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

Gavin R. Putland *

Version 1; February 8, 2020

A comprehensive theory of diffraction is developed from three elementary premises: superposition, the form of the wave function due to a monopole source, and the assumption that the wave function had a beginning in time. It is shown that the wave function in a region $\mathcal{R}$, due to sources outside $\mathcal{R}$ (i.e., in $\mathcal{R}'$), is identical to that due to a distribution of sources on the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$, such that the step-change (saltus) in the wave function, in crossing $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, is equal to the original wave function on $\mathcal{S}$. The necessary sources are shown to be spatiotemporal dipoles (STDs), as discovered by D.A.B. Miller (1991). The Kirchhoff integral theorem and consequent diffraction formulae are then obtained by superposing the elemental wave functions due to the sources on $\mathcal{S}$. Whereas Miller justified the STDs by comparison with Kirchhoff's integral, this paper derives the integral (including a near-primary-source correction to Miller's form thereof) solely from the STDs. The case of diffraction by an aperture in an opaque screen can be handled on the assumption that we retain only those sources on the part of $\mathcal{S}$ that spans the aperture. This assumption avoids the notorious inconsistency in Kirchhoff's boundary conditions (1882), but gives the same diffraction integral, and yields the consistent saltus condition of Kottler (1923).

Contents

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

1 Preliminaries

1.1 Wave functions; boundary conditions; Huygens' principle

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

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

That immediately raises three questions:

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

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

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

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

---

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

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

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

Concerning the second question, suppose that we have an aperture (such as a slit or an iris) in an otherwise opaque screen, with a point-source on one side of the screen, and suppose that we want to predict the wave function on the other side. Experience tells us that the edge of the shadow cast by the screen will not be perfectly sharp, even for the point-source, but will be blurred and fringed—an effect known as **diffraction**. The same phenomenon causes even the depths of the shadow to be less than perfectly dark or silent. In this situation, let us choose the surface $\mathcal{S}$ so that one part of it, say $\mathcal{S}_a$, spans the aperture, while the remaining part, say $\mathcal{S}_b$, is on the back (dark side or quiet side) of the screen. Then we might suppose, as a first approximation, that the wave function (or its normal derivative) on $\mathcal{S}_b$ is zero, while the wave function (or its normal derivative) on $\mathcal{S}_a$ is the same as if the screen were not there. The conditions at these boundaries determine the wave function in the region beyond the screen—including those parts in, and near the edge of, the geometric shadow of the screen,[3] where the wave function is certainly *not* the same as if the screen were not there. Thus we might hope to explain and predict diffraction in a quantitative manner.

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

Proposition 3 *Let the geometric surface $\mathcal{S}$ divide the medium into a region $\mathcal{R}$, containing no sources, and a region $\mathcal{R}'$. Let sources in $\mathcal{R}'$ give a certain wave function, hence a certain BC at $\mathcal{S}$. If those sources can be replaced by a distribution of sources on $\mathcal{S}$ which, by themselves, give the original BC at the $\mathcal{R}$ side of $\mathcal{S}$ and a zero BC at the $\mathcal{R}'$ side, then the sum of the contributions from the sources on $\mathcal{S}$ is the original wave function in $\mathcal{R}$, and a zero wave function in $\mathcal{R}'$.*

We call this sum a **surface integral** over $\mathcal{S}$. It is understood that the BC may be a condition on the wave function *or* its normal derivative.

That result raises another question: If we only want to reproduce the wave function in $\mathcal{R}$, why should we bother setting it to zero in $\mathcal{R}'$? The answer has two parts:

(i) A unique distribution of sources on $\mathcal{S}$, *by itself*, obviously gives a unique wave function on each side of $\mathcal{S}$, hence a unique BC on each side. Therefore, if we did not specify the BC on both sides, the required distribution of sources would not be unique. (We have not yet demonstrated that such a distribution is possible.)

(ii) If we can find a continuous distribution of sources on $\mathcal{S}$ that reproduces the original wave function in $\mathcal{R}$, it will show that the wave function in $\mathcal{R}$ is *as if* the incident wave turned every infinitesimal element of the surface $\mathcal{S}$ into a particular **secondary source**; in other words, it will verify and quantify **Huygens' principle** for the surface $\mathcal{S}$ and the region $\mathcal{R}$. If, in addition, that distribution of sources yields a null (zero) wave function in $\mathcal{R}'$, we will have found secondary sources that cause no "backward" or "retrograde" secondary waves.

1.2 Saltus conditions

If the distribution of sources on $\mathcal{S}$ imposes the original value of ψ or ψ_n at the $\mathcal{R}$ side, and a null value at the $\mathcal{R}'$ side, it will generally impose a step-change (or step-*discontinuity*) as we cross $\mathcal{S}$. Very

[3] The *geometric* shadow is the shadow that would be predicted by thinking in terms of *rays* instead of waves.

conveniently, the step-change at any point on $\mathcal{S}$ is due solely to the density (and, if meaningful, the orientation) of sources *at that point* (or to the densities and orientations—plural—if there is more than one kind of source). The step-change is *not* due to any sources at finite (non-infinitesimal) distances from that point, because those distances change by (at most) infinitesimal fractions as we cross the surface (*cf.* Larmor [7], p. 6). Not so conveniently, we cannot draw any similar conclusion about the separate *values* of ψ or ψ_n; getting closer to one point on $\mathcal{S}$ reduces the area around that point that subtends a given solid angle, and therefore does not necessarily dilute the influence of remote points on $\mathcal{S}$. Consequently, the conditions directly imposed by the sources on $\mathcal{S}$ are not the actual values of ψ or ψ_n, but the *changes* or *discontinuities* in those values; such conditions are called **saltus conditions** (from the Latin noun *saltus*, pronounced "SAHL-toos", meaning "jump" or "leap").

So, in order to find sources that impose the original BC at the $\mathcal{R}$ side of $\mathcal{S}$, and a null BC at the $\mathcal{R}'$ side, we must somehow re-express the BCs as a saltus condition. We can do this by starting with the "original" case, adding sources on $\mathcal{S}$ to impose a strategic saltus condition, then invoking superposition. In all cases to be considered, the geometric surface $\mathcal{S}$ divides the medium into a region $\mathcal{R}$, containing no sources, and a region $\mathcal{R}'$. The first case is trivial:

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

At a general point Q on $\mathcal{S}$, let the original wave function be $\psi(Q,t)$. Now let us modify Case 1 by adding sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$ at the point Q, the wave function changes by $-\psi(Q,t)$; or, equivalently, as we cross $\mathcal{S}$ in the other direction at point Q, the wave function changes by $+\psi(Q,t)$. Since the wave function had a beginning, we can take it to be 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 (due to the sufficiency of the ensuing null boundary condition) exerts no further influence in $\mathcal{R}$—or in $\mathcal{R}'$, except by the 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. 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 (saltus) in the wave function is *minus* the original wave function on $\mathcal{S}$. We get the original wave function in $\mathcal{R}'$ and a null wave function in $\mathcal{R}$.

Changing the signs in Case 2 and superposing Case 1, we obtain:[4]

Case 3: We have only sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the saltus in the wave function is equal to the original wave function on $\mathcal{S}$. We get a null wave function in $\mathcal{R}'$ and the original wave function in $\mathcal{R}$.

