Diffraction of Electromagnetic Waves

Ulrich Brosa

Brosa GmbH, Am Brücker Tor 4, 35287 Amöneburg, Germany
and Philipps-Universität, Renthof 6, 35032 Marburg, Germany

Reprint requests to U. B.; E-mail: brosa-gmbh@t-online.de

Z. Naturforsch. **65a,** 1 – 24 (2010); received September 16, 2009 / revised November 17, 2009

The general method to obtain solutions of the Maxwellian equations from scalar representatives is developed and applied to the diffraction of electromagnetic waves. Kirchhoff's integral is modified to provide explicit expressions for these representatives. The respective integrals are then evaluated using the method of stationary phase in two dimensions. Hitherto unknown formulae for the polarization appear as well as for imaging by diffraction. Ready-to-use formulae describing Fresnel diffraction behind a round stop are presented.

Key words: Electromagnetism; Optics; Diffraction; Polarization.
PACS numbers: 41.; 42.; 42.25.Fx; 42.25.Ja

1. Sad State of Theory

The systematic solution of partial differential equations for vector fields is demanding. The Maxwell equations, foundation of optics and electrodynamics, are equations of that kind. As a consequence, the theory of diffraction is essentially still scalar.

It was Gustav Kirchhoff who set the standards stolid till now [1 – 3]. Even in most recent textbooks Kirchhoff's achievements are recited as *the* theory of diffraction [4 – 7], see [8] for a historic review. Kirchhoff was aware of Maxwell's work, but did not esteem it too much. He preferred the scalar Helmholtz equation

$$\nabla_{\mathrm{p}}^2\, \psi(\mathbf{r}_{\mathrm{p}}) + k^2\, \psi(\mathbf{r}_{\mathrm{p}}) = 0, \tag{1}$$

wherein $\mathbf{r}_{\mathrm{p}}$ is to denote the place of the probe, ∇_{p} the nabla operator acting on $\mathbf{r}_{\mathrm{p}}$, and k the wave number. Kirchhoff introduced a scalar function ψ about which he did not know what it was supposed to mean.

Today it is generally accepted that light is an electromagnetic wave described by Maxwell's equations. Nevertheless scientists stick to Kirchhoff's scalar function. They observe that in Cartesian coordinates one may extract from Maxwell's equations one Helmholtz equation for each and every Cartesian component of the electric field **E** and the magnetic field **B** [4, Chap. 3]. The value of this observation is zero. The components of the electromagnetic vectors are coupled through the Maxwellian equations for curls and divergences. Solving the six Helmholtz equations for the components of **E** and **B** independently generates grossly wrong results. Even worse: In problems where Cartesian coordinates do not suit, as diffraction by spheres, there is, for the components of **E** and **B**, no Helmholtz equation at all [9].

The next field of questions arose when Kirchhoff, following Hermann Helmholtz, considered an integral which, seemingly, solves the Helmholtz equation [3, p. 82]

$$\psi(\mathbf{r}_{\mathrm{p}}) = \frac{1}{4\pi} \iint_F \left(\frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|)}{|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|} \partial_n \psi(\mathbf{r}) - \psi(\mathbf{r})\, \partial_n \frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|)}{|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|} \right) \mathrm{d}f, \tag{2}$$

wherein the surface F is to divide the entire three-dimensional space in an inner and an outer part; the probe $\mathbf{r}_{\mathrm{p}}$ resides in the inner part. $\mathbf{r}$ points to a point on F; it incorporates as components the variables of integration. $\mathrm{d}f$ is the respective element of the surface. ∂_n symbolizes a differentiation in the direction of the outer normal on F. The integral is interpreted in the way that primary waves $\psi(\mathbf{r})$ travel through the outer space until they strike F; there they excite secondary spherical waves $\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|)/|\mathbf{r}_{\mathrm{p}} - \mathbf{r}|$ interfering to produce the wanted $\psi(\mathbf{r}_{\mathrm{p}})$.

Arnold Sommerfeld has criticized that simultaneous fixing of boundary values for the function and its derivative as required in Kirchhoff's integral (2) causes contradictions [10 – 12]. But scientists answered and still answer, they would take for $\psi(\mathbf{r})$ a plane wave and differentiate it consistently, without contradiction. These scientists do not understand that diffraction is only possible if part of the primary wave is screened. How should one treat these parts of F? Kirchhoff and his followers demand simultaneously $\psi(\mathbf{r}) = 0$ and $\partial_n \psi(\mathbf{r}) = 0$ on the screened part of F and call it a 'black screen' [3, p. 40]. Yet the Helmholtz equation is elliptic of second degree. It has no real characteristics. Hence, from the Cauchy Kovalevskaya Theorem it follows that the only solution compatible with a black screen is a global zero. Kirchhoff's integral produces no better than order-of-magnitude results.

The third field of questions grew up when Kirchhoff wanted to evaluate the integral (2), but could not do it exactly. Especially, the function in the exponents seemed to be invincible. Kirchhoff substituted for the invincible function its crippled Taylor expansion. The evaluation with linear terms in $\mathbf{r}$ was called Fraunhofer diffraction, whereas truncation after the quadratic terms in $\mathbf{r}$ was said to yield Fresnel diffraction [3, p. 86]. One should assume that this kind of Fresnel diffraction includes Fraunhofer diffraction as the special case in which quadratic terms are negligible. But this does not happen. The formulae derived in this way describe fundamentally different pattern.

All these questions will be answered in the present article.

The general method to decouple Maxwell's equations will be developed in Section 2. The electromagnetic field is represented by three scalar functions a, b, and c which obey separated differential equations. One may solve the separated equations for these mathematical auxiliaries one by one and afterwards calculate the physical fields $\mathbf{E}$ and $\mathbf{B}$ from the auxiliaries by straightforward differentiation. c is the familiar scalar potential, a and b are common factors in the components of the simple and the double vector potential, respectively. This will be demonstrated for all electrodynamics in homogeneous and isotropic materials, in particular for those with finite conductivity. The General Representation Theorem in Section 2 constitutes the first main result in this article.

In electrodynamics, diffraction is just a simple special case for which the representatives a and b suffice.

In all problems with partial differential equations, also in the theory of diffraction, suitable boundary values must be fixed. Instead of the inconsistent black screen we will put up a screen of perfect conductivity. This entails boundary conditions for the electrical field $\mathbf{E}$. When these are converted for a and b, it turns out that the one representative obeys a homogeneous boundary condition of the first kind (aka Dirichlet), whereas the other is subject to a homogenous boundary condition of the second kind (aka Neumann), see Section 3.

To complete the mathematical definition of diffraction, initial values must be given. In Section 4 we will account for them using a Laplace transform. Then the transformed representatives a and b fulfill separated Helmholtz equations. This explains why Kirchhoff's formulas are not entirely wrong.

For the representation theorem the way the representatives are found does not matter. Expansions in terms of partial waves are, for example, feasible, but the aim of this article is to mend Kirchhoff's advances. Since the representatives a and b fulfill Helmholtz equations, one may derive for them integrals similar to (2). As Sommerfeld's criticism must be attended, the spherical wave will be replaced with genuine Green functions for boundary problems of the first and the second kind in Section 5.

Hence, it might seem that, for the true theory of electromagnetic diffraction, expense is double as two integrals must be computed. Yet if for the primary wave a spherical wave is appointed that proceeds from a source at $\mathbf{r}_\text{q}$, both a and b are determined by the same function of $\mathbf{r}_\text{p}$ and $\mathbf{r}_\text{q}$, the only difference being that probe and source are interchanged. The equations (57) through (60) display the second main result in this article.

The invincible function in the integral, mentioned above, can be rewritten to have an intuitive meaning: It is the difference of lengths when the points $\mathbf{r}_\text{q}$ and $\mathbf{r}_\text{p}$ are connected either directly or via a point $\mathbf{r}$ on the screen. These very terms occur in the familiar triangle inequation. It follows from the properties of the triangle function that simple geometry determines the basics of diffraction. For example, the border of shadow can be derived from the zero of the triangle function. The Criterion of Light will be established in Section 6.

Augustin Fresnel introduced certain integrals for the description of diffraction. For a realistic theory of diffraction these integrals are too clumpsy. It is more convienient to use instead the error function known to all statisticians. Although one needs, for diffraction, the error function of a complex argument, it behaves in a similar way as that of a real argument. Aside for greater simplicity, we get from the error function more results, namely diffraction in dissipative materials, which is of paramount importance for practical purposes. The salient properties of the complex error function will be described in Section 7.

There are several methods to evaluate the integral (58), e. g. for great distances of the probe from the screen. In this article we will apply the method of stationary phase to find the integral in the limit of short waves. It is this what Kirchhoff wanted when he expanded his functions up to second powers. We will avoid Taylor's expansion and introduce instead new variables that map the triangle function exactly. The Principle of Utter Exhaust will be introduced in Section 8. It constitutes the third main result in this article.

Utter exhaust (90) will be applied to the corrected Kirchhoff integral (58) in Section 9. It yields the fourth main result in this article, the Universal Formula of Diffraction (105): The diffracted wave equals the primary wave times the complementary error function which accounts through its argument for the specific shape of the edge; the argument is found by a purely algebraic calculation. The formula is universal in so far as it holds for all single diffracting edges. The diffraction by screens with several edges can be derived from it by mere superposition. The universal formula also holds for probes arbitrarily close to the screen. Therefore it can describe the transition from Fresnel to Fraunhofer diffraction. It also holds for dissipative materials; the argument of the error function also accounts for damping. Yet the formula (105) does not describe diffraction of very long waves. Moreover, under certain conditions, we will notice its weakness for great distances of the probe from the screen.

In this article, no graph of calculated fields will be shown. Detailed comparison to or prediction of measurements is here out of scope. Nevertheless, some impressive applications will be outlined. The first is the diffraction by a straight edge in Section 10. Diffraction means breaking beams asunder. Nevertheless, a metallic half screen creates via diffraction an image of the source that is focusing of beams. Astonishing as this result might appear, the same formula contains as a limiting case Sommerfeld's stringent solution describing the diffraction of a plain wave by a half plane, see Section 11. Thus the asymptotic methods developed here have more power than Sommerfeld's stringent integration of Maxwell's equations.

From the diffraction by a single straight edge it is just a small step to the diffraction by a slit, namely a superposition. Nevertheless, the ensuing formula describes both Fresnel and Fraunhofer diffraction and the transition between these regimes, see Section 12.

In Section 13 the universal formula is applied to diffraction by a circular aperture. It results in simple formulae describing Fresnel diffraction behind a round stop, a device that is used in almost all optical instruments.

In the concluding Section 14 the possibly novel results of this work will be listed and a program what is to be done next will be given.

2. The Representatives of Electrodynamics

The purpose of the representation theorem is to replace Maxwell's vector equations for the magnetic and electric fields **B** and **E** with separated differential equations for scalar representatives a, b, and c. As soon as the latter equations are solved one may determine successively the vector fields from the scalar representatives by straightforward differentiation. To understand and to prove the general representation theorem of electrodynamics, three requisites are needed. First, the

Lemma of Triple Curl. *The vector differential equation*

$$\nabla \times \nabla \times \nabla \times \mathbf{v} a(\mathbf{r},t) = -\nabla \times D_t \mathbf{v} a(\mathbf{r},t) \tag{3}$$

can be reduced to the scalar differential equation

$$\nabla^2 a(\mathbf{r},t) = D_t a(\mathbf{r},t) + \begin{cases} f(\mathbf{v}_0\mathbf{r},t) \text{ if } v_1 = 0 \\ f(|\mathbf{v}|,t) \text{ otherwise} \end{cases} \tag{4}$$

if and only if the supporting vector field $\mathbf{v}$ *is preformed as*

$$\mathbf{v} = \mathbf{v}_0 + v_1 \mathbf{r} \tag{5}$$

with arbitrary constants $\mathbf{v}_0$ *and* v_1. D_t *symbolizes an operator possibly including differentiations with respect to time t, but definitely no differentiation with respect to space* $\mathbf{r}$. $f(\cdot,t)$ denotes a free function. [1]

D_t is, for example, $\varepsilon\mu\partial_t^2 + \mu\sigma\partial_t$, see equation (13) below, with constants ε, μ, and σ denoting dielectric constant, magnetic permeability, and conductivity, respectively. $f(\cdot,t)$ is a so-called gauge, a function which can be chosen according to convenience. For present purposes $f(\cdot,t) = 0$ suffices.

The lemma was proven in [13] published in [14], and is now available in a textbook [15]. It is useful for

[1] In this section and the ensuing two, the index p at the position of the probe $\mathbf{r}_\mathrm{p}$ will be omitted.

uncoupling all vector equations which describe physical phenomena in a homogeneous and isotropic space, e. g. in the theory of elasticity and liquidity.

Secondly, it is assumed that the electromagetic field propagates in an isotropic and at least piecewise homogeneous medium. The constitutive relations $\mathbf{D} = \varepsilon\mathbf{E}$ and $\mathbf{H} = \mathbf{B}/\mu$, which relate the force-exerting fields $\mathbf{E}$ and $\mathbf{B}$ with the source-caused fields $\mathbf{D}$ and $\mathbf{H}$, as well as Ohm's law

$$\mathbf{j} = \sigma\mathbf{E}, \tag{6}$$

which relates the electric field $\mathbf{E}$ with the electric current density $\mathbf{j}$, will be mounted in Maxwell's equations from the very start.

Thirdly, as we want to solve initial- and boundary-value problems in time t and three-dimensional space $\mathbf{r}$ for the electric field $\mathbf{E}(\mathbf{r},t)$ and the magnetic field $\mathbf{B}(\mathbf{r},t)$, it is necessary to discriminate between charges and currents that are enforced from outside and those that come about through the free play between inside fields. The charge density $\rho_0(\mathbf{r})$ at the initial time $t = t_0$ is determined by reasons outside the considered system. In conducting materials it decays exponentially. The remainder $\rho(\mathbf{r},t)$ is zero at $t = t_0$ and thus determined by the density of the current because of continuity

$$\begin{aligned}&\partial_t\left(\rho(\mathbf{r},t) + \rho_0(\mathbf{r})\exp\left(\frac{\sigma}{\varepsilon}(t_0 - t)\right)\right)\\ &= -\nabla(\mathbf{j}_\mathrm{e}(\mathbf{r},t) + \mathbf{j}(\mathbf{r},t) + \mathbf{j}_0(\mathbf{r},t)).\end{aligned} \tag{7}$$

The current density, in its turn, is partly generated by external sources $\mathbf{j}_\mathrm{e}(\mathbf{r},t)$. Moreover, in conducting materials the electric field drives inner currents $\mathbf{j}(\mathbf{r},t) + \mathbf{j}_0(\mathbf{r},t)$ according to Ohm's law (6). The latter is the current caused by the decay of $\rho_0(\mathbf{r})$.

