²For more details on conic sections and their properties in optics, see Jenkins and White, *Fundamentals of Optics*, Chapter 4.

sum of distances to two foci, the hyperbola has a constant difference of distances, and the parabola maintains equal distances to a focus and a directrix.

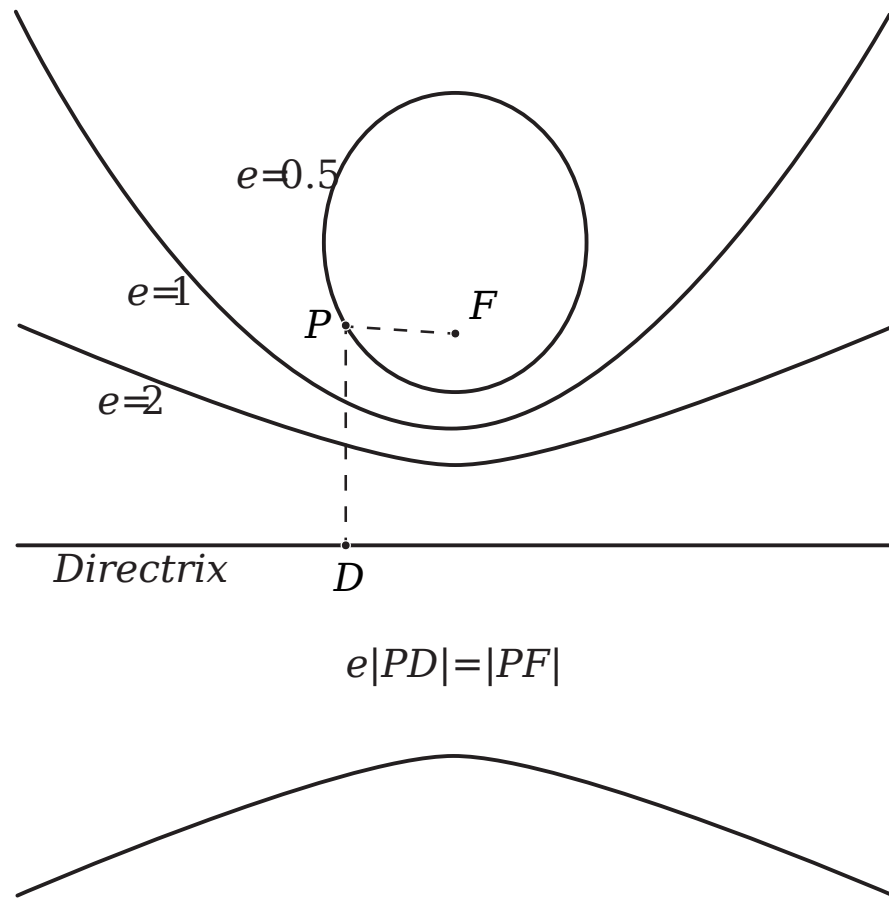


Figure 8: Conic sections through the focus–directrix relation $|PF| = e|PD|$: $e < 1$ gives an ellipse, $e = 1$ a parabola, and $e > 1$ a hyperbola. The horizontal line through D is the common directrix in this drawing.

We start by explaining what in geometrical optics is meant by **perfect imaging**. Let S be a point source. The rays perpendicular to the spherical wave fronts emitted by S radially fan out from S . Due to objects such as lenses etc. the spherical wave fronts are deformed and the directions of the rays are made to deviate from the radial propagation direction. If every ray in a cone from S meets again at P , as in [Figure 9](#), P is a **perfect geometrical image** of S . For an ideal stigmatic system, the wavefront arriving at P also has equal optical phase across the contributing aperture. Fermat's stationarity condition alone does not say that arbitrary rays have equal travel times. By reversing the direction of the rays, S is similarly a perfect image of P . The optical system in which this happens is called **stigmatic for the two points S and P** .

