Optics

Haydn Cheng

May 6, 2025

Contents

1 Geometrical Optics

1.1 Basic Principles

1.1.1 Fermat's Principle

The Fermat's principle states that the paths between two points taken by a beam of light are the ones corresponding to stationary travel times. In fact, Fermat's principle is merely a special case of the least action principle applied to the motion of light. Light take all possible paths between points A and B, but only those with the similar phase construtively interfere.

If a photon leaves point A with internal phase φ_0, then it also arrives B with the same internal phase φ_0. Suppose two photons emitted from A at different times t_1 and t_2 interfere at B simultaneously, then the phase difference when they interfere is $\Delta\varphi = \varphi_{02} - \varphi_{01} = \omega(t_2 - t_1) = \omega\Delta t$. For the phase difference to vanish, we require $\Delta t = 0$, so the time travels is stationary.

1.1.2 Huygens' Principle

The Huygens' principle states that every point on a wavefront can be considered the source of spherical wvaelets with the same wavelength and propagating in the same direction of the original wavefront. The evelope of these wavelets then forms the new wvaefront.

1.2 Imaging Systems

Imaging systems are devices that aim to produce an image of an object. For this to happen we need that all the rays of light leaving from a single point O in the object arrive at a single point I in the image, whatever their point of entry M in the device. To image an actual object, this requirement must hold for every object point and its conjugate image point. By the principle of reversibility, O is the image point of the object point I if light travels back in time.

Since Fermat's principle tells us that light always takes the fastest path from O to I, this means that the optical path of all rays emerging from O and arriving at I must be

the same and equal to an extrema. This is often the way to find the shape of an optical instrument.

In almost all of geometrical optics problems, we employ the paraxial approximation, which assumes that all light rays are nearly parallel to the optical axis, which implies that the angles formed by the light rays with the optical axis is always small enough that the small angle approximations are valid.

Nonideal images are formed in practice becuase of light scattering, due to unwanted reflections and refractions, abbrrations, due to the small angle approximation no longer holding, and diffractions, due to the system only intercepting a portion of the wavefronts emerging from the object.

1.2.1 Sign Conventions

When discussing spherical mirrors and spherical surfaces, we will see that we can subdivide each of them cases. However, only one equation is needed for one case. This is because we adopt the following sign convention to embed the negative signs into distances:

1. The object distance u is positive if the object point is on the same side as the incident light.
2. The image distance v is positive if the image point is on the same side as the outgoing light (this implies a real object).
3. The radius R is positive if the center of curvature is on the same side as the outgoing light.

An easy way to remember to remember the convention of the object distance and the image distance is that they are positive if the object point and the image point are where they are supposed to be at.

1.2.2 Spherical Reflecting Surfaces

Referring to fig. 1.1,[1] there are two possibilities in which light can reflect at a spherical mirror.

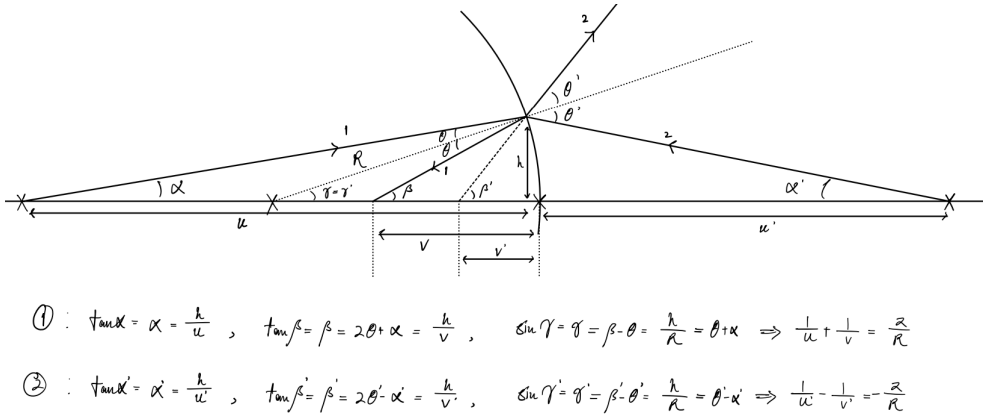


Figure 1.1

[1] Note that the arrows drawn in the figure may be reversed due to light's reversibility.

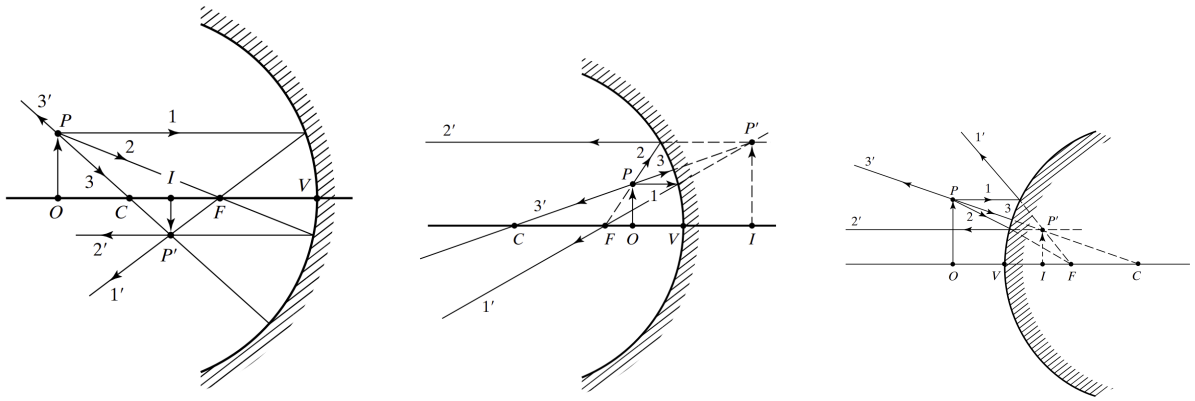
The first possibility, labelled as path 1, is that the object (at this stage we assume that it is real) is situated inside the sphere, which creates a real object. The second possibility, labelled as path 2, is that the object is situated outside the sphere, which would create a virtual object.

However, using the sign convention defined above, we can generalized the two equations governing the two cases as

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R} \equiv \frac{1}{f}. \tag{1.1}$$

The quantity $2/R$ is denoted as $1/f$, since this is universal for all u and v, and defines the image point for an infinity far object.

Figure 1.2 shows the ray diagram for an object reflecting off a spherical surface, the case for a concave lens can be constructed simply by reversing the role of object and image.



(a) (b) (c)

Figure 1.2

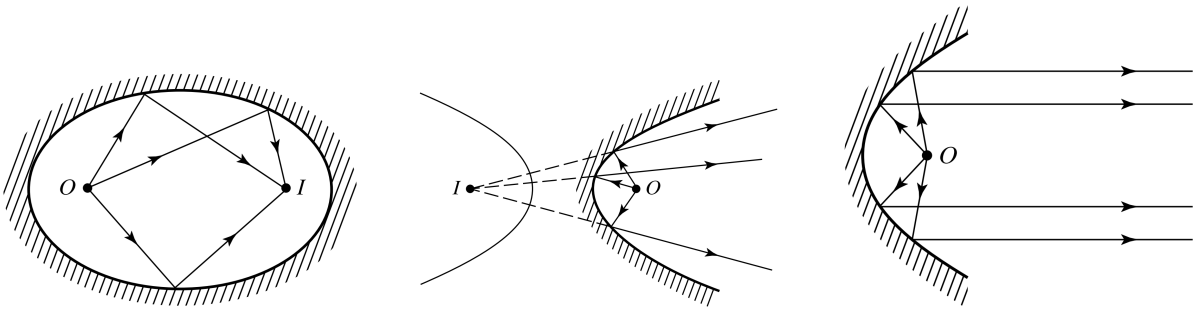
The case where the object distance u is negative, which only occurs when there are more than two imaging systems, can be regarded as if u is real but v and R are negative, which corresponds to switching between the first case and the second case mentioned above. Therefore, it is equivalent as if there is a real object situated at the opposite side of the mirror with the same distance from the mirror. This implies that we can simply treat the image point of the first imaging system as the object point of the second imaging system.

To avoid spherical abberations for large angles, one find, from Fermat's principle, that the exact shape of a perfect reflecting surface is one of the conic sections, as shown in fig. 1.3

Note, however, that the aberration-free imaging so achieved applies only to object point O at the correct distance from the lens and on axis. For nearby points, closer or further from the lens, or off the axis, is not imaged perfectly.

The linear magnification is given by

$$m = \frac{h_i}{h_o} = -\frac{v}{u}. \tag{1.2}$$



(a) (b) (c)

Figure 1.3

The image formation for a spherical mirror is summarized in fig. 1.4, where R and V refer to real and virtual, and O and I refers to object and image.

(a) (b)

Figure 1.4

1.2.3 Spherical Refracting Surfaces

Referring to fig. 1.5,[2] there are two possibilities[3] in which light can refract at the spherical surfaces.

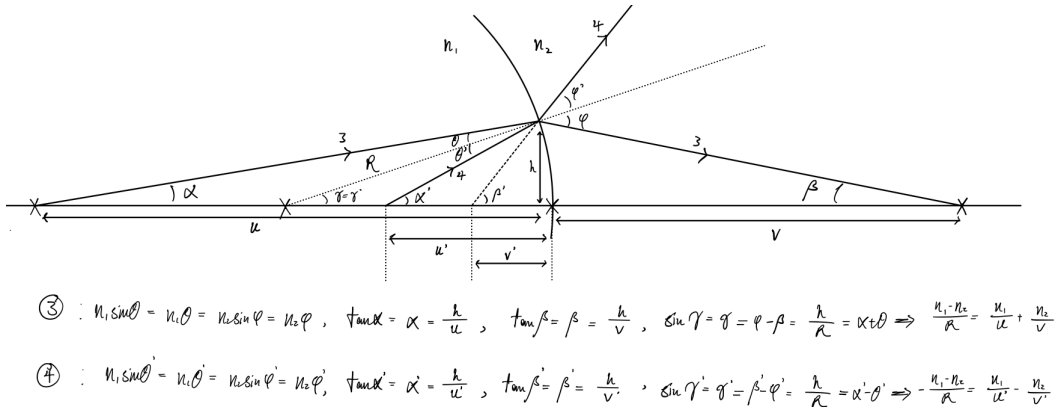


Figure 1.5

[2]See footnote under section 1.2.2.

[3]There are actually four, the other two being if the object is situated at the opposite side of the spherical surface compared to what is shown above. But we have already shown in section 1.2.2 that we can simply add a negative sign to the radius to account for these cases, so we focus our disccusion to the two cases both having their object point inside the sphere.

The first possibility, labelled as path 3, is that the object (at this stage we assume that it is real) is situated behind the center of curvature, which creates a real object. The second possibility, labelled as path 4, is that the object is situated closer to the surface than the center of curvature, which creates a virtual object.

