Diffraction, Part 3

CURVED WAVEFRONTS AND FRESNEL DIFFRACTION

by Dwight E. Neuenschwander, Southern Nazarene University

When Thomas Young presented his wave theory of diffraction before Great Britain's Royal Society on January 16, 1800, it was coolly received. One historian of physics writes[1]

> *"...the wave theory might have suffered sterility and oblivion had not sounder critics revived it in France. In 1815 Augustin Fresnel, a brilliant young military engineer and mathematician, submitted to the Academy a paper on diffraction, which, as was the custom of that learned body, was reported upon by two members—François Arago and Louis Poinsot. The former took up the matter with great enthusiasm and drew Fresnel's attention to the almost identical views of Young published fifteen years previously. Although this was the year of the battle of Waterloo, Fresnel paid a generous tribute to Young and they corresponded frequently until the year 1827 when the death of the former put an end to a career full of great promise."*

Parts 1 and 2 of this series[2,3] considered diffraction produced by plane wave fronts—Fraunhofer diffraction. Its signature phenomena include Young's famous double-slit experiment, which demonstrated the wave nature of light. But of course there is more to the story. An unobstructed wave front radiating from a point source forms an expanding spherical surface. Fresnel diffraction takes into account this spherical shape, and I would be remiss to not discuss it, and show how to derive it, in this summary of the basic elements of diffraction theory.

As in the plane wave paradigm, Huygens' principle forms the working tool for understanding Fresnel diffraction. The principle holds that each infinitesimal patch of surface on a primary wave front may be considered the source of a secondary wave. The signal subsequently detected at a point P beyond the surface is the superposition of those secondary waves that reach P.

Consider a point source S that radiates a monochromatic wave of angular frequency ω, so that the signal leaving S is proportional to $\cos(\omega t)$. It is sufficient to consider a monochromatic harmonic wave because, so long as the wave equation is linear, any wave can

AS IN THE PLANE WAVE PARADIGM, HUYGENS' PRINCIPLE FORMS **the working tool in Fresnel diffraction**

be written as a superposition of harmonics. Picture the spherical wave front when it has expanded to radius ρ, and formed a spherical surface σ. Because of energy conservation, the amplitude of a spherical wave drops off as the inverse of the distance from the source. The spherical wave front at σ carries time-dependence cos [ω(t-ρ/c)] with amplitude ~ 1/ρ, as described by the wave function

$$\psi(\rho,t) = \frac{\mathcal{E}_o}{\rho}\cos(k\rho - \omega t) \tag{1}$$

where $k = 2\pi/\lambda$ denotes the wavenumber, and λ the wavelength. The coefficient $\mathcal{E}_o$, determined by the source's luminosity L, will be considered known; in particular, for a light wave, $\mathcal{E}_o = \sqrt{(L/2\pi\epsilon_o c)}$.[4]

On another spherical surface σ' centered on S but having radius $\rho + r_o$ (see Fig. 1), i.e. a surface that includes point P, Eq. (1) says that for the wave front passing over P,

$$\psi(\rho + r_o,t) = \frac{\mathcal{E}_o}{\rho+r_o}\cos\,[k(\rho + r_o) - \omega t]. \tag{2}$$

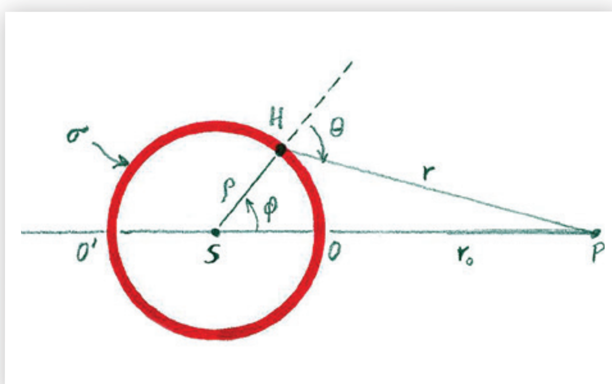

FIG. 1: The geometry of the Fresnel diffraction analysis. Please note the distances $\rho = \overline{SH}$, $r = \overline{HP}$, $r_o = \overline{OP}$, the angles φ and θ, and the annular ring of area $da = (2\pi\rho\sin\varphi)(\rho\, d\varphi)$.

Now return to Huygens' principle. Each little patch of area on the spherical surface σ, such as the one at a typical point H as shown in Fig. 1, serves as a source of a secondary wave. The signal increment $d\psi$ that leaves H at time zero arrives at P at time $t > 0$ with an amplitude modified from that of Eq. (1), is

$$d\psi(H\to P) = \frac{\mathcal{E}_H\, K(\theta)\, da}{r} \cos\left[k(\rho + r) - \omega t\right] \quad (3)$$

where $\mathcal{E}_H$ depends on $\mathcal{E}_o$; relating them forms one of our tasks. The factor $K(\theta)$ (see θ in Fig. 1), called the "obliquity factor," was originally included *ad hoc* by Fresnel because of his intuition that the amplitude of the signal that makes it from H to P depends on θ. To see the need for this, look at the extremes on the surface σ. The da at O will assuredly send signal towards P, but signal radiated from the da located at O' never arrives at P. Evidently, $K(0) = 1$ and $K(\pi) = 0$. A function that exhibits this behavior is

$$K(\theta) = \tfrac{1}{2}(1 + \cos\theta). \quad (4)$$

But until demonstrated rigorously, this expression merely offers a plausible candidate for describing Fresnel diffraction. We may assume, however, that $K(\theta)$ will change very slowly especially for small θ because $\cos\theta \approx 1 - \frac{1}{2}\theta^2$. Although Fresnel's intuition was sound, an obliquity factor derived with rigor had to await some 60 years, until Gustav Kirchhoff published in 1883 a diffraction theory based on the differential equations of waves.

