Consistent derivation of Kirchhoff's integral theorem and diffraction formula and the Maggi-Rubinowicz transformation using high-school math

Gavin R. Putland,* BE PhD (Qld)

Version 0.41; December 27, 2022

A theory of diffraction is constructed on three principles: causality, superposition, and the assumption that the wave function had a beginning. Given a region $\mathcal{R}$ containing no sources and separated from the remaining region $\mathcal{R}'$ by a surface $\mathcal{S}$, these principles suffice to show that if the ("primary") sources in $\mathcal{R}'$ are *replaced* by a distribution of ("secondary") sources on $\mathcal{S}$, such that the step-changes in the wave function and its normal derivative, in crossing from $\mathcal{R}'$ to $\mathcal{R}$, are respectively equal to the original wave function and its normal derivative on $\mathcal{S}$, then we get the original wave function in $\mathcal{R}$ and a null wave function in $\mathcal{R}'$ (no backward secondary waves).

By further assuming the form of the wave function due to a monopole source, we obtain the distribution of monopole and dipole sources over $\mathcal{S}$ that causes the desired step-changes ("saltus conditions"). Adding the contributions from these secondary sources yields an integral expressing the wave function in $\mathcal{R}$ in terms of the boundary conditions at the $\mathcal{R}$ side of $\mathcal{S}$. Combining this formula with the null wave function in $\mathcal{R}'$, we have the Helmholtz integral theorem for general time-dependence (whereas the traditional derivation is for sinusoidal time-dependence). The Kirchhoff integral theorem follows by elementary rules of differentiation. Diffraction by an aperture in a baffle is modeled by letting $\mathcal{S}$ comprise two segments, namely $\mathcal{S}_a$ spanning the aperture and $\mathcal{S}_b$ on the $\mathcal{R}$ side of the baffle, and *assuming* that the baffle simply eliminates the secondary sources on $\mathcal{S}_b$ (without changing those on $\mathcal{S}_a$), with the result that the range of integration is limited to $\mathcal{S}_a$ while the integrand is unchanged. This "secondary-source selection" assumption circumvents the need to assume boundary conditions on both the wave function and its normal derivative, and thereby avoids the notorious inconsistency in Kirchhoff's boundary conditions, but yields the same result. The "secondary-source selection" assumption has a saltus interpretation and is easily shown to be equivalent to Kottler's saltus formulation—involving saltus conditions at $\mathcal{S}_b$—which also has a more intuitive interpretation based on secondary sources.

The case of a *single monopole primary source* leads to two simplifications. First, the spatial derivatives in the Kirchhoff integrand are expressed in terms of angles, yielding the Kirchhoff diffraction formula and its far-field obliquity factor, which reduces to the familiar $(1+\cos\chi)/2$ if $\mathcal{S}$ is a primary wavefront. Second, the Helmholtz form of the integral over an aperture is expressed as a geometrical-optics term plus an integral over the conical boundary of the geometric shadow, and the latter term is converted to an integral along the edge of the aperture (Maggi-Rubinowicz transformation, for general time-dependence) by a simplified method based on a short-wavelength approximation; an exact conversion modeled on Rubinowicz's, yielding the same result, is given in an appendix. A single monopole primary source also allows the monopole and dipole secondary sources on a surface element $d\mathcal{S}$ to be replaced by a single "generalized spatiotemporal dipole" (GSTD), in which the inverted monopole is delayed and attenuated relative to the uninverted one. If $\mathcal{S}$ is a primary wavefront in the far field of the primary source, the GSTD reduces to the "spatiotemporal dipole" proposed by D. A. B. Miller in 1991.

The above results, having been derived for general time-dependence, are then restated for sinusoidal time-dependence, showing that previously mentioned "far-field" and "short-wavelength" approximations are synonymous. A *consistent* introduction to complex numbers and the "cis" representation is included. However, the sinusoidal (monochromatic) diffraction integrals are not cited in the subsequent analysis of Poisson's spot and the field on the axis of a circular aperture; these cases can be described in terms of the unobstructed wave function and a delayed version thereof.

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

Contents

1 Foundations

1.1 Wave functions; boundary conditions vs. saltus 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 the case or light or other electromagnetic waves, the "medium" may be a vacuum.) Let the resulting pattern of waves be described by a function $\psi(P,t)$, called the **wave function**, representing some physical quantity which depends on the position P (the **field point**, or observation point) and the time t.

Assumption 1 *The wave function had a beginning.*

That is, there was a time before which ψ was zero everywhere. This obviously implies that the sources were null before that time. We do not (yet) assume anything else about the sources.

Assumption 2 *Let space be divided into a region $\mathcal{R}$, a region $\mathcal{R}'$, and a surface $\mathcal{S}$ separating the two regions; and let $\mathcal{R}$ contain* none *of the sources.*

This $\mathcal{S}$ is merely a geometric surface, not necessarily a material one; we say "*Let*..." because, given the sources, we can always choose $\mathcal{S}$ to satisfy the stated condition.

For clarity and emphasis, we shall often repeat some or all of Assumption 2 in later statements. This assumption does not specify whether the sources are strictly inside $\mathcal{R}'$ or on $\mathcal{S}$, provided that they are not inside $\mathcal{R}$. For convenience, however, we shall choose $\mathcal{S}$ so that the actual sources (which we shall call "primary" sources) are strictly inside $\mathcal{R}'$. (Later we shall consider hypothetical alternative sources, usually called "secondary" sources, which will be *on* $\mathcal{S}$.)

We do not care if the waves emitted by some "sources" are partly or fully composed of reflections of waves from elsewhere; for our purposes, reflectors are simply sources (and classified as "primary").

Neither do we 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 through the surface $\mathcal{S}$*. When they pass through $\mathcal{S}$, 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 the region $\mathcal{R}$, bounded by the 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:

Q1. What information on the wave function at the boundary surface is sufficient to determine the function throughout the region?

Q2. Why might it be useful to determine the wave function in this way?

Q3. What formula yields the wave function at any point in the region, in terms of the said information about the function at the boundary surface?

Apropos of Q1, 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 $\mathcal{R}$. Its use involves an assumption which should be acknowledged:

Assumption 3 *The wave function $\psi(P,t)$ is differentiable w.r.t. spatial displacements of the point P.*

[1] *Warning*: Some authors in some contexts measure the normal coordinate in the opposite direction, changing the signs of first derivatives w.r.t. that coordinate in their results. In this paper, n is measured *into* the region containing *no sources*.

The specification of ψ_n at every point on a surface for all time, on a specified side of the surface, is called a **boundary condition** (or **BC**) "on" the normal derivative ψ_n. Similarly, the specification of ψ at every point on a surface for all time, on a specified side of the surface, is called a boundary condition on the wave function ψ. A boundary condition at a surface is to be distinguished from a **saltus condition**, which specifies the *step-change*, i.e. the "jump" (Latin: *saltus*), in a quantity as we cross the surface in an agreed "positive" direction: for our surface $\mathcal{S}$, we conveniently choose the direction of n. (Reversing the direction of crossing obviously changes the sign of the "jump".) Thus the specification of the step-change in the wave function or its normal derivative at every point on a surface for all time is called a saltus condition on the wave function or the normal derivative, respectively. Naturally enough, the problem of finding a position-dependent function of a given type (such as a wave function) with given *boundary* conditions is called a **boundary-value problem**, whereas the problem of finding such a function with given *saltus* conditions is called a **saltus problem**.

Now consider the boundary-value problem in which, at every point on $\mathcal{S}$, on the side facing $\mathcal{R}$, *either* the wave function *or* its normal derivative is specified for all time.[2] Suppose that the given boundary 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 $\mathcal{R}$ (Assumption 2), and which satisfy the same BC at $\mathcal{S}$. Changing the signs in one of these cases and adding the other case (applying the principle of **superposition**), we find that the non-zero difference between the two wave 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 two wave functions had beginnings, their difference had a beginning,[3] and its initial entry into $\mathcal{R}$ at any point on $\mathcal{S}$ 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 the region $\mathcal{R}$, bounded by the surface $\mathcal{S}$, contains no sources, the specification for all time of either the wave function or its normal derivative at each point on $\mathcal{S}$, on the side facing $\mathcal{R}$, determines the wave function throughout $\mathcal{R}$.*

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

In principle, Proposition 2 answers Q1. But there's a catch. The surface $\mathcal{S}$ is not assumed to have any particular shape, and is not assumed to be a physical barrier. Hence, in specifying "either the wave function or its normal derivative" at each point on the $\mathcal{R}$ side of $\mathcal{S}$, we must beware of the influence of the BCs on the $\mathcal{R}'$ side at remote points on $\mathcal{S}$, because $\mathcal{S}$ may be locally concave on the $\mathcal{R}'$ side, or even closed around $\mathcal{R}'$, allowing such influence to propagate across $\mathcal{R}'$ from one point on $\mathcal{S}$ to another, and thence *through* $\mathcal{S}$, affecting the BCs on the $\mathcal{R}$ side. Moreover, Proposition 2 nominates only one purpose for which we need only specify "either the wave function or its normal derivative". There might be other purposes for which it is useful to specify both. Indeed, as the next three propositions will show, one such purpose is the avoidance of unwanted influence from the $\mathcal{R}'$ sides of remote points on $\mathcal{S}$.

Proposition 3 *If space is divided into regions $\mathcal{R}$ and $\mathcal{R}'$ and a surface $\mathcal{S}$ separating them, with no sources in $\mathcal{R}$, then, as far as the wave function in $\mathcal{R}'$ is concerned, a distribution of sources or other matter on $\mathcal{S}$ will be indistinguishable from the absence of any such distribution if, and only if, the presence or absence of that distribution gives the same boundary conditions on the wave function* and *its normal derivative at the $\mathcal{R}'$ side of $\mathcal{S}$.*

[2] The wording is meant to allow the wave function to be specified at some points and the normal derivative at other points; but this technicality is not essential for our purposes.

[3] This automatically satisfies any requirement that we specify *initial conditions* (*cf.* Baker & Copson [1], p. 41), because if the time line goes back far enough, the initial values are everywhere zero.

To understand why, we may reason from the simplest example of waves in three dimensions, namely sound in a non-viscous fluid, and apply the result to other cases in which the wave function is observed to behave analogously. For sound, we may take the wave function to be the sound pressure (relative to equilibrium), which has a gradient, which causes an acceleration, which when integrated twice w.r.t. time, over all time up to the present, gives a displacement; and the integrals are well defined because there was a time before which the sound pressure and hence the displacement were everywhere zero (Assumption 1).[4] The convergence or divergence of the displacement gives a compression or rarefaction, hence a change in the pressure; and so the cycle of causation continues. In a three-dimensional region, specifying the wave function implicitly specifies its gradient and all that follows therefrom. On a *surface*, however, specifying the wave function implicitly specifies its gradient in directions tangential to the surface, but not normal to the surface, so that the normal derivative—if relevant—must be specified separately. So a distribution of sources or other matter on a surface is fully characterized by, at most, its effect on the wave function and its normal derivative. And of course both effects are relevant—the former because it may alter the wave function on the surface, and the latter because it may alter the wave function at small distances *off* the surface (and, in the case of sound, alter the acceleration at the surface). Thus, for the distribution on $\mathcal{S}$ to imitate the absence of any such distribution, it is sufficient, and necessary, to imitate the wave function and the normal derivative on the $\mathcal{R}'$ side of $\mathcal{S}$.

It follows from Proposition 3 that if the presence or absence of the distribution on $\mathcal{S}$ gives the same boundary conditions on the wave function *and* its normal derivative, the distribution will give no reflection or echo of the waves from the sources in $\mathcal{R}'$—because the "absence" of the distribution means simply a continuation of the medium, through which the waves continue to propagate, with nothing to cause any reflection or echo.

Returning to Q1, let us consider a superposition of two cases in which, as usual, space is divided into regions $\mathcal{R}$ and $\mathcal{R}'$ and a surface $\mathcal{S}$ between them, with no sources in $\mathcal{R}$.

Case 1: We have the original sources in $\mathcal{R}'$. We get the original wave function in $\mathcal{R}'$ and $\mathcal{R}$.

Let that "original" wave function be $\psi(P,t)$ as usual, and let Q be a general point on $\mathcal{S}$, so that $\psi(Q,t)$ and $\psi_n(Q,t)$ are respectively the wave function and its normal derivative at $\mathcal{S}$. Now *suppose* that we modify Case 1 by adding sources on $\mathcal{S}$ which *somehow* impose the following saltus conditions: as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$ at point Q, the wave function changes by $-\psi(Q,t)$ and its normal derivative changes by $-\psi_n(Q,t)$; or, equivalently, as we cross $\mathcal{S}$ in the other direction, the wave function changes by $+\psi(Q,t)$ and its normal derivative changes by $+\psi_n(Q,t)$. By Assumption 1, the wave function is initially null in $\mathcal{R}$. When the first disturbance arrives at Q from the $\mathcal{R}'$ side, it is nullified on the $\mathcal{R}$ side by the saltus condition, and (by the sufficiency of the zero boundary conditions) exerts no further influence in $\mathcal{R}$—or in $\mathcal{R}'$, except by the same saltus condition in the reverse direction, which merely restores the original boundary conditions on the $\mathcal{R}'$ side from the null conditions on the $\mathcal{R}$ side, giving a perfect imitation of a continuation of the medium (Proposition 3), hence no reflection or echo. This situation continues indefinitely. In summary:

Case 2: We have the original sources in $\mathcal{R}'$ plus sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the step-change in the wave function and its normal derivative are respectively *minus* the original wave function on $\mathcal{S}$ and *minus* the original normal derivative on $\mathcal{S}$. We get the original wave function in $\mathcal{R}'$ and a null wave function in $\mathcal{R}$.[5]

Changing the signs in Case 2 and superposing Case 1, we obtain:

Case 3: We have only sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the step-changes in the wave function and its normal derivative are respectively equal to the original wave

[4] Alternatively, initial conditions on the displacement and velocity determine the two "constants of integration" as functions of position, making the integrals unambiguous.

[5] Obviously this piecewise-defined wave function is not strictly differentiable at $\mathcal{S}$. However, Assumption 3 refers to the *original* wave function, not necessarily the one modified by sources on $\mathcal{S}$. For the latter, we can still take limits of the wave function and its normal derivative as we approach $\mathcal{S}$ from either side, and treat these limits as the boundary conditions.

> function on $\mathcal{S}$ and its normal derivative on $\mathcal{S}$. We get a null wave function in $\mathcal{R}'$ and the original wave function in $\mathcal{R}$.[6]

Cases 2 and 3 respectively establish the following two propositions:

Proposition 4 *Let space be divided into a region $\mathcal{R}$ containing no sources, a region $\mathcal{R}'$, and a surface $\mathcal{S}$ separating the two. Let "primary" sources in $\mathcal{R}'$ give a certain "original" wave function. Now let the primary sources be supplemented by a distribution of sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the changes in the wave function and its normal derivative are respectively equal to* minus *the original wave function on $\mathcal{S}$, and* minus *the original normal derivative on $\mathcal{S}$. Then we get the original wave function in $\mathcal{R}'$ and a null wave function in $\mathcal{R}$.*

Proposition 5 *Let space be divided into a region $\mathcal{R}$ containing no sources, a region $\mathcal{R}'$, and a surface $\mathcal{S}$ separating the two. Let "primary" sources in $\mathcal{R}'$ give a certain "original" wave function. Now let the primary sources be* replaced *by a distribution of "secondary" sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the changes in the wave function and its normal derivative are respectively equal to the original wave function on $\mathcal{S}$ and its normal derivative on $\mathcal{S}$. Then we get a null wave function in $\mathcal{R}'$ and the original wave function in $\mathcal{R}$.*

Obviously the original wave function in $\mathcal{R}$ includes the original boundary conditions at the $\mathcal{R}$ side of $\mathcal{S}$, for both the wave function and its normal derivative. So, under the hypotheses of Proposition 5, the two saltus conditions become boundary conditions at the $\mathcal{R}$ side of $\mathcal{S}$. Either of these BCs, by Proposition 2, would be sufficient to determine the wave function in $\mathcal{R}$. But the attainment of this "sufficient" BC is partly due to the null wave function in $\mathcal{R}'$, which ensures that the BC is not contaminated by waves propagating across across $\mathcal{R}'$ from remote points on the $\mathcal{R}'$ side of $\mathcal{S}$, and which has been obtained by imposing saltus conditions corresponding to *both* BCs (Proposition 5). Thus in practice it helps to know both, although in principle it suffices to know one or the other. That is the more nuanced answer to Q1.

With regard to Q2, we are yet to identify the distribution of sources that yields the desired saltus conditions on $\mathcal{S}$. If we succeed in doing this, we shall have identified secondary sources on $\mathcal{S}$ which, by themselves, give the same wave function in $\mathcal{R}$ as the primary sources, showing that the wave function in $\mathcal{R}$ is *as if* the incident waves turned the surface $\mathcal{S}$ into a particular distribution of secondary sources replacing the primary sources; in other words, for field points in $\mathcal{R}$, we shall have verified and quantified **Huygens' principle** for each point on the surface.[7] Moreover, as the desired saltus conditions give a null wave function in $\mathcal{R}'$, we shall have found secondary sources that cause no "backward" (or "retrograde") secondary waves, ensuring that no such waves can propagate across $\mathcal{R}'$ to another point on $\mathcal{S}$ and thence into $\mathcal{R}$, adding to the wave function in $\mathcal{R}$. The same requirement is expressed in the following statement by Joseph Larmor, in which his "inside" or "interior" is our $\mathcal{R}'$, and his "outside" is our $\mathcal{R}$:

> 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

[6] *Cf.* Miller [17, s.3]. Equivalently, if we change the signs of the surface sources in Case 3 and add back the original sources in Case 1, we recover Case 2; *cf.* Larmor [13], at p.11—without his sources "m_B" (in the region corresponding to our $\mathcal{R}$), which tend to obscure the relevance of his work in our context. Larmor's treatment, however, does not seem to include any step analogous to our Proposition 3.

[7] Notice that the desired secondary sources are *not* segments of the moving wavefronts, but segments of a stationary surface influenced by the passing waves. Compare Huygens' original statement: "that *each particle of matter* in which a wave spreads, ought not to communicate its motion only to the next particle which is in the straight line drawn from the luminous point, but that it also imparts some of it necessarily to all the others which touch it and which oppose themselves to its movement. So it arises that around each particle there is made a wave of which *that particle* is the centre" [9, p.19, emphasis added]. Huygens chooses secondary sources on the same primary wavefront at the same time for the purpose of constructing the "continuation" of the wavefront (the same wavefront at a later time) in the same medium [9, pp. 19, 50–51], but *not* for the purpose of constructing a wavefront reflected or refracted at an interface between two media; for the latter purpose, he chooses secondary sources at various points on the reflecting or refracting surface, although the primary wavefront reaches those points at various times [9, pp. 23–4, 35–7, etc.].

> be an alteration of the physical circumstances, would in fact *add to* and confuse the effect of the assigned true sources inside [14, p.172; emphasis added].