General Representation Theorem of Electrodynamics. *Solutions of the Maxwellian equations*

$$\nabla\times\mathbf{B}(\mathbf{r},t) = \varepsilon\mu\partial_t\mathbf{E}(\mathbf{r},t) + \mu\sigma\mathbf{E}(\mathbf{r},t) + \mu\mathbf{j}_\mathrm{e}(\mathbf{r},t), \tag{8}$$

$$\nabla\times\mathbf{E}(\mathbf{r},t) = -\partial_t\mathbf{B}(\mathbf{r},t), \tag{9}$$

$$\nabla\mathbf{E}(\mathbf{r},t) = \frac{1}{\varepsilon}\left(\rho(\mathbf{r},t) + \rho_0(\mathbf{r})\exp\left(\frac{\sigma}{\varepsilon}(t_0 - t)\right)\right), \tag{10}$$

$$\nabla\mathbf{B}(\mathbf{r},t) = 0 \tag{11}$$

are provided by the solutions of the differential equations for the scalar representatives a, b, and c

$$\begin{aligned}\nabla^2 a(\mathbf{r},t) &= \varepsilon\mu\partial_t^2 a(\mathbf{r},t) + \mu\sigma\partial_t a(\mathbf{r},t)\\ &\quad -\mu\int_{t_0}^{t} j_\mathrm{e}(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau - t)\right)\mathrm{d}\tau,\end{aligned} \tag{12}$$

$$\nabla^2 b(\mathbf{r},t) = \varepsilon\mu\partial_t^2 b(\mathbf{r},t) + \mu\sigma\partial_t b(\mathbf{r},t), \tag{13}$$

$$\nabla^2 c(\mathbf{r}) = -\frac{1}{\varepsilon}\rho_0(\mathbf{r}), \tag{14}$$

if the magnetic and electric fields $\mathbf{B}$ *and* $\mathbf{E}$ *are computed from a, b, and c according to*

$$\mathbf{B}(\mathbf{r},t) = \nabla\times\mathbf{v}\left(\frac{\sigma}{\varepsilon} + \partial_t\right)a(\mathbf{r},t) - \nabla\times\nabla\times\mathbf{v}b(\mathbf{r},t) \tag{15}$$

$$\begin{aligned}\mathbf{E}(\mathbf{r},t) = \frac{1}{\varepsilon\mu}&\nabla\times\nabla\times\mathbf{v}a(\mathbf{r},t) + \nabla\times\mathbf{v}\partial_t b(\mathbf{r},t)\\ &-\nabla c(\mathbf{r})\exp\left(\frac{\sigma}{\varepsilon}(t_0 - t)\right)\\ &-\frac{1}{\varepsilon}\int_{t_0}^{t}\mathbf{j}_\mathrm{e}(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau - t)\right)\mathrm{d}\tau.\end{aligned} \tag{16}$$

The supporting vector field $\mathbf{v}$ *must be parallel to the density of the enforced current*

$$\mathbf{j}_\mathrm{e}(\mathbf{r},t) = \mathbf{v}j_\mathrm{e}(\mathbf{r},t). \tag{17}$$

For the general three-dimensional density, this amounts to choose three linearly independent supporting fields according to (5), to introduce three representatives a, and to solve three equations (12).

The equations (12) through (17) constitute the first major item of this article.

The proof proceeds in three steps since Maxwell's equations constitute a linear system. We compose the general solution from the particular ones. Let us begin with the extraction of nonhomogeneities.

First step: The ansatz

$$\begin{aligned}&\mathbf{E}(\mathbf{r},t) = -\nabla c(\mathbf{r})\exp\left(\frac{\sigma}{\varepsilon}(t_0 - t)\right),\\ &\mathbf{B}(\mathbf{r},t) = \mathbf{0}\end{aligned} \tag{18}$$

is to extract the nonhomogeneity of (10) with $\rho_0(\mathbf{r})$. The ansatz satisfies all Maxwellian equations except the third (10). The third yields the Poisson equation (14) for the scalar potential $c(\mathbf{r})$. Please find ansatz (18) linearly enclosed in the general representation formulae (15) and (16).

The Ohmian current (6) caused by this electric field

$$\mathbf{j}_0(\mathbf{r},t) = -\sigma\nabla c(\mathbf{r})\exp\left(\frac{\sigma}{\varepsilon}(t_0 - t)\right) \tag{19}$$

balances in (7) the term with the charge density $\rho_0(\mathbf{r})$. We discard them both to go on with a simplified equation of continuity and observe that it still facilitates the elimination of $\rho(\mathbf{r},t)$.

Hence, for the remaining nonhomogeneities of Maxwell's equations, we do not miss anything when we differentiate the remainder of (10) with respect to time

$$\nabla\left(\partial_t\mathbf{E}(\mathbf{r},t)+\frac{\sigma}{\varepsilon}\mathbf{E}(\mathbf{r},t)+\frac{1}{\varepsilon}\mathbf{j}_{\mathrm{e}}(\mathbf{r},t)\right)=0. \quad (20)$$

The equation of continuity (7) was used to get rid of the charge density $\rho(\mathbf{r},t)$. Ohm's law (6) was applied to eliminate the current density $\mathbf{j}(\mathbf{r},t)$. Equation (20), however, is guaranteed if the first Maxwellian equation (8) is fulfilled. To see this, one just has to take its divergence.

We can completely forget about the third and fourth Maxwellians (10) and (11) when we represent the magnetic field $\mathbf{B}$ by a vector potential $\mathbf{A}(\mathbf{r},t)$, i. e. $\mathbf{B}(\mathbf{r},t)=\nabla\times\mathbf{A}(\mathbf{r},t)$. This shall be done henceforth. One should, however, keep in mind that mere introduction of a vector potential does not help much. The differential equations for the vector potential are coupled in a similiar way as Maxwell's equations for the electric and magnetic fields. We must construct special vector potentials to obtain uncoupled differential equations.

Second step: The ansatz

$$\mathbf{B}(\mathbf{r},t)=\nabla\times\mathbf{v}\alpha(\mathbf{r},t) \quad (21)$$

introduces, as announced, a special vector potential, $\mathbf{A}(\mathbf{r},t)=\mathbf{v}\alpha(\mathbf{r},t)$, viz. a predetermined vector field times a free scalar function.

It shall be used to extract the nonhomogeneity $\mathbf{j}_{\mathrm{e}}(\mathbf{r},t)$. The first Maxwellian (8) can be written as

$$\begin{aligned}\partial_t\mathbf{E}(\mathbf{r},t)+\frac{\sigma}{\varepsilon}\mathbf{E}(\mathbf{r},t)+\frac{1}{\varepsilon}\mathbf{j}_{\mathrm{e}}(\mathbf{r},t)&\\ =\frac{1}{\varepsilon\mu}\nabla\times\nabla\times\mathbf{v}\alpha(\mathbf{r},t).&\end{aligned} \quad (22)$$

Integrating this equation with respect to time yields a representation of the electric field

$$\begin{aligned}\mathbf{E}(\mathbf{r},t)=\frac{1}{\varepsilon\mu}\nabla\times\nabla\times\mathbf{v}\int_{t_0}^{t}\alpha(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau-t)\right)\mathrm{d}\tau&\\ -\frac{1}{\varepsilon}\int_{t_0}^{t}j_{\mathrm{e}}(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau-t)\right)\mathrm{d}\tau.&\end{aligned} \quad (23)$$

Most people prefer differentiations over integrations. Therefore we redefine the representative

$$a(\mathbf{r},t)=\int_{t_0}^{t}\alpha(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau-t)\right)\mathrm{d}\tau. \quad (24)$$

This transforms (21) to

$$\mathbf{B}(\mathbf{r},t)=\nabla\times\mathbf{v}\left(\frac{\sigma}{\varepsilon}+\partial_t\right)a(\mathbf{r},t) \quad (25)$$

and (23) to

$$\begin{aligned}\mathbf{E}(\mathbf{r},t)=\frac{1}{\varepsilon\mu}\nabla\times\nabla\times\mathbf{v}a(\mathbf{r},t)&\\ -\frac{1}{\varepsilon}\int_{t_0}^{t}\mathbf{j}_{\mathrm{e}}(\mathbf{r},\tau)\exp\left(\frac{\sigma}{\varepsilon}(\tau-t)\right)\mathrm{d}\tau.&\end{aligned} \quad (26)$$

Please find these terms enclosed in the representation formulae (15) and (16).

The only Maxwellian equation (9) not yet fulfilled produces after insertion of (25) and (26) the nonhomogeneous telegraph equation (12). To see this, one has to apply the lemma of triple curl (3). This is possible only if the current density is parallel to the supporting field, i. e. if condition (17) is fulfilled.

Now that we have taken care of the enforced current $\mathbf{j}_{\mathrm{e}}(\mathbf{r},t)$, we may assume for the rest of the proof that only the current caused by the inner electric field via Ohm's law (6) remains. This is not a nonhomogeneity. The first Maxwellian (8) can be simplified to

$$\nabla\times\mathbf{B}(\mathbf{r},t)=\mu\sigma\mathbf{E}(\mathbf{r},t)+\varepsilon\mu\partial_t\mathbf{E}(\mathbf{r},t). \quad (27)$$

Third step: The ansatz

$$\mathbf{E}(\mathbf{r},t)=\nabla\times\mathbf{v}\beta(\mathbf{r},t) \quad (28)$$

inserted into the second Maxwellian (9) produces the representation

$$\mathbf{B}(\mathbf{r},t)=-\nabla\times\nabla\times\mathbf{v}\int_{t_0}^{t}\beta(\mathbf{r},\tau)\mathrm{d}\tau. \quad (29)$$

Again, for calculational convenience we redefine

$$b(\mathbf{r},t)=\int_{t_0}^{t}\beta(\mathbf{r},\tau)\mathrm{d}\tau \quad (30)$$

to obtain from (28) and (29) the still missing terms in the general representation formulae (15) and (16). The only Maxwellian (27) which is yet not satisfied causes after application of the lemma of triple curl (3) the condition (13), which is again a telegraph equation, this time, however, a homogeneous one. Q.E.D.

The general representation theorem copes with almost everything discussed in most monographies on electrodynamics [16]: electrostatics, magnetostatics, also electric discarge, skin effects in conductors, propagation of electromagnetic waves in space, guides and resonating cavities, optics, metallic optics too, radiation from antennae and all kinds of electromagnetic scattering, especially Mie scattering, which appears in thoses monographies as an extremely difficult case. For the theory of supraconductivity, Ohm's law (6) must be replaced with London's equation, but this even simplifies the derivation of a slightly modified representation theorem. Only the electrodynamics of nonisotropic materials cannot be tackled in this way.

Vector potential and double vector potential, dubbed Hertzian vector, are known since long. The furthest reaching representation of electromagnetic fields in terms of these potentials was probably found by Max von Laue [17]. For instance, the term in (16) with the integral over the current density, which bewilders saplings, was given by Laue, but without consideration of conductivity. However, Laue did not disentangle Maxwell's equations. Many scientists subject their vector potentials to so-called Coulomb or Lorentz gauges. These gauges are related to invariances, but initial and boundary data break them. Thus in initial- and boundary-value problems, these vector potentials mislead. What one must use instead are adaptive scalars times fixed vectors. One can imagine suitable vector potentials as scalars riding on prepared vector fields like trains ride on rails. Inspiring in this direction was Peter Debye's simplified solution of Mie's problem using a special Hertzian vector [18]. A remarkably complete list of scalar functions that are useful for the disentanglement of Maxwell's equations was presented by Meixner and Schäfke [19]. Yet this list is valid only for free propagation of harmonic electromagnetic waves. The general principle of representation, i. e. the lemma of triple curl, was obviously unknown to all these scientists. Later on electronic computers spread and absorbed interest. So this gap in mathematical physics was filled only in 1985 [13].

The brightest indication that scientists do not understand the general principle of representation is the lack of a reasonable theory of electromagnetic diffraction. There were attempts, for example [20 – 23], to account for the vector fields, but they produced after longish explications only approximations – if at all. Also, partial wave expansions are not helpful since they converge well only for long waves [19]. By contrast, the theory that will be developed in the following sections is straightforward, yields exact equations, and is easily applied to practical problems. Kirchhoff or his scholars would have done this if they only would have known the approach.

The situation was similar in hydrodynamics. With the same methods as explained here it was possible to derive for the first time turbulence in pipes from the Navier-Stokes equation [24]. In 1989 the author published the prediction that pipe turbulence consists of transients [25]. It was verified experimentally in 2006 [26].

3. Boundary Values on Perfect Conductors

While the representatives a, b, and c obey separated differential equations, they are usually tied together in the conditions on the boundary- and initial values of **B** and **E**. For simplicity, consider the diffraction of electromagnetic waves, where the explicit consideration of sources is not necessary. We do not need a scalar potential c. a and b must solve only homogeneous telegraph equations, cf. (12) and (13).

Waves are diffracted when impeded by a screen. The only way to get on with boundary conditions rather than with conditions of transition is to have the screen made of perfectly conducting material. Then, because of Ohm's law (6) for $\sigma \to \infty$, the electric field **E** cannot maintain any tangent component on the screen. The waves do not penetrate.

Let **t** denote any vector tangent to the screen S not to be confounded with the time t. The boundary conditions follow from the representation (16)

$$\frac{1}{\varepsilon\mu}(\nabla\times\nabla\times\mathbf{v}a(\mathbf{r},t))\mathbf{t}+\partial_t(\nabla\times\mathbf{v}b(\mathbf{r},t))\mathbf{t}=0 \quad \text{for } \mathbf{r}\in S. \tag{31}$$

These are two equations because there are two linearly independent tangential vectors **t** on a two-dimensional boundary. All the more it is surprising that the two unknowns a and b can satisfy the next four equations

$$\begin{aligned}&(\nabla\times\nabla\times\mathbf{v}a(\mathbf{r},t))\mathbf{t}=0 \text{ and}\\&(\nabla\times\mathbf{v}b(\mathbf{r},t))\mathbf{t}=0 \text{ for } \mathbf{r}\in S.\end{aligned} \tag{32}$$

This is possible if the supporting field **v** is parallel or perpendicular to the diffracting screen.

To prove this, define a local Cartesian coordinate system

$$\mathbf{r}=\mathbf{e}_x x+\mathbf{e}_y y+\mathbf{e}_z z, \qquad \mathbf{t}=\mathbf{e}_x t_x+\mathbf{e}_y t_y \tag{33}$$

such that it unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ be parallel to the screen, whereas $\mathbf{e}_z$ pierce it normally. Just the components t_x and t_y of the tangential vector $\mathbf{t}$ are arbitrary though constant. Consequently, only x and y components of curl and double curl need to be considered if the boundary conditions (32) are to be satified.

$\mathbf{v} = \mathbf{e}_z$ is according to (5) an admissible choice for the supporting vector field. The equations (32) then become

$$\begin{aligned}(\nabla\times\mathbf{e}_z b)\mathbf{t} &= \mathbf{e}_x\mathbf{t}\partial_y b - \mathbf{e}_y\mathbf{t}\partial_x b = 0,\\ (\nabla\times\nabla\times\mathbf{e}_z a)\mathbf{t} &= \mathbf{e}_x\mathbf{t}\partial_x\partial_z a + \mathbf{e}_y\mathbf{t}\partial_y\partial_z a = 0.\end{aligned} \tag{34}$$

The differentiations with respect to x and y are inner ones since the tangent vector $\mathbf{t}$ of equation (32) is spanned by $\mathbf{e}_x$ and $\mathbf{e}_y$. Both $\partial_x b = 0$ and $\partial_y b = 0$ are satisfied on the screen if $b = 0$. Equally, both $\partial_x\partial_z a = 0$ and $\partial_y\partial_z a = 0$ are satisfied on the screen if $\partial_z a = 0$.