The conceptual distinction between Eqs. (2) and (3) raises basic questions: Why bother with the Huygens surface σ in the first place? Why not merely go with Eq. (2) and be done with it? After all, integrating Eq. (3) over σ must reproduce Eq. (2).

That's a good question. If spherical waves always traveled unobstructed from S to P then Huygens' principle would indeed be a redundant complication. But neither would there be much interest in diffraction, where the interesting results arise in the interactions of waves and apertures. The superposition of Huygens secondary waves *that get through an aperture* makes possible the calculation of diffraction patterns. The aperture problem is what we intend to solve, taking into account the curvature of the wave front. Thus the spherical wave front of interest will be the one with a radius ρ that places its surface at the aperture.

FRESNEL ZONES

At first glance, our task may appear rather simple. To predict the diffraction pattern produced by an aperture, we could merely integrate Eq. (3) over limits defined by the aperture boundary, it would seem. However, the finite speed of the wave (e.g., c for light) offers a subtle complication. Let's approach it this way: If the source S emits a flash of light of infinitesimal duration, in the reference frame of S that flash arrives at all points on σ at a time we may set to zero. What does the observer at P (assumed to be at rest with respect to S) detect? Nothing at all for $0 < t < r_o/c$. Then at the time r_o/c the first light arrives at P, the light from point O. After the signal from O sweeps by, subsequent signals arrive later at P from other locations on σ. A secondary signal leaving a point on σ at the distance r from P (see Fig. 1) arrives at $t = r/c$. The last light to arrive at P from σ occurs as $t \to (r_o + 2\rho)/c$, when some fraction of the signal from the "back side" of the sphere makes it to P (signal from O′ at θ = π never reaches P). How do we account for this time delay when S emits a sinusoidal wave continuously? Here the *Fresnel zones* fall readily to hand.

The Huygens analysis of spherical wave fronts with strengths that oscillate harmonically proceeds by partitioning σ into a set of Fresnel zones centered on the "north pole" at O. By definition, the vibrations from the inner and outer boundaries of a zone are half a cycle out of phase when they arrive at P. To map them in terms of the r of Fig. 1 and the wavelength λ, let $r_1 = r_o + \lambda/2$, $r_2 = r_o + \lambda$, $r_3 = r_o + 3\lambda/2$, etc. All points on σ for which r lies between r_o and r_1 are within the first Fresnel zone, those for which $r_1 \le r \le r_2$ lie in the second Fresnel zone, and so on.

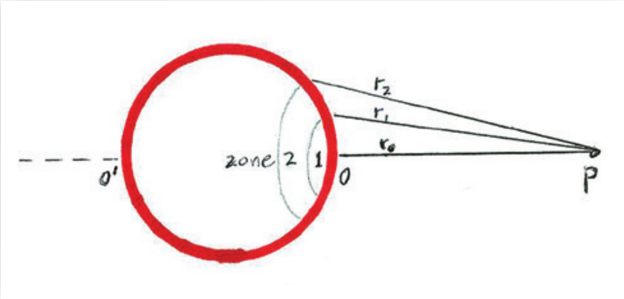

FIG. 2: Fresnel zones.

Consider the situation when S continuously emits a sinusoidal wave. Suppose that at time $t = r_o/c$ a wave crest from O sweeps over P. At the same instant, a wave trough from r_1 will be passing through P, and another wave crest from r_2 will be at P, and so on. At any time, the signals arriving at P from boundaries of adjacent Fresnel zones will be half a cycle out of phase. Their amplitudes at P diminish with larger r_n, thanks to the varying r and the obliquity factor.

Let us prepare to integrate Eq. (3). For an infinitesimal area da consider an annular ring on Fig. 1 for which $da = (2\pi\rho\sin\varphi)(\rho\, d\varphi)$. For fixed ρ, da may be written terms of r with the help of the law of cosines applied to triangle SHP,

$$r^2 = \rho^2 + (\rho + r_o)^2 - 2\rho(\rho + r_o)\cos\varphi, \text{ for} \quad (5)$$

from which

$$da = \frac{2\pi\rho r\, dr}{\rho + r_o}. \quad (6)$$

Now consider an aperture of area Γ, located at O and oriented with the aperture plane perpendicular to the SP axis. The number of Fresnel zones N (assumed for simplicity to be an integer) that pass through the aperture follows from

$$A_1 + A_2 + \cdots A_N = \Gamma \quad (7)$$

where A_n denotes the area of the nth Fresnel zone. Its value follows by integrating Eq. (6) from r_{n-1} to r_n, so that

$$A_n = \frac{\pi\rho\lambda r_o}{\rho + r_o}\left[1 + \frac{\lambda}{4r_o}(2n + 1)\right]. \quad (8)$$