Thus we have specified the distribution of secondary sources on $\mathcal{S}$ that gives the original wave function in $\mathcal{R}$ with no backward secondary waves. (We still need to work out how to meet the specification.) And of course, as we get the original wave function in $\mathcal{R}$ and a null wave function in $\mathcal{R}'$, we also get the original boundary conditions (on the wave function and all its derivatives) at the $\mathcal{R}$ side of $\mathcal{S}$, and null boundary conditions at the $\mathcal{R}'$ side. So the sources that satisfy the saltus condition in Case 3 also satisfy the BCs in Proposition 3, on both sides of $\mathcal{S}$. In summary:

Proposition 4 *Let the geometric surface $\mathcal{S}$ divide the medium into a region $\mathcal{R}$, containing no sources, and a region $\mathcal{R}'$. Let sources in $\mathcal{R}'$ give a certain wave function in $\mathcal{R}'$ and $\mathcal{R}$. Now let there be a distribution of sources on $\mathcal{S}$ such that, as we cross $\mathcal{S}$ from $\mathcal{R}'$ to $\mathcal{R}$, the saltus in the wave function is equal to the original wave function on $\mathcal{S}$. Then the sources on $\mathcal{S}$, by themselves, give the original wave function in $\mathcal{R}$, and a zero wave function in $\mathcal{R}'$.*

[4] 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 [7], at p. 11.

1.3 Diffraction as a saltus problem

Logically enough, a problem with given boundary conditions is called a **boundary-value problem**, whereas a problem with given saltus conditions is called a **saltus problem**. Case 3, above, relates the boundary condition on the $\mathcal{R}$ side to the saltus condition, and thence to the local sources on $\mathcal{S}$. This case depends on Case 2, whose derivation tacitly exploits the assumption that $\mathcal{S}$ completely separates $\mathcal{R}$ from $\mathcal{R}'$, so that the waves cannot enter $\mathcal{R}$ except via $\mathcal{S}$. If $\mathcal{S}$ did *not* completely separate $\mathcal{R}$ from $\mathcal{R}'$, we would lose the simple relation between the BC and the saltus condition, but retain the local relation between the saltus condition and the sources on $\mathcal{S}$. So, *if we are trying to solve a diffraction problem by means of secondary sources, we are treating it fundamentally as a saltus problem*—and only indirectly, if at all, as a boundary-value problem.

Let us therefore reconsider the problem of diffraction through an aperture in an opaque screen, for which we partitioned the surface $\mathcal{S}$ into segments $\mathcal{S}_a$ and $\mathcal{S}_b$, with $\mathcal{S}_a$ spanning the aperture, and $\mathcal{S}_b$ on the back of the screen. Let us compare the case in which the screen is absent (the *unobstructed* case) with the case in which the screen is present—keeping S and its segments the same in both cases. In the *absence* of the screen, the wave function in $\mathcal{R}$ (but not in $\mathcal{R}'$) is as in Case 3 above, where the sources on $\mathcal{S}$ give a saltus equal to the original wave function on $\mathcal{S}$. Let the wave functions due to the sources on $\mathcal{S}_a$ alone and $\mathcal{S}_b$ alone be, respectively, $\psi_{(a)}(P,t)$ and $\psi_{(b)}(P,t)$; we shall write this as

$$\text{Sources on } \mathcal{S}_a \longrightarrow \psi_{(a)}(P,t) \tag{1}$$

and

$$\text{Sources on } \mathcal{S}_b \longrightarrow \psi_{(b)}(P,t)\,. \tag{2}$$

In the same notation, Case 3 is

$$\text{Sources on } \mathcal{S}_a \;+\; \text{Sources on } \mathcal{S}_b \longrightarrow \begin{cases} \psi(P,t) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}'. \end{cases} \tag{3}$$

In the *presence* of the screen (the partly obstructed case), the simplest assumption we can make is that the screen eliminates the sources on $\mathcal{S}_b$, leaving only the sources on $\mathcal{S}_a$, as in equation (1). In terms of saltus conditions, this assumption means imposing the usual saltus condition (of Case 3) at $\mathcal{S}_a$ only, and no saltus at $\mathcal{S}_b$. In terms of superposing the contributions of the sources by means of a surface integral, this assumption means integrating over $\mathcal{S}_a$ only, instead of the whole surface $\mathcal{S}$. Limiting the range of integration is indeed the conventional approach to the problem—although, unfortunately, it is usually justified by assuming inconsistent boundary conditions instead of a consistent saltus condition! We shall revisit that issue later.

According to our "simplest" assumption, since equation (1) is *with* the screen and equation (3) is *without*, the effect of the screen can be obtained by subtracting (3) from (1). That gives

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

which says that the screen is equivalent to sources on $\mathcal{S}_b$ alone, namely *minus* the sources on $\mathcal{S}_b$ in Case 3, imposing a saltus equal to *minus* the original wave function on $\mathcal{S}_b$ as we cross from $\mathcal{R}'$ to $\mathcal{R}$; the saltus may be inferred from Case 3, or by comparing the two regions in equation (4).

In summary, the integration over $\mathcal{S}_a$ alone is equivalent to the assumption that the screen imposes a saltus equal to *minus* what the wave function would be in the absence of the screen. Friedrich Kottler, in a paper published in 1923, discovered this fact by a more sophisticated method [1, pp. 98–101].

In the title of the present paper, the word "exact" applies to the calculation of the wave function *from* the boundary or saltus conditions, but does not generally apply to the boundary or saltus conditions themselves. That distinction was significant in our first mention of the diffraction problem, in which the assumed BCs were only a "first approximation" obtained by treating the propagation paths from

the source to $\mathcal{S}$ as completely obstructed or completely unobstructed. And it is significant in Kottler's saltus condition—which, while notable for its mathematical consistency [1, p. 101], is not deducible from physical premises [3, p. 503] and is equivalent, as we have just seen, to the heroic assumption that the screen simply eliminates part of the sheet of secondary sources.

1.4 Note on the absence of backward secondary waves

Suppose that there are no sources except on the surface $\mathcal{S}$, which separates regions $\mathcal{R}'$ and $\mathcal{R}$. Then the jump in the value of the wave function as we cross $\mathcal{S}$, from $\mathcal{R}'$ to $\mathcal{R}$, is equal to the boundary value on the $\mathcal{R}$ side if and only if the boundary value on the $\mathcal{R}'$ side is zero, in which case (by Proposition 2) the whole wave function on that side is zero. Conversely, a null wave function on the $\mathcal{R}'$ side gives a null boundary value on that side. Thus the absence of "backward secondary waves" converts a saltus condition, which can be set by the surface sources, into a boundary condition on the "forward" side, which is enough to determine the wave function in the "forward" direction. That is one mathematical justification of the physical intuition that there are no backward secondary waves.

The argument can be made more physical, as in the following statement by Joseph Larmor, where, in the notation of the present paper, the "inside" is $\mathcal{R}'$ and the "outside" is $\mathcal{R}$:

> A state of stress and strain is continually transmitted up to the surface $\mathcal{S}$ from the actual sources inside, and we are to find a distribution of secondary sources that will send it on to the outside as it arrives, without sending anything back. For the sending of a disturbance back into the interior would be an alteration of the physical circumstances, would in fact add to and confuse the effect of the assigned true sources inside [8, p. 172].

Indeed, because $\mathcal{S}$ may be curved, or even closed around $\mathcal{R}'$, any radiation from the secondary sources "back" into $\mathcal{R}'$ could continue across $\mathcal{S}$ into $\mathcal{R}$, altering the wave function in $\mathcal{R}$ and thereby invalidating the setting of its BCs.

And yet, as long as we are trying to reproduce a wave function in a region $\mathcal{R}$ containing no sources, due to sources in a region *completely separated* from $\mathcal{R}$ by a surface $\mathcal{S}$, Proposition 4 says that we need not explicitly concern ourselves with the BCs: if we merely set the prescribed *saltus* in the wave function as we cross $\mathcal{S}$ to the $\mathcal{R}$ side, we will set the correct BCs on that side, "without sending anything back."