Further on Q2, suppose that we have an aperture (e.g., a slit or a circular hole) in an otherwise opaque screen (also called a *baffle*), with a source or sources on one side, 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 baffle will not be perfectly sharp; even if there is only a single point-source, the shadow 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, called $\mathcal{S}_a$, spans the aperture, while the remaining part, called $\mathcal{S}_b$, is on the baffle, on the dark side or quiet side (the side facing $\mathcal{R}$); the subscripts may be remembered as a for *aperture* and b for *baffle*. Then, by Proposition 2, the boundary conditions at the $\mathcal{R}$ sides of $\mathcal{S}_a$ and $\mathcal{S}_b$, on the wave function *or* its normal derivative, determine the wave function in the region $\mathcal{R}$ beyond the baffle; and by Proposition 5, these BCs can be set by secondary sources on $\mathcal{S}$ that impose the corresponding saltus conditions on the wave function *and* its normal derivative, with no other sources, and no baffle (since any contribution from the baffle is treated as a primary source and therefore replaced). As a first approximation, we might suppose that the required secondary sources on $\mathcal{S}_a$ are the same as if the baffle were not there, and that the required secondary sources on $\mathcal{S}_b$ are null; in other words, we might suppose that the baffle simply eliminates the secondary sources on the corresponding part of $\mathcal{S}$. Thus we might hope to explain and predict diffraction in a reasonably accurate *quantitative* manner.

In principle, the answer to Q3 is now clear: to calculate the wave function in $\mathcal{R}$, we simply add up the contributions from the secondary sources on $\mathcal{S}$. These sources are those which impose the required saltus conditions on the wave function and its normal derivative, namely the saltus conditions consistent with the desired boundary conditions at the $\mathcal{R}$ side of $\mathcal{S}$ with null BCs on the other side. It remains to determine the required secondary sources, and thence the sum of their contributions.

1.2 Secondary sources matching saltus conditions

If $\mathcal{S}$ is a sheet of sources, then, as we cross $\mathcal{S}$ at any point Q, the saltus in the wave function ψ or its normal derivative ψ_n is not influenced by any sources at finite (i.e., non-infinitesimal) distances from Q, on $\mathcal{S}$ or elsewhere, because those distances, and hence the effects of the corresponding sources, change by only infinitesimal fractions as we cross $\mathcal{S}$; thus the saltus at any point on $\mathcal{S}$ is due solely to the concentration(s)[8] of sources on an infinitesimal surface element $d\mathcal{S}$ *at that point* (*cf.* Larmor [13], p.6). The boundary conditions generally *are* influenced by sources at finite distances; but these influences are the same on each side of $\mathcal{S}$ and consequently do not affect the saltus, which is the *difference* between the BCs on the two sides. In summary, the saltus conditions imposed by a sheet of sources, unlike the boundary conditions, are strictly **local**. In this sense the saltus conditions are more fundamental, more characteristic of the sources (rather than the geometry of $\mathcal{S}$), and easier to calculate from, or express as, a distribution of sources.

The effects of sources obviously depend on the medium in which they are immersed—about which we have so far assumed nothing. From here on, we confine our attention to the simplest kind of medium:

Assumption 4 *The medium is* ***homogeneous*** *and* ***isotropic****, and perfectly transparent to the waves.*

A homogeneous medium (also called a *uniform* medium) is one whose properties are independent of *location* within the medium. An isotropic medium is one whose properties are independent of *direction*. The properties are likewise said to be homogeneous (uniform) and isotropic, respectively.

We can now introduce the simplest type of source, which will serve as the basic component of more complicated types. Whereas a wave function in a three-dimensional medium is generally a function of *three* spatial coordinates and time, the wave function due to the simplest type of source in the simplest type of medium can be described with only *one* spatial coordinate and time:

[8] Plural if there is more than one kind of source.

Assumption 5 *A **monopole** source with **strength** $f(t)$ generates the wave function*

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

where t is time, r is the radial distance from the source, and c is a constant.

So the *strength* of the source is a function of time. In expression (1), to satisfy the earlier Assumption 3, *the function f must be differentiable*. If (1) is to be valid for arbitrarily small distances—as we shall require, at least for secondary sources—then the source must be of infinitesimal size. Therefore, at least for secondary sources, *a "monopole" will be a point-source or a source of infinitesimal extent*.

Assumption 5 not only defines the strength $f(t)$ in terms of the wave function that it causes,[9] but also asserts four properties of the wave function. First, at distance r from the source, the argument of f is changed from t to $t-r/c$, so that t must be increased—that is, *delayed*—by r/c in order to obtain the same value of the argument, hence the same value of f, as at the source. So r/c is the time taken for the influence of the source to travel the distance r (for all r), whence *c is the propagation speed*.[10] Second, due to the factor $1/r$, the wave function at distance r from the source is *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 for a given value of f satisfies the *inverse-square law*. Third, there are surfaces over which the wave function is uniform. These surfaces, which clearly constitute **wavefronts**, are the surfaces of constant r, i.e. the *spheres* centered on the source. Fourth, the form of the function f (the "waveform") does not change as the wave propagates. This means either (a) that the medium is special in the sense that c is indifferent to the functional form of f, so that if we express f as a sum of components, those components propagate at the same speed and maintain their sum, or (b) that f is special in the sense that it maintains its form as it propagates, notwithstanding the quirks of the medium. For most of this paper we shall assume (a), in which case the medium is described as **non-dispersive**. In summary, the wave function (1) describes *spherical waves* which recede from the source at speed c (maintaining the waveform), and whose squared amplitudes satisfy the inverse-square law.[11]

Now let us investigate the saltus conditions imposed by a continuous sheet of monopoles. Let $M(t)$ be the *strength density* of monopole secondary sources on the surface $\mathcal{S}$, so that the source strength corresponding to a surface element of area $d\mathcal{S}$ is $M(t)\,d\mathcal{S}$. In general, M depends on the position of the element and on time, but only the time-dependence is shown explicitly. By formula (1), the element's contribution to the wave function at a point N, at a distance σ from the element,[12] is

$$d\psi = \frac{1}{\sigma}\, M\big(t - \sigma/c\big)\, d\mathcal{S}\,. \tag{2}$$

Let n be the normal coordinate of N—that is, the perpendicular distance of N from $\mathcal{S}$, with a positive sign if N is on the $\mathcal{R}$ side (as in Fig. 1), and a negative sign if it is on the $\mathcal{R}'$ side. For the purpose of finding the saltus in ψ or ψ_n as N crosses $\mathcal{S}$, we may let σ be so small that $\mathcal{S}$ can be considered flat (as in Fig. 1) and the variation in $M\big(t-\sigma/c\big)$ with σ can be neglected, whether it is due to the variation of σ in the argument of the function M (*cf.* [13], p. 5) or the variation of that function with position on $\mathcal{S}$ (which is why we show only the time-dependence of M); and we may let n be small compared with the permitted range of σ. Equation (2) then becomes

$$d\psi = \frac{1}{\sigma}\, M(t)\, d\mathcal{S}\,. \tag{3}$$

---

[9] This definition follows the old convention used by (e.g.) Baker and Copson [1, p. 42], Born and Wolf [2, p. 421], and Larmor [13, p. 5]. Miller (in [16], p. 1371, and [17], s. 3) uses the denominator $4\pi r$ instead of our r. Fortunately the effects of this clash of conventions will usually cancel out.

[10] The symbol c comes from the Latin *celeritas*, meaning speed. Another way to perceive the wave-like nature of (1) is to let the function f have a certain feature at a certain value of its argument. Setting the argument equal to that value (a constant) and differentiating with respect to t, we get $\frac{dr}{dt} = c$, showing that the "feature" travels in the r direction with speed c.

[11] We could add that because c does not depend on the distance or direction from the source or on the direction of propagation, c is homogeneous and isotropic, and that because the intensity at distance r is inversely proportional to the area of the surface at that distance, there is no absorption of energy by the medium. But these things are already implied by Assumption 4.

[12] We call this distance σ instead of r, because we are reserving r for a distance from a *primary* source.

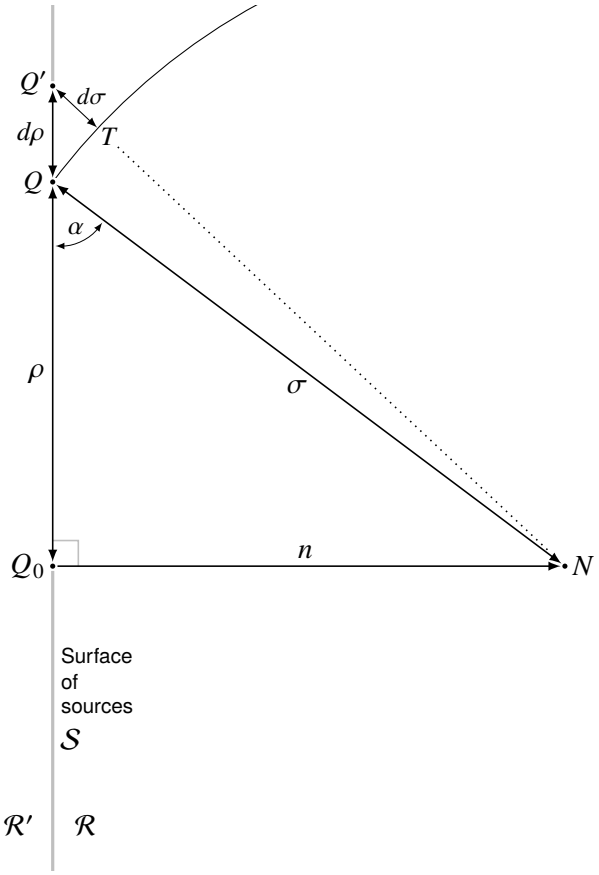

Fig. 1: If σ is sufficiently small, the surface $\mathcal{S}$ (seen edge-on) can be considered flat, and the surface element $d\mathcal{S}$ can be chosen as an annulus with axis Q_0N, inner radius ρ, and width $d\rho$. The arc QT is centered on N.

For such small σ and n, it is clear from the symmetry that changing the sign of n causes no change in σ, so that there is no saltus in $d\psi$ as N crosses $\mathcal{S}$. And for larger σ, there is no saltus in $d\psi$ because $d\mathcal{S}$ is not local. So, as N crosses $\mathcal{S}$, there is no saltus in the contribution to ψ from any $d\mathcal{S}$, near or far, hence no saltus in the total wave function ψ.

For the normal derivative ψ_n, however, the same symmetry has a different effect: changing the sign of n causes no change in the contribution to the normal derivative in the direction *away from* $\mathcal{S}$; but this is the n direction on the $\mathcal{R}$ side, and the $-n$ direction on the $\mathcal{R}'$ side, so that the contribution in the n direction changes sign. To find that contribution, we differentiate (3) w.r.t. n by the chain rule, obtaining

$$d\psi_n = -\frac{1}{\sigma^2}\, M(t)\, \frac{d\sigma}{dn}\, d\mathcal{S}\,. \tag{4}$$

In Fig. 1, by Pythagoras, for both signs of n,

$$\rho^2 + n^2 = \sigma^2. \tag{5}$$

Differentiating this w.r.t. n for constant ρ gives $2n = 2\sigma\frac{d\sigma}{dn}$, whence $\frac{d\sigma}{dn} = \frac{n}{\sigma}$, so that (4) becomes

$$d\psi_n = -\frac{n}{\sigma^3}\, M(t)\, d\mathcal{S}\,. \tag{6}$$

For the supposed *small* values of σ and n, 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 of this element have the same value of σ in (6). The area of this element is

$$d\mathcal{S} = 2\pi\rho\, d\rho\,, \tag{7}$$

so that (6) becomes

$$d\psi_n = -\frac{2\pi n}{\sigma^3}\, M(t)\, \rho\, d\rho\,. \tag{8}$$

For integration, we need to express this in terms of σ or ρ, not a mixture of the two. Differentiating (5) w.r.t. ρ for constant n gives $2\rho = 2\sigma\,\frac{d\sigma}{d\rho}$, hence[13]

$$\rho\, d\rho = \sigma\, d\sigma\,, \tag{9}$$

which may be substituted into (8) to eliminate ρ and obtain $d\psi_n$ in terms of σ:

$$d\psi_n = -2\pi n\, M(t)\, \frac{1}{\sigma^2}\, d\sigma\,. \tag{10}$$

According to the geometry of Fig. 1, the distance σ would range from $|n|$ to ∞ (where the absolute value allows N to be on either side of $\mathcal{S}$, with n measured *into* $\mathcal{R}$), so that the wave function at N would be

$$\psi_n = 2\pi n\, M(t) \int_{|n|}^{\infty} \frac{-1}{\sigma^2}\, d\sigma = 2\pi n\, M(t)\, \frac{1}{\sigma}\bigg|_{\sigma=|n|}^{\sigma\to\infty} = -2\pi\, M(t)\, \frac{n}{|n|} = \begin{cases} -2\pi\, M(t) & \text{for } n>0 \\ 2\pi\, M(t) & \text{for } n<0\,. \end{cases} \tag{11}$$

I say "According to..." and "would..." because the last step (11) puts $\sigma\to\infty$, whereas "the geometry of Fig. 1", like the rest of the preceding argument, requires σ to be *small*. So the value of ψ_n given by (11) is liable to be inaccurate, due to the influence of sources on remote parts of $\mathcal{S}$.

But, in the evaluation of the *saltus* in ψ_n as we cross $\mathcal{S}$, errors in ψ_n due to remote sources do not matter, because at every point on $\mathcal{S}$, these errors are the same on both sides and cancel out. According to (11), as n changes from 0^- to 0^+, the normal derivative ψ_n changes from $2\pi M(t)$ to $-2\pi M(t)$; and the step-change between these values is correct even if the values themselves are not. The step-change is[14]

$$\Delta\psi_n = -4\pi\, M(t)\,. \tag{12}$$

We still need to find a distribution of sources that gives a saltus in ψ. A single sheet of monopoles cannot do this. But now suppose that for every element of $\mathcal{S}$, with area $d\mathcal{S}$, on which the monopole source strength is $M(t)\,d\mathcal{S}$, there is a parallel element, displaced to the $\mathcal{R}'$ side by an infinitesimal perpendicular distance h (that is, at $n=-h$), on which the monopole source strength is $-M(t)\,d\mathcal{S}$. We shall call these sources the *uninverted monopole* and the *inverted monopole*, respectively. Thus there are two parallel sheets of monopole sources with spacing h: the *uninverted sheet* at $n=0$ (coinciding with $\mathcal{S}$), and the *inverted sheet* at $n=-h$.

According to (11), the contribution to ψ_n from either one of these sheets is independent of the distance from the sheet, at least for small distances. It follows that the contributions to ψ_n from the two sheets cancel out as $n\to-h$ from the $\mathcal{R}'$ (negative) side, and as $n\to0$ from the $\mathcal{R}$ side, so that the saltus in ψ_n due to the combined sheets is zero. As usual, the saltus is accurate even if the boundary values given by (11) are not. In this case, however, the boundary values are as accurate as the saltus, in spite of the inaccuracy of the contributions from remote sources on each sheet, because those sources are in pairs with equal and opposite, infinitesimally-spaced elements, so that their contributions cancel at finite distances.

This knowledge of the boundary values of ψ_n on the outer sides of the two sheets is the key to finding the overall saltus in ψ itself. By (12), as we cross the uninverted sheet into $\mathcal{R}$, the normal derivative ψ_n changes by $-4\pi M(t)$. As its value on the $\mathcal{R}$ side is zero, its value between the sheets is $4\pi M(t)$. And if we consider the zero value of ψ_n for $n<h$ and the saltus at the inverted sheet, we get the same value between the sheets. Multiplying that value by h gives the change in ψ as we cross the two sheets into $\mathcal{R}$:

$$\Delta\psi = 4\pi h\, M(t)\,. \tag{13}$$

[13] Alternative explanation: 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 $Q'QT$ are equal, both being complementary to α. So the triangles are similar, with $d\rho/d\sigma = \sigma/\rho$, hence (9).

[14] *Cf.* Larmor [13], p.6, second paragraph.

Now let us choose

$$M(t) = \frac{D(t)}{h} \,, \tag{14}$$

so that the uninverted monopole on the surface element of area $d\mathcal{S}$ has the strength $D(t)\,d\mathcal{S}/h$, while its inverted counterpart, displaced by the distance h in the $-n$ direction, has the strength $-D(t)\,d\mathcal{S}/h$. Together, these two sources are said to constitute a **dipole**, also called a *doublet*,[15] of strength $D(t)\,d\mathcal{S}$ in the n direction, so that $D(t)$ is the dipole strength per unit area of $\mathcal{S}$. Substituting (14) into (13), we find that as we cross this sheet of dipoles into $\mathcal{R}$, the change in ψ is[16]

$$\Delta\psi = 4\pi D(t) \,. \tag{15}$$

As $h \to 0$ and $M(t) \to \infty$ in (14), we may wonder whether we can maintain that $\psi_n = 0$ at the outer sides of the two sheets. But, as ψ_n between the sheets is also proportional to $M(t)$, this does not introduce any fractional error into (13) or (15).

Now by Proposition 5, the wave function in $\mathcal{R}$ due to primary sources in $\mathcal{R}'$ is reproduced if we replace the primary sources by a distribution of secondary sources on $\mathcal{S}$ such that $\Delta\psi_n$ and $\Delta\psi$, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, are equal to the original values of ψ_n and ψ on $\mathcal{S}$. To find these sources, we simply drop the Δ operators in (12) and (15) and solve for the strength densities, obtaining

$$M(t) = -\frac{\psi_n}{4\pi} \tag{16}$$

for the monopole sources, and

$$D(t) = \frac{\psi}{4\pi} \tag{17}$$

for the dipole sources. These monopole and dipole sources on $\mathcal{S}$, by themselves, give the original wave function (including original BCs) in $\mathcal{R}$ and a null wave function (including null BCs) in $\mathcal{R}'$.

A dipole source 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 (16) and (17) really 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 $\mathcal{R}'$.

1.3 Note on electromagnetic waves

Assumption 5 defines a *monopole* source of strength $f(t)$ as a source that generates the wave function

$$\frac{1}{r}\, f\big(t - r/c\big) \,, \tag{18}$$

where r is the distance from the source, *regardless of direction*. This definition is compatible with a *scalar*-valued function f. But if f is *vector*-valued, 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,[17]

$$d\varphi = \frac{1}{4\pi\epsilon_0 r}\, \varrho\big(t - r/c\big)\, dV \tag{19}$$

[15] The term *doublet*, which seems to be older, is used by Baker and Copson [1], Born and Wolf [2, p. 421], and Larmor [13]. This ordinary dipole or doublet, also called a *spatial* dipole, is not to be confused with a *spatiotemporal* dipole [16], which we discuss in Section 3.7.

[16] *Cf.* Larmor [13], p. 6, third paragraph.

[17] *Cf.* Feynman *et al.* [5], vol. 2, Chapter 15, Table 15-1.

and

$$d\mathbf{A} = \frac{1}{4\pi\epsilon_0 c^2 r}\,\mathbf{J}\big(t - r/c\big)\,dV\,, \tag{20}$$

where ϵ_0 is a physical constant called the electric permittivity of a vacuum, and r is the distance from dV *regardless of direction*. Thus the electromagnetic wave functions $d\varphi$ and $d\mathbf{A}$ indeed have the form (18), with the strength functions $\varrho(t)\,dV/4\pi\epsilon_0$ and $\mathbf{J}(t)\,dV/4\pi\epsilon_0 c^2$ respectively.[18]

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 in the present context, in which we have just disclaimed any interest in whether the sources specified by (16) and (17) *could* exist, we need not pursue this issue any further.[19]

2 Helmholtz and Kirchhoff integral theorems

2.1 Derivation

As foreshadowed in the answer to Q3, we must now calculate the wave function in $\mathcal{R}$ by adding up the contributions from the secondary sources on $\mathcal{S}$, whose strength densities $M(t)$ and $D(t)$ are given by (16) and (17). The general term in the sum is the contribution to the wave function at a general field point P in $\mathcal{R}$, due to the sources on a general surface element of area $d\mathcal{S}$, that element being of such small dimensions that it can be regarded as concentrated at a point; let it be at a distance s from P. 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 (16) and (17) are already consistent with that convention.

