UNIVERSITY OF AGRICULTURE, ABEOKUTA, NIGERIA

DEPARTMENT OF PHYSICS

PHS 245: BASIC PHYSICS FOR ENGINEERING APPLICATIONS (2 UNITS)

PART II

BY
DR. RASAQ BELLO

Acknowledgement: This lecture note was adapted from the lecture note on Optics taught at Texas A & M University-Kingsville.

Photons and Light.

There are some effects that light has which can only be described by treating it as a particle. Two of these effects are radiation pressure and momentum. As early as 1619, Kepler proposed that the pressure of sunlight was responsible for blowing back a comet's tail so that it always points away from the Sun. This idea was initially used by proponents of the particle nature of light. For a while it seemed as though this effect might at last establish the superiority of the particle theory over the wave theory of light, but all the experimental efforts failed to detect the force of radiation. It wasn't until Maxwell unified electricity and magnetism that interest in radiation pressure was revived.

When an electromagnetic wave hits a material surface, it interacts with the charges that constitute solid matter. Regardless of whether the wave is partially absorbed or reflected, it exerts a force on those charges, and hence on the surface itself. For example, consider a pulse of a wave, reaching a charge at rest. The electric field will put the charge in motion, then the magnetic field will act perpendicular to its velocity, causing it to be pushed in the original direction of the wave. Thus, the wave exerts pressure and transfers momentum. It is possible to compute the resulting force via classical electromagnetic theory. From Newton's second law, we see that the wave itself must carry momentum. Indeed whenever we have a flow of energy, it is reasonable to expect that there will be an associated momentum.

Radiation Pressure

So how do we describe the radiation pressure? Maxwell showed that the **radiation pressure**, *P*, equals the energy density of the electromagnetic wave. we see that

$$\begin{aligned} P &= u \\ &= u_E + u_B \\ &= \frac{1}{2}\varepsilon_0 E^2 + \frac{1}{2\mu_0} B^2 \\ &= \frac{S}{c} \end{aligned} \tag{1}$$

Since the electric and magnetic fields vary rapidly, *S* varies rapidly. Thus, it is appropriate to use the average radiation pressure,

$$\begin{aligned} \langle P \rangle &= \frac{\langle S \rangle}{c} \\ &= \frac{I}{c} \end{aligned} \tag{2}$$

We can determine the momentum of light in two different ways. If p_V is the momentum per unit volume of the radiation, then during each time interval Δt the force exerted by the beam on an absorbing surface is

$$\begin{aligned} F &= AP \\ &= \frac{\Delta p}{\Delta t} \\ &= \frac{p_V (cA\Delta t)}{\Delta t} \\ &= A\frac{S}{c} \end{aligned} \tag{3}$$

Thus the volume density of electromagnetic momentum is

$$p_V = \frac{S}{c^2} \tag{4}$$

In the photon picture, each quanta is seen to have an energy $E = h\nu$. Using (1) and (4) we see that

$$\begin{aligned} p &= p_V V = \frac{S}{c^2} V \\ &= \frac{P}{c} V = \frac{u}{c} V \\ &= \frac{E}{c} = \frac{h\nu}{c} \\ &= \frac{h}{\lambda} \end{aligned} \tag{5}$$

The vector momentum is written as

$$\vec{p} = \hbar \vec{k} \tag{6}$$

where ***k*** is the propagation vector and $\hbar = h/2\pi$. This is in agreement with special relativity, which states that, in general,

$$E^2 = (pc)^2 + \left(m_0 c^2\right)^2 \tag{7}$$

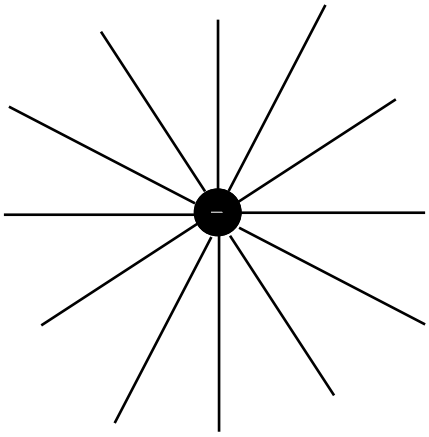
In turn, this implies that $m_\gamma \equiv m_0 = 0$.

Electromagnetic Radiation from Charge Motion

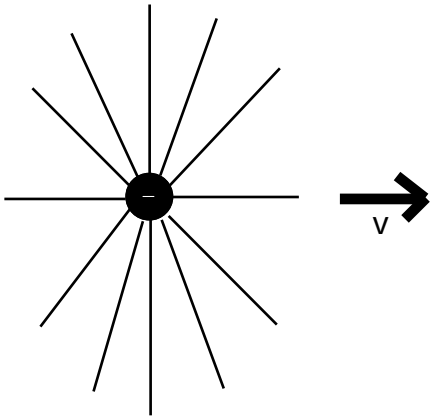
We have seen that all forms of electromagnetic radiation are different aspects of one entity, the electromagnetic wave. Maxwell's equations are independent of wavelength and so suggest no fundamental differences in the different types of radiation. Accordingly, it is reasonable to look for a common source mechanism for all radiation. This common source turns out to be that they are all associated with nonuniformly moving charges.

A stationary charge has a constant electric field and no magnetic field, and hence produces no radiation. A uniformly moving charge has both an electric and magnetic field, but by changing our frame of reference to one co-moving with the charge, the magnetic field can be made to vanish and we would be back to the first case. Thus, we are left with nonuniformly moving charges as the source of electromagnetic radiation.

Let's look at the electric field produced by a single negative charge. If the charge is stationary, then we know that the electric field is spherically symmetric, so that the field lines look like

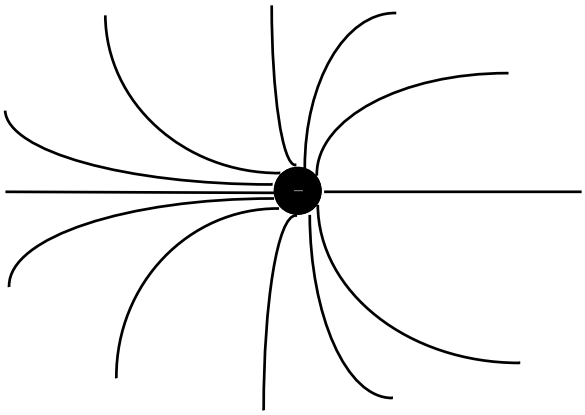


Notice that the field lines are uniformly distributed about the charge. If the charge is moving, then special relativity tells us that the field lines, while still radial and straight, are no longer uniformly distributed. Instead, they undergo length contraction in the direction of motion, so that the field lines look like

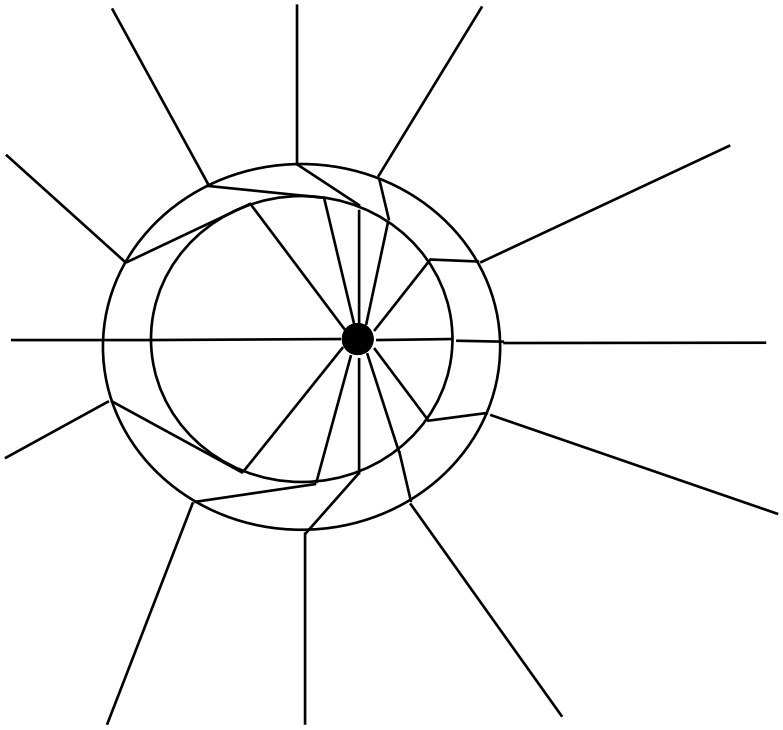


An interesting effect of this is that at near the speed of light, the field lines are compressed into a disk which is perpendicular to the direction of motion.

The situation is different when the charge is accelerating. Now, instead of straight lines, the field lines are curved.



In particular, we can ask what the field line configuration would look like if we had a charge which, at time $t = 0$ begins accelerating, and then at time $t = t_1$ stops accelerating and continues at a constant velocity. For this case, the field lines look like



Notice that near the charge the field lines are compressed according to special relativity, while far away from the charge the field lines are still centered on the charge location at $t = 0$. The two sets of lines are connected by an accelerated region, resulting in a "kink" in the field line distribution. From this kinked region we see that there exists a transverse component of the electric field, $\boldsymbol{E}_T$. This transverse component propagates outward, away from the charge, at the speed of light. Thus, the pulse which contains the kink is not only a function of space, but also of time. From Maxwell's equations, we see that the transverse field is therefore accompanied by a magnetic field.

From Coulomb's law, we know that the radial component of the electric field is proportional to $1/r^2$. By solving Maxwell's equations, it can be shown that the transverse component goes as $1/r$. Thus, at large distances from the charge, the transverse component of the pulse will dominate. We call this the **radiation field**.

Sources of the Radiation Field

Recall that the radiation field is carrying energy with it. Where does this energy come from? By conservation of energy, it must be supplied to the charge by the accelerating force, which is doing work on the charge.