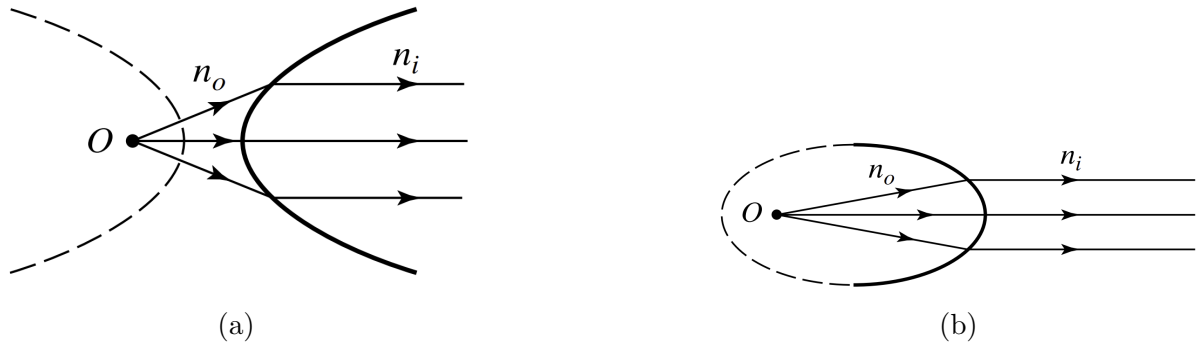
However, using the sign convention defined above, we can generalized the equations governing the cases as

$$\frac{n_1}{u} + \frac{n_2}{v} = \frac{n_1 - n_2}{R}. \tag{1.3}$$

The quantity $(n_1 - n_2)/R$ is universal for all u and v, and defines the image point for an infinity far object, but has no special name for it.

To avoid spherical abberations for large angles, one find, from Fermat's principle, that the eaxt shpe of a perfect refracting surface is the Cartesian ovoid of revolution. In most cases, however, we want to construct a lens that has two spherical surfaces and refracts light rays twice to produce a real image outside of the lens. Thus it is of particular interest to determine the shape of the surface that render every object ray parallel after the first refraction, which is either a hyperbolic or an ellipse depending on the relative magnitudes of the refractive indices, as shown in fig. 1.6.

Note, however, that the aberration-free imaging so achieved applies only to object point O at the correct distance from the lens and on axis. For nearby points, closer or further from the lens, or off the axis, is not imaged perfectly.



(a) (b)

Figure 1.6

The linear magnification is given by

$$m = \frac{h_i}{h_o} = -\frac{n_1 v}{n_2 u}. \tag{1.4}$$

The image formation for a thin lens is summarized in fig. 1.7, where R and V refer to real and virtual, and O and I refers to object and image.

1.2.4 Thin Lens

Lensmaker's Equation

From the two fundamental building blocks, we can construct a lens with two sphereical surfaces. Treating the image of the first surface as the object of the second surface, we

(a) (b)

Figure 1.7

can derive the lensmaker's equation

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{1.5}$$

Thin Lens Equation

Suppose our thin lens has thickness $d(z) \ll u, v$, which depends on height z. The paraxial approximation tells us that $z \ll u, v$ and the path length of light travelled inside the lens is simply $d(z)$. Therefore, using Fermat's principle, we have

$$\begin{aligned} L(z) &= \sqrt{\left(u - \frac{d(z)}{2}\right)^2 + z^2} + nd(z) + \sqrt{\left(v - \frac{d(z)}{2}\right)^2 + z^2} \\ &\approx \left(u - \frac{d(z)}{2}\right)\left(1 + \frac{z^2}{2\left(u - \frac{d(z)}{2}\right)^2}\right) + nd(z) + \left(v - \frac{d(z)}{2}\right)\left(1 + \frac{z^2}{2\left(v - \frac{d(z)}{2}\right)^2}\right) \\ &= u + v + (n-1)d(z) + \frac{z^2}{2}\left(\frac{1}{\left(u - \frac{d(z)}{2}\right)} + \frac{1}{\left(v - \frac{d(z)}{2}\right)}\right) \\ &\approx u + v + (n-1)d(z) + \frac{z^2}{2}\left(\frac{1}{u} + \frac{1}{v}\right) = \text{constant}. \end{aligned} \tag{1.6}$$

Upon differentiation,

$$\begin{aligned} \frac{d}{dz}d(z) &= -\frac{z}{n-1}\left(\frac{1}{u} + \frac{1}{v}\right) \\ \implies d(z) &= d(0) - \frac{z^2}{2(n-1)}\left(\frac{1}{u} + \frac{1}{v}\right). \end{aligned} \tag{1.7}$$

For the lens to work with arbitrary u, v, we require that

$$\frac{1}{u} + \frac{1}{v} = \text{constant} \equiv \frac{1}{f}. \tag{1.8}$$

One can introudce the variables $x = u - f$ and $x' = v - f$, and rewrite the thin lens equation into the Newtonian form

$$xx' = f^2. \tag{1.9}$$

A way to reduce spherical aberrations is to combine convex and concave lens such that the spherical aberration from one tends to cancels that from the other. One can also simply block the light with large incidence angle but this would reduce the image brightness.

Chromatic and Coma Aberration

Chromatic aberration occurs since the refraction index of a material differs for different wavelengths, so light with different wavelengths will focus at different points (logitudinal chromatic aberration) and have different magnifications (lateral chromatic aberration), leading to color frinages and blurred white-light image.

To resolve this issue two lens of opposite powers are often used. The effective focal length f of two thin lenses with focal lengths f_1 and f_2, separated by a distance L is give by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{L}{f_1 f_2} = (n-1)(K_1 + K_2) - L(n-1)^2 K_1 K_2, \tag{1.10}$$

where K_1 and K_2 are quantities related to the radii of curvatures of the lens surfaces.

To correct for chromatic aberration,[4] we require that the effective focal length remain independent of refractive index

$$\frac{d(1/f)}{dn} = K_1 + K_2 - 2LK_1K_2(n-1) = 0 \implies L = \frac{1}{2}(f_1 + f_2). \tag{1.11}$$

Coma aberration refers to when the image of an off-axis point object appears with a comet-like flare, especially for rays where the paraxial approximation breaks down.

chromatic resolving power pedrotti

1.2.5 Plano-Cylindrical Lens

Refer to fig. 1.8a, we see that focusing occurs for rays along a vertical section but not for rays along a horizontal section, where the lens represents no curvature. A point object at infinity thus forms a horizontal line instead of a point at the focus. Figure 1.8b shows the case where the object distance is finite.

Consider the non-focusing light rays at the edge of the lens we can relate the effective length of the lens and the length of the line image, as

$$AB = \left(\frac{u+v}{u}\right) CL. \tag{1.12}$$

[4] Only the transverse chromatic aberration can be fixed but not the longitudinal since the principla planes of the system do not coincide.

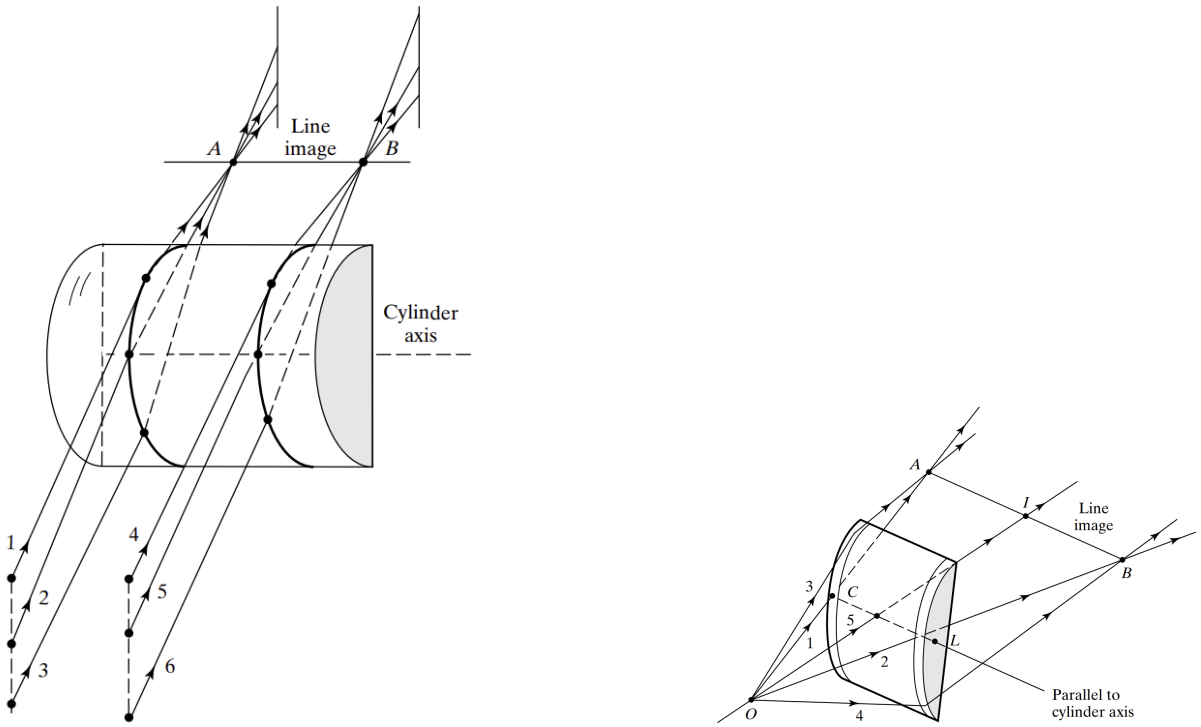


Figure 1.8

1.3 Applications

1.3.1 Terminology

Aperture Stops and Pupils

Aperture stop (AS) is the physical element in an optical system that limits the size of the light cone entering the system from any object point on the optical axis. This is usually the first optical element, but in some cases, such as in fig. 1.9, if the object is placed close enough to the lens then at some point the lens becomes the aperature stop of the system. As another example, in fig. 1.10, the aperature stop is placed behind the lens, which limits the size of the maximum cone of rays.

Entrance pupil (ENP) is the image of the aperture stop as seen from the object side of the system, formed by the optical elements positioned before the aperture stop. If aperture stop is the first element of the optical system then it is also the entrance pupil by this definition.

Exit pupil (EXP) is the image of the aperture stop formed by the optical elements that follow the aperture stop, as seen from the image side.