The element $d\mathcal{S}$ gives the monopole source strength $M(t)\,d\mathcal{S}$ at distance s from P. Its contribution to the wave function at P, according to formula (1), is

$$d\psi_M(P,t) = \frac{1}{s}\,M\big(t - s/c\big)\,d\mathcal{S}\,. \tag{21}$$

The same element $d\mathcal{S}$ gives the dipole source strength $D(t)\,d\mathcal{S}$ facing $\mathcal{R}$. This consists of one monopole of strength $D(t)\,d\mathcal{S}/h$, and another of strength $-D(t)\,d\mathcal{S}/h$, displaced from the first by a distance h in the $-n$ direction, where n is the normal coordinate from $\mathcal{S}$ into $\mathcal{R}$. The respective strengths *per unit area of* $\mathcal{S}$ are $D(t)/h$ and $-D(t)/h$. In Fig. 2, the contribution to the wave function at P from the monopole of strength $D(t)/h$, according to formula (1) again, is

$$\frac{D\big(t - s/c\big)}{hs}\,, \tag{22}$$

so that the contribution from *both* monopoles (that is, from the dipole) is the *change* in expression (22), due to the change in s, as n increases by h (from $-h$ to 0). As h is infinitesimal, the said change in expression (22) is given by

$$h\,\frac{\partial}{\partial n}\left(\frac{D\big(t - s/c\big)}{hs}\right)\,, \quad \text{i.e.} \quad \frac{\partial}{\partial n}\left(\frac{D\big(t - s/c\big)}{s}\right). \tag{23}$$

Multiplying this by $d\mathcal{S}$ gives the contribution to the wave function at P from the dipole sources on $d\mathcal{S}$:

$$d\psi_D(P,t) \;=\; \frac{\partial}{\partial n}\left(\frac{D\big(t - s/c\big)}{s}\right)d\mathcal{S}\,. \tag{24}$$

Adding (24) and (21), we find that the contribution from $d\mathcal{S}$ to the wave function at P, due to both types of sources, is

$$d\psi(P,t) \;=\; \left\{\frac{\partial}{\partial n}\left(\frac{D\big(t - s/c\big)}{s}\right) + \frac{1}{s}\,M\big(t - s/c\big)\right\}d\mathcal{S}\,, \tag{25}$$

[18] Version 0.2 of this paper omitted the factor dV from the stated strength functions, which were therefore per-unit-volume.

[19] For a context in which it matters whether the sources are realizable, see Miller [17], especially s.6.1.

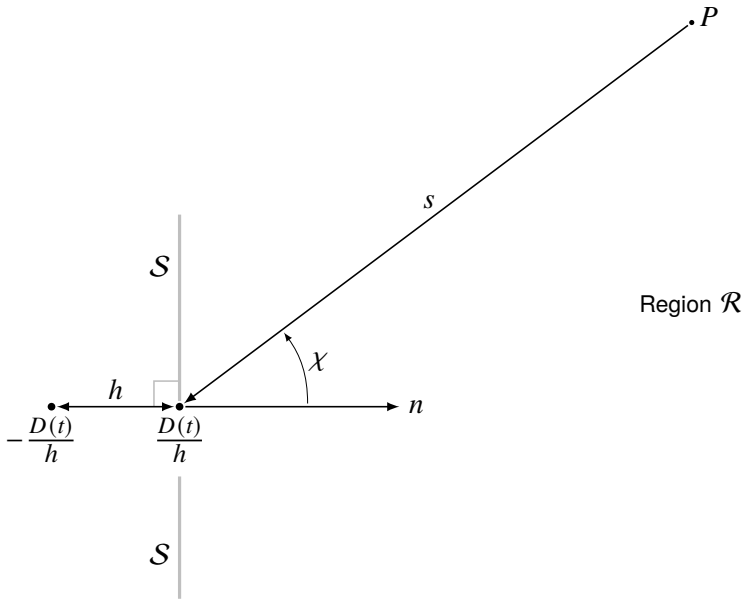

Fig. 2: Paired monopole sources per unit area of a dipole source distribution with strength density $D(t)$ facing $\mathcal{R}$. Gray lines show the orientation of the surface $\mathcal{S}$, which may be curved. The separation h is infinitesimal (so the diagram is not to scale). The normal coordinate n, measured from $\mathcal{S}$ into $\mathcal{R}$, increases from $-h$ to 0 as we move from the inverted source $-D(t)/h$ to the uninverted source $D(t)/h$.

which we then express in terms of the boundary conditions by substituting from (17) and (16): [20]

$$d\psi(P,t) = \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,. \tag{26}$$

The total wave function at P is the sum over the surface $\mathcal{S}$—i.e., the *surface integral over* $\mathcal{S}$—of all the contributions from the elements $d\mathcal{S}$, and is written

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,, \tag{27}$$

where the double integral sign acknowledges the two-dimensional range of integration. This equation (27) is the **Helmholtz formula** in a generalized form;[21] we shall call the right-hand expression the **Helmholtz integral**. Through its derivation, the integral has a simple physical interpretation: comparing (27) with (24) and (21), we see that the first term in the braces in the integrand (with the leading factor $1/4\pi$) represents dipole secondary sources of strength $\psi/4\pi$ per unit area, normal to $\mathcal{S}$, while the second term (with the same factor) represents monopole secondary sources of strength $\psi_n/4\pi$ per unit area.

The arguments of ψ and ψ_n in (27) allow for the propagation time *from* each secondary source. In Fig. 2, however, the arrow for s is drawn in the opposite direction, so that n and s can be construed as coordinates of the same point. This convention does not affect the *sign* of the coordinate s, which is an absolute distance and therefore positive; but it affects the *positive direction* of s, as we shall see when we introduce angles between positive directions of coordinates.

Equation (27) is for P in $\mathcal{R}$. If, on the contrary, P is in $\mathcal{R}'$, then the sources represented by the integral give a null wave function (Proposition 5), so that the integral is zero. Thus we have established the following:

[20] Continuing from footnote 9: If the factor 4π is included in the denominator of equation (1), it influences subsequent equations and eventually cancels itself out in (12) and in (15) to (17). But then it is needed again in (21) to (25), with the result that (26), and therefore (27) and its corollaries, are unchanged—as they must be, because they are expressed in terms of boundary conditions, not strengths. Equations (19) and (20) 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.

[21] The usual form of the Helmholtz formula [2, p.419, eq.7], which dates from 1859 [8, p.23, eq.7d], assumes that the wave function is a sinusoidal function of time. Our "generalized" form avoids this assumption.

Proposition 6 (Helmholtz integral theorem): *If $\psi(P,t)$ is a wave function and $\mathcal{R}$ is a region bounded by a surface $\mathcal{S}$ and containing no sources, then the Helmholtz integral over $\mathcal{S}$ (with the normal into $\mathcal{R}$) gives $\psi(P,t)$ if P is in $\mathcal{R}$, but zero if P is outside $\mathcal{R}$.*

Our derivation of this result takes more space than conventional ones,[22] but does not demand as much mathematical knowledge from the reader.

In Fig. 2, $D(t)$ does not depend on n, but s does (see after eq. 22 above). Hence, in the Helmholtz formula (27), in the first term in the braces, $\psi(t)$ is evaluated at $n=0$, while $\psi\big(t-s/c\big)$ varies with n because s does. But in the second term, both ψ_n and s are simply evaluated at $n=0$. This irregularity can be removed by performing the indicated differentiation in the first term:

$$\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) = \psi\big(t-s/c\big)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{1}{s}\,\frac{\partial}{\partial n}\psi\big(t-s/c\big) \tag{28}$$

$$= \psi\big(t-s/c\big)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'\big(t-s/c\big)\,\frac{\partial s}{\partial n}\,, \tag{29}$$

where the first equality follows from the product rule, and the second from two applications of the chain rule. Substituting (29) back into (27) yields

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\psi\big(t-s/c\big)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'\big(t-s/c\big)\,\frac{\partial s}{\partial n} - \frac{1}{s}\,\psi_n\big(t-s/c\big)\right\} d\mathcal{S}\,, \tag{30}$$

in which the right-hand expression is known as the **Kirchhoff integral**. As this is just another form of the Helmholtz integral,[23] Proposition 6 still applies, and may be restated as follows:

Proposition 7 (Kirchhoff integral theorem): *If $\psi(P,t)$ is a wave function and $\mathcal{R}$ is a region bounded by a surface $\mathcal{S}$ and containing no sources, then the Kirchhoff integral over $\mathcal{S}$ (with the normal into $\mathcal{R}$) gives $\psi(P,t)$ if P is in $\mathcal{R}$, but zero if P is outside $\mathcal{R}$.*

In the Kirchhoff form (30) of the integral, for a given surface element, s is fixed in the arguments of ψ, ψ', and ψ_n, so that (30) can be written

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\left[\frac{\partial\psi}{\partial t}\right]\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{31}$$

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*.[24] Now (31), like (30), contains derivatives of both s and $1/s$ w.r.t. the normal coordinate.[25] But if we apply the chain rule to the first term in the braces, take the common factor $1/s$ outside the braces, and use an overdot to denote (partial) differentiation with respect to t, we obtain the alternative form

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s}\left\{-\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot\psi]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{32}$$

[22] See e.g. Baker & Copson [1, pp. 23–4], who assume sinusoidal time-dependence, and whose normal coordinate ν is in the opposite direction from our n.

[23] In our expression of the Kirchhoff integral, we follow standard practice by *not* assuming that the wave function is a sinusoidal function of time. Had we made this assumption in deriving the Helmholtz integral, our subsequent derivation of the Kirchhoff integral would have been more difficult, because we would have needed to generalize the time-dependence in the Helmholtz integral (as in [1], pp. 36–7, or [2], pp. 420–21), or start over (as in [1], pp. 38–40).

[24] *Cf.* Born & Wolf [2] at pp. 420–21 (especially eq. 13). *Cf.* also Baker & Copson [1, p. 37] and Miller [16, 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.

[25] This feature descends from Kirchhoff himself; see [10, p. 669, eq. 9], where the arguments are explicit, and [11, p. 103, eq. 6], for which the arguments are specified in the subsequent text.

which gives the wave function at any point in $\mathcal{R}$, due to sources outside $\mathcal{R}$, in terms of the wave function and its derivatives and $\frac{\partial s}{\partial n}$ at the boundary surface. Again the path of derivation yields the physical interpretation: the last term in the braces, carried through from (27), represents the monopole secondary sources, while the terms in ψ and $\dot{\psi}$ represent the dipole sources, albeit in a less obvious manner than the first term in (27).[26] We shall find the last form (32) most convenient.

2.2 Application to diffraction by an aperture

Returning to Q2, let us now formalize our "first approximation" (p. 7):

Assumption 6 (secondary-source selection): *For an aperture in an opaque baffle, let the surface $\mathcal{S}$ be the union of two segments, namely $\mathcal{S}_a$ spanning the aperture, and $\mathcal{S}_b$ on the $\mathcal{R}$ side of the baffle. Then the baffle simply eliminates the secondary sources on $\mathcal{S}_b$, leaving the secondary sources on $\mathcal{S}_a$ as if the baffle were not present.*

This assumption, together with Proposition 5, implies that the wave function in $\mathcal{R}$ is found by adding the contributions from the said secondary sources on $\mathcal{S}_a$ only. That means evaluating the Helmholtz integral (27) or the Kirchhoff integral (32) over $\mathcal{S}_a$ instead of the whole of $\mathcal{S}$, with the integrand unchanged. Similarly, as any special case of the Helmholtz or Kirchhoff integral over $\mathcal{S}$ represents the sum of the contributions from certain secondary sources on $\mathcal{S}$, applying the "secondary-source selection" assumption to that case means restricting the summation to *part* of $\mathcal{S}$, with the secondary sources unchanged, and is represented by restricting the range of integration to that part of $\mathcal{S}$, with the integrand unchanged.

While restricting the range of integration to a surface spanning the aperture is indeed a standard method of calculating a wave function affected by diffraction, Assumption 6 is only one of its physical interpretations. There are at least three other interpretations that pass the test of internal consistency, and an older interpretation which fails that test; these are discussed in Appendices A and B.

The accuracy of Assumption 6, unlike that of our other Assumptions and Propositions, depends on the shapes of $\mathcal{S}$ and its segments: the claim that the baffle does not change the secondary sources on $\mathcal{S}_a$ is clearly not realistic if part of $\mathcal{S}_a$ lies in the geometric shadow of part of $\mathcal{S}_b$ w.r.t. any of the primary sources; and the claim that the secondary sources on $\mathcal{S}_b$ simply disappear is clearly not sufficient if the observation point lies in the geometric shadow of part of $\mathcal{S}_b$ w.r.t. any secondary sources on $\mathcal{S}_a$. The explicit adoption of Assumption 6 therefore entails the tacit assumption that we have made reasonable choices of $\mathcal{S}_a$ and $\mathcal{S}_b$ (including a reasonably shaped baffle). It also entails the assumption that no waves reflected from the $\mathcal{R}'$ side of the baffle find their through the aperture into $\mathcal{R}$; it is as if the baffle were perfectly *absorbent*—which, in the case of light waves, means not only opaque but also perfectly *black*.

3 Assumption: Single monopole primary source

So far, we have not assumed anything about the primary sources except that the resulting wave function had a beginning (Assumption 1), which implies that there was a time before which the primary sources were null. From here on, however, the following assumption will apply:

Assumption 7 *The only primary source is a monopole in $\mathcal{R}'$.*

3.1 Relating ψ_r and $\dot{\psi}$

Let the monopole primary source be at point O. Let the point N, associated with the general surface element $d\mathcal{S}$, lie at a distance r from the source, and at a distance s from the field point P (Fig. 3). Then, by Assumption 5, in the absence of any obstruction, the wave function at N has the form of (1):

$$\psi = \frac{1}{r} f\big(t - r/c\big) \,. \tag{33}$$

[26] *Cf.* Baker & Copson [1, pp. 42–3]; Born & Wolf [2, p. 421].

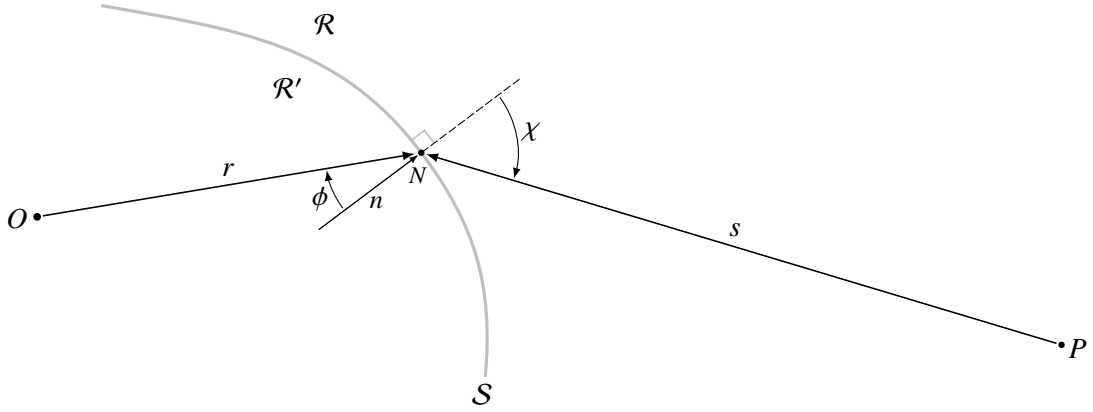

Fig. 3: Distances and angles pertaining to the calculation of the wave function at the field point P in region $\mathcal{R}$ due to a source at O in region $\mathcal{R}'$, by integration over the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$. Angles ϕ and χ are not generally coplanar. For the normal coordinate n, the positive direction is *into* $\mathcal{R}$. In general, n is measured from $\mathcal{S}$ to N; but the diagram depicts the case in which N is *on* $\mathcal{S}$ (that is, $n = 0$), and shows the destination, but not the origin, of n.

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

$$\frac{\partial\psi}{\partial r} = -\frac{\dot{\psi}}{c} - \frac{\psi}{r} \,. \tag{34}$$

In relation (34), which we shall cite below, the time-independent factors $1/c$ and $1/r$ suggest the need for a rule which we shall also cite hereafter (although the rule itself does not depend on Assumption 7). Recall that square brackets enclosing a function of time tell us to subtract a certain delay from t in the argument(s) 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. So the rule is: *Time-independent factors may be taken outside or inside square brackets*. Or, to put it more axiomatically, *delaying and scaling may be done in either order*.

3.2 Kirchhoff diffraction formula

For a given surface element, the distance s is fixed in the arguments of ψ and its derivatives in equation (30), but the resulting equation (32) still involves $\frac{\partial s}{\partial n}$, in which s is obviously allowed to vary with n. Hence we say above that the point N (in Fig. 3) is *associated with* the element $d\mathcal{S}$, rather than *on* it, in order to allow N and consequently r and s to vary with n; only if $n = 0$ will N be *on* $\mathcal{S}$ (as shown in Fig. 3). Treating n as an independent coordinate of N, we have by the chain rule

$$\frac{\partial\psi}{\partial n} = \frac{\partial\psi}{\partial r}\frac{\partial r}{\partial n} \,. \tag{35}$$

If we substitute this into (32) and take the time-independent factor $\frac{\partial r}{\partial n}$ outside the square brackets, the expression in braces in (32) becomes

$$\left\{ -\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot{\psi}]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial r}\right]\frac{\partial r}{\partial n} \right\} . \tag{36}$$

Now 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 N to P (in Fig. 3). Then, from the geometry,

$$\frac{\partial r}{\partial n} = \cos\phi \tag{37}$$

and

$$\frac{\partial s}{\partial n} = -\cos\chi \,, \tag{38}$$

where the derivatives are evaluated at $\mathcal{S}$ $(n=0)$. Substituting (34), (37), and (38) into (36), we get

$$\left\{\frac{1}{s}\,[\psi]\cos\chi + \frac{1}{c}\left[\dot{\psi}\right]\cos\chi + \left[\frac{\dot{\psi}}{c}+\frac{\psi}{r}\right]\cos\phi\right\}. \tag{39}$$

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, (39) can be written

$$\left\{\frac{1}{s}\,[\psi]\cos\chi + \frac{1}{c}\left[\dot{\psi}\right]\cos\chi + \left(\frac{1}{c}\left[\dot{\psi}\right]+\frac{1}{r}\,[\psi]\right)\cos\phi\right\} \tag{40}$$

or, regrouping the terms,

$$\left\{\frac{\cos\phi+\cos\chi}{c}\left[\dot{\psi}\right] + \left(\frac{\cos\phi}{r}+\frac{\cos\chi}{s}\right)[\psi]\right\}. \tag{41}$$

Putting this expression back in (32), we obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s}\left\{\frac{\cos\phi+\cos\chi}{c}\left[\dot{\psi}\right] + \left(\frac{\cos\phi}{r}+\frac{\cos\chi}{s}\right)[\psi]\right\} d\mathcal{S}\,. \tag{42}$$

Equation (42), usually with the surface $\mathcal{S}$ replaced by $\mathcal{S}_a$ as in Section 2.2, 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 analysis of diffraction). We call it the *exact* form to distinguish it from a widely-used approximate form: if (as is typically the case) the distances r and s are sufficiently large, 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 $\left[\dot{\psi}\right]$. That term contains the factor $\cos\phi+\cos\chi$, in which the angles ϕ and χ are measured from the normal to $\mathcal{S}$. Accordingly we might describe this factor, or something proportional to it, as the "obliquity factor". In particular, if we define the obliquity factor as

$$\tfrac{1}{2}\left(\cos\phi+\cos\chi\right), \tag{43}$$

it will have a maximum possible value of 1, for $\phi=\chi=0$ (no obliquity); and it will have a value of zero where ϕ and the supplement of χ are equal. The latter condition applies at geometric *points of reflection* off $\mathcal{S}$—but not only at those points, because reflection also requires the angles to be coplanar; in general, the obliquity factor is zero at any point Q on $\mathcal{S}$ such that O and P lie on a cone with its apex at Q and its axis normal to $\mathcal{S}$ at Q.

