Exact derivation of Kirchhoff's integral theorem and diffraction formula using high-school math

Gavin R. Putland *

Version 4; April 9, 2020

A theory of diffraction is developed from three premises: causality (including the assumption that the wave function had a beginning), superposition, and the form of the wave function due to a monopole source. It is shown that the wave function in a region $\mathcal{R}$, due to sources outside $\mathcal{R}$ (that is, in $\mathcal{R}'$), is identical to that due to a distribution of spatiotemporal dipole (STD) sources on the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$, facing $\mathcal{R}$, with a strength density given by the original wave function on $\mathcal{S}$ (divided by 4π, under the adopted definition of strength). The Kirchhoff integral theorem and consequent diffraction formulae are obtained by superposing the elemental wave functions due to the sources on $\mathcal{S}$. Whereas D.A.B. Miller (1991) justified the STDs by comparison with Kirchhoff's integral, I derive the integral (with a near-primary-source correction to Miller's form) solely from the STDs. Diffraction by an aperture in an opaque screen is modeled by assuming that we retain only the sources on that part of $\mathcal{S}$ which spans the aperture. This assumption avoids the notorious inconsistency in Kirchhoff's boundary conditions (1882), but gives the same diffraction integral. It also yields the consistent saltus conditions of Kottler (1923) and a more direct saltus description of the aperture, giving the same solution in $\mathcal{R}$. Finally, the diffraction integral for a monopole primary source is transformed to a geometrical-optics term plus an integral along the edge of the aperture, and applied to the Poisson-Arago spot. The transformation (due to Maggi and Rubinowicz) is first derived from a far-field/short-wavelength assumption, but later re-derived exactly for general time-dependence, and shown to be a valid approximation for a weakly directional primary source.

Contents

1 Preliminaries 2
- 1.1 Wave functions; boundary conditions; Huygens' principle . . . 2
- 1.2 *Naive* boundary conditions . . . 4
- 1.3 Plan of the paper . . . 4

2 Matching the boundary conditions 5
- 2.1 First attempt: Monopole sources . . . 5
- 2.2 Second attempt: Spatiotemporal dipole (STD) sources . . . 7
- 2.3 Note on electromagnetic waves . . . 9

3 Helmholtz formula; Kirchhoff integral theorem 9
- 3.1 Derivation . . . 9
- 3.2 Application to diffraction by an aperture . . . 11
- 3.3 Exact? Consistent? Well-posed? . . . 12

4 Point-source: Kirchhoff diffraction formula 13

5 Integration over a primary wavefront 15
- 5.1 Spherical wavefront . . . 15
- 5.2 Plane wavefront: Approximation for non-spherical wavefront . . . 15

6 The sinusoidal (monochromatic) case 16
- 6.1 Exact forms . . . 16
- 6.2 Approximate forms . . . 18

7 Notes on Miller (1991) 19

8 Appendices 19
- 8.1 Diffraction as a saltus problem . . . 19
- 8.2 Diffraction as an edge effect: Maggi-Rubinowicz transformation; Poisson's spot . . . 21
- 8.3 Exactness of the transformation; approximation for a directional primary source . . . 28
- 8.4 Revision history . . . 31

9 Conclusion 31

* Melbourne, Australia. No institutional affiliation at the time of writing. Gmail address: grputland.

1 Preliminaries

1.1 Wave functions; boundary conditions; Huygens' principle

Suppose that we have a number of sources of waves (e.g., of light or sound) in a three-dimensional medium. (In this paper, a vacuum counts as a "medium".) Let the resulting combination of waves be described by the function $\psi(P,t)$, representing some physical quantity which depends on the position P and the time t; we shall call ψ the **wave function**. We shall always suppose that the wave function had a **beginning**—in other words, that there was a time before which ψ was zero everywhere. Let $\mathcal{R}$ be a region containing *none* of the sources; let $\mathcal{R}'$ be the remaining region, containing *all* the sources; and let $\mathcal{S}$ be the geometric surface (*not* a physical barrier) separating the two regions. We do not care whether $\mathcal{S}$ is a closed surface with the sources outside and $\mathcal{R}$ inside, or a closed surface with the sources inside and $\mathcal{R}$ outside, or an infinite open surface with the sources on one side and $\mathcal{R}$ on the other; the essential feature of all these cases is that *the waves cannot enter the region $\mathcal{R}$ except by crossing the surface $\mathcal{S}$*. When they cross the surface, their influence, as always, travels sequentially from point to neighboring point, without "action at a distance". Therefore, due to the chain of causation:

Proposition 1 *If a region $\mathcal{R}$, bounded by a surface $\mathcal{S}$, contains no sources, the behavior of the wave function throughout $\mathcal{R}$ is fully determined by its behavior on $\mathcal{S}$.*

That immediately raises three questions:

- What information on the wave function at the boundary surface is sufficient to determine the function throughout the region?
- Why might it be useful to determine the wave function in this way?
- What formula or algorithm yields the wave function at a given point in the region, in terms of the said information about the function at the boundary surface?

Apropos of the first question, the **normal derivative** of the wave function ψ at the surface $\mathcal{S}$ is defined as the derivative of ψ with respect to a coordinate n, which measures the normal (perpendicular) distance from $\mathcal{S}$ into $\mathcal{R}$; this derivative, which we shall write as ψ_n or $\frac{\partial\psi}{\partial n}$, is called a *partial* derivative because it is the derivative of ψ w.r.t. *one* variable while other variables, on which ψ also depends, are held constant.[1] The derivative is to be evaluated at $\mathcal{S}$, on the side of $\mathcal{S}$ that faces $\mathcal{R}$. (We specify the side because there may be a step-change in the derivative at $\mathcal{S}$, as we shall see.)

At every point on $\mathcal{S}$, on the side facing $\mathcal{R}$, let *either* the wave function *or* its normal derivative be specified for all time. We call this specification a **boundary condition** (or **BC**) "on" the wave function or its normal derivative. Suppose that this condition is *not* sufficient to determine the wave function throughout $\mathcal{R}$. Then there are at least two different wave functions defined in $\mathcal{R}$, which can be generated by (different) sets of sources entirely outside $\mathcal{R}$, and which satisfy the same BC on $\mathcal{S}$. By superposition, the non-zero difference between those functions is also a wave function that can be generated by sources entirely outside $\mathcal{R}$; and at every point on $\mathcal{S}$, either this function or its normal derivative is *zero* for all time. But this is a *reductio ad absurdum* because, given that the wave function had a beginning,[2] the initial entry of the disturbance into $\mathcal{R}$ at any point on its boundary must disturb both the wave function and its normal derivative, so that *neither* can be perpetually zero at that point. So we must conclude, contrary to our initial supposition, as follows:

Proposition 2 *If a region $\mathcal{R}$, bounded by a surface $\mathcal{S}$, contains no sources, the specification of either the wave function or its normal derivative at $\mathcal{S}$ (for all time) determines the wave function throughout $\mathcal{R}$.*

[1] *Warning*: Some authors measure n in the opposite direction, changing the signs of all terms in ψ_n. The sign convention for n must be borne in mind when comparing results. In this paper, n is measured *into* the region containing *no sources*.

[2] If we did not assume that the wave function had a beginning, we would also need to specify *initial conditions* on the wave function and its derivative w.r.t. time (*cf.* Baker & Copson [1], p. 41). This is to be expected because, if the time line goes back far enough, the assumption of a beginning *does* specify initial conditions—namely that the initial values were zero.

If it suffices to specify either the wave function or the normal derivative at the boundary, it certainly suffices to specify both. But, because either boundary condition determines the whole wave function throughout $\mathcal{R}$, either boundary condition determines the other; so *we cannot choose both arbitrarily* (*cf.* Baker & Copson [1], pp. 38, 40–42).

Concerning the second question, suppose that we have an aperture (such as a slit or an iris) in an otherwise opaque screen, with a point-source on one side of the screen, and suppose that we want to predict the wave function on the other side. Experience says that the edge of the shadow cast by the screen will not be perfectly sharp, even for the point-source, but will be blurred and fringed—an effect known as **diffraction**. The same phenomenon causes even the depths of the shadow to be less than perfectly dark or silent. In this situation, let us choose the surface $\mathcal{S}$ so that one part of it, say $\mathcal{S}_a$, spans the *a*perture, while the remaining part, say $\mathcal{S}_b$, is on the *b*ack (dark side or quiet side) of the screen. Then we might suppose, as a first approximation, that the wave function (or its normal derivative) on $\mathcal{S}_b$ is zero, while the wave function (or its normal derivative) on $\mathcal{S}_a$ is the same as if the screen were absent. The conditions at these boundaries determine the wave function in the region beyond the screen—including those parts in, and near the edge of, the geometric shadow of the screen, where the wave function is certainly *not* the same as if the screen were absent. (The *geometric* shadow is the shadow that would be predicted by thinking in terms of *rays* instead of waves.) Thus we might hope to explain and predict diffraction in a quantitative manner.

On the third question, recall that we have a wave function $\psi(P,t)$ in regions $\mathcal{R}'$ and $\mathcal{R}$, due to sources in $\mathcal{R}'$ only, hence certain values of ψ and its normal derivative ψ_n on the surface $\mathcal{S}$ that separates $\mathcal{R}'$ and $\mathcal{R}$. Now suppose that, by eliminating the original sources and introducing a suitable continuous distribution of sources over the surface $\mathcal{S}$, we can reproduce the original values of ψ or ψ_n at the $\mathcal{R}$ side of the surface, while setting ψ or ψ_n to *zero* at the $\mathcal{R}'$ side. If we take $\mathcal{R}$ and $\mathcal{R}'$ as excluding $\mathcal{S}$, neither region contains the sources. Hence, by the sufficiency of the boundary condition (BC) in each region, we reproduce the original wave function in $\mathcal{R}$ while setting the wave function to *zero* in $\mathcal{R}'$. That is:

Proposition 3 *Let the geometric surface $\mathcal{S}$ divide the medium into a region $\mathcal{R}$, containing no sources, and a region $\mathcal{R}'$.* Case 1: *Let sources in $\mathcal{R}'$ give a certain ("original") wave function in $\mathcal{R}'$ and $\mathcal{R}$, hence a certain ("original") BC at $\mathcal{S}$.* Case 2: *Let those sources be replaced by a distribution of sources on $\mathcal{S}$ which, by themselves, give the original BC at the $\mathcal{R}$ side of $\mathcal{S}$ and a zero BC at the $\mathcal{R}'$ side. Then we get the original wave function in $\mathcal{R}$ and a zero wave function in $\mathcal{R}'$.*

In Case 2, the wave function in each region is the sum of the contributions from the sources on $\mathcal{S}$; we call this sum a **surface integral** over $\mathcal{S}$. It is understood that the boundary condition may be on the wave function *or* its normal derivative.

The last answer raises a further question: If we only want to reproduce the wave function in $\mathcal{R}$, why set it to zero in $\mathcal{R}'$? There are three reasons. First, a unique distribution of sources on $\mathcal{S}$, *by itself*, obviously gives a unique wave function on each side of $\mathcal{S}$; hence, if we did not specify the wave function on both sides, the required distribution of sources (*if* such a distribution is possible) would not be unique. Second, if we can find a distribution of sources over $\mathcal{S}$ that reproduces the original wave function in $\mathcal{R}$, it will show that the wave function in $\mathcal{R}$ is *as if* the incident wave turned the surface $\mathcal{S}$ into a particular distribution of **secondary sources** replacing the original sources; in other words, it will verify **Huygens' principle** for the surface $\mathcal{S}$ and the region $\mathcal{R}$. If, in addition, that distribution of sources yields a null (zero) wave function in $\mathcal{R}'$, we will have found secondary sources that cause no "backward" or "retrograde" secondary waves. This is critical because $\mathcal{S}$ may be curved, or even closed around $\mathcal{R}'$, so that any radiation from the secondary sources "back" into $\mathcal{R}'$ could continue across $\mathcal{S}$ and into $\mathcal{R}$ again, *adding to* the wave function in $\mathcal{R}$ and thereby invalidating the setting of its BCs. That is the implication of the following statement by Joseph Larmor, in which the "inside" means $\mathcal{R}'$ (in our notation) and the "outside" means $\mathcal{R}$:

> A state of stress and strain is continually transmitted up to the surface $\mathcal{S}$ from the actual sources inside, and we are to find a distribution of secondary sources that will send it on to the outside as

> it arrives, without sending anything back. For the sending of a disturbance back into the interior would be an alteration of the physical circumstances, would in fact *add to* and confuse the effect of the assigned true sources inside [10, p. 172; my emphasis].

Third, as we are about to see, the choice of a null wave function in $\mathcal{R}'$, together with the assumption that the original wave function had a beginning, simplifies the specification of the sources on $\mathcal{S}$.

1.2 *Naive* boundary conditions

For the purposes of this paper, let us define a **naive** wave function or boundary condition as one that would be calculated at each point by considering only the sources within some uniform finite distance of that point, neglecting sources further away.[3] Then:

Proposition 4 *In Case 2 of Proposition 3, the naive BCs at $\mathcal{S}$ are the actual BCs.*

Proof: At all times before the original disturbance enters $\mathcal{R}$, the original BC at the $\mathcal{R}$ side of $\mathcal{S}$ is null, so that, by hypothesis, the sources on $\mathcal{S}$ are those that give a null BC at both sides—at all points on $\mathcal{S}$, at all such times. Those sources at those times must be *null* sources, which also give null *naive* BCs. So Proposition 4 is true at all times before the disturbance enters $\mathcal{R}$. To prove that it remains true thereafter, it suffices to use an inductive argument in which we suppose that the proposition is true up to time t (the "inductive hypothesis"), and deduce that it is true at the later time $t+\Delta t$, where Δt is chosen to be less than the time taken by the waves to travel the finite distance considered in the calculation of the naive wave function or BCs. Let this time be τ_{n}. At time $t+\Delta t$, the actual wave function at each point P, *including the actual BCs* at each point Q on $\mathcal{S}$, cannot differ from the naive wave function (including the naive BCs) except because of waves which have been traveling for longer than τ_{n}, and which were therefore emitted before time t, from sources at points on $\mathcal{S}$. Let Q' denote any such point. Because wave propagation is sequential, those sources cannot have influenced the wave function at P, including the BCs at Q, except via the BCs at the points Q', likewise at time t or earlier. By the inductive hypothesis, those BCs agree with the naive BCs, completing the inductive step. ■

1.3 Plan of the paper

In the next Section, we seek a distribution of sources on $\mathcal{S}$ for which the naive boundary conditions on the wave function are the original value at the $\mathcal{R}$ side and a null value at the $\mathcal{R}'$ side. We find a solution in the form of **spatiotemporal dipoles** (**STDs**), as first described by D.A.B. Miller [12]. In Section 3, we use the STD distribution to express the wave function at a general point in $\mathcal{R}$ (the **field point**) as the surface integral, over $\mathcal{S}$, of an expression involving the wave function and two of its derivatives. The easiest way to model an obstruction is to suppose that it simply eliminates the sources on part of $\mathcal{S}$ and thereby restricts the range of integration. In Section 4 we specialize the integrand for the case of a single primary source. The dominant term of the integrand is seen to contain an **obliquity factor** involving two angles, made by the source and the field point with the normal to the surface of integration. In Section 5, we further specialize the integral for the case considered by Fresnel, in which the surface of integration is a primary wavefront, so that the obliquity factor reduces to a function of latter angle (the former being fixed at zero). We also consider the case in which the primary wavefront may be non-spherical but not too sharply curved. In Section 6, we rewrite our results for the **monochromatic** case, in which the time-dependence of the wave function is *sinusoidal*. In Section 7, we add a near-source correction term to Miller's result for a spherical primary wavefront. This does not require any correction to Miller's spatiotemporal dipoles, because it is derived *from* them. In an Appendix, it is shown that the standard method of modeling diffraction by an aperture in a screen—evaluating the integrand as if the screen were not there, and integrating over the aperture only—is the exact solution to two other problems involving step-changes in the wave-function and its normal derivative. A second Appendix shows that in the case of a single primary source, diffraction can be represented mathematically as what it superficially appears to be: an edge effect. This representation, although initially derived by an approximate method, is shown in a third Appendix to be exact.