Whenever $\lambda \ll r_o$ (typical for visible light and macroscopic apertures), then A_n becomes approximately independent of n:

$$A_n \approx \frac{\pi\rho\lambda r_o}{\rho + r_o}. \quad (9)$$

For instance, with a circular aperture of radius a, Eq. (7) gives

$$N = \frac{\pi a^2}{A_1} = \frac{a^2}{\lambda}\left(\frac{1}{\rho} + \frac{1}{r_o}\right). \tag{10}$$

Numerically, if $\rho = r_o = 1$ m, $\lambda = 600$ nm, and $a = 1$ mm then $N \approx 3$; but if $a = 1$ cm then $N = 333$. Seen another way, if $\rho = r_o$ then the distance r_o that allows only the first Fresnel zone to pass is $r_o = \frac{2a^2}{\lambda}$, which, for $a \sim 1$ mm and $\lambda \sim 500$ nm gives $r_o \sim 4$ m! At larger values for r_o and ρ, only a fraction of the first Fresnel zone, approximately flat, passes through the aperture. The situation reduces to the Fraunhofer diffraction of plane waves for which $\rho \to \infty$.

Before continuing discussion with this or that aperture, let us investigate the contribution at P of just one Fresnel zone. The signal ψ_n arriving at P from the nth Fresnel zone follows by integrating Eq. (3) over that zone only:

$$\psi_n = \frac{2\pi\rho\mathcal{E}_H}{\rho + r_o}\int_{r_{n-1}}^{r_n} K_n(\theta)\cos[k(\rho + r) - \omega t]\, dr \tag{11}$$

where in general θ is a function of r. But $K(\theta)$ varies slowly with θ, so it may be safe to assume that, *within* any single Fresnel zone, its obliquity factor K_n is approximately constant. Then Eq. (11) can be integrated at once:

$$\psi_n = \frac{2\pi\rho\mathcal{E}_H K_n}{k(\rho + r_o)}\left[\sin(\alpha + n\pi) - \sin[\alpha + (n-1)\pi)\right]$$

$$= K_n\,(-1)^{n+1}\,\psi_1 \tag{12}$$

where $k = 2\pi/\lambda$, $\alpha = k(\rho + r_o) - \omega t$, and

$$\psi_1 = \frac{2\rho\lambda\mathcal{E}_H}{\rho + r_o}\sin[k(\rho + r_o) - \omega t]. \tag{13}$$

As we anticipated, the signal $\psi_{(N)}$ that passes through P at time t, coming from Fresnel zones 1 through N, forms an alternating series:

$$\psi_{(N)} = \psi_1 \sum_{n=1}^{N} (-1)^{n+1} K_n(\theta)$$

$$= \psi_1\,[K_1 - K_2 + K_3 - \ldots + (-1)^{N+1}K_N]$$

$$\equiv \psi_1 Z_N. \tag{14}$$

Because K_n varies slowly with θ, we may assume that $K_{n+1} \approx K_n$. In that case, should N be an even number, then $\psi_{(N)} \approx 0$. But for odd N one obliquity factor survives. The K_n are not all *exactly* equal, because $K(\theta)$ *does* diminish with increasing θ. This ambiguity presents us with a technical problem. Consider, for example, the case $N = 5$, for which

$$Z_5 = K_1 - K_2 + K_3 - K_4 + K_5\,. \tag{15}$$

To make use of the near-cancellation of adjacent obliquity factors, should Z_5 be grouped as

$$Z_5 = (K_1 - K_2) + (K_3 - K_4) + K_5 \approx K_5, \tag{16}$$

or as

$$Z_5 = K_1 - (K_2 - K_3) - (K_4 - K_5) \approx K_1, \tag{17}$$

or in some other way? In the absence of an explicit obliquity factor the sum is ambiguous. Pending a justification of a result like Eq. (4), we have to make use of the decreasing nature of $K(\theta)$ and consider averages under various alternative grouping scenarios. Requiring consistency offers as an estimate the median value between the first and last obliquity factors[5]

$$Z_N \approx \frac{1}{2}(K_1 + K_N). \tag{18}$$

The intensity (or "irradiance") at P is proportional to $|\psi|^2$. Slide the detector (move point P) along the SP axis, and a series of intensity maxima and minima appear as $\psi_{(N)}$ passes through successively odd and even values of N.

Since adjacent Fresnel zones tend to cancel one another, Fresnel zone plates can be constructed, typically by photographic reduction. These are masks consisting of concentric opaque rings that block out either the even Fresnel zones or the odd ones. For example, should all the even zones be blocked and an aperture allow no more than 10 zones to pass, then the total signal arriving at P would be that of the odd-numbered zones 1 through 9. Assuming all their obliquity factors to be near unity,

$$\psi \approx \psi_1 + \psi_3 + \psi_5 + \psi_7 + \psi_9. \tag{19}$$

Since each $\psi_n \approx \psi_1$ for small n, in this example $\psi \approx 5\psi_1$, giving an intensity 25 times that of the first Fresnel zone by itself![6]

A special case of an aperture would be no aperture at all! Then N becomes very large. The number of Fresnel zones across the entire sphere is $N_{\text{all}} = 4\rho/\lambda$, which follows by counting the Fresnel zones between O ($r = r_o$) and O' ($r = r_o + 2\rho$). For instance, $N \sim 10^5$ for a sphere of 1 cm radius with $\lambda = 400$ nm. At least for waves in the optical portion of the spectrum, including all the Fresnel zones across the entire sphere suggests taking the limit as $N \to \infty$. Because $K_\infty \approx K(\pi) = 0$, it follows from Eq. (18) that $Z_\infty \approx \frac{1}{2}K_1$.