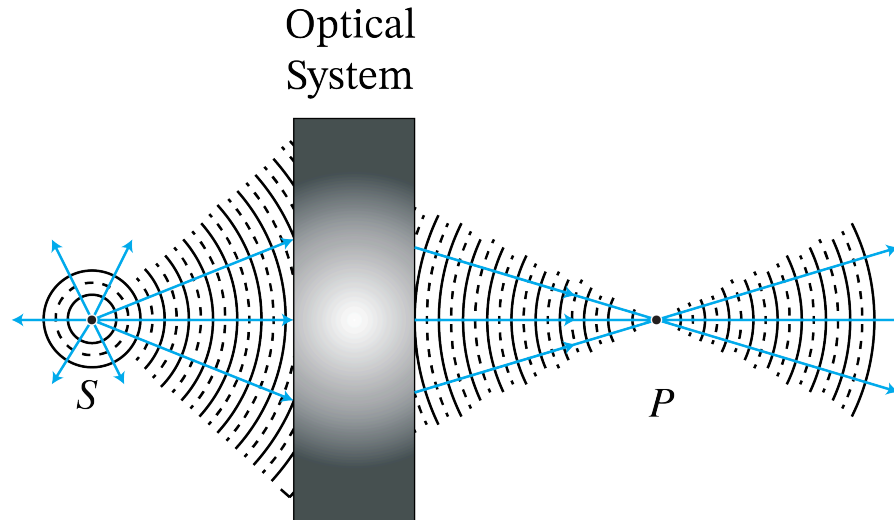


Figure 9: Perfect imaging: a cone of rays which diverge from S and all intersect in point P . The rays continue after P .

Remark. The concept of a perfect image point exists only in geometrical optics. In reality, finite apertures of lenses and other imaging systems cause diffraction due to which image points are never perfect but blurred.

We summarize the main examples of stigmatic systems.

1. Perfect focusing and imaging by refraction. A parallel bundle of rays propagating in a medium with refractive index n_2 can be focused into a point F in a medium with refractive index n_1 . If $n_2 > n_1$, the interface between the media should be a hyperbola with focus F , whereas if $n_2 < n_1$ the interface should be an ellipse with focus F . By reversing the rays we obtain perfect collimation. Therefore, a point S in air can be perfectly imaged onto a point F in air by inserting a piece of glass in between them with hyperbolic surfaces. These properties are derived in Problem 2.2.

2. Perfect focusing of parallel rays by a mirror. A bundle of parallel rays in air can be focused into a point F by a mirror of parabolic shape with F as focus. This is derived in Problem 2.3. By reversing the arrows, we get (within geometrical optics) a perfectly parallel beam. Parabolic mirrors are used everywhere, from automobile headlights to radio telescopes.

Remark.

Although we found that conic surfaces give perfect imaging for a certain pair of points, other points do *not* have perfect images in the sense that for a certain cone of rays, all rays are refracted (or reflected) to the same point.

External sources in recommended order

- [KhanAcademy - Geometrical Optics](#): Playlist on elementary geometrical optics.
- [Yale Courses - 16. Ray or Geometrical Optics I](#) - Lecture by Ramamurti Shankar
- [Yale Courses - 17. Ray or Geometrical Optics II](#) - Lecture by Ramamurti Shankar

5. GAUSSIAN GEOMETRICAL OPTICS

We have seen that, although by using lenses or mirrors which have surfaces that are conic sections we can perfectly image a certain pair of points, for other points the image is not perfect. The imperfections are caused by rays that make larger angles with the **optical axis**, i.e. with the symmetry axis of the system. Rays for which these angles are small are called **paraxial rays**. Because for paraxial rays the angles of incidence and transmission at the surfaces of the lenses are small, the sines of the angles in Snell's Law are replaced by the angles themselves:

$$n_i \theta_i = n_t \theta_t \quad (\text{paraxial rays only}). \quad (12)$$

This approximation greatly simplifies the calculations. At first paraxial order, a smooth rotationally symmetric surface can be replaced by a sphere with the same curvature at its vertex. Higher-order terms determine its aberrations outside that approximation. Spherical surfaces are not only simpler in the derivations but they are also much easier to manufacture. Hence in the optical industry spherical surfaces are used a lot. To reduce imaging errors caused by non-paraxial rays one applies two strategies: 1. adding more spherical surfaces; 2. replacing one of the spherical surfaces (typically the last before image space) by a non-sphere.

Note

In Gaussian geometrical optics only paraxial rays and spherical surfaces are considered. In Gaussian geometrical optics every point has a perfect image.

