

Diffraction, Part 1

HUYGENS' PRINCIPLE AND YOUNG'S EXPERIMENT

by Dwight E. Neuenschwander, Southern Nazarene University



WAVES ARE DIFFRACTED around a rocky outcrop. Photo by Wing-Chi Poon.

Like a test pilot pushing an airplane prototype to the limits of its capabilities, physics seeks Nature's fundamental limits and then pushes them as far as possible. For instance, special relativity implies that no information can travel faster than the speed of light. Quantum mechanics holds that the more precisely we know a particle's momentum, the less precisely we can know its location. The second law of thermodynamics asserts that the efficiency of any two-temperature engine cannot exceed that of the idealized Carnot engine, and the third law of thermodynamics implies that a system cannot be cooled to absolute zero in a finite number of steps. In optics, diffraction limits the maximum resolution of a telescope and the minimum size of a computer chip circuit made by mask-and-etch techniques.

Diffraction is the spreading out of a wave into regions that would otherwise be in shadow. A beam of light, even one as tightly concentrated as that emerging from a laser, spreads out as it travels, creating a sizable spot when it reaches the opposite side of a room. Such displays of diffraction offer clear evidence that light, sound, and other signals can behave as if they are waves.

Diffraction can be casually observed in water waves sweeping around a buoy, or in sound waves passing through a doorway. To observe diffraction with visible light takes more care. Ever since antiquity, people have wondered "What is light?" By the time of Robert Hooke (1635–1703), Christiaan Huygens (1629–1695), and Isaac Newton (1642–1727), responses to the question had been reduced to a binary choice: Is light a wave, or is it a beam of particles? Hooke and Huygens argued for waves, Newton for particles. Hooke pointed out that producing a specific color in thin-film interference requires choosing a specific film thickness, analogous to tuning the harmonics of an organ pipe by adjusting its length. Clearly something periodic occurs with light. Newton observed the variable thin-film interference we now call "Newton's rings."^[1] However, he argued that periodicity is not *inherent* in light, but is *put into* light by its interaction with matter.

Noting the blazing speed of light, Huygens argued that, although bulk matter does not travel so fast, a wave might be able to, as illus-

Diffraction is the spreading out of a wave

INTO REGIONS THAT WOULD OTHERWISE BE IN SHADOW

trated with a row of marbles struck sharply at one end. The marbles do not move in bulk, but an impulse propagates swiftly down the row. Huygens developed a general model for wave propagation around a basic assumption we now call Huygens' principle. It asserts that each point on a wave front serves as the source of a subsequent wave (Fig. 1a).

The laws of geometrical optics follow from Huygens' construction. Significantly, Huygens' principle also predicted that waves necessarily diffract (Fig. 1b), as point sources on a wave front near the edge of an aperture send radiation into regions that otherwise would be in shadow.

Settling the argument in favor of waves required a demonstration of diffraction. Francesco Grimaldi (1618–1663) was evidently the first to do such a demonstration when he placed thin objects such as needles in beams of light and observed fine fringes at the edges of the objects' shadows. But the evidence was not universally convincing, especially not to Newton, whose large influence helped to advance the particle model over the wave model.

Consensus began to build in 1801 when Thomas Young (1773–1829) conducted experiments similar to Grimaldi's using a "slip of paper about one-thirtieth of an inch in breadth." Satisfied that diffraction was real, he then set out to measure the wavelength of light by allowing a beam of monochromatic light to pass through two small apertures. Upon emerging, the light from one aperture spread out by diffraction and overlapped

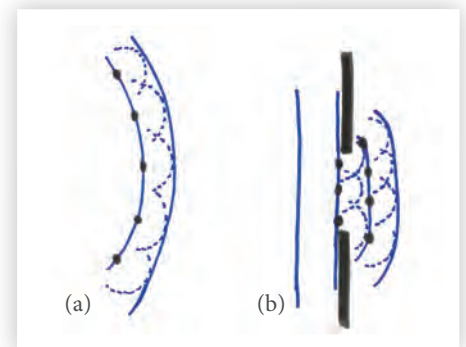


FIG. 1: (a) Huygens' principle. Note that the Huygens construction neglects the backward-propagating wave, which had to be reconciled in mathematical versions of the principle using propagator theory. (b) Huygens' principle predicts diffraction.

with the light diffracting from the other aperture. At a given point on a screen placed opposite the apertures, the total signal was a superposition of the undulating light from both apertures, which produced interference. From the intensity pattern, Young measured the wavelength of visible light. It turned out to be macroscopically tiny (400–750 nm), explaining why the diffraction of visible light is not casually obvious.