Among the most simple types of acceleration mechanisms is that of creating an oscillating electric dipole. Consider two charges, one positive and one negative. Let the positive charge be stationary, while the negative charge oscillates linearly with simple harmonic motion. If the angular frequency of the oscillation is ω, then the dipole moment has a magnitude of

$$p(t) = p_0 \cos \omega t \tag{8}$$

At $t = 0$, the dipole has a magnitude $p = p_0 = qd$, where d is the initial maximum separation between the centers of the two charges. Recall that the dipole moment is actually a vector pointing from the negative charge to the positive charge. Then as the negative charge moves away from the positive charge, the field lines separate from each other and move away from the dipole. As the two charges come closer, the field lines emanate from a smaller space, until finally, when the positive and negative charges overlap, the field lines on each other. Thus, as the dipole oscillates, alternating sets of closed field lines are generated, propagating outward away from the dipole source.

Very near the dipole source, the electric field has the form of a static electric dipole. Farther out, in the region where the closed loops form, the field is complicated, with five different terms contributing to the field strength. Far from the dipole, where the transverse component dominates, the electric field can again be easily described. At this distance a fixed wavelength has become established, and the electric and magnetic fields have become transverse, mutually perpendicular and in phase. Specifically, we find that

$$E = \frac{p_0 k^2 \sin\theta}{4\pi\varepsilon_0} \frac{\cos(kr - \omega t)}{r} \tag{9}$$

and $B = E/c$, as usual. The region where the electric field is defined by (9) is known as the **wave**, or **radiation zone**.

Another form of radiation is **synchrotron radiation**. This is generated whenever an unbound charged particle travels on any sort of curved path. The frequency of the orbit determines the frequency of the emission, which also contains higher harmonics.

A charged particle slowly revolving in a circular orbit radiates a donut shaped pattern perpendicular to the acceleration. Since this acceleration is centripetal acceleration, the radiation pattern forms lobes in front and behind the direction of motion. As the speed of the particle is increased, the symmetrical pattern that is evident in the particle's rest frame becomes more and more distorted, creating a large lobe before the particle and a smaller lobe behind it. At speeds approaching that of light, the particle radiates essentially along a narrow cone pointing tangent to the orbit in the instantaneous direction of the velocity, $\boldsymbol{v}$. Thus, for $v \approx c$ the radiation will be very strongly polarized in the plane of the motion.

Matter and the Radiation Field

Probably the most important form of radiation comes from bound charges. Much of the chemical and optical behavior of matter is determined by the outer, or valence, electrons. The remainder of the electron cloud usually forms an unresponsive shell around, and tightly bound to, the nucleus. The net effect of the closed shells is to reduce the effective potential generated by the nucleus. As for the valence electrons, we know with some certainty that light is emitted during readjustments in the outer charge distribution of the electron cloud.

Usually an atom exists with its associated electrons arranged in a stable configuration that corresponds to the lowest energy distribution. This energy distribution is known as the **ground state** configuration. Any mechanism that puts energy into the atom will alter the ground state. According to quantum mechanics, the electrons of an atom can only exist in certain specific configurations corresponding to certain discrete values of energy. In addition to the ground state, there are higher energy levels, known as **excited states**, each associated with a specific energy and a specific cloud configuration. When one or more electrons occupies a level higher than its ground state level, the atom is said to be excited.

At low temperatures, atoms tend to be in their ground state; at progressively higher temperatures, more and more of them will become excited through atomic collisions. This sort of mechanism is indicative of a class of relatively gentle excitations - glow discharge, flame, spark, and so forth - which energize only the outermost unpaired valence electrons.

When enough energy is imparted to an atom, whatever the cause, the atom can react by suddenly ascending from a lower to a higher energy level. The electron will usually make a very rapid transition, a quantum jump, from its ground state orbital configuration to one of the well defined excited states. As a rule, the amount of energy taken up in the process equals the energy difference between the initial and final states, and since this is specific and well defined, the amount of energy that can be absorbed by an atom is quantized. This state of atomic excitation is a short-lived resonance phenomenon. Usually, after about 10^{-8} or 10^{-9} seconds, the excited atom spontaneously relaxes back to a lower state, losing excitation energy along the way. This energy readjustment can occur by way of the emission of light, or (especially in dense materials) by conversion to thermal energy through interatomic collisions within the medium.

If the atomic transition is accompanied by the emission of light, the energy of the photon exactly matches the quantized energy decrease of the atom. That corresponds to a specific frequency, by way of $\Delta E = h\nu$, a frequency associated with both the photon and the atomic transition between the two particular states. This is said to be a **resonance frequency**.

The emission spectra of single atoms or low-pressure gases, whose atoms do not interact appreciably, consist of sharp lines, that is, fairly well defined frequencies characteristic of the atoms. There is always some frequency broadening due to atomic motion, collisions, and so forth, so it is never precisely monochromatic. Generally however, the atomic transition from one level to another is characterized by the emission of a well-defined narrow range of frequencies. On the other hand, the spectra of solids and liquids, in which the atoms now interact with each other, is broadened into wide frequency bands. When two atoms are brought close together, the result is a slight shift in their respective energy levels, because they act upon each other. The many interacting atoms in a solid create a tremendous number of such shifted levels, in effect spreading out each of their original levels, blurring them into essentially continuous bands. Light emitted from a large assemblage of randomly oriented independent atoms will consist of wavetrains in all directions. Each one of these will bear no particular consistent phase relation with any of the others, nor will they share a common polarization.

Interference

In order to understand interference, recall that optical disturbances are described by second order, homogeneous, linear partial differential equations. This means that they obey the *principle of superposition*. Thus, the resultant electric field intensity, $\boldsymbol{E}$, at a point in space where two or more waves overlap is equal to the vector sum of the individual constituent disturbances. This leads us to say that **optical interference may be considered as an interaction of two or more light waves which yield a resulting flux that is different from the scalar sum of the component fluxes**.

We have previously considered the problem of the superposition of two scalar waves, and these results will again be applicable here. However, light is a vector phenomenon; both the electric and magnetic field are vector quantities. Understanding this added level of complexity is crucial to understanding many optical phenomena.

Starting with the principle of superposition, the electric field intensity at a particular point in space is generated by the various fields, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, ..., of the constituent sources,

$$\vec{E} = \vec{E}_1 + \vec{E}_2 + \vec{E}_3 + \cdots \tag{10}$$

For the sake of simplicity, consider two point sources, S_1 and S_2, emitting monochromatic waves of the same frequency in a homogenous medium. Let their separation a be much greater than λ. Locate the point of observation, P, far enough away from the sources so

that at *P* the wavefronts will be planes. For now, consider only linearly polarized waves of the form

$$\vec{E}_1(\vec{r},t) = \vec{E}_{0,1}\cos\left(\bar{k}_1\cdot\bar{r}-\omega t+\varepsilon_1\right) \tag{11}$$

and

$$\vec{E}_2(\vec{r},t) = \vec{E}_{0,2}\cos\left(\bar{k}_2\cdot\bar{r}-\omega t+\varepsilon_2\right). \tag{12}$$

The irradiance at *P* is given by

$$\begin{aligned} I &= \varepsilon v\left\langle \vec{E}^2\right\rangle \\ &\propto \left\langle \vec{E}^2\right\rangle \end{aligned}. \tag{13}$$

Recognizing that $\left\langle \vec{E}^2\right\rangle$ is the time average of the square of the magnitude of the electric field intensity, we see that

$$\begin{aligned} I &= \left\langle \vec{E}^2\right\rangle \\ &= \left\langle \left(\vec{E}_1+\vec{E}_2\right)^2\right\rangle \\ &= \left\langle \left(\vec{E}_1+\vec{E}_2\right)\cdot\left(\vec{E}_1+\vec{E}_2\right)\right\rangle \\ &= \left\langle \vec{E}_1^2\right\rangle + \left\langle \vec{E}_2^2\right\rangle + 2\left\langle \vec{E}_1\cdot\vec{E}_2\right\rangle \\ &= I_1 + I_2 + I_{12} \end{aligned} \quad . \tag{14}$$

The last term is known as the **interference term**. For the waves described by (11) and (12), this can be evaluated as follows. First, consider the effect of $\vec{E}_1\cdot\vec{E}_2$:

$$\begin{aligned} \vec{E}_1\cdot\vec{E}_2 &= \vec{E}_{0,1}\cdot\vec{E}_{0,2}\cos\left(\vec{k}_1\cdot\vec{r}-\omega t+\varepsilon_1\right)\cos\left(\vec{k}_2\cdot\vec{r}-\omega t+\varepsilon_2\right) \\ &= \vec{E}_{0,1}\cdot\vec{E}_{0,2}\left[\cos\left(\vec{k}_1\cdot\vec{r}+\varepsilon_1\right)\cos\omega t+\sin\left(\vec{k}_1\cdot\vec{r}+\varepsilon_1\right)\sin\omega t\right]. \\ &\quad\times\left[\cos\left(\vec{k}_2\cdot\vec{r}+\varepsilon_2\right)\cos\omega t+\sin\left(\vec{k}_2\cdot\vec{r}+\varepsilon_2\right)\sin\omega t\right] \end{aligned} \tag{15}$$

Recall that the time average of a function *f(t)*, taken over an interval *T*, is

$$\left\langle f(t)\right\rangle = \frac{1}{T}\int_t^{t+T} f(t')\,dt' .$$

The period τ of a harmonic function is $2\pi/\omega$; for this problem $T >> \tau$. After multiplying out and averaging equation (15) we have

$$\left\langle \vec{E}_1 \cdot \vec{E}_2 \right\rangle = \tfrac{1}{2} \vec{E}_{0,1} \cdot \vec{E}_{0,2} \cos\left(\vec{k}_1 \cdot r + \varepsilon_1 - \vec{k}_2 \cdot \vec{r} - \varepsilon_2 \right), \tag{16}$$

where we used the fact that $\left\langle \cos^2 \omega t \right\rangle = \tfrac{1}{2}$, $\left\langle \sin^2 \omega t \right\rangle = \tfrac{1}{2}$, and $\left\langle \cos \omega t \sin \omega t \right\rangle = 0$. The interference term is then

$$I_{12} = \vec{E}_{0,1} \cdot \vec{E}_{0,2} \cos \delta, \tag{17}$$

where $\delta = \vec{k}_1 \cdot r + \varepsilon_1 - \vec{k}_2 \cdot \vec{r} - \varepsilon_2$ is the **phase difference**. It comes from the combined path length and the initial phase angle difference.