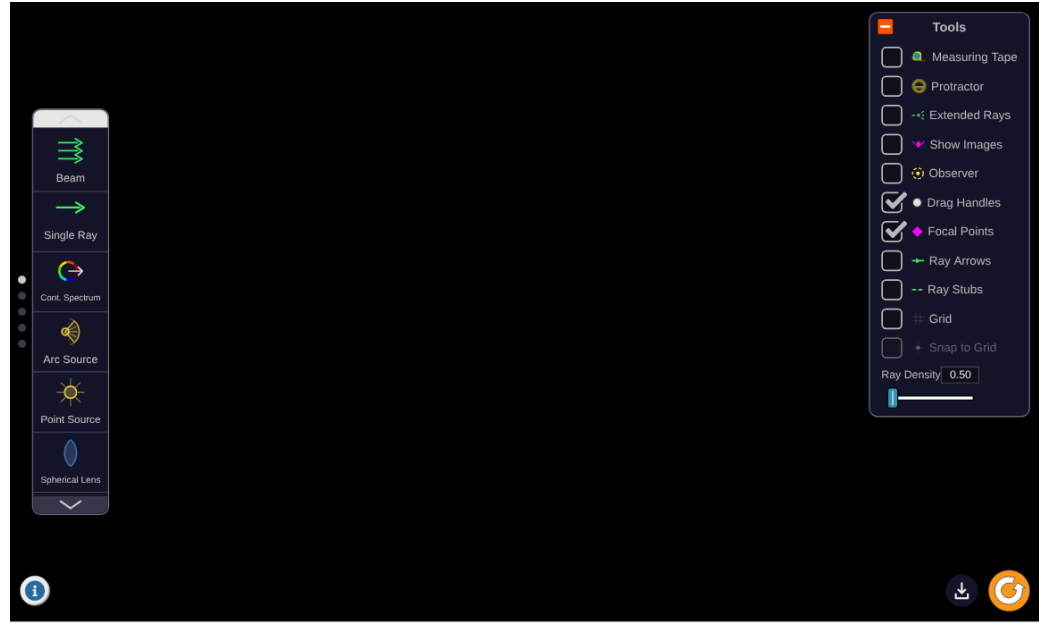


Figure 10: An open optical bench: rays, parallel and divergent beams, lenses, mirrors, and refracting interfaces. Build the single-surface and thin-lens geometries of this section and watch where the rays meet.

Interactive simulation: [Optics Lab](#)

5.a. Gaussian Imaging by a Single Spherical Surface

We will first show that within Gaussian optics a single spherical surface between two media with refractive indices $n_1 < n_2$ images all points perfectly (Figure 11). The sphere has radius R and center C which is inside medium 2. We consider a point object S to the left of the surface. We draw a ray from S perpendicular to the surface. The point of intersection is V . Since for this ray the angle of incidence with the local normal on the surface vanishes, the ray continues into the second medium without refraction and passes through the center C of the sphere. Next we draw a ray that hits the spherical surface at some point A and draw the refracted ray in medium 2 using Snell's law in the paraxial form Eq. (12). Note that the angles of incidence and transmission must be measured with respect to the local normal at A , i.e. with respect to CA . We assume that this ray intersects the first ray at point P . We will show that within the approximation of Gaussian geometrical optics, all rays from S pass through P . Furthermore, with respect to a coordinate system (y, z) with origin at V , the z -axis pointing from V to C and the y -axis positive upwards as shown in Figure 11, we have:

$$-\frac{n_1}{s_o} + \frac{n_2}{s_i} = \text{cal } P, \quad (13)$$

where

$$\text{cal } P = \frac{n_2 - n_1}{R}, \quad (14)$$

is called the power of the surface and where s_o and s_i are the z -coordinates of S and P , respectively, hence $s_o < 0$ and $s_i > 0$ in Figure 11.

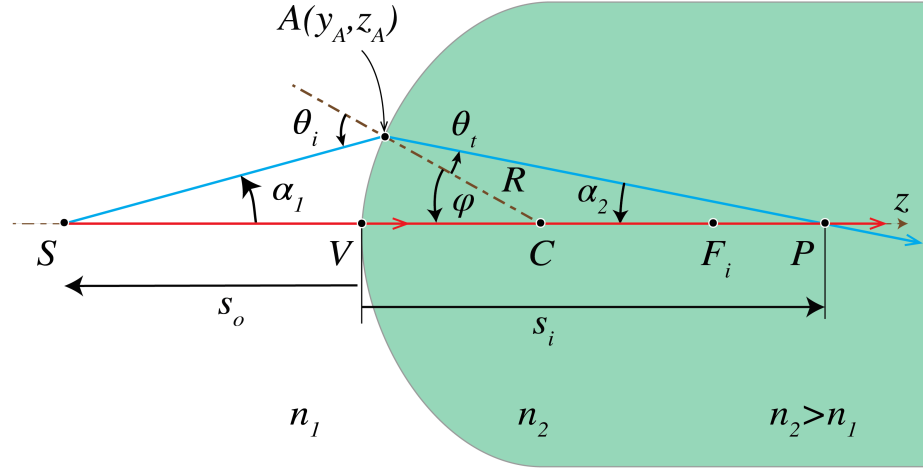


Figure 11: Imaging by a spherical interface between two media with refractive indices $n_2 > n_1$.

Proof.

