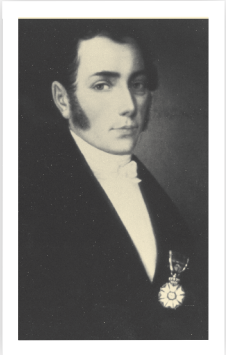
Diffraction, Part 2

MULTIPLE POINT SOURCES, APERTURES, AND DIFFRACTION LIMITS

by Dwight E. Neuenschwander, Southern Nazarene University



JOSEPH VON FRAUNHOFER (also Frauenhofer, 1787-1826). The namesake of diffraction by plane waves, Fraunhofer invented the diffraction grating, which transformed spectroscopy. With his spectrometer Fraunhofer discovered the dark lines in the solar spectrum, which now bear his name. Photo credit: Bavarian Academy of Sciences, courtesy AIP Emilio Segre Visual Archives.

If you have ever looked at a streetlight through an umbrella's fabric and seen a neat array of tiny bright spots; noticed thin streaks of light emanating from images of small, bright lights in photographs; or wondered how the metallic mesh in a microwave oven door allows visible light but not microwaves to pass, then you have encountered diffraction. Many physics experiments, from spectroscopy to measurements of the wavelength of laser light, employ diffraction. Diffraction is the signature phenomena of wave motion.[1]

In Part 1 of this series on diffraction,[2] we met Huygens' principle and applied it to the interference produced by two slits, modeled as point sources that coherently radiate equal-amplitude and equal-wavelength harmonic waves. The corresponding experiment, first done by Thomas Young in 1801, demonstrated that light is a wave, or, as we would say today, that light *behaves* as a wave in this situation. In our analysis of the Young experiment we observe waves sufficiently far from their source to make the wave-front curvature negligible across an aperture. Diffraction with such plane waves is called "Fraunhofer diffraction."

Working within the Fraunhofer paradigm, here we extend Young's experiment to multiple point sources. We will go to the limit of an infinite number of contiguous infinitesimal point sources to derive the diffraction patterns produced by a single slit as well as its complement, an opaque ribbon. That will put us in the position to consider double slits of finite width as a better model of Young's apparatus. The result illustrates the array theorem, which says that the image produced by an array of N identical apertures equals the diffraction pattern of one aperture times the interference pattern of N point sources. We will also consider diffraction from all four edges of a rectangular aperture and from a circular aperture.

Working within the Fraunhofer paradigm,

HERE WE EXTEND YOUNG'S EXPERIMENT TO MULTIPLE POINT SOURCES

Before leaving plane waves we will discuss what it means to say that the image on the screen is the Fourier transform of the aperture. For Fraunhofer diffraction, Fourier transforms provide the link between waves, aperture, and image.

INTERFERENCE FROM *N* POINT SOURCES

Consider three point sources equally spaced and separated by the distance a. Place a screen some distance z away (Fig. 1). Let locations on the screen be mapped by the coordinate y (or alternatively, the angle θ), where $y = 0$ ($\theta = 0$) describes the point opposite the source array's midpoint. Assume the three sources emit coherently, and their waves leave in phase.

A portion of the signal from each source arrives at location y on the screen at time t. The part of the total signal coming from source 1 travels along the ray of length r_1. The signal from source

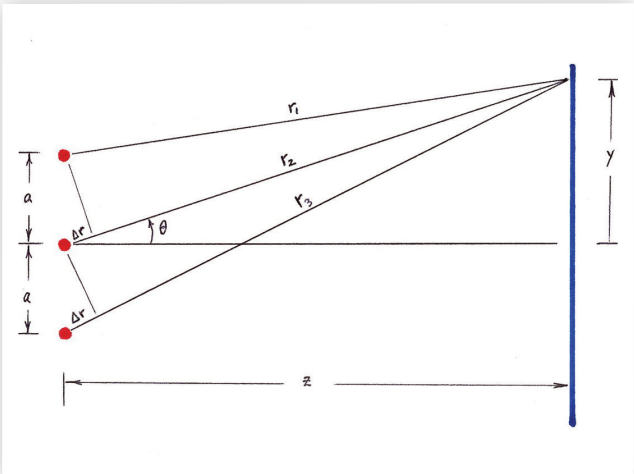


FIG. 1: The geometry of the three-point-source interference experiment. In practice, $z >> a$ and y, so that $\sin\theta << 1$. Note that the lines of length r_1, r_2, and r_3 are then approximately parallel near the slit.

2 travels the distance $r_2 = r_1 + \Delta r$, and the ray from source 3 has length $r_3 = r_1 + 2\Delta r$. Due to the different path lengths, the three signals acquire phase differences, causing interference when the signals are added together at y. Assuming equal amplitudes ψ_o upon arrival,[3] the total wave function $\psi(y,t)$ at y is

$$\psi(y,t) = \psi_o\,[\cos(\alpha) + \cos(\alpha + \delta) + \cos(\alpha + 2\delta)], \qquad (1)$$

where $\alpha = kr_1 - \omega t$ and $\delta = k\Delta r$. Typically, Δr is too small to measure directly with a meterstick, but z is large, so θ is small, and $\Delta r = a \sin\theta \approx a \tan\theta = ay/z$.