1.5 Plan of the paper

Accordingly, the plan for the rest of the paper is as follows. In Section 2 we seek a distribution of sources on the surface $\mathcal{S}$ that yields the original value of ψ at the $\mathcal{R}$ side of the surface, with a zero value at the other side. Since we are really interested in the difference between the two, we can if necessary tolerate an unknown additive constant common to both (but it turns out not to be necessary). We find a solution in the form of **spatiotemporal dipoles** (**STDs**), as first described by D.A.B. Miller [10]. In Section 3, we use the STD distribution to express the wave function at a general point in $\mathcal{R}$ (the **field point**) as the surface integral, over $\mathcal{S}$, of an expression involving the wave function and two of its derivatives. We specialize the integrand for the case of a single primary source in Section 4. The dominant term of the integrand is seen to contain an **obliquity factor** involving two angles, made by the source and the field point with the normal to the surface of integration. In Section 5, we further specialize the integral for the case considered by Fresnel, in which the surface of integration is a primary wavefront, so that the obliquity factor reduces to a function of latter angle (the former being fixed at zero). We also consider the case which the primary wavefront may be non-spherical but not too sharply curved. In Section 6, we rewrite our results for the **monochromatic** case, in which the time-dependence of the wave function is *sinusoidal*. In Section 7, we add a near-source correction term to Miller's result for a spherical primary wavefront. This does not require any matching correction to Miller's spatiotemporal dipoles, because it is derived *from* them.

2 Matching the boundary-saltus conditions

2.1 First attempt: Monopole sources

Consider the function

$$f(t - s/c)\,, \tag{5}$$

where t is time, s is a coordinate measuring distance, and c is a constant. The *argument* of the function is $t-s/c$. Let some local feature of the function occur where the argument has a particular value. If we set the argument equal to that value (a constant) and differentiate with respect to t, we obtain $\frac{ds}{dt} = c$, indicating that the feature moves in the direction of s at speed c (indeed the symbol comes from the Latin *celeritas*, meaning speed). Thus the function (5) describes a wave or waves moving at that speed, and may therefore be called a wave function.

In the argument of the function f above, the presence of the $-s/c$ term means that the time t must be increased—that is, *delayed*—by s/c in order to obtain the same argument, hence the same function value, as if that term were absent. So the function (5) may be understood as $f(t)$ delayed by the time s/c, which is the time taken by the waves to travel the distance s. This delayed function may be written $[f(t)]$ or simply $[f]$, where the square brackets indicate that the contents are to be delayed by the propagation time. We shall use this notation later in the paper. Accordingly, we shall avoid using square brackets like ordinary parentheses.

In three dimensions, a wave function is generally a function of three spatial coordinates and time. But special cases may reduce the number of coordinates. In particular, in a *homogeneous* and *isotropic* medium,[5] we shall define a **monopole** source of **strength** $f(t)$ as a source generating the wave function

$$\frac{1}{r}\, f(t - r/c)\,, \tag{6}$$

where r is the radial distance from the source.[6] A wave function of form (6) describes *spherical* waves receding from the source at speed c. Note that we have defined the "strength" as a function of time. We shall need the assumption that the function is differentiable and therefore continuous. In (6), the function is not only delayed by the propagation time for the distance r, but also attenuated so that its amplitude is inversely proportional to that distance. Consequently, if the *intensity* (power per unit area) is proportional to the square of the wave function, the intensity satisfies the inverse-square law.

Now let $M(t)$ be the *strength density* of monopole sources on the surface $\mathcal{S}$, so that the source strength corresponding to a surface element of area $d\mathcal{S}$ is $M(t)\,d\mathcal{S}$. By formula (6), this element's contribution to the wave function at a distance x from the element is

$$d\psi = \frac{1}{x}\, M(t - x/c)\, d\mathcal{S}\,. \tag{7}$$

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

For a flat $\mathcal{S}$ and a uniform $M(t)$, we may take the element $d\mathcal{S}$ as 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 values of x and M in (7). Hence, if we are to remain consistent with our assumptions, we cannot have $\rho \to \infty$ or $x \to \infty$. The area of the chosen surface element is

$$d\mathcal{S} = 2\pi\rho\, d\rho\,, \tag{8}$$

[5] A *homogeneous* (or *uniform*) medium is one whose properties, including c, are independent of location within the medium. An *isotropic* medium is one whose properties are independent of direction. The properties of a homogeneous or isotropic medium are likewise said to be homogeneous or isotropic, respectively.

[6] In defining the *strength*, I follow the old convention used by (e.g.) Baker & Copson [1, p. 42], Born & Wolf [2, p. 421], and Larmor [7, p. 5]. Miller [10, p. 1371] implicitly uses the denominator $4\pi r$ instead of my r. Fortunately the effects of this clash of conventions will cancel out.

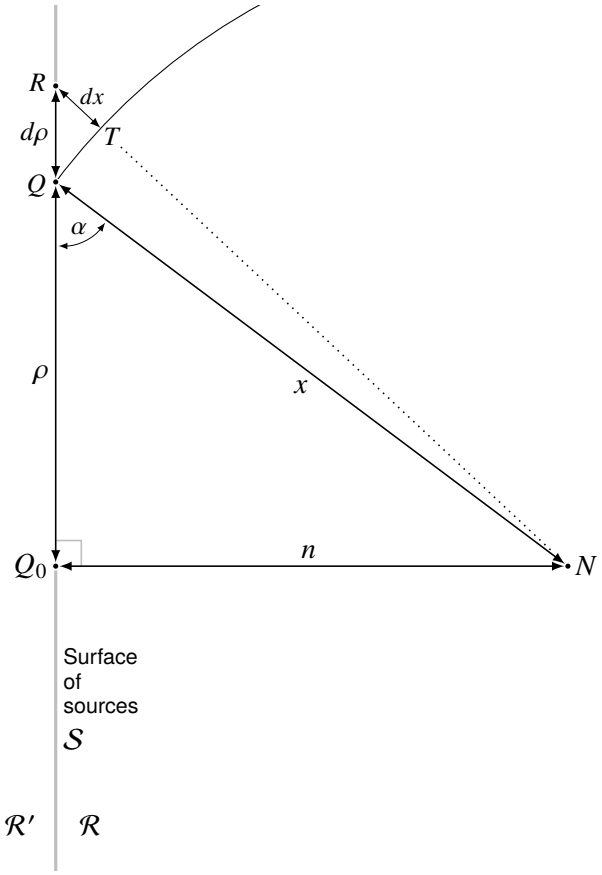

Fig. 1: If the normal coordinate n is so small that the surface $\mathcal{S}$ (seen edge-on) can be considered flat, we can take the surface element $d\mathcal{S}$ as an annulus with axis Q_0N, inner radius ρ, and width $d\rho$.

so that (7) becomes

$$d\psi = \frac{2\pi}{x}\, M(t - x/c)\, \rho\, d\rho\,. \tag{9}$$

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

$$\rho\, d\rho = x\, dx\,. \tag{10}$$

With this substitution, (9) reduces to

$$d\psi = 2\pi\, M(t - x/c)\, dx\,. \tag{11}$$

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

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

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

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

I say "According to" and "should" and "would", because we do not know the geometry of $\mathcal{S}$ or the behavior of M for $x \gg n$; and this matters because there is no factor making the integrand negligible for

large x. So (12) does not yet give a usable value of ψ. Neither does it yield a usable *saltus* in ψ, because (i) there is nothing to make the integrand infinite as $n \to 0$, and (ii) the uncertainty for $x \gg n$ cannot cause any discontinuity in ψ as we cross $\mathcal{S}$, because the effect of this crossing on x is infinitesimal.

But I chose a sheet of monopoles as a "first attempt", not because it had any chance of succeeding, but because it is a building-block with which we might construct something better. For the final product I am indebted to Miller [10]; but my method of assembling it is different from his.