It suffices to show that P is independent of the ray, i.e. of A . We will do this by expressing s_i in terms of s_o and showing that the result is independent of A . Let α_1 and α_2 be the angles of the rays SA and AP with the z -axis as shown in Figure 11. Let θ_i be the angle of incidence of ray SA with the local normal CA on the surface and θ_t be the angle of refraction. By considering the angles in triangle ΔSCA we find

$$\theta_i = \alpha_1 + \phi. \quad (15)$$

Similarly, from ΔCPA we find

$$\theta_t = -\alpha_2 + \phi. \quad (16)$$

By substitution into the paraxial version of Snell's Law Eq. (12), we obtain

$$n_1 \alpha_1 + n_2 \alpha_2 = (n_2 - n_1) \phi. \quad (17)$$

Let y_A and z_A be the coordinates of point A . Since $s_o < 0$ and $s_i > 0$ we have

$$\alpha_1 \approx \tan(\alpha_1) = \frac{y_A}{z_A - s_o}, \quad \alpha_2 \approx \tan(\alpha_2) = \frac{y_A}{s_i - z_A}. \quad (18)$$

Furthermore,

$$\phi \approx \sin \phi \approx \frac{y_A}{R}. \quad (19)$$

which is small for paraxial rays. Hence,

$$z_A = R - R \cos \phi = R - R \left(1 - \frac{\phi^2}{2} \right) = \frac{R}{2} \phi^2 \approx 0, \quad (20)$$

because it is second order in y_A and therefore is neglected in the paraxial approximation. Then, Eq. (18) becomes

$$\alpha_1 = -\frac{y_A}{s_o}, \quad \alpha_2 = \frac{y_A}{s_i}. \quad (21)$$

By substituting Eq. (21) and Eq. (19) into Eq. (17) we find

$$-\frac{n_1}{s_o}y_A + \frac{n_2}{s_i}y_A = \frac{n_2 - n_1}{R}y_A, \quad (22)$$

or

$$-\frac{n_1}{s_o} + \frac{n_2}{s_i} = \frac{n_2 - n_1}{R}, \quad (23)$$

which is Eq. (13). It implies that s_i , and hence P , is independent of y_A , i.e. of the ray chosen. Therefore, P is a perfect image within the approximation of Gaussian geometrical optics.

When $s_o \rightarrow -\infty$, the incident rays are parallel to the z -axis in medium 1 and the corresponding image point F_i is called the **second focal point** or **image focal point**. Its z -coordinate is given by:

$$f_i = \frac{n_2}{\text{cal } P} = \frac{n_2 R}{n_2 - n_1}, \quad (24)$$

The absolute value of f_i is called the **second focal length** or **image focal length**. For $R > 0$, the signed value f_i is negative when $n_2 < n_1$. When $s_i \rightarrow \infty$, the rays after refraction are parallel to the z -axis and we get $s_o \rightarrow -n_1 R / (n_2 - n_1)$. The object point for which the rays in medium 2 are parallel to the z -axis is called the **first focal point** or **object focal point** F_o . Its z -coordinate is:

$$f_o = -\frac{n_1}{\text{cal } P} = -\frac{n_1 R}{n_2 - n_1}. \quad (25)$$

The absolute value $|f_o|$ of f_o is called the **front focal length** or **object focal length**.

With Eq. (24) and Eq. (25), Eq. (13) can be rewritten as:

$$-\frac{n_1}{s_o} + \frac{n_2}{s_i} = \text{cal } P = \frac{n_2}{f_i} = -\frac{n_1}{f_o}. \quad (26)$$

5.b. Virtual Images and Virtual Objects of a Single Spherical Surface

If we adopt the sign convention listed in Table 1 below, it turns out that Eq. (13) holds generally. So far we have considered a convex surface of which the center C is to the right of the surface, but Eq. (13) applies also to a concave surface of which the center is to the left of the surface, provided that the radius R is chosen negative. The convention for the sign of the radius is illustrated in Figure 12.

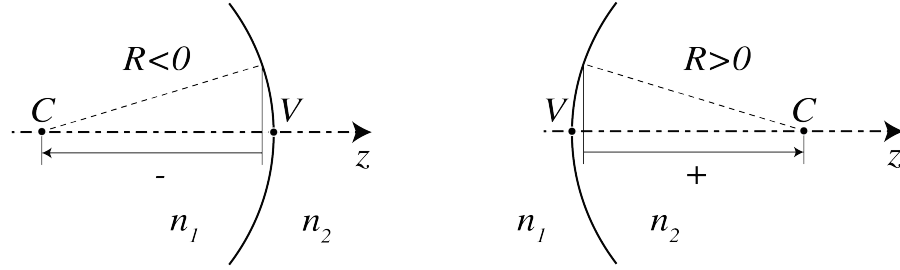


Figure 12: Sign convention for the radius R of a spherical surface

If the power $\mathcal{P}$ given by Eq. (14) is positive, then the surface makes bundles of incident rays convergent or less divergent. If the power is negative, incident bundles are made divergent or less convergent. The power of the surface can be negative for two reasons:

