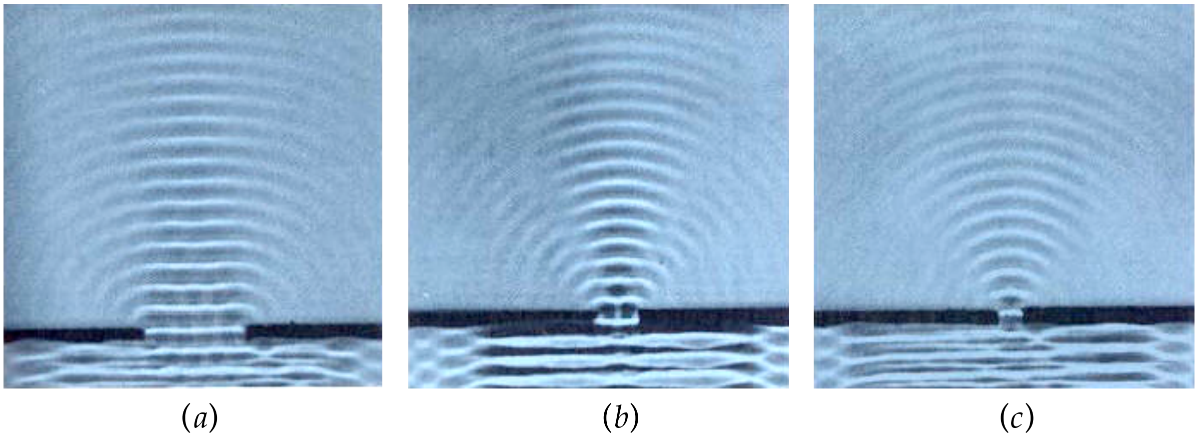
Wave Optics 2

The only cure for boredom is curiosity. There is no cure for curiosity.
— Dorothy Parker

Overview: obstructing part of a wave causes diffraction

When part of a wavefront is obstructed, the progress of the wave energy into the region beyond the obstruction is determined (according to Huygens's principle) by the waves emitted from points on the unobstructed part of the wavefront. (The rest of the energy is either reflected back or absorbed by the obstruction.) The waves from the unobstructed points spread out, sending some energy into the geometric shadow region. This energy deviates from the directions of the original rays, "bending" into the shadow. This is the phenomenon of **diffraction**. It is a general property of waves, including water waves.



Shown are water waves in a ripple tank, viewed from above. The waves, which have the parallel wave fronts characteristic of plane waves, move up the page, striking the dark wall; the wall's opening allows a small part of the wave to pass through. In (a) the opening is about 4 times the wavelength; in (b) it is twice the wavelength; in (c) is is equal to the wavelength. In (a) nearly all of the wave energy goes in the original direction and there is a fairly clear shadow effect. In (b) a significant amount of the energy deviates into the shadow region. In (c) even more is deviated into the shadow. Note that in (b) there are directions where none of the energy goes; the directional intensity pattern has both maxima and minima.

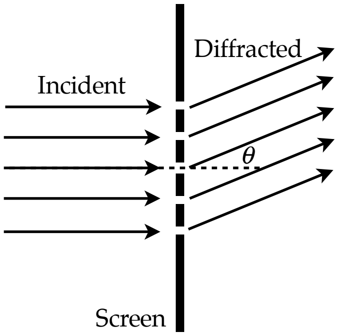
If the size of obstacles or apertures encountered by the wave is large compared to the wavelength, nearly all of the energy goes in straight lines along the original rays. This is why we see fairly sharp shadows of everyday objects illuminated by visible light, and why the ray approximation give good results in many cases. But if one looks carefully enough it will be found that the shadows do not really have sharp boundaries. And in situations where the aperture or obstacle size is comparable to the wavelength, diffraction becomes a major influence in distributing the energy.

The simplest situation mathematically (which we will analyze in detail here) is one where the incoming waves can be approximated by plane waves, and where the intensity pattern is detected far enough from the obstacle or aperture that the waves can again be treated as plane waves. This case is **Fraunhofer diffraction**.