[3] This concept does not seem to have a standard name. For brevity, it needs a name; and for clarity, the name ought to convey a hint of skepticism.

2 Matching the boundary conditions

2.1 First attempt: Monopole sources

Consider the function

$$f(t - s/c)\,, \tag{1}$$

where t is time, s is a coordinate measuring distance, and c is a constant. The *argument* of the function is $t-s/c$. Let some local feature of the function occur where the argument has a particular value. If we set the argument equal to that value (a constant) and differentiate with respect to t, we obtain $\frac{ds}{dt} = c$, indicating that the feature moves in the direction of s at speed c (indeed the symbol c comes from the Latin *celeritas*, meaning speed). Thus the function (1) describes a wave or waves moving at that speed, and may therefore be called a wave function.

In the argument of the function f above, the presence of the $-s/c$ term means that the time t must be increased—that is, *delayed*—by s/c in order to obtain the same argument, hence the same function value, as if that term were absent. So the function (1) may be understood as $f(t)$ delayed by the time s/c, which is the time taken by the waves to travel the distance s. This delayed function may be written $[f(t)]$ or simply $[f]$, where the square brackets indicate that the contents are to be delayed by the propagation time. We shall use this notation later in the paper. Accordingly, we shall avoid using square brackets as ordinary parentheses.

In three dimensions, a wave function is generally a function of three spatial coordinates and time. But special cases may reduce the number of coordinates. From now on, let us suppose that the medium is **homogeneous** (meaning that its properties, including c, are *independent of location* within the medium) and **isotropic** (meaning that its properties are *independent of direction*).[4] In such a medium, let us define a **monopole** source of **strength** $f(t)$ as a point-source giving the wave function

$$\frac{1}{r}\, f(t - r/c)\,, \tag{2}$$

where r is the radial distance from the source.[5] A wave function of form (2) describes **spherical waves** receding from the source at speed c, and is specified in terms of a *single* spatial coordinate r. Note that we have defined the "strength" of the source as a function of time. We shall need the assumption that the function is differentiable and therefore continuous. In (2), the function is not only delayed by the propagation time for the distance r, but also attenuated so that its amplitude is inversely proportional to that distance. Consequently, if the *intensity* (power per unit area) is proportional to the square of the wave function, the intensity satisfies the inverse-square law (for a given argument of the function).

Now let $M(t)$ be the *strength density* of monopole sources on the surface $\mathcal{S}$, so that the source strength corresponding to a surface element of area $d\mathcal{S}$ is $M(t)\,d\mathcal{S}$. By formula (2), this element's contribution to the wave function at a distance x from the element is

$$d\psi = \frac{1}{x}\, M(t - x/c)\, d\mathcal{S}\,. \tag{3}$$

(We call this distance x instead of r, because we are reserving r for something else.) For the purpose of finding boundary conditions at the surface $\mathcal{S}$, it suffices to consider a field point N at an arbitrarily small distance n from $\mathcal{S}$. Suppose that $\mathcal{S}$ is smooth. Then, for sufficiently small n, the surface $\mathcal{S}$ may be assumed flat and $M(t)$ may be assumed uniform on $\mathcal{S}$, over distances that are not very large compared with n. But we cannot let these distances go to infinity without invalidating our assumptions.

For a flat $\mathcal{S}$ and a uniform $M(t)$, we may choose the element $d\mathcal{S}$ to be an annulus of radius ρ and infinitesimal width $d\rho$, centered on the foot of the perpendicular from N to $\mathcal{S}$ (Fig. 1), because all parts

[4] A homogeneous medium is sometimes called a *uniform* medium. The properties of a homogeneous or isotropic medium are likewise said to be homogeneous or isotropic, respectively.

[5] In defining the *strength*, I follow the old convention used by (e.g.) Baker & Copson [1, p. 42], Born & Wolf [2, p. 421], and Larmor [9, p. 5]. Miller [12, p. 1371] implicitly uses the denominator $4\pi r$ instead of my r. Fortunately the effects of this clash of conventions will usually cancel out.

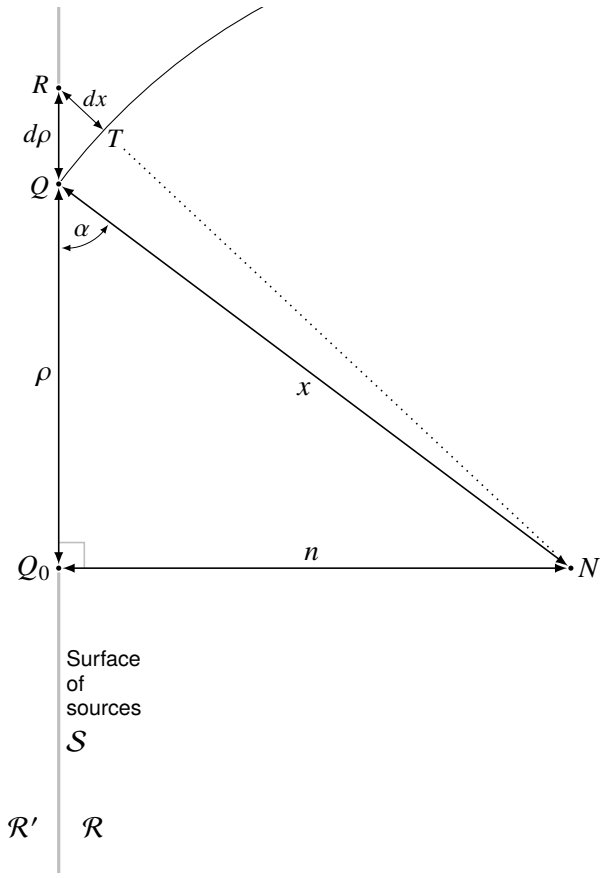

Fig. 1: If the normal coordinate n is so small that the surface $\mathcal{S}$ (seen edge-on) can be considered flat, we can take the surface element $d\mathcal{S}$ as an annulus with axis Q_0N, inner radius ρ, and width $d\rho$.

of this element have the same values of x and M in (3). Hence, if we are to remain consistent with our assumptions, we cannot have $\rho \to \infty$ or $x \to \infty$. The area of the chosen surface element is

$$d\mathcal{S} = 2\pi\rho\, d\rho\,, \tag{4}$$

so that (3) becomes

$$d\psi = \frac{2\pi}{x}\, M(t - x/c)\, \rho\, d\rho\,. \tag{5}$$

To integrate this, we want it all in terms of x or ρ, not a mixture of the two. In Fig. 1, for infinitesimal $d\rho$, the arc QT (centered on N) may be taken as straight, so that the angles QNQ_0 and RQT are equal, both being complementary to α. So the two triangles are similar, with $d\rho/dx = x/\rho$, whence

$$\rho\, d\rho = x\, dx\,. \tag{6}$$

With this substitution, (5) reduces to

$$d\psi = 2\pi\, M(t - x/c)\, dx\,. \tag{7}$$

According to the geometry of Fig. 1, the distance x should range from n to ∞, so that the whole wave function at N should be

$$\psi = \int_n^\infty 2\pi\, M(t - x/c)\, dx\,.$$

This is for positive n. To make it valid on both sides of $\mathcal{S}$, we must replace n by $|n|$, because the derivation assumes that n is measured away from $\mathcal{S}$, whereas n was originally defined as *into* $\mathcal{R}$. With that correction, the wave function at N would become

$$\psi^+ = \int_{|n|}^\infty 2\pi\, M(t - x/c)\, dx\,. \tag{8}$$

I say "According to" and "should" and "would", because we do not know the geometry of $\mathcal{S}$ or the behavior of M for $x \gg n$; and this matters because there is no factor making the integrand negligible for large x. So (8) does not yet give a usable value of ψ. Neither does it allow the boundary condition on ψ^+ to be zero at the $\mathcal{R}'$ side and non-zero at the $\mathcal{R}$ side (as in Proposition 3), because that requires a discontinuity in the integral, which is impossible because (i) there is nothing to make the integrand infinite as $n \to 0$, and (ii) the uncertainty for $x \gg n$ cannot cause any discontinuity in ψ^+ as we cross $\mathcal{S}$, because the effect of this crossing on x is infinitesimal.

But I chose a sheet of monopoles as a "first attempt", not because it had any chance of succeeding, but because it is a building-block with which we might construct something better. For the final product I am indebted to Miller [12]; but my method of assembling it is different from his.

2.2 Second attempt: Spatiotemporal dipole (STD) sources

Suppose that we have a second sheet of monopoles, called the **inverted** monopoles, whose strengths are equal and opposite to those in Fig. 1 and equation (8), and whose positions are displaced by a small distance h in the $-n$ direction (normal to $\mathcal{S}$, toward $\mathcal{R}'$) relative to their uninverted counterparts. And suppose that the strength functions of the inverted monopoles are *delayed* by h/c (the propagation time for the distance h). Then, on the $\mathcal{R}'$ side of both sheets, this delay compensates for the shorter propagation time from the inverted sheet; but on the $\mathcal{R}$ side of both sheets, there is no such compensation, because the delay in the inverted sources adds to the extra propagation time from their sheet, which is the more distant sheet as seen from that side. Hence we might hope that the waves emitted by the two sheets cancel on the $\mathcal{R}'$ side but not on the $\mathcal{R}$ side. We might further hope that by making the strengths of the monopoles *inversely* proportional to h, and taking limits as $h \to 0$, we can get a non-zero wave function on the $\mathcal{R}$ side of the surface $\mathcal{S}$, with a discontinuity in the wave function at the surface.

Now let us see whether our hopes are mathematically verified. The contribution to the wave function from the uninverted sheet of sources is ψ^+, given by equation (8). We need to modify this equation so as to obtain the inverted sheet's contribution, which we shall call ψ^-. For the inversion, we simply change the sign. For the delay, we replace t by $t-h/c$. For the displacement h, we replace n by $n+h$, so that putting $n=-h$ gives the argument that was formerly given by $n=0$. Thus we obtain

$$\psi^- = -\int_{|n+h|}^{\infty} 2\pi\, M\Big(t - \frac{x+h}{c}\Big)\, dx \tag{9}$$

or, changing the variable of integration to $\xi = x + h$,

$$\psi^- = -\int_{|n+h|+h}^{\infty} 2\pi\, M(t - \xi/c)\, d\xi\,. \tag{10}$$

Adding (8) and (10), and renaming the bound variable of integration ("dummy variable") in the latter, we obtain the wave function due to both sheets together:

$$\psi = \psi^+ + \psi^- = \int_{|n|}^{\infty} 2\pi\, M(t - x/c)\, dx - \int_{|n+h|+h}^{\infty} 2\pi\, M(t - x/c)\, dx\,. \tag{11}$$

Here I should acknowledge a technicality that I have glossed over. In the second integral, the surface of integration has moved to the left in Fig. 1, so that a given x (in the argument of M) no longer corresponds to the same ρ. But this does not matter, because (i) M is assumed uniform, and (ii) even if it were not, the change in ρ will vanish as $h \to 0$. So we can indeed treat M as the same function in both integrals, so that the integrands are the same, so that the difference between the integrals is simply the integral over the difference in the ranges of integration:

$$\psi = \int_{|n|}^{|n+h|+h} 2\pi\, M(t - x/c)\, dx\,. \tag{12}$$

On the $\mathcal{R}'$ side of both sheets (the left side in Fig. 1), both n and $n+h$ are negative, so that $|n| = -n$ and $|n+h| = -n-h$, whence both limits of integration simplify to $-n$, so the "range" of integration collapses

to nothing; thus, *on the $\mathcal{R}'$ side of both sheets, the wave function is zero*—as hoped. Meanwhile, on the $\mathcal{R}$ side of both sheets (the right side in Fig. 1), both n and $n+h$ are positive, so that $|n| = n$ and $|n+h| = n+h$; hence:

$$\text{In } \mathcal{R}, \quad \psi = \int_n^{n+2h} 2\pi M(t - x/c)\, dx\,. \tag{13}$$

In (13), the range of integration is limited to small values of x, suggesting that the uncertainty in the integrand for larger x is moot. But the same equation tells us that, in a certain direction from each point on $\mathcal{S}$, the contributions from the two sheets do *not* cancel; and this effect could extend to the observation point N from remote points on $\mathcal{S}$. We can ignore this effect on the understanding that we are actually working towards a *naive* wave function, hence *naive* boundary conditions.[6]

Under the standing assumption that the strength M is continuous, there exists a value ν, within the range of integration in (13), such that the integral is equal to the width of the range multiplied by the integrand evaluated at ν; that is:

$$\text{In } \mathcal{R}, \quad \psi = 2h \cdot 2\pi M(t - \nu/c)\,, \quad \text{where } n < \nu < n + 2h\,. \tag{14}$$

To see what happens when the strength density of each sheet is inversely proportional to h, let us write

$$M(t) = \frac{D(t)}{h}\,, \tag{15}$$

so that (14) becomes

$$\text{In } \mathcal{R}, \quad \psi = 4\pi D(t - \nu/c)\,, \quad \text{where } n < \nu < n + 2h \tag{16}$$

or, in the limit as $h \to 0$ (hence $\nu \to n$):

$$\text{In } \mathcal{R}, \quad \psi = 4\pi D(t - n/c)\,. \tag{17}$$

In summary, for small $|n|$, the wave function due to both sheets is

$$\psi = \begin{cases} 4\pi D(t - n/c) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}'. \end{cases} \tag{18}$$

In the limit as $n \to 0$ from the $\mathcal{R}$ side, this gives

$$D(t) = \frac{\psi}{4\pi}\,, \tag{19}$$

where ψ is to be evaluated at the $\mathcal{R}$ side of $\mathcal{S}$—that is, as a (naive) boundary condition. And in the limit as $n \to 0$ from the $\mathcal{R}'$ side, equation (18) gives a null BC on ψ. Thus we have identified sources on $\mathcal{S}$ for which (at least) the naive BCs on ψ are the original at the $\mathcal{R}$ side and null at the $\mathcal{R}'$ side, so that, by Propositions 3 and 4, we get the original wave function throughout $\mathcal{R}$ and a null wave function throughout $\mathcal{R}'$. Hence we also get the original BC on ψ_n at the $\mathcal{R}$ side. For future reference, however, we differentiate (18) w.r.t. n on the $\mathcal{R}$ side, obtaining

$$\psi_n = -\frac{4\pi}{c} D'(t - n/c)\,, \tag{20}$$

whence, in the limit as $n \to 0$,

$$D'(t) = -\frac{c\psi_n}{4\pi}\,. \tag{21}$$

The quantity $D(t)$, given in terms of the required BC by equation (19), is the *strength density of spatiotemporal dipole sources facing $\mathcal{R}$*. In general, a **spatiotemporal dipole** (**STD**) of strength $f(t)$,

[6] This essential qualification was overlooked in Version 1 of the paper. Other flaws in Version 1 are acknowledged in the footnotes of Version 2. See Appendix 8.4 for a brief revision history.

in the direction of the coordinate n, consists of a monopole source of strength $f(t)/h$, and another monopole source of strength $-f(t-h/c)/h$, displaced from the first by a distance h in the $-n$ direction, where $h \to 0$. If the strength *density* of such sources on a surface $\mathcal{S}$ is $D(t)$, then each element $d\mathcal{S}$ carries monopoles of strengths $D(t)\,d\mathcal{S}/h$ and $-D(t-h/c)\,d\mathcal{S}/h$, separated by the distance h. Hence the monopole strengths *per unit area of* $\mathcal{S}$ are $D(t)/h$ and $-D(t-h/c)/h$, as above.