Parallel and Equal Amplitudes

We can simplify our results in the case of parallel amplitudes, $\vec{E}_{0,1} || \vec{E}_{0,2}$. In this case, equation (17) reduces to

$$I_{12} = E_{0,1} E_{0,2} \cos \delta. \tag{18}$$

Using the fact that

$$I_1 = \left\langle \vec{E}_1^2 \right\rangle = \frac{E_{0,1}^2}{2}$$

and

$$I_2 = \left\langle \vec{E}_2^2 \right\rangle = \frac{E_{0,2}^2}{2},$$

this can be rewritten as

$$I_{12} = 2\sqrt{I_1 I_2} \cos \delta, \tag{19}$$

so that the total irradiance becomes

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta. \tag{20}$$

This reaches a maximum value of

$$\begin{aligned} I_{\max} &= I_1 + I_2 + 2\sqrt{I_1 I_2} \\ &= \left(\sqrt{I_1} + \sqrt{I_2}\right)^2 \end{aligned} \tag{21}$$

when $\delta = 2m\pi$, where $m = 0, \pm 1, \pm 2, \ldots$. In this case the distributions are said to be *in phase*. This is known as **total constructive interference**. When $0 < \cos\delta < 1$, the waves are *out of phase*, $I_1 + I_2 < I < I_{\max}$, and the condition is known as **constructive interference**. At $\delta = \pi/2$, $\cos\delta = 0$, the optical disturbances are said to be **90° out of phase**. For $0 > \cos\delta > -1$ we get **destructive interference**, $I_1 + I_2 > I > I_{\min}$. The minimum occurs when

$$\begin{aligned} I_{\min} &= I_1 + I_2 - 2\sqrt{I_1 I_2} \\ &= \left(\sqrt{I_1} - \sqrt{I_2}\right)^2 \end{aligned} \tag{22}$$

when $\delta = (2m+1)\pi$, where $m = 0, \pm 1, \pm 2, \ldots$. This is known as **total destructive interference**.

Another special case is when the amplitudes of both waves are equal. In this case the irradiances from both sources are equal, so let $I_1 = I_2 = I_0$. Equation (22) can then be written as

$$\begin{aligned} I &= 2I_0(1 + \cos\delta) \\ &= 4I_0 \cos^2 \frac{\delta}{2} \end{aligned} \tag{23}$$

from which it follows that $I_{\min} = 0$ and $I_{\max} = 4I_0$.

Conditions for Interference

We have now discussed how two waves overlap to create an interference, or **fringe**, **pattern**. In order for this pattern to be observed, the two sources do not need to be in phase with each other. If there is some constant initial phase difference between the two sources, the resulting interference pattern will be identical to the original pattern, although it will be shifted in terms of the location of the minima and maxima. Such sources are said to be **coherent**. Remember that conventional quasimonochromatic sources produce light which is a mix of photon wavetrains. At each illuminated point in space there is a net field which oscillates through approximately a million cycles, which averages 10 ns or less, before it randomly changes phase. This interval over which the light wave resembles a sine function is a measure of its **temporal coherence**. Since the average time interval during which the wavetrain oscillates in a predictable manner is given by the **coherence**

time, we see that the longer the coherence time, the greater the temporal coherence of the source.
In a similar way, if we observe the light wave from a fixed point in space, we see that it appears to be fairly sinusoidal for some number of oscillations between abrupt phase changes. The corresponding spatial extent over which the light wave oscillates in a regular, predictable way has already been identified as the **coherence length**. If we view the light beam as a progression of well defined sinusoidal wavegroups of average length Δx_C, whose phases are uncorrelated to one another, then we find that normal coherence lengths range from several millimeters for standard laboratory discharge tubes up to tens of kilometers for some lasers.

Two ordinary sources will normally maintain a constant relative phase for a time no greater than Δt_C, so the interference pattern that they produce will randomly shift around in space at an extremely rapid rate, averaging out and making it impractical to observe. Until the advent of the laser, it was generally accepted that no two individual sources would ever produce an observable interference pattern. The coherence time of lasers, however, is long enough so that interference of two independent lasers has been detected electronically.

If two beams are to interfere to produce a stable pattern, they must have nearly the same frequency. A significant frequency difference would result in a rapidly varying time dependent phase difference, which would cause I_{12} to average out to zero during the detection interval. If the sources both emit white light, the component reds will interfere with the reds, and the blues with the blues. A great many overlapping monochromatic patterns will be produces which combine to create a total white light pattern. This final pattern will not be as sharp or extensive as a monochromatic pattern, but white light will produce observable interference.

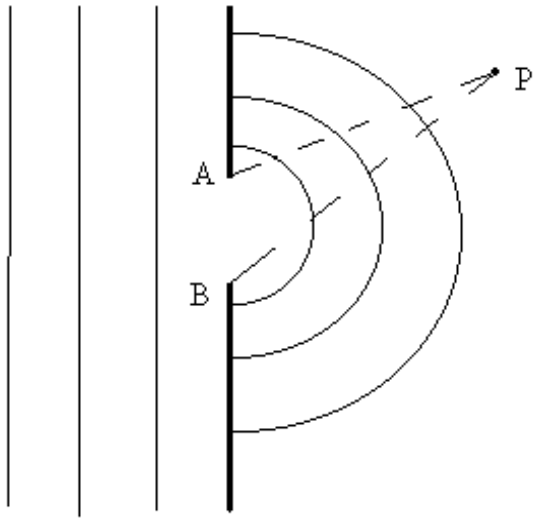
Diffraction

If we look at the shadow cast by an opaque object, we would find that it is very intricate. In fact, the shadow would consist of bright and dark regions which are not expected from everyday geometrical optics. This is known as **diffraction**, and it was first shown in the 1600s to be **a general characteristic of wave phenomenon which occurs whenever a portion of a wavefront is obstructed in some way**. In particular, if a wave encounters an obstacle, then diffraction occurs when a region of the wavefront is altered in amplitude or phase.

It is important to realize that there is not physical difference between interference and diffraction. However, it is traditional to consider a phenomenon as interference when it involves the superposition of only a few waves, and as diffraction when a large number of waves are involved. Another aspect that is important to understand is the fact that every optical instrument only uses a portion of the full incident wavefront. Because of this, diffraction plays a significant role in the detailed understanding of the light train through

the device. Even in all of the potential defects in the lens system were eliminated, the ultimate sharpness of the image would be limited by diffraction.
In order to begin to understand diffraction, let's return to Huygen's principle. Recall that this told us that each point on a wavefront can be viewed as a source of secondary spherical wavelets. From this, the progress of the wavefront as it moves through space can theoretically be determined. At any particular time, the shape of the wavefront is made up from the envelope of the secondary wavelets. There is a problem with this approach. In only considering the envelope of the secondary wavelets, Huygen's principle ignores most of the secondary wavelet and retains only the portion which is common to the envelope. As a result of this, Huygen's principle is unable to account for the details of the diffraction process. An example of this can be seen by comparing radio and visible light waves. Radio waves are seen to "bend" around large objects, such as buildings and telephone poles, but visible light creates a fairly distinct shadow. Huygen's principle is independent of any wavelength consideration and predicts the same wavefront configuration in both situations.

This problem was resolved when Fresnel added to Huygen's principle with the idea of interference. The resulting principle, known as the **Huygens-Fresnel principle**, states that **every unobstructed point of a wavefront, at a given instant in time, serves as a source of spherical secondary wavelets, with the same frequency as that of the primary wave. The amplitude of the optical field at any point beyond is the superposition of all these wavelets, taking into consideration their amplitudes and relative phases**. As an example of this, consider the following drawing



Define the maximum optical path length difference as $\Lambda_{\max} = \left|\overline{AP} - \overline{BP}\right|$. Assume that $\overline{AB} \geq \Lambda_{\max}$. Then when $\lambda >> \overline{AB}$, we also have that $\lambda >> \Lambda_{\max}$. Since the waves were initially in phase, they must all interfere constructively, no matter where *P* happens to be.

On the other hand, when $\lambda << \overline{AB}$, the area where $\lambda >> \Lambda_{\max}$ is limited to a small region extending out directly in from of the aperture, and it is only there that all of the wavelets interfere constructively. Beyond this region, some of the wavelets can interfere destructively. This is the geometric shadow. Remember that the idealized geometric shadow corresponds to $\lambda \to 0$.

Fraunhofer Diffraction

Consider the case where the point of observation is very distant from the array line and $R >> D$. Then $r(y)$ does not deviate very significantly from R. In this case, Eq. (14.8) becomes

$$E = \frac{\varepsilon_L}{R} \int_{-D/2}^{D/2} \sin\left[kr(y) - \omega t\right] dy \tag{24}$$

If we write r as an explicit function of y, we get

$$r = R - y\sin\theta + \frac{y^2}{2R}\cos^2\theta + \cdots \tag{25}$$

where θ is measured from the x-z plane. The non-linear terms in y can be ignored when their contribution to the phase is insignificant. This is true whenever $\left(\pi D^2 / 4\lambda R\right)\cos^2\theta + \cdots$ is negligible; a condition that is satisfied for all values of θ whenever R is large. This is known as the **Fraunhofer condition**, where the distance r is linear in y. In turn, this leads to the fact that the distance to the point of observation, and thus the phase, can be written as a linear function of the aperture variables.