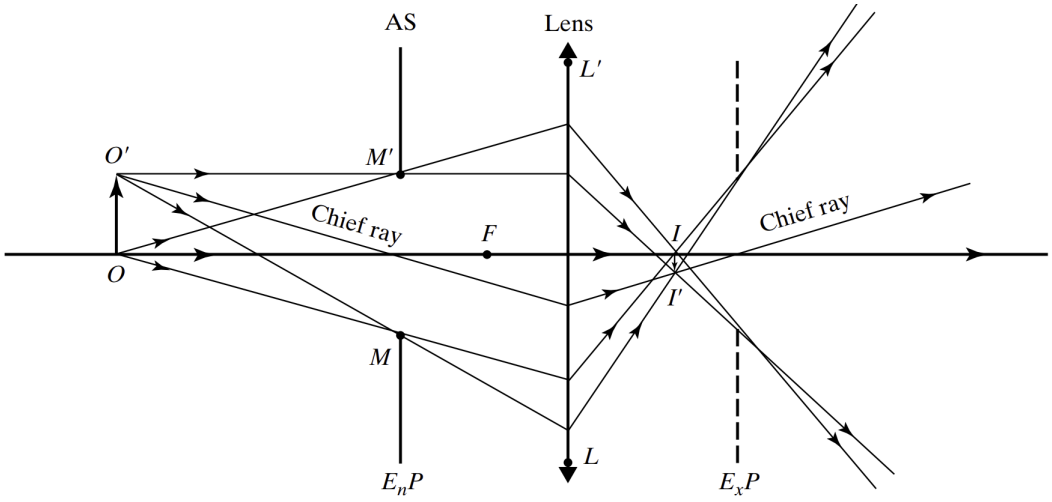


Figure 1.9

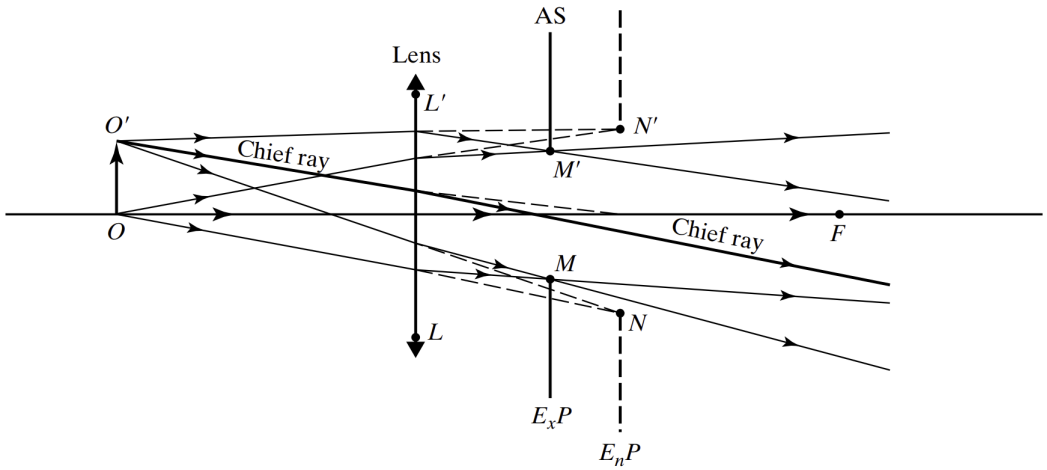


Figure 1.10

The aperture stop of an optical system is not always obvious. To find the aperature stop in fig. 1.11, we have to compare the size of the entrance pupil assuming that the candidate component is the aperture stop, and we have the aperature A is the aperture stop of the system since the rim of its image subtends the smallest angle at O.

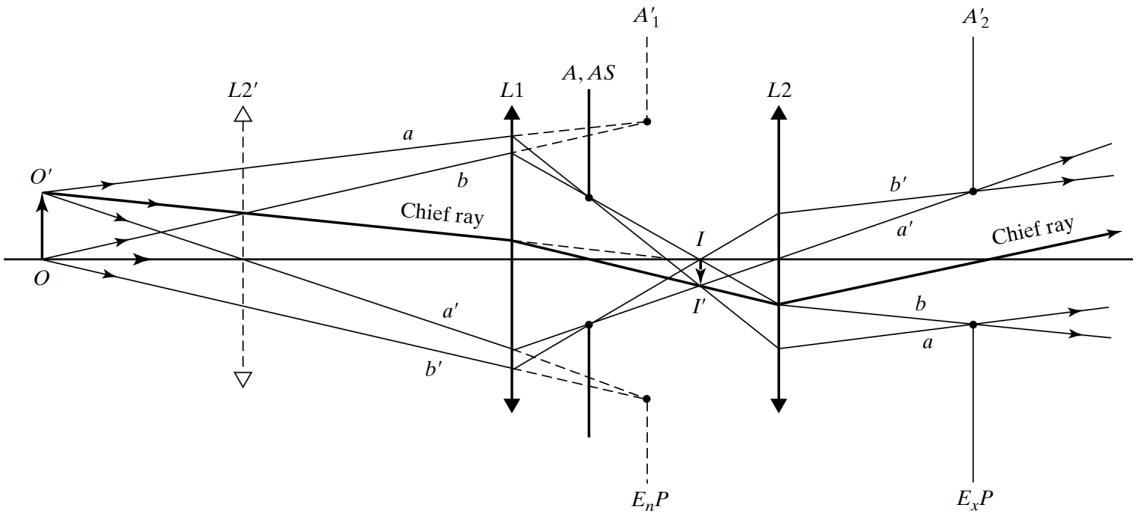


Figure 1.11

Chief Ray and Marginal Ray

Chief ray is the ray from an object point that passes through the axial point in the plane of the entrance pupil, while marginal ray is that passes through the rim of the entrance pupil.

Field Stops and Windows

Field stop (FS) is the physical stop positioned in the optical system that limits the angular field of view, which restricts the extent of off-axis light rays that contribute to the formation of the image. In figs. 1.12 and 1.13, for example, only a portion of light ray from points below T can reachs the lens, and points below V cannot reach the lens at all, so the image formed will be dimmer as we approach the edge.

Entrance window (ENW) is the image of the field stop formed by the optical elements placed before the field stop.

Exit window (EXW) is the image of the field stop formed by the optical elements that come after the field stop.

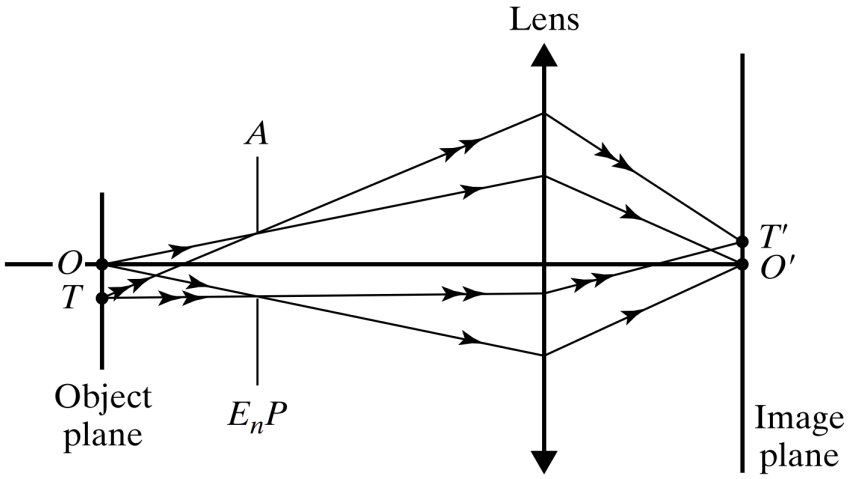


Figure 1.12

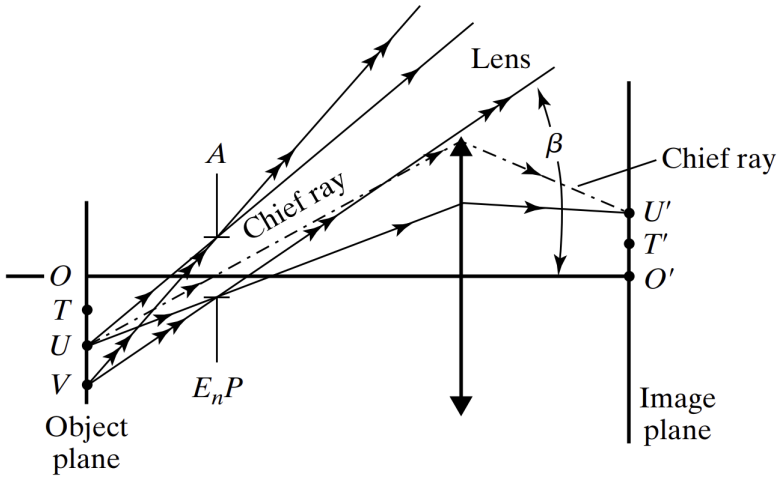


Figure 1.13

Figure 1.14 shows the general situation with all the components we have discussed above.

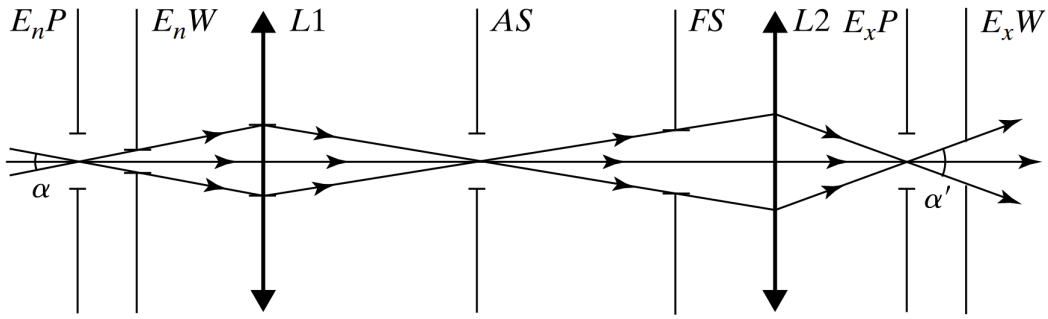


Figure 1.14

Example: Terminology.

Question: An optical system (see fig. 1.15) is made up of a positive thin lens L_1 of diameter 6 cm and focal length $f_1 = 6\,\text{cm}$, a negative lens L_2 of diameter 6 cm and focal length $f_2 = -10\,\text{cm}$, and an aperture A of diameter 3 cm. The aperture A is located 3 cm in front of lens L_1, which is located 4 cm in front of lens L_2. An object OP of height 3 cm is located 18 cm to the left of L_1.

1. Determine which element (A, L_1 or L_2) serves as the aperture stop.
2. Determine the size and location of the entrance and exit pupils.
3. Determine the location and size of the intermediate image of OP formed by L_1 and the final image formed by the system.
4. Draw a diagram of the optical system and locate the two pupils, intermediate image $O'P'$ and the final image $O''P''$.
5. Draw the chief ray from object point P to its conjugate in the final image, P".

Solution:

1. Elements A and L_1 have no "optics" to their left, so each subtends a half-angle directly. For element A, this is $\theta_A = 1.5/15 = 0.1\,\text{rad}$. For element L_1, this is $\theta_{L_1} = 3/18 = 0.17\,\text{rad}$.
   For L_2, we have to first find its image through L_1 by the thin lens equation, which is found to be 12 cm to the right of L_1, so the angle subtended is $\theta_{L_2} = 9/30 = 0.3\,\text{rad}$.
   Comparing the half-angles, we have the element A serving the aperture stop.
2. Since there are no optics to the left of the aperture stop, element A also serves as the entrance pupil. To locate the exit pupil, we find the image of A through L_1 and L_2 as usual, which gives the exit pupil locating 5 cm to the left of L_2. The size is found by multiplying the magnifications at L_1 and L_2 by its original size, which is found to be unchanged, *i.e.,* 3 cm.
3. Using the thin lens equation, $O'P'$ is inverted, 1.5 cm long, and 5 cm to the right of L_2, while $O''P''$ is inverted, 3 cm long, and 10 cm from L_2.
4. The final drawing is shown in fig. 1.16.
5. The chief ray from point P to its conjugate point P'' is shown in fig. 1.16. Note that it leaves P, passes through M, the center of AS and ENP, underoges refraction at L_1, heads for $O'P'$, refracts again at L_2 before reaching $O'P'$ and heads for $O''P''$, the final image. Note also that the segment of the chief ray from L_2 to P'', if traced backward, will appear to be coming from point N, the center of the exit pupil EXP. THus the chief ray involves the centers of AS, ENP and EXP, as defined.

