A tautological theory of diffraction

Gavin R. Putland [1]

December 5, 2019

Three characteristic features of diffraction integrals, namely the quarter-cycle phase advance of the secondary sources, the proportionality of their strengths to the wave number, and the values of the obliquity factor at the limits of its domain, are explained by a one-dimensional integral identity that assumes nothing about the spatial variation of the wave function. If we subsequently assume uniform spherical primary waves, we can then relate the integrand to the wave function on the surface of integration. If that surface is a primary wavefront, the integrand becomes consistent with spherical secondary waves, but still contains an unknown obliquity factor and its derivative. If we try the standard obliquity factor, the integrand agrees *exactly* (not only in the high-frequency limit) with the Kirchhoff theory, and yields a near-source correction to the much-cited integrand of D.A.B. Miller (1991).

1 Introduction

It is well known [1, p. 22] that in a diffraction integral, the phases of the secondary waves must be taken as advanced by a quarter-cycle relative to the primary wave, while their amplitudes must be taken as inversely proportional to the wavelength λ. Born & Wolf [2, p. 416] describe these adjustments as "devoid of any physical significance" after deriving them for the degenerate case of propagation in "free space"—that is, the unobstructed case, in which the surface of integration completely surrounds the source. Their derivation begins by assuming:

(i) that the primary and secondary waves decay with amplitudes inversely proportional to their distances from their respective sources [2, p. 413]; and

(ii) that the *inclination factor* $K(\chi)$, by which the contribution from each surface element is weighted, is greatest on the line of sight, and falls to zero where the *diffraction angle* χ reaches $\pi/2$ [2, p. 414]; and (tacitly) that $K(\chi)$ remains zero for $\pi/2 < \chi \leq \pi$.

Specifically, χ is the angle between the normal to the primary wavefront and the direction of the observation point. The authors choose a primary wavefront as the surface of integration and employ Fresnel zones in their argument.

It is evident from the authors' own reasoning [2, pp. 413–4], and I shall confirm in the present paper, that assumption (ii) is unnecessarily strict: it is enough that the inclination factor falls smoothly to zero at the point of maximum propagation delay, and the fall need not be monotonic if it is smooth. But I shall also show that, for the purpose of explaining the preprocessing of the secondary waves (the quarter-cycle advance and the $1/\lambda$ factor), assumption (i) is redundant, and we need not restrict the surface of integration. Moreover, we can obtain a more general form of the integral without assuming sinusoidal waves, and we need not mention Fresnel zones even *after* we introduce that assumption. The integrand has two interrelated terms, of which one exhibits the preprocessing and the inclination factor, and the other is usually negligible.

[1] Melbourne, Australia. No institutional affiliation at time of writing. Gmail address: grputland.

2 A one-dimensional integral identity

Let t be time, and let $\psi(t)$ be the wave function at point P due to a source at point O. Let us specify a "ray velocity" (generally a function of location and direction) by which we can assign "propagation times" to paths. (The "scare quotes" will be explained in due course.) Let Σ be a geometric surface (not a physical interface). Let Q be a general point on Σ, and let the "propagation time" of the path OQP be a minimum at $Q = Q_0$; for any other position of Q, a signal taking the path OQP suffers a "delay" relative to one taking the path OQ_0P. Let S be the area of that part of the surface Σ for which the "delay" is less than τ, so that τ can be written $\tau(S)$. Let the maximum delay be T, at $S = S_T$. Then

$$\tau(0) = 0\,; \tag{1}$$
$$\tau(S_T) = T\,. \tag{2}$$

The problem is to express $\psi(t)$ as an integral over Σ. As a first trial, we might suppose that the integrand has the form of $\psi(t)$, delayed by τ to allow for the extra propagation time, and weighted to allow for the extra distance and the inclination factor. Then we might hope that the problem has been reduced to finding the weight function. Because the integrand is a product, suggesting integration by parts, let us write the weight function in differentiated form as $g'(S)$. Then the trial integrand is $g'(S)\,\psi\big(t-\tau(S)\big)$, to be integrated over the full range of S. By parts, and using (1) and (2), we obtain

$$\int_0^{S_T} g'(S)\,\psi\big(t-\tau(S)\big)\,dS \;=\; g(S_T)\,\psi(t-T) - g(0)\,\psi(t) + \int_0^{S_T} g(S)\,\psi'\big(t-\tau(S)\big)\,\tau'(S)\,dS\,,$$