As Young’s experiment illustrates, interference occurs because diffraction spreads the two emerging beams of light into overlapping regions. But diffraction itself relies on interference between secondary waves emitted from points on the primary wave fronts, as envisioned by Huygens’ principle. Interference and diffraction are not independent phenomena.

The pattern of water surface waves diffracting around a buoy carries information about the buoy’s size and shape. Diffraction measurements are thus scattering experiments consisting of three components: waves, an obstacle or aperture with which the waves interact, and an image formed on a screen or film or charged-coupled device. If you know any two of these elements you can infer the third. For instance, if you have measurements of the aperture and the image (which is essentially the Fourier transform of the aperture or obstacle), you can deduce attributes of the radiation, such as the wavelength in Young’s experiment. Conversely, diffraction was crucial in deducing the helical structure of DNA, as articulated in 1953 by Francis Crick and James Watson, who were guided by an X-ray photograph of the DNA (the obstacle) taken by Rosalind Franklin 1951. From the X-rays’ wavelength and measurements made from the image, the double-helix structure could be inferred. That hint, and knowing the lengths of adenine-thymine bonds and guanine-cytosine bonds, allowed Crick and Watson to assemble a detailed model of DNA.

FRAUNHOFER AND FRESNEL DIFFRACTION

As a wave travels from sources to screen, how does its energy get redistributed? What we see on the screen is not the wave function itself but its intensity—the time-averaged power carried by the total wave function. Since a wave’s energy is proportional to its amplitude squared, the amplitude will vary with distance. Let us see how this works.

Consider a point source emitting a wave. If no waveguides, obstacles, or apertures interact with the wave, then its rays (normal to the wave fronts) propagate radially outward in all directions from the source point. By definition, the luminosity L of that source is its time-averaged radiated power. The energy carried by the wave, emitted by the point source during an infinitesimal time interval dt , spreads out over the area of a spherical surface with a radius that grows as the wave moves outward. The intensity I of the signal is the local time-averaged power per unit area received by a detector.[2] For an observer located a distance r from an unobstructed point source, the relation between received intensity and emitted luminosity is

$$I = \frac{L}{4\pi r^2} \quad (1)$$

For light understood as a wave in an electromagnetic field, the intensity must also be related to the electric and magnetic field amplitudes. As shown in electrodynamics, the observer of an electromagnetic field receives an intensity numerically equal to the time average of the magnitude of Poynting’s vector $\mathbf{S} = \mathbf{E} \times \mathbf{B} / \mu_0$, where \mathbf{E} and \mathbf{B} denote the electric and magnetic fields, respectively, with μ_0 being the permeability of vacuum. Since waves can be written as a superposition of harmonics, one should consider a monochromatic sinusoidal wave, where \mathbf{E} and

\mathbf{B} have amplitudes E_0 and B_0 . The harmonic variation in space and time enters through the factor $\cos(kr - \omega t)$, in which k is the wavenumber, ω the angular frequency, r the distance from the source, and t the time. Noting that \mathbf{E} and \mathbf{B} are in phase and mutually perpendicular, time-averaging $|\mathbf{S}|$ over one period yields

$$\begin{aligned} \langle |\mathbf{S}| \rangle &= \frac{1}{\mu_0} \int_0^T E_0 B_0 \cos^2(kr - \omega t) \frac{dt}{T} \\ &= \frac{1}{2} E_0 B_0 / \mu_0 \end{aligned} \quad (2)$$

The period is $T = 2\pi/\omega$, and the square brackets denote a time average. Comparing this result to Eq. (1), we see that the amplitude for spherically propagating radiation varies spatially as $1/r$.