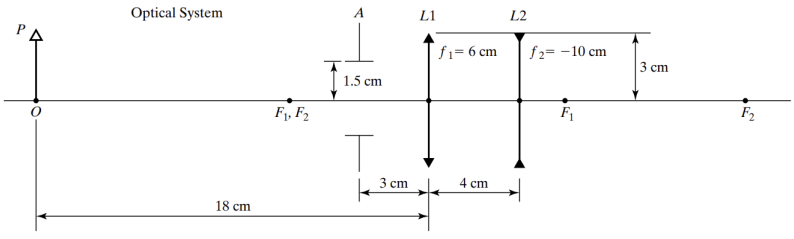


Figure 1.15

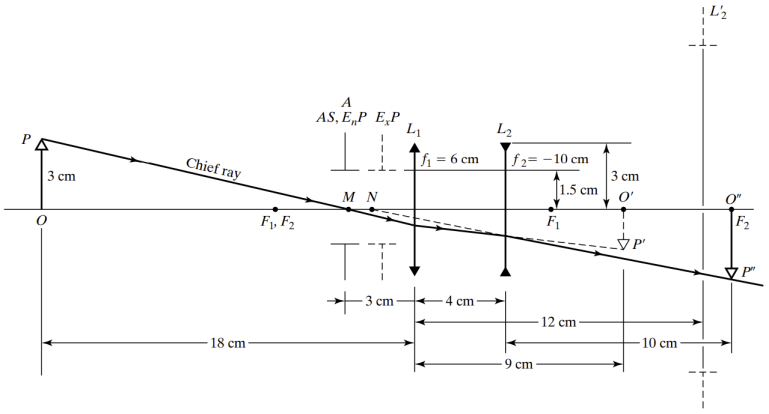


Figure 1.16

1.3.2 Pinhole Camera

Figure 1.17 shows the simple mechanism of a pinhole camera. Since the aperture stop is small, it requires a long exposure time. On the other hand, the depth of field is unlimited, and can focus object from any distance. In general, the aperture size is inversely related to the depth of field, since a smaller aperture produces a smaller blur circle (the image of a point source when it is slightly out of focus). Because the individual point sources spread less when the aperture is reduced, the system can tolerate more diviation from the perfect focus position before the blur becomes noticeable.

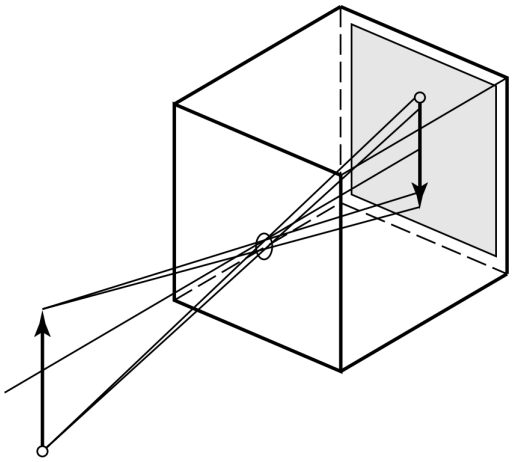


Figure 1.17

1.3.3 The Eye

For simplicity, a human eye can be regarded as a lens with adjustable focal length, and the closest point that can be imaged is called the Puntum Proximum (PP), and is often taken to be 25 cm.

If somebody's Puntum Proximum is at 70 cm, then we would need to add a convex lens such that an object positioned at 25 cm appear to be at 70 cm.

1.3.4 Magnifying Lens

The maximum angular size of an object with height L is $\theta_0 = L/PP$. There are two configurations in which a convex lens can be placed to produce different magnfications.[5]

If the object is placed at the focal point of the lens so that a virtual image is created at $v = \infty$ with height L', then the angular size of the object is given by

$$\theta = \frac{L'}{v} = \frac{L}{f}, \tag{1.13}$$

so the magnification is

$$M = \frac{\theta}{\theta_0} = \frac{L/f}{L/PP} = \frac{PP}{f}. \tag{1.14}$$

If the object is placed closer to the lens then at the focal point, so that a virtual image is formed at $v = -PP$, then the angular size of the image would be

$$\theta = \frac{L'}{v} = \left(\frac{vL}{u}\right)\left(\frac{1}{v}\right) = \frac{L}{u}, \tag{1.15}$$

so the magnification is

$$M = \frac{\theta}{\theta_0} = \frac{L/u}{L/PP} = \frac{PP}{u} = PP\left(\frac{1}{f} + \frac{1}{PP}\right) = 1 + \frac{PP}{f}. \tag{1.16}$$

1.3.5 Refracting Telescope

The purpose of a telescope is to take a real object at infinity and make a real image at infinity with a larger angular size. Refering to fig. 1.18, it consists of two convex lens, with an objective lens with focal length f_o and an eyepiece with focal length f_e. The two lens are arranged so that the focal point of the objective and the eyepiece coincides.

[5]Angular magnification refers to the ratio of angular size while linear magnification refers to the ratio of height. We will be predominately using angular magnification hereafter and it is assumed that we are refering to angular magnificiation when we say magnification, unless otherwise specified. This choice is justified becuase when the image is at infinity the lateral magnifications also approches infinity and is not very useful.

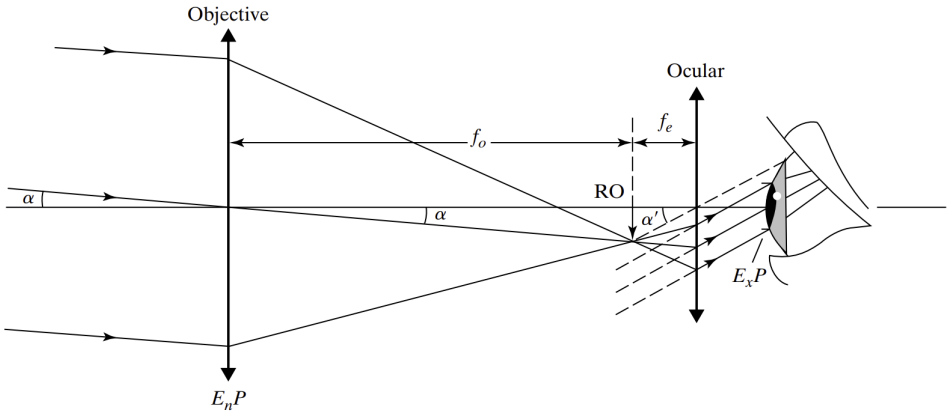


Figure 1.18

From the figure we find the magnification as

$$M = \frac{\alpha'}{\alpha} = \frac{f_o}{f_e}. \tag{1.17}$$

At the focal plane one can place a masking device to block the light of one of the sources but not the other. This technique, called chronography is used to image planets around other stars, which is usually much fainter than its host star.

1.3.6 Compound Microscope

The goal of a microscope is to make a large image at the Puntum Proximum of an object that is very small. Similar to the telescope it uses an objective lens of focal length f_o and an eyepiece of focal length f_e, but the object we are trying to image is not at infinity but close to the objective (see fig. 1.19).

Since the image is positioned at the Puntum Proximum, the angular magnification is the same as the linear magnification and is given by

$$M = \frac{PP}{f_{\text{eff.}}} = \frac{PP(f_e + f_o - d)}{} \tag{1.18}$$

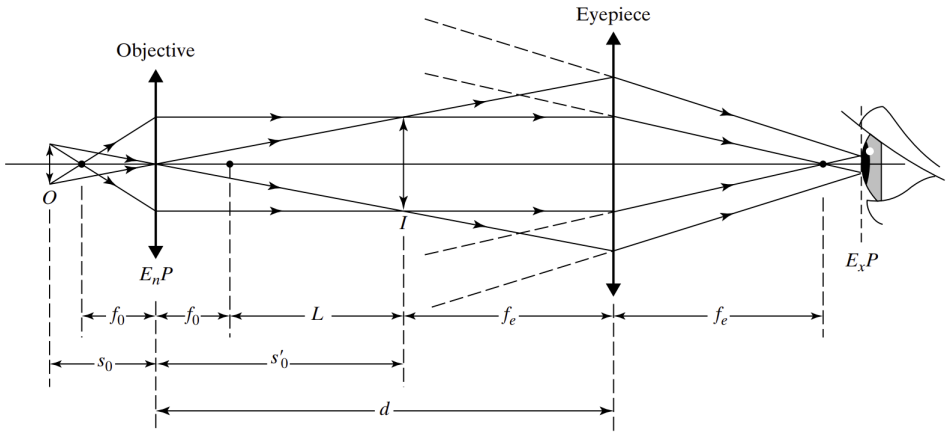


Figure 1.19

1.3.7 Prism

With some angle tracing, we can find the deviation angle of a light passing through a prism with refractive index n, apex angle α, incident angle θ_1 and refracted angle θ_2 as

$$\begin{aligned}\delta &= \theta_1 + \theta_2 - \left(\sin^{-1}\left(\frac{\sin\theta_1}{n}\right) + \sin^{-1}\left(\frac{\sin\theta_2}{n}\right)\right) \\ &= \theta_1 - \alpha + \sin^{-1}\left(n\sin\left(\alpha - \sin^{-1}\left(\frac{\sin\theta_1}{n}\right)\right)\right).\end{aligned} \tag{1.19}$$

To find the minimum deviation angle δ, we can set the derivative to be zero. Alternatively, we can use the reversibility of light and argue that by symmetry the light must be parallel to the base of the prism while travelling inside the prism. More formally, since δ is a symmetric function of θ_1 and θ_2, *i.e.*, $\delta(\theta_1, \theta_2) = \delta(\theta_2, \theta_1)$, so at minimum we must have $\theta_1 = \theta_2$, therefore the minimum deviation angle can be found as $\delta_{\text{min.}} = 2\sin^{-1}(n\sin(\alpha/2)) - \alpha$.

It is this minimum deviation angle that leads to the formation of halo and rainbow, since more light will cross the crystals or raindrop with a deflection angle close to the minimum angle.

2 Interference and diffraction

To begin, we must be noted that interference and diffraction are not fundamentally different processes, they are both interaction between light as a manifestation of its wave properties. In most contexts, interfrences refer to the interaction between finite number of light waves while diffraciton refer to infinite number of light waves.