Our task is to sum the three cosines in an interpretable way. Phasor diagrams fall readily to hand. Pretend that Eq. (1) is the horizontal component of a vector sum, add the vectors graphically, and then evaluate the horizontal component of the resultant to obtain $\psi(y,t)$, as in Fig. 2.

The condition for the resultant to achieve maximum magnitude requires the three vectors to be collinear, corresponding to places on the screen where $\delta = 2n\pi$, with $n = 0,1,2,3\ldots$. In these places, $\psi = 3\psi_o\cos(\alpha)$, a primary maximum. Since one source by itself produces an intensity I_o proportional to $|\psi_o|^2$, the intensity of a primary maximum will be $9I_o$. Zero intensity occurs when the three phasors close back on themselves to make a triangle. This requires δ to be one-third of a rotation or a whole number of rotations beyond that, in other words, $\delta = 2\pi(n + ⅓)$, where, again, $n = 0,1,2,3,\ldots$. When two of the three phasors cancel out, so that $\delta = (2n+1)\pi$, then $I = I_o$, a secondary maximum.

Imagine a phasor diagram in which the value of δ can be controlled by turning a knob while the intensity is monitored. Steadily increasing δ corresponds to sliding the observation point along the y-axis. When the three phasors are initially lined up with $\delta = 0$, the intensity is $9I_o$. As δ (and y) increases, the intensity first declines, reaching zero at $\delta = 2\pi/3$. The intensity then increases to I_o at $\delta = \pi$, decreases again to zero when $\delta = 4\pi/3$, and increases back to $9I_o$ when $\delta = 2\pi$. This pattern repeats periodically thereafter and is symmetric at about $y = 0$. The intensity distribution, as a function of position y on the screen, is shown schematically in Fig. 3a.

With four equally spaced point sources emitting coherently, the four vectors in the phasor diagram predict principal intensity maxima of $16I_o$ when all four vectors are collinear, or $\delta = 2\pi n$. Zero intensity occurs when all phasors cancel at $\delta = (n+1)\pi/2$. Secondary maxima, when three of the four waves cancel leaving $I = I_o$, occur where $\delta = 2\pi(n + ⅓)$. (Fig. 3b)

One may continue such analyses with $N = 5, 6, 7,\ldots$ coherent, equally spaced point sources. The primary maxima occur for $\delta = 2\pi n$, the same as occurs with two point sources. On the screen, in between adjacent primary maxima of intensity N^2I_o, one finds $N-2$ secondary maxima of intensity I_o and $N-1$ minima of zero intensity.

When the number of sources per centimeter reaches into the hundreds or thousands, the array is called a diffraction grating. Made with fine parallel lines on a transparent sheet (etched into glass, or created by photolithography on plastic), each line scatters the incoming light and serves as a point source. The energy emerging on the far side of the grating becomes concentrated in the well-separated primary maxima intensity peaks, leaving negligible the secondary maxima and making such gratings useful for spectroscopy by sending principal maxima of different wavelengths to widely separated angles.

Suppose that along a line of fixed width w the number of point sources N is allowed to become arbitrarily large. The spacing between adjacent sources must, of course, grow smaller. In the limit as $N \to \infty$, the array becomes a slit of finite width. The geometry is similar to that in Fig. 1. Let the slit be mapped with a y'-axis, with edges at $y' = \pm w/2$. Let the slit and screen be separated by the distance z and locations on the screen be mapped with a y-axis. By Huygens' principle, a segment of width dy' of the harmonic plane wave passing through the slit becomes the source of a new wave $d\psi$. In terms of complex numbers, which like phasors have an amplitude and a phase,[4] when this little wave arrives on the screen at time t it has the form

$$d\psi = dA\, e^{i(kr - \omega t)}, \qquad (2)$$

where k is the wavenumber and ω the angular frequency of the harmonic wave, with

$$r = \sqrt{z^2 + (y - y')^2}. \qquad (3)$$

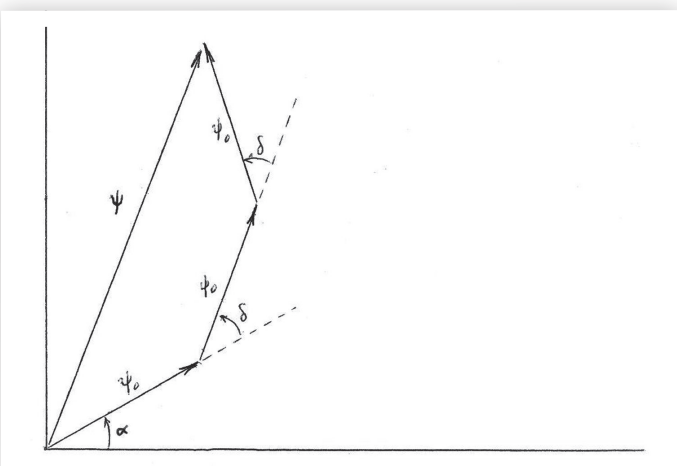


FIG. 2: Phasor diagram construction for three coherent point sources, where Eq. (1) is considered a component of the vector sum $\psi = \psi_1 + \psi_2 + \psi_3$.