2.2 Second attempt: Spatiotemporal dipole (STD) sources

Suppose that we have a second sheet of monopoles, called the **inverted** monopoles, whose strengths are equal and opposite to those in Fig. 1 and equation (12), and whose positions are displaced by a small distance h in the $-n$ direction (normal to $\mathcal{S}$, toward $\mathcal{R}'$) relative to their uninverted counterparts. Then the wave function due to both sheets will be a *difference* between two integrals like the one in (12). We might hope that the range over which the integrand is uncertain will cancel out in the subtraction, eliminating the uncertainty.

Hence we might hope that by appropriately delaying or advancing the inverted sources in time (with a compensating change in the variable of integration), we can make the ranges of integration cancel completely on the $\mathcal{R}'$ side of both sheets, but incompletely on the $\mathcal{R}$ side of both sheets, so that the wave function on the $\mathcal{R}$ side of both sheets is proportional to the difference (or "overhang") of the ranges of integration, hence proportional to that small displacement h. Hence we might hope that by making the strengths of the inverted and uninverted monopoles *inversely* proportional to h, and taking limits as $h \to 0$, we can get a zero value of the wave function at the $\mathcal{R}'$ side of $\mathcal{S}$ and a non-zero value at the $\mathcal{R}$ side, with a discontinuity at $\mathcal{S}$.

According to (12), we cannot obtain such a discontinuity in ψ^+ from a *single* sheet of sources with *finite* strength density. But the same equation should lead us to expect a discontinuity in the normal derivative ψ^+_n, because ψ^+ is a function of $|n|$, which in turn has a slope-discontinuity at the origin. Indeed, as a consequence of the fundamental theorem of calculus, the derivative of the definite integral in (12) w.r.t. its *lower* limit is *minus* the integrand evaluated at that limit; that is,

$$\psi^+_{|n|} = -2\pi\, M\Big(t - \tfrac{|n|}{c}\Big)\,. \tag{13}$$

Hence, by the chain rule,

$$\psi^+_n = -2\pi\, M\Big(t - \tfrac{|n|}{c}\Big)\, \tfrac{d|n|}{dn}\;; \tag{14}$$

and the last factor changes sign at $n = 0$, giving a step-change in the slope of the wave function w.r.t. n. With two opposing sheets of sources, together contributing two opposing changes in the slope, we hope to construct (so to speak) a *ramp* connecting two different values of the wave function. By making the strength densities inversely proportional to the width (h) of the ramp, we hope to make the slope of the ramp inversely proportional to its width, so that its height is fixed. Then, if we let the width approach zero, the ramp will approach a *step* of that height.

The right-hand side of (13) has the form of (5) with the distance coordinate $|n|$; it describes a wave traveling away from the sheet of uninverted sources, in both directions, at speed c. Recall that the sheet of inverted sources is displaced to the $\mathcal{R}'$ side by a distance h. Hence, if we want the wave from the inverted sheet to cancel that from the uninverted sheet on the $\mathcal{R}'$ side of both sheets, the necessary time-shift in the inverted sources is obvious: they must be *delayed* by h/c (the propagation time for the distance h), in order to compensate for the extra propagation time from the more distant uninverted sheet. Meanwhile, on the $\mathcal{R}$ side of both sheets, there will be no such cancellation, because the delay in the inverted sources will add to, rather than compensate for, the extra propagation time from the *inverted* sheet, which is the more distant sheet as seen from that side.

So let us see whether our hopes are mathematically verified. Recall that the contribution to the wave function from the uninverted sheet of sources is ψ^+, given by equation (12). We need to modify this equation so as to obtain the inverted sheet's contribution, which we shall call ψ^-. For the inversion,

we simply change the sign. For the delay, we replace t by $t-h/c$. For the displacement h, we replace n by $n+h$, so that putting $n=-h$ gives the argument that was formerly given by $n=0$. Thus we obtain

$$\psi^- = -\int_{|n+h|}^{\infty} 2\pi\, M\Big(t - \frac{x+h}{c}\Big)\, dx \tag{15}$$

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

$$\psi^- = -\int_{|n+h|+h}^{\infty} 2\pi\, M(t - \xi/c)\, d\xi\,. \tag{16}$$

Adding (12) and (16), and renaming the bound variable of integration ("dummy variable") in the latter, we obtain the wave function due to both sheets together:

$$\psi \;=\; \psi^+ + \psi^- \;=\; \int_{|n|}^{\infty} 2\pi\, M(t - x/c)\, dx \;-\; \int_{|n+h|+h}^{\infty} 2\pi\, M(t - x/c)\, dx\,. \tag{17}$$

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

$$\psi \;=\; \int_{|n|}^{|n+h|+h} 2\pi\, M(t - x/c)\, dx\,. \tag{18}$$

On the $\mathcal{R}'$ side of both sheets (the left side in Fig. 1), both n and $n+h$ are negative, so that $|n| = -n$ and $|n+h| = -n-h$, whence both limits of integration simplify to $-n$, so the "range" of integration collapses to nothing; thus, on the $\mathcal{R}'$ side of both sheets, the wave function is zero—as hoped. Meanwhile, on the $\mathcal{R}$ side of both sheets (the right side in Fig. 1), both n and $n+h$ are positive, so that $|n| = n$ and $|n+h| = n+h$; hence:

$$\text{In } \mathcal{R}, \quad \psi \;=\; \int_{n}^{n+2h} 2\pi\, M(t - x/c)\, dx\,. \tag{19}$$

Thus the range of integration is limited to small values of x, and the uncertainty in the integrand for larger x is moot—as hoped. Under the standing assumption that the strength M is continuous, there exists a value ν, within the range of integration, such that the integral is equal to the width of the range multiplied by the integrand evaluated at ν; that is:

$$\text{In } \mathcal{R}, \quad \psi \;=\; 2h \cdot 2\pi\, M(t - \nu/c)\,, \quad \text{where } \; n < \nu < n + 2h\,. \tag{20}$$

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

$$M(t) = \frac{D(t)}{h}\,, \tag{21}$$

so that (20) becomes

$$\text{In } \mathcal{R}, \quad \psi \;=\; 4\pi\, D(t - \nu/c)\,, \quad \text{where } \; n < \nu < n + 2h \tag{22}$$

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

$$\text{In } \mathcal{R}, \quad \psi \;=\; 4\pi\, D(t - n/c)\,. \tag{23}$$

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

$$\psi \;=\; \begin{cases} 4\pi\, D(t - n/c) & \text{in } \mathcal{R} \\ 0 & \text{in } \mathcal{R}'. \end{cases} \tag{24}$$

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

$$D(t) = \frac{\psi}{4\pi}\,, \tag{25}$$

where ψ is to be evaluated at the $\mathcal{R}$ side of $\mathcal{S}$—that is, as a BC. And in the limit as $n \to 0$ from the $\mathcal{R}'$ side, (24) gives a null BC on ψ (hence, by Proposition 3, a null wave function throughout $\mathcal{R}'$ and a null BC on ψ_n). By setting the boundary values of ψ, we have satisfied the saltus condition in Proposition 4; and by Proposition 3 or 4, we get the original wave function throughout $\mathcal{R}$, hence the original BC on ψ_n at the $\mathcal{R}$ side. For future reference, however, we differentiate (24) w.r.t. n on the $\mathcal{R}$ side, obtaining

$$\psi_n = -\frac{4\pi}{c}\,D'(t - n/c)\,, \tag{26}$$

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

$$D'(t) = -\frac{c\psi_n}{4\pi}\,. \tag{27}$$

The quantity $D(t)$, given in terms of the required BC by equation (25), is the *strength density of spatiotemporal dipole sources facing* $\mathcal{R}$. In general, a **spatiotemporal dipole** (**STD**) of strength $f(t)$, in the direction of the coordinate n, consists of a monopole source of strength $f(t)/h$, and another monopole source of strength $-f(t - h/c)/h$, displaced from the first by a distance h in the $-n$ direction, where $h \to 0$. If the strength *density* of such sources on a surface $\mathcal{S}$ is $D(t)$, then each element $d\mathcal{S}$ carries monopoles of strengths $D(t)\,d\mathcal{S}/h$ and $-D(t - h/c)\,d\mathcal{S}/h$, separated by the distance h. Hence the monopole strengths *per unit area of* $\mathcal{S}$ are $D(t)/h$ and $-D(t - h/c)/h$, as above.