For the interaction to be visible, the light waves must be coherent, *i.e.,* has a constant phase difference which is indepdent of time, *i.e.,* has the same frequency. This is acheived either by genearting them with the same plane wave (*i.e.,* spherical wave with distance from the origin far enough), or by a laser. If this is not the case, then the phase differences between the electric fields will vary randomly and average out to no visible pattern.

Since the sources are usually generated by the same plane wave, the electric fields can be assumed to be parallel to each other. Even if this is not the case, unpolarized light (*e.g.,* natural light) can still interfere as long as the lights are coherent, so that the effect of unparallel lights only decrease the intensity of the interference pattern, but the shape would remain unchanged.

For genearl consideration, we consider the interference of two electric fields

$$\mathbf{E}_1 = \mathbf{E}_{01}\cos(\mathbf{k}_1\cdot\mathbf{r}-\omega t+\epsilon_1) \quad \text{and} \quad \mathbf{E}_2 = \mathbf{E}_{02}\cos(\mathbf{k}_2\cdot\mathbf{r}-\omega t+\epsilon_2). \tag{2.1}$$

The resultant intensity is

$$\begin{aligned} I &= \langle(\mathbf{E}_1+\mathbf{E}_2)\cdot(\mathbf{E}_1+\mathbf{E}_2)\rangle = \langle|\mathbf{E}_1|^2+|\mathbf{E}_2|^2+2\mathbf{E}_1\cdot\mathbf{E}_2\rangle \\ &= I_1+I_2+(\mathbf{E}_{01}\cdot\mathbf{E}_{02})\left\langle\cos\left((\mathbf{k}_2-\mathbf{k}_1)\cdot\mathbf{r}+\varphi_2-\varphi_1\right)\right\rangle. \end{aligned} \tag{2.2}$$

When $\mathbf{E}_{01} \parallel \mathbf{E}_{02}$ and $E_0 \equiv E_{01} = E_{02}$, we have

$$I = I_1+I_2+\sqrt{I_1I_2}\langle\cos\alpha\rangle = 4I_0\cos^2\left(\frac{\alpha}{2}\right) \tag{2.3}$$

where α is the phase difference between two electric fields and at the last step we assumed it to be constant in time.

2.1 Fraunhofer diffraction

The Fraunhofer diffraction (or far-field limit) occus when the phase from each point in the diffracting object varies linearly with the transverse position, so that the diffracted wavefronts from an aperature or a slit can be approximated as planar, which is typically met when a focusing lens is used, so that the screen is effectively at infinity, or when $D \gg a^2/\lambda$, where D is the distance from the aperature or slit to the screen and a is the charateristic size of the aperature or the slit.[1] This can be seen from eq. (2.17) since the width of the central maximum is given by $W = 2D\lambda/a$, and we require it to be much greater than a to acheive far-field limit. The source must also be placed sufficiently far away such that it hits the aperature or the slit in planar manner.

As a result, for a certain point on the screen, all the lights arrive approximately in parallel and the electric field has the same amplitude $E_0(\theta) = E_0(0)/\sqrt{\cos\theta}$ for cylindrical wave (or 2D spherical wave),[2] usually we will further neglect this dependency and have $E(\theta) = E(0) \equiv E_0$. Note also that we are not using the electric fields at the aperature or the slit, since due to the singularity at the source one could not obatin the realtionship between the electric fields at the screen and at the source by simple means, except for the simple plane wave case where the electric field amplitude remain constnat througout.

In practice, far-field limit is usually acheived by putting a convex lens between the aperature and the screen, with the screen being the focal plane of the lens. This has the effect of bringing what we would observe at infinity (if there is no lens) closer to the focal plane of the lens. Essentially, what we want to observe is the interference of the parallel beams (at infinity), and the lens change the point of convergence from infinity to a point on the focal plane for easier observation. Illustrated in fig. 2.1, we can see that the parallel rays before the lens is what hit the point P' (not shown in the figure) in an ideal screen situated at infinity. Putting the lens causes the rays to converge to a closer point P while maintaining the phase difference to be unchanged thanks to Fermat's principle.

[1] Aperature and slit are not fundamentally different objects and slit is simply a special extreme case of an aperature, just as interference and diffraction are not fundamentally different processes, diffraction is simply a special name for interference when light passes through an aperature.

[2] Although Huygen's principle says that the secondary wavelets are spherical in nature, the wavefronts of the spherical waves created by a cylindrical source (the slit) is cylindrical in nature, thus the $1/\sqrt{r}$ decaying factor instead of $1/r$. But in doing so we must replace the electric field per unit area to the electric field per unit lenght. Equivalently, the changing of $1/r$ to $1/\sqrt{r}$ is due to one of the two integrations needed in a double integration.

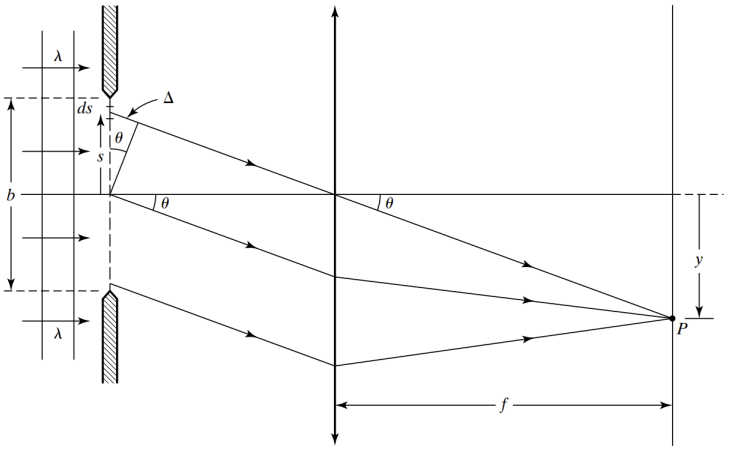


Figure 2.1

2.1.1 Interference

Double Slits

To find the interference pattern $I(\theta)$, we have

$$E(\theta) = E_0 e^{i(kr_1-\omega t)} + E_0 e^{i(kr_2-\omega t)} = 2E_0 \left(e^{i(k(r_1-r_2)/2)} + e^{-i(k(r_1-r_2)/2)}\right) e^{i(k(r_1+r_2)/2-\omega t)}, \tag{2.4}$$

where E_0 is the electric field produced by the single slit at $x = 0$ on the screen, which is not to be confused with $E(0)$, which is the electric field produced by both of the slits at $x = 0$.

We have also assumed the parallel of the electric fields, thus used scalar addition, and we have changed $\mathbf{k} \cdot \mathbf{r}$ to just kr since $\mathbf{k}$ and $\mathbf{r}$ are parallel if we use the coordinate systems where the origins are located at the two slits when doing the two dot products. Note that we do not have to use the same coordinate systems throughout, as long as the same coordinate system is used to evaluate one dot product.

$$I(\theta) = \langle |E(y)|^2 \rangle = 4E_0^2 \cos^2\left(\frac{kd\sin\theta}{2}\right) = 4I_0 \cos^2\left(\frac{kdx}{2D}\right) = 4I_0 \cos^2\left(\frac{\alpha}{2}\right), \tag{2.5}$$

where $I_0 \equiv E_0^2$, but $I(0) = (2E_0)^2 = 4I_0$, as expected. The interference pattern are evenly spaced fringes with separation $\Delta x = \lambda D/d$, shown in fig. 2.2.

The minima occurs at

$$\frac{\alpha}{2} = (m + \frac{1}{2})\pi, \tag{2.6}$$

and the maxima occur at

$$\frac{\alpha}{2} = m\pi. \tag{2.7}$$

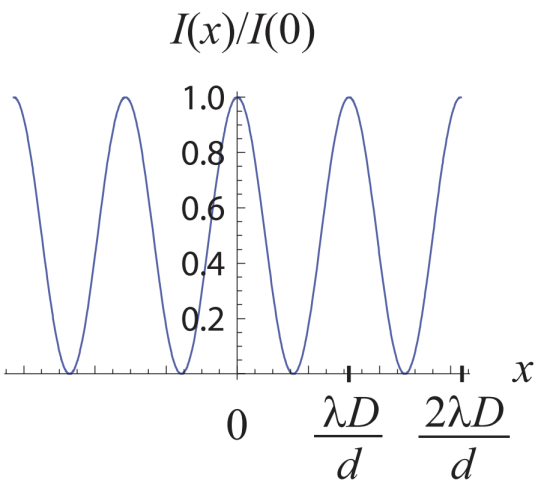


Figure 2.2

N-Slits

We start with finding the electric field

$$E(\theta) = E_0 e^{i(kr_1-\omega t)} \sum_{n=1}^{N} e^{i(k(n-1)d\sin\theta)} = E_0 e^{i(kr_1-\omega t)} \left(\frac{Z^N-1}{Z-1}\right)$$
$$= E_0 e^{i(kr_1-\omega t)} \frac{Z^{N/2}(Z^{N/2}-Z^{-N/2})}{Z^{1/2}(Z^{1/2}-Z^{-1/2})} = E_0 e^{i(kr_1-\omega t+(N-1)kd\sin\theta/2)} \left(\frac{\sin\left(Nkd\sin\theta/2\right)}{\sin\left(kd\sin\theta/2\right)}\right), \tag{2.8}$$

where $Z \equiv e^{ikd\sin\theta} = e^{i\alpha}$. So the intensity is

$$I(\theta) = \langle |E(x)|^2 \rangle = I_0 \left(\frac{\sin\left(Nkd\sin\theta/2\right)}{\sin\left(kd\sin\theta/2\right)}\right)^2 = I_0 \left(\frac{\sin\left(Nkdx/2D\right)}{\sin\left(kdx/2D\right)}\right)^2 = I_0 \left(\frac{\sin\left(N\alpha/2\right)}{\sin\left(\alpha/2\right)}\right)^2. \tag{2.9}$$

Here $I(0) = (NE_0)^2 = N^2 I_0$, which can be verified from the above equation by taking the limit as $\alpha \to 0$. Note that although $\theta \ll 1$, kd is not necessarily much smaller than 1, so the expression can not be simplified further.