Returning to Eq. (25), we see that Eq. (24) becomes

$$E = \frac{\varepsilon_L}{R} \int_{-D/2}^{D/2} \sin\left[k\left(R - y\sin\theta\right) - \omega t\right] dy$$

which can be integrated to yield

$$E = \frac{\varepsilon_L D}{R} \frac{\sin\left[\dfrac{kD}{2}\sin\theta\right]}{\dfrac{kD}{2}\sin\theta} \sin\left(kR - \omega t\right). \tag{26}$$

In order to simplify (26), define

$$\beta = \frac{kD}{2}\sin\theta \tag{27}$$

Then

$$E = \frac{\varepsilon_L D}{R}\frac{\sin\beta}{\beta}\sin\left(kR-\omega t\right).$$

From *E*, the irradiance can be determined to be

$$\begin{aligned} I(\theta) &= \left\langle E^2 \right\rangle \\ &= \left(\frac{\varepsilon_L D}{R}\right)^2\left(\frac{\sin\beta}{\beta}\right)^2\left\langle \sin^2\left(kR-\omega t\right)\right\rangle . \\ &= \frac{1}{2}\left(\frac{\varepsilon_L D}{R}\right)^2\left(\frac{\sin\beta}{\beta}\right)^2 \end{aligned} \tag{28}$$

Notice that when $\theta = 0$, $\left(\sin\beta/\beta\right) = 1$ and $I(\theta)$ is a maximum. This maximum is known as the **principal maximum**, and the **irrandiance resulting from an idealized coherent line source in the Fraunhofer approximation** becomes

$$I(\theta) = I(0)\left(\frac{\sin\beta}{\beta}\right)^2. \tag{29}$$

Since $\beta = \frac{kD}{2}\sin\theta$, when $D >> \lambda$, the irradiance drops extremely rapidly as θ deviates from zero. Also, when $D >> \lambda$, the source, which is a relatively long coherent line source, can be viewed as a single point emitter radiating predominately in the forward direction. When the opposite is true, namely that $\lambda >> D$, then β is small and the irradiance remains essentially constant for all values of θ. This means that the line source more closely resembles a point source emitting spherical waves.

The Circular Aperture

Fraunhofer diffraction through a circular aperture can be found in a manner similar to that used for the rectangular aperture. In this case, instead of using rectangular coordinates, the symmetry of the situation dictates the use of cylindrical coordinates. Thus Eq. (26) becomes

$$E = \frac{\varepsilon_A e^{i(kR-\omega t)}}{R}\int_0^a\int_0^{2\pi} e^{i\left(kq\rho/R\right)\cos(\phi-\Phi)}\rho\, d\rho\, d\phi, \tag{30}$$

where

$$z = \rho\cos\phi \quad y = \rho\sin\phi$$
$$Z = q\cos\Phi \quad Y = q\sin\Phi \,.$$

From symmetry, the result must not depend on Φ, so we can set it to zero in Eq. (30). Consider the azimuthal integral first. The quantity

$$J_0(u) = \frac{1}{2\pi}\int_0^{2\pi} e^{iu\cos v}\,dv \tag{31}$$

is known as a **Bessel function of the first kind**. Comparing it to the azimuthal integral in (30), we see that

$$E = \frac{\varepsilon_A e^{i(kR-\omega t)}}{R} 2\pi \int_0^a J_0\left(\frac{kq\rho}{R}\right)\rho\,d\rho \tag{32}$$

Using the recurrence relationship for Bessel functions,

$$\int_0^u u'\,J_0(u')du' = uJ_1(u)$$

where

$$J_1(u) = \frac{-i}{2\pi}\int_0^{2\pi} e^{i(v+u\cos v)}dv\,,$$

this can be evaluated as

$$E = \frac{\varepsilon_A e^{i(kR-\omega t)}}{R} 2\pi a^2\left(\frac{1}{ka\sin\theta}\right)J_1(ka\sin\theta) \tag{33}$$

where the relationship $\sin\theta = \frac{q}{R}$ was used. The irradiance becomes

$$I(\theta) = 2\left(\frac{\varepsilon_A A}{R}\right)^2\left[\frac{J_1(ka\sin\theta)}{ka\sin\theta}\right]^2. \tag{34}$$

At the center of the aperture, the irradiance is

$$I(0) = \frac{1}{2}\left(\frac{\varepsilon_A A}{R}\right)^2 \tag{35}$$

and so Eq. (35) becomes

$$I(\theta) = I(0)\left[\frac{2J_1(ka\sin\theta)}{ka\sin\theta}\right]^2. \tag{36}$$

Resolution of circular images

The center of the aperture has a large circular maximum. This maximum is known as the **Airy disk**. The size of the Airy disk can be used to determine the maximum resolution of a lens system. For simplicity, consider two incoherent distant point sources of equal irradiance. The radius of the Airy disk is given by

$$q_1 = 1.22\frac{f\lambda}{D}. \tag{37}$$

If $\Delta\theta$ is the corresponding angular measure, then, using the fact that

$$\frac{q_1}{f} = \sin\Delta\theta \approx \Delta\theta,$$

we find that

$$\Delta\theta = 1.22\frac{\lambda}{D}$$

The Airy disk for each source will be spread out over a half width $\Delta\theta$, centered on the geometric image point. If the angular separation of the two points is $\Delta\varphi$, and if $\Delta\varphi >> \Delta\theta$ the images will be distinct and easily resolved. As the two sources approach each other, their respective images would also approach each other, overlap, and blend into a single set of fringes. We can use Lord Rayleigh's criterion to determine when the two objects are just resolved. This criterion states that the resolution of two fringes of equal flux density requires that the principal maximum of one coincide with the first minimum of the other. Using this criterion, the two objects are **just resolved** when the center of one Airy disk falls on the first minimum of the Airy pattern of the other object. Thus, the **angular limit of resolution** is

$$(\Delta\varphi)_{\min} = 1.22\frac{\lambda}{D}. \tag{38}$$

Fresnel Diffraction

In Fraunhofer diffraction, the diffracting system was relatively small and the point of observation was very distant. This allowed the potential problems associated with the Huygens-Fresnel principle to be completely passed over. Now we are concerned with the near-field region, which extends right up to the diffracting element itself.

We must reconsider the Huygens-Fresnel principle more closely to understand what is happening in this region. Recall that we can envision every point on the primary wavefront as a continuous emitter of spherical secondary wavelets. However, if each wavelet is radiating uniformly in all directions, then there would be a wave traveling back towards the source in addition to the normal outgoing wave. Since no such wave is found experimentally, we must somehow modify the radiation pattern of the secondary emitters. This can be done by introducing the **obliquity**, or **inclination factor**, $K(\theta)$. The obliquity is used to describe the directionality of the secondary emission. Kirchoff was the first person to analytically define the obliquity as

$$K(\theta) = \tfrac{1}{2}(1+\cos\theta) \tag{39}$$

where θ is the angle made with the normal, ***k***, to the primary wavefront.

Consider a spherical wave emitted from a point *S* at a time $t = 0$. A time *t'* later, the wave has a radius of ρ and is described by

$$E = \frac{\mathcal{E}_0}{\rho}\cos(k\rho - \omega t'). \tag{40}$$

We can divide the wavefront into a series of annular regions. The boundaries of the various regions correspond to the intersections of the wavefront with a series of spheres centered at some observation point, *P*, with radii given by

$$R = r_0 + \sum_{n=1}^{\infty}\frac{n\lambda}{2}$$

where r_0 is the minimum distance from *P* to the wavefront. These spheres are known as the **Fresnel**, or **half period**, **zones**. Since each zone is finite in extent, we can define a ring shaped differential area element *dS* associated with the zone. All of the point sources within *dS* are coherent, and we can assume that each one radiates in phase with the primary wave. Thus, in any zone each of the secondary wavelets travel a distance *r* to reach *P* at a time *t*, and all of the wavelets arrive there with the same phase, $k(\rho + r) - \omega t$.

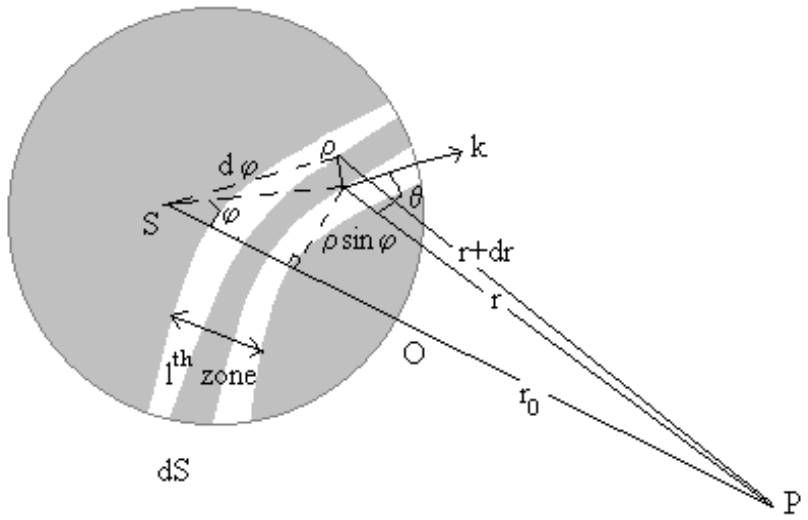
We can assume that the source strength per unit area ε_A of the secondary emitters on *dS* is proportional to the amplitude of the primary wave,

$$\varepsilon_A = Q\frac{\varepsilon_0}{\rho}.$$

The contribution to the optical disturbance at *P* from the secondary sources on *dS* is therefore

$$dE = K\frac{\varepsilon_A}{\rho}\cos\left[k(\rho + r) - \omega t\right]dS. \tag{41}$$

The obliquity factor must vary slowly and thus can be assumed to be constant over a single Fresnel zone. Consider the following drawing



The area element *dS* is seen to be

$$dS = 2\pi\rho(\rho\sin\varphi)d\varphi$$

which, combined with the law of cosines

$$r^2 = \rho^2 + (\rho + r_0)^2 - 2\rho(\rho + r_0)\cos\varphi$$
$$\Rightarrow 2rdr = 2\rho(\rho + r_0)\sin\varphi d\varphi$$

yields

$$dS = 2\pi \frac{\rho}{\rho + r_0} r dr \tag{46}$$

Substituting this into Eq. (45) and integrating yields

$$\begin{aligned} E_\ell &= 2\pi K_\ell \frac{\varepsilon_A}{(\rho + r_0)} \int_{r_{\ell-1}}^{r_\ell} \cos\left[k(\rho + r_0) - \omega t\right] dr \\ &= -\frac{K_\ell \varepsilon_A \rho \lambda}{(\rho + r_0)} \left[\sin(k\rho + kr - \omega t)\right]_{r=r_{\ell-1}}^{r=r_\ell} \\ &= (-1)^{\ell+1} \frac{2K_\ell \varepsilon_A \rho \lambda}{(\rho + r_0)} \sin\left[k(\rho + r_0) - \omega t\right] \end{aligned} \tag{47}$$