A spatiotemporal dipole is not to be confused with an *ordinary* or *spatial* dipole, also called a *doublet*,[7] in which the inverted monopole is *not* time-shifted relative to the uninverted one.

A dipole source (spatiotemporal or not) comprises two opposing infinite monopole sources separated by an infinitesimal distance. We must admit that this concept strains credulity. But we are not claiming that the sources specified by (19) actually exist, or even that they *could* exist. We have only shown that the wave function in $\mathcal{R}$ is *as if* it had been generated by the specified distribution of sources, and that the same distribution would give a null wave function in $\mathcal{R}'$.

2.3 Note on electromagnetic waves

Recall that we defined a *monopole* source of strength $f(t)$ as a source that generates the wave function

$$\frac{1}{r}\,f(t-r/c)\,, \tag{22}$$

where r is the distance from the source, *regardless of direction*. Obviously this definition is compatible with a *scalar* function f. But if f is a *vector*, this definition requires not only the magnitude of the wave function, but also its direction, to be independent of the direction of propagation. That might seem to exclude electromagnetic waves, for which the electric and magnetic fields are transverse to the direction of propagation and therefore not independent of it. However, it is possible to describe electromagnetic waves in terms of two other wave functions known as the *electric scalar potential*, denoted by φ, and the *magnetic vector potential*, denoted by $\mathbf{A}$. For a volume element dV carrying a scalar charge density $\varrho(t)$ and a vector current density $\mathbf{J}(t)$, the contributions to φ and $\mathbf{A}$ are, respectively,[8]

$$d\varphi = \frac{1}{4\pi\epsilon_0 r}\,\varrho(t-r/c)\,dV \tag{23}$$

and

$$d\mathbf{A} = \frac{1}{4\pi\epsilon_0 c^2 r}\,\mathbf{J}(t-r/c)\,dV\,, \tag{24}$$

where ϵ_0 is a physical constant (called the electric permittivity of a vacuum), and r is the distance from dV, *regardless of direction*. Of course, elemental sources of these kinds cannot be arranged arbitrarily, because charge must be conserved as it moves under the influence of the fields. But, subject to that constraint, with appropriate definitions of f, the elemental sources' contributions to the wave functions do indeed have the form of (22).

3 Helmholtz formula; Kirchhoff integral theorem

3.1 Derivation

Having found the secondary sources that match the BCs in $\mathcal{R}$, we must now express the wave function in $\mathcal{R}$ *in terms of the sources*, and thence in terms of the BCs. The first step is to find the infinitesimal contribution to the wave function at a general field point P in $\mathcal{R}$, due to the secondary sources on a general surface element $d\mathcal{S}$. We shall call this contribution $d\psi(P,t)$, specifying the place P as an argument. *From here on, if the place of evaluation of ψ or any of its derivatives is not specified, we take it to be at $d\mathcal{S}$, on the side facing $\mathcal{R}$.* Equations (19) and (21) are already consistent with that convention.

[7] The term *doublet*, which seems to be older, is used by Baker & Copson [1], Born & Wolf [2, p. 421], and Larmor [9].

[8] *Cf.* Feynman *et al.* [5], vol. 2, Chapter 15, Table 15-1. Here we use a dotless $\mathbf{J}$ to avoid any suggestion of differentiation.

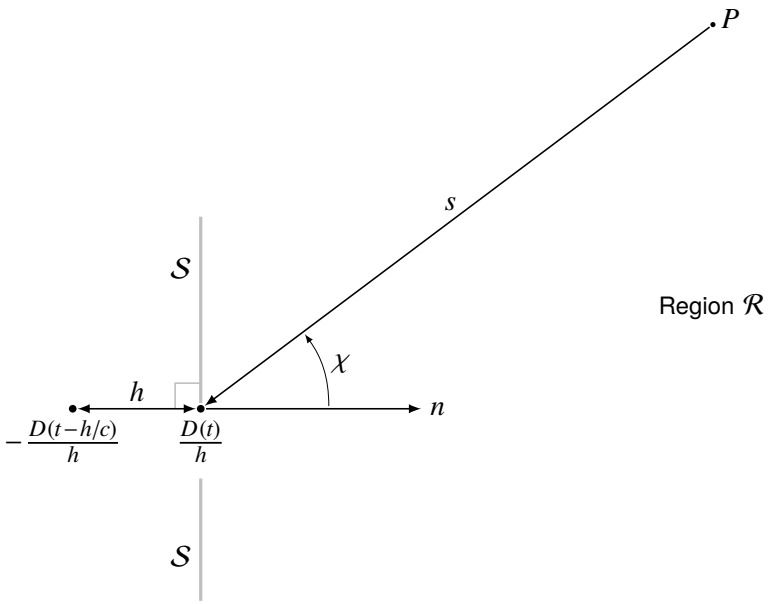

Fig. 2: Paired monopole sources per unit area of a spatiotemporal-dipole source distribution with strength density $D(t)$ facing $\mathcal{R}$. Gray lines show the orientation of the surface $\mathcal{S}$, which may be curved. The separation h is infinitesimal (so the diagram is not to scale). The normal coordinate n increases by h as we move from the inverted monopole (left) to the uninverted one.

For the STD strength density $D(t)$, equivalent monopoles *per unit area of* $\mathcal{S}$ are as shown in Fig. 2. The contribution to the wave function at P from the uninverted monopole (on the right) is

$$\frac{D(t-s/c)}{hs}\,. \tag{25}$$

The contribution from both monopoles together is the *change* in the above expression due to n increasing by h, *and* t increasing by h/c. As h is infinitesimal, that change is

$$h\frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{hs}\right) + \frac{h}{c}\frac{\partial}{\partial t}\left(\frac{D(t-s/c)}{hs}\right) \tag{26}$$

$$= \frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{s}\right) + \frac{1}{cs}D'(t-s/c)\,. \tag{27}$$

This is per unit area of $\mathcal{S}$; to find the contribution from the elemental area $d\mathcal{S}$, we simply multiply by $d\mathcal{S}$, obtaining[9]

$$d\psi(P,t) = \left\{\frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{s}\right) + \frac{1}{cs}D'(t-s/c)\right\}d\mathcal{S} \tag{28}$$

or, upon substitution from (19) and (21),

$$d\psi(P,t) = \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\psi_n(t-s/c)\right\}d\mathcal{S}\,. \tag{29}$$

The total wave function at P is the sum over the surface $\mathcal{S}$—in other words, the *surface integral over* $\mathcal{S}$—of all the contributions $d\psi(P,t)$ from the elements $d\mathcal{S}$, and is written

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\psi_n(t-s/c)\right\}d\mathcal{S}\,, \tag{30}$$

[9] Continuing from footnote 5: If the factor 4π is included in the denominator of equation (2), it influences subsequent equations and eventually cancels out in (16) to (21). But then it is needed again in (25) to (28), with the result that (29), and therefore (30) and its corollaries, are unchanged! Equations (23) and (24) hint at why one might include such a factor in the wave function due to a source with "unit strength". For better or worse, we have not used that convention here.

where the double integral sign acknowledges that the range of integration is two-dimensional. Equation (30) is the **Helmholtz formula** in its most general form.[10] In the integrand, ψ and ψ_n are understood to be evaluated at the element $d\mathcal{S}$, and their arguments indicate that they are delayed by the propagation time for the distance s from the element $d\mathcal{S}$ to the field point P. For historical reasons, the arrow for s in Fig. 2 is drawn in the opposite direction, so that s can be construed as a coordinate of a general point on $\mathcal{S}$ (like r in Fig. 3, below). This convention does not affect the *sign* of s, because s is an absolute distance and therefore positive; but it affects what we call the positive *direction* of s (as we shall see later, when we define angles between "positive directions" of coordinates).

If we want the integrand in (30) to be in terms of BCs on ψ, we must eliminate derivatives of expressions other than ψ. In the first term in the braces, the fraction to be differentiated depends on n, *not* through the normal derivative ψ_n (since ψ is evaluated at $n=0$), but through s (which decreases as n increases by h in Fig. 2) and therefore through the s-dependent delay in $\psi(t)$. So the first term is

$$\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) = \psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{1}{s}\,\frac{\partial}{\partial n}\psi(t-s/c) \tag{31}$$

$$= \psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'(t-s/c)\,\frac{\partial s}{\partial n}\,, \tag{32}$$

where the first equality follows from the product rule, and the second from two applications of the chain rule. Substituting (32) into the Helmholtz formula (30) yields

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi}\left\{\psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'(t-s/c)\,\frac{\partial s}{\partial n} - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,. \tag{33}$$

This result, known as the **Kirchhoff integral theorem**, gives the wave function at any point in the region $\mathcal{R}$ due to sources outside $\mathcal{R}$, in terms of the wave function and its derivatives at the $\mathcal{R}$ side of the boundary surface $\mathcal{S}$. It is more commonly written as

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\left[\frac{\partial\psi}{\partial t}\right]\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{34}$$

where *square brackets indicate that the enclosed function is to be delayed by the propagation time from the surface element $d\mathcal{S}$ to the field point.*[11] If we apply the chain rule to the first term in the braces, take the factor $1/s$ outside, and use an overdot to denote (partial) differentiation w.r.t. t, we obtain the alternative form

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi s}\left\{-\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot{\psi}]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{35}$$

which we shall find convenient in Section 4.

3.2 Application to diffraction by an aperture

So we arrive at a catch-22. We established very early (Proposition 2) that the wave function throughout $\mathcal{R}$ is determined by *either* the wave function *or* its normal derivative at $\mathcal{S}$, so that each BC is determined by the other through the complete wave function. But, in order to calculate that wave function using (33) or (34) or (35), we need to know *both* the wave function ψ *and* its normal derivative ψ_n at $\mathcal{S}$. In a non-trivial problem, how can we deduce one BC from the other without first knowing the very wave function that we are trying to find? For that matter, how can we know even *one* BC without first knowing the whole wave function?

[10] The form given here allows arbitrary time-dependence of the wave function. The form usually found in textbooks, which dates from 1859, assumes sinusoidal time-dependence.

[11] *Cf.* Born & Wolf [2] at pp. 420–21 (especially eq. 13). *Cf.* also Baker & Copson [1, p. 37] and Miller [12, eq. 2], who use r instead of s (among other notational differences). Baker & Copson, in a later example [1, p. 40, last eq.], give a different sign because the normal coordinate (which they call ν in this case) is measured in the other direction.

In the case of diffraction by an aperture in an opaque screen, we have already proposed to choose $\mathcal{S}$ as the union of two segments, namely $\mathcal{S}_a$ spanning the aperture, and $\mathcal{S}_b$ on the back of the screen. For want of a better option, we might now proceed as follows. First, knowing the primary sources, we might calculate the wave function *as if the screen were not there*, and thence calculate the corresponding BCs at $\mathcal{S}$, and the secondary sources on $\mathcal{S}$ that give those BCs. Then we might *assume* that the screen simply eliminates the secondary sources on $\mathcal{S}_b$. In terms of superposing the contributions of the sources by means of a surface integral, this assumption means integrating over $\mathcal{S}_a$ only, instead of the whole surface $\mathcal{S}$.

Limiting the range of integration is indeed the conventional approach to the problem. But, for historical reasons, the conventional *explanation* of it is very different from the one just given. Here we have derived Kirchhoff's integral by means of secondary sources. But Kirchhoff himself, in 1882, arrived at his integral theorem by a different path, which obliged him to deal directly with boundary conditions. So, for want of a better option, he supposed that the wave function *and* its normal derivative on $\mathcal{S}_b$ were zero, while the wave function *and* its normal derivative on $\mathcal{S}_a$ were as if the screen were absent. His approach and ours lead to the same approximation: that the integrand is as if the screen were not there, while the integral is over the aperture only. And the approximation turns out to be remarkably accurate—indeed, usually far *more* accurate than the boundary conditions on which it is apparently based [16], because the errors in the BCs tend to "average out" in the integration.

3.3 Exact? Consistent? Well-posed?

In the title of this paper, the word "exact" applies to the expression of the wave function in terms of the boundary conditions or secondary sources, but does not generally apply to the BCs or secondary sources themselves. That distinction was significant in our first mention of the diffraction problem, in which the assumed BCs were only a "first approximation" obtained by treating the propagation paths from the source to $\mathcal{S}$ as completely obstructed or completely unobstructed. Kirchhoff took that approximation further, applying it to the BCs on both the wave function and its normal derivative. And we have reproduced Kirchhoff's result by making the heroic assumption that the screen simply eliminates part of the sheet of secondary sources, without changing the secondary sources that remain.

By assuming that the integrand in the aperture (interpreted in terms of boundary conditions or secondary sources) is the same as if the screen were absent, we neglect any reflection or scattering into the aperture from the edge or back of the screen, so that the back of the screen must be absorbent. For that reason, in optical examples, the screen is typically described as not merely opaque, but *black*.

As the BC on *either* the wave function *or* its normal derivative at $\mathcal{S}$ determines the wave function in $\mathcal{R}$, either BC determines the other. Kirchhoff tried to set both at once. It would be an unbelievable fluke if his settings were consistent; and indeed it was proven by Poincaré, no later than 1892, that they are not, except in the trivial case of a null wave function.[12] Here I offer a proof that is somewhat less rigorous, but much simpler.

Obviously, if the BCs are to be consistent, the wave function predicted by inserting those BCs into the Kirchhoff integral must give back the same BCs. We have seen that we get the same wave function in $\mathcal{R}$ by retaining the secondary sources on $\mathcal{S}_a$ only. The Kirchhoff BCs then allege that non-zero sources on this part of $\mathcal{S}$ give null BCs everywhere on the rest of $\mathcal{S}$, for both the wave function and its normal derivative, for all time. That is implausible on its face, and is easily reduced to an absurdity if we allow one very "unassuming" assumption: that in the absence of the screen, there is at least one surface element $d\Sigma$ on the perimeter of $\mathcal{S}_a$ where the wave function is non-zero for at least one moment in time.[13] At that moment, the prescribed STDs on $d\Sigma$ are non-zero and therefore impose a non-zero step-discontinuity in the wave function at $d\Sigma$. Hence, as we approach $d\Sigma$ along $\mathcal{S}_b$ (on which there are no sources), the normal derivative tends to infinity—whereas Kirchhoff's assumed BCs say it is zero.

[12] Poincaré's proof is for sinusoidal time-dependence. It has been explained in English by Buchwald & Yeang [4, pp. 485–90] and, more tersely, by Baker & Copson [1, pp. 71–2].

[13] This is certainly the case if the wave function is due to a single monopole source.

There has been much ado about this inconsistency [4, 11, 16]. In Kirchhoff's defense, it is perhaps enough to say that (i) he could hardly have done anything else at the time, (ii) he offered a theory that stood up better than its predecessors under the test of experiment, and (iii) this is supposed to be the definition of progress in physics. It is definitely *not* enough to say that his BCs were only meant to be approximate and therefore only needed to be approximately consistent, because in fact they are not even approximately consistent. We have just seen that as we approach the edge of the aperture ($\mathcal{S}_a$) along the back of the screen ($\mathcal{S}_b$), the predicted normal derivative is wildly divergent from the assumption on which it supposedly depends. The wave function predicted by the Kirchhoff integral (or an equivalent) in the aperture can also differ substantially from the assumed boundary condition [11, Fig. 3(b)]. Hence it is sometimes said that the Kirchhoff BCs give more accurate results than we have any right to expect [16]. However, if a boundary-surface integral of a function of the BCs is to be accurate, it is not necessary that the BCs themselves be accurate at every point; it suffices that the errors in the BCs "average out", nearly enough, in the composition of the integrand and the subsequent integration (*cf.* [16], at s. 4.2). It is no cause for astonishment that the latter condition is often satisfied without the former.