where a prime ($'$) indicates that a function is to be differentiated with respect to its argument. If this equation is to yield $\psi(t)$ as an integral, we must have

$$g(S_T) = 0\,. \tag{3}$$

Solving for $\psi(t)$ then requires dividing through by $g(0)$ or, equivalently, scaling $g(S)$ so that

$$g(0) = 1\,. \tag{4}$$

Choosing the latter option, which of course maintains condition (3), we get

$$\psi(t) \;=\; \int_0^{S_T} \Big\{ g(S)\,\tau'(S)\,\psi'\big(t-\tau(S)\big) \;-\; g'(S)\,\psi\big(t-\tau(S)\big) \Big\}\,dS\,. \tag{5}$$

So the trial yielded a useful error: equation (5) expresses $\psi(t)$ as a surface integral, albeit one more complicated than the one we first tried. The integrand has two terms. The second term is in the form we tried, but with a minus sign in front of it, compensating for the typically negative sign of $g'(S)$; cf. (4) and (3). And in the *first* term, the wave function is differentiated and its weight is not $g'(S)$ or even $g(S)$, but $g(S)\,\tau'(S)$.

In the first term, by the chain rule, $\psi'\big(t-\tau(S)\big)$ is the derivative of $\psi\big(t-\tau(S)\big)$ not only with respect to its argument, but also w.r.t. time t. So, notwithstanding the delay, we can always treat ψ' and $\partial\psi/\partial t$ as interchangeable.

In the important case of *sinusoidal time dependence* with the time-dependent factor $e^{-i\omega t}$, either differentiation is equivalent to multiplication by $-i\omega$, so that (5) becomes

$$\psi(t) \;=\; \int_0^{S_T} \Big\{ -i\omega\, g(S)\,\tau'(S) \;-\; g'(S) \Big\}\,\psi\big(t-\tau(S)\big)\,dS\,. \tag{6}$$

This is a surface integral in the form that we first tried, with the weight function in braces. For a smooth g, if the frequency is sufficiently high (a plausible assumption in optics), we can neglect the second term in the braces. The remaining term then contains the factor $-i\omega$, which accounts for the quarter-cycle phase advance and the inverse proportionality to wavelength. That term also contains the factor $g(S)$ and therefore, according to (3), falls to zero at maximum “delay”, as expected of the inclination factor. But we have not delimited the surface that defines that maximum, and nowhere have we assumed that the fall in $g(S)$, or in $g(S)\,\tau'(S)$, is monotonic in S or any other variable.

Even (5) suggests that if ψ varies much faster than g, then the first term in the braces, containing the factor $g(S)$, is dominant. Indeed, noting that $\tau'(S)\,dS = d\tau$ and $g'(S)\,dS = dg$ and applying conditions (1) to (4), we readily obtain

$$\psi(t) \;=\int_0^T g(S)\,\psi'(t-\tau)\,d\tau \;+\; \psi(t-\tau)\,. \tag{7}$$

This not only confirms the dominance of the first term at high frequencies, but also shows that as T and τ approach zero (or, alternatively, as $\psi'\to 0$), the first term approaches zero while the second approaches $\psi(t)$. Letting T and τ approach zero handles the case in which the surface Σ shrinks to a point (whereas $\psi'\to 0$ in the *low*-frequency limit).

For the rest of this paper, the main line of argument will deal with general time dependence, and sinusoidal time dependence will be treated more or less parenthetically as required.

While equation (5) expresses the wave function at P as an integral over the surface Σ, this does *not* mean that the integrand has anything to do with the wave function *on that surface*. Under conditions (1) to (4), equation (5) is simply an identity, the truth of which does not depend on the meanings of the functions. Under the further condition that ψ has the time-dependent factor $e^{-i\omega t}$, equation (6) is likewise an identity. Neither result depends on the physical reality of the “ray velocity” or the consequent “propagation times” and “delays” (hence the scare quotes). Neither result depends on the amplitude of the wave function at points on Σ, or at any other points except P. Thus, for example, the ray velocity could be fictitious, and the so-called wave function could satisfy the diffusion equation instead of the wave equation, and both results—including the factor $-i\omega$ in the sinusoidal case—would still be true!