If there are a total of *m* zones on the wavefront, then the sum of the optical disturbances from all *m* zones at *P* is

$$\begin{aligned} E &= E_1 + E_2 + E_3 + \cdots + E_m \\ &= |E_1| - |E_2| + |E_3| - \cdots \pm |E_m| \end{aligned} \tag{48}$$

If *m* is odd, the series can be written in one of two ways. The first way is

$$\begin{aligned} E = \frac{|E_1|}{2} &+ \left(\frac{|E_1|}{2} - |E_2| + \frac{|E_3|}{2}\right) + \left(\frac{|E_3|}{2} - |E_4| + \frac{|E_5|}{2}\right) + \cdots \\ &+ \left(\frac{|E_{m-2}|}{2} - |E_{m-1}| + \frac{|E_m|}{2}\right) + \frac{|E_m|}{2} \end{aligned} \tag{49}$$

while the second is

$$\begin{aligned} E = |E_1| &- \frac{|E_2|}{2} - \left(\frac{|E_2|}{2} - |E_3| + \frac{|E_4|}{2}\right) - \left(\frac{|E_4|}{2} - |E_5| + \frac{|E_6|}{2}\right) + \cdots \\ &+ \left(\frac{|E_{m-3}|}{2} - |E_{m-2}| + \frac{|E_{m-1}|}{2}\right) - \frac{|E_{m-1}|}{2} + |E_m| \end{aligned} \tag{50}$$

This means that either $|E_\ell| > \dfrac{|E_{\ell-1}| + |E_{\ell+1}|}{2}$ or $|E_\ell| < \dfrac{|E_{\ell-1}| + |E_{\ell+1}|}{2}$. Using Eqs. (49) and (50), these conditions become

$$E < \frac{|E_1| + |E_m|}{2} \tag{51}$$

and

$$E > \frac{|E_1| + |E_m|}{2} \tag{52}$$

from which is can be concluded that

$$E \approx \frac{|E_1|}{2} + \frac{|E_m|}{2} \tag{53}$$

If *m* is even, similar arguments leads to a result of

$$E \approx \frac{|E_1|}{2} - \frac{|E_m|}{2} \tag{54}$$

Fresnel showed that the last contributing zone satisfied

$$K(\theta) = 0 \quad \text{for} \quad \frac{\pi}{2} \le \theta \le \pi$$

so that (53) and (54) both reduces to

$$E \approx \frac{|E_1|}{2} \tag{55}$$

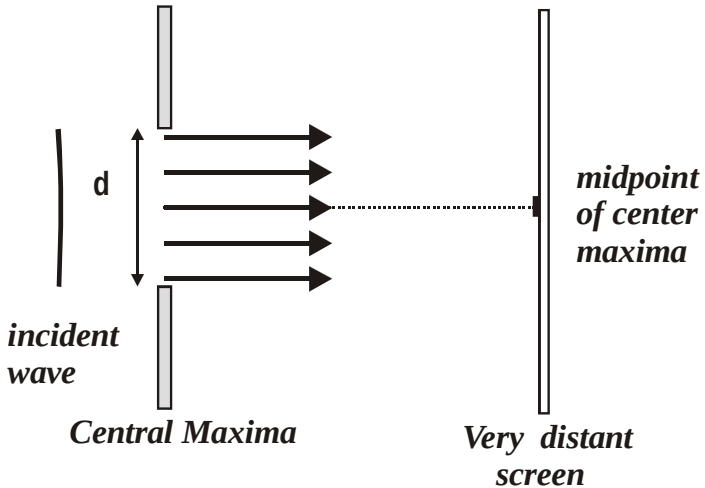
Thus, we see that **the optical disturbance generated by the entire unobstructed wavefront is approximately equal to half the contribution from the first zone**.

Single Slit Diffraction: A single slit will also form an interference pattern when light passes through it.

Each part of the slit acts as source of waves. This is described in Huygen's principle.

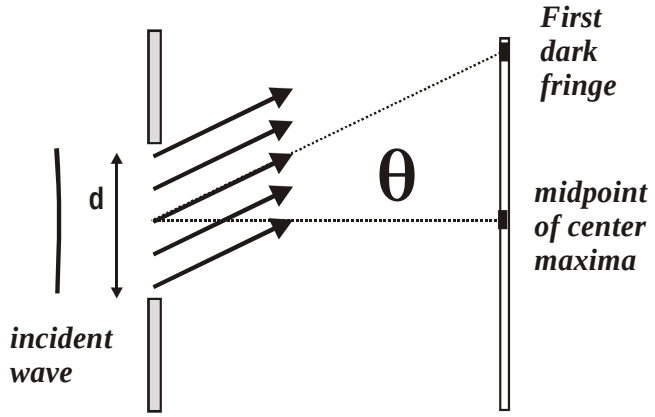
> ***Huygen's principle ≡ Every point on a wave front acts as a source of tiny wavelets that move forward with the same speed as the wave. The wave front at a later instant of time is the surface that is tangent to the wavelets.***

You can imagine that across the width of the slit, little wavelets originate and travel through the slit. These waves pass through the slit and form a bright central fringe on the

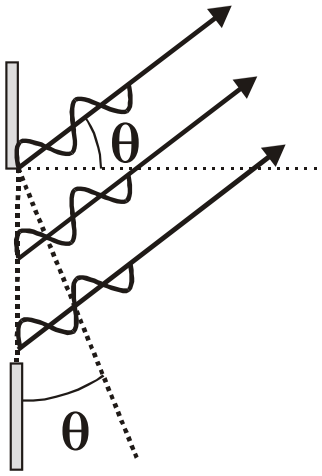


screen, which is at a far distance from the slit. This distance is so far that all the waves are essentially parallel to one other. All the wavelets travel the same distance and arrive at the screen in phase with each other and we get constructive interference. This creates a bright central fringe at the center of the screen directly opposite the slit.

The wavelets that originate in the slit can also interfere destructively. Here's a drawing that shows the very thing.



Wavelet fromation for dark fringe



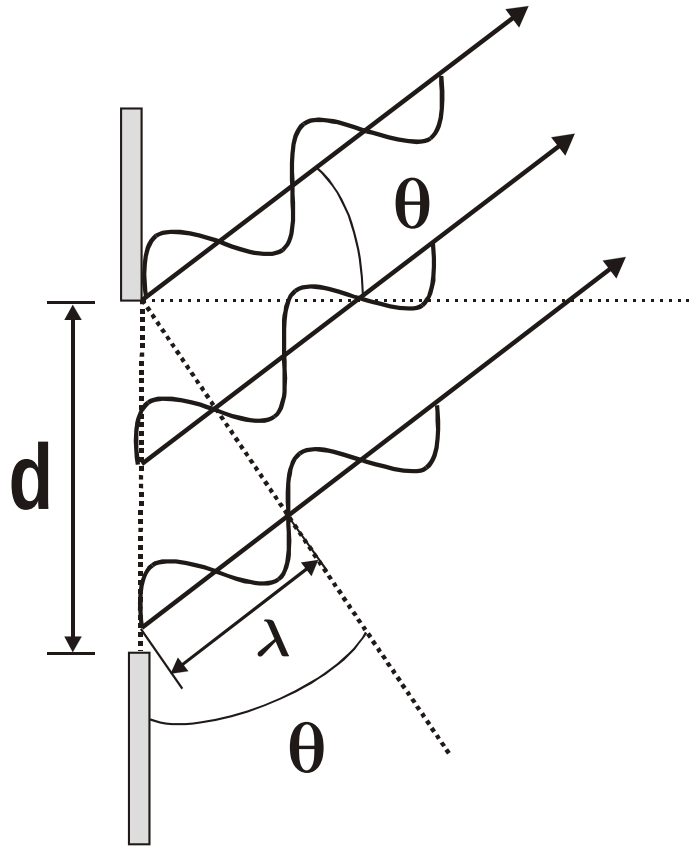
How interference takes place

Light from one part of slit interferes with light from another part of slit, forming the patterns. Again the cause of the interference is the path difference for the waves (wavelets in this case).

The patterns that form can be described in this way:

There will be a bright central fringe surrounded by two dark fringes, then a set of weaker bright fringes, a set of dark fringes, a set of weaker (than the central fringe) bright fringes, and so on.

Let's look at the geometry of the thing.



The same equation is used with single slit diffraction as with the double slit diffraction, except that the angle we get, θ, is the angle from the center of the slit to the center of the dark fringe.

$$d\sin\theta = m\lambda$$

This equation describes destructive interference.

> d is the width of the slit, θ is the angle to the center of the dark fringe, m is the integer order number, and λ is the wavelength of the light.

We can analyze it the same way we did the double slit deal to find the spacings between the central fringe and the dark fringes.