Generally the supporting vector field $\mathbf{v}$ (5) is not constant. However, if it is normal to the surface, we can construe its length as a factor of the representative and repeat the preceding calculation. Hence, the following theorem:

Theorem on Boundary Conditions. *A supporting vector field* $\mathbf{v}$ *normal to the surface S of a perfect conductor induces homogeneous boundary conditions of the first kind for the representative b, whereas the representative a must fulfill homogeneous boundary conditions of the second kind*

$$b(\mathbf{r},t) = 0 \ \textit{and}\ \partial_n|\mathbf{v}|a(\mathbf{r},t) = 0 \ \textit{for}\ \mathbf{r}\in S. \tag{35}$$

Oppositely, a supporting vector field tangent to the surface induces homogeneous boundary conditions of the first kind for the representative a, whereas the representative b must fulfill homogeneous boundary conditions of the second kind

$$a(\mathbf{r},t) = 0 \ \textit{and}\ \partial_n b(\mathbf{r},t) = 0 \ \textit{for}\ \mathbf{r}\in S. \tag{36}$$

∂_n *denotes differentiation along the normal on S.*

For convenience of reference, the author bundled the essentials of this section in one theorem. Its second part still has to be proven. With the local coordinate system introduced above, $\mathbf{v} = \mathbf{e}_x$ is according to (5) an admissible choice for the supporting vector field, too. The equations (32) are now

$$\begin{aligned}(\nabla\times\mathbf{e}_x b)\mathbf{t} &= \mathbf{e}_y\mathbf{t}\partial_z b = 0,\\ (\nabla\times\nabla\times\mathbf{e}_x a)\mathbf{t} &= \mathbf{e}_x\mathbf{t}(\partial_x^2 - \varepsilon\mu\partial_t^2 - \mu\sigma\partial_t)a\\ &\quad + \mathbf{e}_y\mathbf{t}\partial_x\partial_y a = 0.\end{aligned} \tag{37}$$

For the first component of the latter equation the homogenous telegraph equation (12) was exerted. Again, the differentiations with respect to x and y are inner ones. The same is true for the differentiations with respect to t because boundary conditions must hold for all times. Hence, $\partial_z b = 0$ and $a = 0$ on the screen. When this is written without Cartesian coordinates, the second part of the above theorem emerges. A correction with the length of the supporting vector is not necessary here because $\mathbf{v}$ as defined in (5) does not vary in the direction of the normal if it is perpendicular to that normal. Q.E.D.

Screens do not enclose radiation. Much of it spreads in open space. Thus the boundary conditions need to be completed, namely by retardation

$$\begin{aligned}&a \ \text{or}\ b(\mathbf{r},t) \sim\\ &\left(\frac{f_{a\ \text{or}\ b}(|\mathbf{r}| - t/\sqrt{\varepsilon\mu})}{|\mathbf{r}|} + O(\sigma)\right)\mathrm{e}^{-\sigma t/2\varepsilon}\\ &\text{for}\ |\mathbf{r}|\to\infty\end{aligned} \tag{38}$$

meaning that waves trail away in nirvana and never return. The functions f_a and f_b may depend on the direction of the radiation, but they depend on the distance $|\mathbf{r}|$ only through the compound argument $|\mathbf{r}| - t/\sqrt{\varepsilon\mu}$. $O(\sigma)$ is E. Landau's order symbol to appraise neglected terms on the right-hand side, see e. g. [27, Section 1.1]. Notice: σ denotes here the conductivity of the propagating medium, for example air.

Hence, we may have boundary conditions that do not couple the representatives. Yet to profit from the theorem we need supporting vector fields that are *everywhere* either normal or tangent to the screen. According to (5) this can be achieved for four types of screens: for plane ones, for parts of spheres, for parts of cones, and for parts of cylinders, i. e. for cylinders with arbitrary cross sections. In this article we will be busy enough to cope with diffraction by plane screens and will use the theorem with the tangent supporting vector field (36).

4. Initial Values Transformed

We account for arbitrary initial values using a Fourier or rather a Laplace transform. All fields are proportional to $\exp(-\mathrm{i}\omega t)$.

$$\mathbf{B}(\mathbf{r},t) = \mathbf{B}_k(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}, \quad \mathbf{E}(\mathbf{r},t) = \mathbf{E}_k(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}, \tag{39}$$

$$a(\mathbf{r},t) = a_k(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}, \quad b(\mathbf{r},t) = b_k(\mathbf{r})\mathrm{e}^{-\mathrm{i}\omega t}. \tag{40}$$

The telegraph equations (12) and (13) become thus Helmholtz equations

$$\nabla^2 a_k(\mathbf{r})+k^2 a_k(\mathbf{r})=0, \quad \nabla^2 b_k(\mathbf{r})+k^2 b_k(\mathbf{r})=0. \tag{41}$$

The wave number k depends on the frequency ω according to

$$k^2=\varepsilon\mu\omega^2+\mathrm{i}\mu\sigma\omega. \tag{42}$$

One of the two parameters, ω or k, can be chosen as real. In the latter case, a negative imaginary part of ω describes fading with time $t \to \infty$. This is expressed through

$$\omega=\frac{k}{\sqrt{\varepsilon\mu}}-\mathrm{i}\frac{\sigma}{2\varepsilon}+O(\sigma^2) \tag{43}$$

and corresponds to equation (38). In the first case, a positive imaginary part of k describes attenuation in space as $|\mathbf{r}| \to \infty$. Modulus and phase of k can be read from

$$k=\sqrt[4]{\varepsilon^2\mu^2\omega^4+\mu^2\sigma^2\omega^2}\exp\left(\frac{\mathrm{i}}{2}\arctan\frac{\sigma}{\varepsilon\omega}\right). \tag{44}$$

The point to be made is that the phase of k varies only between 0 and $\pi/4$. This will matter in the discussions of Section 7.

Finally, we must translate retardation (38) into the language of Fourier transforms. Both representatives a_k and b_k behave as leaving spherical waves

$$a \text{ or } b_k(\mathbf{r})=F_{a \text{ or } b}\frac{\exp(\mathrm{i}k|\mathbf{r}|)}{|\mathbf{r}|}+O(|\mathbf{r}|^{-2}) \text{ for } |\mathbf{r}|\to\infty \tag{45}$$

at which the real parts of k and ω are supposed to carry the same sign. The scattering amplitudes F_a and F_b may depend on the wave number k and the direction of the radiation, but not on $|\mathbf{r}|$. Though Sommerfeld preferred to write this as $\partial_{|\mathbf{r}|}a_k(\mathbf{r})=\mathrm{i}ka_k(\mathbf{r})+O(|\mathbf{r}|^{-2})$ etc., the author persists in calling (45) Sommerfeld's radiation condition.

5. Kirchhoff's Theory Corrected

Altogether we found for a plane screen that the representatives $a_k(\mathbf{r})$ and $b_k(\mathbf{r})$ can be computed separately. Both satisfy Helmholtz equations (41) and exhibit the same behaviour in infinity (45). Yet on the screen $a_k(\mathbf{r})$ must solve a boundary-value problem of the first kind, whereas $b_k(\mathbf{r})$ is subject to boundary conditions of the second kind, see equations (36).[2]

Modifications of Kirchhoff's integral (2) inaugurated by Sommerfeld are useful to solve both boundary-value problems. Instead of the spherical wave $\exp(\mathrm{i}k|\mathbf{r}_\mathrm{p}-\mathbf{r}|)/|\mathbf{r}_\mathrm{p}-\mathbf{r}|$ flexible Green functions $G(\mathbf{r}_\mathrm{p},\mathbf{r})$

$$\psi(\mathbf{r}_\mathrm{p})=\frac{1}{4\pi}\int\!\!\int_F \big(G(\mathbf{r}_\mathrm{p},\mathbf{r})\partial_n\psi(\mathbf{r})-\psi(\mathbf{r})\partial_n G(\mathbf{r}_\mathrm{p},\mathbf{r})\big)\,\mathrm{d}f \tag{46}$$

are introduced. Kirchhoff's formula remains valid if these Green functions fulfill the Helmholtz equation (1), if they have the same singularity as the spherical wave

$$\partial_{|\mathbf{r}_\mathrm{p}-\mathbf{r}|}G(\mathbf{r}_\mathrm{p},\mathbf{r})=-(\mathbf{r}_\mathrm{p}-\mathbf{r})^{-2}+O(|\mathbf{r}_\mathrm{p}-\mathbf{r}|^{-1}) \quad \text{for } \mathbf{r}_\mathrm{p}\to\mathbf{r} \tag{47}$$

and respect the radiation condition as in equation (45).

In a boundary-value problem of the first kind, the values of the function $\psi(\mathbf{r})$ itself are known on the plane, but no information on the values of the derivative $\partial_n\psi(\mathbf{r})$ is available beforehand. Hence, we need a Green function that is zero on the plane lest the unknown values matter. We find from (46)

$$\psi(\mathbf{r}_\mathrm{p})=-\frac{1}{4\pi}\int\!\!\int_F \partial_n G_1(\mathbf{r}_\mathrm{p},\mathbf{r})\psi(\mathbf{r})\,\mathrm{d}f \quad \text{if } G_1(\mathbf{r}_\mathrm{p},\mathbf{r})=0 \text{ for } \mathbf{r}\in F. \tag{48}$$

On the other hand, if the values of the derivative are known, we must do without the function itself. For the solution of the boundary-value problem of the second kind we need a second Green function

$$\psi(\mathbf{r}_\mathrm{p})=\frac{1}{4\pi}\int\!\!\int_F G_2(\mathbf{r}_\mathrm{p},\mathbf{r})\partial_n\psi(\mathbf{r})\,\mathrm{d}f \quad \text{if } \partial_n G_2(\mathbf{r}_\mathrm{p},\mathbf{r})=0 \text{ for } \mathbf{r}\in F. \tag{49}$$

For F being a plane characterized in Cartesian coordinates (33) by, say, $z=0$, both Green functions are easily found as a spherical wave around the point of measurement $\mathbf{r}_\mathrm{p}$ plus its mirrored image on the other side of the plane

$$G_{1,2}(\mathbf{r}_\mathrm{p},\mathbf{r})=\frac{\exp(\mathrm{i}k|\mathbf{r}_\mathrm{p}-\mathbf{r}|)}{|\mathbf{r}_\mathrm{p}-\mathbf{r}|}\mp\frac{\exp(\mathrm{i}k|\mathbf{r}_\mathrm{m}-\mathbf{r}|)}{|\mathbf{r}_\mathrm{m}-\mathbf{r}|} \tag{50}$$

[2] From now on, the index p at the position of the probe $\mathbf{r}_\mathrm{p}$ is indispensable again as there are other locations which must be discriminated, viz. the position of the source $\mathbf{r}_\mathrm{q}$ and arbitrary points on the screen $\mathbf{r}$.

with

$$\begin{aligned}\mathbf{r}_\mathrm{p} &= \mathbf{e}_x x_\mathrm{p} + \mathbf{e}_y y_\mathrm{p} + \mathbf{e}_z z_\mathrm{p},\\ \mathbf{r}_\mathrm{m} &= \mathbf{e}_x x_\mathrm{p} + \mathbf{e}_y y_\mathrm{p} - \mathbf{e}_z z_\mathrm{p}.\end{aligned} \tag{51}$$

Both functions own the correct singularity as in (47) and respect the radiation condition (45).

Furthermore, it shall be assumed that the primary wave is created by a point-like source at $\mathbf{r}_\mathrm{q}$ behind the plane

$$\psi(\mathbf{r}) = \frac{\exp(\mathrm{i}k|\mathbf{r}-\mathbf{r}_\mathrm{q}|)}{|\mathbf{r}-\mathbf{r}_\mathrm{q}|}. \tag{52}$$

This kind of primary wave is to be prefered over the usual plane wave because it respects a radiation condition of type (45) and ensures thus the validity of the generalized Kirchhoff formula (46). Also it is closer to experiments as it is simpler to produce an approximate spherical wave than an approximate plain wave. The beloved plain wave can be obtained from the spherical wave by a straightforward limiting transition $|\mathbf{r}_\mathrm{q}| \longrightarrow \infty$. Radiation from higher multipoles can be derived by differentiation with respect to $\mathbf{r}_\mathrm{q}$ and the general primary wave by superposition.

Writing down the integral for the boundary conditions of the first kind (48) in Cartesian coordinates ends up with

$$\begin{aligned}\psi(\mathbf{r}_\mathrm{p}) = {} & \frac{\exp(\mathrm{i}kR_\mathrm{pq})}{R_\mathrm{pq}} \frac{-z_\mathrm{p}R_\mathrm{pq}}{2\pi}\\ & \cdot \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\mathrm{i}k - 1/R_\mathrm{p}(x,y)}{R_\mathrm{p}^2(x,y)R_\mathrm{q}(x,y)} \mathrm{e}^{\mathrm{i}k\Delta(x,y)}\mathrm{d}x\mathrm{d}y\end{aligned} \tag{53}$$

and the integral for the boundary conditions of the second kind (49) becomes

$$\begin{aligned}\psi(\mathbf{r}_\mathrm{p}) = {} & \frac{\exp(\mathrm{i}kR_\mathrm{pq})}{R_\mathrm{pq}} \frac{z_\mathrm{q}R_\mathrm{pq}}{2\pi}\\ & \cdot \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{\mathrm{i}k - 1/R_\mathrm{q}(x,y)}{R_\mathrm{q}^2(x,y)R_\mathrm{p}(x,y)} \mathrm{e}^{\mathrm{i}k\Delta(x,y)}\mathrm{d}x\mathrm{d}y\end{aligned} \tag{54}$$

with

$$\begin{aligned}R_\mathrm{p}(x,y) &= |\mathbf{r}_\mathrm{p} - \mathbf{r}(x,y)|\\ &= \sqrt{(x-x_\mathrm{p})^2 + (y-y_\mathrm{p})^2 + z_\mathrm{p}^2},\\ R_\mathrm{q}(x,y) &= |\mathbf{r}(x,y) - \mathbf{r}_\mathrm{q}|\\ &= \sqrt{(x-x_\mathrm{q})^2 + (y-y_\mathrm{q})^2 + z_\mathrm{q}^2},\\ R_\mathrm{pq} &= |\mathbf{r}_\mathrm{p} - \mathbf{r}_\mathrm{q}|\\ &= \sqrt{(x_\mathrm{p}-x_\mathrm{q})^2 + (y_\mathrm{p}-y_\mathrm{q})^2 + (z_\mathrm{p}-z_\mathrm{q})^2},\end{aligned} \tag{55}$$

and the *triangle function*

$$\Delta(x,y) = R_\mathrm{p}(x,y) + R_\mathrm{q}(x,y) - R_\mathrm{pq}, \tag{56}$$

which measures the difference of lengths when the points $\mathbf{r}_\mathrm{p}$ and $\mathbf{r}_\mathrm{q}$ are either directy connected or via an arbitrary point $\mathbf{r} = \mathbf{e}_x x + \mathbf{e}_y y$ on the intermediate plane.

It is not by accident that the ψ's in the equations (53) and (54) are identical. The factors behind the spherical wave $\exp(\mathrm{i}kR_\mathrm{pq})/R_\mathrm{pq}$ including the double integrals extending from $-\infty$ to $+\infty$ have both the value 1; a proof of this fact will be given in Section 11. The formulae (53) and (54) are just different mathematical realizations of Huygens' principle: A wave propagating from a source point that strikes a plane excites there secondary waves interfering to reproduce the original wave.

Diffraction happens only if a screen covers parts of the plane. Using the boundary conditions in the theorem with the tangent supporting field (36) yields explicit expressions for the representatives

$$\begin{aligned}a_k(\mathbf{r}_\mathrm{p}) &= \frac{\exp(\mathrm{i}kR_\mathrm{pq})}{R_\mathrm{pq}} \mathit{II}(\mathbf{r}_\mathrm{p},\mathbf{r}_\mathrm{q}),\\ b_k(\mathbf{r}_\mathrm{p}) &= \frac{\exp(\mathrm{i}kR_\mathrm{pq})}{R_\mathrm{pq}} \mathit{II}(\mathbf{r}_\mathrm{q},\mathbf{r}_\mathrm{p})\end{aligned} \tag{57}$$

with the double integral

$$\begin{aligned}\mathit{II}(\mathbf{r}_\mathrm{p},\mathbf{r}_\mathrm{q}) = {} & \frac{-|z_\mathrm{p}|R_\mathrm{pq}}{2\pi}\\ & \cdot \iint_D \frac{\mathrm{i}k - 1/R_\mathrm{p}(\xi,\eta)}{R_\mathrm{p}^2(\xi,\eta)R_\mathrm{q}(\xi,\eta)} \frac{\partial(x,y)}{\partial(\xi,\eta)} \mathrm{e}^{\mathrm{i}k\Delta(\xi,\eta)}\mathrm{d}\xi\mathrm{d}\eta.\end{aligned} \tag{58}$$

To be general enough for all applications, arbitrary variables ξ and η shall substitute the Cartesian ones $x(\xi,\eta)$ and $y(\xi,\eta)$. The functional determinant $\partial(x,y)/\partial(\xi,\eta)$ enters to transform the element of the surface $\mathrm{d}f = \mathrm{d}x\mathrm{d}y$. The integral II extends only over the aperture or diaphragm D in the surface F. The arbitrariness of ξ and η shall be used for a simple description of the aperture in the way that freely varying ξ with a fixed η depicts an edge. The aperture is described, for example, by $-\infty < \xi < \infty$ and $-\eta_- < \eta < \eta_+$.