The more complicated case, where the spherical nature of the wavefronts must be taken into account, is **Fresnel diffraction**. We will not carry out the difficult mathematical analysis for that case, but will discuss some of its characteristic phenomena.

Fraunhofer diffraction by rectangular slits

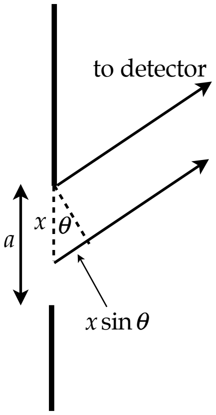
To illustrate the phenomenon and the method of analysis, we consider an opaque screen (shown from the side) in which has been cut a set of identical parallel long narrow rectangular slits running perpendicular to the page. Light is incident on the screen from the left, at normal incidence. Since the light consists of plane waves the rays are parallel. After passing through the slits, the light is diffracted. We seek the intensity of the light moving off at angle θ relative to the original direction.



To calculate the intensity at a distant detector we must find the E-field at that point. This field is the superposition of the fields from all the point sources (according to Huygens's principle) on the parts of the wavefront that pass through the slits. These sources emit waves that begin in phase at the slits, but they travel different distances to the detector, so they arrive with different phases and interfere to produce the detected intensity.

First we consider waves emanating from various point *within* one of the slits. Let the slits have width a. Let the distance to the detector from the *top* edge of the *top* slit be R. (See the diagram.) Now consider the waves emanating from an infinitesimal part of the wavefront of size dx, located vertically at the distance x from the top of the slit. The waves in this infinitesimal part travel distance $R + x\sin\theta$ to the detector, so they all arrive in phase with each other. Their total contribution to the E-field at the detector is thus proportional to dx. We describe their contribution to the detected E-field by

$$dE_1(x) = K\,dx \cdot \cos[k(R + x\sin\theta) - \omega t].$$



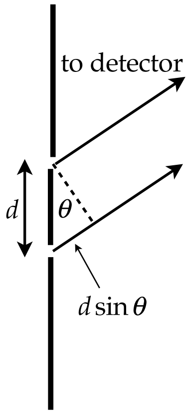
(The subscript 1 indicates the top slit.) Here K is a constant depending on the intensity of the incident light. To find the total E-field at the detector due to light from the entire top slit, we integrate this expression over x from 0 to the width a of the slit. If we had only one slit, this would give the total E-field at the detector.

But we have N identical parallel slits, each pair separated by distance d. Consider light from corresponding points in the top two slits. As the diagram shows, the light from the second slit travels an extra distance $d\sin\theta$ to get to the detector. Thus the wave from the corresponding part of the second slit is described by

$$dE_2(x) = K\,dx \cdot \cos[k(R + x\sin\theta + d\sin\theta) - \omega t].$$

For the nth slit the contribution is

$$dE_n(x) = K\,dx \cdot \cos[k(R + x\sin\theta + (n-1)d\sin\theta) - \omega t].$$



The total E-field at the detector from all the slits is therefore

$$E(\theta)=K\int_0^a dx\cdot\sum_{n=1}^{N}\cos[k(R+x\sin\theta+(n-1)d\sin\theta)-\omega t].$$

This gives us the field we need. The problem is to evaluate the sum and the integral. It looks formidable, but it turns out not to be too hard, if we use the Euler trick.

We replace the cosines by complex exponentials and use the factoring properties of exponentials:

$$E^c(\theta)=Ke^{i(kR-\omega t)}\int_0^a dx\,e^{ikx\sin\theta}\cdot\sum_{n=1}^{N}e^{ik(n-1)d\sin\theta}\;.$$

Carrying out the sum and the integral are relatively straightforward exercises.

> The integral is just an exponential. The sum is a geometric series, for which $\sum_{n=0}^{N}x^n=\frac{1-x^N}{1-x}$.