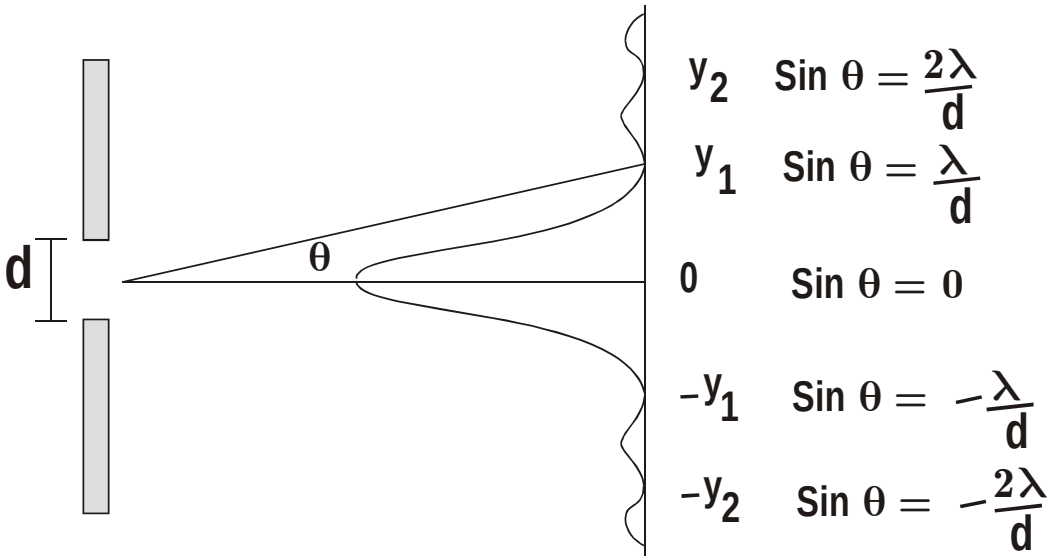
$$\sin\theta = \frac{\lambda}{d} \qquad \sin\theta \cong \tan\theta = \frac{y}{L}$$

$$\frac{y}{L} = \frac{\lambda}{d} \qquad y = \frac{\lambda L}{d}$$

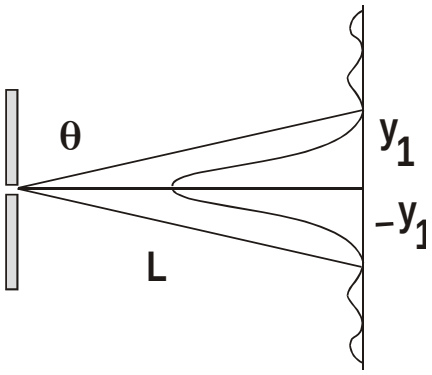
This general case is:

$$x_m \approx \frac{m\lambda L}{d}$$

While this is the same equation as for a double slit deal, for a single slit it gives you the distance from the center of the bright central fringe to the ***desired dark fringe***.



- 575 nm light passes through a slit of width 0.250 mm. An observing screen is set up 3.00 m away. (a) Find the position of the first dark fringe. (b) What is the width of the central maxima?



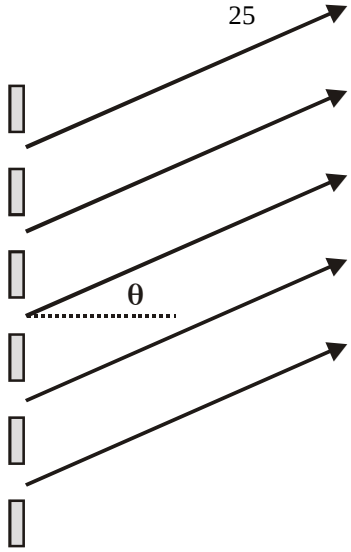
(a) This is the first minima, so ***m*** = 1. The spacing is given by:

$$x_m \approx \frac{m\lambda L}{d}$$

$$y_1 = \frac{(1)575\,x\,10^{-9}\,\cancel{m}\,(3.00\,m)}{0.250\,x\,10^{-3}\,\cancel{m}} = 6900\,x\,10^{-6}\,m = \boxed{6.90\,mm}$$

Diffraction Grating:

Diffraction gratings are a recent invention (well, a lot more recent than the old double slit deal). Basically, a diffraction grating is a piece of transparent material that has parallel cuts scribed in it. The scribings are so small that you can't really see them. At any rate, the grating has a very large number of equally spaced parallel slits cut into it. This would be on the order of hundreds to several thousand lines per centimeter.



The grating acts like a double slit setup. It produces a large number of very bright, sharp fringes separated from one another by fairly wide minima.

Maxima are given by the same equation as we have seen before:

$$d \sin\theta = m\lambda$$

m is the order number, d is the spacing between the slits, λ is the wavelength of the light, and θ is the angle formed by a normal to the grating to a line at the center of the fringe.

- Light from a distant star enters a telescope and then passes through a diffraction grating onto a screen. A first order red line appears on the screen at an angle of 25.93°. The lines of the grating are separated by 1.50 x 10^{-6} m. What is the wavelength of the light?

$$d \sin\theta = m\lambda \qquad \lambda = d \sin\theta$$

$$\lambda = \left(1.50\, x 10^{-6} m\right) \sin 25.93^{\circ} \quad = 0.656\, x 10^{-6} m$$

$$\lambda = 656\, x 10^{-9} m \quad = 656\, nm$$

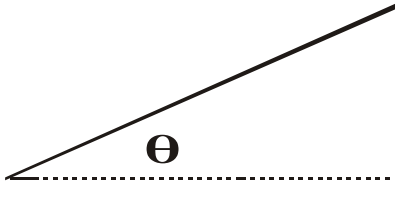
- 632.8 nm laser beam passes through a diffraction grating that has 6 000.0 lines per centimeter. An observing screen is set up 3.00 m away. Find separation of the maxima.

The slit separation is the inverse of slit density.

$$d = \frac{1}{6\,000} cm \quad = 1.667\, x 10^{-4} cm \quad = 1.667\, x 10^{-6} m$$

$$d \sin\theta = m\lambda \qquad \sin\theta = \frac{\lambda}{d}$$

The angle θ is not small and we cannot make the assumption that the sine of θ is equal to the tangent.

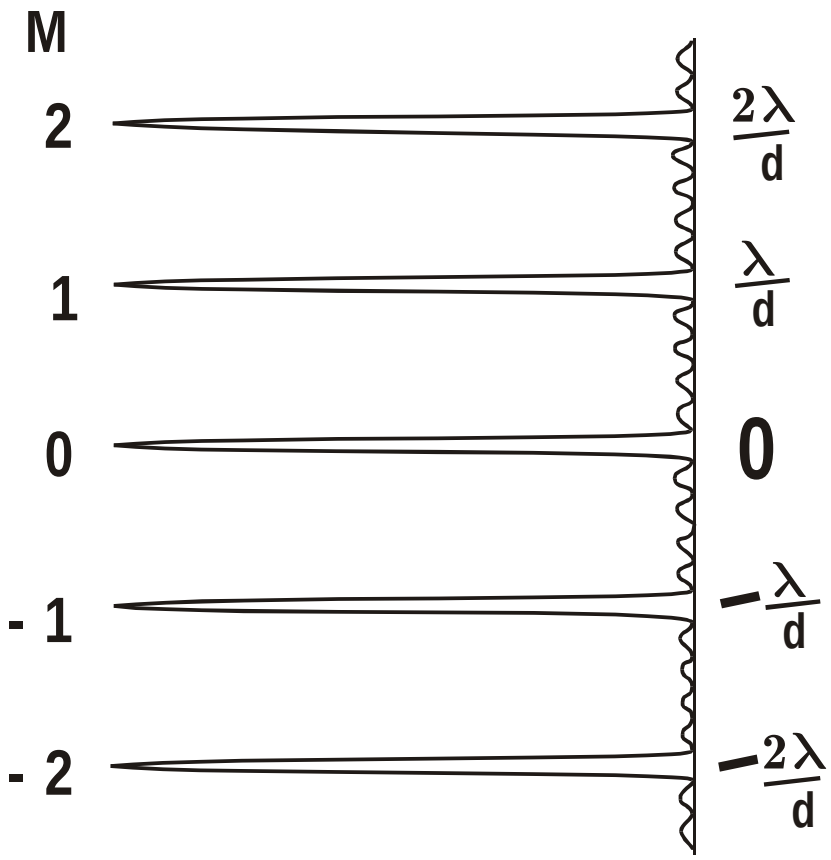


We can find θ from the equation and then use the tangent to find ***y***.

$$\sin\theta = \frac{\lambda}{d} \quad = \frac{632.8\,x\,10^{-9}\,\cancel{m}}{1.667\,x\,10^{-6}\,\cancel{m}} \quad = 379.6\,x\,10^{-3} \quad = 0.3796 \qquad \theta = 22.3^{o}$$

$$\tan\theta = \frac{y}{L} \qquad y = (\tan\theta)L \quad = \left(\tan 22.3^{o}\right)(3.00\,m) \quad = \boxed{1.23\,m}$$

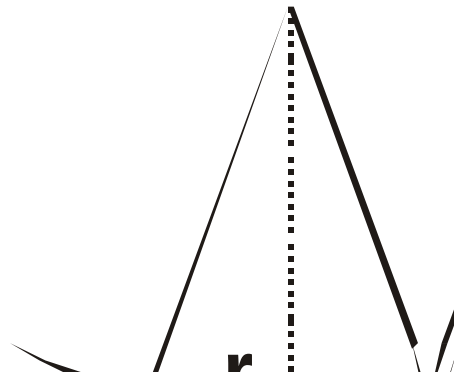
The pattern of maxima from a diffraction grating looks like this:



The maxima are very bright and sharp, they are also widely separated from one another. For this reason, the diffraction grating is preferred to double slits. So diffraction gratings are better because:

- Get very sharp maxima
- Get very wide dark areas

With a few easy to make measurements, one can easily calculate the wavelength of the light.



Propagation of Light using Geometry

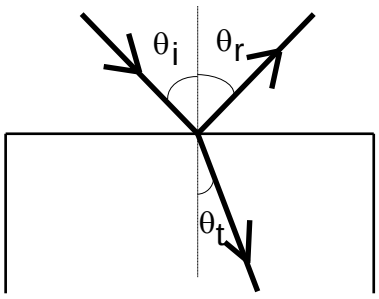
The treatment of light as wave motion allows for a region of approximation in which the wavelength is considered to be negligible compared with the dimensions of the relevant components of the optical system. This region of approximation is called **geometrical optics**. When the wave character of the light may not be ignored, the field is known as **physical optics**. Since the wavelength of light is very small compared to ordinary objects, early unrefined observations of the behavior of a light beam passing through apertures or around obstacles in its path could be handled by geometrical optics.