With Euler’s formula

$$e^{i\theta} = \cos\theta + i\sin\theta \quad (3)$$

where $i^2 = -1$, complex numbers can alternatively be employed to describe a harmonic wave function. $\psi_0 \cos(kr - \omega t)$ gets replaced with $\psi_0 e^{i(kr - \omega t)}$, with the understanding that the real part of the complex number describes the physical signal. In the language of complex variables, the result of Eq. (2) will be retained if we write

$$I = \frac{1}{2} |\mathbf{E}^* \times \mathbf{B}| / \mu_0 = \frac{1}{2} E_0 B_0 / \mu_0 \quad (4)$$

with $*$ denoting the complex conjugate.

When the distance r in Eq. (1) is sufficiently large (the “far field”), the wave front passing over the observer is locally indistinguishable from a plane. Diffraction produced by plane waves is called Fraunhofer diffraction, named after spectroscopist Joseph Fraunhofer (1787–1826). When the spherical curvature of the wave front is taken into account (the “near field”), the ensuing diffraction is called Fresnel diffraction after Augustin Fresnel (1788–1827), who, independent of Young, extensively developed the wave model of light, beginning around 1815.

In the remainder of the present article we will examine Young’s experiment. This exercise offers a prototype in terms of plane waves for procedures and ways of thinking that will be employed in this series on diffraction. In the next article, we will stay for awhile within the Fraunhofer paradigm, extend Young’s experiment to multiple point sources, and let the sources blend continuously into a single slit aperture. We will derive the diffraction pattern produced by a single slit and relate it to the pattern produced by an opaque ribbon. Since Young’s apparatus involved a pair of narrow slits and not point sources, we will study the diffraction pattern produced by two identical slits. The result will illustrate the elegant array theorem, which holds that the image produced by an array of N identical apertures equals the diffraction pattern of one aperture multiplied by the interference pattern of N point sources.

Then we will consider the diffraction produced by rectangular and circular apertures. The latter is important because human beings view the world through circular apertures—the pupils of our eyes—not to mention the circular apertures in cameras, telescopes, and microscopes. Before leaving Fraunhofer diffraction we will explain what it means to say that the image on the screen is the Fourier transform of the aperture.

Moving on to the complications that arise when taking into account the spherical curvature of wave fronts, we will study Fresnel diffraction. Here we will meet a classy tool for reducing complicated definite integrals to distances on a chart—the Cornu spiral.[3]

Diffraction is an old subject, now part of the furniture of physics. Still, it would be difficult to find many contemporary research topics or high-tech applications that do not depend on it in some way. We should not take for granted its elegant and useful analytic tools. Today we rely on those tools, which were mostly developed during the nineteenth century.

YOUNG'S EXPERIMENT

Between 1801 and 1803 Thomas Young, a British medical doctor, Egyptologist, and physicist (although the term “physicist” did not exist in 1800), performed a double-slit experiment repeated in introductory physics laboratories to this day. Young’s experiment cleanly demonstrated the wave nature of light. That was how light was subsequently understood until Einstein’s light quanta paper in 1905 presented the utility of a complementary particle model. From the wave-particle duality of quantum mechanics that matured in the 1920s, we now know that Nature does not neatly partition all of its structures and interactions into the binary categories of particle or wave. Instead of asserting that light is a wave, we now say that in some situations light behaves as if it were a wave, but in other situations it behaves as if it were a beam of particles. Particles and waves are analogies for describing light (and electrons—nature is symmetric in these matters). In discussing diffraction one operates within the wave paradigm.

Turning to the details of Young’s experiment, we can conceptualize the experiment as the diffraction produced by two point sources emitting coherently, in phase, and with equal amplitude. The experimental setup is shown schematically in Fig. 2.

The two point sources are separated by distance a , with the screen located distance D away. Field points on the screen are mapped by the coordinate s , with $s = 0$ describing the point on the screen opposite the midpoint between the two sources (the sources and s axis are coplanar). To emit coherently means the phase difference between the sources upon emission remains constant, which is realized in practice by having a single beam of monochromatic light illuminate both apertures. To emit in phase means the phase difference at emission equals zero.