The enthusing result is that one needs to evaluate only one integral for the full electromagnetic theory of diffraction with all possible polarizations. There is scarcely more work to be done than in the theory for scalar waves. The addititional work consists in some elementary differentiations as prescribed in the representation formulae (15) and (16) simplified for Fourier transforms as

$$\mathbf{B}_k(\mathbf{r}_\mathrm{p}) = \left(\frac{\sigma}{\varepsilon} - \mathrm{i}\omega\right) \nabla_\mathrm{p} \times \mathbf{t} a_k(\mathbf{r}_\mathrm{p}) - \nabla_\mathrm{p} \times \nabla_\mathrm{p} \times \mathbf{t} b_k(\mathbf{r}_\mathrm{p}), \tag{59}$$

$$\mathbf{E}_k(\mathbf{r}_\mathrm{p}) = \frac{1}{\varepsilon\mu} \nabla_\mathrm{p} \times \nabla_\mathrm{p} \times \mathbf{t} a_k(\mathbf{r}_\mathrm{p}) - \mathrm{i}\omega \nabla_\mathrm{p} \times \mathbf{t} b_k(\mathbf{r}_\mathrm{p}) \tag{60}$$

with a supporting vector field $\mathbf{v} = \mathbf{t}$ that consists of a constant tangent as in (33).

The equations (57) through (60) constitute the second major item of this article. The integral (58) can be evaluated using various techniques. The author will discuss in the following sections only one: the method of stationary phase applicable for short waves with $\Re k \to \infty$ and $\Im k > 0$ as conditions on the real and imaginary parts of the wave number k, respectively.

While equations (57) and (58) solve mathematically well-posed boundary-value problems without any error, they do not describe physical diffraction exactly. When a conducting screen diffracts a wave, reflection cannot be avoided. Most of the reflected wave stays in the outer space, $z_\mathrm{p} < 0$, but it is also diffracted. A tiny fraction of reflection invades inner space $z_\mathrm{p} > 0$. We will learn to handle this in Section 11.

6. Light and Shadow

The triangle function rules the diffraction of short waves. In its definition (56) interest was focused on the point of the screen $\mathbf{r}$. Yet the triangle function also depends on the points of probe $\mathbf{r}_\mathrm{p}$ and source $\mathbf{r}_\mathrm{q}$

$$\Delta(\mathbf{r}, \mathbf{r}_\mathrm{p}, \mathbf{r}_\mathrm{q}) = |\mathbf{r}_\mathrm{p} - \mathbf{r}| + |\mathbf{r} - \mathbf{r}_\mathrm{q}| - |\mathbf{r}_\mathrm{p} - \mathbf{r}_\mathrm{q}|. \tag{61}$$

The author decided to locate

$$\begin{aligned} &\text{screen } \mathbf{r} = \mathbf{e}_x x + \mathbf{e}_y y + \mathbf{e}_z z \text{ at } z = 0, \\ &\text{probe } \mathbf{r}_\mathrm{p} = \mathbf{e}_x x_\mathrm{p} + \mathbf{e}_y y_\mathrm{p} + \mathbf{e}_z z_\mathrm{p} \text{ at } z_\mathrm{p} > 0, \\ &\text{source } \mathbf{r}_\mathrm{q} = \mathbf{e}_x x_\mathrm{q} + \mathbf{e}_y y_\mathrm{q} + \mathbf{e}_z z_\mathrm{q} \text{ at } z_\mathrm{q} < 0. \end{aligned} \tag{62}$$

The triangle function is positive except at that point $\mathbf{r} = \mathbf{r}_\mathrm{s} = \mathbf{e}_x x_\mathrm{s} + \mathbf{e}_y y_\mathrm{s}$, where the straight line between probe and source pierces the plane

$$x_\mathrm{s} + \mathrm{i}y_\mathrm{s} = \frac{z_\mathrm{p}(x_\mathrm{q} + \mathrm{i}y_\mathrm{q}) - z_\mathrm{q}(x_\mathrm{p} + \mathrm{i}y_\mathrm{p})}{z_\mathrm{p} - z_\mathrm{q}}. \tag{63}$$

Complex notation is prefered because it eases transformation to other coordinate systems, see below. The point $\mathbf{r}_\mathrm{s}$ is the location of the absolute minimum of the triangle function, and all the more it is a *stationary point.*

The consideration holds for fixed points of source and probe. However, if only $\mathbf{r}_\mathrm{q}$ is fixed whereas $\mathbf{r}_\mathrm{p}$ varies while $\mathbf{r} = \mathbf{r}_\mathrm{s}$ slides on the edge of the screen, then

$$\Delta(\mathbf{r}, \mathbf{r}_\mathrm{p}, \mathbf{r}_\mathrm{q}) = 0 \text{ if } \mathbf{r} \text{ on the edge} \tag{64}$$

determines as function of $\mathbf{r}_\mathrm{p}$ a surface, namely the *border of shadow*.

For calculating diffraction, we will need the root of the triangle function. The author utilizes the ambiguity of the root to demand

$$\sqrt{\Delta(\mathbf{r}, \mathbf{r}_\mathrm{p}, \mathbf{r}_\mathrm{q})} = \begin{cases} +|\sqrt{\Delta(\mathbf{r}, \mathbf{r}_\mathrm{p}, \mathbf{r}_\mathrm{q})}| \text{ if } \mathbf{r}_\mathrm{p} \text{ in the shadow,} \\ -|\sqrt{\Delta(\mathbf{r}, \mathbf{r}_\mathrm{p}, \mathbf{r}_\mathrm{q})}| \text{ if } \mathbf{r}_\mathrm{p} \text{ in the light.} \end{cases} \tag{65}$$

It is cogent to assign different signs to the dark and the bright if the function is to be differentiable. The triangle function is an analytic function which depends quadratically on its variables around its minimum defined by (64). Therefore, omitting the signs in (65) would induce a similar discontinuity as in the assignment $\sqrt{x^2} = |x|$. The absolute assignment of the sign, on the contrary, is arbitrary since diffraction either by a screen or its complement is equal; remember Babinet's principle [28, § 11.3].

To decide where there is light or shadow, a handy criterion is needed. There is light on the probe if the screen does not impede the straight connection between source $\mathbf{r}_\mathrm{q}$ and probe $\mathbf{r}_\mathrm{p}$. Thus the positive sign in (65) is to be taken if the stationary point $x_\mathrm{s} + \mathrm{i}y_\mathrm{s}$ in (63) misses the aperture.

Criterion of Light. *When one uses transformed coordinates ξ, η adapted to the screen such that the aperture is described by $\eta_- < \eta < \eta_+$ while ξ varies freely, the negative sign of triangle function's root (65) has to be taken if*

$$\eta_- < \eta_\mathrm{s} < \eta_+. \tag{66}$$

η_s *is calculated from the stationary point (63) via coordinate transformation.*

For example on a screen with a circular aperture, cylindrical coordinates $\rho, \varphi, z = 0$ suit. The aperture is defined by $\rho < \rho_0$ with ρ_0 as the radius of the stop, while φ varies freely. The transformation to cylindrical coordinates is facilitated through

$$\begin{aligned} &x_\mathrm{s} + \mathrm{i}y_\mathrm{s} = \rho_\mathrm{s}\mathrm{e}^{\mathrm{i}\varphi_\mathrm{s}}, \\ &x_\mathrm{p} + \mathrm{i}y_\mathrm{p} = \rho_\mathrm{p}\mathrm{e}^{\mathrm{i}\varphi_\mathrm{p}}, \quad x_\mathrm{q} + \mathrm{i}y_\mathrm{q} = \rho_\mathrm{q}\mathrm{e}^{\mathrm{i}\varphi_\mathrm{q}} \end{aligned} \tag{67}$$

with the ρ's as axial distances and the φ's as azimuthal angles. Insertion into (63) produces two equivalent formulas for the axial distance

$$\begin{aligned} \rho_\mathrm{s} &= \frac{\sqrt{(z_\mathrm{p}\rho_\mathrm{q} - z_\mathrm{q}\rho_\mathrm{p})^2 + 4z_\mathrm{p}z_\mathrm{q}\rho_\mathrm{p}\rho_\mathrm{q}S^2}}{z_\mathrm{p} - z_\mathrm{q}} \\ &= \frac{\sqrt{(z_\mathrm{p}\rho_\mathrm{q} + z_\mathrm{q}\rho_\mathrm{p})^2 - 4z_\mathrm{p}z_\mathrm{q}\rho_\mathrm{p}\rho_\mathrm{q}C^2}}{z_\mathrm{p} - z_\mathrm{q}} \end{aligned} \tag{68}$$

with the abbreviations

$$C = \cos\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2}, \quad S = \sin\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2}. \tag{69}$$

$\rho_\mathrm{s} > \rho_0$ is thus the criterion for the domain of shadow, i. e. for the positive sign in (65).

Instead of applying elementary geometry, as was done in this section, one may calculate the stationary point $\xi_\mathrm{s}, \eta_\mathrm{s}$ by simultaneous solution of the equations

$$\begin{aligned} &\partial_\xi \Delta(\xi, \eta) = 0, \quad \partial_\eta \Delta(\xi, \eta) = 0 \\ &\leftrightarrow \xi = \xi_\mathrm{s}, \quad \eta = \eta_\mathrm{s}. \end{aligned} \tag{70}$$

The result is, of course, the same as that given in (63) with subsequent transformation of coordinates, but the computational effort is larger. The author displays the equations (70) only to ease comprehension of the astonishing equation (88) which will appear in Section 8.

7. Using the Error Function for Diffraction

In the theory of diffraction, Fresnel integrals

$$\begin{aligned} C_1(z) &= \sqrt{\frac{2}{\pi}}\int_0^z \cos w^2 \mathrm{d}w, \\ S_1(z) &= \sqrt{\frac{2}{\pi}}\int_0^z \sin w^2 \mathrm{d}w \end{aligned} \tag{71}$$

are still custom, but the error function or rather the *complementary error function* is handier [29, 30]

$$\mathrm{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_z^\infty \mathrm{e}^{-w^2}\mathrm{d}w. \tag{72}$$

All statisticians become perfect opticians when they are willing to handle their favorite function with complex argument.

The error function comprises the Fresnel integrals in a similar fashion as the exponential function contains sine and cosine

$$\mathrm{erfc}(\sqrt{-\mathrm{i}}z) = 1 - \sqrt{-\mathrm{i}}\sqrt{2}((C_1(z) + \mathrm{i}S_1(z)). \tag{73}$$

If z is assumed as real, $\sqrt{-\mathrm{i}}z$ varies on the second main diagonal of the complex plane since $\sqrt{-\mathrm{i}} = (1 - \mathrm{i})/\sqrt{2} = \exp(-\mathrm{i}\pi/4)$. The features known for real argument remain if the complex argument is enclosed between the first and the second main diagonals of the complex plane; expressed by a relation between imaginary and real parts: $|\Im z| \leq |\Re z|$. This is exactly what we need for optics, see (44). For large negative real parts $\Re z \to -\infty$ of the argument the complementary error function starts at the value 2, assumes the value 1 at the origin $z = 0$, and attains the value 0 for large positive real parts $\Re z \to +\infty$. In the crudest approximation, one may think of the complementary error function as 2 for negative arguments and 0 for positive ones. It is a switch.

The asymptotic expansion of the error function familiar on the real axis remains valid in the wedge between the main diagonals which contains the real axis

$$\mathrm{erfc}(z) = \frac{1}{\sqrt{\pi}z}\mathrm{e}^{-z^2}(1 + O(|z|^{-2}) \text{ for } \Re z \to +\infty. \tag{74}$$

The asymptotic expansion on the other side of the complex plane $\Re z < 0$ follows from

$$\mathrm{erfc}(z) = 2 - \mathrm{erfc}(-z). \tag{75}$$

The only refinement due to complexity is that the complementary error function decreases monotonously on the real axis, whereas it takes complex values and both real and imaginary parts oscillate when the argument becomes complex.

8. The Method of Stationary Phase for Two-Dimensional Integrals

Let us recall the asymptotic calculation of one-dimensional integrals

$$I(k, \eta_-, \eta_+) = \int_{\eta_-}^{\eta_+} A(\eta)\mathrm{e}^{\mathrm{i}k\Delta(\eta)}\mathrm{d}\eta \tag{76}$$

for $\Re k \to \infty$ with $\Im k > 0$. We assume that the real function $\Delta(\eta) \geq 0$ is stationary for some $\eta = \eta_\mathrm{s}$, i. e.

$$\Delta(\eta) = \frac{\Delta_{\eta\eta}}{2}(\eta - \eta_\mathrm{s})^2 + O((\eta - \eta_\mathrm{s})^3). \tag{77}$$

The indices at Δ indicate that the function be differentiated twice and the result evaluated at $\eta = \eta_\mathrm{s}$.

The familiar approach in the method of stationary phase is to introduce a new variable

$$v = \sqrt{\frac{\Delta_{\eta\eta}}{2}}(\eta - \eta_\mathrm{s}) \tag{78}$$

and to forget the higher-order terms $O((\eta - \eta_\mathrm{s})^3)$ in equation (77). The function

$$\delta(v) \sim \sqrt{\frac{k}{\pi \mathrm{i}}}\mathrm{e}^{\mathrm{i}kv^2} \quad \text{for } \Re k \to \infty, \quad \Im k > 0 \tag{79}$$

may be construed as a representation of Dirac's delta function. Thus the integral (76) yields

$$I(k, -\infty, +\infty) \sim \sqrt{\frac{2\pi \mathrm{i}}{k\Delta_{\eta\eta}}}A(\eta_\mathrm{s}). \tag{80}$$

The amplitude $A(\eta)$ appears as a constant.