So far this result has been shown to hold only for odd N. The case of even N must also be worked out. For large N one finds the same result[5], which might be expected because the difference between one more or one fewer Fresnel zone makes no difference as $N \to \infty$. Therefore all Fresnel zones contributing to the signal at P (no aperture) yields the elegant result

$$\psi_{all} \approx \frac{1}{2}\psi_1 \tag{20}$$

where $K_1 \approx 1$ has been used. About one-quarter of the intensity of the unobstructed wave comes from the first Fresnel zone.

It is interesting to note from Eq. (14), that by separating the first Fresnel zone's contribution from all others,

$$\psi_{(N)} = \psi_1 + \psi_{(N>1)} \tag{21}$$

then for large N and from Eq. (20) it follows that $\psi_{(N>1)} \approx -\frac{1}{2}\psi_1$. Blocking out the first zone of an otherwise unobstructed wave produces a wave function inverted relative to ψ_1 and carrying half its amplitude. This is the Fresnel diffraction version of the famous Poisson spot, mentioned earlier[3] in the context of its Fraunhofer analog that applied Babinet's principle.

Now we can relate the $\mathcal{E}_H$ of Eq. (3) to the $\mathcal{E}_o$ of Eq. (2). Using Eqs. (13) and (20) and $K_1 \approx 1$, the unobstructed wave function arriving at P, due to all the Huygens secondary sources on σ, is

$$\psi_{all} = \frac{\mathcal{E}_H \rho\lambda}{\rho + r_o} \sin[k(\rho + r_o) - \omega t]. \qquad (22)$$

Comparing this amplitude to that of Eq. (2) requires

$$\mathcal{E}_H = \frac{\mathcal{E}_o}{\rho\lambda} \qquad (23)$$

so that, for the unobstructed wave,

$$\psi_{all} \approx \frac{\mathcal{E}_o}{\rho + r_o} \sin[k(\rho + r_o) - \omega t]. \qquad (24)$$

There remains the difficulty of the $\pi/2$ phase shift between the cosine in Eq. (2) and the sine in Eq. (24). We leave its resolution as a question for the Kirchhoff theory (see Appendix) and turn to diffraction with specific apertures.

FRESNEL DIFFRACTION WITH AN APERTURE

Interference between the Huygens sources emanating from the spherical surface σ produces the wave function that arrives at P. Apply Eqs. (3) and (23), switch to complex notation for the harmonic dependence, and we obtain

$$d\psi = \frac{\mathcal{E}_o K(\theta) da}{\rho\lambda r} e^{i[k(\rho + r) - \omega t]} \qquad (25)$$

which will be integrated to find the contribution to the signal at P that comes from the Fresnel zones allowed pass through the aperture. In Eq. (25) the area *da* is a patch of area on the sphere σ, but the aperture is presumably a plane. However, where the sphere σ intersects the aperture in the plane of the aperture, the aperture boundary forms the limits of integration. Towards this end we construct an *xy* coordinate system in the aperture plane, with the origin lying at the intersection of that plane and the SP axis. Note in Fig. 3 the distinction between ρ and ρ_o, the latter being the distance from S to the origin of the aperture plane.

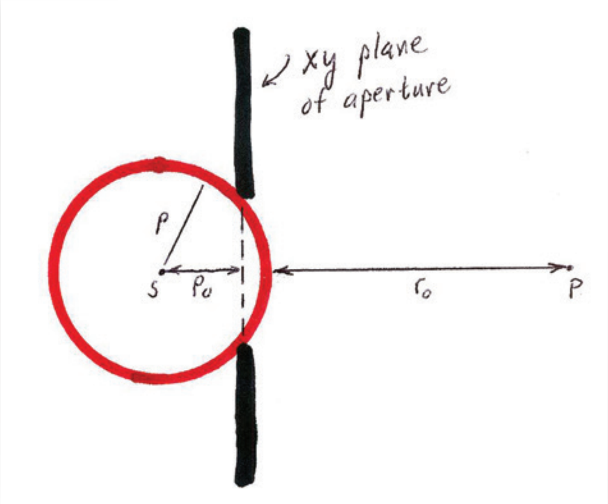

FIG. 3: A spherical wave front passing through the aperture of area Γ, and the area da.

In Eq. (25) we may approximate $\rho \approx \rho_o$ and $r \approx r_o$ in the *amplitude*. However, the *phase* is far more sensitive to small differences in distances. From Fig. 3, the Pythagorean theorem followed by the binomial expansion gives

$$\rho = (\rho_o^2 + x^2 + y^2)^{\frac{1}{2}} \approx \rho_o + \frac{x^2 + y^2}{2\rho_o} \qquad (26)$$

and

$$r = (r_o^2 + x^2 + y^2)^{\frac{1}{2}} \approx r_o + \frac{x^2 + y^2}{2r_o}. \qquad (27)$$

If only a small number of Fresnel zones are allowed to pass through the aperture, the approximation $K(\theta) \approx 1$ may be used for each contributing zone.