1. $R > 0$ and $n_1 > n_2$, or
2. $R < 0$ and $n_1 < n_2$, but the effect of the two cases is the same. For any object to the left of the surface: $s_o < 0$, Eq. (26) and a negative power imply that $s_i < 0$, which suggests that the image is to the left of the surface. Indeed, in both cases, the diverging ray bundle emitted by S is made more strongly divergent by the surface. By extending these rays in image space back to object space (without refraction at the surface), they are seen to intersect in a point P to the left of the surface. This implies that for an observer at the right of the surface it looks as if the diverging rays in image space are emitted by P . Because there is no actual concentration of light intensity at P , it is called a **virtual image**, in contrast with the **real images** that occur to the right of the surface and where there is an actual concentration of light energy. We have in this case $f_o > 0$ and $f_i < 0$, which means that the object and image focal points are to the right and left, respectively, of the surface.

Note that also when the power is positive, a virtual image can occur, namely when the object S is in between the object focal point F_o and the surface. Then the bundle of rays from S is so strongly diverging that the surface cannot convert it into a convergent bundle and hence again the rays in image space seem to come from a point P to the left of the surface. This agrees with the fact that when $\mathcal{P} > 0$ and $f_o < s_o < 0$, Eq. (26) implies that $s_i < 0$.

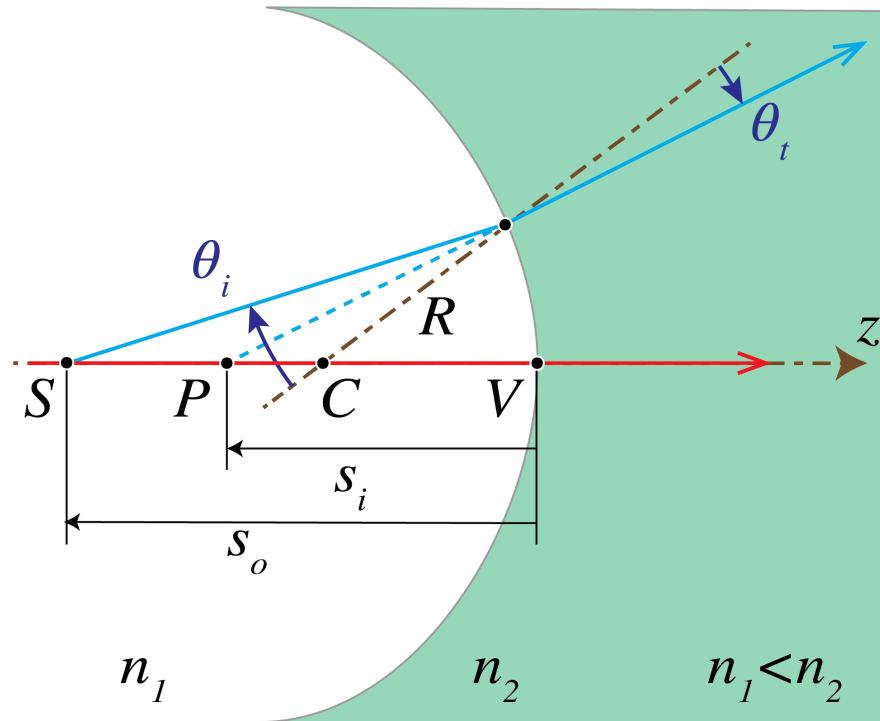


Figure 13: Imaging by a concave surface ($R < 0$) with $n_2 > n_1$. All image points are to the left of the surface, i.e. are virtual ($s_i < 0$).

Finally we look at a case in which there is a bundle of convergent rays incident on the surface from the left. If extended into the right medium without refraction, these rays would intersect in a point S . Since this point is not actually present, it is called a **virtual object point**, in contrast to **real object points** which are to the left of the surface. The coordinate of a virtual object point is positive: $s_o > 0$. One may wonder why we look at this case. The reason is that if we have several spherical surfaces behind each other, we can compute the image of an object point by first determining the intermediate image by the leftmost surface and then use this intermediate image as the object for the next surface and so on. In such a case, it can easily happen that an intermediate image is to the right of the next surface and hence is a virtual object for that surface. In the case of Figure 14 at the left, the power is positive, hence the convergent bundle of incident rays is made even more convergent which leads to a real image point. Indeed when $s_o > 0$ and $\text{cal } P > 0$ then Eq. (13) implies that always $s_i > 0$. At the right of Figure 14 the power is negative but is not sufficiently strong to turn the convergent incident bundle into a divergent bundle. So the image is still real. However, the image will be virtual when the virtual object S is to the right of F_o (which in this case is to the right of the surface) since then the bundle of rays converges so weakly that the surface turns it into a divergent bundle.