If the limits of integration $\eta_\pm$ are $\pm\infty$, this is correct, but for finite limits the local approximation (78) induces systematic errors. What we have to use instead is the *global map* $\eta \to v$

$$v = \sqrt{\Delta(\eta)}. \tag{81}$$

To rewrite the integral from the variable η to the variable v, we must calculate the differential $\mathrm{d}\eta = (\mathrm{d}\eta/\mathrm{d}v)\mathrm{d}v = (\mathrm{d}v/\mathrm{d}\eta)^{-1}\mathrm{d}v$. Because of (79) it is sufficient to know the value of the differential for $v = 0$ corresponding to $\eta = \eta_\mathrm{s}$. Thus the value of the differential of the global map (81) to be used in the integral is the same as the differential of the local approximation (78). The peculiarity of the global map appears only in the limits:

$$I(k, \eta_-, \eta_+) \sim \sqrt{\frac{2\pi \mathrm{i}}{k\Delta_{\eta\eta}}}A(\eta_\mathrm{s}) \cdot \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta(\eta_-)} - \operatorname{erfc}\sqrt{-\mathrm{i}k\Delta(\eta_+)}}{2}. \tag{82}$$

While the preceding is not familiar, it is known [27, Section 2.9]. Yet in the theory of diffraction one needs to evaluate two-dimensional integrals

$$\mathit{II}(k, \eta_-, \eta_+) = \int_{\eta_-}^{\eta_+}\int_{-\infty}^{\infty} A(\xi, \eta)\mathrm{e}^{\mathrm{i}k\Delta(\xi,\eta)}\mathrm{d}\xi\mathrm{d}\eta. \tag{83}$$

Again it is assumed that the real function $\Delta(\xi, \eta) \geq 0$ is stationary at $\xi_\mathrm{s}, \eta_\mathrm{s}$

$$\begin{aligned}\Delta(\xi, \eta) = {} & \frac{\Delta_{\xi\xi}}{2}(\xi - \xi_\mathrm{s})^2 + \Delta_{\xi\eta}(\xi - \xi_\mathrm{s})(\eta - \eta_\mathrm{s}) \\ & + \frac{\Delta_{\eta\eta}}{2}(\eta - \eta_\mathrm{s})^2 + O(|\xi - \xi_\mathrm{s}|^3 + |\eta - \eta_\mathrm{s}|^3).\end{aligned} \tag{84}$$

It seems to be a hitherto unsolved problem to find a suitable two-dimensional global map $\xi, \eta \to u, v$ such that

$$u^2 + v^2 = \Delta(\xi, \eta). \tag{85}$$

Here is the solution: The map is

$$u = \sqrt{\Delta(\xi, \eta) - \Delta_\mathrm{s}(\eta)}, \tag{86}$$

$$v = \sqrt{\Delta_\mathrm{s}(\eta)}. \tag{87}$$

The function $\Delta_\mathrm{s}(\eta)$ is determined by the

Principle of Utter Exhaust. *Eliminate ξ from the derivative*

$$\partial_\xi \Delta(\xi, \eta) = 0 \leftrightarrow \xi = \Xi_\mathrm{s}(\eta) \tag{88}$$

to find the exhausting dependence $\xi = \Xi_\mathrm{s}(\eta)$. Insert the exhausting dependence into the function $\Delta(\xi, \eta)$ to obtain the exhausting function

$$\Delta_\mathrm{s}(\eta) = \Delta(\Xi_\mathrm{s}(\eta), \eta). \tag{89}$$

The integral (83) is, for $\Re k \to \infty$ and $\Im k > 0$, asymptotically equal to

$$\mathit{II}(k, \eta_-, \eta_+) \sim \frac{2\pi \mathrm{i} A(\xi_\mathrm{s}, \eta_\mathrm{s})}{k\sqrt{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta_{\xi\eta}^2}} \cdot \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_\mathrm{s}(\eta_-)} - \operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_\mathrm{s}(\eta_+)}}{2}. \tag{90}$$

Utter exhaust follows from the following indispensable requirements:

$$\xi = \xi_\mathrm{s}, \eta = \eta_\mathrm{s} \text{ be mapped to } u = 0, \quad v = 0, \tag{91}$$

$$u, v \text{ be real for all } \xi, \eta, \tag{92}$$

$$v^2 = f(\eta) \text{ be a function of } \eta \text{ only.} \tag{93}$$

We need the third requirement (93) to retain the simplicity of the limits in the integral II (83) when mapping ξ, η to u, v. Because of equation (85) and the requirement (92) the function $f(\eta)$ must never exceed

$\Delta(\xi,\eta)$ whatever value ξ takes. Nevertheless, for every η there must exist $\xi = \Xi_{\rm s}(\eta)$ such that equality is reached: $f(\eta) = \Delta(\Xi_{\rm s}(\eta),\eta)$. Otherwise $u = \sqrt{\Delta(\xi,\eta) - f(\eta)}$ cannot take the value 0, a contradiction to the requirement (91). In other words, $f(\eta)$ and $\Delta(\xi,\eta)$ coincide at that value of $\xi = \Xi_{\rm s}(\eta)$, where $\Delta(\xi,\eta)$ gets stationary with respect to ξ. Hence $\Xi_{\rm s}(\eta)$ is determined by elimination of ξ in the condition (88).

The final formula (90) can be understood when compared to formula (82). The double integration generates the factor $\pi{\rm i}/k$, the square of $\sqrt{\pi{\rm i}/k}$. The determinant in $2/\sqrt{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta^2_{\xi\eta}}$ of the map $\xi,\eta \to u,v$ replaces the differential in $\sqrt{2/\Delta_{\eta\eta}}$ of the map $\eta \to v$. Q.E.D.

The principle was named as of utter exhaust because the function $\Delta_{\rm s}(\eta)$ is the largest possible function of η only that fulfills $\Delta_{\rm s}(\eta) \le \Delta(\xi,\eta)$; it takes from the two-variable function as much as a one-variable function can afford, cf. equation (86).

A further intuitive interpretation follows from a comparison of equation (88) with the equations (70). The stationary point is where the function $\Delta(\xi,\eta)$ takes its absolute extremum. On the exhausting dependence, by contrast, $\Delta(\xi,\eta)$ is extremized only with respect to the one variable ξ, whereas the other variable η is fixed. In optics, $\Delta(\xi,\eta)$ essentially measures the distance between points. The absolutely shortest connection is of course a straight line. But the exhausting function $\Delta_{\rm s}(\eta)$ measures a conditionally shortest distance, namely if the connecting line is forced to touch the edge of the diffracting screen. One can determine the exhausting function experimentally using a ribbon of rubber and a lubricated model of the edge.

The difficult part of utter exhaust is the elimination according to equation (88). It is therefore gratifying to possess linear approximations of equations (86) and (87), similar to the one-dimensional case (78). These approximations are

$$u \approx \sqrt{\frac{\Delta_{\xi\xi}}{2}}(\xi - \xi_{\rm s}) + \frac{\Delta_{\xi\eta}}{\sqrt{2\Delta_{\xi\xi}}}(\eta - \eta_{\rm s}) \tag{94}$$

$$v \approx \sqrt{\frac{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta^2_{\xi\eta}}{2\Delta_{\xi\xi}}}(\eta - \eta_{\rm s}). \tag{95}$$

They were found by utter exhaust (88) applied to the quadratic terms on the right-hand side of (84); searching principal axes of the ellipse is not a good idea. For the integral *II* (83), equation (90) can be used when the exhausting function is approximated as

$$\Delta_{\rm s}(\eta_{+{\rm or}-}) \approx \frac{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta^2_{\xi\eta}}{2\Delta_{\xi\xi}}(\eta_{+\ {\rm or}\ -} - \eta_{\rm s})^2. \tag{96}$$

The relation to the border of shadow discussed in Section 6 appears here at first sight.

The method of stationary phase is sometimes blamed as not being mathematical stringent. It is argued that certain integrals do not converge and thus certain errors cannot be estimated. The criticism does not apply here. Guided by mother nature, we made a theory where the parameter k of asymptoticity has a positive imaginary part $\Im k > 0$. This guarantees the convergence of those integrals. We can even write the complete asymptotic series.

Augustin Fresnel invented the *zone construction* in 1816 to prove that light travels as a wave, but this was just a semi-quantitive idea [28, § 8.2; 31, p. 247]. Its mathematical solution is the principle of utter exhaust presented only now.

9. The Universal Formula of Diffraction

To make use of stationary phase for diffraction, the amplitude in the integral (58)

$$A(\xi,\eta) \sim \frac{-z_{\rm p}R_{\rm pq}}{2\pi}\frac{{\rm i}k}{R^2_{\rm p}(\xi_{\rm s},\eta_{\rm s})R_{\rm q}(\xi_{\rm s},\eta_{\rm s})}\frac{\partial(x,y)}{\partial(\xi,\eta)} \tag{97}$$
$$\text{for } \Re k \to \infty, \quad \Im k > 0$$

is to be evaluated at the point of stationarity $\xi_{\rm s}, \eta_{\rm s}$ and multiplied by the factor $2\pi{\rm i}/k\sqrt{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta^2_{\xi\eta}}$ of equation (90).

Trivially, $R_{\rm p}$ and $R_{\rm q}$ are just the two fractions of $R_{\rm pq}$

$$R_{\rm p}(\xi_{\rm s},\eta_{\rm s}) = \frac{z_{\rm p}}{z_{\rm p} - z_{\rm q}}R_{\rm pq},\ R_{\rm q}(\xi_{\rm s},\eta_{\rm s}) = \frac{-z_{\rm q}}{z_{\rm p} - z_{\rm q}}R_{\rm pq}. \tag{98}$$

Distances as they are, they do not depend on the coordinate system. This is different for the determinant $\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta^2_{\xi\eta}$. Let us calculate it first in Cartesian coordinates. Differentiating the triangle function (56) twice and evaluating it at the stationary point gives

$$\Delta_{xx} = \frac{(z_{\rm p} - z_{\rm q})^2((y_{\rm p} - y_{\rm q})^2 + (z_{\rm p} - z_{\rm q})^2)}{-z_{\rm p}z_{\rm q}R^3_{\rm pq}},$$

$$\Delta_{yy} = \frac{(z_{\mathrm{p}}-z_{\mathrm{q}})^2((x_{\mathrm{p}}-x_{\mathrm{q}})^2+(z_{\mathrm{p}}-z_{\mathrm{q}})^2)}{-z_{\mathrm{p}}z_{\mathrm{q}}R_{\mathrm{pq}}^3},$$
$$\Delta_{xy} = \frac{-(z_{\mathrm{p}}-z_{\mathrm{q}})^2(x_{\mathrm{p}}-x_{\mathrm{q}})(y_{\mathrm{p}}-y_{\mathrm{q}})}{-z_{\mathrm{p}}z_{\mathrm{q}}R_{\mathrm{pq}}^3}. \tag{99}$$

Consequently

$$\Delta_{xx}\Delta_{yy} - \Delta_{xy}^2 = \frac{(z_{\mathrm{p}}-z_{\mathrm{q}})^6}{z_{\mathrm{p}}^2 z_{\mathrm{q}}^2 R_{\mathrm{pq}}^4}. \tag{100}$$

The root of this determinant is transformed multiplying it by the functional determinant taken also at the point of stationarity

$$\sqrt{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta_{\xi\eta}^2} = \sqrt{\Delta_{xx}\Delta_{yy} - \Delta_{xy}^2}\,\frac{\partial(x,y)}{\partial(\xi,\eta)}. \tag{101}$$

Additional terms do not occur. They would contain first derivatives of the triangle function which are zero at the stationary point, see conditions (70).

The result is astonishing:

$$\frac{2\pi\mathrm{i}A(\xi_{\mathrm{s}},\eta_{\mathrm{s}})}{k\sqrt{\Delta_{\xi\xi}\Delta_{\eta\eta} - \Delta_{\xi\eta}^2}} = 1 + O(k^{-1}) \tag{102}$$

though it should be noted that it is -1 if the functional determinant is negative, i. e. if the coordinates ξ, η form a left-handed system. The final formulae for the representatives derived from (90) are thus simple and identical

$$a_k(\mathbf{r}_{\mathrm{p}}) = \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}} \cdot \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}(\eta_-)} - \operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}(\eta_+)}}{2} + O(k^{-1}), \tag{103}$$

$$b_k(\mathbf{r}_{\mathrm{p}}) = \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}} \cdot \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}(\eta_-)} - \operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}(\eta_+)}}{2} + O(k^{-1}). \tag{104}$$

Further simplifaction is possible. If the upper edge is removed, i. e. $\eta_+ \to +\infty$, the exhausting function is supposed to increase beyond measure $\Delta_{\mathrm{s}}(\eta_+) \to +\infty$ and, according to the asymptotics (74), the formulae (103) and (104) are reduced to the

Universal Formula of Diffraction. *The diffracted wave is the primary wave times a universal function describing the change from light to shadow.*

$$a \text{ or } b_k(\mathbf{r}_{\mathrm{p}}) = \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}}\,\frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}}}{2} + O(k^{-1}). \tag{105}$$

The geometry of the diffracting edge enters just through the argument of that function. This argument is calculated from the exhausting function (89) evaluated at the position of the edge: $\Delta_{\mathrm{s}} = \Delta_{\mathrm{s}}(\eta_-)$.

From here one may return to the apparently more general formulae above due to the linearity of Maxwell's equations and a superposition of their solutions. So Δ_{s} may be understood as an abbreviation for $\Delta_{\mathrm{s}}(\eta_-)$ or $\Delta_{\mathrm{s}}(\eta_+)$ as required. By the way linearity: Factors on the right-hand sides that do not depend on $\mathbf{r}_{\mathrm{p}}$ are always free. The author omitted them. This is all the more tolerable since the representatives get effective only if taken together with the supporting vector field $\mathbf{t}$ in equations (59) and (60). According to (33) the tangent vector $\mathbf{t}$ contains the components t_x and t_y. They can be chosen to adjust strength and polarization of the primary wave according to experimental givens.

Formula (105) constitutes the main achievement of this article.

One may check the accuracy of (105) by insertion into the Helmholtz equations (41). One finds that (105) satisfies these equations in the order of k^2 identically, but the terms proportional to $k^{3/2}$ cancel only if the exhausting function Δ_{s} satisfies an equation of eikonal type.

$$(\nabla_{\mathrm{p}}\Delta_{\mathrm{s}})^2 = -2\frac{\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{q}}}{R_{\mathrm{pq}}}\nabla_{\mathrm{p}}\Delta_{\mathrm{s}}. \tag{106}$$

Indeed, the fulfillment of this equation follows from the definition of the triangle function (61) and the condition of utter exhaust (88). Hence the largest erroneous term is $O(k)$. But since the order of the Helmholtz operator is k^2, the relative error is $O(k^{-1})$, as expected.

Checking the fulfillment of the boundary conditions yields at first sight a worse result, namely a relative error of $O(k^{-1/2})$. We see this from the asymptotic expansion of the complementary error function (74). This asymptotic expression applies because, directly behind the screen, the probe resides in the shadow where there is, according to (65), $\sqrt{\Delta_{\mathrm{s}}} > 0$. Thus the argument of the error function $\sqrt{-\mathrm{i}k\Delta_{\mathrm{s}}}$ lies in the right-hand side of the complex plane. Hence,

$$a_k(\mathbf{r}_{\mathrm{p}}) = \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}}\,\frac{\exp(\mathrm{i}k\Delta_{\mathrm{s}})}{2\sqrt{-\mathrm{i}\pi k\Delta_{\mathrm{s}}}} + O(k^{-3/2}) \tag{107}$$

instead of an exact zero as required in the boundary conditions (36). A similar result holds for the representative $b_k(\mathbf{r}_{\mathrm{p}})$. To verify its boundary condition, one must calculate its normal derivative, i. e. its derivative with respect to z. While this derivative generally is $O(k)$, its value on the screen is $O(k^{1/2})$ such that the relative error is $O(k^{-1/2})$.

At the second view, both results are better than expected. Firstly, in Kirchhoff's scalar theory of diffraction, the probe must not approach the screen since Kirchhoff's functions become singular [3, p. 9]. In the present theory, all expressions stay regular. As we will see in Section 11, this facilitates a simple trick to suppress errors on the boundaries. Secondly, formula (105), even as it is now, satisfies the boundary conditions not exactly, but accurately. The reason is the product $k\Delta_\mathrm{s}$ in the denominator of (107). The exhausting function Δ_s measures, in a slightly unusual way, the distance of the probe from the border of shadow. The distance between that border and the screen is the largest possible distance available in a diffracting system. Hence, for short waves, both the wave number k and the distance Δ_s are big making (107) negligibly small.