Moreover, we have seen that we get the same result as Kirchhoff by assuming that we retain the secondary sources over the aperture only. This assumption—whatever may be said about its physical accuracy—does not involve any mathematical inconsistency.

A problem is said to be **well-posed** if a solution exists, is unique, and is a continuous function of the data; otherwise the problem is said to be **ill-posed**.[14] The problem of satisfying the Kirchhoff BCs is ill-posed in the sense that no solution exists, because the BCs, being inconsistent, cannot be simultaneously satisfied. But, as we have seen, the Kirchhoff integral over the aperture (with the unobstructed boundary conditions) is the solution of another problem, which is well-posed.

4 Point-source: Kirchhoff diffraction formula

In this paper, square brackets enclosing a function of time tell us to subtract s/c from t in the argument of the function. It makes no difference whether we do this before or after we multiply that function by a time-independent factor, because the latter factor has no argument that can be affected by the subtraction. Consequently, *time-independent factors may be taken inside or outside the square brackets*.

Now suppose that the wave function given by equations (33) to (35) has only *one* primary source, namely a single monopole ("point-source") at the point O outside $\mathcal{R}$. Let the coordinate r measure the distance from O to the general surface element $d\mathcal{S}$ at point Q (Fig. 3). Treating n as an independent coordinate of Q, we have by the chain rule

$$\frac{\partial\psi}{\partial n} = \frac{\partial\psi}{\partial r}\,\frac{\partial r}{\partial n}\,, \tag{36}$$

where the derivatives w.r.t. n are evaluated at $n = 0$ (so that Q is indeed on $\mathcal{S}$, as shown). Because the second factor on the right is time-independent, it can be taken outside the square brackets in (35), so that the expression in braces becomes

$$\left\{-\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot{\psi}]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial r}\right]\frac{\partial r}{\partial n}\right\}. \tag{37}$$

Let ϕ be the angle between the positive directions of n and r, and χ the angle between the positive directions of n and $-s$, the last direction being from Q to P in Fig. 3. Then, from the geometry,

$$\frac{\partial r}{\partial n} = \cos\phi \tag{38}$$

and

$$\frac{\partial s}{\partial n} = -\cos\chi\,, \tag{39}$$

[14] Obviously the existence and uniqueness of a solution may depend on what we call a "solution". For example, if an equation is satisfied by several values of a variable, there is not a unique *value* that satisfies it, but there is a unique *set* of values that satisfy it. And if there is *no* value that satisfies it, there is still a unique *set* of values that satisfy it—namely the empty set!

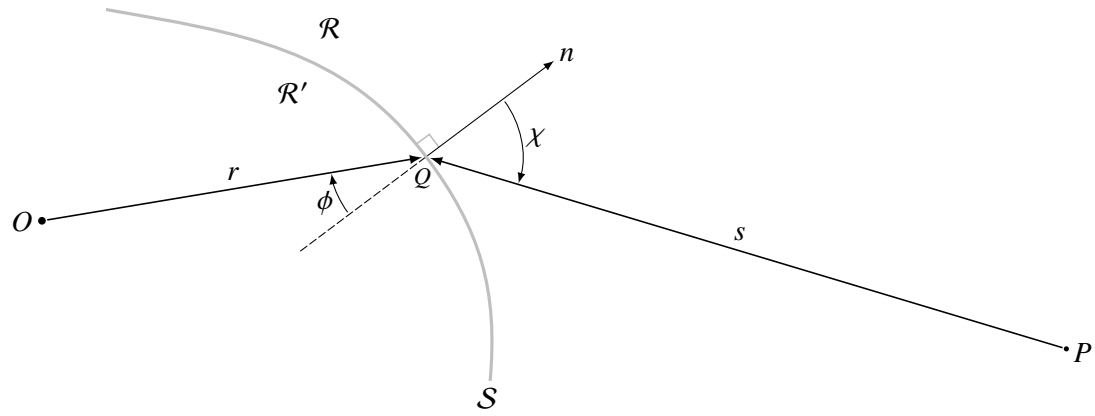

Fig. 3: Distances and angles pertaining to the calculation of the wave function at the field point P in region $\mathcal{R}$ due to a source at O in region $\mathcal{R}'$, by integration over the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$. The coordinate n is measured normal to $\mathcal{S}$, into $\mathcal{R}$. The angles ϕ and χ are not necessarily coplanar.

where again the derivatives are evaluated at $n = 0$. As the primary source is a single monopole, the wave function on the surface $\mathcal{S}$ has the form (2):

$$\psi = \frac{1}{r} f(t - r/c) \,. \tag{40}$$

By evaluating and comparing the partial derivatives of this w.r.t. r and t, we readily obtain

$$\frac{\partial \psi}{\partial r} = -\frac{\dot{\psi}}{c} - \frac{\psi}{r} \,. \tag{41}$$

Substituting (38), (39), and (41) into (37), we get

$$\left\{ \frac{1}{s}[\psi]\cos\chi + \frac{1}{c}[\dot{\psi}]\cos\chi + \left[\frac{\dot{\psi}}{c} + \frac{\psi}{r}\right]\cos\phi \right\} . \tag{42}$$

Because the delayed sum is the sum of the terms delayed separately, and because a time-independent factor can be taken outside the delay operation, (42) can be written

$$\left\{ \frac{1}{s}[\psi]\cos\chi + \frac{1}{c}[\dot{\psi}]\cos\chi + \left(\frac{1}{c}[\dot{\psi}] + \frac{1}{r}[\psi]\right)\cos\phi \right\} \tag{43}$$

or, regrouping the terms,

$$\left\{ \frac{\cos\phi + \cos\chi}{c}[\dot{\psi}] + \left(\frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right)[\psi] \right\} . \tag{44}$$

Putting this expression back in (35), we obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{\cos\phi + \cos\chi}{c}[\dot{\psi}] + \left(\frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right)[\psi] \right\} d\mathcal{S} \,. \tag{45}$$

Equation (45) is the exact form of the **Kirchhoff diffraction formula** (also known as the *Huygens-Kirchhoff* diffraction formula, because it quantifies Huygens' principle for a monopole primary source; or the *Fresnel-Kirchhoff* diffraction formula, because it generalizes Fresnel's treatment of the subject). We call it the *exact* form in order to distinguish it from a more common approximate form: if the distances from the primary source to $\mathcal{S}$ and from $\mathcal{S}$ to P are sufficiently large (as is typically the case), then ψ/r and ψ/s are very small compared with $\dot{\psi}/c$, so that we can neglect the second term in the braces, retaining only the term in $[\dot{\psi}]$. That term contains the factor $\cos\phi + \cos\chi$, in which the angles ϕ and χ are relative to the normal to $\mathcal{S}$; we might therefore describe this factor (or something proportional to it) as the "obliquity factor".

5 Integration over a primary wavefront

5.1 Spherical wavefront

Recall that ϕ is the angle between the positive directions of n and r. This is the angle between the normal to $\mathcal{S}$ and the radius of the (spherical) primary wavefront, hence the angle between the normal to $\mathcal{S}$ and the *normal* to the primary wavefront. So if $\mathcal{S}$ *is* a primary wavefront, we simply put $\phi = 0$ in (45), obtaining

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} [\dot{\psi}] + \left(\frac{1}{r} + \frac{\cos\chi}{s} \right) [\psi] \right\} d\mathcal{S} \,. \tag{46}$$

Equation (46) is the exact form of the **Huygens-Fresnel-Kirchhoff diffraction formula**; it gives a precise mathematical form to the following monumental statement by Augustin Fresnel, which we would now call the **Huygens-Fresnel principle**:

> *The vibrations at each point in the wave-front may be considered as the sum of the elementary motions which at any one instant are sent to that point from all parts of this same wave in any one of its previous positions...* [6, p. 108].

In (46), for sufficiently large r and s, we may again neglect the second term in the braces and retain only the term in $[\dot{\psi}]$, in which the obliquity factor has been reduced to a function of the single obliquity angle χ. We now see that if we define the **obliquity factor** in (46) as

$$\tfrac{1}{2}(1+\cos\chi)\,, \tag{47}$$

then it will have a maximum of 1 for $\chi = 0$ (direct forward secondary waves), and a minimum of 0 for $\chi = 180°$ (backward secondary waves). Hence, in (45) and (46), it might be preferable to define the obliquity factor as *half* the numerator of the first fraction in the braces.

5.2 Plane wavefront: Approximation for non-spherical wavefront

The curvature of the primary wavefront enters into the derivation of (45) via the radius r. Now suppose that the primary waves are **plane waves**. Then for integration over a general surface, we simply put $r \to \infty$ in (45) and obtain

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{\cos\phi + \cos\chi}{c} [\dot{\psi}] + \frac{\cos\chi}{s} [\psi] \right\} d\mathcal{S} \,. \tag{48}$$

And for integration over a primary wavefront, we put $r \to \infty$ in (46), *or* $\phi = 0$ in (48), and obtain

$$\psi(P,t) = \iint\limits_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} [\dot{\psi}] + \frac{\cos\chi}{s} [\psi] \right\} d\mathcal{S} \,. \tag{49}$$

These two results are good approximations for large r, whether s is large or not; in other words, they are valid in the **far field** of the primary source, for both the far field and the **near field** of the surface of integration.[15]

As (48) and (49) are exact for plane waves and good approximations for large r, we can reasonably expect them to be good approximations for primary wavefronts with *non-spherical* curvature, provided

[15] The term *far field* also has a stronger meaning, namely that the curvatures of the constant-r and constant-s surfaces can be neglected when calculating differences in path lengths. Diffraction under those conditions is called *Fraunhofer diffraction*. We are *not* using that meaning here.

that the curvature is sufficiently gradual. (In a homogeneous, isotropic medium, a non-spherical wavefront typically comes from an initially plane or spherical wavefront that has been reflected or refracted at the interface with a different medium.) In such cases, ϕ is to be understood as the angle between the normals of the primary wavefront and the surface of integration.

Similarly, although the Kirchhoff formula (45) and the Huygens-Fresnel-Kirchhoff formula (46) have been derived for monopole sources, which by definition are omnidirectional, we can expect both formulae to be good approximations for directional sources provided that the sources give recognizable spherical wavefronts with a sufficiently gradual variation of intensity over each wavefront. And, as above, we can expect the approximations to hold even if the wavefront also has non-spherical curvature, provided that the curvature is sufficiently gradual.

6 The sinusoidal (monochromatic) case

6.1 Exact forms

If the function f in equation (2) has the **sinusoidal** form[16]

$$f(t) = A\cos(\omega t + \theta) \tag{50}$$

where A (the *peak amplitude*), ω (the *angular frequency* or *radian frequency*), and θ (the *phase angle*) are constants, then the wave function due to the monopole source in (2) becomes[17]

$$\frac{A}{r}\cos\left(\omega(t - r/c) + \theta\right), \tag{51}$$

which is usually written

$$\frac{A}{r}\cos\left(\omega t - kr + \theta\right), \tag{52}$$

where

$$k = \omega/c\,. \tag{53}$$

We call k the *wavenumber* (or, more precisely, the *angular* or *radian* wavenumber). In (52), we see that the argument of the cosine function changes by 2π if t changes by $2\pi/\omega$ or r changes by $(-)2\pi/k$. Thus the *period* of the undulation is

$$T = 2\pi/\omega\,, \tag{54}$$

and the *wavelength* is

$$\lambda = 2\pi/k\,. \tag{55}$$

The *linear* frequency or *cycle* frequency, usually called simply the frequency, often represented by the Greek ν (not to be confused with the English v), is the reciprocal of the period:

$$\nu = \frac{\omega}{2\pi}\,. \tag{56}$$

We shall have occasion to mention the *angular* or *radian* wavelength, also called the *reduced* wavelength, given by

$$\lambda\!\!\!^{-} = \frac{\lambda}{2\pi} = 1/k\,. \tag{57}$$

For a given location, hence a fixed r, the wave function (52) has the same form as (50), but with different values of the peak amplitude and phase angle. The contributions to the wave function at that point due to any other sinusoidal monopole sources with the same ω are also of that form, so that their

[16] A cosine function is sinusoidal, since $\cos\xi = \sin(\xi + \pi/2)$.

[17] If this form applies for all t, it is obviously inconsistent with the assumption that the wave function had a beginning. Perhaps the simplest workaround is to suppose that there has been enough time, since all the sinusoidal sources started up, for the waves to fill the region of interest, and for any start-up effects to leave that region or fade away. The essential feature of any workaround is that "equation" (60) be true, or at least sufficiently accurate.

total contribution to the wave function at that point is a sum of functions of form (50), with the same ω (but generally not the same constants A and θ). A pattern of sinusoidal waves of the same ω (hence the same frequency) is described as **monochromatic**, meaning "of one color", because that is the implication in the case of light waves.

Now it is readily seen that *the sum of any number of functions of form* (50)*, with the same frequency, is a function of the same form with the same frequency.*[18] Indeed, in the Cartesian xy plane, function (50) is the x component of a vector of length A making an angle $\omega t + \theta$ with the x axis, hence the x component of a vector of length A rotating at angular frequency ω from an initial angle θ. The sum of any number of functions of that form, with that frequency, is the sum of the x-components of the associated vectors, which is the x component of the sum of the vectors; and that vector sum is itself a vector rotating at the same frequency, so its x component is a function of the same form and frequency.

It is therefore useful to consider a wave function whose value at the $\mathcal{R}$ side of the surface element $d\mathcal{S}$ is of the form

$$\psi = A\cos(\omega t + \theta)\,, \tag{58}$$

where A and θ are independent of time but vary from point to point. From (58),

$$\frac{\partial\psi}{\partial t} = -\omega A\sin(\omega t + \theta) = \omega A\cos\left(\omega t + \theta + \pi/2\right)\,. \tag{59}$$

So, for a function of form (58), the operator $\partial/\partial t$ (or an overdot) is equivalent to multiplication by ω combined with a phase advance of a quarter-cycle ($\pi/2$ radians; $90°$). This can be written

$$\frac{\partial}{\partial t} = j\omega\,, \tag{60}$$

where ω (as usual) is a simple multiplicative factor, and j *is an operator that produces a quarter-cycle phase advance*. The "equality" in this result represents an equivalence between operators.

A double application of the j operator, obviously represented by the operator j^2, produces a $180°$ phase advance, which is equivalent to a change of sign. So the j operator has the interesting property that $j^2 = -1$. This suggests that there is a more sophisticated way of understanding j. For the purposes of this paper, however, it suffices to think of j as an operator!

The j notation is widely used by electrical engineers. Physicists, in contrast, tend to use the symbol i to denote a phase *lag* of a quarter-cycle.[19] Hence a physicist, in order to represent a quarter-cycle *advance*, would usually write $-i$, because the combination of a quarter-cycle lag and a sign-change is equivalent to a quarter-cycle advance. In this paper we use the electrical engineering convention.[20] Our results may be converted to the physics convention by simply substituting $-i$ for j.