3 Uniform spherical waves

To relate the integrand to the wave function on Σ, we must assume something about the spatial variation of the wave function. From here on, we shall assume a wave function of the form

$$\psi = \frac{A}{r}\, f(r-ct)\,, \tag{8}$$

where A is a constant, r is the distance from the source, $f(\cdots)$ is a twice-differentiable function, and c is the constant (homogeneous and isotropic) ray speed and wave-normal speed.[2] This form describes outward-propagating uniform spherical waves and satisfies the wave equation.

As r is the distance from the source at O to a general field point, let s be the distance from that point to the observation point P, and let the direct distance from O to P be $a{+}b$ (written as a sum because we shall later break it into two steps). Then, putting $r=a{+}b$ in (8), we obtain the wave function at P, which we have called $\psi(t)$:

$$\psi(t) = \frac{A}{a{+}b}\, f(a+b-ct)\,. \tag{9}$$

[2] In general, the ray velocity is the radial velocity of the elementary wavefront [4]. In an isotropic medium, the elementary wavefront is spherical, so that the radial velocity coincides with the normal velocity.

Now let us restrict the "general field point" to the general point Q on Σ, and let us further stipulate that the line OP passes through Σ, so that the length of the shortest path from O to P via Σ is simply $a{+}b$. Since τ is the delay due to traveling via Q instead of by that shortest path, we have

$$c\tau = r + s - (a{+}b)\,. \tag{10}$$

Hence, by (9) and (10),

$$\begin{aligned}\psi(t{-}\tau) &= \frac{A}{a{+}b}\, f(a + b + c\tau - ct)\\ &= \frac{A}{a{+}b}\, f(r + s - ct)\\ &= \frac{r}{a{+}b}\,[\psi]\,,\end{aligned} \tag{11}$$

where

$$[\psi] = \frac{A}{r}\, f\big(r - c(t{-}s/c)\big)\,. \tag{12}$$

Comparing (12) with (8), we see that $[\psi]$ *is the wave function at* Q*, delayed by the propagation time from* Q *to* P. We shall use this notation from here on.

Substituting (11), which comes from the form of the wave function, into our one-dimensional integral identity (5), we can now relate the wave function at P to the wave function on the surface Σ:

$$\psi(t) \;=\; \int_0^{S_T} \frac{r}{a{+}b}\,\Big\{ g(S)\,\tau'(S)\,[\psi'] \;-\; g'(S)\,[\psi]\Big\}\, dS\,. \tag{13}$$

This looks unenlightening because we have not yet assumed anything about Σ from which we could deduce anything about $\tau'(S)$. Given conditions (1)–(4) and (8), with a matching notion of propagation times, equation (13) is applicable to a general surface Σ pierced by the line of sight OP, even if the surface does not completely surround the source at O. Changing Σ will generally change the values of S_T and T and shift the zero of $g(S)$, but the foregoing definitions and conditions have already allowed for that.

In (13), by conditions (2) and (3), the weight of $[\psi']$ (that is, the weight that matters in the high-frequency limit) is zero where the delay τ, hence the total propagation time, is maximized.[3] *If* the surface Σ is smooth and completely encloses the source, the maximum in propagation time is a stationarity. By Fermat's principle, the point on Σ where this stationarity occurs is a point of reflection (or would be, if Σ were reflective). In an isotropic medium, a necessary condition for a point of reflection is that the angle of incidence equals the angle of departure, both angles being measured from the normal to Σ. In summary, where the weight is zero, the angle of incidence equals the angle of departure. *This is consistent with the Fresnel-Kirchhoff diffraction formula* [2, p. 422], in which the integrand contains a factor given by the cosine of the angle of incidence minus the cosine of the angle of departure.

In the special case in which the surface Σ is a spherical wavefront centered on O, the point of reflection is the point at which the primary wavefront faces directly away from P, so that any secondary waves would be retrograde. In this case, setting the weight of ψ' to zero at the point of reflection can be understood as suppressing retrograde secondary waves.

4 Restricting the surface of integration

We now confine Σ to a spherical wavefront, but otherwise revert to the conditions of (13): that Σ is pierced by the line of sight but may not completely surround the source. Let the wavefront

[3] That is, point Q is positioned so that the propagation time of the two-step path OQP is maximized, subject to the constraint that the propagation times of the steps OQ and QP are minimized.