For applications it is worthwile to calculate from (105) the field strengths according to (59) and (60). For simplicity we let $a_k(\mathbf{r}_\mathrm{p}) = 0$. Then the electric field $\mathbf{E}_k(\mathbf{r}_\mathrm{p})$ stems only from the simple curl in (60) and the magnetic field $\mathbf{B}_k(\mathbf{r}_\mathrm{p})$ only from the double curl in (59).

$$\mathbf{E}_k(\mathbf{r}_\mathrm{p})\frac{R_\mathrm{pq}}{\exp(\mathrm{i}kR_\mathrm{pq})} = \frac{(\mathbf{r}_\mathrm{p}-\mathbf{r}_\mathrm{q})\times\mathbf{t}}{R_\mathrm{pq}}\frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_\mathrm{s}}}{2}\omega k + \frac{\nabla_\mathrm{p}\Delta_\mathrm{s}\times\mathbf{t}}{\sqrt{\Delta_\mathrm{s}}}\exp(\mathrm{i}k\Delta_\mathrm{s})\sqrt{\frac{\mathrm{i}\omega^2 k}{4\pi}} + O(k), \tag{108}$$

$$\begin{aligned}\mathbf{B}_k(\mathbf{r}_\mathrm{p})\frac{R_\mathrm{pq}}{\exp(\mathrm{i}kR_\mathrm{pq})} = {} & \frac{(\mathbf{r}_\mathrm{p}-\mathbf{r}_\mathrm{q})((\mathbf{r}_\mathrm{p}-\mathbf{r}_\mathrm{q})\mathbf{t})-\mathbf{t}R_\mathrm{pq}^2}{R_\mathrm{pq}^2}\frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_\mathrm{s}}}{2}k^2 \\ & + \frac{\nabla_\mathrm{p}\Delta_\mathrm{s}((\mathbf{r}_\mathrm{p}-\mathbf{r}_\mathrm{q})\mathbf{t})+(\mathbf{r}_\mathrm{p}-\mathbf{r}_\mathrm{q}+R_\mathrm{pq}\nabla_\mathrm{p}\Delta_\mathrm{s})(\nabla_\mathrm{p}\Delta_\mathrm{s}\mathbf{t})}{R_\mathrm{pq}\sqrt{\Delta_\mathrm{s}}}\exp(\mathrm{i}k\Delta_\mathrm{s})\sqrt{\frac{\mathrm{i}k^3}{4\pi}} + O(k).\end{aligned} \tag{109}$$

The ubiquitous primary wave was drawn to the left-hand sides to direct attention to the nontrivial terms on the right-hand sides.

We learn from these equations that diffraction of electromagnetic waves without polarization does not exist. Yet the contributions to the right-hand sides with the complementary error function describe just the vector properties of the primary wave. The changes of polarization effected by the screen are described by the contributions with the exponential function. The contributions with the error function are proportional to k^2 since $\omega = O(k)$, cf. equation (43), whereas those with the exponential function are proportional only to $k^{3/2}$. Nevertheless, one should not be deluded that polarization by diffraction is always a lower-order effect. In the shadow the error function decreases dramatically. It ceases to be $O(1)$ and weakens to be only $O(k^{-1/2})$ as indicated in equation (107). In the shadow both contributions to the right-hand sides of (108) and (109) reach comparable sizes. We must expect hitherto unseen effects.

All terms in (108) and (109) remain finite in the entire domain of solution except, of course, immediately at the edge where Bremmer's effect takes place. The singularities on the border of shadow caused by $\sqrt{\Delta_\mathrm{s}}$ in the denominators are cancelled by the derivatives $\nabla_\mathrm{p}\Delta_\mathrm{s}$ in the nominators since Δ_s varies quadratically in the vicinity of the border.

For radiation with low frequencies, e. g. microwaves, the electromagnetic fields can be measured directly. However, the Maxwell equations are real equations for real observables. Thus, we must extract from the fields (108) and (109) their real parts prior to comparison with experimental data. In optics with its much higher frequencies, the classic measurements are all calorimetric ones. One collects the energy flux impinging on a given surface for times much longer than inverses of these frequencies. The observable is here the time average of the Pointing vector $\mathbf{S}$

$$\begin{aligned}\bar{\mathbf{S}}(\mathbf{r}_\mathrm{p}) &= \lim_{\tau\to\infty}\frac{1}{\tau}\int_0^\tau \Re\mathbf{E}(\mathbf{r}_\mathrm{p},t)\times\Re\mathbf{B}(\mathbf{r}_\mathrm{p},t)\frac{\mathrm{d}t}{\mu} \\ &= \frac{1}{2\mu}\Re(\mathbf{E}_k(\mathbf{r}_\mathrm{p})\times\mathbf{B}_k^*(\mathbf{r}_\mathrm{p})).\end{aligned} \tag{110}$$

The relation between time-dependent and time-independent fields is as defined in equations (39). The asterisk denotes complex conjugation; one can affix it either to the magnetic or to the electric field without affecting the result. The a priori decomposition of the complex fields (108) and (109) is not needed here.

10. Diffraction by a Straight Edge

The first application is, of course, diffraction by a straight edge. For its description, the Cartesian coordinate system is best. Hence, $\xi = x$, $\eta = y$, and

$$-\infty < x < +\infty, \quad -\infty < y < y_- = \text{const}, \quad z = 0 \quad (111)$$

defines the screen. y_- is the location of the edge.

The exhausting dependence is found from equations (88) and (56)

$$\partial_x R_\mathrm{p}(x,y) = -\partial_x R_\mathrm{q}(x,y). \qquad (112)$$

Squaring this equation and cancelling identical terms on both sides results in a quadratic equation for x

$$(x_\mathrm{p}-x)^2((y_\mathrm{q}-y)^2+z_\mathrm{q}^2)) = (x_\mathrm{q}-x)^2((y_\mathrm{p}-y)^2+z_\mathrm{p}^2)). \qquad (113)$$

y must be considered as given. One solution of this equation is a mathematical ghost. It appears because the original equation (112) was squared to get rid of the roots. The other solution, however, does solve the original. It is the searched-for exhausting dependence $\xi = \Xi_\mathrm{s}(\eta)$ (88) in Cartesian coordinates:

$$\begin{aligned} x &= X_\mathrm{s}(y) \\ &= \frac{x_\mathrm{p}\sqrt{(y_\mathrm{q}-y)^2+z_\mathrm{q}^2} + x_\mathrm{q}\sqrt{(y_\mathrm{p}-y)^2+z_\mathrm{p}^2}}{\sqrt{(y_\mathrm{q}-y)^2+z_\mathrm{q}^2} + \sqrt{(y_\mathrm{p}-y)^2+z_\mathrm{p}^2}}. \end{aligned} \qquad (114)$$

Inserting this dependence in the triangle function (56) gives the exhausting function of the straight edge

$$\begin{aligned} &\Delta_\mathrm{s}(y) = \\ &\sqrt{(x_\mathrm{p}-x_\mathrm{q})^2 + (\sqrt{(y_\mathrm{p}-y)^2+z_\mathrm{p}^2} + \sqrt{(y_\mathrm{q}-y)^2+z_\mathrm{q}^2})^2} \\ &-\sqrt{(x_\mathrm{p}-x_\mathrm{q})^2+(y_\mathrm{p}-y_\mathrm{q})^2+(z_\mathrm{p}-z_\mathrm{q})^2}. \end{aligned} \qquad (115)$$

We take this function at the edge $y = y_-$ and need to determine the sign of $\sqrt{\Delta_\mathrm{s}} = \sqrt{\Delta_\mathrm{s}(y_-)}$, i. e. where there is shadow or light. Shadow prevails, according to the criterion (66), when $y_\mathrm{s} < y_-$. With y_s from equation (63) we find

$$y_\mathrm{p} < y_- + (y_- - y_\mathrm{q})\frac{z_\mathrm{p}}{-z_\mathrm{q}}. \qquad (116)$$

Hence, equation (65) reads here

$$\sqrt{\Delta_\mathrm{s}} = \begin{cases} +|\sqrt{\Delta_\mathrm{s}}| \text{ if } y_\mathrm{p} < y_- - (y_- - y_\mathrm{q})z_\mathrm{p}/z_\mathrm{q} \\ \qquad\quad \text{with } z_\mathrm{p} > 0, \\ -|\sqrt{\Delta_\mathrm{s}}| \text{ elsewhere.} \end{cases} \qquad (117)$$

What remains to be done is to insert this in the universal formula of diffraction (105).

11. Imaging by Diffraction

We have now enough experience to perform the simple trick that was announced in Section 9. To improve the fulfillment of boundary conditions, one inserts a mirrored picture of the source in the same way as Sommerfeld did when he calculated suitable Green functions, see Section 5. If $\mathbf{r}_\mathrm{q}$ is the position of the source, $\mathbf{r}_\mathrm{m}$ has the same coordinates x_q and y_q, but the opposite value of z_q as explained in (51). Let the distances R_pq and R_pm be defined according to (55):

$$\begin{aligned} R_\mathrm{pq} &= \sqrt{(x_\mathrm{p}-x_\mathrm{q})^2+(y_\mathrm{p}-y_\mathrm{q})^2+(z_\mathrm{p}-z_\mathrm{q})^2}, \\ R_\mathrm{pm} &= \sqrt{(x_\mathrm{p}-x_\mathrm{q})^2+(y_\mathrm{p}-y_\mathrm{q})^2+(z_\mathrm{p}+z_\mathrm{q})^2}. \end{aligned} \qquad (118)$$

Hence, original and mirrored distance become equal, $\lim R_\mathrm{pq} = \lim R_\mathrm{pm}$, if the probe approaches the screen, $z_\mathrm{p} \to 0$; it does not matter if the approach takes place in inner or outer space, $z_\mathrm{p} > 0$ or $z_\mathrm{p} < 0$.

However, we must be careful when we introduce the mirrored exhausting function. In the definition of the original exhausting function taken at the edge $y = y_-$

$$\begin{aligned} \Delta_\mathrm{sq} = \Big[&(x_\mathrm{p}-x_\mathrm{q})^2 + \Big(\sqrt{(y_\mathrm{p}-y_-)^2+z_\mathrm{p}^2} \\ &+ \sqrt{(y_\mathrm{q}-y_-)^2+z_\mathrm{q}^2}\Big)^2\Big]^{1/2} \\ &- \sqrt{(x_\mathrm{p}-x_\mathrm{q})^2+(y_\mathrm{p}-y_\mathrm{q})^2+(z_\mathrm{p}-z_\mathrm{q})^2}, \end{aligned} \qquad (119)$$

$$\sqrt{\Delta_\mathrm{sq}} = \begin{cases} +|\sqrt{\Delta_\mathrm{sq}}| \text{ if } y_\mathrm{p} < y_- - (y_- - y_\mathrm{q})z_\mathrm{p}/z_\mathrm{q} \\ \qquad\quad \text{with } z_\mathrm{p} > 0, \\ -|\sqrt{\Delta_\mathrm{sq}}| \text{ elsewhere.} \end{cases}$$

the assignment of signs of the root belongs to the definition, cf. equations (115) and (117). In about three quarters of space, the root of the original exhausting function is negative, viz. for all negative values of z_p and, if y_p is sufficiently large, for positive values of z_p, too. The source lights the major part of space.

By contrast, the mirrored exhausting function must be defined as

$$\begin{aligned} \Delta_\mathrm{sm} = \Big[&(x_\mathrm{p}-x_\mathrm{q})^2 + \Big(\sqrt{(y_\mathrm{p}-y_-)^2+z_\mathrm{p}^2} \\ &+ \sqrt{(y_\mathrm{q}-y_-)^2+z_\mathrm{q}^2}\Big)^2\Big]^{1/2} \\ &- \sqrt{(x_\mathrm{p}-x_\mathrm{q})^2+(y_\mathrm{p}-y_\mathrm{q})^2+(z_\mathrm{p}+z_\mathrm{q})^2}, \end{aligned} \qquad (120)$$

$$\sqrt{\Delta_{\mathrm{sm}}} = \begin{cases} -|\sqrt{\Delta_{\mathrm{sm}}}| \text{ if } y_{\mathrm{p}} < y_- + (y_- - y_{\mathrm{q}})z_{\mathrm{p}}/z_{\mathrm{q}} \\ \qquad \text{with } z_{\mathrm{p}} < 0, \\ +|\sqrt{\Delta_{\mathrm{sm}}}| \text{ elsewhere.} \end{cases}$$

Almost everything follows from (119) replacing z_{q} with $-z_{\mathrm{q}}$. The only exception is the assignment of signs of the root. Using the freedom mentioned in the discussion behind equation (65), the author reversed it. In about three quarters of space, the root of the mirrored exhausting function is positive, viz. for all positive values of z_{p} and, if y_{p} is sufficiently large, for negative values of z_{p}, too. Shadow prevails.

Yet immediately over and under the screen, the signs of original and mirrored functions are the same, $\lim \Delta_{\mathrm{sq}} = \lim \Delta_{\mathrm{sm}}$, if the probe approaches the screen, $z_{\mathrm{p}} \to 0$; it does not matter if the approach takes place in inner or outer space, $z_{\mathrm{p}} > 0$ or $z_{\mathrm{p}} < 0$.

With these functions it is straightforward to construct solutions that satisfy the boundary conditions (36) exactly. Instead of (105) we obtain

$$\begin{aligned} a_k(\mathbf{r}_{\mathrm{p}}) = {} & \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}} \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{sq}}}}{2} \\ & - \frac{\exp(\mathrm{i}kR_{\mathrm{pm}})}{R_{\mathrm{pm}}} \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{sm}}}}{2} + O(k^{-1}), \end{aligned} \tag{121}$$

$$\begin{aligned} b_k(\mathbf{r}_{\mathrm{p}}) = {} & \frac{\exp(\mathrm{i}kR_{\mathrm{pq}})}{R_{\mathrm{pq}}} \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{sq}}}}{2} \\ & + \frac{\exp(\mathrm{i}kR_{\mathrm{pm}})}{R_{\mathrm{pm}}} \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\Delta_{\mathrm{sm}}}}{2} + O(k^{-1}). \end{aligned} \tag{122}$$

The symbol $O(k^{-1})$ was retained to remember that the differential equations (41) are not exactly fulfilled.

The second terms on the right-hand sides of equations (121) and (122) considered isolated appear fantastic. They describe ghostly radiation from a source at $\mathbf{r}_{\mathrm{m}}$ that becomes bright only when it permeates the screen. Yet when they are considered in cooperation with the first terms, it is recognized that they describe the unavoidable reflection that is diffracted in a similar way as the primary wave. The solutions (121) and (122) are valid in entire space.

From these solutions it follows that radiation emitted at $\mathbf{r}_{\mathrm{q}}$ and diffracted by a screen causes a second singularity at $\mathbf{r}_{\mathrm{m}}$. In other words, diffraction creates a sharp image. Of course, the image is as weak as the multiplicative error function indicates, but it should be observable because the singularity sticks out.

Folks might feel this prediction as daring. They might not appreciate the immense power of the methods developed here. The approximate solution (121) and (122) contains, as a limiting case, the only exact solution of diffraction problems known so far, namely Sommerfeld's celebrated stringent solution [10].