Therefore, the minima occurs when the numerator is zero while the denominator is not, *i.e.,*

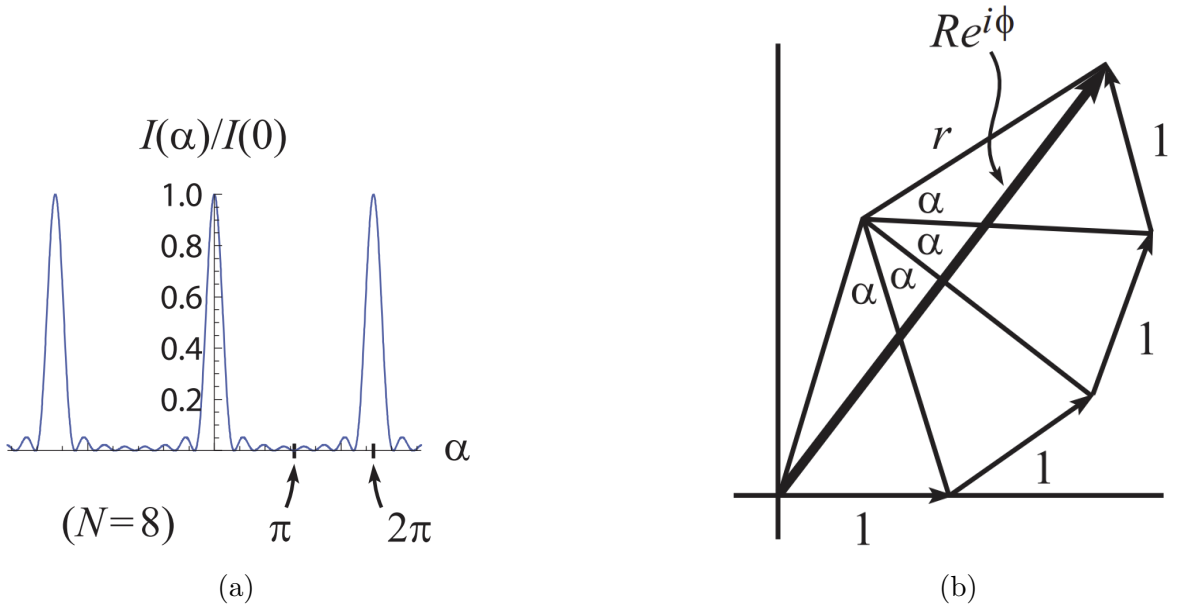
$$\frac{N\alpha}{2} = m\pi \implies d\sin\theta = \frac{m\lambda}{N}, \quad m \neq nN, \tag{2.10}$$

while the global maximum occurs when both the numerator and the denomiator are zero, *i.e.,*

$$\frac{\alpha}{2} = n\pi \implies d\sin\theta = n\lambda \tag{2.11}$$

which is when the lights from all slits are in phase.

The local minima can be found by differentiation, which gives $\tan(N\alpha/2) = N\tan(\alpha/2)$. For large N, the solutions for α genearlly gives symmetric solutions, and the intensity is the smallest for the local maximum centered between the global maxima, as shown in fig. 2.3a for the $N = 8$.



(a) (b)

Figure 2.3

To visualize the interference we can make use of the phasor diagram as shown in fig. 2.3b, where the phase difference between each successive electric field is α and the length of each electric field vector is 1. Consider the two types of isosceles triangles in the figure, we have $R = 2r\sin(N\alpha/2)$ and $1 = 2r\sin(\alpha/2)$, which implies $R = \sin(N\alpha/2)/\sin(\alpha/2)$.

This diagram also visualize the pairwise cancellation of electric field when minima occurs at $N\alpha/2 = m\pi$ as shown in fig. 2.4, where the figures are drawn slightly off to make it easier to see the cancellation. The maxima occur approximately between the minima since the overall phase difference is now integer multiple of π instead of 2π, so the resultant electric field vector has the longest length, instead of being cancelled out, as shown in fig. 2.5 for $N = 50$. The maxmima don't occur exactly at the diameters since the circle shrinks as the vectors wrap around futher as α increases so there are competing effect, but this effect is small when N is large as the circle circle hardly changes sizes.

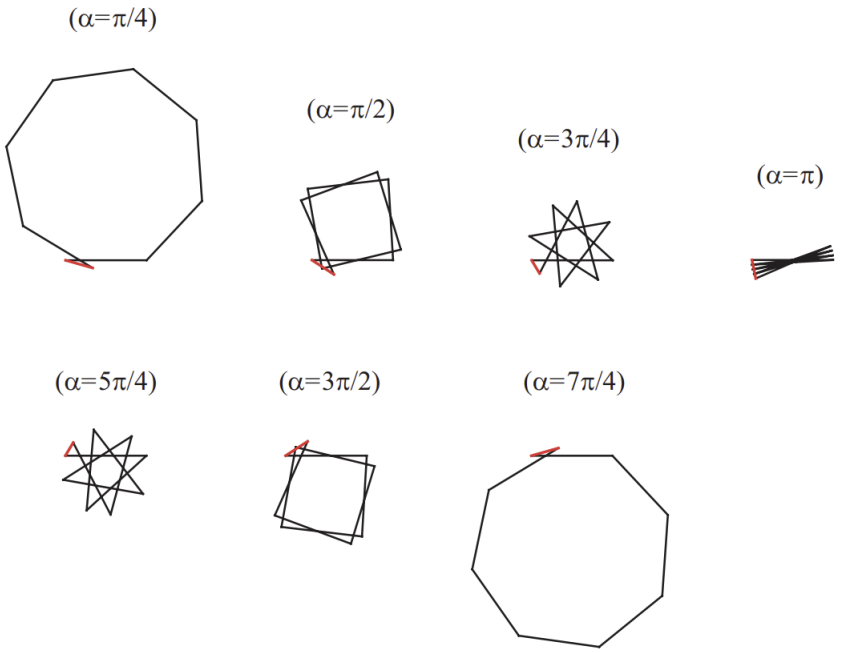


Figure 2.4

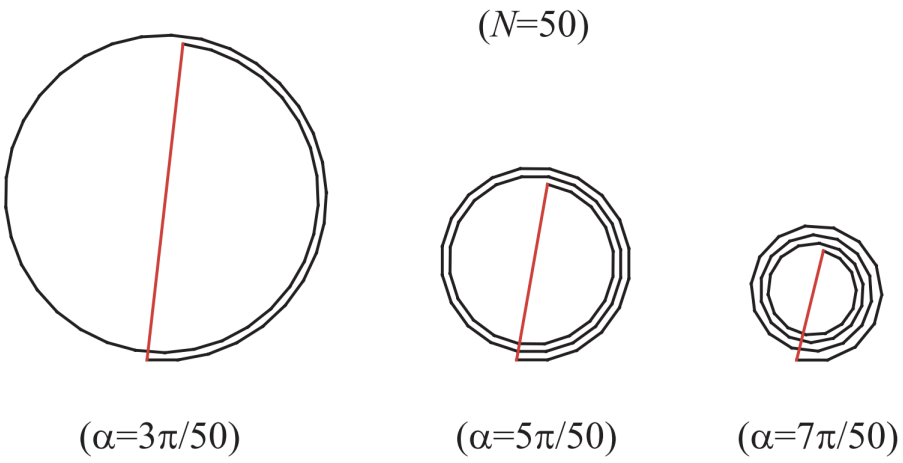


Figure 2.5

Diffraction grating is essentially N slits interference with $N \gg 1$, since then the intensity of the local maximum is negligible, since it scales with $1/N^2$, so the maxima are only observed when $\alpha = 2n\pi$, or $\lambda = d \sin\theta$.

The angular diffraction of a grating is defined as

$$\frac{d\theta}{d\lambda} = \frac{m}{d\cos\theta}, \tag{2.12}$$

which measures how much the maxima angle changes for a small change in wavelength.

A diffration grating can be used to make a spectrometer. Firstly an objective lens image the lamp's arc onto a narrow entrance slit, creating a small and well-defined line source.

A collimating lens then converts the diverging rays from the slit into a paralle beam, and strikes normally at the diffraction grating. Lights with different wavelengths then construtively interfere and forming maxima intensity at different angles as shown in fig. 2.1.

The objective lens is important since the lamp is not a point source, so each point on the lamp would form its own spectrum, blurring over one another. Collimated light is essential so the rays incident on the grating is parallel. Also the whole diffraction grating should be illuminated to maximize the resolving power R.

The resolving power of a grating is defined in general by

$$R = \frac{\lambda}{\Delta\lambda_{\min}}, \tag{2.13}$$

where $\Delta\lambda_{\min}$ is the smallest wavelength difference that the instrument can just resolve at wavelength λ.

The change in diffraction angle $\Delta\theta$ when the wavelength is changed by a small amount $\Delta\lambda$ is given by

$$\Delta\theta = \frac{m\Delta\lambda}{d\cos\theta}. \tag{2.14}$$

On the other hand the angular width of a principal maximum is given by

$$\delta\theta = \frac{\lambda}{Nd\cos\theta}. \tag{2.15}$$

From the Rayleigh criterion we state that two wavelengths λ and $\lambda + \Delta\lambda$ are just resolved when the angular separation of their maxima equals the half-width of a single maximum

$$\Delta\theta = \delta\theta \implies \frac{m\Delta\lambda}{d\cos\theta} = \frac{\lambda}{Nd\cos\theta} \implies \Delta\lambda_{\min} = \frac{\lambda}{mN} \implies R = \frac{\lambda}{\Delta\lambda_{\min}} = mN. \tag{2.16}$$

2.1.2 Diffraction

Single Slit Diffraction

If the slit width a is not negligible, then for a single slit the intensity can be found by treating every points in the slit as a single slit with slit separation $d = a/N \to 0$ and number of slits $N \to \infty$, so from eq. (2.9)

$$I(\theta) = \frac{I(0)}{N^2}\left(\frac{\sin(Nkdx/2D)}{\sin(kdx/2D)}\right)^2 = I(0)\left(\frac{\sin(\beta/2)}{\beta/2}\right)^2, \quad \beta = ka\sin\theta, \tag{2.17}$$

was shown in fig. 2.6a. Therefore, minima occurs when

$$\frac{\beta}{2} = m\pi \implies \lambda = ma\sin\theta \approx ma\theta, \tag{2.18}$$

which is when the electric fields from the top and bottom of the slits are in phase, since then if we divide the slits into two halfs (for $m = 1$) (or four quarters for $m = 2$ *etc.*), we can always find a corresponding electric field is out of phase for any electric field, achieving pairwise cancellation. The maxima condition is now $\tan(\beta/2) = \beta/2$, which occur roughly halfway between the zeros.

More rigorously, we can find the intensity profile by summing the electric field by the infinite slits with width dx and taken the averaged modulus squared to find the intensity, we start with

$$E(\theta) = \int_{-a/2}^{a/2} (E_0' dx) e^{ikx \sin\theta} = \frac{E_0'}{-ik \sin\theta} \left(e^{ik(a/2)\sin\theta} - e^{-ik(a/2)\sin\theta} \right) = E_0' a \left(\frac{\sin(\beta/2)}{\beta/2} \right), \tag{2.19}$$

where E_0' is the electric field per unit length across the slit produced at $x = 0$ on the screen, and thus $E_0' a$ is the electric field of the slit produced at $x = 0$. Therefore averaging the modulus squared give the same result as above.

In fact the electric field can be obtained by the fourier transforming the transimitivity function $T(x)$

$$E(\theta) = E_0' \int_{-\infty}^{+\infty} T(y) e^{-ikx \sin\theta} dx = \tilde{T}(k \sin\theta). \tag{2.20}$$

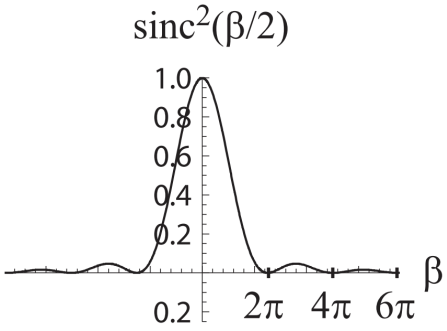
From the theory of Fourier transform, this means that if T has a large componnet with spatial frequency $k \sin\theta$, then the electric field at angle θ will be larger, which is true as when the spatial frequency of T is $k \sin\theta = 2\pi/(\lambda/\sin\theta)$, then from fig. 2.7a we see that the light generally constructively interfere with one another.