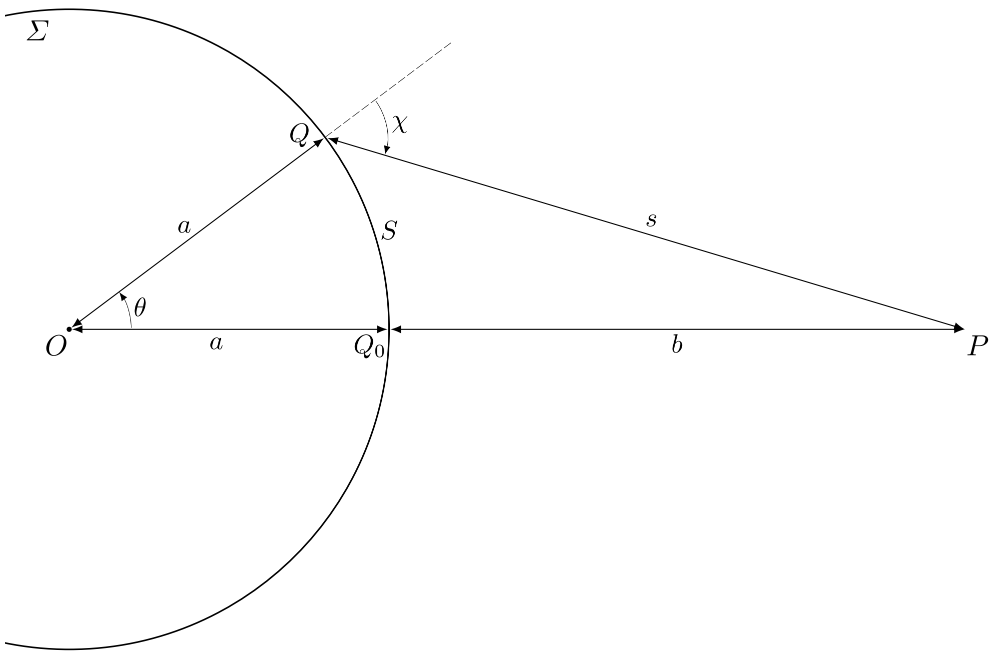

Fig. 1: Spherical wavefront Σ of radius a, emitted from point O in a homogeneous isotropic medium, to be observed at point P. Capital S is the area swept by the arc Q_0Q as it rotates about the axis OP. Lower-case s is the length QP. Both S and s (and the diffraction angle χ) can be considered functions of θ while a and b are fixed.

have radius a, so that

$$r = a \tag{14}$$

on Σ. Then b becomes the shortest distance from Σ to P, and (10) reduces to $c\tau = s - b$.

To obtain a specific example of $\tau'(S)$, let us consider the axisymmetric case, in which S is the area of the spherical cap whose edge is at distance s from the observation point P (see Fig. 1, which shows a cross-section through the axis, i.e. through the line of sight). In triangle QOP, by the cosine rule,

$$s^2 = a^2 + (a{+}b)^2 - 2a(a{+}b)\cos\theta\,. \tag{15}$$

Differentiation w.r.t. θ gives

$$2s\,ds = 2a(a{+}b)\sin\theta\;d\theta\,, \tag{16}$$

which we conveniently multiply by πa, obtaining

$$2\pi as\,ds = (a{+}b)\,ad\theta\;2\pi a\sin\theta \tag{17}$$

$$= (a{+}b)\,dS\,. \tag{18}$$

Putting $ds = c\,d\tau$ and solving for $d\tau/dS$ yields

$$\tau'(S) = \frac{a+b}{2\pi cas}\,. \tag{19}$$

If we define an *obliquity factor* as an inclination factor that has been normalized so as to take a value of 1 on the line of sight, then $g(S)$, according to (3) and (4), has the essential properties of an obliquity factor. It might therefore be useful to rewrite $g(S)$ and $g'(S)$ as related functions of the obliquity χ (hitherto called the diffraction angle). The present geometry indeed allows us to do so. If (say)

$$g(S) = \gamma(\chi)\,, \tag{20}$$

the constraints on γ corresponding to (3) and (4) are

$$\gamma(\chi_T) = 0\,, \tag{21}$$

$$\gamma(0) = 1\,, \tag{22}$$

where χ_T is the maximum value of χ on Σ (we do not yet assume that $\chi_T = \pi$). Differentiating (20) w.r.t. S gives

$$g'(S) = \gamma'(\chi)\,\frac{d\chi}{dS}\,, \tag{23}$$

which is not yet in terms of χ alone. In Fig. 1 we have previously applied the cosine rule to θ. If we now apply it to χ (whose cosine is minus that of its supplement) and differentiate w.r.t. s (the only other variable), we readily obtain

$$as\,\sin\chi\;d\chi \;=\; (s + a\cos\chi)\,ds\,. \tag{24}$$

Then if we multiply equations (18), (23), and (24), we can solve for $g'(S)$ in the desired form:

$$g'(S) = \frac{a+b}{2\pi as\,\sin\chi}\left(\frac{\cos\chi}{s} + \frac{1}{a}\right)\gamma'(\chi)\,. \tag{25}$$