Within the approximation represented by geometrical optics, light is understood to travel out from its source along straight lines or **rays**. The ray is simply the path along which energy is transmitted from one point to another in an optical system. The ray is a useful, although abstract, construct; perhaps the best approximation to a ray of light is a pencil-like laser beam. When a light ray traverses an optical system consisting of several homogeneous media in sequence, the optical path is a sequence of straight-line segments. The laws of geometrical optics that describe the subsequent direction of the rays are succinctly stated as:

Law of Reflection: When a ray of light is reflected at an interface dividing two uniform media, the reflected ray remains within the **plane of incidence**, and the angle of reflection equals the angle of incidence. The plane of incidence includes the incident ray and the normal to the point of incidence.

Law of Refraction (Snell's Law): When a ray of light is refracted at an interface dividing two uniform media, the transmitted ray remains within the plane of incidence and the sine of the angle of refraction is directly proportional to the sine of the angle of incidence.

These laws can be visually seen in the following figure



$$\theta_i = \theta_r$$

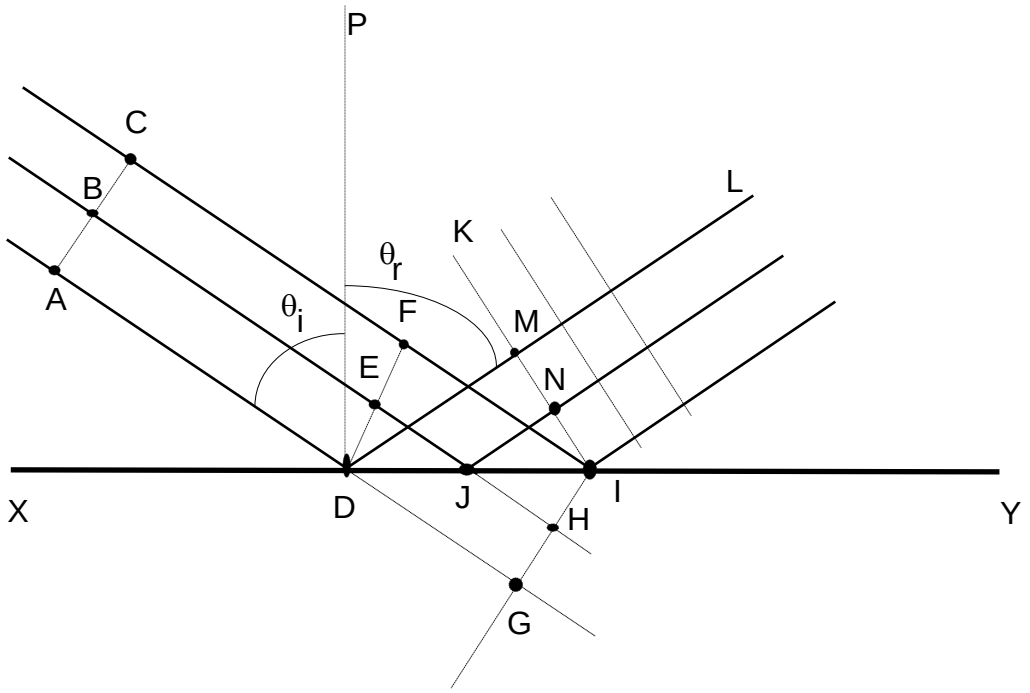
$$\frac{\sin\theta_i}{\sin\theta_t} = \text{constant}$$

Huygens' Principle

The Dutch physicist Christian Huygens imagined each point of a propagating disturbance as capable of originating new pulses that contributed to the disturbance an instant later. To show how his model of light propagation implied the laws of geometrical optics, he

formulated a principle which says that **each point on the leading surface of a wave disturbance may be regarded as a secondary source of spherical waves, which themselves progress with the speed of light in the medium and whose envelope at later times constitutes the new wavefront**. Notice that the new wavefront is tangent to each wavelet at a single point. According to Huygens, the remainder of each wavelet is to be disregarded in the application of the principle. In so disregarding the effectiveness of the overlapping wavelets, Huygens avoided the possibility of diffraction of the light into the region of geometric shadow. Huygens also ignored the wavefront formed by the back half of the wavelets, since these wavefronts implied a light disturbance traveling in the opposite direction. Despite weaknesses in this model, remedied later by Fresnel and others, Huygens was able to apply his principle to prove the laws of both refection and refraction.

Law of Reflection from Huygen's Principle

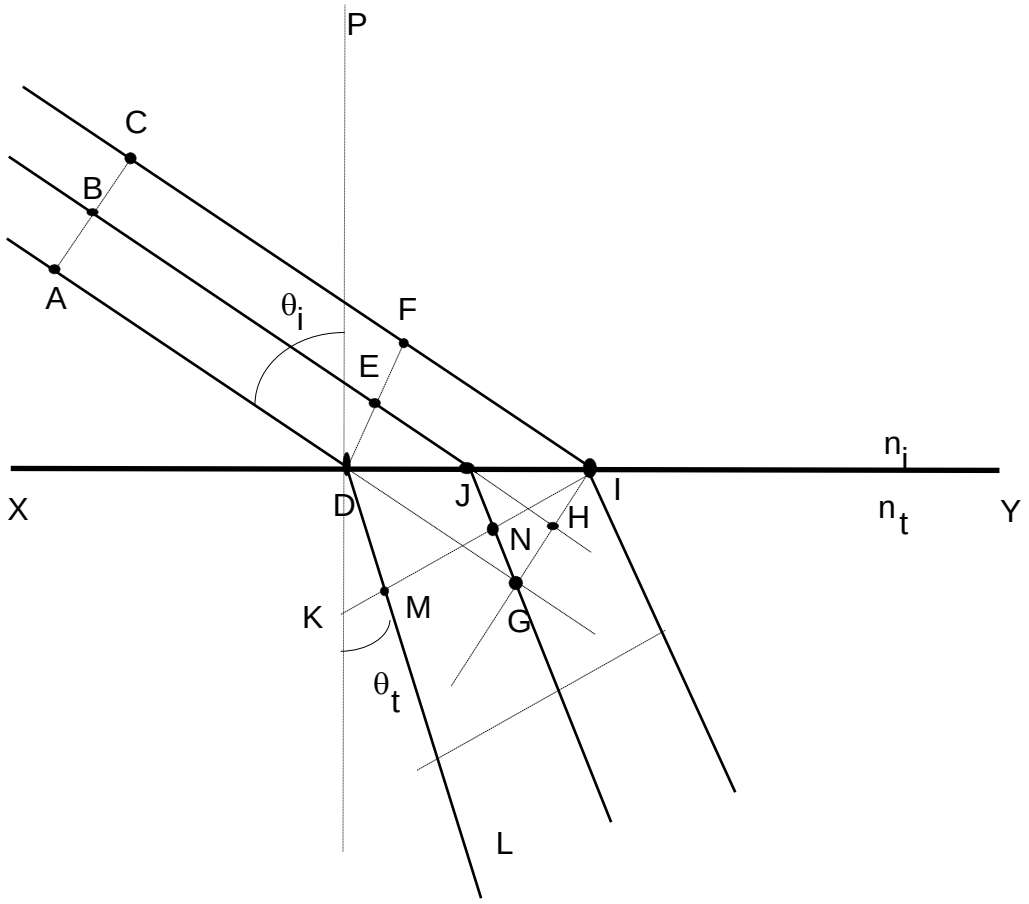


The figure illustrates Huygen's construction for a narrow, parallel beam of light to prove the law of reflection. Huygen's principle must be modified to accommodate the case in which a wavefront, such as *AC*, encounters a plane interface, such as *XY*, at an angle. Here the angle of incidence of the rays *AD*, *BE*, and *CF* relative to the perpendicular *PD* is θ_i. Since points along the plane wavefront do not arrive at the interface simultaneously, allowance is made for these differences in constructing the wavelets that determine the reflected wavefront. If the interface *XY* were not present, the Huygens construction would produce the wavefront *GI* at the instance ray *CF* reached the interface at *I*. The intrusion of the reflecting surface, however, means that during the same time interval required for ray *CF* to progress from *F* to *I*, ray *BE* has progressed from *E* to *J* and then a distance equivalent to *JH* after reflection. Thus a wavelet of radius *JH* centered at *J* is

drawn above the reflecting surface. Similarly, a wavelet of radius *DG* is drawn centered at *D* to represent the propagation after reflection of the lower part of the beam. The new wavefront, which must now be tangent to these wavelets at points *M* and *N*, and include the point *I*, is shown as *KI* in the figure. A representative reflected ray is *DL*, shown perpendicular to the reflected wavefront. The normal *PD* drawn for this ray is used to define angles of incidence and reflection for the beam. The construction clearly shows the equivalence between the angles of incidence and reflection.

Law of Refraction using Huygen's Principle

Similarly, we can use a Huygens construction to illustrate the law of refraction.