The above results can also be seen from fig. 2.6b, where the digaram is constructed by noting that the phase diffrence between the elctric fields at the top and the bottom of the slit is given by β, then the ratio of intensity is given by the ratio of the length between the straight line and the arc, which gives $I(x)/I(0) = (2r \sin(\beta/2)/r\beta)^2$, which is the same as the above equation. The minima condition $\beta = 2\pi$ corresponds to the fact that the arc in fig. 2.6b turns into a full circle, so the resultant electric field is zero.

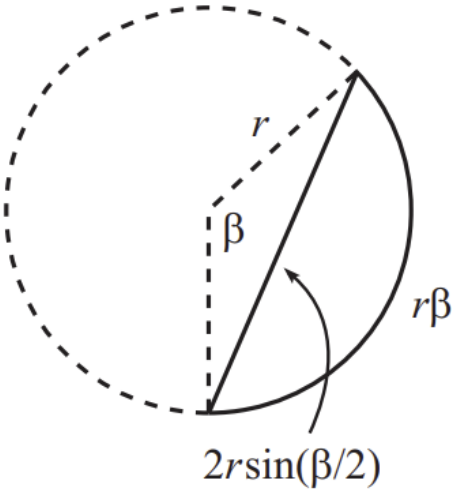
Example: Four Times the Light?

Question: It can be easily proved that $I(0) \propto a^2$. This means that if we double a, then $I(0)$ increases by a factor of 4. Does this make sense and does it mean that if we double the width of the slit, then 4 times as much light makes it through?

Solution: The answers are yes and no, respectively. The intensity increases by a factor of 4 since this is what we get mathematically but the energy does not increases by a factor of 4 since intensity is power divided by area, so energy is the area under graph of the intensity profile. So while $I(0)$ quadruple, since the intensity profile is thinner as a gets larger, exactly by a factor of $1/2$, the energy

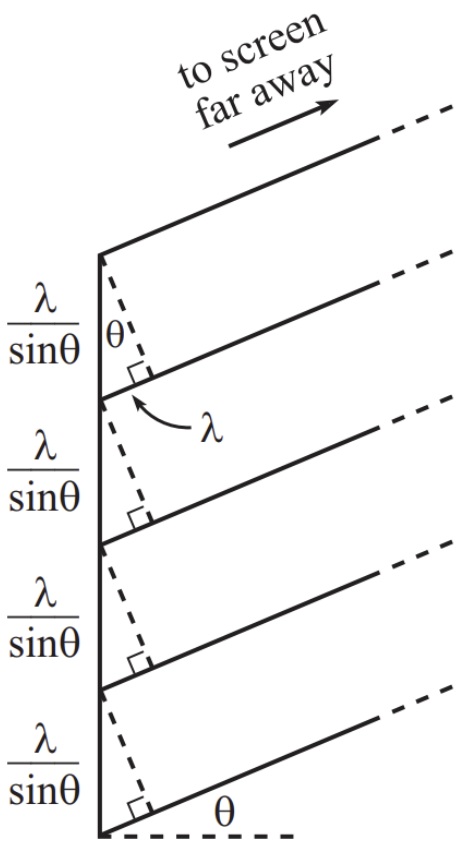


(a)

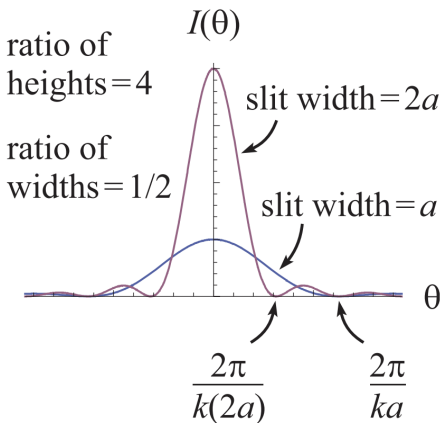


(b)

Figure 2.6



(a)



(b)

Figure 2.7

only doubles. The situation is illustrated in fig. 2.7b.

Example: Increasing or Decresing Intensity?

Question: As we make a larger, will the intensity at a particular point on the screen that is reasonable distance off to the side increases or decrease?

Solution: On one hand, increasing a will aloow more light through the slit, so the intensity should increase. But on the other hand, increasing a will make the diffraction pattern narrower, so the intensity should decrease. In turns out that these two effects exactly cancel, so the intensity remained unchanged. For the fact that $I(\theta) \propto I(0)\sin^2(ka\theta/2)/k^2a^2\theta^2$, but since $I(0) \propto a^2$, thaking the average over a few oscillation of θ, we have $\langle I(\theta)\rangle \propto 1/k^2\theta^2$, independent of a.

N-Slits Diffraction

By integrating the contribution of every single points, we obtain the intensity of N-slits diffraction as

$$I(\theta) = I(0)\left(\frac{\sin\alpha/2}{N\alpha/2}\frac{\sin\beta/2}{\beta/2}\right)^2, \tag{2.21}$$

simply the product of the results for the two separate cases we have discussed, so the diffraction effect simply acts as a modulating envelope for the N-slits interference, as shown in fig. 2.8.

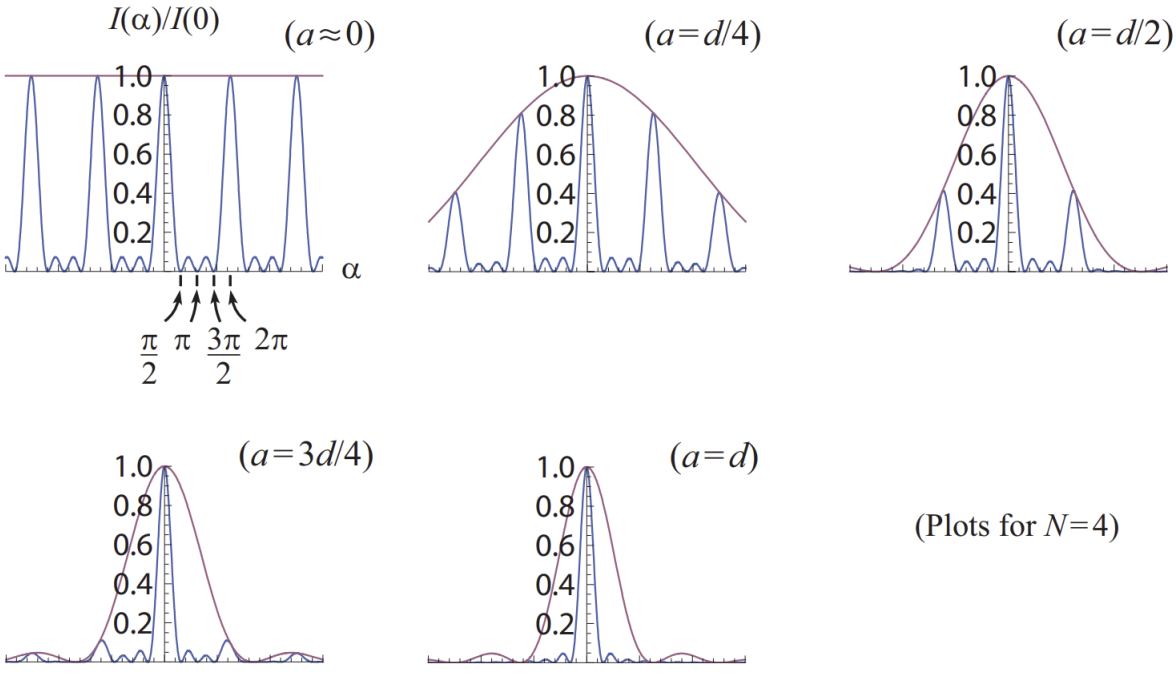


Figure 2.8

Circular Aperature Diffraction

Using the geometry shown in fig. 2.9, the electric field is

$$E(\theta) = \int (E_0'' dA) e^{iks\sin\theta} = 2E_0'' \int_{-R}^{R} e^{iks\sin\theta}\sqrt{R^2 - s^2}ds = \frac{2E_0''\pi J_1(\gamma)}{k\gamma}, \quad \gamma = kR\sin\theta, \tag{2.22}$$

where E_0'' is the electric field per unit area over the aperature produced at $x = 0$ on the screen and $J_1(\gamma)$ is the first-order Bessel function of the first kind, with the first order term equals to $\gamma/2$.

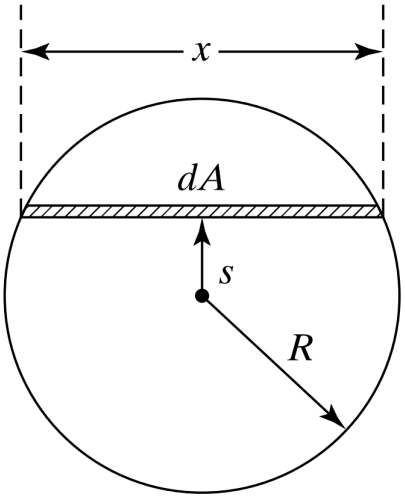


Figure 2.9

The first minimum occurs when $\gamma = 3.832$, or $D\sin\theta = 1.22\lambda$. Rayleigh's criterion for just-resolvable images requires that the central maximum of one image coincedes with the first minimum of the another image, *i.e.,* $\Delta\theta = 1.22\lambda/D$, where D is usually the diameter of the lens if we are viewing two starts through a microscope.

2.1.3 Rayleigh Criterion

The Rayleigh Criterion states that two diffraction-limited point images can be resolved when the central maximum of one coincides with the first minimum of the other.

Example: Rayleigh Criterion.

Question: A spy satellite orbits 200 km above the surface of the Earth and observes light at wavelength 550 nm. Estimate the size of the main mirror on the spy satellite if it is able to read the leahlines of a tabloid newspaper, with a font thickness of 3 cm.
A digital camera is used to record the image at the focus of the spy satellite's main mirror. If the main mirror of the spy satellite has a focal length of 80 m, estimate the maaximum distance between the centers of adjacent pixels for the camera to fully resolve the newspaper headlines.

Solution: Using the Rayleigh Criterion we have

$$\theta = \frac{3\,\text{cm}}{200\,\text{km}} = 1.22\frac{550\,\text{nm}}{D} \implies D = 4.47\,\text{m}. \tag{2.23}$$

As a visual aid, consider two sets of parallel beam, separated by angle θ forming an

2.2 Optical Interferometry

Before extending our discussion beyond Far-field limit, we will devote this secion to the application of interference of light, maining with optical interferoemters.