In rule (60), the order of j and ω may be switched. In general, for a function of form (58), the j operator adds $\pi/2$ to the argument of the cosine function, or adds $\pi/2$ to the term ωt, whereas multiplication by a time-independent factor involves no such argument or term; so the order of the two operations makes no difference. In this respect, the operator j behaves like any other time-independent factor. Hence we can divide (60) through by c and apply (53), obtaining the operational equivalence

$$\frac{1}{c}\frac{\partial}{\partial t} = jk\,. \tag{61}$$

Now recall that another property of time-independent factors, such as k, is that they can be taken inside or outside the square brackets. Similarly, for a function of form (58), the j operator adds $\pi/2$ to the argument of the cosine function, whereas the square brackets subtract s/c from t within that

[18] This was first shown by Fresnel, by an analytical method, in a "supplement" dated January 1818 [7, vol. 1, at pp. 489–92]; see [14] for context. The time t was measured in periods. He repeated the demonstration in his prize memoir on diffraction [6, at pp. 103–5], with an influential change of notation: in the former document, the wavelength was called d; in the latter, it was called λ.

[19] Exceptions to both rules are Kottler [8] and Miller [12], who use i to denote a quarter-cycle *advance*.

[20] In my last paper [13], for better or worse, I used the physics convention.

argument, and the order of these operations makes no difference. Consequently, *the j operator, like a time-independent factor, may be taken inside or outside the square brackets*. This rule will now prove especially useful in combination with (61).

Applying rule (61) to equation (45)—or, more precisely, taking $1/c$ inside the square brackets, applying (61), then taking jk outside the square brackets, and finally taking $[\psi]$ outside the braces, on the understanding that j still operates on $[\psi]$—we obtain the exact monochromatic form of the Kirchhoff diffraction formula, which concerns spherical primary wavefronts and an arbitrary surface of integration:

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(\cos\phi + \cos\chi\right) + \frac{\cos\phi}{r} + \frac{\cos\chi}{s} \right\} d\mathcal{S}\,. \tag{62}$$

Similarly applying (61) in (46), *or* putting $\phi = 0$ in (62), we get the exact monochromatic form of the Huygens-Fresnel-Kirchhoff formula, in which the integration is over a spherical primary wavefront:[21]

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(1 + \cos\chi\right) + \frac{1}{r} + \frac{\cos\chi}{s} \right\} d\mathcal{S}\,. \tag{63}$$

For *plane* primary waves, we apply (61) in (48) and (49), or let r become infinite in (62) and (63), obtaining

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(\cos\phi + \cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{64}$$

for integration over a general surface, and

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(1 + \cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{65}$$

for integration over a primary wavefront. Of course (65) is also obtainable from (64) by setting $\phi = 0$.

6.2 Approximate forms

We have noted that in equations (45) and (46), if the distances are sufficiently large, we can neglect the term in $[\psi]$ and retain only the term in $[\dot{\psi}]$. In the sinusoidal case, the term in $[\dot{\psi}]$ becomes the term in k, allowing us to clarify the meaning of "sufficiently large". In (62) and (63), if $1/r$ and $1/s$ are very small compared with k—that is, *if r and s are very large compared with the radian wavelength*—then we can neglect the terms in r and s. It is also convenient to recall (55) and put $k = 2\pi/\lambda$. Equations (62) and (63) then reduce to

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{j}{\lambda} \frac{\cos\phi + \cos\chi}{2s} [\psi]\, d\mathcal{S} \tag{66}$$

if $\mathcal{S}$ is a general surface, and

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{j}{\lambda} \frac{1 + \cos\chi}{2s} [\psi]\, d\mathcal{S} \tag{67}$$

if $\mathcal{S}$ is a primary wavefront; and again the second result follows from the first by setting $\phi = 0$. Both results indicate that the secondary sources are advanced in phase by 90° relative to the primary wave function at $d\mathcal{S}$, that their strengths are inversely proportional to the wavelength, and that the amplitudes of the secondary waves decay with distance like $1/s$. The remaining factors in the integrand are the

[21] The corresponding result in Baker & Copson [1, top of p. 33] has a sign error in the last term in the braces (when the sign outside the integral is accounted for). The error, which is found in all editions, can be confirmed by working from their previous equation [bottom of their p. 32]. Larmor [10, p. 172, last eq.] and Miller [12, eq. 6] omit the second-last term in the braces (see Section 7 below), but agree with me on the sign of the last term.

primary wave function (delayed) and the obliquity factor. The $1/r$ dependence of the primary wave function can be made explicit by substituting expression (51) for ψ (*cf.* Appendix 8.2).

Because (66) and (67) assume that s is large, they apply only to the *far field* of the surface $\mathcal{S}$. In other respects they have the same applicability as (64) and (65)—that is, to the far field of the primary source, including primary wavefronts of arbitrary shape but sufficiently gradual curvature, and with sufficiently gradual variation of intensity over the wavefront. In the sinusoidal case, it is clear that a "sufficiently gradual" curvature is one for which the radii of curvature are very large compared with the radian wavelength $\lambda\!\!\!^{-}$, and we might guess that a "sufficiently gradual" variation of intensity is one that gives negligible variation over distances comparable with $\lambda\!\!\!^{-}$.

7 Notes on Miller (1991)

The integrand given by Professor Miller [12, eq. 6] is apparently meant to be "exact for uniform spherical or plane wave fronts" [12, before eq. 6]. But, when translated into the notation of the present paper,[22] his integrand agrees with the plane-wave formula (65) above—not with the more general spherical-wave formula (63) above. The reason is that he neglects the $1/r$ decay in the magnitude of the primary wave, with the result that his equation (4), which corresponds to our (41), lacks the second term on the right.[23]

As Miller notes, the $(\cos\chi)/s$ term in (63) is "near-field" in the sense that it is significant close to the surface (wavefront) of integration. He retains this term but omits the $1/r$ term, which is near-field in the sense that it becomes significant if the surface of integration is close to the primary source—so close that $1/r$ is not negligible compared with k (*cf.* Baker & Copson [1, p. 33], who include the $1/r$ term, which they call $1/r_0$). I concede, however, that the $1/r$ term might reasonably be considered less important, because the condition under which we cannot neglect that term is also a condition under which we cannot neglect variations in the intensity of the primary wavefront due to proximity to the edge of the aperture or (as in Miller's numerical example) directionality of the primary source.

The need for this correction term was revealed by a very different method in my last paper [13], the conclusion of which began:

> While I strongly suspect that Miller's spatiotemporal-dipole interpretation of diffraction can be reconciled with my near-source correction term, I leave the investigation of that question for another paper and probably another author. . .

Indeed it can be reconciled, because Miller's STDs yield formula (34), which in turn yields the $1/r$ term in (46) and (63). Moreover, equation (46) is for general time-dependence, confirming Miller's statement that the validity of spatiotemporal dipoles is not limited to the monochromatic case [12, endnote 12].

So this is the "other paper", although I was not confident that I would be its author. One reason for my doubt may be gleaned from my vague statement that the correction term relates to "the curvature of the surface of dipoles, which in turn implies a departure from a simple proportionality between the strength of the secondary source and the area of the surface element" [13, p. 8]. As it turns out, I have sidestepped that complication by treating the strength densities of monopole sources as per unit area "of $\mathcal{S}$", even if that does not exactly correspond to unit area of a parallel sheet at a distance h from $\mathcal{S}$.

8 Appendices

8.1 Diffraction as a saltus problem

The conventional method of modeling diffraction by an aperture in a screen is to apply Kirchhoff's integral theorem, evaluating the integrand as if the screen were not there, and integrating over the aperture only. Kirchhoff himself arrived at this method via inconsistent boundary conditions. We have

[22] Miller's ϕ is my ψ; his r is my s; his i (not $-i$) is my j; and his θ is my χ.

[23] Larmor had made the same approximation: in [10], on p. 172, the second-last equation is equivalent to Miller's (4), with the result that the last equation agrees with my (49), not my (46).

seen that the inconsistency can be circumvented by choosing a surface $\mathcal{S}$ that covers the aperture and the back of the screen, evaluating the secondary sources on $\mathcal{S}$ as if the screen were absent, and assuming that the screen simply eliminates the secondary sources on the corresponding segment of $\mathcal{S}$ (which we have called $\mathcal{S}_b$), leaving the sources on the segment spanning the aperture (which we have called $\mathcal{S}_a$).

As we cross a sheet of sources $\mathcal{S}$ at any point, the *step-change* in the wave function ψ or its normal derivative ψ_n is due solely to the sources on an infinitesimal surface element $d\mathcal{S}$ *at that point*. It is *not* due to any sources at finite (i.e., non-infinitesimal) distances from that point, because those distances change by only infinitesimal fractions as we cross $\mathcal{S}$ (*cf.* Larmor [9], p. 6). The imposition of such a step-change is called a **saltus condition** (from the Latin noun *saltus*, pronounced "SAHL-toos", meaning "jump" or "leap"), and a problem with given saltus conditions is called a **saltus problem**—whereas a problem with given *boundary* conditions is called a *boundary-value problem*.

Obviously the step-change in ψ or ψ_n as we cross $\mathcal{S}$ is given by the difference between the boundary conditions (on ψ or ψ_n, respectively) at the two sides. But the BCs may be naive, because any influence from remote sources will be the same on the two sides of the surface and will cancel in the subtraction. As the BCs at the two sides can be influenced by remote sources, while the difference between them cannot, the saltus conditions imposed by a sheet of sources may be considered more fundamental than the boundary conditions. In particular, if we have a distribution of sources on a surface $\mathcal{S}$, comprising segments $\mathcal{S}_a$ and $\mathcal{S}_b$, the *saltus* conditions imposed by the sources on $\mathcal{S}_a$ remain the same if the sources on $\mathcal{S}_b$ are taken away, although the *boundary* conditions generally change.

So, when we try to solve the above diffraction problem by retaining the secondary sources on $\mathcal{S}_a$ only, we are actually solving a saltus problem in which we retain the saltus conditions due to the sources on $\mathcal{S}_a$, with no saltus on $\mathcal{S}_b$. And the saltus conditions due to the sources on $\mathcal{S}_a$ are the same as if the sources on $\mathcal{S}_b$ were retained, as in Case 2 of Proposition 3, namely: the step-changes in ψ and ψ_n as we cross $\mathcal{S}$ to the $\mathcal{R}$ side are respectively equal to the original *values* of ψ and ψ_n on $\mathcal{S}$. Indeed, the same applies to the higher-order normal derivatives. However, because the sources in Case 2 are specified by *two* boundary conditions—one on each side—they are specified to the same degree by the saltus conditions on the wave function and its (first) normal derivative. In the case of STD sources, the two "conditions" are encoded in the strength density and the orientation of the STDs.

In summary, when we try to solve a diffraction problem by applying Kirchhoff's integral theorem, evaluating the integrand as if the screen were not there, and integrating over the surface segment $\mathcal{S}_a$ that spans the aperture, we are really solving a saltus problem in which the step-changes in the wave function and its normal derivative, as we cross $\mathcal{S}_a$ in the original direction of propagation, are respectively equal to the original *values* of the wave function and its normal derivative on $\mathcal{S}_a$ in the absence of the screen (with no step-changes elsewhere).

Ironically, the first published formulation of a diffraction problem as a saltus problem was not the above, but a slightly more complex formulation, which we shall now obtain by superposition. For brevity, let us represent Cases 1 and 2 of Proposition 3 as follows:

$$\text{Case 1:} \qquad \text{Sources in } \mathcal{R}' \longrightarrow \psi(P,t)\,;$$

$$\text{Case 2:} \qquad \text{Sources on } \mathcal{S}_a + \text{Sources on } \mathcal{S}_b \longrightarrow \begin{cases} \psi(P,t) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}'. \end{cases}$$

Obviously the sources in $\mathcal{R}'$ are "primary" while those on $\mathcal{S}_a$ and $\mathcal{S}_b$ are "secondary". Now let Case 3 be our proposed solution to the diffraction problem, with $\mathcal{S}_a$ spanning the aperture, $\mathcal{S}_b$ covering the back of the screen, and $\psi_{(a)}(P,t)$ being the wave function. Then, in the same notation, we have

$$\text{Case 3:} \qquad \text{Sources on } \mathcal{S}_a \longrightarrow \psi_{(a)}(P,t)\,.$$

In Case 3, the sources are still taken to be unobstructed; the presence of the screen is not modeled except by the absence of sources on $\mathcal{S}_b$. Adding Cases 1 and 3 and subtracting Case 2, we get

$$\text{Case 4:} \qquad \text{Sources in } \mathcal{R}' - \text{Sources on } \mathcal{S}_b \longrightarrow \begin{cases} \psi_{(a)}(P,t) & \text{in } \mathcal{R} \\ \psi(P,t) + \psi_{(a)}(P,t) & \text{in } \mathcal{R}'. \end{cases}$$

So our solution to the diffraction problem in $\mathcal{R}$—which agrees with Kirchhoff's in $\mathcal{R}$—models the screen as being equivalent to sources on $\mathcal{S}_b$ alone, namely *minus* the sources on $\mathcal{S}_b$ in Case 2. This implies that the step-changes imposed by the screen on the wave function and its normal derivative, as we cross from $\mathcal{R}'$ to $\mathcal{R}$, are *minus* the wave function and normal derivative, respectively, evaluated on $\mathcal{S}_b$ in the absence of the screen; the step-changes may be confirmed by comparing the two regions in Case 4.

In summary, Kirchhoff's inconsistent solution to the diffraction problem is the solution to a saltus problem in which the screen imposes step-changes, in the wave function and its normal derivative, equal to *minus* what the wave function and its normal derivative would be in the absence of the screen. Friedrich Kottler, in a paper published in 1923 [8, p. 411], discovered this fact by a more difficult path, involving the application of the Helmholtz formula to a more complicated surface.[24]

Kottler's saltus formulation, although notable for its mathematical consistency [1, p. 101], is not deducible from physical premises [4, p. 503]. The same of course can be said for our two equivalent formulations of the diffraction problem in $\mathcal{R}$: (i) that the screen eliminates the secondary sources on its segment of the surface of integration, retaining the secondary sources in the aperture, and (ii) that a surface segment spanning the aperture imposes saltus conditions in which the step-changes in the wave function and its normal derivative are equal to the original values of the wave function and its normal derivative in the absence of the screen. All three formulations are consistent, well-posed, and exact in their mathematical reasoning, but make physical assumptions that do *not* purport to be exact.

8.2 Diffraction as an edge effect: Maggi-Rubinowicz transformation; Poisson's spot

Recall that the Kirchhoff diffraction formula (45) expresses the wave function at point P in region $\mathcal{R}$, due to a *monopole* source at point O in the remaining region $\mathcal{R}'$, as an integral over the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$, in which the integrand is the per-unit-area contribution from secondary sources on $\mathcal{S}$. If we introduce an opaque screen with an aperture, partition $\mathcal{S}$ into segment $\mathcal{S}_a$ spanning the aperture and segment $\mathcal{S}_b$ on the back of the screen, and *assume* that the screen simply eliminates the secondary sources on $\mathcal{S}_b$, we conclude that the wave function at P is obtained by integrating over $\mathcal{S}_a$ only, with the same integrand as if the screen were not present. It was shown by Gian Antonio Maggi in 1888, and independently by Wojciech ("Adalbert") Rubinowicz in 1917, that the surface integral over the aperture can be transformed to an integral along the *edge* of the aperture, plus the "unobstructed" wave function *if* the ray line OP is within the aperture. Rubinowicz [15, pp. 257, 262–3] interpreted the transformation as a revival and revision of Thomas Young's idea [17, pp. 26, 42–4] that diffraction was due to interference between the "direct" ray (if any) and a ray or rays somehow deflected from the edge(s) of the aperture or obstruction.

In principle, the possibility of so transforming the integral is obvious enough. The goal is to find the wave function transmitted through an aperture in a screen, from primary sources on one side of the screen. If this can be done by integrating over the aperture only, with the same integrand as if the screen were absent, then the transmitted wave function depends only on the primary sources and the aperture. But the aperture is fully specified by its *edge*,[25] together with some sign convention indicating which side of the edge is the aperture side.