Now we are ready to integrate Eq. (25) over the aperture. In terms using rectangular coordinates,

$$\psi \approx \frac{\mathcal{E}_o}{\rho_o \lambda r_o} e^{i[k(\rho_o + r_o) - \omega t]} \int_\Gamma e^{i\pi\beta(x^2 + y^2)/2} da \qquad (28)$$

where

$$\beta \equiv \frac{2(\rho_o + r_o)}{\lambda \rho_o r_o}. \qquad (29)$$

From Eq. (1), the wave that *would* have arrived as a spherical wave front at P in the absence of an aperture would have been described by

$$\psi_o \approx \frac{\mathcal{E}_o}{\rho_o + r_o} e^{i[k(\rho_o + r_o) - \omega t]}. \qquad (30)$$

Writing $\mathcal{E}_o = \psi_o(\rho_o + r_o)e^{-i[k(\rho_o + r_o) - \omega t]}$ and using Eq. (29) turns Eq. (28) into

$$\psi = \psi_o \frac{\beta}{2} \int_\Gamma e^{i(x^2 + y^2)\pi\beta/2} da. \qquad (31)$$

The integral modifies the amplitude and phase of ψ_o thanks to the superposition of the portion of σ not blocked by the aperture.

Consider a rectangular aperture. The limits on Eq. (31) become

$$\psi = \psi_o \frac{\beta}{2} \int_{x_1}^{x_2} e^{ix^2\pi\beta/2} dx \int_{y_1}^{y_2} e^{iy^2\pi\beta/2} dy. \qquad (32)$$

Let $u = x\sqrt{\beta}$ and $v = y\sqrt{\beta}$. Then

$$\psi = \frac{1}{2}\psi_o \int_{v_1}^{v_2} e^{i\pi v^2/2} dv \int_{u_1}^{u_2} e^{i\pi u^2/2} du. \qquad (33)$$

Introduce the *Fresnel integral*[7]

$$\mathcal{Z}(W) \equiv \int_0^W e^{\frac{i\pi s^2}{2}} ds$$

$$\equiv \mathcal{C}(w) + i\mathcal{S}(w) \qquad (34)$$

where

$$\mathcal{C}(w) \equiv \int_0^w \cos\left(\frac{\pi s^2}{2}\right) ds \qquad (35)$$

and

$$\mathcal{S}(w) \equiv \int_0^w \sin\left(\frac{\pi s^2}{2}\right) ds. \qquad (36)$$

Now Eq. (33) may be written

$$\psi = \frac{1}{2}\psi_o[\mathcal{Z}(v_2) - \mathcal{Z}(v_1)][\mathcal{Z}(u_2) - \mathcal{Z}(u_1)]. \qquad (37)$$

$\mathcal{Z} = \mathcal{C} + i\mathcal{S}$ is a complex number. Starting at $w = 0$, compute successive values of $\mathcal{C}(w)$ and $\mathcal{S}(w)$, then plot them on the complex plane. This procedure generates the *Cornu spiral*, a phasor diagram (Fig. 4) named after the French physicist Alfred Cornu (1841-1902).

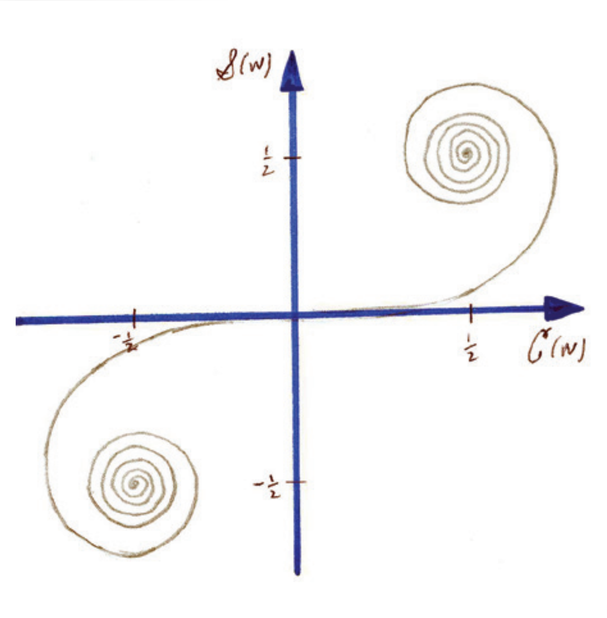

FIG. 4: The Cornu spiral (schematic).

To graphically determine $\mathcal{Z}(b) - \mathcal{Z}(a) \equiv Re^{i\delta}$ draw a straight line connecting $\mathcal{Z}(b)$ and $\mathcal{Z}(a)$. The length of this line is the magnitude $|\mathcal{Z}(b) - \mathcal{Z}(a)| = R$, and δ is the angle the line makes with the real axis.

From the definitions of $\mathcal{C}(w)$ and $\mathcal{S}(w)$, one may easily show that $\mathcal{C}(-w) = -\mathcal{C}(w)$ and $\mathcal{S}(-w) = -\mathcal{S}(w)$, and that $\mathcal{C}(\infty) = \mathcal{S}(\infty) = \frac{1}{2}$.[8] For a consistency check consider no aperture at all: The limits on both x and y go from $-\infty$ to $+\infty$, and Eq. (35) gives the expected intensity, $|\psi|^2/|\psi_o|^2 = 1$.