2.2.1 Michelson Interferometer

The set up and the equivalent optical system are shown in fig. 2.10, where M_1' is constructed by imaging M_1 with respect to the mirror on the beam splitter BS. Therefore the phase difference between the two beams from the source S is $\delta = 2kd\cos\theta + \pi/2$, and the intensity profile is given by $I = 4I_0\cos^2(\delta/2)$. So the condition for dark fringes is given by $2d\cos\theta = m\lambda$, where the order of the central dark fringe is given by $m_{\max} = 2d/\lambda$. If we want we could invert the ordering of the fringes by defining $p = m_{\max} - m$, so that the order increase as we get further away from the center, but usually one would not bother to do this as the ordering is arbitrary anyways.

As we decrease d by moving M_1 away, so that M_1' is closer to M_2 (here we assume initially M_1' is in front of M_2 as shown in fig. 2.10), $m_{\max}$ becomes smaller and would goes to zero as $d = 0$, where there is no interference pattern, but then it would gets larger again when M_1 is move further away. In fact, we almost always only concern about the change in m (this is also why the absolute value of m is not important), given by $\Delta m = 2\Delta d/\lambda$.

One application of the Michelson interferometer is to measure the difference in wavelength between two closed spaced components of a spectral line, such as the two yellow sodium lines. Suppose will adjust d to d_1 such that the two circular interference patterns produced by the two lines coincide with each other at $m_1 = m_1' + N$, then as we adjust d we will see a field of uniform brightness when $m_{1/2} = m_{1/2}' + N + 1/2$, and at d_2 we have $m_2 = m_2' + N + 1$, using $m_{\max} = 2d/\lambda$, we have $\Delta\lambda = \lambda^2/2\Delta d$.

Twyman-Green interferometer, shown in fig. 2.11 is a modified version of the Michelson interferometer, which when gives no interference pattern when nothing is placed between it, thus it can be used to test the quality of optical instruments such as prism P or another lens, in which case the mirror M_1 has to be replaced by a spherical mirror that can reflect the refracted rays back along themselves. If there are imperfections in the optical system, then distortions will appear in the interference pattern.

Furthermore, the modified Michelson interfermoeter shown in fig. 2.12 can be used to measure the thickness d of the thin film. Monochromatic light channeled from a light source LS through a fiber optic light pipe LP to a right-angled beam-splitting prism BS, which transmits one beam to a flat mirror M and the other to the film surface. After reflection, each is transmitted by the beam splitter into a microscope MS, where they are

allowed to interfere. For normal incidence, bright fringes satisfy $2nt + \pi = m\lambda$, when the film is not present, we have $2n\Delta t = 2d = \lambda\Delta m = (\lambda\Delta x)/x$, which implies $d = (\lambda\Delta x)/2x$.

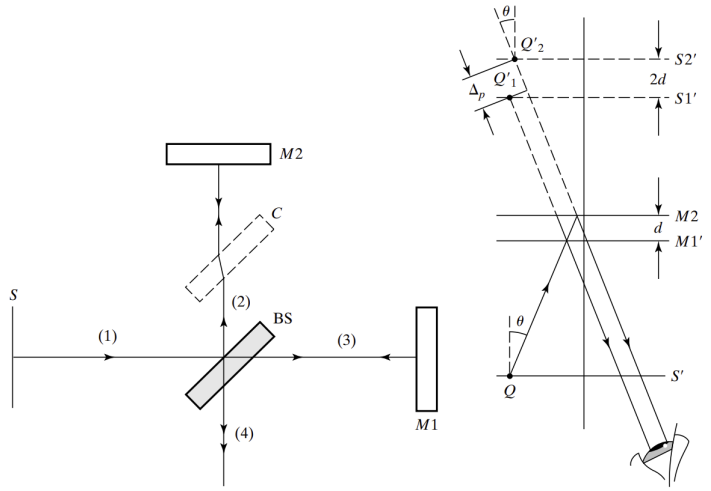


Figure 2.10

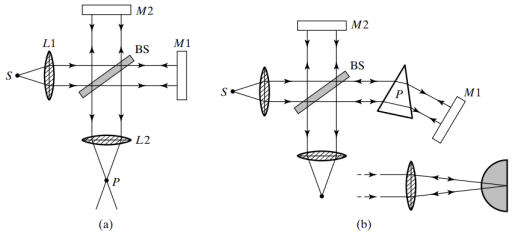
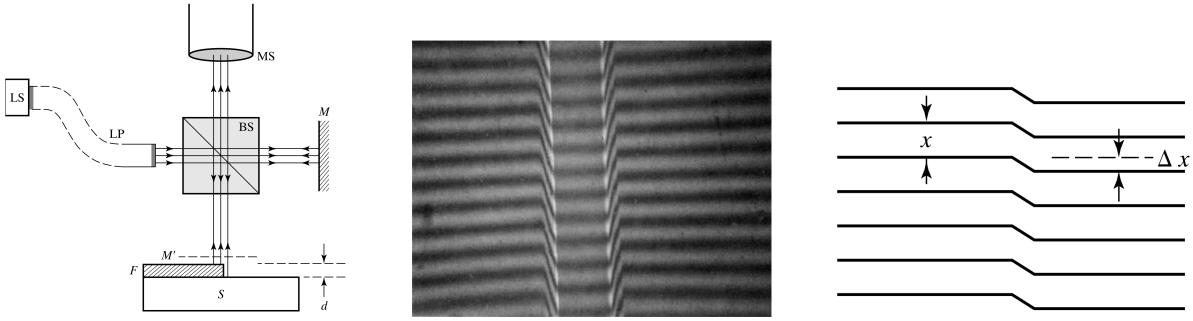


Figure 2.11



(a) (b) (c)

Figure 2.12

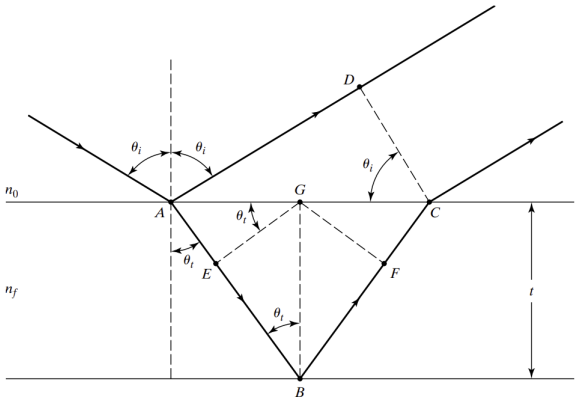


Figure 2.13

stokes relation

2.2.2 Fabry-Perot Interferometer

The two examples below lay the foundation when analysing the fabry-perot interfermoeter.

Example: Thin-film Interference.

Question: Find the condition for constructive interference for the situation shown in fig. 2.13.

Solution: The path difference is

$$\Delta = n_f(AB + BC) - n_0(AD) = \frac{2n_f t}{\cos\theta_t} + \frac{2n_0 t\cos\theta_i}{\sin\theta_t} = 2n_f t\cos\theta_t, \tag{2.24}$$

where we have used the Snell's law to eliminate θ_i.
So the constructive interference condition is $\Delta + \lambda/2 = m\lambda$.

Example: Multiple-Beam Interference in a Parallel Plate.

Question: Refer to fig. 2.14, find the condition for constructive interference in the transmitted beam.

Solution: At the first mirror, the beam (1) reflects at the mirror-medium interface, thus will not acquire a π phase shift, while the beams (2), (3), *etc.* reflects at the medium-mirror interface, thus will acquire a π phase shift, so we have $r = -r'$.
On the other hand, Stokes' relation gives $tt' = 1 - r^2 = 1 - r'^2$.
Letting $\delta = 2kn_f t\cos\theta_t$ to be the phase difference beteween each successive reflected beam, we have

$$E_1 = (rE_0)e^{i\omega t}, \quad E_2 = (tt'r'E_0)e^{i(\omega t-\delta)}, \quad E_3 = (tt'r'^3E_0)e^{i(\omega t-2\delta)} \quad etc., \tag{2.25}$$

where the negative sign in the first reflected beam comes frmo the π phase shift of reflection. Summing the reflected lights, we have

$$
\begin{aligned}
E_R &= \sum_{N=1}^{\infty} E_N = rE_0 e^{i\omega t} + \sum_{N=2}^{\infty} tt' E_0 r'^{(2N-3)} e^{i(\omega t-(N-1)\delta)} \\
&= E_0 e^{i\omega t}\left(r + \frac{(1-r^2)r'e^{-i\delta}}{1-r'^2 e^{-i\delta}}\right) = E_0 e^{i\omega t}\left(\frac{r(1-e^{-i\delta})}{1-r^2 e^{-i\delta}}\right),
\end{aligned} \tag{2.26}
$$

so the intensity is

$$
\begin{aligned}
I_R &= \langle |E_R|^2 \rangle = E_R^* E_R = E_0^2 r^2 \left(\frac{e^{-i\omega t}(1-e^{i\delta})}{1-r^2 e^{i\delta}}\right)\left(\frac{e^{i\omega t}(1-e^{-i\delta})}{1-r^2 e^{-i\delta}}\right) \\
&= \left(\frac{2r^2(1-\cos\delta)}{1+r^4-2r^2\cos\delta}\right) I_i, \quad I_I = E_0^2.
\end{aligned} \tag{2.27}
$$

Similar treatment of the transmitted beams leads to the resultant transmitted intensity

$$
I_T = \left(\frac{(1-r^2)^2}{1+r^4-2r^2\cos\delta}\right) I_I \implies I_R + I_T = I_I. \tag{2.28}
$$

A maximum in transmitted intensity occurs when $\cos\delta = 1$, or $k\delta = 2n_f t\cos\theta_t = m\lambda$, which is when the second reflected beam and all subsequent reflected beams are in phase with one another but exactly out of phase with the first reflected beam due to the π phase shift. Substituting into the reflected and transmitted intensity we found that the reflected beam intensity vanishes and the transmitted beam intensity is the same as the incidence beam.

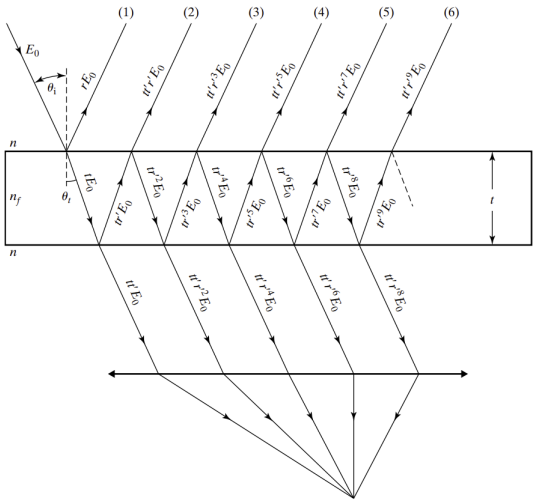


Figure 2.14

2.3 Near-field Limit

In Near-field limit, we can no longer assume that $E(\theta) = E_0$ for all points from the sources, we have to take into account the $1/\sqrt{r}$ dependence in the amplitudes. Moreover, the pathlengths does not thake the nice form $d\sin\theta$, but is more complicated.