If r is "sufficiently large" in the above sense, we say that the surface element $d\mathcal{S}$ is in the **far field** of the primary source. If s is likewise "sufficiently large", we say that the observation point P is in the far field of the secondary sources on $d\mathcal{S}$. Neglecting the term in $[\psi]$ is a *far-field approximation* by both criteria.[27] This approximation is *not* equivalent to neglecting secondary sources of one type (monopole or dipole). The substitution (37) applies to the last term in (36), corresponding to the last term in (32), which represents the monopoles; and the substitution (38) applies to the other terms, which represent the dipoles. Thus, in formula (42), the terms in $\cos\phi$ represent the monopoles while the terms in $\cos\chi$ represent the dipoles, and the far-field approximation must still represent both kinds of secondary sources in order to give the required obliquity factor (43).

3.3 If $\mathcal{S}$ is a spherical primary 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

[27] 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 between path lengths of interfering waves. Diffraction under those conditions is called *Fraunhofer diffraction*. We are *not* using that meaning here.

$\mathcal{S}$ and the *normal* to the primary wavefront. So if $\mathcal{S}$ *is* a primary wavefront—or, more precisely, a fixed surface coinciding with successive primary wavefronts at successive times—we simply put $\phi = 0$ in (42), obtaining

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} \left[\dot{\psi}\right] + \left(\frac{1}{r} + \frac{\cos\chi}{s} \right) [\psi] \right\} d\mathcal{S} \,. \tag{44}$$

Equation (44) 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].

Again in practice the surface of integration tends to be limited to a segment spanning the aperture, as in Section 2.2. We call (44), like (42), the *exact* form to distinguish it from a widely-used far-field approximation: for sufficiently large r and s, we may again neglect the second term in the braces and retain only the term in $\left[\dot{\psi}\right]$, in which the obliquity factor (43) has become

$$\tfrac{1}{2}\left(1+\cos\chi\right) . \tag{45}$$

This has a maximum value of 1 for $\chi = 0$ (direct forward secondary waves), and a minimum of 0 for $\chi = 180°$ (backward secondary waves). In (45), as in (43), the $\cos\chi$ term corresponds to the dipole secondary sources.

For transverse waves, the factor $1+\cos\chi$ in expression (45) was derived from mechanical assumptions by George Gabriel Stokes, in a paper read in 1849 and printed in 1851.[28] For longitudinal waves, this factor is an obvious consequence of Fresnel's earlier suggestion that each secondary wave consists of a compression wave and a velocity wave which concur in the forward direction but cancel in the reverse direction [6, p.109, p.110*n*], although Fresnel himself, to my knowledge, never made this consequence explicit, presumably because of the rapid progress of his transverse-wave-based theory (*cf.* [6], p.109*n*).

3.4 If the primary wavefront is plane (or nearly so)

The curvature of the primary wavefront enters into the derivation of (42) 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 (42) and obtain

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

And for integration over a primary wavefront, we put $r \to \infty$ in (44), *or* $\phi = 0$ in (46), and obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} \left[\dot{\psi}\right] + \frac{\cos\chi}{s} [\psi] \right\} d\mathcal{S} \,. \tag{47}$$

The last two results are good approximations if r is so large that ψ/r, in equations (34) to (44), is very small compared with $\dot{\psi}/c$, whether s is large or not; that is, 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. Being exact for plane waves and good approximations for large r, they should also be good approximations for primary wavefronts with *non-spherical* curvature[29] if the curvature is sufficiently gradual—that is, if the wavefronts are sufficiently "nearly" plane. In such cases, ϕ is to be understood as the angle between the normals of the primary wavefront and the surface of integration.

[28] Stokes [22], p.31, eq. (45); his θ is our χ. For a review see Buchwald & Yeang [4], pp. 469–72.

[29] 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.

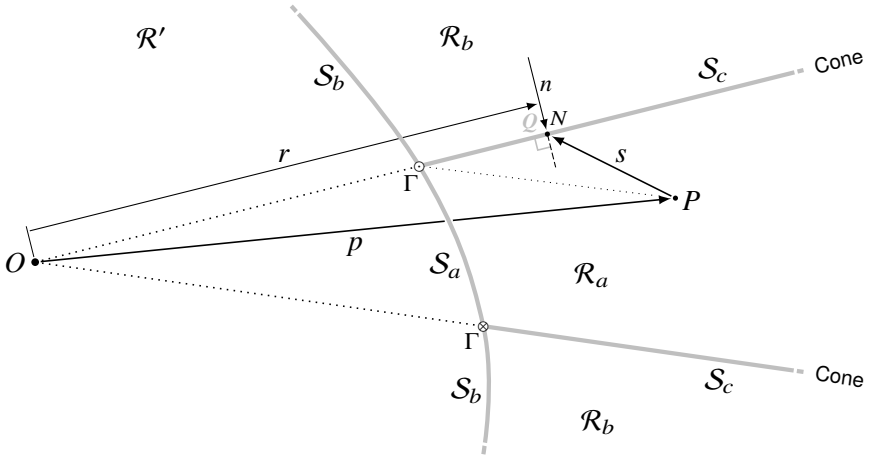

Fig. 4: Cross-section of the conical surface $\mathcal{S}_c$ projected by point O from the curve Γ, which divides the surface $\mathcal{S}$ (not labeled) into segments $\mathcal{S}_a$ and $\mathcal{S}_b$ (shown in cross-section). The cone divides the region $\mathcal{R}$ (not labeled) into subregions $\mathcal{R}_a$ 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 distances n and s (and r, in the present Section 3.5) are coordinates of N; here, as in Fig. 3, we depict the case in which $n = 0$, and show the destination, but not the origin, of n. (The point Q and the dotted line from P to Γ are not needed until Appendix C.)

3.5 The Maggi-Rubinowicz transformation of the Helmholtz integral over $\mathcal{S}_a$

By formula (27), the wave function at a field point P due to a set of (primary) sources is given by the Helmholtz integral over any surface $\mathcal{S}$ that divides the space into two regions, one containing P and the other containing all the sources. Thus the integral is the same for any surface $\mathcal{S}$ that separates P from the sources. In particular, if a segment of $\mathcal{S}$, say $\mathcal{S}_a$, is allowed to vary while the rest of $\mathcal{S}$, say $\mathcal{S}_b$, is fixed, then the integral over $\mathcal{S}$ is the same for all choices of $\mathcal{S}_a$ that maintain the separation. It follows (by subtraction of the integral over $\mathcal{S}_b$) that the integral over $\mathcal{S}_a$ *alone* is the same for all choices of $\mathcal{S}_a$ that maintain the separation, although these choices of $\mathcal{S}_a$ may differ everywhere except at their common edge, where they meet $\mathcal{S}_b$. In other words, given the sources and the separation requirement, and given a specification of the edge (also indicating which side of it is the $\mathcal{S}_a$ side), the Helmholtz integral over the segment $\mathcal{S}_a$ (that is, according to Assumption 6, the wave function admitted through an aperture spanned by $\mathcal{S}_a$) depends only on its edge: *same edge* $\Longrightarrow$ *same integral*. The integral over $\mathcal{S}_a$ therefore ought to be convertible to an expression involving only the (primary) sources and the edge. We shall perform the conversion under Assumption 7—that is, for a single monopole primary source.

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 $\mathcal{S}_b$; thus Γ is the edge (also called the *rim*) of $\mathcal{S}_a$. Now we make a new construction: let the surface $\mathcal{S}_c$ be the *geometric shadow of the curve* Γ w.r.t. the source at O; in other words, let $\mathcal{S}_c$ be the infinite cone (not generally circular) passing through Γ from the apex O, with its tip cut off at Γ (Fig. 4). Consistent with the "tacit assumption" that the choices of $\mathcal{S}_a$ and $\mathcal{S}_b$ are "reasonable" (Section 2.2), we assume that the only place where $\mathcal{S}$ intersects or touches the cone is 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 $\mathcal{S}_a$ spans the aperture and $\mathcal{S}_b$ is the $\mathcal{R}$ side of the baffle, 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 baffle (and the subscripts are a for *aperture*, b for *baffle*, and c for *cone*).

For brevity, let us represent the integrand in (27) by H (for *H*elmholtz). Then, applying Proposition 6 to the region $\mathcal{R}_a$ bounded by the surface comprising $\mathcal{S}_a$ and $\mathcal{S}_c$, for the single source at O, we have

$$\iint_{\mathcal{S}_a} H\,d\mathcal{S} + \iint_{\mathcal{S}_c} H\,d\mathcal{S} = \begin{cases} \psi(P,t) & \text{if } P \text{ is in } \mathcal{R}_a \\ 0 & \text{if } P \text{ is in } \mathcal{R}_b \text{ or } \mathcal{R}'. \end{cases} \tag{48}$$

According to Assumption 6, the first term on the left (the integral over $\mathcal{S}_a$) is the wave function admitted through an aperture spanned by $\mathcal{S}_a$; let us denote this by $\psi_{(a)}(P,t)$. On the right, $\psi(P,t)$ is the *unobstructed* wave function. So, if we now ignore $\mathcal{R}'$ and limit our attention to $\mathcal{R}_a$ and $\mathcal{R}_b$ (that is, to $\mathcal{R}$), the right-hand side of (48) is the wave function in $\mathcal{R}$ as predicted by *geometrical* optics; let us denote this by $\psi_{(g)}(P,t)$. If we then define

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} H\,d\mathcal{S}\,, \tag{49}$$

equation (48) can be abbreviated and rearranged as

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) + \psi_{(d)}(P,t)\,, \tag{50}$$

showing that $\psi_{(d)}(P,t)$ is the effect of *diffraction* (hence the subscript); it is what must be added to the wave function as predicted by geometrical optics, in order to obtain the wave function admitted through the aperture as predicted by Assumption 6.

In (50), the geometric term $\psi_{(g)}(P,t)$ is obviously consistent with our expectation that the wave function admitted through the aperture should depend only on the primary source(s) and edge of the aperture. The correction term for diffraction (49) is also consistent with that expectation, since the surface of integration ($\mathcal{S}_c$) is the geometric shadow of the edge (Γ) w.r.t. the only primary source.

Moreover, as we shall now show, the integral over the surface $\mathcal{S}_c$ in (49) can be transformed to an integral over the curve Γ. Rubinowicz [20, pp. 257, 262–3] interpreted the transformation as a revival and revision of Thomas Young's idea [23, 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.

We shall derive this "edge integral" or "rim integral" in a form like Rubinowicz's—approximately at first, and exactly in Appendix C. When the integrand H is written out in full, equation (49) becomes

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n\big(t-s/c\big)\right\} d\mathcal{S}\,. \tag{51}$$

But, because $\mathcal{S}_c$ is conical with its apex at O, and because the only source is a monopole at O, so that ψ (being the unobstructed wave function at N) depends on the distance r but not on its direction, we have

$$\psi_n = 0\,, \tag{52}$$

eliminating the second term in (51); as regards secondary sources, that means *eliminating the monopoles and retaining the dipoles*. If we then put the first term in square-bracket notation, (51) is reduced to

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\,\frac{\partial}{\partial n}\left(\frac{[\psi]}{s}\right) d\mathcal{S}\,. \tag{53}$$

Now if τ is the total propagation time from O to P via N, we can apply the chain rule, obtaining

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{1}{4\pi}\,\frac{d}{d\tau}\left(\frac{[\psi]}{s}\right)\frac{\partial\tau}{\partial n}\,d\mathcal{S}\,. \tag{54}$$

Here the differentiation w.r.t. τ is for movement in the n direction, in which r is constant (see Fig. 4), so the variation in τ is due solely to the variation in s. In that situation we have $d\tau = ds/c$, so that

$$\frac{\partial\tau}{\partial n} = \frac{1}{c}\,\frac{\partial s}{\partial n} = \frac{1}{c}\cos(n,s)\,, \tag{55}$$

where (n,s) means the angle between the positives directions of n and s (and the cosine function encourages this interpretation). Substituting (55) into (54), we obtain

$$\psi_{(d)}(P,t) \;=\; -\iint\limits_{\mathcal{S}_c} \frac{\cos(n,s)}{4\pi c}\,\frac{d}{d\tau}\left(\frac{[\psi]}{s}\right) d\mathcal{S}\,. \tag{56}$$

Now we introduce a *short-wavelength approximation*—not that it makes any difference to the result (see Appendix C), but because it shortens the derivation. If, as is typically the case, the wave function undulates over distances ("wavelengths") that are very small compared with r and s, then the argument $[\psi]/s$ does not significantly vary with small changes in r and s, except through the consequent changes in the total propagation delay τ. It follows that the derivative w.r.t. τ in (56) does not significantly change if, instead of holding r constant and varying n, we hold n constant and vary r—in other words, if the point N in Fig. 4, instead of moving perpendicular to the cone, moves along a generating line of the cone. Under the latter conditions, τ increases by $d\tau$ as r increases by dr, so that (56) can be written

$$\psi_{(d)}(P,t) = -\iint_{\mathcal{S}_c} \frac{\cos(n,s)}{4\pi c}\,\frac{1}{d\tau}\left\{\frac{[\psi]}{s}\bigg|_{r+dr} - \frac{[\psi]}{s}\bigg|_{r}\right\} d\mathcal{S}\,. \tag{57}$$

In the summation over the conical surface $\mathcal{S}_c$, the second term in the parentheses for one value of r cancels with the first term for that value minus dr, except on the part of the cone for which $r-dr$ is out of bounds. That part, which we shall call $\delta\mathcal{S}_c$, is the band of width dr adjacent to the curve Γ, where the cone meets the rim of the aperture. Thus the integral over $\mathcal{S}_c$ reduces to the integral over $\delta\mathcal{S}_c$ without the first term in the parentheses; that is,

$$\psi_{(d)}(P,t) = \iint_{\delta\mathcal{S}_c} \frac{\cos(n,s)}{4\pi c}\,\frac{1}{d\tau}\,\frac{[\psi]}{s}\,d\mathcal{S}\,. \tag{58}$$

Let ℓ be a coordinate measuring arc length along Γ, and let $(r, d\ell)$ be the angle between the positive directions of r and $d\ell$. Then the area element is

$$d\mathcal{S} = dr\,d\ell\,\sin(r,d\ell)\,. \tag{59}$$

Substituting this into (58) gives an integral w.r.t. ℓ along Γ:

$$\psi_{(d)}(P,t) = \int_{\Gamma} \frac{\cos(n,s)}{4\pi}\,\frac{[\psi]}{s}\,\frac{dr\,\sin(r,d\ell)}{c\,d\tau}\,d\ell\,. \tag{60}$$

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

$$c\,d\tau = dr + ds = dr\left(1+\tfrac{ds}{dr}\right) = dr\left(1+\cos(r,s)\right), \tag{61}$$

and substitute this into (60), obtaining the desired transformation:

$$\psi_{(d)}(P,t) = \int_{\Gamma} \frac{1}{4\pi s}\,\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,[\psi]\,d\ell\,. \tag{62}$$

Here ψ is the unobstructed wave function at the edge Γ, given by (33), which when delayed for the distance from Γ to P becomes

$$[\psi] = \frac{1}{r}\,f\big(t-(r+s)/c\big)\,. \tag{63}$$

Substituting this into (62) gives the alternative form[30]

$$\psi_{(d)}(P,t) = \int_{\Gamma} \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,. \tag{64}$$

Hence, by (50), the complete wave function is

$$\psi_{(a)}(P,t) = \psi_{(g)}(P,t) + \int_{\Gamma} \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,, \tag{65}$$

[30] Continuing from footnotes 9 and 20: If the factor 4π were included in the denominator of (33), it would leave (62) unchanged, but would contribute another factor 4π to the denominator in (64)—which is one reason why we *don't* use that convention here.

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 O 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 geometrically illuminated region,[31] and coordinate ℓ is measured along Γ. For the stated source strength $f(t)$, the geometrical-optics term $\psi_{(g)}(P,t)$ is given by

$$\psi_{(g)}(P,t) \;=\; \begin{cases} \frac{1}{p}\, f\big(t-p/c\big) & \text{in } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}_b\,, \end{cases} \tag{66}$$

where p is the direct distance from O to P.

In (64) and (65), 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; the latter decay factor is also seen in (62). But instead of an "obliquity factor" we now have

$$\frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)}\,, \tag{67}$$

in which the denominator 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 to the source grazes the edge of the aperture. This allows the integral to be discontinuous at the edge of the geometric shadow, as it must be, in order to compensate for the discontinuity in the geometric term $\psi_{(g)}(P,t)$; the sum of the two terms is $\psi_{(a)}(P,t)$, which was initially expressed as a *surface* integral over $\mathcal{S}_a$, in which the integrand gives no cause for any discontinuity at non-zero distances from the edge.

The derivation of (65) does not require Γ to be traversed in a particular direction. But, with a view to rewriting (65) in a *vector* notation, let us now agree that Γ is traversed clockwise about $\mathcal{S}_a$ as seen from O (which typically means counterclockwise as seen from the field point P; *cf.* Fig. 4, on p. 19).[32] 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 $\hat{\mathbf{r}}\times\hat{\mathbf{t}}$ 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 (65). So that factor may be replaced by $\hat{\mathbf{s}}\cdot\hat{\mathbf{r}}\times\hat{\mathbf{t}}$, or $\hat{\mathbf{s}}\times\hat{\mathbf{r}}\cdot\hat{\mathbf{t}}$. More obviously, $\cos(r,s)$ can be written $\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}$. Thus (65) becomes

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) \;+\; \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\;\frac{\hat{\mathbf{s}}\times\hat{\mathbf{r}}\cdot\hat{\mathbf{t}}}{1+\hat{\mathbf{r}}\cdot\hat{\mathbf{s}}}\;d\ell\,. \tag{68}$$

Multiplying the numerator and denominator of the integrand by rs and defining $\mathbf{r}=r\hat{\mathbf{r}}$ and $\mathbf{s}=s\hat{\mathbf{s}}$, we get the result in the computationally convenient form

$$\psi_{(a)}(P,t) \;=\; \psi_{(g)}(P,t) \;+\; \int_\Gamma \frac{f(t-(r+s)/c)}{4\pi rs}\;\frac{\mathbf{s}\times\mathbf{r}\cdot\hat{\mathbf{t}}}{rs+\mathbf{r}\cdot\mathbf{s}}\;d\ell\,. \tag{69}$$

3.6 Results applicable to a directional primary source?

If the single primary source is *extended*—that is, if its dimensions are not negligible compared with those of the spatial fluctuations of the wave function ("wavelengths")—then the waves from different parts of the source may interfere more or less constructively in different directions, so that the radiation is more or less concentrated in different directions, in which case we describe the source as **directional**.

[31] 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 [20, p. 261] and Born & Wolf [2, p. 502, eq. 12].

[32] Version 0.3 and earlier versions of this paper used the contrary convention, so that the subsequent cross products were in the reverse order.

Directionality *by itself* could be represented by allowing the strength function f in equation (33) to depend on directional coordinates as well as time. This *by itself* would not invalidate relation (34) and the rest of the derivation of the Kirchhoff diffraction formula (42) and its special cases.

Directionality, however, is not the only possible consequence of extendedness. Because an extended source, by definition, is not confined to a single point, it causes an ambiguity in r in equations (33) through (42), and in ϕ in equations (37) through (42). The ambiguity in r is especially problematic because it affects the $1/r$ decay of amplitude with distance in (33) and thence *does* invalidate relation (34).