To map the wave function and intensity pattern off the SP axis—a main objective of diffraction theory— make an $x'y'$ plane parallel to the aperture plane and passing through P, and locate the origin at P. To examine the wave function at $(x',y') = (a',b')$, instead of moving P move the aperture in the reverse direction: Merely translate the origin of the aperture coordinates by the amount $\Delta x = -a'$ and $\Delta y = -b'$. Adjust the limits on the Fresnel integrals, use Eq. (37), and $\psi(a',b')$ results.

A circular aperture of radius a suggests mapping the plane of the aperture in polar coordinates, so that $da = 2\pi r dr$. Now Eq. (31) becomes

$$\psi = \pi\beta\psi_o \int_0^a e^{i\pi\beta r^2/2}\, r\, dr. \tag{38}$$

Let $z = \pi\beta r^2/2$, so that

$$\psi = \psi_o \int_0^{\beta\pi a^2/2} e^{iz}\, dz$$

$$= 2\psi_o e^{i\pi\beta a^2/4} \sin\left(\frac{\pi a^2 \beta}{4}\right). \tag{39}$$

Minima (maxima) occur when $\pi a^2 \beta/4 = m\pi$ where m denotes a positive integer (half an odd integer), which produces a result analogous to the thin lens equation of geometrical optics,

$$\frac{1}{\rho_o} + \frac{1}{r_o} = \frac{1}{f_m} \tag{40}$$

where $f_m = a^2/2m\lambda$. This exhibits serious chromatic aberration!

There remains the task to confirm with a derivation the behavior of the obliquity factor, $K(\theta)$, and put the whole Huygens-Fresnel approach on a less cobbled-together and more rigorous basis. The main ideas are described in the Appendix, highlighting the principal elements of Kirchhoff's diffraction theory.

APPENDIX: KIRCHHOFF'S DIFFRACTION THEORY

The German physicist Gustav Kirchhoff (1824-1887) derived a theory of diffraction based on the linear wave equation.[9] It uses a standard tool of field theory, Gauss's divergence theorem, which says the integral of the divergence of a vector field **A** over a volume V equals the flux of **A** through the closed surface Σ that forms the boundary of V:

$$\int_V \nabla \cdot \boldsymbol{A}\, d^3r = \oint_\Sigma \boldsymbol{A} \cdot \hat{\boldsymbol{n}}\, da \tag{A1}$$

where $\hat{\boldsymbol{n}}$ denotes the outward-pointing unit vector normal to Σ on the patch of area da. Consider the case $\boldsymbol{A} = \zeta\, \nabla\eta$ for some scalar fields ζ and η. The divergence theorem gives

$$\int_V [\zeta\nabla^2\eta + (\nabla\zeta)\cdot(\nabla\eta)]\, d^3r = \oint_\Sigma \zeta \frac{\partial\eta}{\partial n}\, da \tag{A2}$$

where $\frac{\partial\zeta}{\partial n} \equiv (\nabla\zeta)\cdot\hat{\boldsymbol{n}}$ denotes the directional derivative. Interchange ζ and η and subtract the two versions to obtain "Green's identity,"[10]

$$\int_V [\zeta\nabla^2\eta - \eta\nabla^2\zeta]\, d^3r = \oint_\Sigma \left[\zeta\frac{\partial\eta}{\partial n} - \eta\frac{\partial\zeta}{\partial n}\right] da. \tag{A3}$$

Now connect this fancy mathematics to some physics. For wave propagation let η be a solution to the inhomogeneous wave equation,

$$\nabla^2\eta(\boldsymbol{r},t) - \frac{1}{c^2}\frac{\partial^2\eta(\boldsymbol{r},t)}{\partial t^2} = -4\pi J(\boldsymbol{r},t) \tag{A4}$$

where J denotes a source density. Maxwell's equations lead to such a wave equation. The total signal is a superposition of harmonics, where each harmonic carries some angular frequency ω. Writing $\eta(\boldsymbol{r},t) = \eta(\boldsymbol{r})e^{-i\omega t}$ and similarly for J, the wave equation becomes the inhomogeneous Helmholtz equation for $\eta(\boldsymbol{r})$,

$$\nabla^2\eta(\boldsymbol{r}) + k^2\eta(\boldsymbol{r}) = -4\pi J(\boldsymbol{r}) \tag{A5}$$

where $k = \omega/c$. The Helmholtz equation can be solved by the method of Green's functions, for which a function $G(\mathbf{r} - \mathbf{r}')$ is found that solves the Helmholtz equation when J gets replaced by a point source,

$$\nabla^2 G + k^2 G = -4\pi\delta^3(\boldsymbol{r} - \boldsymbol{r}'). \tag{A6}$$