We will simply quote the result (after making some use of Euler's formula):

$$E^c(\theta)=Ke^{i(kR-\omega t)}aN\frac{e^{iN\beta}}{e^{i\beta}}\cdot\frac{\sin\alpha}{\alpha}\cdot\frac{\sin N\beta}{N\sin\beta}.$$

Here we have introduced two new angles to shorten the writing:

$$\alpha=\tfrac{1}{2}ka\sin\theta,\;\;\beta=\tfrac{1}{2}kd\sin\theta.$$

We are now ready to calculate the intensity. We multiply this expression for $E^c(\theta)$ by its complex conjugate to obtain the squared amplitude of the E-field; the intensity is a constant times this. We find

$$I(\theta)=\text{const}\cdot\left(\frac{\sin\alpha}{\alpha}\right)^2\cdot\left(\frac{\sin N\beta}{N\sin\beta}\right)^2.$$

> The complex exponentials all disappear when multiplied by their complex conjugates.

To evaluate the constant, we note that as $\theta\to 0$ (which means both $\alpha\to 0$ and $\beta\to 0$ so we can use the small angle approximations) the two factors in the formula both approach 1, so the constant is just $I(0)$. This gives us the final answer:

Fraunhofer pattern for N slits	$I(\theta)=I(0)\cdot\left(\frac{\sin\alpha}{\alpha}\right)^2\cdot\left(\frac{\sin N\beta}{N\sin\beta}\right)^2$ $\alpha=\tfrac{1}{2}ka\sin\theta,\;\beta=\tfrac{1}{2}kd\sin\theta$

This is a complicated formula, but it contains a great deal of information. The intensity at the detector depends not only on the angle θ at which the light is detected, but also on the wavelength of the light (represented by $k=2\pi/\lambda$), the width of each slit (a), the number of slits (N), and their separation (d).

> The total amplitude of the waves is proportional to the area through which the light passes, which is proportional to the slit width a and the number N of slits. So $I(0)$ is proportional to $(Na)^2$.

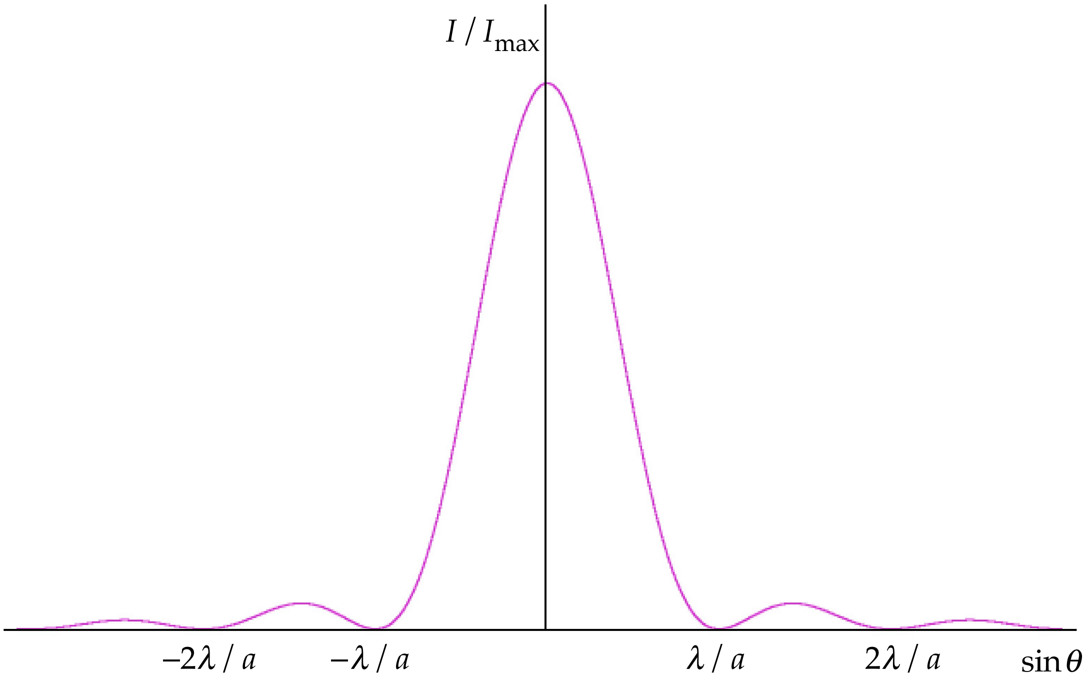
We will examine the pattern described by the formula for various values of N.