We might attempt to salvage (33) through (42) as approximations by supposing that the extendedness is small. Indeed, if the dimensions of the primary source are small relative to the distance from any part of the source to any part of the surface of integration (that is, relative to r in all its ambiguity), then the ambiguities in r and ϕ are similarly small, and the $1/r$ decay factors for the various parts of the source may be taken as a common factor for small variations in r, so that (34) is still a good approximation.

Extendedness, however, is not the only possible cause of directionality or of departure from a $1/r$ decay. We have seen in the derivation of (42) that a dipole secondary source, although infinitesimal in size, gives the directional factor $\cos\chi$; and the $\cos\chi\,/\,s$ term in (42), combined with the leading factor $1/4\pi s$, gives a term that decays like $1/s^2$ instead of $1/s$. Similarly, a dipole primary source gives an amplitude that departs from the $1/r$ decay assumed in the derivation of (34). As the departure of a dipole secondary source from a $1/s$ decay is small if ψ/s is small relative to $\dot{\psi}/c$ (in eq. 42), so the departure of a dipole primary source from a $1/r$ decay is small if $f(t)\,/\,r$ is small relative to $f'(t)\,/\,c$. The latter condition, divided through by r and appropriately delayed, is none other than the "far-field" condition under which we can neglect the term proportional to $1/r$ in equations (34) through (44), obtaining (46) and (47). So r (with its ambiguity) must be far-field in this sense, as well as being large compared with the source, if (34) and hence (42) and its special cases are to remain good approximations for a dipole-like primary source; but in that case the "special cases" include (46) and (47), which hold for non-spherical primary wavefronts with sufficiently gradual curvature.

Another consequence of extendedness is that the waves emitted by different parts of the source may be *incoherent* (not synchronized), in which case, in a given direction, the constructiveness or destructiveness of the interference varies with time. In principle, we have already allowed for this effect by using time-dependent wave functions. Conceptually, however, it may be easier to think of the diffraction-affected wave function due to an extended source as the superposition of all the diffraction-affected wave functions due to the point-sources (coherent or incoherent, as the case may be) that make up the extended source.

If we want to express the wave function due to an extended source as the sum of a geometrical-optics term and a diffraction-correction term, as in (50), the concept of adding the wave functions due to the constituent point-sources of the extended source is especially helpful, because (i) each point on the source projects a different geometric shadow, and (ii) the correction term is large and discontinuous at the edge of that shadow, so that any attempt to apply a single correction term to the composite geometrical-optics function will be inaccurate in and adjacent to the penumbral region of the geometric shadow.

For comprehensiveness, however, let us consider the Maggi-Rubinowicz transformation of the correction term (62) for a *non*-extended directional source, such as a dipole. The transformation relies on two properties of the wave function: first, that we can neglect the ψ_n term in (51); and second, that the primary wave function decays like $1/r$. The first property remains a good approximation if the primary wave function is only weakly directional, which is certainly the case for a dipole. The second is not relied upon in the short-wavelength derivation in Section 3.5, which assumes that r does not affect the wave function except through τ. But in the exact derivation in Appendix C (below), a function involving f is integrated w.r.t. r along a generating line of the cone $\mathcal{S}_c$, which is in a fixed direction from the source. In this integration it does not matter if f varies with direction, provided that the primary wave function decays like $1/r$ in any given direction—which, for a dipole source, is a good approximation if r is "far-field" as defined above.

3.7 Generalized spatiotemporal-dipole secondary sources

In the Helmholtz formula (27), the expression in braces (i.e., 4π times the integrand) is

$$\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\,, \tag{70}$$

in which the second term (with the sign) represents a monopole strength density $-\psi_n$, and the first term represents a dipole strength density ψ in the n direction; the dipole source per unit area of $\mathcal{S}$ is a *spatial* dipole consisting of an "uninverted" monopole with strength ψ/h at $n{=}0$, and an "inverted" monopole with strength $-\psi/h$ at $n{=}-h$, with $h{\to}0$.

Now let us modify the spatial dipole by delaying the strength function of the inverted monopole by τ_h, and reducing its magnitude by a fraction α_h (e.g., no reduction if $\alpha_h = 0$), where τ_h and α_h, like h, are infinitesimal. Thus, compared with the uninverted monopole, the inverted monopole is recessed by the distance h, delayed by the time τ_h, and attenuated by the fraction α_h. At the field point P, the wave function due to the uninverted monopole is

$$\frac{\psi(t-s/c)}{hs}\,. \tag{71}$$

So the wave function due to the modified dipole is the total change in expression (71) due to n increasing by h, and t increasing[33] by τ_h, and the magnitude increasing by α_h times its final value. Since h and τ_h are infinitesimal, that total change is

$$h\,\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{hs}\right) + \tau_h\,\frac{\partial}{\partial t}\left(\frac{\psi(t-s/c)}{hs}\right) + \alpha_h\,\frac{\psi(t-s/c)}{hs}\,, \tag{72}$$

i.e.

$$\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) + \frac{\tau_h}{hs}\,\dot\psi(t-s/c) + \frac{\alpha_h}{hs}\,\psi(t-s/c)\,, \tag{73}$$

which will agree with (70) if and only if, on $\mathcal{S}$,

$$\frac{\tau_h}{h}\,\dot\psi + \frac{\alpha_h}{h}\,\psi = -\psi_n\,. \tag{74}$$

This is the sufficient and necessary condition for the modified dipoles to give the same secondary waves as the original dipoles and monopoles. And it does not look helpful.

But now let us invoke Assumption 7 (a single monopole primary source). As before, let r be the coordinate measured from the primary source to N (Fig. 3, p. 16), and let ϕ be the angle between the positive directions of n and r. Applying the chain rule to (74) gives

$$\frac{\tau_h}{h}\,\dot\psi + \frac{\alpha_h}{h}\,\psi = -\frac{\partial\psi}{\partial r}\,\frac{\partial r}{\partial n} \tag{75}$$

or, upon substitution from (34) and (37),

$$\frac{\tau_h}{h}\,\dot\psi + \frac{\alpha_h}{h}\,\psi = \left(\frac{\dot\psi}{c} + \frac{\psi}{r}\right)\cos\phi\,. \tag{76}$$

To satisfy this for all $\dot\psi$ and ψ, we equate the coefficients of $\dot\psi$, obtaining

$$\tau_h = \frac{h}{c}\,\cos\phi\,, \tag{77}$$

and equate the coefficients of ψ, obtaining

$$\alpha_h = \frac{h}{r}\,\cos\phi\,, \tag{78}$$

[33] If we introduce a delay u, the numerator of (71) becomes $\psi(t-u-s/c)$. In the change from the inverted monopole to the uninverted monopole, u changes from τ_h to 0, which has the same effect on the function as if t *increases* by τ_h.

so that the parameters of our "modified" dipole are uniquely determined. For reasons which will soon be apparent, we shall call this "modified" dipole a **generalized spatiotemporal dipole** (**GSTD**). The Helmholtz integrand in (27) may then be understood as a distribution of GSTDs on $\mathcal{S}$, oriented normal to $\mathcal{S}$, the first term in the braces representing the spatial aspect (equal and opposite monopoles) and the second term (in ψ_n) representing the modifications (delay and attenuation of the inverted monopole). Hence, in the Kirchhoff diffraction formula (42), the terms in $\cos\chi$ represent the spatial aspect while the terms in $\cos\phi$ represent the modifications; and in the integral (51) over the cone $\mathcal{S}_c$ with its apex at the primary source, the disappearance of the ψ_n term means that the dipoles become simply spatial—as is confirmed by putting $\phi = 90°$ in equations (77) and (78), which then give no delay and no attenuation.

According to (77), the delay of the inverted monopole is such that the waves from the two monopoles are synchronized (with opposing amplitudes) in the direction of the primary source, and in the cone of directions which make the same angle ϕ with the negative direction of n; this cone includes the direction of specular reflection off $\mathcal{S}$. And according to (78), the attenuation of the inverted monopole is such that the waves from the two monopoles cancel at a distance r in any of these directions (including at the primary source); at that distance, the greater proximity of the inverted monopole compensates for the reduced strength. Thus there are two ways in which the GSTDs suppress "backward" secondary waves: individually, they suppress secondary waves at particular distances in particular directions; collectively, they are equivalent to the sources in Proposition 5 and therefore, under the conditions of that proposition, suppress secondary waves throughout $\mathcal{R}'$.

If $\mathcal{S}$ coincides with a primary wavefront (as in Section 3.3), we have ϕ=0 in (77), so that the delay τ_h becomes h/c, which is simply the time taken for the waves emitted by the uninverted monopole to reach the inverted monopole. The latter is in the $-n$ direction, which is therefore the direction in which the waves from the two monopoles are synchronized (and cancel at distance r); the "cone of directions" collapses to its axis.

If the primary wavefront is plane (as in Section 3.4), we have $r \to \infty$ in (78), so that $\alpha_h = 0$; the inverted monopole is not attenuated, and the cancellation of the waves from the two monopoles (in the cone at angle ϕ to the $-n$ direction) becomes a far-field effect.

So if $\mathcal{S}$ coincides with a primary wavefront *and* is plane, as in equation (47), the inverted monopole is delayed by h/c and unattenuated, so that the waves from the two monopoles cancel in the $-n$ direction in the far field. This case is what David A. B. Miller [16] called a **spatiotemporal dipole**. We have "generalized" it in two ways: by allowing the delay of the inverted monopole to be less than h/c, so that the direction of cancellation need not be normal to $\mathcal{S}$; and by allowing the inverted monopole to be attenuated, so that the cancellation may occur at a finite distance. Together, these modifications allow the surface of integration $\mathcal{S}$ to be of a general shape and orientation and at a general distance from the primary source.

Although Miller applied his spatiotemporal-dipole theory to "uniform spherical or plane wave fronts" [16, after eq. 5], it is in fact a plain-wave (far-field) approximation in that it neglects the $1/r$ decay in the magnitude of the primary wave, with the result that his equation (4), which corresponds to our (34), lacks the second term on the right.[34] In (44) above, we see that this is a good approximation if ψ/r is very small compared with $\dot{\psi}/c$, which is our usual criterion for the far field of the primary source. To make it exact for all r, we need the inverted monopoles to be attenuated in accordance with (78).

The need for the $1/r$ term in (44) was revealed by a very different method in my 2019 paper "A tautological theory of diffraction" [18], 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...

We have just seen that the reconciliation requires a modification of the spatiotemporal dipole, namely an attenuation of the inverted monopole.

[34] Larmor had made the same approximation: in [14], on p.172, the equation second from the bottom agrees with Miller's (4), with the result that the bottom equation agrees with our (47), not our (44).

So this is the "other paper". Why did I not expect to be its author? One reason 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" [18, p.8]. My concern was that if the surface $\mathcal{S}$ is spherical, the element displaced in the $-n$ direction from $\mathcal{S}$ does not have exactly the same area as the corresponding element of $\mathcal{S}$. In the present paper, I have sidestepped that complication by treating the monopole strengths 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 $\mathcal{S}$. I notice, however, that if the inverted and uninverted monopoles had the same strength per unit area of their respective sheets, this would *not* give the required attenuation, because their strengths would then be proportional to their respective values of r^2, whereas condition (78) makes them proportional to their respective values of r.

4 The sinusoidal (monochromatic) case

4.1 Terminology

If the function f in equation (1) has the **sinusoidal** form [35]

$$f(t) = A\cos(\omega t + \epsilon) \tag{79}$$

where A (the *peak amplitude*), ω (the *angular frequency* or *radian frequency*), and ϵ (the *phase angle*) are constants, then the wave function due to the monopole source in (1) becomes [36]

$$\frac{A}{r}\cos\left(\omega(t - r/c) + \epsilon\right), \tag{80}$$

which is usually written

$$\frac{A}{r}\cos(\omega t - kr + \epsilon), \tag{81}$$

where

$$k = \omega/c\,. \tag{82}$$

We call k the *wavenumber* (or, more precisely, the *angular* or *radian* wavenumber). In (81), 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

$$T = 2\pi/\omega\,, \tag{83}$$

and the *wavelength* is

$$\lambda = 2\pi/k\,. \tag{84}$$

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:

$$\nu = \frac{\omega}{2\pi}\,. \tag{85}$$

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

$$\lambda\!\!\!^{-} = \frac{\lambda}{2\pi} = 1/k\,. \tag{86}$$

Although (80) has been presented as a special case of (1), there is also a sense in which sinusoidal time-dependence yields a generalization: by allowing c to depend on ω, we can handle *dispersive* media (see the discussion after Assumption 5, above). In this paper, however, we shall continue to derive results for sinusoidal time-dependence as special cases of the corresponding results for general time-dependence; the possibility of generalizing c, although ever present, will not be emphasized.

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

[36] 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 depart from that region or fade away.

4.2 Vector representation

At a fixed point, hence a fixed r, the wave function (81) has the same form as (79), except that the peak amplitude and the phase angle depend on r. The contributions to the wave function at that point due to any other sinusoidal monopole sources with the same ω (e.g., reflectors) are also of that form, so that their total contribution to the wave function at that point is a sum of functions of form (79), with the same ω (but generally not the same constants A and ϵ). 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* (79)*, with the same frequency, is a function of the same form with the same frequency*.[37] Indeed, in the Cartesian xy plane, function (79) is the x component of a vector of length A making an angle $\omega t + \epsilon$ with the x axis—that is, rotating at angular frequency ω from an initial angle ϵ. 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 realistic and useful to consider a wave function of the form

$$\psi = A \cos(\omega t + \epsilon) \,, \tag{87}$$

where A and ϵ are independent of time but vary from point to point.

4.3 Complex representation

The need for what we now call *complex numbers* emerged in the 16th century, when the newly discovered formulae for the roots of cubic equations gave intermediate expressions that were sometimes not real although the corresponding roots were real. Nowadays we are more ambitious: we also want to solve equations that don't have real roots by the same methods as those that do, represent physical quantities that cannot be adequately modeled by real numbers while retaining the ability to represent those that can, and represent real functions as functions of a more general kind whenever the latter are more convenient to work with.[38] For any of these purposes, we need the real numbers to be a subset of the complex numbers, and we need the definitions of addition and multiplication of complex numbers to reduce to addition and multiplication of real numbers if the operands happen to belong to that subset.

As the real numbers can be represented by points on a number line, the obvious way to make them a subset of the complex numbers is to identify the number line with the x axis in a Cartesian plane (the "complex plane"), and to define a complex number as any point in that plane. Such a point can be specified by the rectangular coordinates (x, y), where x is called the *real part*, and y (although also a real number) is called the *imaginary part*. The same point can also be specified by the polar coordinates (ρ, θ), such that $x = \rho \cos\theta$ and $y = \rho \sin\theta$, where ρ is called the *modulus* and θ is called the *argument*. If z is a complex number, its modulus is written $|z|$, and is synonymous with the absolute value if z happens to be real. The complex number whose modulus is that of z, but whose argument is *minus* that of z, is called the *complex conjugate* of z and denoted by $\bar{z}$ or z^*. A complex number whose real part is zero is described as *purely imaginary*.

Viewed against the complex plane, addition of real numbers becomes a special case of addition of plane vectors. So, for the definition of complex addition we may say—and by convention we do say—that complex numbers add like their position vectors: *the real part of the sum is the sum of the real parts, and the imaginary part of the sum is the sum of the imaginary parts*. The closure, commutative, associative,

[37] This was first shown by Fresnel, by an analytical method, in a "supplement" dated January 1818 [7, vol. 1, at pp. 489–92]; see [19] 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.

[38] I include the following brief introduction to complex numbers not only for readers who aren't familiar with them, but also for those who are, because it is unusual, and because it registers a further protest against inconsistent introductions.

identity, and inverse laws of complex addition then follow immediately from the corresponding laws of real addition, applied separately to the real and imaginary parts.

In the complex plane, if we consider the multiplication of two non-zero real numbers with all possible combinations of signs, we notice that in all cases *the modulus of the product is the product of the moduli, and the argument of the product is the* sum *of the arguments*. If either operand is zero, the same rule holds except for the inevitable indeterminacy of the argument of zero. So we adopt this rule as the definition of complex multiplication. The closure, commutative, associative, identity, and inverse laws of complex multiplication then follow immediately from the corresponding laws of real multiplication (applied to the moduli) and of real addition (applied to the arguments).

If we then define i (called the *imaginary unit*) as the complex number with rectangular coordinates $(0, 1)$, hence the modulus 1 and the argument $\pi/2$, we find by the above definitions that $i^2 = -1$ (thus we have found a square root of -1), and that the number with rectangular coordinates (x, y) is $x+iy$ (which is the usual way of writing it). From the same definitions it is easy to show *geometrically* that complex multiplication is distributive over complex addition. From the distributive law and the others, we obtain the following formula, which gives the product of two complex numbers in rectangular form without the need to convert them to polar form:

$$(x+iy)(u+iv) = (xu - yv) + i(yu + xv)\,. \tag{88}$$

For example, the product $(\cos\alpha + i\sin\alpha)(\cos\beta + i\sin\beta)$ is given by (88) as

$$\big(\cos\alpha\,\cos\beta - \sin\alpha\,\sin\beta\big) + i\big(\sin\alpha\,\cos\beta + \cos\alpha\,\sin\beta\big) \tag{89}$$

but is given by the polar definition of multiplication as

$$\cos(\alpha+\beta) + i\sin(\alpha+\beta)\,. \tag{90}$$

Equating the real parts of the last two expressions gives the expansion formula for the cosine of a sum, and equating the imaginary parts gives the formula for the sine of a sum; this, as far as I know, is the fastest way to obtain these "angle-sum identities". But perhaps I digress.[39]

The function $A\cos(\omega t + \epsilon)$, which we have usefully characterized as the x component of a vector with length A making an angle $\omega t + \epsilon$ with the x axis, can now be described as the real part of a complex number with modulus A and argument $\omega t + \epsilon$. For brevity we shall write this number as $A\,\mathrm{cis}(\omega t + \epsilon)$, where, in general,

$$\mathrm{cis}\,\theta = \cos\theta + i\sin\theta\,, \tag{91}$$

and the function name "cis", coined by William Rowan Hamilton and published posthumously in 1866, is obviously an acronym for the operator $(\cos + i\sin)$.[40] By definition (91) we have

$$\mathrm{cis}\,0 = 1\,, \tag{92}$$

and by the polar definition of multiplication we have

$$\mathrm{cis}(\alpha+\beta) = \mathrm{cis}\,\alpha\;\mathrm{cis}\,\beta\,. \tag{93}$$

[39] Nearly all introductions to complex numbers give the rectangular definition of multiplication first, and then use the angle-sum identities—obtained from elsewhere—to establish the polar definition. This makes the initial definition of complex multiplication look completely arbitrary, unless it is presented in the form (88), in which case it appears to be derived—and sometimes *is* derived—by prematurely assuming that operations on i follow the rules of real algebra although its defining property $(i^2 = -1)$ does not! Moreover, introductions that define multiplication by (88) begin by defining a complex number as being of the form $x+iy$, which (in the absence of a prior geometric definition of i) fails to explain why x and y should be measured in perpendicular directions, so that the subsequent identification of x and y with Cartesian coordinates is also highly arbitrary. Here I have tried to show that the construction of the complex numbers can be both heuristic and rigorous.

[40] The etymology suggests that "cis" should be pronounced "kiss"; but the pronunciation of other words beginning with "ci..." suggests "siss", which seems to have prevailed.

Definition (91) also yields the rule

$$\frac{d}{d\theta}\big(\operatorname{cis}\theta\big) = i\operatorname{cis}\theta\,, \tag{94}$$

which is extremely convenient because it *reduces differentiation to multiplication by a constant*. We shall see that (93) leads to a similarly convenient rule which *reduces a delay to multiplication by a constant*.