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

A dipole source (spatiotemporal or not) comprises two opposing infinite monopole sources separated by an infinitesimal distance. Obviously this concept strains credulity. But we are not claiming that the sources specified by (25) actually exist, or even that they *could* exist. We have merely shown that the wave function in $\mathcal{R}$ is *as if* it had been generated by the specified distribution of sources, and that the same distribution would give a null wave function in $\mathcal{R}'$.

2.3 Note on electromagnetic waves

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

$$\frac{1}{r}\,f(t - r/c)\,, \tag{28}$$

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

$$d\varphi = \frac{1}{4\pi\epsilon_0 r}\,\varrho(t - r/c)\,dV \tag{29}$$

and

$$d\mathbf{A} = \frac{1}{4\pi\epsilon_0 c^2 r}\,\mathbf{j}(t - r/c)\,dV\,, \tag{30}$$

where ϵ_0 is a physical constant (called the electric permittivity of a vacuum), and r is the distance from dV, *regardless of direction*. Of course, elemental sources of these kinds cannot be arranged arbitrarily, because charge must be conserved. But, subject to that constraint, with appropriate definitions of f, their contributions to the wave functions do indeed have the form of (28).

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

[8] *Cf.* Feynman *et al.* [4], vol. 2, Chapter 15, Table 15-1.

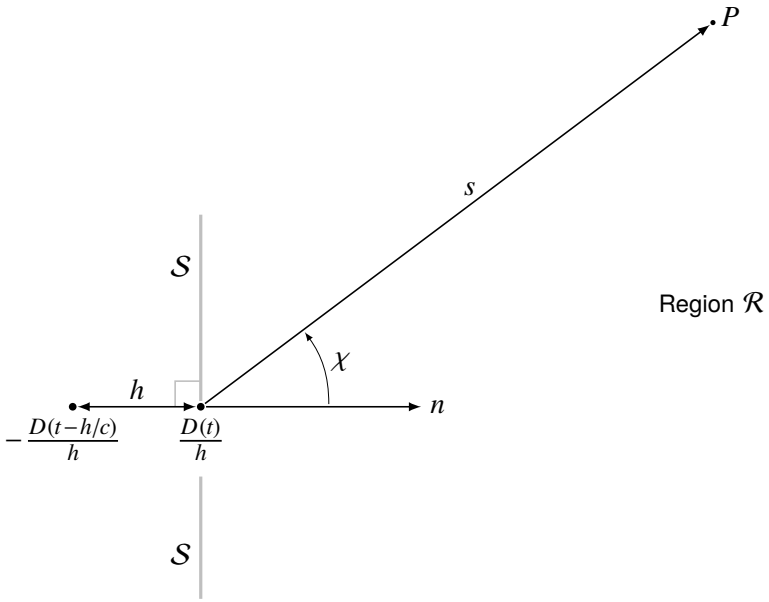

Fig. 2: Paired monopole sources per unit area of a spatiotemporal-dipole source distribution with strength density $D(t)$ facing $\mathcal{R}$. Gray lines show the orientation of the surface $\mathcal{S}$, which may be curved. The separation h is infinitesimal (so the diagram is not to scale). The normal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

3 Helmholtz formula; Kirchhoff integral theorem

3.1 Derivation

Having found the secondary sources that match the BCs in $\mathcal{R}$, we must now express the wave function in $\mathcal{R}$ *in terms of the sources*, and thence in terms of the BCs. The first step is to find the infinitesimal contribution to the wave function at a general field point P in $\mathcal{R}$, due to the secondary sources on a general surface element $d\mathcal{S}$. We shall call this contribution $d\psi(P,t)$, specifying the place P as an argument. *From here on, if the place of evaluation of ψ or any of its derivatives is not specified, we take it to be at $d\mathcal{S}$, on the side facing $\mathcal{R}$.* Equations (25) and (27) are already consistent with that convention.

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

$$\frac{D(t-s/c)}{hs}\,. \tag{31}$$

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

$$h\frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{hs}\right) + \frac{h}{c}\frac{\partial}{\partial t}\left(\frac{D(t-s/c)}{hs}\right) \tag{32}$$

$$= \frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{s}\right) + \frac{1}{cs}D'(t-s/c)\,. \tag{33}$$

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

$$d\psi(P,t) = \left\{\frac{\partial}{\partial n}\left(\frac{D(t-s/c)}{s}\right) + \frac{1}{cs}D'(t-s/c)\right\}d\mathcal{S} \tag{34}$$

[9] Continuing from footnote 6: If the factor 4π is included in the denominator of equation (6), it influences subsequent equations and eventually cancels out in (22) to (27). But then it is needed again in (31) to (34), with the result that (35), and therefore (36) and its corollaries, are unchanged! Equations (29) and (30) 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.

or, upon substitution from (25) and (27),

$$d\psi(P,t) \;=\; \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,. \tag{35}$$

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

$$\psi(P,t) \;=\; \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,, \tag{36}$$

where the double integral sign acknowledges that the range of integration is two-dimensional. Equation (36) is the **Helmholtz formula** in its most general form.[10] In the integrand, ψ and ψ_n are understood to be evaluated at the element $d\mathcal{S}$, and their arguments indicate that they are delayed by the propagation time for the distance s (from the element $d\mathcal{S}$ to the field point P).

If we want the integrand in (36) to be in terms of BCs on ψ, we must eliminate derivatives of expressions other than ψ. In the first term in the braces, the fraction to be differentiated depends on n, *not* through the normal derivative ψ_n (since ψ is evaluated at $n=0$), but through s (which decreases as n increases by h in Fig. 2) and therefore through the s-dependent delay in $\psi(t)$. So the first term is

$$\frac{\partial}{\partial n}\left(\frac{\psi(t-s/c)}{s}\right) \;=\; \psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) + \frac{1}{s}\,\frac{\partial}{\partial n}\psi(t-s/c) \tag{37}$$

$$\;=\; \psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'(t-s/c)\,\frac{\partial s}{\partial n}\,, \tag{38}$$

where the first equality follows from the product rule, and the second from two applications of the chain rule. Substituting (38) into the Helmholtz formula (36) yields

$$\psi(P,t) \;=\; \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{\psi(t-s/c)\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\,\psi'(t-s/c)\,\frac{\partial s}{\partial n} - \frac{1}{s}\,\psi_n(t-s/c)\right\} d\mathcal{S}\,. \tag{39}$$

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

$$\psi(P,t) \;=\; \iint_{\mathcal{S}} \frac{1}{4\pi}\left\{[\psi]\,\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\left[\frac{\partial\psi}{\partial t}\right]\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{40}$$

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

$$\psi(P,t) \;=\; \iint_{\mathcal{S}} \frac{1}{4\pi s}\left\{-\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot\psi]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial n}\right]\right\} d\mathcal{S}\,, \tag{41}$$

which we shall find convenient in Section 4.

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

[11] *Cf.* Born & Wolf [2] at pp. 420–21 (especially eq. 13). *Cf.* also Baker & Copson [1, p. 37] and Miller [10, 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.

3.2 Application to diffraction by an aperture

So we arrive at a catch-22. We established very early (Proposition 2) that the wave function throughout $\mathcal{R}$ is determined by *either* the wave function *or* its normal derivative at $\mathcal{S}$, so that each BC is determined by the other through the complete wave function. At yet, in order to calculate that wave function using (39) or (40) or (41), we need to know *both* the wave function ψ *and* its normal derivative ψ_n at $\mathcal{S}$.