Assembling the pieces, we substitute (14), (19), (20), and (25) into the "unenlightening" equation (13) and obtain

$$\psi(t) \;=\; \int_0^{S_T} \frac{1}{2\pi s}\left\{\frac{1}{c}\,\gamma(\chi)\,[\psi'] \;-\; \frac{1}{\sin\chi}\left(\frac{\cos\chi}{s} + \frac{1}{a}\right)\gamma'(\chi)\,[\psi]\right\} dS\,, \tag{26}$$

which is somewhat more enlightening in that the integrand contains the factor $1/s$, consistent with the production of spherical secondary waves that decay in the usual manner as they travel from the surface element towards the observation point. The corresponding $1/r$ decay of the primary wave is accounted for in the definition of $[\psi]$; see (12) above.

If $\psi(t)$ has the time-dependent factor $e^{-i\omega t}$ (that is, if A in equation (8) is allowed to be complex and $f(r{-}ct) = e^{ik(r-ct)} = e^{i(kr-\omega t)}$, where k is the wave number ω/c), then $\psi' = -i\omega\psi$, so that (26) becomes

$$\psi(t) \;=\; \int_0^{S_T} \frac{[\psi]}{2\pi s}\left\{-ik\,\gamma(\chi) \;-\; \frac{1}{\sin\chi}\left(\frac{\cos\chi}{s} + \frac{1}{a}\right)\gamma'(\chi)\right\} dS\,. \tag{27}$$

Again we have the decay factor $1/s$ outside the braces. The obliquity factor appears in the first term in the braces, which is the high-frequency term. If we retain only that term, the integrand becomes proportional to $-ik$, consistent with the existence of secondary sources whose phases are advanced by a quarter-cycle, and whose strengths are proportional to the wave number and therefore inversely proportional to the wavelength.

Even so, equations (26) and (27) are still tautological in the sense that they are necessary conditions of propagation of uniform spherical waves in "free space".

5 Remaining indeterminacy of the integrand

For uniform spherical waves, equation (13) holds for a general $g(S)$ that satisfies (3) and (4). Under the further condition that Σ is a cap on a spherical wavefront, axisymmetric about OP, equation (26) holds for a general $\gamma(\chi)$ that satisfies (21) and (22). Consequently, equation (26) cannot tell us which particular $\gamma(\chi)$, if any, is the "correct" obliquity factor (whatever that means). It cannot even determine χ_T, the upper limit of the domain of $\gamma(\chi)$. But, for a given $\gamma(\chi)$, it determines *both* terms of the integrand.

For a pertinent example, let Σ be the entire spherical wavefront in Fig. 1 (that is, $\chi_T = \pi$), and let us try the Stokes obliquity factor[4]

$$\gamma(\chi) = \tfrac{1}{2}(1 + \cos\chi)\,, \tag{28}$$

[4] On the history see Darrigol [3] at pp. 273–4.

whence

$$\gamma'(\chi) = -\tfrac{1}{2}\sin\chi\,. \tag{29}$$

Substituting these functions into (26) yields

$$\psi(t) = \iint\limits_{\Sigma} \frac{1}{4\pi s}\left\{\frac{1}{c}\left(1+\cos\chi\right)[\psi'] + \left(\frac{\cos\chi}{s} + \frac{1}{a}\right)[\psi]\right\} dS\,, \tag{30}$$

which we write as a surface integral over Σ because all the variable quantities in the integrand, namely s, χ, $[\psi']$, $[\psi]$, and (implicitly) τ, now depend on the position on Σ. For the sinusoidal case, we put $\psi' = -i\omega\psi$ and $\omega/c = k$ in (30), *or* substitute (28) and (29) into (27), obtaining

$$\psi(t) = \iint\limits_{\Sigma} \frac{[\psi]}{4\pi s}\left\{-ik\left(1+\cos\chi\right) + \frac{\cos\chi}{s} + \frac{1}{a}\right\} dS\,. \tag{31}$$