The Dirac delta function $\delta^3(\boldsymbol{r} - \boldsymbol{r}')$ is the density of a point source (think of a point mass): It vanishes everywhere except at the source itself, where it blows up. Yet its integral is finite (e.g., the integral over all the particle's density equals its mass, whatever its distribution). Thus the Dirac delta may be operationally defined as

$$\int_V f(\boldsymbol{r})\delta^3(\boldsymbol{r}-\boldsymbol{a})d^3r = f(\boldsymbol{a}) \tag{A7}$$

provided $\boldsymbol{a}$ lies within V; otherwise the integral vanishes. Once G is found, then η follows from the original source density J, by Green's theorem,[11]

$$\eta(\boldsymbol{r}) = \int J(\boldsymbol{r}')G(\boldsymbol{r}-\boldsymbol{r}')d^3r' \tag{A8}$$

where the integral is over all space. As shown in electrodynamics textbooks, the Green's function for the Helmholtz equation, with outward-traveling spherical waves, is a damped oscillation:[12]

$$G(\boldsymbol{r}-\boldsymbol{r}') = \frac{e^{ikR}}{R} \tag{A9}$$

where $R = |\boldsymbol{r}-\boldsymbol{r}'|$. In Eq. (A3) let $\eta = G$, and let ζ be the "optical disturbance," e.g., the electric potential ψ of the electromagnetic field. With the help of Eq. (A5) in source-free regions, and Eqs. (A6) and (A7), Eq. (A3) becomes

$$\psi(\boldsymbol{r}) = \frac{1}{4\pi}\oint_\Sigma \left[G\frac{\partial\psi}{\partial n} - \psi\frac{\partial G}{\partial n}\right]da. \tag{A10}$$

For the surface Σ consider the arrangement of Fig. 5. Σ does not enclose the original wave source S, but it does enclose the point P, and the point H lies on the surface of Σ.

From Eq. (2), on Σ at H,

$$\psi = \frac{\mathcal{E}_o}{\rho}e^{i(k\rho-\omega t)} \tag{A11}$$

so that

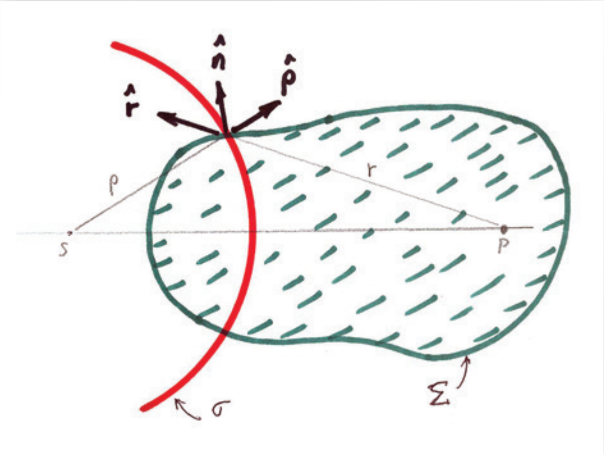

FIG. 5: The closed surface Σ used in Eq. (A10).

$$\left.\frac{\partial\psi}{\partial n}\right|_\Sigma = (\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{\rho}})\left[\frac{ik}{\rho}-\frac{1}{\rho^2}\right]\mathcal{E}_o e^{i(k\rho-\omega t)}. \tag{A12}$$

Similarly, by Eq. (A9),

$$\left.\frac{\partial G}{\partial n}\right|_\Sigma = (\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}})\left[\frac{ik}{r}-\frac{1}{r^2}\right]e^{ikr}. \tag{A13}$$

Now Eq. (A10) becomes

$$\psi(\boldsymbol{r}) = \frac{\mathcal{E}_o}{4\pi}\oint_\Sigma \left[\frac{e^{ik(\rho+r)}}{r}\left(\frac{ik}{\rho}-\frac{1}{\rho^2}\right)(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{\rho}}) - \frac{e^{ik(\rho+r)}}{\rho}\left(\frac{ik}{r}-\frac{1}{r^2}\right)(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}})\right]da. \tag{A14}$$

If $\lambda << r$ and ρ, then we may neglect the $1/r^2$ and $1/\rho^2$ terms. Eq. (A14) becomes

$$\psi(\boldsymbol{r}) \approx \frac{-i\mathcal{E}_o e^{-i\omega t}}{4\pi}\oint_\Sigma \frac{e^{ik(\rho+r)}}{r\rho}\left(\frac{(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}})-(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{\rho}})}{2}\right)da. \tag{A15}$$

The differential form of Eq. (A15) says that

$$d\psi(\boldsymbol{r}) \approx \frac{\mathcal{E}_o}{\lambda}\frac{e^{i[k(\rho+r)-\omega t-\frac{\pi}{2}]}}{r\rho}\left(\frac{(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}})-(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{\rho}})}{2}\right) \tag{A16}$$

where $-i = e^{-i\pi/2}$ has been used. Eq. (3), when rewritten as a complex harmonic, appears as

$$d\psi = \frac{\mathcal{E}_o}{\rho r\lambda}K(\theta)\, e^{i[k(\rho+r)-\omega t]}da \tag{A17}$$

where Eq. (23) has also been used. Comparing Eqs. (A16) and (A17), the former has the correct phase, and comparing the coefficients shows the obliquity factor to be

$$K(\theta) = \frac{1}{2}[(\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{r}}) - (\hat{\boldsymbol{n}}\cdot\hat{\boldsymbol{\rho}})]. \tag{A18}$$

It remains to evaluate the dot products between the unit vectors, expressing them in terms of observables. Consider for this application the doubly-connected surface Σ (Fig. 6), whose outer surface Σ_2 encloses point P, and whose inner surface Σ_1 excludes the source point S. A study of Fig. 6 shows that Eq. (A18) reduces to Eq. (4). //