Single slit pattern

If $N = 1$ the last factor in the formula is 1 and we have

Fraunhofer single slit pattern	$I(\theta) = I(0)\cdot\left(\frac{\sin\alpha}{\alpha}\right)^2$

Since $\sin\alpha/\alpha$ has its largest value (1) when $\alpha = 0$ (i.e., when $\theta = 0$), we see that $I(0)$ is the largest intensity for any angle, so we call it $I_{\max}$. The curve shows $I(\theta)$:



The minima in this pattern occur when α is a multiple of π (except zero), i.e., when

$$\text{Minima: } a\sin\theta = \pm\lambda, \pm 2\lambda, \ldots$$

As the slit is made narrower the angular distance between the minima on both sides of the central peak becomes larger: a narrower slit gives a wider pattern, and vice versa.

If the aperture is *circular*, the formula is more complicated, but the general features are the same. The intensity pattern is a bright central circle peaked at $\theta = 0$, surrounded by concentric alternating dark and bright circular bands of decreasing intensity. The first dark band occurs at an angle given approximately by

$$\text{Circular opening: } \theta_1 \approx 1.22\lambda/D$$

where D is the diameter of the circular opening.

This formula is important in determining the resolution of optical instruments.

Multiple slits

If $N > 1$ we must take into account the second factor in the general formula. It gives zero intensity whenever its numerator vanishes but its denominator does not. This occurs if $N\beta$ is a multiple of π but β is not, that is, if

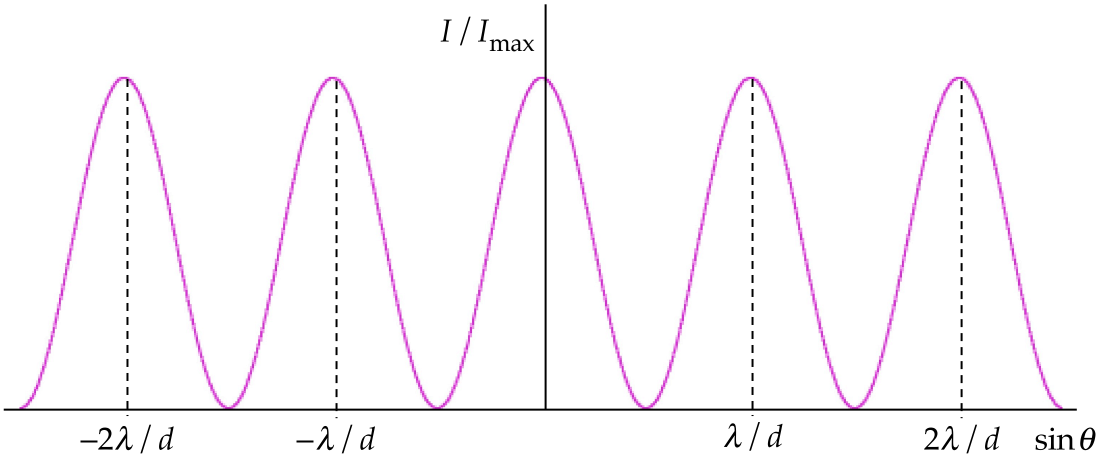
$$N\beta = \pm\pi, \pm 2\pi, \ldots, \pm(N-1)\pi, \pm(N+1)\pi, \ldots$$

These locate the *minima* in the pattern. When β is a multiple of π (including zero), l'Hospital's rule shows that the factor is 1 and we have a **principal maximum**. The condition $\beta = m\pi$ is usually written out as follows:

Principal maxima	$d \sin\theta = m\lambda,\ m = 0, \pm 1, \pm 2 \ldots.$

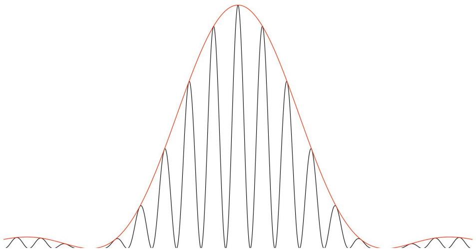
The main feature of the pattern consists of the principal maxima, where the peak intensity is the maximum allowed by the $(\sin\alpha / \alpha)^2$ factor. There are $N-1$ minima and $N-2$ smaller "subsidiary" maxima between successive principal maxima.