6 Agreement with Kirchhoff theory

In order to impute any physical significance to any particular choice of $\gamma(\chi)$, we must depart from the assumption that the same free-space wave function (8) applies at both the surface of integration and the observation point. A more permissive assumption is that the wave function satisfies the wave equation in a region $\mathcal{R}$ enclosed by a surface Σ, with all sources outside $\mathcal{R}$. Under the latter assumption, according to the Kirchhoff integral theorem [2, pp. 418–21], the wave function at a general point P inside $\mathcal{R}$ is given by

$$\psi = \iint\limits_{\Sigma} \frac{1}{4\pi}\left\{[\psi]\frac{\partial}{\partial n}\left(\frac{1}{s}\right) - \frac{1}{cs}\left[\frac{\partial\psi}{\partial t}\right]\frac{\partial s}{\partial n} - \frac{1}{s}\left[\frac{\partial\psi}{\partial n}\right]\right\} dS\,, \tag{32}$$

where n is a coordinate measured normal to Σ and into $\mathcal{R}$ (that is, away from the region containing the sources), and square brackets indicate that the enclosed quantity has been retarded (delayed) by the propagation time from dS to P.

To apply this theorem in the present context, in which the only source is *inside* Σ, we must define $\mathcal{R}$ as the region outside, so that n is measured "into" $\mathcal{R}$ but out of Σ (and still away from the region containing the source).[5] Then, in order to enclose $\mathcal{R}$, the surface Σ must be augmented by another closed surface outside, over which the integral must be zero in order not to disturb the equality in (32). That condition can be arranged, with tolerable generality, by letting the outer surface be so far way that the disturbance has not reached it yet!

The initial formulation of (32), with $\mathcal{R}$ and P inside Σ, sources outside, and n measured inward, is also applicable to a non-degenerate point-source diffraction problem, if (e.g.) *part* of Σ is the portion of a spherical wavefront admitted by an aperture in an opaque screen, so that the wave function on that part of Σ can be estimated from (8), while the remainder of Σ is chosen so that the wave function thereon can be neglected because (e.g.) these parts are on the back of the screen or are sufficiently far away.[6] In this situation we still assume the wave function (8), as promised, but only in the immediate vicinity of the wavefront; we no longer assume that it applies at the observation point P (as it would if the screen were not present).

[5] These are the conditions assumed by Miller [5, eq. 2], except that (among other notational differences) he uses the label V for the volume *inside* the closed surface. Miller had to change the signs given by Stratton [6, p. 427, eq. 22], because Stratton's corresponding integral ("*general solution* of the *homogeneous* equation") likewise measures the normal coordinate outward from the closed surface [6, p. 424] but takes the sources and observation point as being the other way around (so that his normal is *towards* the region containing the sources). Darrigol [3, p. 277, eq. 55] gives the same signs as Stratton, but expresses the result in terms of gradients and a vector surface element.

[6] Born & Wolf [2, pp. 421–3]. In this case the signs given by Stratton [6, p. 427, eq. 22], Baker & Copson [1, p. 40, last eq.], and apparently Darrigol [3, p. 277, eq. 55] measure the normal coordinate in the opposite direction.

In either case, we can write the left-hand side of (32) as $\psi(t)$, since this originally referred to the wave function at P. We shall also find it convenient to apply the chain rule to the first term in (32) and then take out the common factor $1/s$. Thus (32) becomes

$$\psi(t) = \iint\limits_{\Sigma} \frac{1}{4\pi s}\left\{-\frac{1}{s}[\psi]\frac{\partial s}{\partial n} - \frac{1}{c}\left[\frac{\partial\psi}{\partial t}\right]\frac{\partial s}{\partial n} - \left[\frac{\partial\psi}{\partial n}\right]\right\} dS\,. \tag{33}$$

To the extent that Σ coincides with a sphere centered on O (as in Fig. 1), we can take the normal coordinate n as synonymous with the radial coordinate r. Hence, from (8),

$$\frac{\partial\psi}{\partial n} = -\frac{1}{c}\frac{\partial\psi}{\partial t} - \frac{\psi}{n}\,; \tag{34}$$

and from the geometry,

$$\frac{\partial s}{\partial n} = -\cos\chi\,. \tag{35}$$

Substituting (34) and (35) into (33), with $n = a$ and $\partial\psi/\partial t = \psi'$, we recover equation (30)—whose derivation was based on nothing more than a one-dimensional identity, a free-space wave function, and a hypothetical obliquity factor. As before, (31) follows from (30) in the case of a time-dependent factor $e^{-i\omega t}$.