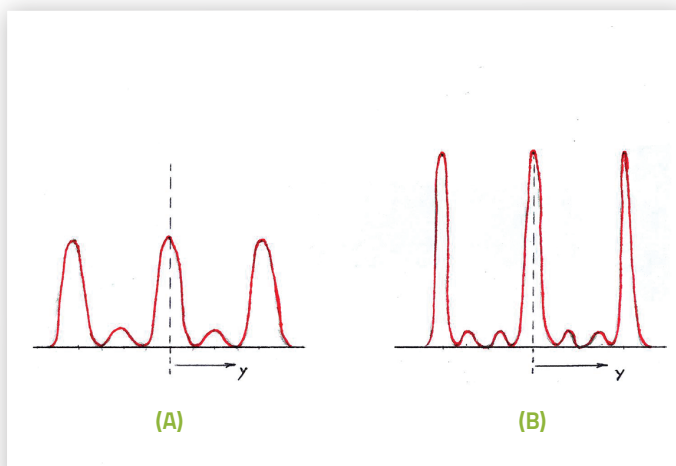


FIG. 3: The intensity pattern of (a) three and (b) four equally spaced coherent point sources.

The amplitude of $d\psi$ can be written as a fraction dy'/w of the amplitude A_o that would have been emitted if the entire signal passing through the slit were instead sent from a point, so that $dA = A_o(dy'/w)$. At a location y on the screen and at a time t, the total wave function $\psi(y,t)$ is the superposition of all the infinitesimal Huygens waves,

$$\psi(y,t) = \frac{A_o}{w}\int_{-w/2}^{w/2} e^{i(kr-\omega t)}dy'. \tag{4}$$

Assuming $z >> y$ and w, to the first order in small parameters, Eq. (3) yields

$$r \approx z - y\,y'/z. \tag{5}$$

Denoting $kz - \omega t = \gamma$, Eq. (4) becomes

$$\begin{aligned}\psi(y,t) &\approx \frac{A_o}{w}e^{i\gamma}\int_{-w/2}^{w/2} e^{-ikyy'/z}dy' \\ &= A_o e^{i\gamma}\frac{\sin(\beta/2)}{\beta/2}\end{aligned} \tag{6}$$

where $\sin(x) = (e^{ix} - e^{-ix})/2i$ has been used and $\beta \equiv kwy/z = 2\pi wy/\lambda z$, with λ the wavelength. The intensity distribution on the screen is

$$I(y) = I_o\left(\frac{\sin\frac{\beta}{2}}{\beta/2}\right)^2 \tag{7}$$

as illustrated in Fig. 4a. Incidentally, the combination $(\sin x)/x$ occurs so frequently in diffraction problems that it has been dignified with the name "sinc x."

The first minimum occurs when $\beta = 2\pi$, where $y = \lambda z/w$, giving $2\lambda z/w$ for the width of the central diffraction peak. The width of the intensity pattern is inversely proportional to the width of the aperture—a hallmark of diffraction.

In terms of the small angle θ (Fig. 1), the first condition for a minimum requires $w \sin\theta = \lambda$, which answers two questions. First, how is the diffraction of visible light, as revealed in such experiments, consistent with the observation that we do not readily notice optical diffraction in everyday life? Second, why do some radio telescopes use metallic meshes for their parabolic reflectors, and how can the doors of microwave ovens include a mesh that allows one to see inside without the microwaves escaping?

In response to the first question, Young's experiment shows that the wavelengths of visible light lie in the range 400–750 nm, thousands of times smaller than apertures encountered in everyday life—indeed, a thousand times smaller than the diameter of a human hair (~0.1 mm)! Thus when $\lambda << w$, then $\sin\theta << 1$, and all the minima from diffraction at a slit's edge crowd together in the forward direction. That makes the separated minima of diffraction hard to see, so the result approaches a sharp shadow.

In reply to the second question, the wavelength of microwaves is larger than the holes in the mesh: $\lambda > w$, so $w \sin\theta = \lambda$ becomes the absurd relation $\sin\theta > 1$. The waves cannot pass through the opening, making it a reflector.