Properties (92) to (94) of the function $\operatorname{cis}\theta$ would be expected if this function were identical with $e^{i\theta}$, and indeed it is easily shown that if the two functions share these properties, they must be identical.[41] Hence $\operatorname{cis}\theta$ is usually written as $e^{i\theta}$. I say this in order to facilitate comparison with references; otherwise we could equally well express our results in the "cis" notation and not be bothered with the meaning of imaginary exponents.

We see in (91) that reversing the sign of θ does not change the real part of $\operatorname{cis}\theta$ (although it changes the sign of the imaginary part). Hence the wave function (81), which is the real part of

$$\frac{A}{r}\operatorname{cis}(\omega t - kr + \epsilon)\,, \tag{95}$$

is also the real part of

$$\frac{A}{r}\operatorname{cis}(kr - \omega t - \epsilon)\,; \tag{96}$$

the latter is usually preferred in optics because it tends to emphasize the manageably rapid spatial variation rather than the unmanageably rapid time-variation. By repeated application of property (93), expression (95) can be written $\frac{A}{r}\operatorname{cis}\epsilon\operatorname{cis}(-kr)\operatorname{cis}\omega t$, which has the form

$$B\operatorname{cis}\epsilon\operatorname{cis}\omega t \tag{97}$$

where B is independent of time. Similarly, expression (96) can be written as $\frac{A}{r}\operatorname{cis}(-\epsilon)\operatorname{cis}kr\operatorname{cis}(-\omega t)$, which has the form

$$B^*\operatorname{cis}(-\epsilon)\operatorname{cis}(-\omega t) \tag{98}$$

where B^* (the conjugate of B) is independent of time. Expressions (97) and (98) have the same real part, and in both expressions the angle ϵ is *added* to ωt and represents a phase *advance*.[42] But this advance is equivalent to a multiplication by $\operatorname{cis}\epsilon$ in the former case, in which the time-dependent factor is $\operatorname{cis}\omega t = e^{i\omega t}$, and equivalent to a multiplication by $\operatorname{cis}(-\epsilon)$ in the latter case, in which the time-dependent factor is $\operatorname{cis}(-\omega t) = e^{-i\omega t}$. For example, a phase advance of a quarter-cycle ($\pi/2$ radians; 90°) is equivalent to a multiplication by $\operatorname{cis}\frac{\pi}{2} = i$ in the former case, and a multiplication by $\operatorname{cis}\frac{-\pi}{2} = -i$ in the latter. And in optics the latter convention, which follows from (96), is usually preferred, so that a factor $-i$ indicates a 90° phase advance.[43]

Now, using properties (92) to (94) and equation (82), we can readily establish that if $g(t)$ is a function with the time-dependent factor $\operatorname{cis}(-\omega t) = e^{-i\omega t}$ (any other factors being independent of time), then

$$\dot{g}/c \;=\; -ikg\,; \tag{99}$$

$$\frac{1}{c}\left[\frac{\partial g}{\partial t}\right] = \frac{1}{c}\big[\dot{g}\big] \;=\; -ik\big[g\big]\,; \tag{100}$$

$$g\big(t - s/c\big) = \big[g(t)\big] \;=\; e^{iks}\,g(t)\,, \tag{101}$$

---

[41] From (94), if we transpose $i\operatorname{cis}\theta$ to the left-hand side, multiply through by $e^{-i\theta}$, and recognize the derivative of a product, we get $\frac{d}{d\theta}\big(\operatorname{cis}\theta\, e^{-i\theta}\big) = 0$. Now integrate w.r.t. θ, use $\theta = 0$ to find the constant of integration, and multiply through by $e^{i\theta}$.

[42] E.g., (98) may be written $B^*\operatorname{cis}\big(-(\omega t + \epsilon)\big)$.

[43] Exceptions include Kottler [12] and Miller [16], who use the time-dependent factor $e^{i\omega t}$, so that i in their results corresponds to $-i$ in ours. Electrical engineers do likewise, except that they usually call the imaginary unit j because they use i for electric current. Thus electrical engineers normally use the time-dependent factor $e^{j\omega t}$, and our results may be converted to their convention by writing $-j$ for our i.

where (101) is the promised rule reducing a delay to multiplication by a constant.[44] By rules (99) to (101), the results that we have obtained for general time-dependence can be rewritten for the special case of sinusoidal time-dependence.

4.4 Restatement of results

Having derived the Helmholtz formula (27) in a form that allows general time-dependence, we can now recover the traditional, monochromatic form. Applying rule (101) in (27) gives

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{e^{iks}\psi}{s}\right) - \frac{1}{s}\,e^{iks}\psi_n\right\} d\mathcal{S}\,. \tag{102}$$

But because ψ is evaluated at $n=0$, we can take it outside the differentiation w.r.t. n. Thus we obtain what is *usually* called the Helmholtz formula:[45]

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\psi\,\frac{\partial}{\partial n}\left(\frac{e^{iks}}{s}\right) - \frac{e^{iks}}{s}\,\psi_n\right\} d\mathcal{S}\,. \tag{103}$$

The Kirchhoff integral (31), unlike the Helmholtz integral, is not usually expressed in monochromatic form; but, applying rule (100) in (31) gives

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{ik}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S} \tag{104}$$

$$= \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial}{\partial s}\left(\frac{1}{s}\right)\frac{\partial s}{\partial n} + \frac{ik}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S} \tag{105}$$

$$= \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial s}{\partial n}\left(\frac{\partial}{\partial s}\left(\frac{1}{s}\right) + \frac{ik}{s}\right) - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{106}$$

which agrees with an intermediate expression given by Baker and Copson.[46] The same result is obtainable from (103) by using the chain and product rules on the first term, taking the common factor e^{iks} out of the resulting two terms, and then applying (101) in reverse (in two places).

Applying (100) in (42), we obtain the exact monochromatic form of the Kirchhoff diffraction formula, which concerns spherical primary wavefronts and an arbitrary surface of integration:

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s}\left\{-ik\left(\cos\phi + \cos\chi\right) + \frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right\} d\mathcal{S}\,. \tag{107}$$

---

[44] To verify rules (99) to (101), it suffices to consider $g(t) = \mathrm{cis}(-\omega t)$, since any time-independent factors are carried through without change. So for (99) we have

$$\dot{g}/c = \frac{1}{c}\,\frac{d}{dt}\,\mathrm{cis}(-\omega t) = \frac{1}{c}\,\frac{d}{d(-\omega t)}\,\mathrm{cis}(-\omega t)\,\frac{d}{dt}\,(-\omega t) = \frac{1}{c}\,i\,\mathrm{cis}(-\omega t)\cdot(-\omega) = -i\,\frac{\omega}{c}\,\mathrm{cis}(-\omega t) = -ikg\,.$$

And for (100), remembering that we can interchange delaying and scaling, we have

$$\frac{1}{c}\left[\dot{g}\right] = \left[\dot{g}/c\right] = \left[-ikg\right] \qquad \text{by (99)}$$
$$= -ik\left[g\right].$$

And for (101) we have

$$g\big(t - s/c\big) = \mathrm{cis}\left(-\omega(t - s/c)\right) = \mathrm{cis}(\omega s/c - \omega t) = \mathrm{cis}(ks - \omega t) = \mathrm{cis}\,ks\,\mathrm{cis}(-\omega t) = e^{iks}\,g(t)\,.$$

[45] *Cf.* Born & Wolf [2], p. 419, eq. (7), where their U is our ψ; Baker & Copson [1], p. 24, eq. (4.24), where their r is our s, their v is our ψ, and their ν is our $-n$ (causing the sign reversals).

[46] See [1], p. 36, above the line "or, finally," where their u is our ψ and their r is our s.

Similarly applying (100) in (44), *or* putting $\phi = 0$ in (107), we get the exact monochromatic form of the Huygens-Fresnel-Kirchhoff formula, in which the integration is over a spherical primary wavefront: [47]

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ -ik\left(1+\cos\chi\right) + \frac{1}{r} + \frac{\cos\chi}{s} \right\} d\mathcal{S} \,. \tag{108}$$

For *plane* primary waves, we apply rule (100) in (46) and (47), or put $r \to \infty$ in (107) and (108), obtaining

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ -ik\left(\cos\phi+\cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{109}$$

for integration over a general surface, and

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ -ik\left(1+\cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{110}$$

for integration over a primary wavefront. [48] Of course (110) is also obtainable from (109) by setting $\phi = 0$.

Applying (101), with $r+s$ instead of s, in (64), we obtain the monochromatic form of the Maggi-Rubinowicz transformation of the diffraction-correction term, which we then divide by $f(t)$—losing the time-dependence—to get the result for a *unit* primary source: [49]

$$\psi_{(d)}(P) = \int_{\Gamma} \frac{e^{ik(r+s)}}{4\pi rs} \frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)} \, d\ell \,. \tag{111}$$

The same treatment of the geometrical-optics term (66) gives

$$\psi_{(g)}(P) = \begin{cases} e^{ikp}/p & \text{in } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}_b \,, \end{cases} \tag{112}$$

where p is the direct distance from the source to P. The complete wave function for a unit source is the sum of the two terms, which we slightly rearrange in imitation of Rubinowicz [20, p.262, eq. 5]:

$$\psi_{(a)}(P) = \psi_{(g)}(P) + \frac{1}{4\pi} \int_{\Gamma} \frac{e^{ikr}}{r} \frac{e^{iks}}{s} \frac{\cos(n,s)\,\sin(r,d\ell)}{1+\cos(r,s)} \, d\ell \,. \tag{113}$$

The same treatment of the vector form (69) of the Maggi-Rubinowicz transformation gives [50]

$$\psi_{(a)}(P) = \psi_{(g)}(P) + \int_{\Gamma} \frac{e^{ik(r+s)}}{4\pi rs} \frac{\mathbf{s}\times\mathbf{r}\cdot\hat{\mathbf{t}}}{rs+\mathbf{r}\cdot\mathbf{s}} \, d\ell \,. \tag{114}$$

[47] 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 [14, p.172, last eq.] and Miller [16, eq. 6] omit the second-last term in the braces (see Section 3.7 above), but agree with me on the sign of the last term.

[48] As Miller [16] neglects the curvature of the wavefront, his (6) agrees with our (110) instead of our (108) when we allow for the differing conventions (his ϕ is our ψ; his r is our s; his i corresponds to our $-i$; and his θ is our χ).

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

[50] *Cf.* Baker and Copson [1, p.79, eq. 2.111], who consider *minus* our diffraction-correction integral, without the leading factor $1/4\pi$; and Kottler [12, p.420, eq.7], whose $d\bar{\mathbf{s}}_Q$ is our $\hat{\mathbf{t}}\,d\ell$, and whose i, like Miller's, corresponds to our $-i$.

4.5 Meaning of "far field" and "short wavelength"

We have defined the *far field* of a primary source as the region in which

$$\left|\frac{\psi}{r}\right| \ll \left|\frac{\dot{\psi}}{c}\right| . \tag{115}$$

For the monochromatic case, by rule (99), this becomes $\left|\psi/r\right| \ll \left|-ik\psi\right|$ or, as the modulus of a product is the product of the moduli, $|\psi|\left|\frac{1}{r}\right| \ll |-i|\,|k|\,|\psi|$, which simplifies to

$$\frac{1}{r} \ll k \tag{116}$$

or, if we take reciprocals,

$$r \gg \lambda\!\!\!^{-} . \tag{117}$$

So the *far-field* condition for a primary source is that *the distance from the source is very large compared with the radian wavelength.*

This is the condition under which we can neglect the curvature of the primary wavefront in the Kirchhoff diffraction formula (42) or its Huygens-Fresnel form (44), obtaining (46) or (47), respectively. (Hence, for a non-spherical primary wavefront, which does not have a unique radius of curvature, we would expect (46) or (47) to be applicable if the *smallest* radius of curvature is very large compared with the radian wavelength.) In the monochromatic forms of the same formulae, the condition can be seen directly: if (116) applies, then (107) and (108) can be approximated by (109) and (110), respectively. This is also the condition under which we can apply all eight results, and the Maggi-Rubinowicz transformation, to a dipole-like primary source (Section 3.6). And by equation (76), it is the condition under which we can neglect the attenuation (but not the delay) of the inverted monopoles in the GSTD secondary sources. In the case of integration over a primary wavefront, this becomes the condition under which the GSTDs may be approximated by Miller's spatiotemporal dipoles.

Similarly, the far field of a secondary source, or of an element of the surface of integration, is the region in which

$$\left|\frac{\psi}{s}\right| \ll \left|\frac{\dot{\psi}}{c}\right| . \tag{118}$$

For the monochromatic case, this condition becomes

$$\frac{1}{s} \ll k \tag{119}$$

or, if we take reciprocals,

$$s \gg \lambda\!\!\!^{-} , \tag{120}$$

which means that the distance from the surface element to the field point is very large compared with the radian wavelength.

Now suppose that, by the above definitions, the surface of integration $\mathcal{S}$ is in the far field of the primary source *and* the field point is in the far field of each surface element. Then, by (115) & (118), we can neglect the $[\psi]$ term and retain only the $[\dot{\psi}]$ term in the Kirchhoff diffraction formula (42) and its Huygens-Fresnel form (44), obtaining

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{[\dot{\psi}]}{4\pi cs}\left(\cos\phi + \cos\chi\right) d\mathcal{S} \tag{121}$$

and

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{[\dot{\psi}]}{4\pi cs}\left(1 + \cos\chi\right) d\mathcal{S} , \tag{122}$$

respectively. And by (116) and (119), we can neglect the real terms in the braces in the corresponding monochromatic formulae (107) and (108), obtaining respectively

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{-ik[\psi]}{4\pi s}\left(\cos\phi+\cos\chi\right)d\mathcal{S} \tag{123}$$

and

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{-ik[\psi]}{4\pi s}\left(1+\cos\chi\right)d\mathcal{S}\,, \tag{124}$$

which also follow from (121) and (122) by rule (100). In all four cases, the so-called obliquity factor depends on the twin far-field assumptions. To express the last four results in terms of the monopole primary source, we substitute (63) into (121) and (122), obtaining respectively

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{f'\left(t-(r+s)/c\right)}{4\pi crs}\left(\cos\phi+\cos\chi\right)d\mathcal{S} \tag{125}$$

and

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{f'\left(t-(r+s)/c\right)}{4\pi crs}\left(1+\cos\chi\right)d\mathcal{S}\,, \tag{126}$$

and then make the same substitution in (123) and (124), apply rule (101) with $r+s$ instead of s, and divide by $f(t)$ to get the corresponding results for a unit primary source:

$$\psi(P) \approx \iint_{\mathcal{S}} \frac{-ik\,e^{ik(r+s)}}{4\pi rs}\left(\cos\phi+\cos\chi\right)d\mathcal{S} \tag{127}$$

and

$$\psi(P) \approx \iint_{\mathcal{S}} \frac{-ik\,e^{ik(r+s)}}{4\pi rs}\left(1+\cos\chi\right)d\mathcal{S}\,. \tag{128}$$

The last two results indicate that the secondary sources are advanced in phase by 90° relative to the primary wave function at $d\mathcal{S}$, and that their strengths are proportional to k, hence inversely proportional to the wavelength. Of the last eight equations, each one with the factor $\left(1+\cos\chi\right)$ applies if $\mathcal{S}$ coincides with a primary wavefront, and follows from its predecessor by setting $\phi=0$.

In (127) and (128), the constant $-ik$ can be taken outside the integral; and in (128), as r is uniform over $\mathcal{S}$, the factors e^{ikr} and $1/4\pi r$ can likewise be taken outside. By (84), we can also put $k=2\pi/\lambda$ in (127) and (128). Then, to obtain the field for a primary source of strength A, we multiply by A, obtaining

$$\psi(P) \approx -\frac{iA}{2\lambda}\iint_{\mathcal{S}} \frac{e^{ik(r+s)}}{rs}\left(\cos\phi+\cos\chi\right)d\mathcal{S} \tag{129}$$

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

$$\psi(P) \approx -\frac{iA}{2\lambda}\frac{e^{ikr}}{r}\iint_{\mathcal{S}} \frac{e^{iks}}{s}\left(1+\cos\chi\right)d\mathcal{S} \tag{130}$$

if $\mathcal{S}$ is a primary wavefront.[51] The approximation (129), rather than the corresponding exact form (107), is what is usually called the *Fresnel-Kirchhoff diffraction formula*.[52]

Together, the two far-field conditions are equivalent to the condition that the radian wavelength is very small compared with r and s. This is none other than the "short-wavelength" condition under which have derived the Maggi-Rubinowicz transformation of the diffraction-correction integral (Section 3.5). Thus our short-wavelength derivation may also be described as a far-field derivation.

[51] *Cf.* Born & Wolf [2], pp. 422–3, eqs. (17) and (18), respectively. If we multiply our (128) by an assumed source strength $e^{-i\omega t}=e^{-ikct}$, the result agrees with Baker & Copson [1], p.33, eq. (4.61); and if we substitute that source strength for A in our equation (129), the result agrees with [1], p.73, eq. (1.33).

[52] Baker & Copson [1], p.73; Born & Wolf [2], p. 422.

5 Notes on Fresnel diffraction

5.1 Basic equations; plane baffle

The **Fresnel approximation** for diffraction refers to a single monopole primary source (Assumption 5), and assumes the following additional conditions:

(F1) The surface of integration is in the far field of the primary source by criterion (115), and the field point (observation point) is in the far field of the secondary sources by criterion (118). For sinusoidal waves, these criteria become (117) and (120), meaning that the distances r and s are very large compared with the radian wavelength.[53]

(F2) For the purpose of calculating the *amplitudes* of the interfering secondary waves, or at least those secondary waves that contribute significantly to the integral, we can neglect variations in the distance-related attenuation factors $1/r$ and $1/s$, and ignore obliquity.

(F3) For the purpose of calculating the *phase differences* between the interfering secondary waves, or at least those secondary waves that contribute significantly to the integral, the relative propagation delay from the primary source to the observation point via the point Q on the surface of integration ("relative" to propagation along the line of sight) is proportional, nearly enough, to the square of the distance of point Q from the line of sight.

The obvious way to satisfy (F2) is to make the aperture small (compared with the distances to the primary source and the observation point) and close to the line of sight, and roughly perpendicular to it. Concerning (F3), as the relative propagation delay must vary smoothly and symmetrically as Q passes through the line of sight, the distance-squared law is a good approximation for surface elements sufficiently close to the line of sight. Accordingly, (F2) and (F3) are described as *paraxial* approximations, where *paraxial* means close to the axis (of symmetry of the propagation time), which is the line of sight.

We have seen that under (F1), the Kirchhoff diffraction formula (42), for primary-source strength $f(t)$, reduces to (125). If the integration is restricted to the aperture $\mathcal{S}_a$, this becomes

$$\psi_{(a)}(P,t) \approx \iint_{\mathcal{S}_a} \frac{f'\big(t-(r+s)/c\big)}{4\pi crs}\big(\cos\phi+\cos\chi\big)\,d\mathcal{S}\,. \tag{131}$$

The corresponding Maggi-Rubinowicz vector form is (69), which we repeat for convenience:

$$\psi_{(a)}(P,t) = \psi_{(g)}(P,t) + \int_{\Gamma} \frac{f(t-(r+s)/c)}{4\pi rs}\,\frac{\mathbf{s}\times\mathbf{r}\cdot\hat{\mathbf{t}}}{rs+\mathbf{r}\cdot\mathbf{s}}\,d\ell\,; \tag{69}$$

this too has been derived by assuming condition (F1), although (as shown in Appendix C) it does not make any difference.