This poses a difficulty when we want to calculate the wave function transmitted through an aperture in an opaque screen. In that case we partitioned $\mathcal{S}$ into two segments, namely $\mathcal{S}_a$ spanning the aperture, and $\mathcal{S}_b$ on the back of the screen; and we proposed to estimate the wave function beyond the aperture by calculating the secondary sources as if the screen were not there, then integrating over $\mathcal{S}_a$ only. We first justified that trick by supposing that the screen simply eliminates the secondary sources on $\mathcal{S}_b$. We then showed, in equation (4), that integrating over only the aperture is equivalent to assuming that the screen imposes an inverted saltus condition. But Kirchhoff himself, in 1882, arrived at his integral theorem by a different path, which obliged him to deal directly with boundary conditions. So, to obtain the integral over the aperture only, he supposed that the wave function *and* its normal derivative on $\mathcal{S}_b$ were zero, while the wave function *and* its normal derivative on $\mathcal{S}_a$ were as if the screen were not there.

All three approaches lead to the same approximation: that the integrand is as if the screen were not there, while the integral is over the aperture only. And the approximation turns out to be remarkably accurate—indeed, usually far *more* accurate than the boundary conditions on which it is apparently based [13], because the errors in the BCs tend to "average out" in the integration.

3.3 Consistency and well-posedness

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

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

There has been much ado about this inconsistency [3, 9, 13]. In Kirchhoff's defense, it is perhaps enough to say that (i) he could hardly have done anything else at the time, (ii) he offered a theory that stood up better than its predecessors under the test of experiment, and (iii) this is supposed to be the definition of progress in physics. It is definitely *not* enough to say that his BCs were only meant to be approximate and therefore only needed to be approximately consistent, because in fact they are not even approximately consistent. We have just seen that as we approach the edge of the aperture ($\mathcal{S}_a$) along the back of the screen ($\mathcal{S}_b$), the predicted normal derivative is wildly divergent from the assumption on which it supposedly depends. The wave function predicted by the Kirchhoff integral (or an equivalent) in the aperture can also differ substantially from the assumed boundary condition [9, Fig. 3(b)]. Hence it is sometimes said that the Kirchhoff BCs give more accurate results than we have any right to expect [13]. However, if a boundary-surface integral of a function of the BCs is to be accurate, it is not necessary that

[12] Poincaré's proof is for sinusoidal time-dependence. It has been explained in English by Buchwald & Yeang [3, pp. 485–90] and, more tersely, by Baker & Copson [1, pp. 71–2].

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

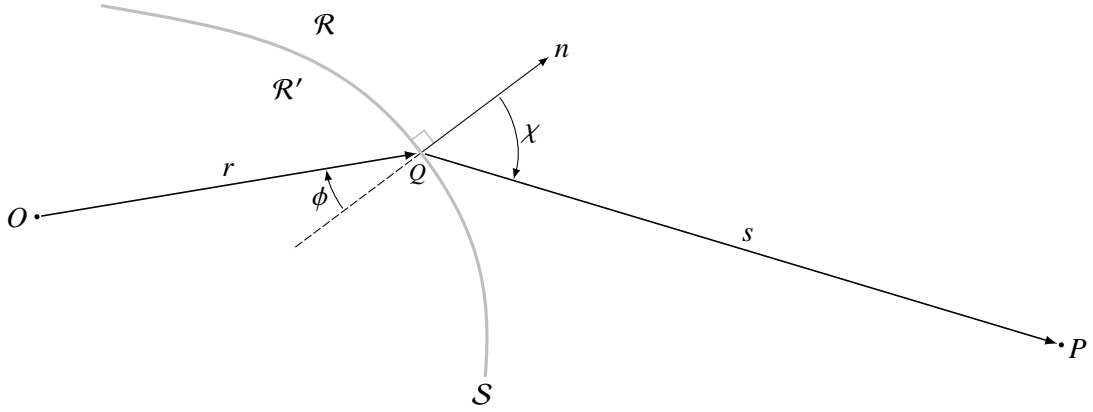

Fig. 3: Distances and angles pertaining to the calculation of the wave function at the field point P in region $\mathcal{R}$ due to a source at O in region $\mathcal{R}'$, by integration over the surface $\mathcal{S}$ separating $\mathcal{R}'$ and $\mathcal{R}$. The coordinate n is measured normal to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

the BCs themselves be accurate at every point; it suffices that the errors in the BCs "average out", nearly enough, in the composition of the integrand and the subsequent integration (*cf.* [13], at s. 4.2). It is no cause for astonishment that the latter condition is often satisfied without the former.

Moreover, we have already seen two other ways to obtain the Kirchhoff integral over $\mathcal{S}_a$, neither of which involves any mathematical inconsistency (whatever may be said about their physical accuracy). One way is simply to retain the secondary sources over $\mathcal{S}_a$ only. The other, as shown by equation (4), is to retain the original (primary) sources and add sources on $\mathcal{S}_b$ to impose an inverted saltus condition. We have noted that the latter formulation, if expressed purely as a saltus problem, is due to Kottler; however, his path to the solution was more difficult, involving the application of the Helmholtz formula to a more complicated surface [1, pp. 98–101].

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

4 Point-source: Kirchhoff diffraction formula

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

Now suppose that the wave function given by equations (39) to (41) has only *one* primary source, namely a single monopole at the point O outside $\mathcal{R}$. Let the coordinate r measure the distance from O to the general surface element $d\mathcal{S}$ (at point Q in Fig. 3). By the chain rule,

$$\frac{\partial\psi}{\partial n} = \frac{\partial\psi}{\partial r}\,\frac{\partial r}{\partial n}\;; \tag{42}$$

and the second factor on the right is time-independent, so that it can be taken outside the square brackets. Then the expression in braces in (41) becomes

$$\left\{-\frac{1}{s}\,[\psi]\,\frac{\partial s}{\partial n} - \frac{1}{c}\,[\dot{\psi}]\,\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial r}\right]\frac{\partial r}{\partial n}\right\}. \tag{43}$$

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

Let ϕ be the angle between the positive directions of n and r (Fig. 3). Then, treating n as the independent variable, we have

$$\frac{\partial r}{\partial n} = \cos\phi\,. \tag{44}$$

Similarly, from Fig. 3,

$$\frac{\partial s}{\partial n} = -\cos\chi\,, \tag{45}$$

where χ is the angle between the positive directions of n and s, and the minus sign arises because n is a coordinate of the point *from* which we measure s (whereas in the preceding equation, n is a coordinate of the point *to* which we measure r). As the primary source is a single monopole, the wave function has the form (6):

$$\psi = \frac{1}{r}\,f(t - r/c)\,. \tag{46}$$

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

$$\frac{\partial\psi}{\partial r} = -\frac{\dot\psi}{c} - \frac{\psi}{r}\,. \tag{47}$$

Substituting (44), (45), and (47) into (43), we get

$$\left\{\frac{1}{s}\,[\psi]\cos\chi + \frac{1}{c}\,[\dot\psi]\cos\chi + \left[\frac{\dot\psi}{c} + \frac{\psi}{r}\right]\cos\phi\right\}. \tag{48}$$

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, (48) be written

$$\left\{\frac{1}{s}\,[\psi]\cos\chi + \frac{1}{c}\,[\dot\psi]\cos\chi + \left(\frac{1}{c}\,[\dot\psi] + \frac{1}{r}\,[\psi]\right)\cos\phi\right\} \tag{49}$$

or, regrouping the terms,

$$\left\{\frac{\cos\phi + \cos\chi}{c}\,[\dot\psi] + \left(\frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right)[\psi]\right\}. \tag{50}$$

Putting this expression back in (41), we obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s}\left\{\frac{\cos\phi + \cos\chi}{c}\,[\dot\psi] + \left(\frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right)[\psi]\right\} d\mathcal{S}\,. \tag{51}$$

Equation (51) is the exact form of the **Kirchhoff diffraction formula** (also known as the *Huygens-Kirchhoff* diffraction formula, because it quantifies Huygens' principle for a monopole primary source; or the *Fresnel-Kirchhoff* diffraction formula, because it generalizes Fresnel's treatment of the subject). We call it the *exact* form in order to distinguish it from a more common approximate form: if the distances from the primary source to $\mathcal{S}$ and from $\mathcal{S}$ to P are sufficiently large (as is typically the case), then ψ/r and ψ/s are very small compared with $\dot\psi/c$, so that we can neglect the second term in the braces, retaining only the term in $[\dot\psi]$. That term contains the factor $\cos\phi + \cos\chi$, in which the angles ϕ and χ are relative to the normal to $\mathcal{S}$; we might therefore describe this factor (or something proportional to it) as the "obliquity factor".