That $w \sin\theta = \lambda$ describes the first minimum in single-slit diffraction may seem counterintuitive at first. This relation for the single slit often follows soon after a discussion of $a \sin\theta = ½\lambda$, the condition for the first minimum in two-point-source interference. Why the factor of ½ that distinguishes these cases when both describe minima? When we muse over it, $w \sin\theta = \lambda$ for a minimum holds because of $a \sin\theta = ½\lambda$. To see this, subdivide a plane wave passing through the slit into, say, 600 Huygens sources. When the signals from sources 1 and 300 cancel out, then so do those from sources 2 and 301, 3 and 302, and so on, all the way through sources 299 and 600. Pairwise cancellation occurs, and, for each pair of sources, $(w/2)\sin\theta = \lambda/2$.

Point sources are an idealization; real sources have finite size. Consider, then, the diffraction produced by two identical slits, each of width w, whose centers are separated by the distance a. The signal arriving at a point $y = z \sin\theta$ on the screen can be predicted by adjusting the integration limits in Eq. (6):

$$\psi(y,t) = \frac{A_o}{w}e^{i\gamma}\left[\int_{(a-w)/2}^{(a+w)/2} e^{iky'\sin\theta}dy' + \int_{(-a-w)/2}^{(-a+w)/2} e^{iky'\sin\theta}dy'\right]. \tag{8}$$

Upon evaluating the integrals and squaring the result to obtain the intensity on the screen, we find (Fig. 4b),

$$I(y) = I_o\cos^2\left(\frac{\delta}{2}\right)\left(\frac{\sin\frac{\chi}{2}}{\frac{\chi}{2}}\right)^2 \tag{9}$$

where $\delta = ka \sin\theta$ and $\chi = kw \sin\theta$. We recognize $\cos^2(\delta/2)$ as the two-

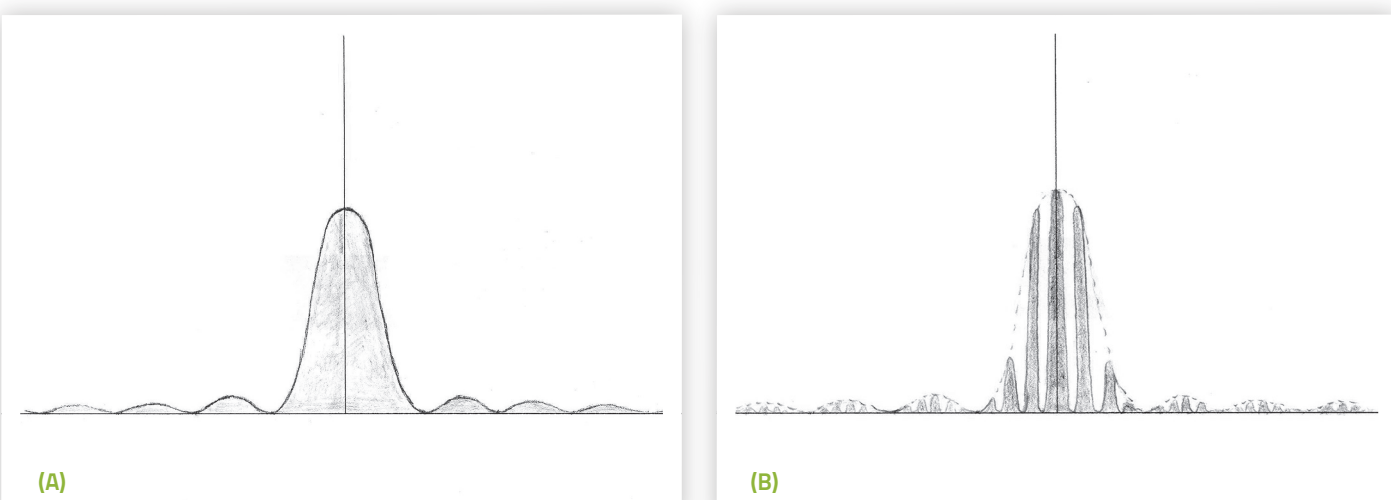


FIG. 4: The Fraunhofer diffraction pattern of an (a) single and (b) double slit.

point-source interference distribution and $\text{sinc}^2(\chi/2)$ as the single-slit diffraction distribution. This result for two identical slits illustrates a larger result, the array theorem: the Fraunhofer intensity pattern produced by N identical apertures equals the diffraction pattern of one aperture multiplied by the interference pattern of N point sources.[5]

What happens when we take away the slit and replace it with an opaque ribbon of the same width? We can conceptualize the situation as follows. Begin with an opaque sheet and cut out an aperture. By itself, the aperture (e.g., a slit) produces on the screen the wave function ψ_{ap}; by itself the cut-out piece (e.g., a ribbon) produces on the screen a wave function ψ_{co}. When the cutout fills the aperture, the total wave function on the screen vanishes, $\psi_{tot} = 0$. By the superposition principle, $\psi_{tot} = \psi_{ap} + \psi_{co}$. Therefore $\psi_{ap} = -\psi_{co}$. But the intensity goes as the wave function squared; hence $I_{ap} = I_{co}$, a marvelous result known as Babinet's principle.[6] An experiment you can easily do, as my students have done many times, is to measure the diameter of a hair from your own head using a laser beam. One industrial application of this principle is the manufacture of fine wire in which the diameter is continuously checked by passing the wire through a laser beam while monitoring the diffraction pattern.