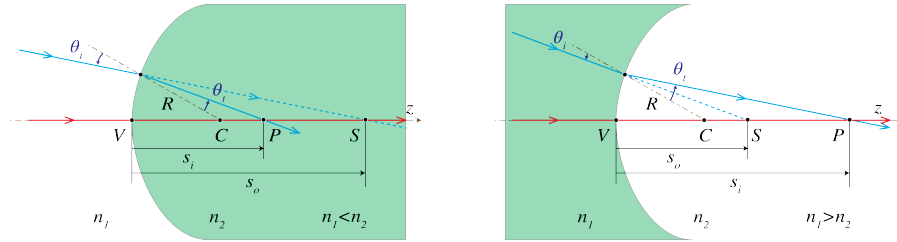


Figure 14: Imaging of a virtual object S by a spherical interface with $R > 0$ between two media with refractive indices $n_1 > n_2$ (left) and $n_2 > n_1$ (right).

In conclusion: provided the sign convention listed in Table 1 is used, formula Eq. (13) can always be used to determine the image of a given object by a spherical surface.

The convention for s_o, f_o, s_i, f_i follows from the fact that these are z -coordinates with the origin at vertex V of the spherical surface (or the center of the thin lens) and the positive z -axis is pointing to the right. The convention for the y -coordinate follows from the fact that the y -axis is positive upwards.

quantity	positive	negative
s_o, s_i, f_o, f_i	corresponding point is to the right of vertex	corresponding point is to the left of vertex
y_o, y_i	object, image point above optical axis	object, image point below optical axis
R	center of curvature right of vertex	center of curvature left of vertex

Table 1: Sign convention for spherical surfaces and thin lenses

5.c. Stops

An element such as the rim of a lens or a diaphragm which determines the set of rays that can contribute to the image is called the **aperture stop**. An ordinary camera has a variable diaphragm.

The **entrance pupil** is the image of the aperture stop by all elements to the left of the aperture stop. In constructing the entrance pupil, rays are used which propagate from the right to the left. The image can be real or virtual. If there are no lenses between object and aperture stop, the aperture stop itself is the entrance pupil. Similarly, the **exit pupil** is the image of the aperture stop by all elements to the right of it. This image can be real or virtual. The entrance pupil determines for a given object the cone of rays in object space that contribute to the image, while the cone of rays leaving the exit pupil is the one taking part in image formation (see Figure 15).

For any object point, the **chief ray** is the ray in the cone that passes through the center of the entrance pupil, and hence also through the centers of the aperture stop and the exit pupil. A **marginal ray** is the ray that for an object point on the optical axis passes through the rim of the entrance pupil (and hence also through the rims of the aperture stop and the exit pupil).

The entrance pupil controls the cone of light accepted from an object point. For a camera focused on a distant scene, the nominal ***f*-number** is the effective focal length f divided by the entrance-pupil diameter D_{entrance} :

$$f\text{-number} = \frac{f}{D_{\text{entrance}}}. \quad (27)$$

For example, $f/2$ means $f = 2D_{\text{entrance}}$. At the same transmission and scene conditions, irradiance is approximately inversely proportional to the square of the f -number, so $f/1.4$ needs about half the exposure time of $f/2$. Finite-distance imaging and transmission losses can modify this simple comparison.