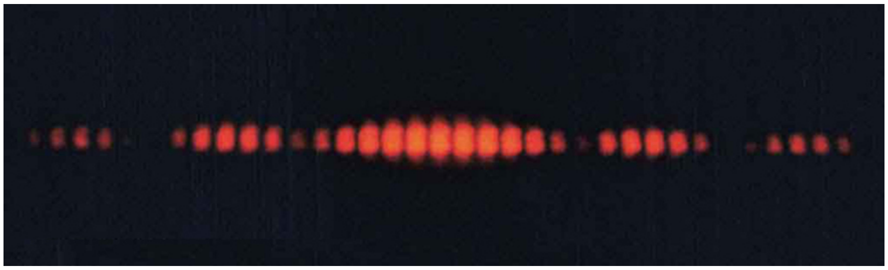
For $N = 2$ there are only principal maxima, with one minimum between each pair. This pattern is shown (we assume a is so small that $\sin\alpha / \alpha \approx 1$).



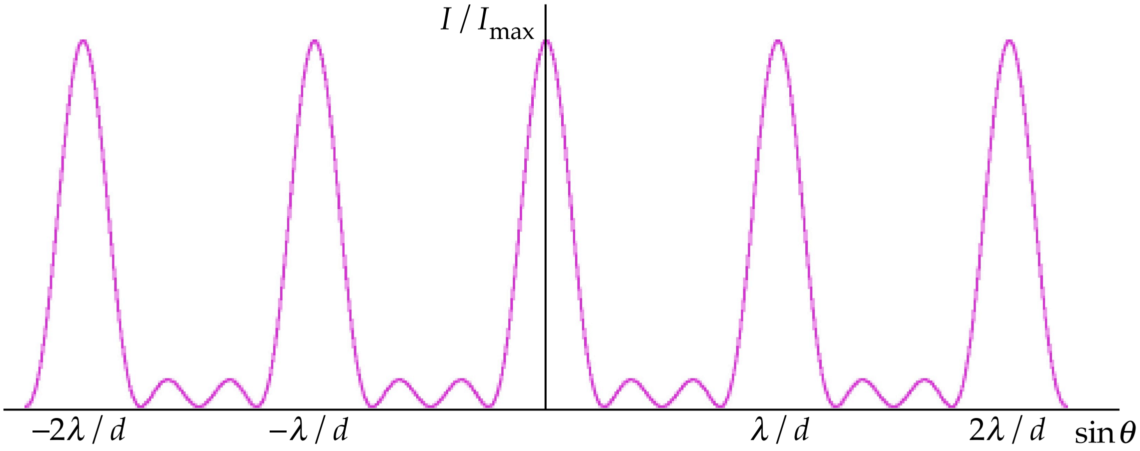
As the separation d between the slits increases, the peaks move closer together.

If the slit width a is not very small, then the intensity pattern is "modulated" by the $(\sin\alpha / \alpha)^2$ term, shown as the colored curve in the drawing. The principal maximum peaks to either side of the central one (at $\theta = 0$) are correspondingly reduced in height. For the drawing it was assumed that $d = 6a$, so the 6th principal maximum is missing because it coincides with the first minimum in the envelope curve. A photograph of this pattern, made by red laser light passing through vertical slits, is shown below:





The pattern for $N = 4$ shows the subsidiary maxima and the fact that the principal peaks are narrower. For given d the principal maxima occur at the same angles, regardless of the value of N. (Again we have assumed that a is very small.)



As the number of slits N is increased, three things happen:

1. The principal maxima become narrower. The half-width (the angular distance from a principal maximum peak to the next minimum) is approximately λ / Nd.
2. The principal maxima become brighter, since $I_{\max}$ is proportional to N^2.
3. The fraction of the energy in the subsidiary maxima becomes smaller; for very large N they become so faint that only the principal maxima are visible.