Most readers will have performed in a general physics lab the kinds of diffraction experiments we have been describing. The apparatus used in such experiments typically consists of a small laser with a beam diameter sufficient to illuminate the edges of a slit but not its ends. Now let's shorten the slit or pull the laser back (its beam, too, spreads by diffraction) to illuminate all the edges.

FRAUNHOFER DIFFRACTION PRODUCED BY AN APERTURE

Consider an aperture of area Γ, uniformly illuminated by monochromatic plane waves. Let the plane of the aperture be mapped with an $x'y'$ coordinate system, and let the screen be mapped with xy coordinates (the $x'y'$ and xy planes are held parallel, separated by distance z). According to Huygens' principle, each infinitesimal patch of area $dx'dy'$ on the wave front in the aperture radiates a wave $d\psi$. Upon the arrival of this wave increment at the screen, it has amplitude dA and an acquired phase, so that

$$d\psi(x,y,t) = dA\ e^{i(kr-\omega t)}, \tag{10}$$

where

$$r = \sqrt{(x-x')^2 + (y-y')^2 + z^2}. \tag{11}$$

At location (x,y) on the screen at time t, the total wave function $\psi(x,y,t)$ is the superposition of the infinitesimal waves:

$$\psi(x,y,t) = A_o \int e^{i(kr-\omega t)} \frac{dx'dy'}{\Gamma}, \tag{12}$$

where the limits, still to be put on the integral, will describe the aperture. When r is large compared to other length scales, then to first order in small quantities,

$$r \approx r_o - (xx' + yy')/r_o, \tag{13}$$

where $r_o = [x^2 + y^2 + z^2]^{1/2}$. Denoting $kr_o - \omega t = \gamma$, Eq. (12) becomes

$$\psi(y,t) \approx A_o e^{i\gamma} \int e^{-ik(xx'+yy')/r_o} \frac{dx'dy'}{\Gamma} = A_o\, e^{i\gamma}\ \xi(x,y) \quad , \tag{14}$$

in which we have introduced the modulation factor

$$\xi(x,y) \equiv \int e^{-ik(xx'+yy')/r_o} \frac{dx'dy'}{\Gamma} \tag{15}$$

with integration limits over Γ understood. If all the signals coming through the aperture were instead concentrated at the $x'y'$ origin, then $A_o e^{i\gamma}$ would be the signal arriving on the screen, where it would produce the uniform intensity $I_o \sim A_o^2$. The factor $\xi(x,y)$ describes the effect of the aperture's finite size and shape in redistributing the signal across the screen. For a rectangular aperture, the integration limits on x' go from $-a/2$ to $+a/2$, y' goes from $-b/2$ to $+b/2$, and Eq. (15) becomes

$$\xi(x,y) \equiv \int_{-a/2}^{a/2} \int_{-b/2}^{b/2} e^{-ik(xx'+yy')/r_o} \frac{dx'dy'}{ab} \tag{16}$$

and therefore

$$I(x,y) = I_o \left(\frac{\sin\frac{\alpha}{2}}{\alpha/2}\right)^2 \left(\frac{\sin\frac{\beta}{2}}{\beta/2}\right)^2, \tag{17}$$

where $\alpha = kax/r_o$ and $\beta = kby/r_o$ (Fig. 5a).

When the aperture is a circle of radius a, in the evaluation of ξ it becomes convenient to switch to polar coordinates ($x' = \rho'\cos\varphi'$, $y' = \rho'\sin\varphi'$) and similarly for the screen coordinates:

$$\xi(\rho,\varphi) = \frac{1}{\pi a^2} \int_0^a \rho' d\rho' \int_0^{2\pi} e^{-ik\rho\rho' \cos(\varphi'-\varphi)/r_o}\, d\varphi'. \tag{18}$$

Noting that the system has rotational symmetry about the aperture axis, we may evaluate ξ at $\varphi = 0$. The definite integral over φ' offers an instance from the family of Bessel functions, which are damped, oscillatory, nonperiodic functions that often appear as solutions to differential equations in cylindrical coordinates. The acoustics of brass, percussion, and woodwind musical instruments, as well as the electrostatics of cylindrical charge distributions, are among the many examples in which Bessel functions appear. Most of us probably first encounter Bessel functions of order μ, denoted $J_\mu(\rho)$, as the solutions of the differential equation,[7,8]

$$\frac{d^2F(\rho)}{d\rho^2} + \frac{1}{\rho}\frac{dF(\rho)}{d\rho} + \left(1 - \frac{\mu^2}{\rho^2}\right) F(\rho) = 0, \tag{19}$$

which does not arise explicitly in our discussion here but offers a backdrop to it. Equation (19) emerges as the radial part of the linear homogeneous wave equation in cylindrical coordinates,[9]

$$\nabla^2\psi - \frac{1}{v^2}\frac{\partial^2\psi}{\partial t^2} = 0, \tag{20}$$

which describes locally any wave traveling with speed v and without dispersion, for which $\psi(x,y,z,t) = \psi(r \pm vt)$. An elegant integral representation of the Bessel function of order μ exists:[10]

$$J_\mu(x) = \frac{i^{-\mu}}{2\pi} \int_0^{2\pi} e^{ix\cos\theta} \cos(\mu\theta) d\theta. \tag{21}$$