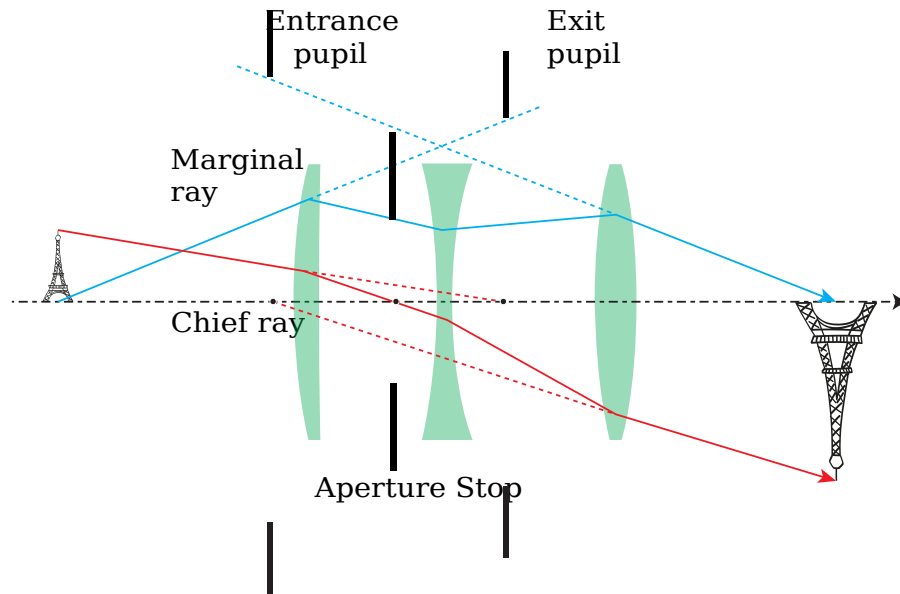


Figure 15: Aperture stop between the first and second lenses. The entrance pupil is the stop's image seen from object space, and the exit pupil is its image seen from image space. The chief and marginal rays are shown.

6. BEYOND GAUSSIAN GEOMETRICAL OPTICS

6.a. Aberrations

For designing advanced optical systems Gaussian geometrical optics is not sufficient. Instead non-paraxial rays, and among them also non-meridional rays, must be traced using software based on Snell's Law with the sine of the angles of incidence and refraction. Often many thousands of rays are traced to evaluate the quality of an image. It is then found that in general the non-paraxial rays do not intersect at the ideal Gaussian image point. Instead of a single spot, a spot diagram is found which is more or less confined. The deviation from an ideal point image is quantified in terms of **aberrations**. One distinguishes between monochromatic and chromatic aberrations. The latter are caused by the fact that the refractive index depends on wavelength. Recall that in paraxial geometrical optics Snell's Law [Eq. \(11\)](#) is replaced by: $n_i \theta_i = n_t \theta_t$, i.e. $\sin \theta_i$ and $\sin \theta_t$ are replaced by the linear terms. If instead one

retains the first two terms of the Taylor series of the sine, the errors in the image can be quantified by five monochromatic **primary (Seidel) aberrations**. The best known is **spherical aberration**, which is caused by the fact that for a convergent spherical lens, the rays that make a large angle with the optical axis are focused closer to the lens than the paraxial rays (see [Figure 16](#)).

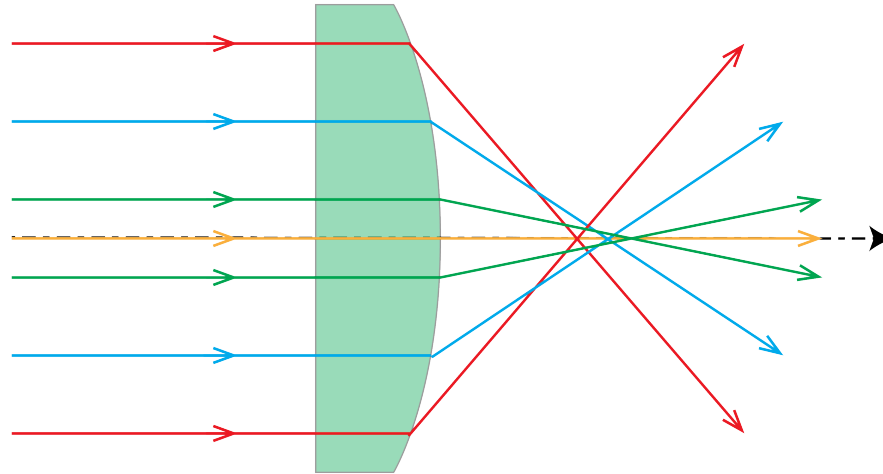


Figure 16: Spherical aberration of a planar-convex lens.

Distortion is one of the five primary aberrations. It causes deformation of images due to the fact that the magnification depends on the distance of the object point to the optical axis.

For high-quality imaging, the aberrations have to be reduced by adding more lenses and optimizing the curvatures of the surfaces, the thicknesses of the lenses and the distances between them. For high-quality systems, a lens with an aspherical surface is sometimes used. Systems with very small aberrations are extremely expensive, in particular if the field of view is large, as is the case in lithographic imaging systems used in the manufacturing of integrated circuits as illustrated by the lithographic system in [Figure 17](#).