Sommerfeld's solution deals with the diffraction of a plane wave. Using the definitions in (55) and (118), $R_{\mathrm{pq}} = |\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{q}}|$ and $R_{\mathrm{pm}} = |\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{m}}|$, the spherical waves in front of the error functions in (121) and (122) can be expanded as

$$\begin{aligned} & \frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{q}}|)}{|\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{q}}|} = \\ & \quad \frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{q}}|)}{|\mathbf{r}_{\mathrm{q}}|} \exp\left(\mathrm{i}k\frac{-\mathbf{r}_{\mathrm{q}}}{|\mathbf{r}_{\mathrm{q}}|}\mathbf{r}_{\mathrm{p}}\right) + O(|\mathbf{r}_{\mathrm{q}}|^{-2}), \end{aligned} \tag{123}$$

$$\begin{aligned} & \frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{m}}|)}{|\mathbf{r}_{\mathrm{p}} - \mathbf{r}_{\mathrm{m}}|} = \\ & \quad \frac{\exp(\mathrm{i}k|\mathbf{r}_{\mathrm{q}}|)}{|\mathbf{r}_{\mathrm{q}}|} \exp\left(\mathrm{i}k\frac{-\mathbf{r}_{\mathrm{m}}}{|\mathbf{r}_{\mathrm{q}}|}\mathbf{r}_{\mathrm{p}}\right) + O(|\mathbf{r}_{\mathrm{q}}|^{-2}). \end{aligned} \tag{124}$$

The first factors on the right-hand sides do not depend on the position of the probe $\mathbf{r}_{\mathrm{p}}$. They are absorbed in the normalizations of (121) and (122). When the source is moved to infinity, $|\mathbf{r}_{\mathrm{m}}| = |\mathbf{r}_{\mathrm{q}}| \to \infty$, only the plane waves remain. Their fronts are defined by the normal vectors

$$\mathbf{c}_{\mathrm{q}} = \lim_{|\mathbf{r}_{\mathrm{q}}| \to \infty} \frac{-\mathbf{r}_{\mathrm{q}}}{|\mathbf{r}_{\mathrm{q}}|}, \quad \mathbf{c}_{\mathrm{m}} = \lim_{|\mathbf{r}_{\mathrm{q}}| \to \infty} \frac{-\mathbf{r}_{\mathrm{m}}}{|\mathbf{r}_{\mathrm{q}}|} \tag{125}$$

with directional cosines as components

$$\begin{aligned} \mathbf{c}_{\mathrm{q}} &= \mathbf{e}_x c_x + \mathbf{e}_y c_y + \mathbf{e}_z c_z, \\ \mathbf{c}_{\mathrm{m}} &= \mathbf{e}_x c_x + \mathbf{e}_y c_y - \mathbf{e}_z c_z. \end{aligned} \tag{126}$$

Likewise the exhausting function for the plane-wave problem is defined as the limit $\delta_{\mathrm{s}}(y) = \lim \Delta_{\mathrm{s}}(y)$ for $|\mathbf{r}_{\mathrm{q}}| \to \infty$. The straightforward calculation transforms (115) to

$$\delta_{\mathrm{s}}(y) = \sqrt{c_y^2 + c_z^2}\sqrt{(y_{\mathrm{p}} - y)^2 + z_{\mathrm{p}}^2} - c_y(y_{\mathrm{p}} - y) - c_z z_{\mathrm{p}}. \tag{127}$$

Taking the value on the edge $y = y_-$ yields the analog of (119)

$$\begin{aligned} & \delta_{\mathrm{sq}} = \sqrt{c_y^2 + c_z^2}\sqrt{(y_{\mathrm{p}} - y_-)^2 + z_{\mathrm{p}}^2} - c_y(y_{\mathrm{p}} - y_-) - c_z z_{\mathrm{p}}, \\ & \sqrt{\delta_{\mathrm{sq}}} = \begin{cases} +|\sqrt{\delta_{\mathrm{sq}}}| \text{ if } y_{\mathrm{p}} < y_- + z_{\mathrm{p}} c_y / c_z \text{ with } z_{\mathrm{p}} > 0, \\ -|\sqrt{\delta_{\mathrm{sq}}}| \text{ elsewhere.} \end{cases} \end{aligned} \tag{128}$$

and the mirrored exhausting function as analog of (120)

$$\delta_{\mathrm{sm}} = \sqrt{c_y^2+c_z^2}\sqrt{(y_{\mathrm{p}}-y_-)^2+z_{\mathrm{p}}^2} - c_y(y_{\mathrm{p}}-y_-) + c_z z_{\mathrm{p}},$$
$$\sqrt{\delta_{\mathrm{sm}}} = \begin{cases} -|\sqrt{\delta_{\mathrm{sm}}}| & \text{if } y_{\mathrm{p}} < y_- - z_{\mathrm{p}}c_y/c_z \\ & \text{with } z_{\mathrm{p}} < 0, \\ +|\sqrt{\delta_{\mathrm{sm}}}| & \text{elsewhere.} \end{cases} \tag{129}$$

Thus the analogs of (121) and (122) are

$$a_k(\mathbf{r}_{\mathrm{p}}) = \exp(\mathrm{i}k\mathbf{c}_{\mathrm{q}}\mathbf{r}_{\mathrm{p}})\frac{\mathrm{erfc}\sqrt{-\mathrm{i}k\delta_{\mathrm{sq}}}}{2} - \exp(\mathrm{i}k\mathbf{c}_{\mathrm{m}}\mathbf{r}_{\mathrm{p}})\frac{\mathrm{erfc}\sqrt{-\mathrm{i}k\delta_{\mathrm{sm}}}}{2}, \tag{130}$$

$$b_k(\mathbf{r}_{\mathrm{p}}) = \exp(\mathrm{i}k\mathbf{c}_{\mathrm{q}}\mathbf{r}_{\mathrm{p}})\frac{\mathrm{erfc}\sqrt{-\mathrm{i}k\delta_{\mathrm{sq}}}}{2} + \exp(\mathrm{i}k\mathbf{c}_{\mathrm{m}}\mathbf{r}_{\mathrm{p}})\frac{\mathrm{erfc}\sqrt{-\mathrm{i}k\delta_{\mathrm{sm}}}}{2}. \tag{131}$$

These functions form an *exact description of diffraction*. Not only the boundary conditions are satisfied perfectly, also the Helmholtz equations (41) as direct checking shows. One may be surprised at this, but it becomes comprehensible when it is noticed that in this problem with the infinite edge and the primary plane wave, the only length is provided by the position of the probe $\mathbf{r}_{\mathrm{p}}$. One can introduce new variables of position instead of $\mathbf{r}_{\mathrm{p}}$ by scaling the latter with with the inverse of the wave number k. Hence, if the solution is known for one wave number, here for $k \longrightarrow \infty$, it is known for all wave numbers.

Based on this argument, one supplements the proof of the fact that the two integrals in (53) and (54) are equal to $\mp 2\pi/z_{\mathrm{p}}R_{\mathrm{pq}}$, respectively. There is no edge at all. The only length is the distance between source and probe. One evaluates the integrals for $k \longrightarrow \infty$ using stationary phase according to Section 8 and introduces thereafter scaled variables to show that the result does not depend on k.

Sommerfeld's solution is a special case of (130) and (131) for flat incidence, i. e. $c_x = 0$ thus $\sqrt{c_y^2+c_z^2} = 1$. In this case, $b_k(\mathbf{r}_{\mathrm{p}})$ is proportional to the magnetic field $\mathbf{B}_k(\mathbf{r}_{\mathrm{p}}) = -\mathbf{e}_x k^2 b_k(\mathbf{r}_{\mathrm{p}})$. The factor $-k^2$ is swallowed by normalization. In the problem with the other polarization, the electric field $\mathbf{E}_k(\mathbf{r}_{\mathrm{p}})$ is proportional to $\mathbf{e}_x a_k(\mathbf{r}_{\mathrm{p}})$. Therefore, Sommerfeld got along without the representation theorem of Section 2. Moreover, Sommerfeld neither knew the method of stationary phase for two-dimensional integrals nor did he use the standardized complementary error function although it does not take much work to show that Sommerfeld's function is equivalent. Sommerfeld found his function via an ingenious contour integration, an approach which does not seem to admit generalization. Finally, it played a role that Sommerfeld was Felix Klein's pupil. Klein was the most influential advocate of *uniformization* meaning that roots had to be avoided because of their ambiguities and to be replaced with suitable parametrizations necessitating transcendental functions. So Sommerfeld introduced, instead of the roots in (128) and (129), artifical angles, sines and cosines. Probably this *rootophobia* is the reason why Sommerfelds exceedingly important solution – it is the 'harmonic oscillator' of diffraction – was never sufficiently appreciated and is not presented in most modern textbooks.

Though Sommerfeld's solution holds only for flat incidence $c_x = 0$, it is easily generalized for skew incidence due to translational invariance. This was done long before this article was drafted [28, § 11.6]. Moreover, the solutions (130) and (131) can be conceived as Fourier components and used to build the most general diffraction at the straight edge. For example, diffraction of a spherical wave, which is described by equations (121) and (122) only asymptotically, was calculated by Macdonald [32]. Yet Macdonald's formulae are so difficult to survey that imaging by diffraction might have escaped notice. The present approach is simpler and generalizable.

12. Diffraction by a Slit

Let us clarify next one of the most lugubrious offences of traditional theory: the difference between Fresnel and Fraunhofer diffraction. To anticipate the answer, Fraunhofer diffraction is a misnamer. In the proper sense of *diffraction*, Fraunhofer diffraction does not exist. There is at best *interference*. It occurs only when a primary wave strikes an aperture with at least two opposite edges. The thus excited secondary waves interfere to generate a pattern now named Fraunhofer diffraction. The formulae derived in this article are powerful enough to describe the gradual transition from Fresnel diffraction to Fraunhofer interference. All is included in the equations (103) and (104).

To understand how the transition proceeds, the simplest example suffices: (i) Damping negligible; imaginary part $\Im k \to 0$; the wave number k is a positive number. (ii) Diffraction by a straight slit with edges at $y_\pm = \pm y_0$; $2y_0$ is a positive number meaning the width of the slit. (iii) Normal incidence of a plane wave; thus $c_x = c_y = 0$, $c_z = 1$, see (125)–(127); in equation (104), the plane wave replaces the spherical wave; consequently the exhausting function Δ_s gives way to its descendant δ_s (127)

$$b_k(\mathbf{r}_\mathrm{p}) = \exp(\mathrm{i}kz_\mathrm{p}) \\ \cdot \frac{\operatorname{erfc}\sqrt{-\mathrm{i}k\delta_\mathrm{s}(y_-)} - \operatorname{erfc}\sqrt{-\mathrm{i}k\delta_\mathrm{s}(y_+)}}{2} + O(k^{-1}) \tag{132}$$

with

$$\begin{aligned} \delta_\mathrm{s}(y_\pm) &= \sqrt{(y_\mathrm{p} \mp y_0)^2 + z_\mathrm{p}^2} - z_\mathrm{p}, \\ \sqrt{\delta_\mathrm{s}(y_\pm)} &= \begin{cases} +|\sqrt{\delta_\mathrm{s}(y_\pm)}| \text{ if } y_\mathrm{p} < \pm y_0, \\ -|\sqrt{\delta_\mathrm{s}(y_\pm)}| \text{ elsewhere.} \end{cases} \end{aligned} \tag{133}$$

(iv) The representative $a_k(\mathbf{r}_\mathrm{p})$ is disregarded. (v) Effects of the reflected wave are neglected; $z_\mathrm{p} > 0$ is implied.

Equations (132) and (133) are still powerfull as the probe can be moved freely within the inner space $z_\mathrm{p} > 0$. Especially the equations remain valid when the probe approaches the screen $z_\mathrm{p} \to 0$. However, in most classical experiments the probe is far away from the screen, $z_\mathrm{p} \to \infty$. Therefore, the exhausting function (133) can be expanded in terms of inverse powers of z_p

$$\delta_\mathrm{s}(y_\pm) = \frac{y_\mathrm{p}^2 \mp 2y_\mathrm{p}y_0 + y_0^2}{2z_\mathrm{p}} \left[1 + O\left(\left(\frac{y_\mathrm{p} \mp y_0}{z_\mathrm{p}}\right)^2\right)\right]. \tag{134}$$

It is worthwhile to mention that using the approximated exhausting function (96) yields the same result. The linearization (94) and (95) spoils the approach to the screen. In the error functions of (132), which oscillate quickly, the first two terms on the right-hand side of (134) must be retained because only the second term will yield specific results. But in algebraic expressions, the leading term will be enough, for example

$$\frac{1}{|\sqrt{\delta_\mathrm{s}(\pm y_0)}|} = \frac{\sqrt{2z_\mathrm{p}}}{|y_\mathrm{p}|} \left[1 + O\left(\frac{y_0}{y_\mathrm{p}}\right) + O\left(\left(\frac{y_\mathrm{p}}{z_\mathrm{p}}\right)^2\right)\right]. \tag{135}$$

To pacify simple-minded scientists' dismay at the error functions in (132), let us return to elementary functions using the asymptotic expansion (74). Because of (133) we must do it differently for $y_\mathrm{p} \to -\infty$ and $y_\mathrm{p} \to +\infty$. For the latter case, equation (75) has to be considered before (74) can be applied

$$\operatorname{erfc}\sqrt{-\mathrm{i}k\delta_\mathrm{s}(\pm y_0)} \sim 2 - \frac{\exp(\mathrm{i}k\delta(\pm y_0))}{\sqrt{-\mathrm{i}\pi k}|\sqrt{\delta(\pm y_0)}|} \\ \text{for } y_\mathrm{p} \to +\infty. \tag{136}$$

The minus signs from the last line of (133) and of the argument in (75) cancel. Opticians interprete this formula as representing the primary wave by the 2 and a secondary wave radiated from the edge by the exponential function. Therefore, it is a formula for the domain of light. This seems to be a contradiction to physics since the domain of light is limited to $|y_\mathrm{p}| < y_0$ while we suppose here $y_\mathrm{p} \to +\infty$. The contradiction is quashed since the two leading 2's cancel in (132).

Calculation based on (134) to (136) yields

$$b_k(\mathbf{r}_\mathrm{p}) \sim \\ \sqrt{-\mathrm{i}} \exp\left[\mathrm{i}k\left(z_\mathrm{p} + \frac{y_\mathrm{p}^2}{2z_\mathrm{p}}\right)\right] \sqrt{\frac{2ky_0^2}{\pi z_\mathrm{p}}} \frac{\sin(ky_0y_\mathrm{p}/z_\mathrm{p})}{ky_0y_\mathrm{p}/z_\mathrm{p}}. \tag{137}$$

Notice the sine is brought about by the interference of the two terms in (132). The formula is symmetric with respect to y_p. It holds both for positive and negative values. A simple calculation for $y_\mathrm{p} \to -\infty$ along the lines just sketched confirms this.

For a comparison with experiments, we need the time-averaged Pointing vector (110). The probe usually is an absorbing layer spanned perpendicularly to the energy flux of the primary wave. The representative in (137) produces via (59) and (60) with $\mathbf{t} = \mathbf{e}_x$ the electromagnetic fields

$$\begin{aligned} \mathbf{B}_k(\mathbf{r}_\mathrm{p}) &= -\mathbf{e}_x k^2 b_k(\mathbf{r}_\mathrm{p}) \\ \mathbf{E}_k(\mathbf{r}_\mathrm{p}) &= \mathbf{e}_y \omega k b_k(\mathbf{r}_\mathrm{p})(1 + O(z_\mathrm{p}^{-1})) + \mathbf{e}_z \ldots . \end{aligned} \tag{138}$$

The dots behind $\mathbf{e}_z$ substitute a function we presently do not need to know. The measured component of the energy flux is thus

$$\mathbf{e}_z\bar{\mathbf{S}}(\mathbf{r}_\mathrm{p}) \sim \frac{\omega k^3}{2\mu}|b_k(\mathbf{r}_\mathrm{p})|^2. \tag{139}$$

$\bar{\mathbf{S}}_0 = \mathbf{e}_z\omega k^3/(2\mu)$ is, because of the normalization chosen in (132), the energy flux of the primary wave. Straightforward evaluation of (139) with (137) gives

$$\mathbf{e}_z\bar{\mathbf{S}}(\mathbf{r}_\mathrm{p}) \sim |\bar{\mathbf{S}}_0| 2y_0 \frac{ky_0}{\pi z_\mathrm{p}} \left(\frac{\sin(ky_0y_\mathrm{p}/z_\mathrm{p})}{ky_0y_\mathrm{p}/z_\mathrm{p}}\right)^2. \tag{140}$$

This is the most fabulous formula of traditional diffraction theory – Kirchhoff's formula of Fraunhofer *diffraction*. However, Kirchhoff found only the factor with the squared sine. Historically, various normalizing factors were adjusted a posteriori, but the fabulous formula was never perceived as an equation among vectors.