5 Integration over a primary wavefront

5.1 Spherical wavefront

Recall that ϕ is the angle between the positive directions of n and r. This is the angle between the normal to $\mathcal{S}$ and the radius of the (spherical) primary wavefront, hence the angle between the normal

to $\mathcal{S}$ and the normal to the primary wavefront. So if $\mathcal{S}$ *is* a primary wavefront, we simply put $\phi = 0$ in (51), obtaining

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} [\dot{\psi}] + \left(\frac{1}{r} + \frac{\cos\chi}{s} \right) [\psi] \right\} d\mathcal{S} \,. \tag{52}$$

Equation (52) 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...* [5, p. 108].

In (52), for sufficiently large r and s, we may again neglect the second term in the braces and retain only the term in $[\dot{\psi}]$, in which the obliquity factor has been reduced to a function of the single obliquity angle χ. We now see that if we define the **obliquity factor** in (52) as

$$\tfrac{1}{2}(1+\cos\chi), \tag{53}$$

then it will have a maximum of 1 for $\chi = 0$ (direct forward secondary waves), and a minimum of 0 for $\chi = 180°$ (backward secondary waves). Hence, in (51) and (52), it might be preferable to define the obliquity factor as *half* the numerator of the first fraction in the braces.

5.2 Plane wavefront: Approximation for non-spherical wavefront

The curvature of the primary wavefront enters into the derivation of (51) 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 (51) and obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{\cos\phi+\cos\chi}{c} [\dot{\psi}] + \frac{\cos\chi}{s} [\psi] \right\} d\mathcal{S} \,. \tag{54}$$

And for integration over a primary wavefront, we put $r \to \infty$ in (52), *or* $\phi = 0$ in (54), and obtain

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{1}{4\pi s} \left\{ \frac{1+\cos\chi}{c} [\dot{\psi}] + \frac{\cos\chi}{s} [\psi] \right\} d\mathcal{S} \,. \tag{55}$$

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

As (54) and (55) are exact for plane waves and good approximations for large r, we can reasonably expect them to be good approximations for primary wavefronts with *non-spherical* curvature, provided that the curvature is sufficiently gradual. (In a homogeneous, isotropic medium, a non-spherical wavefront typically comes from an initially plane or spherical wavefront that has been reflected or refracted at the interface with a different medium.) In such cases, ϕ is to be understood as the angle between the normals of the primary wavefront and the surface of integration.

Similarly, although the Kirchhoff formula (51) and the Huygens-Fresnel-Kirchhoff formula (52) have been derived for monopole sources, which by definition are omnidirectional, we can expect both formulae to be good approximations for directional sources provided that the sources give recognizable spherical wavefronts with a sufficiently gradual variation of intensity over each wavefront. And, as above, we can expect the approximations to hold even if the wavefront also has non-spherical curvature, provided that the curvature is sufficiently gradual.

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

6 The sinusoidal (monochromatic) case

6.1 Exact forms

If the function f in equation (6) has the **sinusoidal** form[16]

$$f(t) = A\cos(\omega t + \theta) \tag{56}$$

where A (the *peak amplitude*), ω (the *angular frequency* or *radian frequency*), and θ (the *phase angle*) are constants, then the wave function due to the monopole source in (6) becomes[17]

$$\frac{A}{r}\cos\left(\omega(t - r/c) + \theta\right), \tag{57}$$

which is usually written

$$\frac{A}{r}\cos\left(\omega t - kr + \theta\right), \tag{58}$$

where

$$k = \omega/c\,. \tag{59}$$

We call k the *wavenumber* (or, more precisely, the *angular* or *radian* wavenumber). In (58), 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

$$T = 2\pi/\omega\,, \tag{60}$$

and the *wavelength* is

$$\lambda = 2\pi/k\,. \tag{61}$$

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:

$$\nu = \frac{\omega}{2\pi}\,. \tag{62}$$

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

$$\lambda\!\!\!^{-} = \frac{\lambda}{2\pi} = 1/k\,. \tag{63}$$

For a given location, hence a fixed r, the wave function (58) has the same form as (56), but with different values of the peak amplitude and phase angle. The contributions to the wave function at that point due to any other sinusoidal monopole sources with the same ω are also of that form, so that their total contribution to the wave function at that point is a sum of functions of form (56), with the same ω (but generally not the same constants A and θ). A pattern of sinusoidal waves of the same ω (hence the same frequency) is described as **monochromatic**, meaning "of one color", because that is the implication in the case of light waves.

Now it is readily seen that *the sum of any number of functions of form* (56)*, with the same frequency, is a function of the same form with the same frequency.*[18] Indeed, in the Cartesian xy plane, function (56) is the x component of a vector of length A making an angle $\omega t + \theta$ with the x axis, hence the x component of a vector of length A rotating at angular frequency ω from an initial angle θ. The sum

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

[17] If this form applies for all t, it is obviously incompatible with the assumption that the wave function had a beginning. Perhaps the simplest workaround is to suppose that there has been enough time, since all the sinusoidal sources started up, for the waves to fill the region of interest, and for any start-up effects to leave that region or fade away. The essential feature of any workaround is that "equation" (66) be true, or at least sufficiently accurate.

[18] This was first shown by Fresnel, by an analytical method, in a "supplement" dated January 1818 [6, vol. 1, at pp. 489–92]; see [12] for context. The time t was measured in periods. He repeated the demonstration in his prize memoir on diffraction [5, 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.

of any number of functions of that form, with that frequency, is the sum of the x-components of the associated vectors, which is the x component of the sum of the vectors; and that vector sum is itself a vector rotating at the same frequency, so its x component is a function of the same form and frequency.

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

$$\psi = A\cos(\omega t + \theta)\,, \tag{64}$$

where A and θ are independent of time but vary from point to point. From (64),

$$\begin{aligned}\frac{\partial\psi}{\partial t} &= -\omega A\sin(\omega t+\theta)\\ &= \omega A\cos\left(\omega t+\theta+\pi/2\right).\end{aligned} \tag{65}$$

So, for a function of form (64), the operator $\partial/\partial t$ (or an overdot) is equivalent to multiplication by ω combined with a phase advance of a quarter-cycle ($\pi/2$ radians; $90°$). This can be written

$$\frac{\partial}{\partial t} = j\omega\,, \tag{66}$$

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

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

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

In rule (66), the order of j and ω may be switched. In general, for a function of form (64), the j operator adds $\pi/2$ to the argument of the cosine function, or adds $\pi/2$ to the term ωt, whereas multiplication by a time-independent factor involves no such argument or term; so the order of the two operations makes no difference. In this respect, the operator j behaves like any other time-independent factor. Hence we can divide (66) through by c and apply (59), obtaining the operational equivalence

$$\frac{1}{c}\,\frac{\partial}{\partial t} = jk\,. \tag{67}$$

Now recall that another property of time-independent factors, such as k, is that they can be taken inside or outside the square brackets. Similarly, for a function of form (64), the j operator adds $\pi/2$ to the argument of the cosine function, whereas the square brackets subtract s/c from t within that argument, and the order of these operations makes no difference. Consequently, *the j operator, like a time-independent factor, may be taken inside or outside the square brackets*. This rule will now prove especially useful in combination with (67).