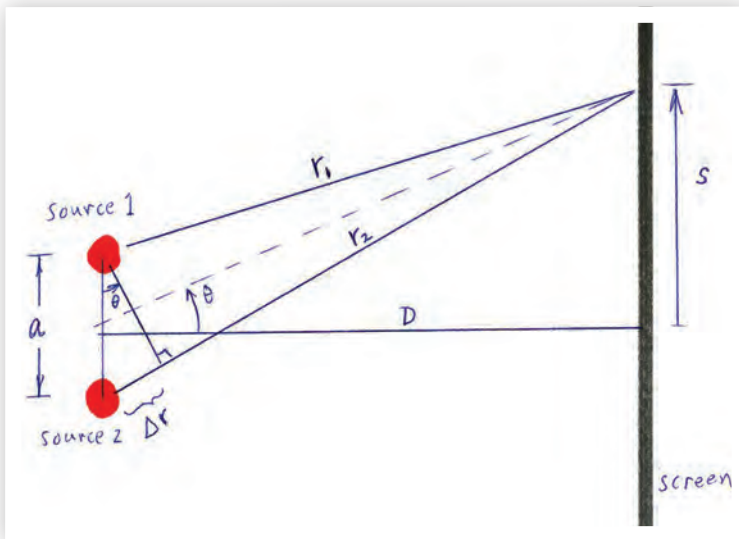


FIG. 2: Schematic of Young’s experiment. In practice, $D \gg a$ and s , so that $\sin\theta \ll 1$. Note that the lines of length r_1 and r_2 are then approximately parallel.

After leaving the apertures, the signals from both apertures spread out and add together to make a pattern on the screen. At time t some of the net signal arrives at a location on the screen having the coordinate s . The part of the total signal arriving there that came from source one traveled along a ray of length r_1 , and the signal that came from source two traveled the distance $r_2 = r_1 + \Delta r$. Upon their arrival at s , the separate signals have acquired a phase difference, causing interference when they add together. Assuming equal amplitudes upon emission, by superposition the total wave function $\psi(s, t)$ ($= |\mathbf{E}|$ or $|\mathbf{B}|$) arriving on the screen at location s at time t is

$$\psi(s, t) = \psi_0(r_1)\cos(kr_1 - \omega t) + \psi_0(r_2)\cos(kr_2 - \omega t) \quad (5)$$

with $\psi_0(r_1)$ and $\psi_0(r_2)$ denoting the amplitudes of the waves where they arrive at the screen. The amplitudes vary as $1/r$, but $(1/r_2) - (1/r_1) \sim \Delta r/r_1^2$, which is negligible. Abbreviating $kr_1 - \omega t \equiv \alpha$, Eq. (5) may be written

$$\psi(s, t) = \psi_0 [\cos(\alpha) + \cos(\alpha + \delta)] \quad (6)$$

which introduces a phase shift δ . For the Young experiment δ is acquired from the path difference,

$$\delta = k\Delta r. \quad (7)$$

More generally, δ can also be caused by various other mechanisms, such as half-cycle phase shifts due to reflections or a time delay from the waves passing through different refractive media.

Our mathematical problem is to write the sum of the two cosines in a way that can be interpreted. We may choose between three equivalent methods: trigonometry identities, complex numbers, or phasor diagrams. Let us pursue all three to illustrate the diversity of techniques that will prove useful in various diffraction problems.