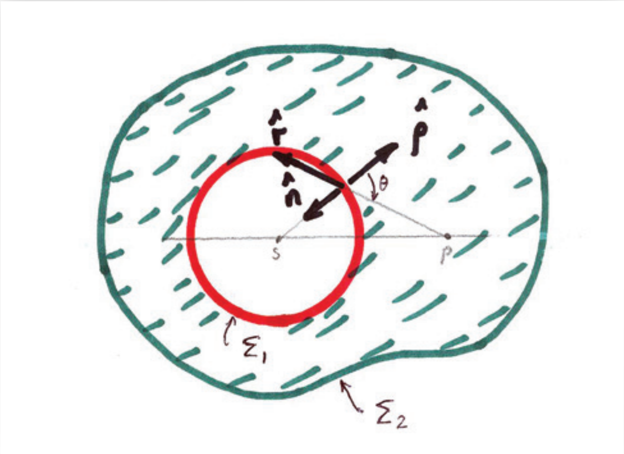

FIG. 6: The doubly-connected closed surface Σ made by inner surface Σ_1 and outer surface Σ_2.

ACKNOWLEDGEMENT

Thanks to Toni Sauncy for carefully reading a draft of this manuscript and making useful suggestions.

REFERENCES AND NOTES

[1] W.P.D. Wightman, *The Growth of Scientific Ideas* (Yale Univ. Press, New Haven, CT, 1953), pp. 146–150.

[2] "Elegant Connections in Physics, Diffraction Part 1: Huygens' Principle and Young's Interference," *SPS Observer*, Winter 2012–13, pp. 20–23.

[3] "Elegant Connections in Physics, Diffraction, Part 2: Multiple Point Sources, Apertures, and Diffraction Limits," *SPS Observer*, Spring 2013, pp. 18–23.

[4] At a distance r from a point source of luminosity (power output) L, the intensity I (power per unit area received) relates to L according to $I = L/4\pi r^2$. For light waves, the intensity is also given by the time-average of the Poynting's vector $\mathbf{S} = (\mathbf{E}\times\mathbf{B})/\mu_o$. Since $\mathbf{E}$ and $\mathbf{B}$ are perpendicular, $|\mathbf{E}| = c|\mathbf{B}|$ and $1/c^2 = \mu_o\, \varepsilon_o$, the result follows.

[5] Eugene Hecht, *Optics,* 4th ed. (Addison–Wesley, San Francisco, CA, 2002), pp. 485–489; Francis A. Jenkins and Harvey E. White, *Fundamentals of Optics* (McGraw–Hill, New York, 1950), p. 352. The arguments go like this: For odd N the series may be grouped as

$$Z_N = \frac{K_1}{2} + \left(\frac{K_1}{2} - K_2 + \frac{K_3}{2}\right) + \left(\frac{K_3}{2} - K_4 + \frac{K_5}{2}\right) + \ldots$$

$$+ \left(\frac{K_{N-2}}{2} - K_{N-1} + \frac{K_N}{2}\right) + \frac{K_N}{2}$$

$$= \frac{K_1}{2} + \left(\frac{K_1 + K_3}{2} - K_2\right) + \left(\frac{K_3 + K_5}{2} - K_4\right) + \ldots + \frac{K_N}{2}$$

$$\equiv \frac{K_1 + K_N}{2} + R \qquad \text{(a)}$$

where R denotes a "remainder." If the nth obliquity factor is greater than the average of its two neighboring ones, then $R < 0$ and $Z_N < ½(K_1 + K_N)$. A different grouping gives

$$Z_N = K_1 + K_N - \frac{K_2 + K_{N-1}}{2} - T \qquad \text{(b)}$$

with remainder T. Again, if the nth obliquity factor is larger than the mean of its neighbors, then $T < 0$. Because $K_{n+1} - K_n$ is small, for N not very large inequality Eq. (b) gives $Z_N > ½(K_1 + K_N)$. Consistency between the two inequalities requires $Z_N \approx ½(K_1 + K_N)$. The same conclusion holds when the nth obliquity factor is less than the mean of its neighbors.

[6] Another way to describe the effect of the obliquity factor is through the "vibration curve," a phasor diagram. See Hecht (Ref. 5), p 489; Jenkins and White (Ref. 5), p. 286; Bruno Rossi, *Optics* (Addison-Wesley, Reading, MA, 1965), p. 162.

[7] Fresnel integrals and the Cornu spiral are well tabulated numerically and graphically, e.g., see Hecht (Ref. 5), pp. 497–509; Jenkins and White (Ref. 5), pp. 358–366; Rossi (Ref. 6), pp. 189–198.

[8] This result requires the Gaussian integral,

$$\int_{-\infty}^{+\infty} e^{-ax^2}\,dx = \sqrt{\frac{\pi}{a}}.$$

[9] Gustav Kirchhoff, "Zür Theorie der Lichtstrahlen" ("On the Theory of Light Rays"), *Annalen der Physik-Berlin* **254** (4), 663–695 (1883). See also Hecht (Ref. 5), p. 510.

[10] George Green (1793–1841) was an amazing fellow, brilliant and self-taught in mathematics. Despite being stuck in the family's bakery and grain milling business, Green laid down some fundamental results in vector calculus and differential equations that would be used later by James Maxwell in his theory of the electromagnetic field.

[11] "Discrete Sources and the Continuum, and the Functions of Dirac and Green," *SPS Newsletter* (October 1996), pp.10–12.

[12] E.g., see J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1975), p. 224.