In terms of a Bessel function, Eq. (18) becomes

$$\xi(\rho) = \frac{2\pi}{\pi a^2} \int_0^a \rho' J_o\left(\frac{k\rho\rho'}{r_o}\right) d\rho'. \tag{22}$$

Now make a change of variable and borrow a result from tabulated integrals of Bessel functions:[11]

$$\int_0^{\eta} x J_o(x) dx = \eta J_1(\eta) \tag{23}$$

to obtain $\xi(\rho) = 2J_1(\kappa)/\kappa$, where $\kappa = k\rho a/r_o$, to find

$$I = 4I_o \left(\frac{J_1(\kappa)}{\kappa}\right)^2. \tag{24}$$

This pattern features a central peak of intensity maximum (called an Airy disk for the central bright spot that appears on the screen), surrounded by a dark circular line at the intensity minimum and circular ripples of intensity maxima and minima outside the first minimum (Fig. 5b). The first minimum occurs when κ is the first root of $J_1(\kappa) = 0$, or $\kappa = 3.823$.[7,8,10] In terms of the angle θ, $ka \sin\theta = 3.823$, which is typically expressed by relating the aperture diameter to wavelength:

$$(2a)\ \sin\theta = 1.22\lambda. \tag{25}$$

When the circular aperture is replaced by a circular disk of the same diameter, Babinet's principle predicts a bright spot directly behind the disk, in the region that would, but for diffraction, be in shadow! This "unexpected" bright spot is called Poisson's spot, named after Siméon D. Poisson, thanks to an incident of intellectual dispute that occurred between him and Augustin Fresnel in 1818—therein lies an amusing story.[12]

If two sources close together are viewed from a large distance away, diffraction sets a limit on how well they can be resolved, that is to say, distinguished from a single source. A simple criterion put forward by Lord Rayleigh says that the two circular apertures can be resolved if the center of the Airy spot from one aperture lands on top of the first minimum from the other aperture.[13] The critical angle at which this occurs satisfies Eq. (25). If $\theta > \sin^{-1}(1.22\lambda/2a)$, then the two apertures can be resolved. Familiar applications include distinguishing two stars (e.g., the double star Mizar in the handle of the Big Dipper), and the two taillights of a distant car, when the pupil of your eye is the aperture. Diffraction places a limit on what we can expect from an optical system consisting of radiation, an aperture, and an image.

FRAUNHOFER DIFFRACTION AS A FOURIER TRANSFORM

We can formally generalize the aperture-image, diffraction-limited relationship in the Fraunhofer far-field zone. The crucial insight comes from an appreciation of Fourier transforms that gives one the choice between describing a signal in terms of its spatial-temporal variation, or in terms of its mixture of harmonic frequencies.

Any signal that is periodic with period T in time, or period λ in space, can be written as a harmonic series of sines and cosines, with each contributing frequency an integral multiple of the fundamental temporal frequency $\omega = 2\pi/T$ or of the fundamental spatial frequency $k = 2\pi/\lambda$. For instance, imagine freezing a wave in space at the instant $t = 0$. One side of Fourier's theorem says that, given the amplitudes A_n and B_n, one synthesizes $\psi(x)$ according to

$$\psi(x) = \frac{A_o}{2} + \sum_{n=1}^{\infty} [A_n cos(nkx) + B_n sin(nkx)]. \tag{26}$$

For this to be meaningful the inverse problem must be solved, and the theorem goes on to show that, given $\psi(x)$, the amplitudes of the harmonics are

$$A_n = \frac{2}{\lambda}\int_0^{\lambda} \psi(x) cos(nkx) dx, \tag{27}$$

and similarly for B_n, with the sine replacing the cosine. A nonperiodic signal, or wave pulse, can be modeled by first giving it a finite period and then taking the limit as the period goes to infinity. Using complex forms of the trig functions through $e^{i\theta} = \cos\theta + i \sin\theta$, and assuming the superposition principle holds, a nonperiodic wave can be written as a superposition of harmonics.[14] However, because such a wave is unbounded, it no longer has a fundamental frequency. The sum over frequencies becomes continuous. For instance, looking again at the signal frozen in space, in this limit and in the language of complex numbers Eq. (26) becomes

$$\psi(x) = \int_{-\infty}^{+\infty} \zeta(k) e^{ikx} \frac{dk}{\sqrt{2\pi}}, \tag{28}$$