Diffraction gratings

The bright and sharp principal maxima for very large N can be used to measure wavelengths of the light. Knowing d, one measures the the angular location θ of the maxima and uses the equation given above for the principal maxima. A set of slits (or rulings) used for this purpose is called a **diffraction grating**.

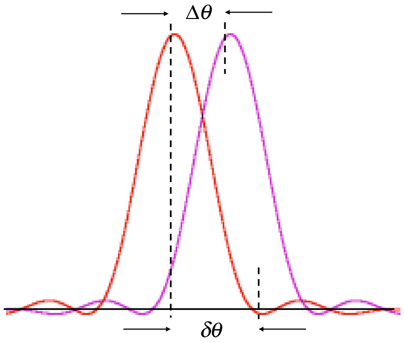
Since the light emitted by many sources consists of discrete wavelengths (spectral “lines”), some of which may have nearly the same wavelength, a desirable property of a grating is its ability to distinguish ("resolve") two nearly equal wavelengths. This is called the **resolving power** of the grating, defined as

$$R = \lambda_{av} / \Delta\lambda .$$

Here λ_{av} is the average wavelength in the region and $\Delta\lambda$ is the difference between the closest wavelengths that the grating can clearly distinguish as separate.

Under what circumstances can we tell that they are separate? We note that in the intensity pattern made by a grating one spectral line forms a diffraction pattern, with two characteristics: the location of the maximum intensity (“the peak”) and the spread of the pattern around that location (“the width”). The problem is that the patterns for nearly equal wavelengths may overlap too much, so we cannot tell from the pattern that there are two separate wavelengths.

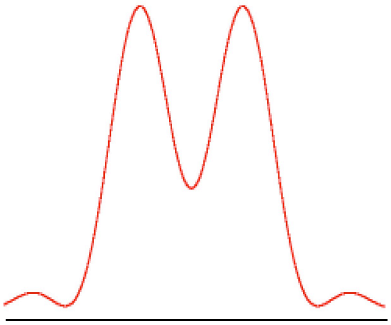
Consider principal maxima produced by two wavelengths, show in the diagram. The angular separation of their peaks is $\Delta\theta$. Each peak has a half-width $\delta\theta$, the angular distance from the peak to the first minimum on either side. If $\Delta\theta$ is too small the peaks overlap too much to be distinguished.



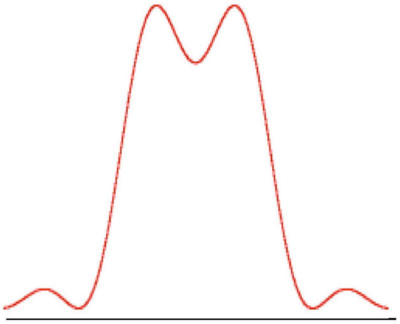
Rayleigh's Criterion. The criterion proposed by Rayleigh specifies how large $\Delta\theta$ must be if we are to distinguish the peaks as separate:

Rayleigh's criterion	Two diffraction peaks are resolved if the peak separation $\Delta\theta$ is at least equal to the peak half-width $\delta\theta$.

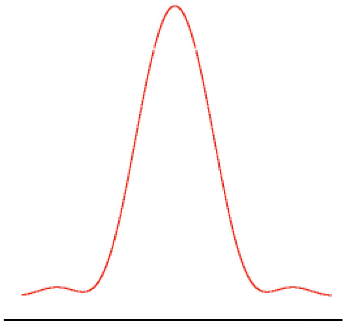
Shown below are plots of the total intensity (which is what a detector will record) for the two peaks in the drawing above, in three cases:



$\Delta\theta > \delta\theta$



$\Delta\theta = \delta\theta$



$\Delta\theta < \delta\theta$

In the left case the peaks are far apart and well resolved; in the center case they are barely resolved; in the right case they are too close together to be resolved.

> Rayleigh's criterion applies to the resolution of diffraction patterns, however they are formed. Here we are concerned with diffraction gratings, but the same criterion applies to the patterns of different points in an image formed by an optical instrument.