To apply (F2), we replace r and s by constants, say a and b, in the denominators $4\pi crs$ and $4\pi rs$, and replace $\cos\phi$ and $\cos\chi$ by 1. But we do not (yet) replace r and s in the denominator $rs+\mathbf{r}\cdot\mathbf{s}$, which has a minimum of zero when $\mathbf{r}$ and $\mathbf{s}$ are opposite; any shift in that minimum could cause gross changes in the integral near the boundary of the geometric shadow. So (131) and (69) become

$$\psi_{(a)}(P,t) \approx \iint_{\mathcal{S}_a} \frac{f'\big(t-(r+s)/c\big)}{2\pi cab}\,d\mathcal{S} \tag{132}$$

and

$$\psi_{(a)}(P,t) \approx \psi_{(g)}(P,t) + \int_{\Gamma} \frac{f(t-(r+s)/c)}{4\pi ab}\,\frac{\mathbf{s}\times\mathbf{r}\cdot\hat{\mathbf{t}}}{rs+\mathbf{r}\cdot\mathbf{s}}\,d\ell\,. \tag{133}$$

[53] At this writing, the *Wikipedia* article on "Fresnel diffraction" describes the associated approximations as "*near*-field" because it uses the term "far field" in a more restrictive sense, namely the sense of *Fraunhofer* diffraction (see footnote 27, above). But the article goes on to describe limits on how "near" we can go.

To apply (F3) while maintaining compliance with (F2), we must be more specific about the geometry. In Cartesian coordinates (x, y, z), let the primary source O be at $(0, 0, -a)$ and the observation point P at $(0, 0, b)$, so that the line of sight is along the z axis; and let the surface $\mathcal{S}$, comprising segments $\mathcal{S}_a$ spanning the aperture and $\mathcal{S}_b$ on the P side of the baffle, be the plane $z = 0$ (the xy plane); and let (ρ, θ) be the polar coordinates of the point Q in that plane, so that $x = \rho\cos\theta$, $y = \rho\sin\theta$, $\rho^2 = x^2 + y^2$, and the corresponding right-handed cylindrical coordinates are (ρ, θ, z). Then, by Pythagoras,

$$r = OQ = \sqrt{a^2 + \rho^2} = a\sqrt{1 + \left(\tfrac{\rho}{a}\right)^2} \approx a\left(1 + \tfrac{1}{2}\left(\tfrac{\rho}{a}\right)^2\right) \tag{134}$$

if ρ/a is sufficiently small,[54] whence

$$r \approx a + \frac{\rho^2}{2a}\,. \tag{135}$$

Similarly, if ρ/b is sufficiently small,

$$s \approx b + \frac{\rho^2}{2b}\,. \tag{136}$$

Indeed, if Q is sufficiently close to the z axis, $r \approx a$ and $s \approx b$, consistent with (F2), even if $\mathcal{S}$ is slightly curved and/or slightly tilted w.r.t. the xy plane. Moreover, adding (135) and (136) gives

$$r + s \approx a + b + \frac{a+b}{2ab}\rho^2\,, \tag{137}$$

which is consistent with (F3), since $a + b$ is the propagation distance along the line of sight; and again the consistency is maintained if $\mathcal{S}$ is slightly curved and/or slightly tilted, because the coefficient of ρ^2 changes only slightly with small changes in a or b. Now, using the cylindrical coordinates, let us write a vector as a column whose elements are the components (with dimensions of length) in the directions of ρ, θ, and z, in that order. Then we have

$$\mathbf{s} = \begin{bmatrix} \rho \\ 0 \\ -b \end{bmatrix}\,, \quad \mathbf{r} = \begin{bmatrix} \rho \\ 0 \\ a \end{bmatrix}\,, \quad \text{and} \quad \hat{\mathbf{t}}\,d\ell = \begin{bmatrix} d\rho \\ \rho\,d\theta \\ 0 \end{bmatrix}\,, \tag{138}$$

hence (expanding the determinant of the three columns along the middle row)

$$\mathbf{s} \times \mathbf{r} \cdot \hat{\mathbf{t}}\,d\ell = -(a+b)\rho^2\,d\theta\,, \tag{139}$$

and

$$\mathbf{r}\cdot\mathbf{s} = \rho^2 - ab\,. \tag{140}$$

Multiplying (135) by (136) and neglecting the term in ρ^4 (which would be wrong anyway), we get[55]

$$rs \approx ab + \frac{a^2+b^2}{2ab}\rho^2\,, \tag{141}$$

to which we may add (140), obtaining

$$rs + \mathbf{r}\cdot\mathbf{s} \approx \frac{(a+b)^2\rho^2}{2ab}\,, \tag{142}$$

where the right-hand side has a minimum of zero when $\mathbf{r}$ and $\mathbf{s}$ are opposite (i.e., when $\rho = 0$), as required. Substituting (137) into (132) yields

$$\psi_{(a)}(P,t) \approx \frac{1}{2\pi cab}\iint_{\mathcal{S}_a} f'\left(t - \tfrac{a+b}{c}\left(1 + \tfrac{\rho^2}{2ab}\right)\right)\,d\mathcal{S}\,; \tag{143}$$

[54] For small h, $\sqrt{1+h} \approx 1 + h/2$, because $(1 + h/2)^2 = 1 + h + h^2/4 \approx 1 + h$.

[55] Confirmation: By analogy with (134),

$$rs = a\sqrt{1 + \left(\tfrac{\rho}{a}\right)^2}\;b\sqrt{1 + \left(\tfrac{\rho}{b}\right)^2} = ab\sqrt{\left(1 + \left(\tfrac{\rho}{a}\right)^2\right)\left(1 + \left(\tfrac{\rho}{b}\right)^2\right)} = ab\sqrt{1 + \frac{\rho^2}{a^2} + \frac{\rho^2}{b^2} + \frac{\rho^4}{a^2b^2}}\,,\ \text{etc.}$$

and substituting (137), (139), and (142) into (133) yields

$$\psi_{(a)}(P,t) \;\approx\; \psi_{(g)}(P,t) \;-\; \frac{1}{2\pi(a+b)} \int_{\Gamma} f\Big(t - \tfrac{a+b}{c}\big(1 + \tfrac{\rho^2}{2ab}\big)\Big)\, d\theta\,, \tag{144}$$

where the geometrical-optics term $\psi_{(g)}(P,t)$ is given by (66) with $p = a + b$:

$$\psi_{(g)}(P,t) \;=\; \begin{cases} \frac{1}{a+b}\, f\big(t - \frac{a+b}{c}\big) & \text{in } \mathcal{R}_a \\ 0 & \text{in } \mathcal{R}_b\,. \end{cases} \tag{145}$$

5.2 Maggi-Rubinowicz treatment of Poisson's spot

Now let ρ be constant, so that Γ is the circle of radius ρ in the xy plane, centered on the origin. And let the "baffle" be the interior of the circle—that is, the corresponding *disk*—so that the surface segment $\mathcal{S}_a$, spanning the "aperture", is the rest of the xy plane. As the disk obstructs the line of sight, we have $\psi_{(g)}(P,t) = 0$. As ρ is constant, so is the integrand in (144), so that the integral is simply the integrand multiplied by the range of θ. That range has magnitude 2π; but, because the aperture must be circled counterclockwise as seen from the field point (from which the xy plane is seen as it is usually drawn), and because the aperture is the *exterior* of the disk, the interior must be circled clockwise, so that the range of θ is *minus* 2π. With these simplifications, (144) becomes

$$\psi_{(a)}(P,t) \;\approx\; \tfrac{1}{a+b}\, f\big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2cab}\,\rho^2\big)\,, \tag{146}$$

which is simply the *unobstructed* wave function with the additional delay $\frac{a+b}{2cab}\,\rho^2$, due to propagation via the edge of the disk instead of by the line of sight; *cf.* (137). As ρ approaches zero, this additional delay also approaches zero; thus the result correctly predicts that the obstructed wave function approaches the unobstructed wave function as the obstruction shrinks to a point, and this confirms that disk has been circled in the right direction. The additional delay, by itself, does not cause any change in the intensity, or in the mixture of frequencies, of the incident radiation. In the case of light, that means no change in the brightness or color. Thus, subject to the accuracy of the Fresnel approximation, the center of the geometric shadow of the disk will be illuminated with the same brightness and color as if the disk were not there!

Other points in the geometric shadow can be modeled by allowing the center of the disk to be off the origin. Then ρ varies with θ, and the additional path length $\frac{a+b}{2ab}\,\rho^2$ (due to deviation of the path from the line of sight) consequently varies over Γ, so that the contributions to the integral tend to cancel out, at least if they are spread over a large number of cycles. In optical experiments, for reasonable dimensions of the apparatus, the variation in the path length tends to be minuscule in percentage terms, but large compared with the wavelength. For example, if $a = b = 500\,\text{mm}$ and $\rho = 5\,\text{mm}$, the additional path length will be $0.05\,\text{mm}$, which is 0.005% of the line-of-sight path length, but 100 wavelengths @ $\lambda = 500\,\text{nm}$; hence any considerable decentering of the disk will cause the path length to vary by many wavelengths, so that the contributions to the integral will indeed be spread over many cycles and will tend to cancel out. Consequently the center of the geometric shadow, illuminated as if the disk were not there, will appear as a bright spot in an otherwise dark 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 [19] and the references therein).

Returning to the case in which ρ is constant, let us now consider its *complement*, in which the aperture and the baffle are interchanged: the aperture, indicated by the subscript (a'), is the interior of the circle and the baffle is the exterior, so that the field point is illuminated through the center of a *circular aperture*. In the latter case the geometrical-optics term is the upper option in (145), and the circle is traversed *counterclockwise*, so that the range of θ is $+2\pi$. With these substitutions, (144) reduces to

$$\psi_{(a')}(P,t) \;\approx\; \tfrac{1}{a+b}\, f\big(t - \tfrac{a+b}{c}\big) \;-\; \tfrac{1}{a+b}\, f\big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2cab}\,\rho^2\big)\,. \tag{147}$$

Adding (146) and (147), we find that $\psi_{(a)}(P,t)+\psi_{(a')}(P,t)$ is the unobstructed wave function—as it must be, because the union of the two apertures is the entire xy plane.

According to (147), the wave function on the axis of the circular aperture is the difference between the unobstructed wave function and a delayed version thereof, the delay being due to the slightly greater length of the path via the edge of the aperture. Again, as ρ approaches zero, this delay approaches zero, so that the difference approaches zero. So (147) predicts that the field point, although it retains a line of sight to the source, gets darker as the aperture becomes very small—matching the common observation that as a small aperture gets smaller, the diffraction pattern gets bigger, spreading less light over a larger area.

For *monochromatic* light, the two terms in (147) cancel not only if the delay is zero, but also if the delay is a whole number of cycles; but they reinforce each other if the delay is an odd number of half-cycles. That is, the terms cancel if the path lengths via the center and edge of the aperture differ by a whole number of wavelengths, but they reinforce each other if the path lengths differ by an odd number of *half*-wavelengths. Hence, if the aperture is sufficiently large, there will be a sequence of alternating nodes and antinodes along the axis of the beam admitted through the aperture, because as the field point moves away from the source, the path difference becomes smaller, passing through a succession of even and odd multiples of the half-wavelength.

For *white* light, the variation in the path difference along the axis 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 aperture is sufficiently large, there will be so many visible wavelengths at 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.[56]

At a field point *off* the axis of the circular aperture, the contributions to the rim integral in (144) again tend to cancel out, so that the wave function is roughly as predicted by geometrical optics.

In general, bright spots in the depths of 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 to the Maggi-Rubinowicz integral from that part of the edge 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 total contribution from the edge integral becomes smaller and smaller, and the wave function is approximated more and more accurately by the geometric term $\psi_{(g)}(P,t)$—in conformity with the principle that geometrical optics (ray optics) is the limit of physical optics (wave optics) as the wavelength approaches zero [1, pp. 79–84]. This principle is more apparent from the Maggi-Rubinowicz form of the diffraction integral, in which the geometrical-optics contribution is shown as a separate term, than from a surface-integral form.

That being said, let us see what we can deduce from the surface integral in (143) concerning the disk and the circular aperture, for an observation point on the axis. Exploiting the axial symmetry, we can choose $d\mathcal{S} = dS$, where S (in ordinary italics) is the area inside radius ρ, so that $S = \pi\rho^2$, $dS = 2\pi\rho\, d\rho$, and (most usefully) $\rho^2 = S/\pi$; substituting the last of these relations into (143) gives

$$\psi_{(a)}(P,t) \approx \frac{1}{2\pi cab} \iint_{S_a} f'\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab} S\Big)\, dS\,. \tag{148}$$

Now, by the chain rule,

$$\tfrac{\partial}{\partial S} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab} S\Big) = f'\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab} S\Big)\, \tfrac{\partial}{\partial S}\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab} S\Big) = -\tfrac{a+b}{2\pi cab} f'\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab} S\Big)$$

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

so that

$$f'\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big) = -\tfrac{2\pi cab}{a+b}\,\tfrac{\partial}{\partial S} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big)\,; \tag{149}$$

and substituting this into (148) gives

$$\psi_{(a)}(P,t) \approx -\frac{1}{a+b}\iint\limits_{\mathcal{S}_a} \tfrac{\partial}{\partial S} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big)\,dS\,. \tag{150}$$

For the aperture, we integrate from $S = 0$ to $S = \pi\rho^2$, obtaining

$$\psi_{(a')}(P,t) \approx -\frac{1}{a+b}\int_0^{\pi\rho^2} \tfrac{\partial}{\partial S} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big)\,dS = -\frac{1}{a+b} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big)\Big|_0^{\pi\rho^2}, \tag{151}$$

which yields (147) again. This may be subtracted from the unobstructed wave function to obtain (146) again, for the axis of the disk (Poisson's spot).

Alternatively, we might try to find the result for the axis of the disk by working directly from (150). First we would integrate w.r.t. S from $\pi\rho^2$ to ∞, obtaining

$$\psi_{(a')}(P,t) \approx \frac{1}{a+b} f\Big(t - \tfrac{a+b}{c} - \tfrac{a+b}{2\pi cab}S\Big)\Big|_\infty^{\pi\rho^2}, \tag{152}$$

and observe that as $S \to \infty$, the argument of the primary source-strength function f goes to $-\infty$. Hence we might invoke the assumption that f had a beginning, conclude that f is zero for sufficiently large S, and get (146) again. But that would be disingenuous, because the beginning of f can be an arbitrarily long time ago, for which S and hence ρ can be arbitrarily large, hence large enough to cause significant obliquity and significant variations in $1/r$ and $1/s$ in (F2), and significant departures from the distance-squared law in (F3). As $\rho \to \infty$, the factors $1/r$, $1/s$, and $(\cos\phi + \cos\chi)$ in (131) approach zero, so that the associated surface elements do not "contribute significantly to the integral". But that leaves intermediate values of ρ which are large enough to invalidate the distance-squared law (F3) yet not large enough to make the contributions from the surface elements negligible. The simple workaround, as we have seen, is to evaluate the integral from $\pi\rho^2$ to ∞ as the integral from 0 to ∞ (which would be the unobstructed wave function, if our approximations were accurate) *minus* the integral from 0 to $\pi\rho^2$ (for which our approximations are reasonable).

For monochromatic light, we know from (147) that the wave function on the axis of the circular aperture is zero if the propagation delays via the center and edge differ by a whole number of cycles. From the integral in (151), we can state this result another way: the contributions to the wave function on the axis will cancel out if they are spread over a whole number of cycles. This cancellation occurs because, under our approximations, the relative propagation delay at radius ρ and the area inside radius ρ are both proportional to ρ^2, so that equal increments in the propagation delay (in the argument of f) are associated with equal increments of area, giving contributions of equal magnitude; hence each contribution from one half of a cycle is canceled by an equal-and-opposite contribution from the other half of the cycle.

Notice that in the foregoing discussion of Poisson's spot and its complementary case, although we mention monochromatic light, we do not find it necessary to invoke the monochromatic forms of the diffraction integrals; we argue in terms of general time-dependency and particular propagation delays.

6 Appendices

A Physical interpretations of integration over the aperture

To model diffraction by an aperture in a baffle, we have chosen a surface $\mathcal{S}$ comprising two segments, namely $\mathcal{S}_a$ spanning the aperture and $\mathcal{S}_b$ on the $\mathcal{R}$ side of the baffle, and brazenly *assumed* that the wave function in $\mathcal{R}$ is as if the baffle simply eliminated the secondary sources on $\mathcal{S}_b$, leaving the secondary sources on $\mathcal{S}_a$ as if the baffle were not there (Assumption 6, "secondary-source selection", p.15). In the

application of the Helmholtz or Kirchhoff integral, or the Kirchhoff diffraction formula (42), or its special case if $\mathcal{S}$ is a spherical primary wavefront (44) or if the primary wavefront is plane (46, 47), this assumption leads to integration over $\mathcal{S}_a$ instead of the whole of $\mathcal{S}$, with the integrand unchanged. Assumption 6 may therefore be understood as a physical interpretation of the integral over $\mathcal{S}_a$. Designating this interpretation as number **I**, we shall now derive three more.

II. We have seen that integration over $\mathcal{S}_a$ selects the secondary sources on $\mathcal{S}_a$, which impose the saltus conditions of Proposition 5. Thus diffraction through an aperture spanned by $\mathcal{S}_a$ is treated as a saltus problem in which, as we cross $\mathcal{S}_a$ from $\mathcal{R}'$ to $\mathcal{R}$, the step-changes in the wave function and its normal derivative are respectively equal to the unobstructed wave function and its normal derivative. The primary sources are eliminated (for field points in $\mathcal{R}$), and there is no saltus at $\mathcal{S}_b$.

III. The integral over $\mathcal{S}_a$ is expressible as a difference between two other integrals. If H denotes the integrand in (27),

$$\iint_{\mathcal{S}_a} H\,d\mathcal{S} = \iint_{\mathcal{S}} H\,d\mathcal{S} - \iint_{\mathcal{S}_b} H\,d\mathcal{S} \tag{153}$$

$$= \iint_{\mathcal{S}} H\,d\mathcal{S} + \iint_{\mathcal{S}_b} (-H)\,d\mathcal{S}\,, \tag{154}$$

where the integral over $\mathcal{S}$ is recognizable as the *unobstructed* wave function due to the primary sources, and the integral over $\mathcal{S}_b$ represents *inverted* secondary sources on $\mathcal{S}_b$—that is, the sources on $\mathcal{S}_b$ in Proposition 4 instead of Proposition 5. By Proposition 4, if this pattern of secondary sources were spread over all of $\mathcal{S}$, it would *stop* the waves from reaching $\mathcal{R}$, without causing any reflection back into $\mathcal{R}'$. So this interpretation models a baffle with no aperture as a sheet of secondary sources which would block the waves without causing any reflections, and models an aperture as the *absence* of those blocking secondary sources over part of the sheet. This interpretation makes sense in both $\mathcal{R}$ and $\mathcal{R}'$—unlike interpretations I and II, which give a null wave function in $\mathcal{R}'$ if the aperture shrinks to nothing (because all the sources vanish) *or* the aperture spreads over the whole of $\mathcal{S}$ (the case of Proposition 5).