With trig identities, one uses the clever trick of adding and subtracting $\frac{1}{2}\delta$ in the first term,

$$\cos(\alpha) + \cos(\alpha + \delta) = \cos[\alpha + \frac{1}{2}\delta - \frac{1}{2}\delta] + \cos[\alpha + \frac{1}{2}\delta + \frac{1}{2}\delta] \quad (8)$$

then abbreviates $\alpha + \frac{1}{2}\delta \equiv \varphi$. By the trig identities for the cosine of a sum or a difference, this becomes

$$\psi(y, t) = 2\psi_0 \cos(\frac{1}{2}\delta) \cos(\varphi + \frac{1}{2}\delta). \quad (9)$$

Using complex numbers Eq. (6) gets replaced with

$$\begin{aligned} \psi_0 [e^{i\alpha} + e^{i(\alpha+\delta)}] &= \psi_0 e^{i(\alpha+\delta/2)} [e^{-i\delta/2} + e^{i\delta/2}] \\ &= 2\psi_0 e^{i(\alpha+\delta/2)} \cos(\frac{1}{2}\delta). \end{aligned}$$

The real part corresponds to the physical signal:

$$\psi(s, t) = 2\psi_0 \cos(\frac{1}{2}\delta) \cos(\varphi + \frac{1}{2}\delta). \quad (10)$$

For the phasor diagram approach, pretend the cosine terms in Eq. (6) are x components of vectors. Vectors that represent such signals are called phasors.[3] One adds the phasors via usual vector addition and finds the x component of the resultant to get the total wave function (see Fig. 3).

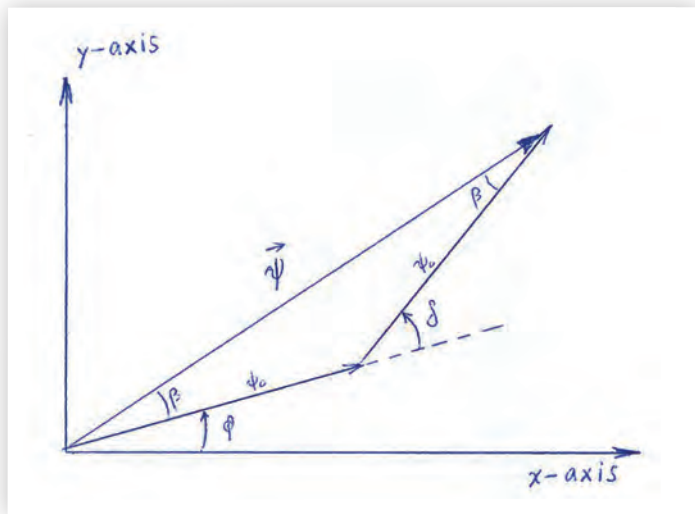


FIG. 3: Phasor diagram construction. Equation (6) is considered the x component of this vector sum.

Apply the law of cosines to $\psi = \psi_1 + \psi_2$:

$$\begin{aligned} |\psi|^2 &= \psi_0^2 + \psi_0^2 - 2\psi_0^2 \cos(\pi - \delta) \\ &= 2\psi_0^2 (1 + \cos \delta). \end{aligned} \quad (11)$$

The trig identity $1 + \cos \delta = 2\cos^2(\frac{1}{2}\delta)$ yields $|\psi|^2 = 4\psi_0^2 \cos^2(\frac{1}{2}\delta)$. The x component of the resultant, the physical total wave function, becomes

$$|\psi_x| = 2\psi_0 \cos(\frac{1}{2}\delta) \cos(\beta + \phi). \quad (12)$$

The triangle formed by ψ_1 , ψ_2 , and ψ shows that $2\beta + (\pi - \delta) = \pi$, and thus

$$\psi(s, t) = 2\psi_0 \cos(\frac{1}{2}\delta) \cos(\phi + \frac{1}{2}\delta). \quad (13)$$

All three approaches give the same answer, as they must, since they are different ways of doing the same thing. However you compute it, the total wave function arriving at s oscillates harmonically in time, weighted by the effective amplitude:

$$\psi_{\text{eff}} = 2\psi_0 \cos(\frac{1}{2}\delta). \quad (14)$$

By virtue of Eq. (2), the intensity is (see Fig. 4)