In short, if we somehow choose the correct obliquity factor $\gamma(\chi)$, the "tautological" result (26) agrees exactly—not only in the high-frequency limit, but exactly—with the Kirchhoff theory.

7 Small correction to Miller (1991)

The integrand in (31) is

$$\frac{[\psi]}{4\pi s}\left\{-ik\left(1+\cos\chi\right) + \frac{\cos\chi}{s} + \frac{1}{a}\right\}. \tag{36}$$

If the integrand given by Professor Miller [5, eq. 6] is translated into the notation and convention of the present paper,[7] it agrees with the above except that the term $1/a$ is missing. The reason is straightforward: he neglects the $1/r$ decay in the magnitude of the primary wave, with the result that his equation (4), which corresponds to my (34), lacks the second term on the right. The corresponding result in Baker & Copson [1, top of p. 33] has a term matching my $1/a$; but in all editions the sign of the term corresponding to my $(\cos\chi)/s$ disagrees with Miller and me (when the sign outside the integral is accounted for).

As Miller notes, the $(\cos\chi)/s$ term is "near-field" in that it becomes significant when the observation point is so close to the surface of integration that $1/s$ is not negligible compared with k. The $1/a$ term is near-field in a different sense: it becomes significant if the surface of integration is so close to the primary source that $1/a$ is not negligible compared with k [1, p. 33]. Because this is a condition under which we cannot neglect aperture-edge effects or (as in Miller's numerical example) non-uniform intensity of the primary wavefront, one might reasonably argue that the second near-field term is less important than the first.

Miller's interpretation of the integrand in terms of spatiotemporal dipoles is not necessarily invalidated by the $1/a$ term, because that 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.

I should also declare that the progress of this investigation in my own mind was not nearly as orderly as the way it is presented here. Nevertheless, I can report that the term missing from Miller's equation (6) first came to my notice, not through the Kirchhoff theory or any reference, but through my equation (30), as derived from the "tautological" theory and the Stokes obliquity

[7] Miller's ϕ is my ψ; his r is my s; i from his time-dependent factor $e^{+i\omega t}$ becomes $-i$; and his θ is my χ.

factor. I was merely trying to reproduce Miller's result; and had I "succeeded", I would not have bothered to seek further confirmation. However minor the missing term may be, it is, no doubt, piquant that an advanced theory can be corrected by an elementary one.

8 Conclusion

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; the intended subject of my present note, which is already longer than I first envisaged, was not the hidden physical meaning of diffraction integrals, but their hidden mathematical inevitability.

We have seen that a one-dimensional integral identity involving three functions yields a result in the form of a diffraction integral, provided that one function has the minimum properties of a propagation delay (relative to the path of least time) and another has the minimum properties expected of an obliquity factor, while the third is the wave function at the observation point. The integrand has two terms but is well approximated at high frequencies by one term, in which the wave function appears in a differentiated form. In the case of sinusoidal time-dependence, that differentiation gives a phase advance of a quarter-cycle and an amplitude inversely proportional to the wavelength. In the case of uniform spherical primary waves, with a primary wavefront as the surface of integration, the integrand shows the expected decay of the secondary waves with distance, and the delay and the obliquity factor become functions of the diffraction angle. The obliquity factor is still indeterminate; but if the correct factor is inserted, both terms in the integrand are then fully determined—and agree exactly with the Kirchhoff theory.

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] O. Darrigol, *A History of Optics: From Greek Antiquity to the Nineteenth Century*, Oxford, 2012.

[4] A. J. de Witte, "Equivalence of Huygens' principle and Fermat's principle in ray geometry", *American Journal of Physics*, vol. 27, no. 5 (May 1959), pp. 293–301. *Erratum*: In Fig. 7(b), each instance of "ray" should be "normal" (noted in vol. 27, no. 6, p. 387).

[5] 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.

[6] J.A. Stratton, *Electromagnetic Theory*, New York: McGraw-Hill, 1941.