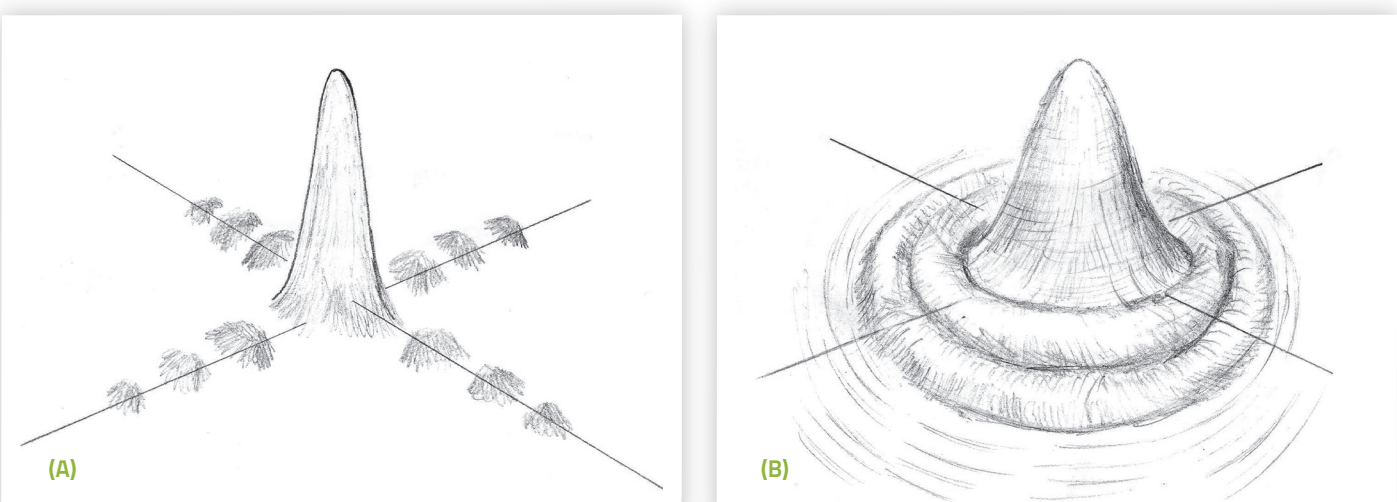


FIG. 5: The diffraction of (a) a square aperture and (b) a circular aperture.

where $\zeta(k)$ is the amplitude of the harmonic of wavenumber k (see below). Conversely, given $\psi(x)$ one can find the spectrum of amplitudes $\zeta(k)$. For the inverse problem, define $\zeta_n = ½(A_n - iB_n)$, then take the limit as $\lambda \to \infty$. Equation (27) and its counterpart for B_n give together

$$\zeta(k) = \int_{-\infty}^{+\infty} \psi(x) e^{-ikx} \frac{dx}{\sqrt{2\pi}}. \tag{29}$$

Go back now to the integral of Eq. (6) and set $t = 0$, so that before being evaluated the integral says

$$\psi(y) = \frac{A_o}{w} e^{ikz} \int_{-w/2}^{w/2} e^{iky y'/z} dy'. \tag{30}$$

Let $\kappa = k/zy'$, allow A_o to become a function ζ of y', and hence a function of κ (so it gets pulled back behind the integral sign), absorb remaining constants into the amplitude $\zeta(\kappa)$, and integrate over the entire κ-axis by allowing $w \to \infty$. Now Eq. (30) becomes

$$\psi(y) = \int_{-\infty}^{+\infty} \zeta(\kappa) e^{iy\kappa} d\kappa, \tag{31}$$

the synthesis side of a Fourier transform. Given the distribution of amplitudes $\zeta(\kappa)$ along the κ (and thus y') axis, one can do the integral and find the wave function on the screen. For instance, if $\zeta(\kappa) = A_o$ for $w/2 \le y' \le w/2$ and $\zeta(\kappa) = 0$ for $|y'| > w/2$, then the result following Eq. (6) is recovered. The smaller the value of w, the wider the spread of the pattern on the screen, and vice versa. That is the diffraction limit.

BEYOND THE DIFFRACTION LIMIT

Diffraction ultimately limits the resolution of a telescope, microscope, camera, or eye. In less-familiar settings, the ability of a computer chip manufacturer to further reduce the size of circuits that are made by mask-and-etch techniques [15] is also diffraction limited. This limit must be appreciated if we are to have a realistic understanding of what is possible, in principle, for imaging.

However, the diffraction limit is not absolute, being susceptible to loopholes offered by quantum mechanics. One example of a quantum loophole is the tunneling effect, in which a particle's probability to pass through a classically forbidden potential barrier is not zero but is exponentially damped. Another example is the Hawking mechanism of black hole evaporation, where, due to the Heisenberg uncertainty principle, vacuum fluctuations just outside a black hole's event horizon can occasionally produce a particle–antiparticle pair and one member of the pair escapes.