Here we must take into account a different speed of light in the upper and lower media. If the speed of light in vacuum is *c*, we express the speed in the upper medium by the ratio c/n_i, where n_i is the refractive index. Similarly, the speed of light in the lower medium is c/n_t. The points *D*, *E* and *F* on the incident wavefront arrive at points *D*, *J* and *I* of the plane interface *XY* at different times. In the absence of the refracting surface, the wavefront *GI* is formed at the instant ray *DF* reaches *I*. During the progress of ray *CF* from *F* to *I* in time *t*, however, the ray *AD* has entered the lower medium, where the speed

is different. Thus if the distance DG is $v_i t$, a wavelet of radius $v_t t$ is constructed with center at D. The radius DM can also be expressed as

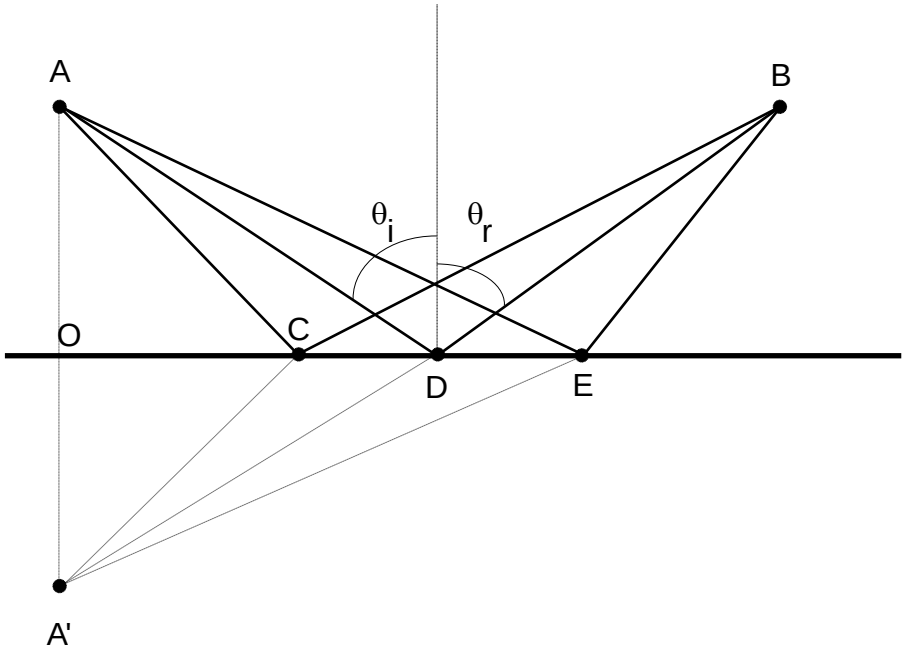
$$\begin{aligned} DM &= v_t t \\ &= v_t \left(\frac{DG}{v_i} \right) \\ &= \left(\frac{n_i}{n_t} \right) DG \end{aligned}$$

Similarly, a wavelet of radius $(n_i/n_t)JH$ is drawn centered at J. The new wavefront KI includes point I on the interface and is tangent to the two wavelets at points M and N. The geometric relationship between the angles θ_i and θ_t, formed by the representative incident ray AD and refracted ray DL, is **Snell's law**, which may be expressed as

$$n_i \sin \theta_i = n_t \sin \theta_t \tag{56}$$

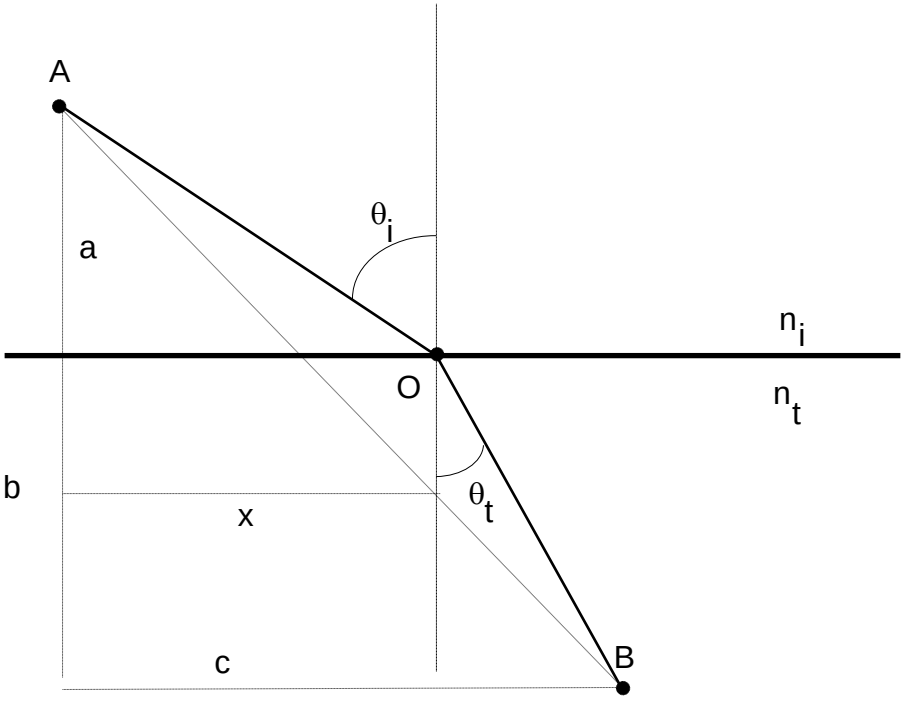
Fermat's Principle

The laws of geometrical optics can also be derived from a different fundamental hypothesis. Let us suppose that nature is economical, and thus requires that the time required for light to travel from point A to B is the minimum time required. To prove the law of reflection, we use the fact that, for propagation in the same medium, the velocity is a constant and this minimizing the time is the same as minimizing the distance traveled. Consider the following drawing



Three possible paths from *A* to *B* are shown. Let's look at the arbitrary path *ACB*. If point *A'* is constructed on the perpendicular *AO* such that *AO* = *A'O*, the right triangles *AOC* and *A'OC* are equal. Thus *AC* = *A'C* and the distance traveled by the ray of light from *A* to *B* via *C* is the same distance from *A'* to *B* via *C*. The shortest distance from *A'* to *B* is obviously the straight line *A'DB*, so the path *ADB* is the correct choice taken by the actual light ray. Geometry shows that for this path, $\theta_i = \theta_r$. Also note that to maintain *A'DB* as a single straight line, the reflected ray must remain within the plane of incidence.

We can also prove the law of refraction. If the light travels more slowly in the second medium, light bends at the interface so as to take a path that favors a shorter time in the second medium, thereby minimizing the overall transit time from *A* to *B*.



Mathematically, we are required to minimize the total time

$$t = \frac{AO}{v_i} + \frac{OB}{v_t} = \frac{\sqrt{a^2 + x^2}}{v_i} + \frac{\sqrt{b^2 + (c - x)^2}}{v_t} \tag{57}$$

Since other choices of path change the position of the point *O* and therefore the distance *x*, we can minimize the time by setting $dt/dx = 0$:

$$\begin{aligned} 0 &= \frac{dt}{dx} \\ &= \frac{x}{v_i\sqrt{a^2+x^2}} - \frac{c-x}{v_t\sqrt{b^2+(c-x)^2}} \\ &= \frac{\sin\theta_i}{v_i} - \frac{\sin\theta_t}{v_t} \end{aligned} \tag{58}$$

where the last step used the relationships shown in the figure. Introducing the refractive indices of the media, we arrive at Snell's law

$$n_i \sin\theta_i = n_t \sin\theta_t \tag{59}$$

Fermat's principle, like that of Huygens, required refinement to achieve more general applicability. Situations exist where the actual path taken by a light ray may represent a maximum time or even one of many possible paths, all requiring equal time. As an example of the latter case, consider light propagating from one focus to the other inside an ellipsoidal mirror, along any of an infinite number of possible paths. Since the ellipse is the locus of all points whose combined distances from the two foci remain constant, all paths are indeed of equal time. A more precise statement of Fermat's principle, which requires merely an extremum relative to neighboring paths, may be given as follows: **The actual path taken by a light ray in its propagation between two given points in an optical system is such as to make its optical path equal, in the first approximation, to other paths closely adjacent to the actual one**.

With this formulation, Fermat's principle falls in the class of problems called variational calculus, a technique which determines the form of a function that minimizes a definite integral. In optics, the definite integral is the integral of the time required for the transit of a light ray from starting to finishing points.

Optical Path Length

Suppose that we have a stratified material composed of m layers, each having a different index of refraction. The transit time across the layers will then be

$$\begin{aligned} t &= \frac{s_1}{v_1} + \frac{s_2}{v_2} + \cdots + \frac{s_m}{v_m} \\ &= \sum_{i=1}^{m} \frac{s_i}{v_i} \\ &= \frac{1}{c}\sum_{i=1}^{m} n_i s_i \end{aligned} \tag{60}$$

where the summation is called the **optical path length** traversed by the ray. Clearly for an inhomogeneous medium where *n* is a function of position, the summation must be changed to an integral

$$(OPL) = \int n(s)ds$$

Since the optical path length is related to the time, we can restate Fermat's principle again as **a light ray in going from point *A* to point *B* must traverse an optical path length that is stationary with respect to variations of that path**.

Optical Reversibility

Consider applying Fermat's principle to an optical system. Since the time must be minimized, we see that the same path is predicted regardless of whether we start at *A* and travel to *B*, or start at *B* and travel to *A*. In general, any actual ray of light in an optical system, if reversed in direction, will retrace the same path backward. Before discussing the formation of images in a general way, let's look at the simplest — and experimentally, the most accessible — case of images formed by plane mirrors. In this context it is important to distinguish between **specular reflection** from a perfectly smooth surface and **diffuse reflection** from a granular or rough surface. Specular reflection occurs when all the rays of a parallel beam incident on the surface obey the law of reflection from a plane surface and therefore reflect as a parallel beam. In the case of diffuse reflection, although the law of reflection holds locally, the microscopically granular surface results in reflected rays in various directions and thus a diffuse scattering of the originally parallel rays of light. Every plane surface will produce some such scattering, since a perfectly smooth surface is not obtainable in reality. In many cases, however, the diffuse scattering is small and we can approximate the reflection as specular reflection.

Consider the specular reflection of a single light ray from the *x-y* plane. By the law of reflection, the reflected ray remains within the plane of incidence, making equal angles with the normal at the point of contact. If the path is resolved into components, it is clear that the direction of the incident ray is altered only by reflection along the *z* direction, and then in such a way that its *z* component is simply reversed. If the direction of the incident ray is described by its unit vector $\hat{r}_1 = (x, y, z)$, then the reflection causes

$$\hat{r}_1 = (x, y, z) \rightarrow \hat{r}_2(x, y, -z) \tag{61}$$

It follows that if a ray is incident from such a direction as to reflect sequentially from all three coordinate planes, then

$$\hat{r}_1 = (x, y, z) \rightarrow \hat{r}_2(-x, -y, -z) \tag{62}$$

and the ray returns precisely parallel to the line of its original approach. A network of such corner reflectors ensures the exact return of a beam of light.