Applying rule (67) to equation (51)—or, more precisely, taking $1/c$ inside the square brackets, applying (67), then taking jk outside the square brackets, and finally taking $[\psi]$ outside the braces, on the understanding that j still operates on $[\psi]$—we obtain the exact monochromatic form of the Kirchhoff diffraction formula, which concerns spherical primary wavefronts and an arbitrary surface of integration:

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s}\left\{jk\left(\cos\phi+\cos\chi\right) + \frac{\cos\phi}{r} + \frac{\cos\chi}{s}\right\} d\mathcal{S}\,. \tag{68}$$

[19] An exception to both rules is Miller in reference [10], where he uses i to denote a quarter-cycle *advance*.

[20] In my last paper [11], for better or worse, I used the physics convention.

Similarly applying (67) in (52), *or* putting $\phi = 0$ in (68), we get the exact monochromatic form of the Huygens-Fresnel-Kirchhoff formula, in which the integration is over a spherical primary wavefront:[21]

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(1+\cos\chi\right) + \frac{1}{r} + \frac{\cos\chi}{s} \right\} d\mathcal{S}\,. \tag{69}$$

For *plane* primary waves, we apply (67) in (54) and (55), or let r become infinite in (68) and (69), obtaining

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(\cos\phi+\cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{70}$$

for integration over a general surface, and

$$\psi(P,t) = \iint_{\mathcal{S}} \frac{[\psi]}{4\pi s} \left\{ jk\left(1+\cos\chi\right) + \frac{\cos\chi}{s} \right\} d\mathcal{S} \tag{71}$$

for integration over a primary wavefront. Of course (71) is also obtainable from (70) by setting $\phi = 0$.

6.2 Approximate forms

We have noted that in equations (51) and (52), if the distances are sufficiently large, we can neglect the term in $[\psi]$ and retain only the term in $[\dot\psi]$. In the sinusoidal case, the term in $[\dot\psi]$ becomes the term in k, allowing us to clarify the meaning of "sufficiently large". In (68) and (69), if $1/r$ and $1/s$ are very small compared with k—that is, *if r and s are very large compared with the radian wavelength*—then we can neglect the terms in r and s. It is also convenient to recall (61) and put $k = 2\pi/\lambda$. Equations (68) and (69) then reduce to

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{j}{\lambda} \frac{\cos\phi+\cos\chi}{2s} [\psi]\, d\mathcal{S} \tag{72}$$

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

$$\psi(P,t) \approx \iint_{\mathcal{S}} \frac{j}{\lambda} \frac{1+\cos\chi}{2s} [\psi]\, d\mathcal{S} \tag{73}$$

if $\mathcal{S}$ is a primary wavefront; and again the second result follows from the first by setting $\phi = 0$. Both results indicate that the secondary sources are advanced in phase by 90° relative to the primary wave function at $d\mathcal{S}$, that their strengths are inversely proportional to the wavelength, and that the amplitudes of the secondary waves decay with distance like $1/s$. The remaining factors in the integrand are the primary wave function (delayed) and the obliquity factor. The $1/r$ dependence of the primary wave function can be made explicit by substituting expression (57) for ψ.

Because (72) and (73) assume that s is large, they apply only to the *far field* of the surface $\mathcal{S}$. In other respects they have the same applicability as (70) and (71)—that is, to the far field of the primary source, including primary wavefronts of arbitrary shape but sufficiently gradual curvature, and with sufficiently gradual variation of intensity over the wavefront. In the sinusoidal case, it is clear that a "sufficiently gradual" curvature is one for which the radii of curvature are very large compared with the radian wavelength λ, and we might guess that a "sufficiently gradual" variation of intensity is one that gives negligible variation over distances comparable with λ.

[21] The corresponding result in Baker & Copson [1, top of p. 33] has a sign error in the last term in the braces (when the sign outside the integral is accounted for). The error, which is found in all editions, can be confirmed by working from their previous equation [1, bottom of p. 32]. Larmor [8, p. 172, last eq.] and Miller [10, eq. 6] omit the second-last term in the braces (see Section 7 below), but agree with me on the sign of the last term.

7 Notes on Miller (1991)

7.1 Near-source term

The integrand given by Professor Miller [10, eq. 6] is apparently meant to be "exact for uniform spherical or plane wave fronts" [10, before eq. 6]. However, when translated into the notation of the present paper,[22] his integrand agrees with the plane-wave formula (71) above—not with the more general spherical-wave formula (69) above. The reason is that he neglects the $1/r$ decay in the magnitude of the primary wave, with the result that his equation (4), which corresponds to our (47), lacks the second term on the right.[23]

As Miller notes, the $(\cos\chi)/s$ term in (69) is "near-field" in the sense that it is significant close to the surface (wavefront) of integration. He retains this term but omits the $1/r$ term, which is near-field in the sense that it becomes significant if the surface of integration is close to the primary source—so close that $1/r$ is not negligible compared with k (*cf.* Baker & Copson [1, p. 33], who include the $1/r$ term, which they call $1/r_0$). I concede, however, that the $1/r$ term might reasonably be considered less important, because the condition under which we cannot neglect that term is also a condition under which we cannot neglect aperture-edge effects or (as in Miller's numerical example) any directional variation in the intensity of the primary wavefront.

7.2 Answering my own question

The conclusion of my last paper [11, p. 9] began:

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

Indeed it can be reconciled, because Miller's STDs yield formula (40), which in turn yields the $1/r$ term in (52) and (69). Moreover, equation (52) is for general time-dependence, confirming Miller's statement that the validity of spatiotemporal dipoles is not limited to the monochromatic case [10, endnote 12].

So this is the "other paper", although I was not confident that I would be its author. One reason for my doubt may be gleaned from my vague statement that the correction term relates to "the curvature of the surface of dipoles, which in turn implies a departure from a simple proportionality between the strength of the secondary source and the area of the surface element" [11, p. 8]. As it turns out, I have sidestepped that complication by treating the strength densities of monopole sources as per unit area "of $\mathcal{S}$", even if that does not exactly correspond to unit area of a parallel sheet at a distance h from $\mathcal{S}$.

8 Conclusion

Whereas the presentation of Kirchhoff's integral theorem and its corollaries has hitherto been thought to require sophomore-level mathematics, this paper develops the same theory from concepts that are either familiar to high-school graduates or easily introduced to them as needed. The approach is both heuristic and traditional, proceeding from the sequential nature of wave propagation, through superposition, Huygens' principle, saltus conditions, and their imposition by spatiotemporal-dipole sources, leading to the Kirchhoff integral theorem, followed by the case of the monopole primary source, obliquity factors, Fresnel's choice of the primary wavefront as the surface of integration, and special cases for large distances. The assumption of sinusoidal time-dependence is not essential, but is treated as another special case; this has been done without complex numbers by defining j as a phase-shift operator. In the problem of diffraction by an aperture in an opaque screen, in which the inconsistencies in Kirchhoff's boundary conditions have hitherto been either tolerated, or circumvented by introducing

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

[23] Larmor had made the same approximation: in [8], on p. 172, the second-last equation is equivalent to Miller's (4), with the result that the last equation agrees with my (55), not my (52).

further complexity, the simpler approach of the present paper avoids the cause of the inconsistency. For these reasons I entertain the hope that the essential mathematical theory of Huygens' principle and diffraction has been rendered more accessible.

References

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

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

[3] J. Z. Buchwald & 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.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.)