We can reach the same conclusion without assuming that the integral over $\mathcal{S}_a$ *alone* has any physical meaning. In the unobstructed case (from which we get the integrand), the integral of the contributions of the secondary sources over *all* of $\mathcal{S}$ is the wave function at P, which, by Proposition 3, is not altered by any change in $\mathcal{S}$, provided that all the primary sources remain on one side and P remains on the other. In particular, the integral over $\mathcal{S}$ is not affected by any change in $\mathcal{S}_a$ that leaves $\mathcal{S}_b$, and hence the edge of $\mathcal{S}_a$, unchanged. It follows (by subtraction of the integral over $\mathcal{S}_b$) that the integral over $\mathcal{S}_a$ is not affected by any change in $\mathcal{S}_a$ that leaves the edge unchanged (and which, as before, leaves P and the primary sources on their respective sides); in other words, the integral over $\mathcal{S}_a$ depends only on the edge and on the primary sources.

[24] For a summary see Baker & Copson [1, pp. 98–101].

[25] Provided of course that the edge is well defined—e.g., because the local thickness of the screen is negligible.

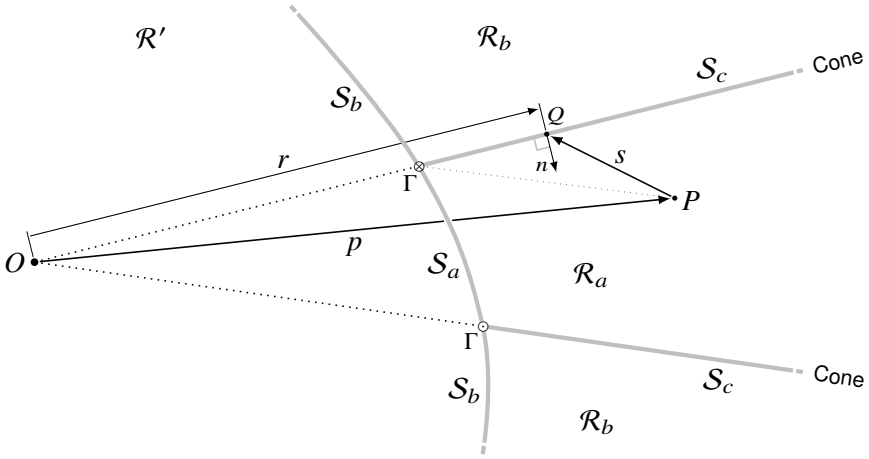

Fig. 4: Cross-section of the conical surface $\mathcal{S}_c$ projected from point O by the curve Γ, which divides the surface $\mathcal{S}$ (not labeled) into segments $\mathcal{S}_a$ and $\mathcal{S}_b$ (shown in cross-section). The cone divides the region $\mathcal{R}$ (not labeled) into subregions $\mathcal{R}_a$ and $\mathcal{R}_b$. The field point P and the direction of the normal coordinate n may independently lie outside the plane of the page, and Γ does not necessarily intersect that plane at right angles. (The dotted line from Γ to P is not needed until Appendix 8.3.)

Several methods of transforming the Helmholtz or Kirchhoff surface integral to an edge integral (or "rim" integral) have been discovered,[26] and none of them reasonably qualifies as "high-school math". So here I propose to show how the transformation can be simplified using the underlying STD sources and the principle of superposition.

Let $\mathcal{S}$ be the surface separating the region $\mathcal{R}$, containing no sources, from the remaining region $\mathcal{R}'$, containing a single monopole source of strength $f(t)$ at point O. Let Γ be a curve on $\mathcal{S}$, dividing $\mathcal{S}$ into segments $\mathcal{S}_a$ and $\mathcal{S}_b$. All this is familiar except that we have specified the primary source and named the "edge" or "rim" of $\mathcal{S}_a$. But now we introduce a new feature and a new assumption. With respect to the point-source at O, let the surface $\mathcal{S}_c$ be the *geometric shadow of the curve* Γ; in other words, let $\mathcal{S}_c$ be the infinite cone (not necessarily circular) passing through Γ, with the apex O, and with its tip cut off at Γ (Fig. 4). And let us assume that $\mathcal{S}$ does not intersect or touch the cone except at Γ, so that the cone divides $\mathcal{R}$ into two subregions: $\mathcal{R}_a$, bounded by $\mathcal{S}_a$ and $\mathcal{S}_c$; and $\mathcal{R}_b$, bounded by $\mathcal{S}_b$ and $\mathcal{S}_c$. Thus, if $\mathcal{S}_a$ is the aperture and $\mathcal{S}_b$ is the back of the screen, then $\mathcal{R}_a$ is the part of $\mathcal{R}$ illuminated by the aperture according to geometrical optics, and $\mathcal{R}_b$ is the geometric shadow of the screen (and the subscripts are a for "aperture", b for "back", and c for "cone").

Continuing the sequence of case numbers, let Case 5 be the unobstructed monopole source at O, giving the wave function $\psi(P,t)$. This is written

$$\text{Case 5:} \qquad \text{Monopole at } O \longrightarrow \psi(P,t)\,.$$

We have established that we get the same wave function in $\mathcal{R}$, and a null wave function in $\mathcal{R}'$, if the original source is replaced by STDs on $\mathcal{S}$, facing $\mathcal{R}$, with a strength density given by the original wave function (divided by 4π). Let us abbreviate this as

$$\text{Case 6:} \qquad \text{STDs on } \mathcal{S}_a \;+\; \text{STDs on } \mathcal{S}_b \longrightarrow \begin{cases} \psi(P,t) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}'. \end{cases}$$

Applying the same rule to the surface comprising $\mathcal{S}_a$ and $\mathcal{S}_c$, we get[27]

$$\text{STDs on } \mathcal{S}_a \;+\; \text{STDs on } \mathcal{S}_c \longrightarrow \begin{cases} \psi(P,t) & \text{in } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}' \text{ and } \mathcal{R}_b\,, \end{cases} \tag{68}$$

[26] See e.g. Baker & Copson [1, pp. 74–9], Born & Wolf [2, pp. 499–503], and Buchwald & Yeang [4, pp. 497–503].

[27] This surface is not smooth at Γ, but may be approximated with arbitrary accuracy by a surface that *is* smooth at Γ. Moreover, the surface area over which the assumption of smoothness is violated has zero area and therefore makes zero contribution to the wave function.

where the "STDs on $\mathcal{S}_c$" are facing $\mathcal{R}_a$. This can be written

$$\text{Case 7:} \qquad \text{STDs on } \mathcal{S}_a \;+\; \text{STDs on } \mathcal{S}_c \;\longrightarrow\; \begin{cases} \psi_{(g)}(P,t) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}', \end{cases}$$

where

$$\psi_{(g)}(P,t) \;=\; \begin{cases} \psi(P,t) & \text{in } \mathcal{R}' \text{ and } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}_b\,. \end{cases} \tag{69}$$

In the diffraction problem, with $\mathcal{S}_a$ spanning the aperture and $\mathcal{S}_b$ covering the back of the screen, $\mathcal{R}_b$ is the geometric shadow of the screen and $\psi(P,t)$ is the unobstructed wave function (as in Case 5), so that $\psi_{(g)}(P,t)$ is the wave function in $\mathcal{R}$ as predicted by *geometrical* optics (hence the subscript).

If we now confine our attention to the region $\mathcal{R}$ (which comprises subregions $\mathcal{R}_a$ and $\mathcal{R}_b$ and contains no sources), our standard solution to the diffraction problem, denoted by $\psi_{(a)}(P,t)$, is given by the STD sources on $\mathcal{S}_a$ only:

$$\text{Case 8:} \qquad \text{STDs on } \mathcal{S}_a \;\longrightarrow\; \psi_{(a)}(P,t)\,.$$

Subtracting Case 7 from Case 8 (and ignoring $\mathcal{R}'$) shows the effect of the cone in $\mathcal{R}$:

$$\text{Case } C\text{:} \qquad -(\text{STDs on } \mathcal{S}_c) \;\longrightarrow\; \psi_{(a)}(P,t) - \psi_{(g)}(P,t)\,.$$

Let the wave function due to *minus* the "STDs on $\mathcal{S}_c$" be denoted by $\psi_{(d)}(P,t)$. Setting this equal to the right-hand side of Case C and rearranging the equation, we obtain

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) + \psi_{(d)}(P,t)\,. \tag{70}$$

Thus the wave function transmitted by the aperture is expressed as the sum of two terms, of which the first is the wave function predicted by geometrical optics, and the second, $\psi_{(d)}(P,t)$, is the correction for *diffraction* (hence the subscript).

That is one advantage of working from Case C instead of the more direct Case 8. Another is that the surface $\mathcal{S}_c$ is special, allowing a simplification of the associated integral. The term $\psi_{(d)}(P,t)$ is the wave function due to *minus* the "STDs on $\mathcal{S}_c$", and is therefore given by *minus* the Helmholtz integral (30) over $\mathcal{S}_c$:

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,, \tag{71}$$

where ψ and ψ_n are evaluated at the element $d\mathcal{S}$, and the normal coordinate n is measured into $\mathcal{R}_a$ (the direction faced by the STDs). But, because $\mathcal{S}_c$ is conical with its apex at O, and because ψ depends on distance from the source at O but not on direction from O, we can put $\psi_n = 0$, eliminating the second term in the braces. Then, if we let $f(t)$ be the strength of the monopole primary source at O, the unobstructed wave function $\psi(t)$ is given by (40), so that

$$\psi(t-s/c) = \frac{1}{r}\,f(t-r/c-s/c)\,. \tag{72}$$

With these substitutions, equation (71) reduces to[28]

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\,\frac{\partial}{\partial n}\left(\frac{f(t-\tau)}{rs}\right) d\mathcal{S}\,, \tag{73}$$

where τ is the total propagation time[29] from O to P via $d\mathcal{S}$:

$$\tau = \frac{r+s}{c}\,. \tag{74}$$

[28] Continuing from footnotes 5 and 9: If the factor 4π were included in the denominator of (40), it would leave (71) unchanged, but would contribute another factor 4π to the denominator in (73)—which is one reason why we have *not* used that convention here.

[29] This differs from the meaning of τ in my last paper [13].

In (73), in the differentiation w.r.t. n at the conical surface, the coordinate r may be treated as constant because n is measured normal to the cone while r is the distance from the apex. That geometric argument does not apply to s. However, under the **far-field approximation**, we can assume that the derivative w.r.t. n is due solely to the variation in the propagation time τ, and not due to the variation in any other spatial coordinate *except* through the consequent variation in the propagation time.[30] Hence we can treat the quantity differentiated w.r.t. n as a function of τ alone, so that, by the chain rule,

$$\psi_{(d)}(P,t) \approx -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\,\frac{d}{d\tau}\left(\frac{f(t-\tau)}{rs}\right)\frac{\partial\tau}{\partial n}\,d\mathcal{S}\,. \tag{75}$$

From (74), treating r as constant, we have

$$\frac{\partial\tau}{\partial n} = \frac{1}{c}\,\frac{\partial s}{\partial n} = \frac{1}{c}\,\cos(n,s)\,, \tag{76}$$

where (n,s) means the angle between the positive directions of n and s (and the cosine function makes it hard to escape that interpretation). Substituting this into (75) gives

$$\psi_{(d)}(P,t) \approx -\iint\limits_{\mathcal{S}_c} \frac{\cos(n,s)}{4\pi c}\,\frac{d}{d\tau}\left(\frac{f(t-\tau)}{rs}\right)d\mathcal{S}\,. \tag{77}$$

So the problem is to transform this integral over the surface $\mathcal{S}_c$ to an integral along the boundary curve Γ, which is also the rim of the aperture. Consider the derivative w.r.t. τ in the integrand. One way to increase τ by $d\tau$ is to *delay* $f(t)$ *at the source* by $d\tau$. In that scenario, τ increases by $d\tau$ everywhere, without changing the range of integration, so that differentiating an integral over the fixed surface $\mathcal{S}_c$ w.r.t. τ is analogous to differentiating a sum over a fixed number of terms. Hence we can take the $d/d\tau$ operator outside the integral sign, obtaining

$$\psi_{(d)}(P,t) \approx -\frac{d}{d\tau}\iint\limits_{\mathcal{S}_c} \frac{\cos(n,s)}{4\pi crs}\,f(t-\tau)\,d\mathcal{S}\,. \tag{78}$$

Another way to increase τ by $d\tau$ is to shift every point Q on the conical surface $\mathcal{S}_c$ in the r direction (that is, away from the apex O) by such a distance that the total length of the path OQP increases by $c\,d\tau$. If we think of the segment $\mathcal{S}_c$ as a sleeve on the cone, we slide the sleeve along the cone—not by a uniform distance, but by whatever infinitesimal distance gives the uniform increase $c\,d\tau$ in the path length. Let the surface segment $d\mathcal{S}_c$ be the change in $\mathcal{S}_c$, with each area element $d\mathcal{S}$ of $d\mathcal{S}_c$ being construed as negative if (as we expect) $\mathcal{S}_c$ contracts or retreats. Then (78), becomes

$$\psi_{(d)}(P,t) \approx -\frac{1}{d\tau}\iint\limits_{d\mathcal{S}_c} \frac{\cos(n,s)}{4\pi crs}\,f(t-\tau)\,d\mathcal{S}\,. \tag{79}$$

This $d\mathcal{S}_c$ is an infinitesimally thin slice of cone bordering on the curve Γ. Now let ℓ be a coordinate measuring arc length along Γ. For a general value of ℓ, let dr be the width of $d\mathcal{S}_c$ in the r direction, and let $(r, d\ell)$ be the angle between positive directions of r and $d\ell$. Then the area element is

$$d\mathcal{S} = -dr\,d\ell\,\sin(r,d\ell)\,, \tag{80}$$

where the minus sign is needed because $\mathcal{S}_c$ retreats for positive dr. Substituting this into (79) and taking the factor $1/d\tau$ inside the integral sign, we get an integral w.r.t. ℓ, which is therefore along Γ:

$$\psi_{(d)}(P,t) \approx \int\limits_{\Gamma} \frac{\cos(n,s)\,\sin(r,d\ell)}{4\pi rs}\,f(t-\tau)\,\frac{dr}{c\,d\tau}\,d\ell\,. \tag{81}$$

[30] Here again, "far-field" is meant in the sense that s is very large compared with the spatial oscillations of the wave function, and *not* necessarily in the stronger Fraunhofer sense (*cf.* footnote 15).

To eliminate differentials other than $d\ell$, we note that

$$c\,d\tau = dr + ds = dr\left(1 + \frac{ds}{dr}\right) = dr\,(1 + \cos(r,s))\,. \tag{82}$$

Substituting this into (81), and writing out τ according to (74), we obtain the desired transformation:[31]

$$\psi_{(d)}(P,t) = \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,. \tag{83}$$

Hence, by (70), the complete wave function is[32]

$$\psi_{(a)}(P,t) = \psi_{(g)}(P,t) + \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,, \tag{84}$$

where (to recap) P is the field point or observation point, t is time, Γ is the edge or rim of the aperture, $f(t)$ is the strength of the source at point O, coordinate r is measured from the source to the edge element $d\ell$, coordinate s is measured from P to $d\ell$, coordinate n is measured normal to the cone defined by O and Γ, into the geometric shadow,[33] and coordinate ℓ is measured along Γ. For the source strength $f(t)$, the geometrical-optics term $\psi_{(g)}(P,t)$ is given by

$$\psi_{(g)}(P,t) = \begin{cases} \frac{1}{p}\,f(t-p/c) & \text{in } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}_b\,, \end{cases} \tag{85}$$

where p is the direct distance from O to P.

The equality in (83) is written as exact rather than approximate because, although we have obtained this result from a far-field approximation, earlier authors (mentioned above and in the footnotes) found equivalent results by retaining the exact Helmholtz or Kirchhoff formula (see Appendix 8.3).