One can move the source and image close together into the near-field region, eschewing the Fraunhofer far-field region. In this setting, if the apertures are about the same size as the object being imaged, the latter can be imaged at a resolution below the diffraction limit. For example, if light is sent one photon at a time through a thin extruded optical fiber with a diameter on the nanometer scale, and if those photons hit a molecule and make it fluoresce, then an array of fluorescing molecules offers an image in which the diffraction resolution limit has been bypassed.[16]

Another method for beating the diffraction limit uses stimulated emission, whereby an originally excited molecule is induced to de-excite by the passage of an appropriately tuned incoming photon. One begins this process by passing a laser pulse through a lens to make the light mimic a point source as closely as possible, then shining the light on a sample. This excitation pulse excites target molecules into fluorescence. The pulse is followed by another one whose frequency is tuned to cause stimulated emission, de-exciting those molecules. To circumvent the diffraction limit, the de-excitation pulse is split into two beams with a relative offset $\Delta\kappa$ [the κ of Eq. (24)]. The beams' two Airy disks land outside the Airy disk of the original excitation pulse. The central peak survives the de-excitation, leaving a pattern smaller than it would have been if limited by diffraction.[17]

In the next installment we will consider Fresnel diffraction, wave physics in the near zone that takes into account the curvature of wave fronts. //

ACKNOWLEDGMENT

I am grateful to Thomas Olsen for carefully reviewing a draft of this manuscript, making several suggestions that resulted in its improvement.

REFERENCES AND NOTES

[1] Richard P. Feynman, Robert B. Leighton, and Matthew Sands, *The Feynman Lectures on Physics* (Addison-Wesley, Reading, MA, 1965), Vol. III, Ch. 1.
[2] "Elegant Connections in Physics, Part 1: Huygens' Principle and Young's Interference," *The SPS Observer*, Winter 2012, pp. 20–23.
[3] We neglect the diminishment of amplitude with distance, as discussed in Ref. 2.
[4] Introductory physics textbooks make excellent use of phasor diagrams in the single slit application; see, e.g., David Halliday, Robert Resnick, and Kenneth S. Krane, *Physics*, 4th ed. (extended version) (Wiley, New York, 1992), Vol. 2, pp. 972–974; Paul A. Tipler, Physics (Worth, New York, 1976), pp. 588–592. Richard Feynman made intriguing use of phasors in a set of public lectures on quantum electrodynamics: *QED: The Strange Theory of Light and Matter* (Princeton Univ. Press, Princeton, NJ, 1985).
[5] For example, see Eugene Hecht, *Optics*, 4th ed. (Addison-Wesley, San Francisco, CA, 2002), pp. 543–544.
[6] Babinet's principle does not hold rigorously in all cases. For example, if the aperture is a point source, then there is no screen big enough to make $\psi_{tot} = 0$ everywhere. See Hecht, Ref. 5, pp. 508–509, who writes "The real beauty of Babinet's principle is most evident when applied to Fraunhofer diffraction,…where the patterns from complementary screens are almost identical."
[7] The Bessel functions themselves are constructed by assuming a power series solution of Eq. (19), which is how most students of physics first encounter them. See differential equations texts or, e.g., J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), pp. 102–108.
[8] Frank Bowman, *Introduction to Bessel Functions* (Dover, New York, 1958).
[9] ∇^2 denotes the Laplacian operator. Equation (19) is the ρ-dependent part of the wave equation, in cylindrical coordinates, after ψ has been factored into a product of three independent functions of ρ, φ, and z.
[10] Milton Abramowitz and Irene A. Stegun, Eds., *Handbook of Mathematical Functions* 9th printing (Dover, New York, 1970), Eq. 9.1.21 on p. 360.
[11] Bowman, Ref. 8, p. 9.
[12] Hecht, Ref. 5, p. 494.
[13] E.g., Hecht, Ref. 5, p. 472.
[14] See, for example, "Elegant Connections in Physics: Synthesis & Analysis, Part 4. Fourier Series and Fourier Transforms," *SPS Newsletter* (Feb/March 1998), pp. 9–13.
[15] In mask-and-etch techniques, one photographs the circuit diagram, reduces it, and with it masks the top layer of a semiconducting wafer, then dissolves away the top layer of unmasked material. Then a new layer of raw material is overlayed and the process repeated.
[16] "Breaking the Diffraction Limit of Light, Part 1," in *The Everyday Scientist*, http://blog.everydayscientist.com/?p=184;/.
[17] Stefan W. Hell and Jan Wichmann, "Breaking the diffraction resolution limit by stimulated emission: Stimulated-emission-depleted fluorescence microscopy," *Opt. Lett.* **19** (11), 780–782 (1994).