Now back to gratings. We consider two wavelengths, λ and $\lambda + \Delta\lambda$. Their mth principal maxima will occur at angles θ and $\theta + \Delta\theta$, respectively, where

$$d\sin\theta = m\lambda, \; d\sin(\theta + \Delta\theta) = m(\lambda + \Delta\lambda)$$

Here the integer m is called the **order number**, and the central maximum at $\theta = 0$ (for all wavelengths) is the case $m = 0$. If the peaks are barely resolved, then $\theta + \Delta\theta$ must also be the location of the first minimum beyond the mth principal maximum for wavelength λ. This minimum occurs when $\beta = (m + 1/N)\pi$, so we have

$$d\sin(\theta + \Delta\theta) = (m + 1/N)\lambda$$

Comparing with the second equation above, we see that $m\Delta\lambda = \lambda / N$, from which we find a formula for the resolving power:

Resolving power of a grating	$R = \lambda / \Delta\lambda = mN$

This shows that to get better resolution it is important to have many rulings, which is why professional gratings typically have $N \sim 10^5$. The spectrum is spread out more in higher orders, which is why R depends on m; but because $\sin\theta$ cannot exceed 1 there is a limit to how large m can be.

Resolution in Optical Instruments

Common optical instruments use circular openings through which light passes. As we have seen, the angular half-width of the Fraunhofer pattern of a circular opening is

$$\delta\theta = 1.22\lambda / D$$

where D is the diameter of the opening. When such instruments are used to make images, points in the original object become diffraction patterns in the image. If these patterns overlap too much the image does not reveal that the original had separate points: the points are not *resolved*. Here are photographs to show the situation.

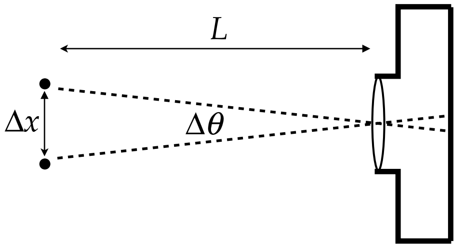
$\Delta\theta < \delta\theta$

$\Delta\theta > \delta\theta$

$\Delta\theta = \delta\theta$

In applying Rayleigh's criterion to such instruments, one takes $\Delta\theta$ to be the angle between the rays from the two objects as they enter the instrument through the center of the opening. For the images to be resolved, this angle must be at least as large as $\delta\theta$.

Consider a simple camera, as shown. Two points in the object are separated by distance Δx; the object is at distance L from the cameral lens. The angular separation of the image peaks on the film is the same as the angle subtended at the lens by the object points. For the images to be resolved, Rayleigh's criterion requires $\Delta\theta \geq 1.22\lambda / D$. If $\Delta x << L$ we have by the small angle approximations $\Delta\theta \approx \Delta x / L$, so



$$\Delta x \geq 1.22\lambda L / D.$$

Rayleigh's criterion specifies a limit on how close two object points can be and still give resolved images.

This restriction applies also to microscopes and telescopes, where it can be a serious limitation on the ability of the instrument to provide magnified images.

Fresnel diffraction

As often happens in science, the earliest example of diffraction to be studied by scientists was not the simplest analytically. Around 1815 Fresnel gave a mathematical theory of diffraction of light waves as they pass by a straight edge obstacle, showing that there are "fringes" (bands of alternating higher and lower intensity) near the edge of the shadow, while some light actually gets into the shadow region where ray geometry would predict it to be dark.

In his analysis, Fresnel could not use the plane wave approximation of Fraunhofer. As a result, the mathematics is much more complicated, but his more general version accounts for phenomena where the source and detector are *not* far from the apertures or obstacles. Because of its mathematical complexity, we will not discuss **Fresnel diffraction** quantitatively.

The mathematician Poisson (who thought the wave theory was wrong) showed that Fresnel's method also predicted a bright spot in the center of the shadow of an opaque disk, and a dark spot in the center of the pattern of light passing through a circular hole. These "absurd" predictions were soon confirmed experimentally by Arago and Fresnel, giving added strength to proponents of the wave theory.

> Fresnel was also the first — decades before Maxwell’s electromagnetic wave theory of light — to give the correct formulas (given in the notes “Light”) for the intensities of the reflected and refracted waves at an interface. He also invented a flattened lens, which bears his name, still used in theatrical lighting and lighthouses.