IV. The sources on $\mathcal{S}_b$ in III impose the following saltus conditions: as we cross $\mathcal{S}_b$ from $\mathcal{R}'$ to $\mathcal{R}$, the step-changes in the wave function and its normal derivative are, respectively, the unobstructed wave function and its normal derivative *with their signs reversed*. Thus diffraction by the aperture is treated as a saltus problem involving the primary sources and inverted saltus conditions at $\mathcal{S}_b$. This in fact was the first saltus interpretation of diffraction to be published—by Friedrich Kottler, in 1923 [12, pp. 410–11]. Kottler's derivation, however, was more difficult and less general: it applied the Helmholtz formula to a more complicated surface, with a single sinusoidal monopole primary source.[57]

Interpretations I to IV, although mathematically correct, are not rigorously deducible from physical laws. Of the four, the most plausible on physical grounds, at least for electromagnetic waves, is III. If a baffle obstructs electromagnetic waves, it is as if the incident waves *induce* secondary sources in the material of the baffle, which tend to oppose the primary waves (Lenz's law); and there are indeed no such sources across the aperture, where there is no such material in which to induce them.[58] But this does not imply that the induced sources include the proportions of monopoles and dipoles that suppress reflections, and common experience indicates that significant reflections are the norm. Moreover, a nearly complete suppression of transmission requires a non-zero thickness, which means either that the aperture is not precisely defined (if the edge of the baffle is blunt) or that the parts of the baffle near the aperture are too thin to suppress transmission (if the edge of the baffle is sharp). Efficient suppression of reflections also requires a non-zero thickness, so that the waves sent back from induced sources can come from a range of depths, allowing destructive interference. Obviously, if the integral over the aperture is to give accurate results, the dimensions of the aperture must be large relative to the uncertainties caused by the non-zero thickness or imperfect opacity of the baffle.

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

[58] This is the basis of Feynman's explanation of "Diffraction of light by a screen" at the end of his lecture on "The Origin of the Refractive Index" [5, vol. 1, Chapter 31].

So far, we have derived or interpreted the terms of the Kirchhoff integrand (32) as contributions from monopole and dipole secondary sources on the surface $\mathcal{S}$. On its face, however, the integrand refers not to sources, nor to related saltus values, but to *boundary values* of the wave function, its time-derivative, and its normal derivative. That is how it was interpreted by Kirchhoff himself, who obtained his integral theorem as a purely mathematical property of wave functions, without our preliminary constructions involving secondary sources. So he faced a catch-22: in a diffraction problem, how are we to know the three boundary values in the integrand without knowing the whole wave function—that is, without knowing the very thing that we seek to calculate? If we know ψ on $\mathcal{S}$ as a function of time, we know $\dot{\psi}$ on $\mathcal{S}$, but how are we to know ψ on $\mathcal{S}$ without knowing it elsewhere? And how are we to know the normal derivative of ψ on $\mathcal{S}$ without knowing ψ on *and off* $\mathcal{S}$? Kirchhoff's solution was to suppose that the wave function *and* its normal derivative on segment $\mathcal{S}_b$ (as we call it) were zero, while the wave function *and* its normal derivative on segment $\mathcal{S}_a$ (as we call it) were as if the baffle were not there. His assumptions led to the same result as ours: that the integral is over the aperture only, while the integrand is as if the baffle were not there.

B Inconsistency of Kirchhoff's approach

As the BC on *either* the wave function *or* its normal derivative at $\mathcal{S}$ determines the wave function in all of $\mathcal{R}$ (Proposition 2), either BC determines the other. Kirchhoff tried to set both *a priori*. It would be an unbelievable fluke if his settings were consistent; and indeed it was proven by Henri Poincaré, no later than 1892, that they are not, except in the trivial case of a null wave function.[59]

The saltus interpretations in Appendix A (above) allow a simpler proof with minimal loss of rigor. By interpretation II, the wave function ψ is the solution to a saltus problem with step-changes in ψ and ψ_n as we cross $\mathcal{S}_a$. Hence, at a typical time, at a typical point on the edge of $\mathcal{S}_a$, the step-change in ψ is non-zero,[60] so that, as we approach this point along the $\mathcal{R}$ side of $\mathcal{S}_b$, the normal derivative tends to infinity—whereas Kirchhoff assumed it is zero. Similarly, under interpretation IV, at a typical time, as we approach a typical edge-point along the $\mathcal{R}$ side of $\mathcal{S}_a$, the normal derivative tends to infinity whereas Kirchhoff assumed it is zero.

There has been much ado about this inconsistency [4, 15, 21]. 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 essence of progress in physics. It is *not* enough to say that his BCs were only meant to be approximate and therefore only needed to be approximately consistent with the wave function calculated from them, because in fact they are not even approximately consistent. We have just seen that as we approach the edge of the aperture along the $\mathcal{R}$ side of $\mathcal{S}$, the predicted normal derivative is wildly divergent from the Kirchhoff assumption from which it can be calculated. Even in the interior of $\mathcal{S}_a$, the assumed boundary value of the wave function can differ substantially from that predicted by the Kirchhoff integral or an equivalent [15, Fig. 3(b)], and from the actual wave function [21, Figure 2], although the latter two may agree better with each other [15, Fig. 3]. Yet, at typical distances from $\mathcal{S}$, the Kirchhoff integral over $\mathcal{S}_a$ gives such accurate results that it continues to be used and taught into the 21st century. This empirical success in spite of inaccurate assumptions is widely seen as an anomaly requiring explanation [21].

In attempting an explanation, I should emphasize—because I have not seen it emphasized elsewhere, although it may have been—that Kirchhoff assumed far more than is necessary. If (in our notation) the integral over $\mathcal{S}$ is to be equal to the integral over $\mathcal{S}_a$, it is sufficient that the integ*ral* over $\mathcal{S}_b$ be zero; it is not necessary that the integ*rand* be zero at every point on $\mathcal{S}_b$. And if the integ*ral* over $\mathcal{S}_a$ is to be the same as for the unobstructed wave function, it is not necessary that the integ*rand* be the same at every point; it is sufficient that the variations of the integ*rand* from the unobstructed wave function "average out" in the integration (which tends to happen; *cf.* [21], s. 4.2). Moreover, even if it were necessary that the

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

[60] This applies to *every* edge-point if the wave function is due to a single monopole primary source.

integr*and*s agree at every point, it would still not be necessary that they agree term-by-term; it would be sufficient that the *sum* of the terms agree. But Kirchhoff assumed that the boundary conditions agree with the null wave function (on $\mathcal{S}_b$) or the unobstructed wave function (on $\mathcal{S}_a$) not only point-by-point, but also term-by-term. The inconsistency of Kirchhoff's theory therefore concerns details that do not need to be correct in order to give correct results. In those circumstances, it is no cause for astonishment that his conclusion is better than his assumptions. That is especially the case when, as shown in Appendix A (above), we can reach the same conclusion from other assumptions which, because of their different subject-matter, are not at risk of contradicting the result.

C Exactness of the Maggi-Rubinowicz transformation

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 (53) and (63) as

$$\psi_{(d)}(P,t) \;=\; -\iint_{\mathcal{S}_c} \frac{1}{4\pi r}\,\frac{\partial}{\partial n}\left(\frac{f(t-(r+s)/c)}{s}\right) d\mathcal{S}\,, \tag{155}$$

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 (respectively) to a general point N at a normal distance n from $\mathcal{S}_c$, measured away from the shadow, and the derivative is evaluated at $n=0$, which means that N is placed *on* $\mathcal{S}_c$ (as shown in Fig. 4 on p. 19). The factor r has been taken outside the differentiation because it is independent of n for infinitesimal n. For the same reason, from now on, we may treat r as a coordinate of Q instead of N, where Q is the foot of the perpendicular from N to $\mathcal{S}_c$. If $n=0$ (as in Fig. 4), N coincides with Q.

We have seen that the surface integral (155) can be transformed to the edge integral (64) 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 that of Gian Antonio Maggi (1888), or sinusoidal time-dependence, like those of Wojciech "Adalbert" Rubinowicz (1917), Friedrich Kottler (1923), and Born & Wolf (1959–2002).[61] We shall now rework the transformation exactly. Our method is essentially that of Rubinowicz, except that we allow general time-dependence.

As before, let the curve Γ be the edge of the aperture and 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");[62] 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. 5) is

$$d\mathcal{S} \;=\; \frac{r}{r_a}\,d\ell\,\sin(r,d\ell)\,dr\,; \tag{156}$$

the magnitude of this area is greater, by a factor r/r_a, than that given by (59) at Γ, because the distance between neighboring generators of the cone is proportional to r. Substituting (156) into (155) gives

$$\psi_{(d)}(P,t) \;=\; -\int_{\Gamma}\int_{s_a}^{\infty} \frac{\sin(r,d\ell)}{4\pi r_a}\,\frac{\partial}{\partial n}\left(\frac{f(t-(r+s)/c)}{s}\right) dr\,d\ell\,, \tag{157}$$

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. 5); and the outer integration over Γ adds up the contributions from all the strips.

[61] 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] profess to assume that the dimensions of the aperture are large compared with the wavelength but small compared with the distances to the source and the field point, and that "the angles of incidence and of diffraction are small." But they do not actually invoke any of these assumptions. Indeed, these assumptions imply the weaker condition that the wavelength is small compared with r and s, which we *do* invoke in Section 3.5 above, greatly simplifying the mathematics!

[62] Born & Wolf [2, pp. 502–3] use a subscript 1 instead of a.

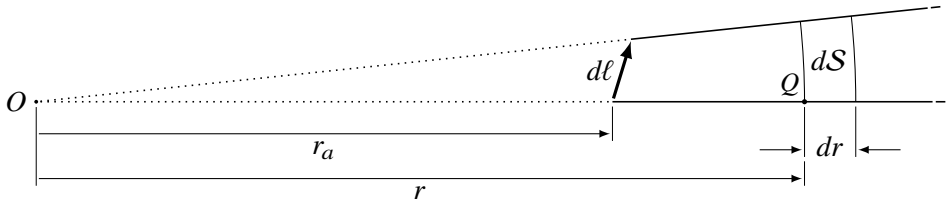

Fig. 5: Area element $d\mathcal{S}$ of the conical surface $\mathcal{S}_c$ (looking in the $-n$ direction in Fig. 4, p. 19).

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

$$\frac{\partial}{\partial n}\left(\frac{f(t-(r+s)/c)}{s}\right) = \frac{\partial}{\partial s}\left(\frac{f(t-(r+s)/c)}{s}\right)\frac{\partial s}{\partial n} = \frac{\partial}{\partial s}\left(\frac{f(t-(r+s)/c)}{s}\right)\cos(n,s)\,. \tag{158}$$

Substituting this into (157) and multiplying the numerator and denominator by s, we get

$$\psi_{(d)}(P,t) = -\int_\Gamma\int_{s_a}^{\infty} \frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\,\frac{1}{s}\frac{\partial}{\partial s}\left(\frac{f(t-(r+s)/c)}{s}\right)dr\,d\ell\,. \tag{159}$$

As noted by Rubinowicz [20, p. 261], the factor $s\cos(n,s)$ is the scalar component of the vector PQ in the n direction (see Fig. 4); and this direction and the component are independent of r for given ℓ. 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

$$\psi_{(d)}(P,t) = \int_\Gamma \frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\int_{s_a}^{\infty}\left\{\frac{f'\big(t-(r+s)/c\big)}{cs^2}+\frac{f\big(t-(r+s)/c\big)}{s^3}\right\}dr\,d\ell\,, \tag{160}$$

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

$$s^2 = (r-r_a)^2 + {s_a}^2 + 2(r-r_a)s_a\cos(r_a,s_a)\,. \tag{161}$$

Hence we could take the square root of the right-hand side, substitute it for every instance of s in the braces in (160), 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

$$-\frac{1}{s^2\left(1+\frac{ds}{dr}\right)}\,\frac{d}{dr}\,f\left(t-\frac{r+s}{c}\right). \tag{162}$$

Now we invoke (161) by differentiating it w.r.t. r (obviously with r_a and s_a constant), to obtain

$$s\,\frac{ds}{dr} = G(r)\,, \tag{163}$$

where, for brevity,

$$G(r) = r - r_a + s_a\cos(r_a,s_a)\,, \tag{164}$$

whence[63]

$$G'(r) = 1\,. \tag{165}$$

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

By (163), we can put $\frac{ds}{dr} = G(r)/s$ in (162), which then becomes

$$-\frac{1}{s\big(s+G(r)\big)}\,\frac{d}{dr}\,f\Big(t-\frac{r+s}{c}\Big)\,. \tag{166}$$

By the product rule, this can be recognized as

$$-\frac{d}{dr}\left\{\frac{f\big(t-(r+s)/c\big)}{s\big(s+G(r)\big)}\right\}+f\Big(t-\frac{r+s}{c}\Big)\,\frac{d}{dr}\left\{\frac{1}{s\big(s+G(r)\big)}\right\} \tag{167}$$

or, if we evaluate the last derivative using the chain and product rules and then apply (163) and (165),

$$-\frac{d}{dr}\left\{\frac{f\big(t-(r+s)/c\big)}{s\big(s+G(r)\big)}\right\}-\frac{f\big(t-(r+s)/c\big)}{s^3}\,. \tag{168}$$

Since this is equal to the *first term* in the expression in the braces in (160), the entire inner integrand is equal to

$$-\frac{d}{dr}\left\{\frac{f\big(t-(r+s)/c\big)}{s\big(s+G(r)\big)}\right\}, \tag{169}$$

so that the inner integral is

$$-\left.\frac{f\big(t-(r+s)/c\big)}{s\big(s+G(r)\big)}\right|_{r_a}^{\infty}=-\left.\frac{f\big(t-(r+s)/c\big)}{s\big(s+r-r_a+s_a\cos(r_a,s_a)\big)}\right|_{r_a}^{\infty}=\frac{f\big(t-(r_a+s_a)/c\big)}{{s_a}^2\big(1+\cos(r_a,s_a)\big)}\,, \tag{170}$$

and the outer integral becomes

$$\psi_{(d)}(P,t)=\int_\Gamma\frac{s\cos(n,s)\sin(r,d\ell)}{4\pi r_a}\,\frac{f\big(t-(r_a+s_a)/c\big)}{{s_a}^2\big(1+\cos(r_a,s_a)\big)}\,d\ell\,. \tag{171}$$

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

$$\psi_{(d)}(P,t)=\int_\Gamma\frac{f\big(t-(r+s)/c\big)}{4\pi rs}\,\frac{\cos(n,s)\sin(r,d\ell)}{1+\cos(r,s)}\,d\ell\,. \tag{172}$$

This exact result agrees with (64), 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 have somehow canceled out.

D Acknowledgments

This paper rises from the ashes of a previous one called "Exact derivation of Kirchhoff's integral theorem and diffraction formula using high-school math", which had four versions, dated 8 February to 9 April, 2020. On 3 July 2022, Fabio Baldassi (linkedin.com/in/fabio-baldassi-1a7a87154) sent me a proposed adaptation of that paper, which alerted me to the need to make structural changes to the original in order to clarify the dependencies. In the course of this effort, in mid August, I belatedly noticed that the proof of a crucial proposition contained in versions 2 to 4 was flawed, and that although the proposition appeared to be correct for independent reasons, it would be easier to eliminate the need for the proposition than to repair or replace the proof. One illustration of the lack of clarity in the previous paper was that Mr. Baldassi thought it treated diffraction primarily as a saltus problem. That was not the intention of versions 2 to 4, but was closer to the intention of version 1, and was precisely the intention of a pre-publication draft![64]

[64] Source file dated 26 January 2020.

Although I had abandoned that path because of logical gaps, it seemed feasible in retrospect to bridge the gaps. Having done this, I noticed—in late October, a fortnight after the publication of a partial draft of the present paper—that Miller's spatiotemporal-dipole secondary sources [16], which I had adopted early in the previous paper, and which were general enough (with one caveat) for Miller's purposes, were not general enough for mine. That pulled the remaining support from under my "crucial proposition". Thus, in eleven weeks, my assessment of the previous paper had sunk from "best I've ever written" to "unfit for public consumption". But a new structure was already in place: the "crucial proposition" had been dumped, and spatiotemporal dipoles had been reduced to an afterthought, albeit with a novel generality.

Independently of the required corrections and clarifications, the present paper simplifies the far-field derivation of the Maggi-Rubinowicz transformation for general time-dependence,[65] and presents it as early as possible (Section 3.5). This feature seemed notable enough to warrant a new title.

Version 0.1 of the present paper was shown only to Mr. Baldassi. Version 0.2 was the first to be published. Version 0.3 further simplified Section 3.5, added Sections 3.6 (on directional primary sources), 3.7 (on spatiotemporal dipoles), and 4.1 to 4.5 (on the sinusoidal case), and radically revised Section 1.2 (acknowledging the assumption that we can neglect the variation of M with σ). Version 0.4 added Sections 5.1 and 5.2 (on Fresnel diffraction), and changed the traversal direction, and consequently the order of the cross product, in the vector form of the Maggi-Rubinowicz transformation, in order to match Baker & Copson (whom I had misunderstood) and Kottler (whom they cite). Version 0.41 (current) makes minor textual corrections and improvements.

7 Conclusion

Kirchhoff's integral theorem and its corollaries, usually thought to be based on sophomore-level mathematics, can in fact be developed from concepts that are familiar to high-school graduates or easily introduced to them as needed: the exposition proceeds from the sequential nature of wave propagation, through superposition, boundary conditions, saltus conditions, Huygens' principle, and the specification of the secondary sources, to the Helmholtz and Kirchhoff integral theorems. The assumption of a single monopole primary source then leads to the Kirchhoff diffraction formula, the far-field obliquity factor, Fresnel's choice of the surface of integration, and other special cases. In the problem of diffraction by an aperture—in which the inconsistencies in Kirchhoff's boundary conditions have hitherto been either tolerated, or circumvented by introducing further complexity—working from the secondary sources avoids the cause of the inconsistency and leads to other consistent formulations of the problem. The resulting 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; and the latter term, with the aid of the usual sort of far-field approximation employed in Kirchhoff diffraction calculations, is easily transformed to an edge integral, which agrees with that obtained by Rubinowicz's more difficult, exact method. The assumption of sinusoidal time-dependence is not needed for any of these purposes, but can be added afterwards as a special case. I hope that by this reworking of foundations, the essential mathematical theory of Huygens' principle and diffraction has been rendered more accessible.

[65] The first attempt was published in version 3 of the earlier paper, on 29 March 2020.

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. (Cited page numbers are from the translation.)

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

[8] H. Helmholtz, "Theorie der Luftschwingungen in Röhren mit offenen Enden", *Journal für die reine und angewandte Mathematik*, vol. 57, no. 1 (1859), books.google.com/books?id=7c8GAAAAYAAJ, pp. 1–72.

[9] C. Huygens (1690), tr. S.P. Thompson, *Treatise on Light*, University of Chicago Press, 1912; Project Gutenberg, 2005, gutenberg.org/files/14725/14725-h/14725-h.htm. (See also "Errata in various editions of Huygens' *Treatise on Light*" at *www.grputland.com* or *grputland.blogspot.com*, June 2016.)

[10] G.R. Kirchhoff, "Zur Theorie der Lichtstrahlen", *Annalen der Physik und Chemie*, vol. 18, no. 4 (1883), pp. 663–95; archive.org/details/sim_annalen-der-physik_1883_18_4/page/663.

[11] G.R. Kirchhoff, *Vorlesungen über mathematische Physik*, vol. 2 (ed. K.W.S. Hensel), Leipzig: Teubner, 1891; books.google.com/books?id=CB5WAAAAMAAJ.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] D.A.B. Miller, "On perfect cloaking", *Optics Express*, vol. 14, no. 25 (11 Dec. 2006), pp. 12457–66; doi.org/10.1364/OE.14.012457.

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

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

[20] 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.

[21] 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.)

[22] G.G. Stokes, "On the dynamical theory of diffraction" (read 26 Nov. 1849), *Transactions of the Cambridge Philosophical Society*, vol. 9, part 1 (1851), pp. 1–62; archive.org/stream/transactionsofca09camb#page/n15.

[23] 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.