$$I(s) = 4I(0) \cos^2(\frac{1}{2}\delta). \quad (15)$$

Intensity maxima, $I(s) = 4I(0)$, occur where $\cos(\frac{1}{2}\delta) = \pm 1$ so that $\frac{1}{2}\delta = n\pi$, with $n = 0, 1, 2, \dots$. Intensity minima, $I(s) = 0$, occur where $\cos(\frac{1}{2}\delta) = 0$ and thus $\frac{1}{2}\delta = (2n' + 1)\pi/2$ with $n' = 0, 1, 2, 3, \dots$. For δ due to a path difference, recalling that $k = 2\pi/\lambda$ with λ the wavelength, those results give $\Delta r = n\lambda$ for maxima and $\Delta r = (n' + \frac{1}{2})\lambda$ for minima. In practice Δr is too small to measure directly with a meter stick. But since $\theta \ll 1$, Δr may be written, using the similar triangles in Fig. 2, in terms of easily measurable parameters:

$$\Delta r = a \sin \theta \approx a \tan \theta = a s/D. \quad (16)$$

With this method, Young's experiment facilitates measuring the wavelength of light.

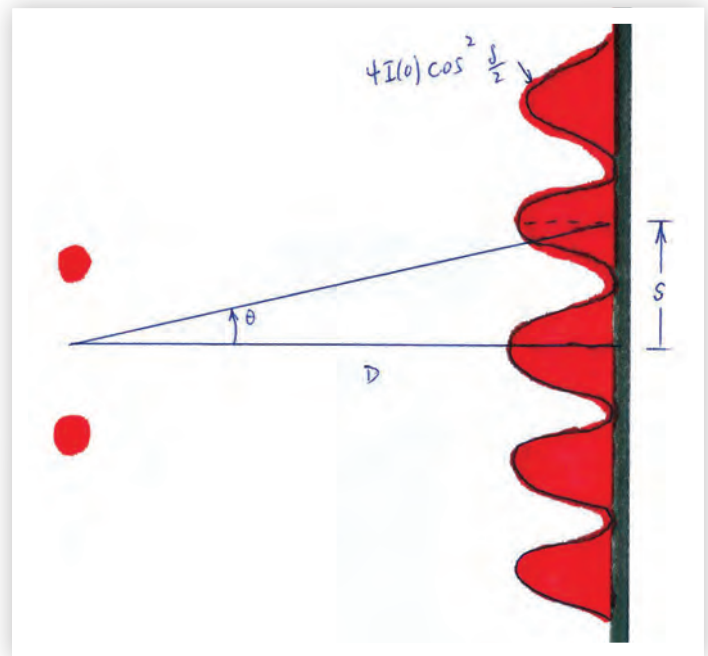


FIG. 4: The intensity pattern in Young's experiment.

The radiation has so far been assumed to be coherent, which means that δ remains constant in time. If the sources emit radiation incoherently (i.e., random emission at the sources), then δ varies randomly with time, so $\cos^2(\frac{1}{2}\delta)$ varies randomly between 0 and 1 and averages to $\frac{1}{2}$. Now Eq. (15) reduces to the uniform illumination $I(s) = 2I(0)$. Two identical candles emitting radiation incoherently merely double the intensity of one, and no interference patterns appear.

In the next installment we will recall that the sources in Young's experiment are not really point sources but have finite sizes. Thus to examine a double slit we must first study in detail diffraction from a single slit. Guided by Huygens' principle, our first intermediate step will be to generalize the interference from two point sources to that of multiple point sources. The Huygens wave front coming through the slit can then be considered the limit of an infinite number of infinitesimally small sources. //

ACKNOWLEDGMENT

I am grateful to Thomas Olsen for carefully reading a draft of this manuscript and offering many useful suggestions that improved it.

BIBLIOGRAPHY & NOTES

Eugene Hecht, *Optics*, 4th ed. (Addison-Wesley, San Francisco, CA, 2002).
William Wightman, *The Growth of Scientific Ideas* (Yale University Press, New Haven, CT, 1953).

- [1] Place a spherical lens on flat glass and shine monochromatic light on it from above. The rings of maxima and minima you see in the curved wedge of air between the lens and the glass sheet are Newton's rings.
- [2] Some authors call I the irradiance because it is the average power per area being absorbed, the opposite of radiance. This term can, however, be confusing. "Irrational" implies "not rational," possibly leading one to ask how irradiance could measure the brightness of a light source.
- [3] The Cornu spiral is an elaborate phasor diagram.