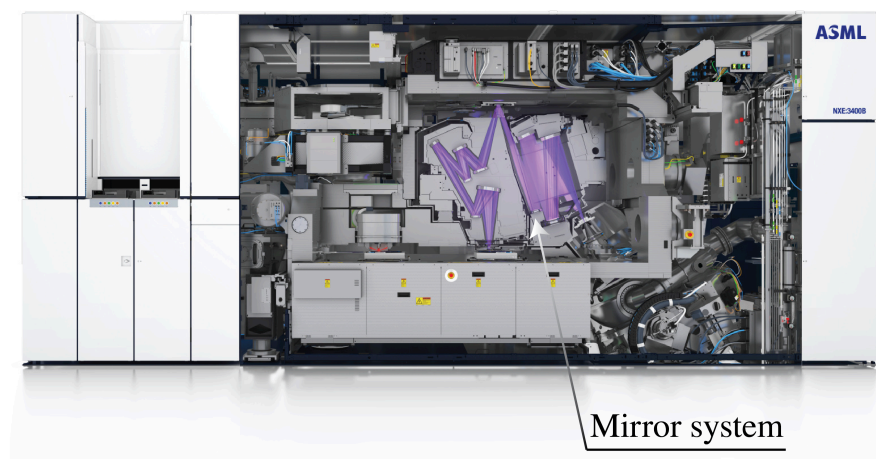


Figure 17: ASML’s TWINSCAN NXE:3400B extreme-ultraviolet lithography system. The purple overlay traces its reflective projection optics; EUV systems use 13.5 nm light and mirrors rather than transmissive lenses ([ASML product information](#)).

6.b. Diffraction

According to a generally accepted criterion formulated first by Rayleigh, aberrations start to deteriorate images considerably if they cause path length differences of more than a quarter of the wavelength. When the aberrations are less than this, the system is called **diffraction limited**.

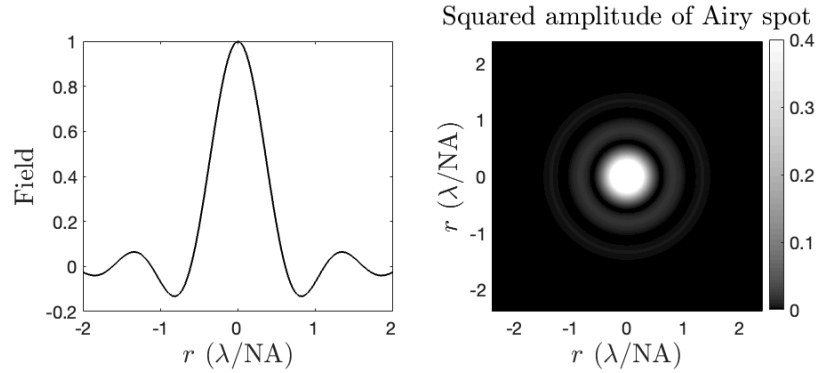


Figure 18: Left: cross section of the field of the Airy pattern. Right: intensity pattern of the Airy pattern.

Even if the wave transmitted by the exit pupil were perfectly spherical (no aberrations), its finite aperture causes diffraction. A point object therefore forms an Airy pattern rather than a point ([Figure 18](#)). For an ideal uniformly illuminated circular pupil, the central intensity spot has approximate full width at half maximum:

$$\text{FWHM} \approx 0.514 \frac{\lambda_0}{\text{NA}}, \quad (28)$$

where λ_0 is the vacuum wavelength and $\text{NA} = n_{\text{image}} \sin \theta$ is the image-space numerical aperture. Here θ is the half-angle of the cone from the exit pupil to the image point; in air and the paraxial limit, $\text{NA} \approx a/L$ for exit-pupil radius a and distance L from that pupil to the image. Diffraction depends on wavelength and therefore lies beyond pure geometrical optics. We treat aperture diffraction in [Scalar Diffraction Optics](#).

7. CHAPTER SUMMARY

- **Fermat’s principle** makes optical path length stationary under small changes to a ray path.
- In **Gaussian geometrical optics**, the paraxial (small angle) approximation enables analytical treatment of imaging by spherical surfaces.
- For a **single spherical interface**, $-n_1/s_o + n_2/s_i = (n_2 - n_1)/R$ with the signed coordinates used here.
- **Real images** form where rays converge; **virtual images** form where rays appear to originate.

- **Spherical aberration** limits imaging quality for rays far from the optical axis; it can be minimized using aspherical surfaces or conic sections.
- **Stops and pupils:** The aperture stop limits the light cone; entrance and exit pupils are its images as seen from object and image space.
- The **f -number** is effective focal length divided by entrance-pupil diameter.
- Geometrical optics breaks down when feature sizes approach the wavelength, leading to **diffraction** effects described by the Airy pattern.