In (83) and (84), the integrand is still consistent with a $1/r$ decay in the amplitude of the primary wave and a $1/s$ decay in the amplitudes of the secondary waves. But the "obliquity factor" (if we may still use the term) has a denominator that vanishes if the directions of r and s are opposite—that is, if $d\ell$ is on a straight line from O to P, so that the line of sight grazes the edge of the aperture. This allows the integral to be discontinuous at the edge of the geometric shadow—as it must be, to compensate for the discontinuity in the geometric term $\psi_{(g)}(P,t)$, because when $\psi_{(a)}(P,t)$ is expressed as a *surface* integral over $\mathcal{S}_a$, there is no cause for any discontinuity at non-zero distances from the edge.

The derivation of (84) does not require Γ to be traversed in a particular direction. But, with a view to rewriting (84) in a *vector* notation, let us now agree that Γ is traversed counterclockwise about $\mathcal{S}_a$ as seen from O (which typically means clockwise as seen from the field point P; *cf.* Fig. 4, on p. 22). Let the unit vectors in the directions of r, s, n, and $d\ell$ be, respectively, $\hat{\mathbf{r}}$, $\hat{\mathbf{s}}$, $\hat{\mathbf{n}}$, and $\hat{\mathbf{t}}$ (where $\hat{\mathbf{t}}$ is so called because it is the unit *tangent* vector to Γ). Then the vector product $\hat{\mathbf{t}}\times\hat{\mathbf{r}}$ has the magnitude $\sin(r,d\ell)$ and the direction normal to both $\hat{\mathbf{r}}$ and Γ, and toward $\mathcal{S}_a$ and $\mathcal{R}_a$—that is, the direction of $\hat{\mathbf{n}}$. Hence, if we take its scalar product with $\hat{\mathbf{s}}$, we get the factor $\cos(n,s)\,\sin(r,d\ell)$ in (84). So that factor may be replaced by $\hat{\mathbf{s}}\cdot\hat{\mathbf{t}}\times\hat{\mathbf{r}}$, or $\hat{\mathbf{r}}\times\hat{\mathbf{s}}\cdot\hat{\mathbf{t}}$. More obviously, $\cos(r,s)$ can be written $\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}$. Thus (84) becomes

$$\psi_{(a)}(P,t) = \psi_{(g)}(P,t) + \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\hat{\mathbf{r}}\times\hat{\mathbf{s}}\cdot\hat{\mathbf{t}}}{1+\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}}\,d\ell\,. \tag{86}$$

Multiplying the numerator and denominator by rs, with $\mathbf{r} = r\hat{\mathbf{r}}$ and $\mathbf{s} = s\hat{\mathbf{s}}$, we get the alternative form[34]

$$\psi_{(a)}(P,t) = \psi_{(g)}(P,t) + \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\mathbf{r}\times\mathbf{s}\cdot\hat{\mathbf{t}}}{rs+\mathbf{r}\cdot\mathbf{s}}\,d\ell\,. \tag{87}$$

[31] *Cf.* Born & Wolf [2, p. 503, eq. 19].

[32] *Cf.* Rubinowicz [15, p. 262, eq. 5].

[33] Buchwald & Yeang, in reviewing Rubinowicz, apparently err by describing $\hat{n}$ as the normal to the "screen aperture" [4, top of p. 502]; *cf.* Rubinowicz [15, p. 261] and Born & Wolf [2, p. 502, eq. 12].

[34] *Cf.* Baker & Copson [1, p. 79, eq. 2.111], and Kottler [8, p. 420, eq. 7], where Kottler's $d\bar{\mathbf{s}}_Q$ is apparently *minus* our $\hat{\mathbf{t}}\,d\ell$.

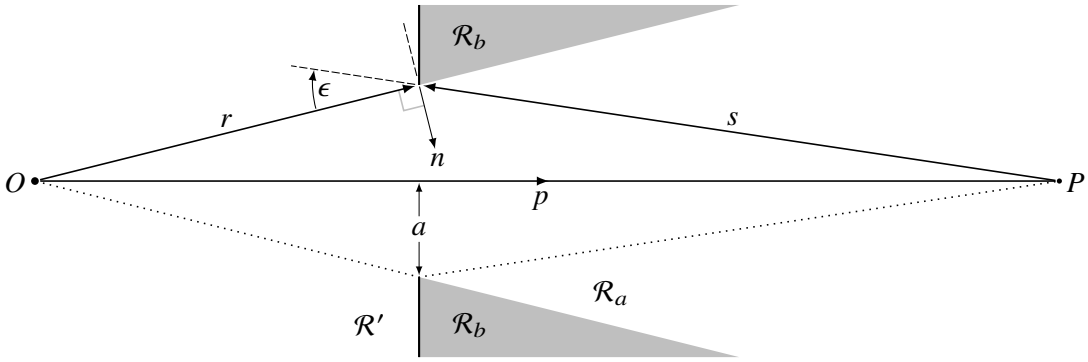

Fig. 5: Distances & angles pertaining to the calculation of the wave function at the field point P due to a monopole source at O, admitted by a small circular aperture of radius a with axis OP. The coordinate n is measured normal to the conical boundary of the geometric shadow $\mathcal{R}_b$, into the geometrically illuminated region $\mathcal{R}_a$.

Let us apply (84) and (85) to a well known pair of optical cases: one in which the aperture is small and circular, of radius a, with its axis on the line of sight (Fig. 5); and the other in which the line of sight is obstructed by a small opaque disk of the same radius, with its axis on the same line, the disk being the entire "screen". The two cases are *complementary*, meaning that the union of the two apertures gives no obstruction, while the union of the two screens gives complete obstruction. In each case, Γ is a circle with its axis on OP. By the axial symmetry, all the distances and angles are uniform over Γ, so that the edge integral is simply the product of the integrand and the length of Γ, that length being $2\pi a$. And the angle between the directions of r and $d\ell$ is 90°, so that $\sin(r, d\ell) = 1$. So (84) is reduced to

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) \;+\; 2\pi a\,\frac{f\big(t-(r{+}s)/c\big)}{4\pi rs}\;\frac{\cos(n,s)}{1+\cos(r,s)}\;. \tag{88}$$

For convenience, since the angle (r, s) is close to π (radians), let ϵ be its supplement (Fig. 5), so that ϵ is small. Then

$$\cos(r,s) = -\cos\epsilon = -\big(1-2\sin^2(\epsilon/2)\big)\,, \tag{89}$$

so that

$$1+\cos(r,s) \;=\; 2\sin^2(\epsilon/2) \approx 2(\epsilon/2)^2 = \tfrac{1}{2}\epsilon^2 \tag{90}$$

(for small ϵ, in radians). Now ϵ, as an exterior angle of the long triangle, is the sum of the interior opposite angles, each of which may be approximated by its sine, so that $\epsilon \approx (a/r + a/s)$. So we have

$$1+\cos(r,s) \approx \tfrac{1}{2}\,(a/r+a/s)^2 = \frac{a^2}{2}\Big(\frac{r{+}s}{rs}\Big)^2. \tag{91}$$

By the axial symmetry, the directions of r, s, and n (in Fig. 5) are coplanar, so that the angle (n, s) is $\frac{\pi}{2} \pm \epsilon$, where we adopt the convention that the upper sign is for the circular aperture (n measured *toward* the axis, as shown) and the lower sign is for the disk (n measured *away from* the axis). Hence, under the same convention,

$$\cos(n,s) = \mp\sin\epsilon \approx \mp\epsilon \approx \mp(a/r+a/s) = \mp\,\frac{a(r{+}s)}{rs}\;. \tag{92}$$

When we substitute (91) and (92) into (88), most factors cancel out and we are left with

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) \;\mp\; \frac{1}{r{+}s}\,f\big(t-(r{+}s)/c\big)\,. \tag{93}$$

For the *circular aperture*, point P is on the axis of the geometrically illuminated region, so we take the upper option in (85) with the upper sign in (93), obtaining

$$\psi_{(a)}(P,t) \;=\; \frac{1}{p}\,f(t-p/c) \;-\; \frac{1}{r{+}s}\,f\big(t-(r{+}s)/c\big)\,, \tag{94}$$

where p is the direct distance from the source at O to the point P, and $r{+}s$ is the distance from O to P via the edge of the aperture. The first term on the right is the *unobstructed* wave function. As the

aperture becomes smaller, $a \to 0$ and $r+s \to p$. In the limit, according to (94), the two terms on the right cancel; that is, as the aperture closes, the field point goes dark. The correctness of the result in this trivial limiting case tends to confirms the sign of the integral in (84) and its subsequent treatment.

For the *disk*, point P is on the axis of the geometric shadow, so we take the lower option in (85) and the lower sign in (93), obtaining

$$\psi_{(a)}(P,t) \;=\; \frac{1}{r+s}\, f(t-(r+s)/c)\;. \tag{95}$$

As $a \to 0$, meaning that the disk shrinks to a point, $r+s \to p$, so that the wave function approaches the unobstructed wave function. Again the limiting case is obvious; but again it has been obtained via (84), helping to confirm the signs.

For the aperture (94), the surface $\mathcal{S}_a$ spans a circle. For the disk (95), $\mathcal{S}_a$ covers the rest of $\mathcal{S}$. So the Helmholtz or Kirchhoff surface integrals for the two cases must add up to the unobstructed wave function, and the transformed wave functions in (94) and (95) must do the same—as they do.

Having checked that equations (94) and (95) satisfy the obvious sanity tests, we must now point out a non-obvious implication. According to (95), the wave function on the axis of the geometric shadow of the disk is the same as the *unobstructed* wave function, except that the propagation distance is increased from p (the direct distance) to $r+s$ (the distance via the edge of the disk). If the radius of the disk is small compared with r and s, the increase in the path length will be very small in percentage terms; but, for reasonable experimental dimensions, it will also be large compared with optical wavelengths.[35] So, on the axis of the shadow, the denominator in (95) will be nearly the same as in the unobstructed case, and the wave function will be nearly the same apart from the propagation delay. Meanwhile, at points well within the shadow but substantially *off* the axis, the contributions to the rim integral in (84), from the various points on the rim, will have propagation delays spread over many cycles, and will therefore tend to cancel out. Thus we conclude that *if a small disk is illuminated by a point-source of light on its axis, there will be a bright spot in the middle of the shadow, where the intensity of the light will be nearly the same as outside the shadow*! This spot—predicted by Poisson using Fresnel's theory, then verified experimentally by Arago, and therefore known as **Poisson's spot** or **Arago's spot**—is the subject of what is probably the most-told anecdote in the history of optics, although its importance in tipping the argument in Fresnel's favor has been greatly exaggerated (see [14] and the references therein).

In the complementary case of the small circular aperture (Fig. 5), at points that are well *off* the axis of the aperture and away from the edge of the geometric shadow, the rim integral in (84) again tends to cancel out, so that the wave function is roughly as predicted by geometrical optics. But *on* the axis, according to (94) with $r+s \approx p$, the wave function is very nearly

$$\frac{1}{p}\, f(t-p/c) \;-\; \frac{1}{p}\, f(t-(r+s)/c) \tag{96}$$

—i.e., very nearly the difference between the unobstructed wave function and a delayed version thereof, this delay being due to the slightly greater length of the path via the edge of the aperture. In the case of *sinusoidal* time-dependence, the two terms cancel if the path lengths differ by a whole number of wavelengths, but reinforce each other if the path lengths differ by an odd number of *half*-wavelengths. In terms of the integral over the surface $\mathcal{S}_a$, this means that the contributions from the various parts of the surface will cancel if their propagation delays range over a whole number of cycles. If $\mathcal{S}_a$ is chosen so that the obliquity factor can be neglected (taken as 1), the cancellation depends on a proportionality between the increment $d\tau$ in the propagation delay, and the area of the surface segment for which the delay is between τ and $\tau+d\tau$; this proportionality can be verified independently (*cf.* [3], pp. 373–6).

For monochromatic light, equation (96) predicts a series of nodes and antinodes along the axis of the beam admitted through the aperture, because the path difference becomes smaller as we move away from the source. For *white* light, this variation in the path difference causes a variation in the wavelength(s) at which the path difference is an even or odd number of half-wavelengths, hence a variation in color along the axis. However, if the path difference is sufficiently large, there will be so many visible wavelengths at

[35] E.g., if $r = s = 500$mm and $a = 5$mm, the path difference will be 0.05mm, i.e. 0.005%, or 100 wavelengths @ $\lambda = 500$nm.

which the path difference is an even or odd multiple of a half-wavelength, and so many such wavelengths detectable by each type of color sensor in the eye, that the light will look white even on the axis. These predictions, again hinted at by Poisson using Fresnel's theory, were elaborated and experimentally verified by Fresnel himself.[36]

Bright spots in shadows, or brighter or darker or differently-colored spots in illuminated areas, occur where the propagation delay from the source to the field point via the edge of the aperture is uniform over a non-zero length of the edge, so that the contributions from the length elements to the integral in (84) interfere constructively. But this circumstance is exceptional. Usually the distribution of propagation times (or path lengths) via the points on the edge is not infinitely concentrated at particular values, but distributed over a continuous range. Hence, as the frequency increases, the contributions from points on the edge are spread over an increasing number of cycles, and cancel out more completely, so that the contribution from the integral in (84) becomes smaller and smaller, and the wave function is approximated more and more accurately by the geometric term $\psi_{(g)}(P,t)$. This conforms to the principle that geometrical optics (ray optics) is the limit of physical optics (wave optics) as the wavelength approaches zero [1, pp. 79–84].

8.3 Exactness of the transformation; approximation for a directional primary source

For a single monopole primary source of strength $f(t)$ at point O, the contribution to the wave function at point P due to diffraction (as distinct from geometrical optics) is given by equations (73) and (74) as

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi r}\,\frac{\partial}{\partial n}\left(\frac{f(t-(r+s)/c)}{s}\right) d\mathcal{S}\,, \tag{97}$$

where the surface $\mathcal{S}_c$ is the conical boundary of the geometric shadow, the coordinates r and s are the distances from O and P to a general point Q on $\mathcal{S}_c$, and n is the normal coordinate measured from $\mathcal{S}_c$ away from the shadow (see Fig. 4, on p. 22). This n is also a coordinate of Q, but the derivative w.r.t. n is evaluated at $n=0$, so that Q is indeed on $\mathcal{S}_c$. The factor r has been taken outside the differentiation because it is insensitive to variations in n.

We have seen that the surface integral (97) can be transformed to the edge integral (83) by means of a far-field/short-wavelength approximation. However, earlier derivations of the same transformation do not depend on that approximation, whether they use general time-dependence, like Maggi, or sinusoidal time-dependence, like Rubinowicz, Kottler, and Born & Wolf.[37] We shall now rework the transformation exactly. Our method is essentially that of Rubinowicz, except that we allow general time-dependence, which paradoxically makes the mathematics more elementary.

As before, let Γ be the edge of the aperture and the edge of $\mathcal{S}_c$, and let ℓ be a coordinate measuring arc length along Γ. At the edge element $d\ell$, let the values of the coordinates r and s be r_a and s_a (with the subscript a for "aperture");[38] these are functions of ℓ. For a given ℓ, hence a given point Q_a on Γ, the line through O and Q_a is a generating line of the cone of $\mathcal{S}_c$; and as r ranges from r_a to ∞ along this generator, the point Q runs from Γ to infinity. Hence, as ℓ varies over Γ, and r varies from r_a to ∞ for each ℓ, the point Q traces out the surface $\mathcal{S}_c$. Accordingly, we shall use ℓ and r as *parameters* of the surface—that is, as coordinates by which any point Q on $\mathcal{S}_c$ can be uniquely identified. In terms of these parameters, the area element of the conical surface (Fig. 6) is

$$d\mathcal{S} \;=\; \frac{r}{r_a}\,d\ell\,\sin(r,d\ell)\,dr\,; \tag{98}$$

[36] See e.g. Buchwald [3] at p. 376, and Fresnel *et al.* [7], vol. 1, pp. 245–6.

[37] On Maggi and Rubinowicz, see Buchwald & Yeang [4, pp. 497–503]. On Kottler, see Baker & Copson [1, pp. 74–9]. Born & Wolf [2, pp. 499–503] assume at the outset that the dimensions of the aperture are large compared with the wavelength and small compared with the distances to the source and the field point; but neither assumption is actually invoked.

[38] Born & Wolf [2, p. 503, eq. 19] use a subscript 1 instead of a.

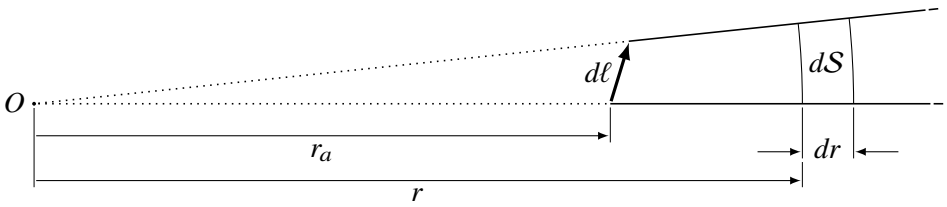

Fig. 6: Area element $d\mathcal{S}$ of the conical surface $\mathcal{S}_c$ (looking in the n direction in Fig. 4, p. 22).

the magnitude of this area is greater, by a factor r/r_a, than that given by (80) at Γ, because the distance between neighboring generators of the cone is proportional to r. Substituting (98) into (97) gives

$$\psi_{(d)}(P,t) \;=\; -\int\!\!\int_{\Gamma}\!\!\int_{s_a}^{\infty} \frac{\sin(r,d\ell)}{4\pi r_a}\,\frac{\partial}{\partial n}\left(\frac{f\big(t-(r+s)/c\big)}{s}\right) dr\,d\ell\,, \tag{99}$$

where the surface integration is shown in two stages: the inner integral w.r.t. r, multiplied by $d\ell$, is the integral over the truncated sectoral strip that constitutes the geometric shadow of the element $d\ell$ (to the right of $d\ell$ in Fig. 6); and the outer integration over Γ adds up the contributions from all the strips.

So, to reduce the double integral to a single integral over Γ, it suffices to evaluate the inner integral analytically. Since r can be treated as constant in the differentiation w.r.t. n, we have by the chain rule

$$\frac{\partial}{\partial n}\left(\frac{f\big(t-(r+s)/c\big)}{s}\right) \;=\; \frac{\partial}{\partial s}\left(\frac{f\big(t-(r+s)/c\big)}{s}\right)\frac{\partial s}{\partial n} \;=\; \frac{\partial}{\partial s}\left(\frac{f\big(t-(r+s)/c\big)}{s}\right)\cos(n,s)\,. \tag{100}$$

Substituting this into (99) and multiplying the numerator and denominator by s, we get

$$\psi_{(d)}(P,t) \;=\; -\int\!\!\int_{\Gamma}\!\!\int_{s_a}^{\infty} \frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\,\frac{1}{s}\,\frac{\partial}{\partial s}\left(\frac{f\big(t-(r+s)/c\big)}{s}\right) dr\,d\ell\,. \tag{101}$$

As noted by Rubinowicz [15, p. 261], the factor $s\cos(n,s)$ is the scalar component of the vector PQ in the n direction; and this direction and the component are independent of r for given ℓ (see Fig. 4). More obviously, for given ℓ, the directions of r and $d\ell$ are fixed, so that $\sin(r,d\ell)$ is fixed. And of course $4\pi r_a$ is constant. Taking these r-independent factors outside the inner integral w.r.t. r, then expanding the derivative w.r.t. s by the product rule, and taking the factor $1/s$ into the expansion, we get

$$\psi_{(d)}(P,t) \;=\; \int_{\Gamma} \frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\int_{s_a}^{\infty}\left\{\frac{f'\big(t-(r+s)/c\big)}{cs^2} + \frac{f\big(t-(r+s)/c\big)}{s^3}\right\} dr\,d\ell\,, \tag{102}$$

in which the leading minus sign has vanished because it has canceled with minus signs in the expansion.

To evaluate the inner integral, we must find a function whose derivative w.r.t. r (taking into account the dependence of s on r) is equal to the expression in braces. The obvious approach is to express s in terms of r by means of geometry. In Fig. 4, the values of r and s at Γ are r_a and s_a. So, applying the cosine rule to the triangle $Q\Gamma P$, with the *exterior* angle $O\Gamma P = (r_a, s_a)$, we have

$$s^2 \;=\; (r-r_a)^2 + {s_a}^2 + 2(r-r_a)s_a\cos(r_a,s_a)\,. \tag{103}$$

Hence we could take the square root of the right-hand side, substitute it for every instance of s in the braces in (102), and try to find an antiderivative! Fortunately there is a tidier approach, for which we are again indebted to Rubinowicz. By the chain rule, the first term in the braces is easily shown to be

$$-\,\frac{1}{s^2\left(1+\frac{ds}{dr}\right)}\,\frac{d}{dr}\,f\left(t-\tfrac{r+s}{c}\right). \tag{104}$$

Now we invoke (103) by differentiating it w.r.t. r (obviously with r_a and s_a constant), to obtain

$$s\,\frac{ds}{dr} \;=\; G(r)\,, \tag{105}$$

where, for brevity,

$$G(r) = r - r_a + s_a \cos(r_a, s_a)\,, \tag{106}$$

whence[39]

$$G'(r) = 1\,. \tag{107}$$

By (105), we can put $\frac{ds}{dr} = G(r)/s$ in (104), which then becomes

$$-\frac{1}{s(s+G(r))}\,\frac{d}{dr}\,f\Big(t - \tfrac{r+s}{c}\Big)\,. \tag{108}$$

By the product rule, this is easily recognized as

$$-\frac{d}{dr}\left\{\frac{f(t-(r+s)/c)}{s(s+G(r))}\right\} + f\Big(t - \tfrac{r+s}{c}\Big)\,\frac{d}{dr}\left\{\frac{1}{s(s+G(r))}\right\} \tag{109}$$

or, if we evaluate the last derivative using the product and chain rules and then apply (105) and (107),

$$-\frac{d}{dr}\left\{\frac{f(t-(r+s)/c)}{s(s+G(r))}\right\} - \frac{f(t-(r+s)/c)}{s^3}\,. \tag{110}$$

Since this is equal to the *first term* in the expression in the braces in (102), the entire inner integrand is equal to

$$-\frac{d}{dr}\left\{\frac{f(t-(r+s)/c)}{s(s+G(r))}\right\}, \tag{111}$$

so that the inner integral is

$$-\left.\frac{f(t-(r+s)/c)}{s(s+G(r))}\right|_{r_a}^{\infty} = -\left.\frac{f(t-(r+s)/c)}{s(s+r-r_a+s_a\cos(r_a,s_a))}\right|_{r_a}^{\infty} = \frac{f(t-(r_a+s_a)/c)}{{s_a}^2(1+\cos(r_a,s_a))}\,, \tag{112}$$

and the outer integral becomes

$$\psi_{(d)}(P,t) = \int_\Gamma \frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\,\frac{f(t-(r_a+s_a)/c)}{{s_a}^2(1+\cos(r_a,s_a))}\,d\ell\,. \tag{113}$$

In this integral, because we have shown that the first factor in the integrand is independent of r, or because the integrand is evaluated on Γ (either reason being sufficient), we can substitute r_a and s_a for r and s. Alternatively, because the path of integration determines r and s for given ℓ, we can simply drop the subscripts a. Choosing the latter option, then canceling a factor s and rearranging the other factors, we obtain

$$\psi_{(d)}(P,t) = \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi r s}\,\frac{\cos(n,s)\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,. \tag{114}$$

This exact result agrees with (83), which we derived by a far-field/short-wavelength approximation. So, in the special case to which this transformation applies—the case of a single monopole primary source—the mathematical errors in our initial derivation somehow canceled out.

In our derivation of the Maggi-Rubinowicz formula, we assume that the unobstructed wave function is of the form

$$\frac{1}{r}\,f(t-r/c)\,, \tag{115}$$

where r is the distance from the source, and f is a function that does not depend on direction. But the non-dependence of f on direction is invoked in only one place, namely equation (71), where we claim that ψ_n is zero because it is measured normal to a cone with its apex at the source. That claim remains a good approximation for a directional f, provided that f varies sufficiently gradually over each wavefront. The *exact* transformation of the cone integral to the rim integral is a matter of integrating a function involving f w.r.t. r *for given* ℓ. As f is arbitrary and ℓ is held constant, this operation is

[39] $G(r)$ is my abbreviation, not Rubinowicz's.

indifferent to whether f varies with ℓ (as s and the angles generally do). The corresponding *approximate* transformation assumes that spatial derivatives of $f(t-\tau)$ are negligible except through the consequent variations in τ. This is clearly the case for sufficiently gradual spatial variations in f due to directionality. The initial transformation of the aperture integral to the geometrical-optics term plus the cone integral (eqs. 70 & 71) assumes a "single monopole source" for the sole purpose of establishing a well-defined geometric shadow; in equation (69), the directionality or otherwise of $\psi(P,t)$ is immaterial.

Thus, for a directional primary source that gives recognizable spherical wavefronts, with a sufficiently gradual variation in intensity over each wavefront, the Maggi-Rubinowicz formula (84), like Kirchhoff's diffraction formula (45), is still approximately correct, and the diffraction can still be well described as an edge effect.

8.4 Revision history

Version 1 of this paper suffered from a *saltus in demonstrando* (see footnote 6 on p. 8 of the current version). This was addressed in Version 2 by introducing the concept of *naive* wave functions and BCs. Version 2 also reversed the arrow for s (for consistency with the references) and relegated some material to Appendix 8.1. Version 3 added Appendix 8.2 and clarified the regions of applicability of the various formulations of the diffraction problem. Version 4 (current) adds Appendices 8.3 & 8.4 and belatedly explains that, from Section 4 onward, n is a coordinate of Q for the purpose of evaluating derivatives w.r.t. n; these are evaluated at $n=0$, which means that Q is *on* the indicated surface.

9 Conclusion

Whereas the presentation of Kirchhoff's integral theorem and its corollaries has hitherto been thought to require sophomore-level mathematics, this paper develops the same theory from concepts that are either familiar to high-school graduates or easily introduced to them as needed. The approach is both heuristic and traditional, proceeding from the sequential nature of wave propagation, through superposition, boundary conditions, Huygens' principle, and spatiotemporal dipole (STD) sources, to the Kirchhoff integral theorem, the monopole primary source, obliquity factors, Fresnel's choice of the primary wavefront as the surface of integration, and special cases for large distances. The assumption of sinusoidal time-dependence is not essential, but is treated as another special case; this has been done without complex numbers by defining j as a phase-shift operator. In the problem of diffraction by an aperture in an opaque screen, in which the inconsistencies in Kirchhoff's boundary conditions have hitherto been either tolerated, or circumvented by introducing further complexity, the simpler approach of the present paper avoids the cause of the inconsistency and leads to other consistent formulations of the problem. By the same approach, the diffraction integral for a single monopole primary source can be expressed as a geometric term plus an integral over the conical boundary of the geometric shadow (with a minus sign for the chosen direction of the normal). This leads to two derivations of the Maggi-Rubinowicz transformation for general time-dependence: an easy one, which invokes a far-field approximation; and a non-obvious one (following Rubinowicz), which does not. For these reasons I entertain the hope that the essential mathematical theory of Huygens' principle and diffraction has been rendered more accessible.

References

[1] B.B. Baker & E.T. Copson, *The Mathematical Theory of Huygens' Principle*, Oxford, 1939.

[2] M. Born & E. Wolf, *Principles of Optics*, 7th Ed., Cambridge, 1999 (reprinted with corrections, 2002).

[3] J.Z. Buchwald, *The Rise of the Wave Theory of Light: Optical Theory and Experiment in the Early Nineteenth Century*, University of Chicago Press, 1989.

[4] J.Z. Buchwald & C.-P. Yeang, "Kirchhoff's theory for optical diffraction, its predecessor and subsequent development: the resilience of an inconsistent theory", *Archive for History of Exact Sciences*, vol. 70, no. 5 (Sep. 2016), pp. 463–511; doi.org/10.1007/s00407-016-0176-1.

[5] R.P. Feynman, R.B. Leighton, & M. Sands, *The Feynman Lectures on Physics*, California Institute of Technology, 1963–2013; feynmanlectures.caltech.edu.

[6] A. Fresnel, "Mémoire sur la diffraction de la lumière" (submitted 29 July 1818, "crowned" 15 March 1819), partly translated as "Fresnel's prize memoir on the diffraction of light", in H. Crew (ed.), *The Wave Theory of Light: Memoirs by Huygens, Young and Fresnel*, American Book Co., 1900, pp. 81–144; archive.org/details/wavetheoryofligh00crewrich/page/81.

[7] A. Fresnel (ed. H. de Senarmont, E. Verdet, & L. Fresnel), *Oeuvres complètes d'Augustin Fresnel* (3 vols.), Paris: Imprimerie Impériale, 1866, 1868, 1870.

[8] F. Kottler, "Zur Theorie der Beugung an schwarzen Schirmen", *Annalen der Physik*, vol. 375, no. 6 (1923), pp. 405–56; doi.org/10.1002/andp.19233750602.

[9] J. Larmor, "On the mathematical expression of the principle of Huygens" (read 8 Jan. 1903), *Proceedings of the London Mathematical Society*, Ser. 2, vol. 1 (1904), pp. 1–13.

[10] J. Larmor, "On the mathematical expression of the principle of Huygens—II" (read 13 Nov. 1919), *Proceedings of the London Mathematical Society*, Ser. 2, vol. 19 (1921), pp. 169–80.

[11] E.W. Marchand & E. Wolf, "Consistent formulation of Kirchhoff's diffraction theory", *Journal of the Optical Society of America*, vol. 56, no. 12 (Dec. 1966), pp. 1712–22; doi.org/10.1364/JOSA.56.001712.

[12] D.A.B. Miller, "Huygens's wave propagation principle corrected", *Optics Letters*, vol. 16, no. 18 (15 Sep. 1991), pp. 1370–72; stanford.edu/~dabm/146.pdf.

[13] G.R. Putland, "A tautological theory of diffraction", doi.org/10.5281/zenodo.3563468, 5 Dec. 2019.

[14] G.R. Putland *et al.*, "Augustin-Jean Fresnel", *Wikipedia* (10 Dec. 2017—).

[15] A. Rubinowicz, "Die Beugungswelle in der Kirchhoffschen Theorie de Beugungserscheinungen", *Annalen der Physik*, vol. 358, no. 12 (1917), pp. 257–78; doi.org/10.1002/andp.19173581202.

[16] J. Saatsi & P. Vickers, "Miraculous success? Inconsistency and untruth in Kirchhoff's diffraction theory", *British J. for the Philosophy of Science*, vol. 62, no. 1 (March 2011), pp. 29–46; jstor.org/stable/41241806. (Pre-publication version, with different pagination: dro.dur.ac.uk/10523/1/10523.pdf.)

[17] T. Young, "On the theory of light and colours" (Bakerian Lecture, read 12 Nov. 1801), *Philosophical Transactions of the Royal Society*, vol. 92 (1802), pp. 12–48; rstl.royalsocietypublishing.org/content/92/12.