Here, by contrast, everything arises from first principles. The last two factors in (140) when integrated over y_p from $-\infty$ to $+\infty$ yield 1. The first two factors describe the primary energy flux squeezed through the slit of width $2y_0$. Equation (140) expresses energy conservation as warranted by Maxwell's equations in a medium without damping.

The fabulous formula is worse than admitted in textbooks. From the estimate in (135) we see it holds only for

$$y_0 \ll |y_\mathrm{p}| \ll z_\mathrm{p}. \tag{141}$$

$y_0 \ll z_\mathrm{p}$ means that the probe is far away from the screen. This is all right because $z_\mathrm{p} \to \infty$ was announced at the start of the calculation. But the extendend condition (141) entails that Kirchhoff's formula provides neither a valid description of the central peak around $|y_\mathrm{p}| \le y_0$ nor of large-angle diffraction $|y_\mathrm{p}| > z_\mathrm{p}$. Only some intermediate wiggles are correctly seized. That the condition $y_0 \ll |y_\mathrm{p}|$ matters, is corroborated through the use of the asymptotic expansion (136). It is utterly wrong at values of the exhausting function close to zero. The triangle function and thus the exhausting function of optics is, according to (64), zero on the border of shadow defined here by $y_0 = |y_\mathrm{p}|$.

The positive outcome of the preceding discussion is better understanding of the working of formula (132) and its prototypes (103) and (104). Let us position the probe first close to an edge. To describe this measurement, one of the terms in (132) is enough. However, we must utilize the error function full-fledged. When the probe is removed from the edge, but kept close to the screen, we are still in the realm of pure diffraction. Still one term of (132) suffices, but we may use now its asymptotic expansion (74). Only in that part of space where distances to the edges are about equal, interferences show up. Both terms in (132) must be kept. When the distances become large, we may use asymptotic expansions, but not at the borders of shadow. With increasing distances from the edges, the diffracted beams spread. In particular this is true close to the borders. The central beam is enclosed between these borders. Therefore, it is always questionable to decribe diffraction and interference in the vicinity of the central beam using the asymptotic expansion (74).

13. Diffraction by a Circular Aperture

After it is confirmed that the new theory encloses correct parts of the traditional theory as special cases, we are free to step into the entirely unknown.

The round stop is the most often installed component in optical instruments. We want to calculate the diffraction it causes, more precisely its Fresnel diffraction, since some information on Fraunhofer interference is already known – Airy's formula [28, § 8.5.2]. The edge of a circular aperture is not straight. Nevertheless, the universal formula of diffraction (105) applies again. All we have to do is to find the suited exhausting function.

The suitable coordinate system is the cylindrical one introduced in Section 6, see (67).

$$-\pi < \varphi \le +\pi, \quad 0 \le \rho < \rho_0 = \mathrm{const}, \quad z = 0 \tag{142}$$

defines a circular aperture in the screen. The exhausting dependence is found from equations (88) and (56)

$$\partial_\varphi R_\mathrm{p}(\rho,\varphi) = -\partial_\varphi R_\mathrm{q}(\rho,\varphi) \tag{143}$$

with

$$\begin{aligned} R_\mathrm{p}(\rho,\varphi) &= \sqrt{\rho_\mathrm{p}^2 + \rho^2 - 2\rho_\mathrm{p}\rho\cos(\varphi_\mathrm{p}-\varphi) + z_\mathrm{p}^2}, \\ R_\mathrm{q}(\rho,\varphi) &= \sqrt{\rho_\mathrm{q}^2 + \rho^2 - 2\rho_\mathrm{q}\rho\cos(\varphi_\mathrm{q}-\varphi) + z_\mathrm{q}^2}. \end{aligned} \tag{144}$$

For normal incidence, $\rho_\mathrm{q} = 0$, one can observe the solutions using the rubber-ribbon experiment explained in Section 8. Actually, there are two:

$$\varphi = \varphi_\mathrm{p} \ \text{ and } \ \varphi = \varphi_\mathrm{p} + \pi. \tag{145}$$

These exhausting dependencies $\varphi = \varPhi_\mathrm{s}(\rho)$, analogues of $\xi = \varXi_\mathrm{s}(\eta)$ in (88), surprise as they do not depend on ρ. Inserting them into the triangle function (56) gives the exhausting function of the circular stop. We need it at the edge $\rho = \rho_0$:

$$\begin{aligned} \Delta_\mathrm{s} = {} & \sqrt{(\rho_\mathrm{p}-\rho_0)^2 + z_\mathrm{p}^2} + \sqrt{\rho_0^2 + z_\mathrm{q}^2} \\ & - \sqrt{\rho_\mathrm{p}^2 + (z_\mathrm{p}-z_\mathrm{q})^2}. \end{aligned} \tag{146}$$

The definition has to be completed with the disposal of the root

$$\sqrt{\Delta_\mathrm{s}} = \begin{cases} +|\sqrt{\Delta_\mathrm{s}}| \text{ if } \rho_\mathrm{p} > \rho_0(z_\mathrm{q} - z_\mathrm{p})/z_\mathrm{q} \\ \qquad\text{with } z_\mathrm{p} > 0, \\ -|\sqrt{\Delta_\mathrm{s}}| \text{ elsewhere,} \end{cases} \tag{147}$$

according to equations (65) and (68). The exhausting function of the second solution in (145) is derived from this replacing ρ_0 with $-\rho_0$. The author names the first solution with the positive ρ_0 *near* because its point on the edge is closer to the probe than in the opposite case. The second solution with the negative ρ_0 is thus called *far*.

For mathematical reasons, only the near solution may dominate. The definition (142) of cylindrical coordinates restricts the range of the angle φ. Yet unrestricted variation of the variable ξ, analog of φ, is a premise of the principle of utter exhaust as declared in Section 8. The violation of this premise is not disastrous as long as the exhausting function $\Delta(\xi,\eta)$ takes a sharp extremum on the edge. In this case, the finiteness of integration causes negligible corrections. But for the round stop $\Delta(\xi,\eta)$ becomes equal for all points on the edge when source and probe both approach the optical axis; there is no extremum of $\Delta(\xi,\eta)$ at all. Consequently the universal formula of diffraction (105) applied to circular apertures needs $\rho_\mathrm{p} > z_\mathrm{p} > 0$ to secure its applicability.

Of course, the primary wave is also diffracted at the far side of the circular stop. There will be similar interferences as those caused by the slit, see Section 12, but the contributions from the far side are here, in the range of validity just declared, small. Fraunhofer interferences cannot be described well. Nevertheless, within the restricted range, there seems to be no rival of the formulas (146) and (147). We can calculate diffraction immediately behind a circular aperture and predict what a curved edge does to polarization.

For general parameters, the exhausting equation (143) cannot be solved in terms of roots. When the transcendental functions sine and cosine are rationalized introducing, for instance, the variables c and s

$$c = \cos\left(\varphi - \frac{\varphi_\mathrm{p} + \varphi_\mathrm{q}}{2}\right),\ s = \sin\left(\varphi - \frac{\varphi_\mathrm{p} + \varphi_\mathrm{q}}{2}\right), \tag{148}$$

where $c^2 + s^2 = 1$, and when the roots are removed by squaring, an algebraic equation of the sixth degree is obtained.

Nevertheless we can find more physically relevant solutions, even for oblique incidence, performing the rubber-ribbon experiment as above. Namely the *one-sided flat situation*, where source and probe sit on the same side of the optical axis, and the *two-sided flat situation*. In both situations, there are near and far solutions:

$$\begin{aligned} &\text{one-sided flat: } \varphi_\mathrm{p} = \varphi_\mathrm{q} \\ &\quad\text{near: } \varphi = \varphi_\mathrm{p} \text{ far: } \varphi = \varphi_\mathrm{p} + \pi, \end{aligned} \tag{149}$$

$$\begin{aligned} &\text{two-sided flat: } \varphi_\mathrm{p} = \varphi_\mathrm{q} + \pi \\ &\quad\text{near: } \varphi = \varphi_\mathrm{p} \text{ far: } \varphi = \varphi_\mathrm{p} + \pi. \end{aligned} \tag{150}$$

But this is not enough when vector fields as in (108) and (109) are to be calculated. For them we must differentiate the exhausting function with respect to z_p, ρ_p, and φ_p. The last differentiation is not feasible when the function is only known for a fixed φ_p.

Mending the shortcoming is easy. The sine S and the cosine C were defined in (69). In the context of (143), they must be considered as given and fixed. The sine s and the cosine c were defined in (148). They contain the angle which is sought as the exhausting dependence $\varphi = \Phi_\mathrm{s}(\rho)$. With the abbreviations

$$\begin{aligned} P(\rho) &= \sqrt{(\rho_\mathrm{p} - \rho)^2 + z_\mathrm{p}^2}/\rho_\mathrm{p}, \\ Q(\rho) &= \sqrt{(\rho_\mathrm{q} - \rho)^2 + z_\mathrm{q}^2}/\rho_\mathrm{q} \end{aligned} \tag{151}$$

the exhausting equation (143) can be written in four equivalent variants

$$\begin{aligned} &\frac{Cs - Sc}{\sqrt{P(\pm\rho)^2 - 2\rho(Cc + Ss \mp 1)/\rho_\mathrm{p}}} = \\ &-\frac{Cs + Sc}{\sqrt{Q(\pm\rho)^2 - 2\rho(Cc - Ss \mp 1)/\rho_\mathrm{q}}}. \end{aligned} \tag{152}$$

Differently to all other equations in this article, signs can be chosen independently. Just under the roots, the signs must be altered coherently.

These variants permit to find solutions in the neighbourhood of (149) and (150) such that they are exact in linear terms. From (69) we have

$$\begin{aligned} \text{one-sided: } S &= O(\varphi_\mathrm{p} - \varphi_\mathrm{q}), \\ C &= 1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q})^2), \\ \text{two-sided: } S &= 1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi)^2), \\ C &= O(\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi). \end{aligned}$$

Inserting the estimates according to (149) and (150):

$$\begin{aligned}
&\text{one-sided near: } \varphi = \varphi_\mathrm{p} + O(\varphi_\mathrm{p} - \varphi_\mathrm{q}),\\
&\text{one-sided far: } \varphi = \varphi_\mathrm{p} + \pi + O(\varphi_\mathrm{p} - \varphi_\mathrm{q}),\\
&\text{two-sided near: } \varphi = \varphi_\mathrm{p} + O(\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi),\\
&\text{two-sided far: } \varphi = \varphi_\mathrm{p} + \pi + O(\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi),
\end{aligned}$$

in the definitions (148) yields:

$$\begin{aligned}
&\text{one-sided near: } s = O(\varphi_\mathrm{p} - \varphi_\mathrm{q}),\\
&\qquad\qquad c = 1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q})^2),\\
&\text{one-sided far: } s = O(\varphi_\mathrm{p} - \varphi_\mathrm{q}),\\
&\qquad\qquad c = -1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q})^2),\\
&\text{two-sided near: } s = 1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi)^2),\\
&\qquad\qquad c = O(\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi),\\
&\text{two-sided far: } s = -1 + O((\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi)^2),\\
&\qquad\qquad c = O(\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi).
\end{aligned}$$

Hence, the nominators in (152) are always small of first order. In the denominators, $P(\pm\rho)$ and $Q(\pm\rho)$ are positive constants. But the expressions $Cc + Ss \mp 1$ and $Cc - Ss \mp 1$ are always small of second order if signs are suitably selected. Thus these expressions can be neglected. Without extertion one finds now the exhausting dependencies for the one-sided flat situation

$$\begin{aligned}
&\text{near: } s = -S\frac{P(\rho) - Q(\rho)}{P(\rho) + Q(\rho)},\\
&\text{far: } s = S\frac{P(-\rho) - Q(-\rho)}{P(-\rho) + Q(-\rho)}
\end{aligned} \tag{153}$$

and for the two-sided flat situation

$$\begin{aligned}
&\text{near: } c = -C\frac{P(\rho) + Q(-\rho)}{P(\rho) - Q(-\rho)},\\
&\text{far: } c = C\frac{P(-\rho) + Q(\rho)}{P(-\rho) - Q(\rho)}.
\end{aligned} \tag{154}$$

The exhausting function of the one-sided flat situation, based on the near solution and taken at the edge $\rho = \rho_0$, is, when abbreviations are restored,

$$\begin{aligned}
\Delta_\mathrm{s} = {} & \sqrt{(\rho_\mathrm{p} - \rho_0)^2 + z_\mathrm{p}^2} + \sqrt{(\rho_\mathrm{q} - \rho_0)^2 + z_\mathrm{q}^2}\\
& - \sqrt{(\rho_\mathrm{p} - \rho_\mathrm{q})^2 + (z_\mathrm{p} - z_\mathrm{q})^2}\\
& + 2\left[\left(\frac{\sqrt{(\rho_\mathrm{p} - \rho_0)^2 + z_\mathrm{p}^2}}{\rho_\mathrm{p}\rho_0} + \frac{\sqrt{(\rho_\mathrm{q} - \rho_0)^2 + z_\mathrm{q}^2}}{\rho_\mathrm{q}\rho_0}\right)^{-1}\right.\\
& \left. - \left(\frac{\sqrt{(\rho_\mathrm{p} - \rho_\mathrm{q})^2 + (z_\mathrm{p} - z_\mathrm{p})^2}}{\rho_\mathrm{p}\rho_\mathrm{q}}\right)^{-1}\right]\sin^2\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2},
\end{aligned} \tag{155}$$

$$\begin{aligned}
&\sqrt{\Delta_\mathrm{s}} = +|\sqrt{\Delta_\mathrm{s}}| \text{ if } \rho_\mathrm{p} >\\
&\frac{\rho_0(z_\mathrm{q} - z_\mathrm{p}) + \rho_\mathrm{q}z_\mathrm{p}}{z_\mathrm{q}}\left[1 + 2\frac{\rho_\mathrm{q}z_\mathrm{p}}{\rho_0(z_\mathrm{p} - z_\mathrm{q})}\sin^2\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2}\right]\\
&\text{with } z_\mathrm{p} > 0,\\
&\sqrt{\Delta_\mathrm{s}} = -|\sqrt{\Delta_\mathrm{s}}| \text{ elsewhere.}
\end{aligned} \tag{156}$$

The equations (68) and (69) were used to determine the border of shadow.

The respective expressions for the two-sided flat situation are

$$\begin{aligned}
\Delta_\mathrm{s} = {} & \sqrt{(\rho_\mathrm{p} - \rho_0)^2 + z_\mathrm{p}^2} + \sqrt{(\rho_\mathrm{q} + \rho_0)^2 + z_\mathrm{q}^2}\\
& - \sqrt{(\rho_\mathrm{p} + \rho_\mathrm{q})^2 + (z_\mathrm{p} - z_\mathrm{q})^2}\\
& + 2\left[\left(\frac{\sqrt{(\rho_\mathrm{p} - \rho_0)^2 + z_\mathrm{p}^2}}{\rho_\mathrm{p}\rho_0} - \frac{\sqrt{(\rho_\mathrm{q} + \rho_0)^2 + z_\mathrm{q}^2}}{\rho_\mathrm{q}\rho_0}\right)^{-1}\right.\\
& \left. + \left(\frac{\sqrt{(\rho_\mathrm{p} + \rho_\mathrm{q})^2 + (z_\mathrm{p} - z_\mathrm{p})^2}}{\rho_\mathrm{p}\rho_\mathrm{q}}\right)^{-1}\right]\cos^2\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2},
\end{aligned} \tag{157}$$

$$\begin{aligned}
&\sqrt{\Delta_\mathrm{s}} = +|\sqrt{\Delta_\mathrm{s}}| \text{ if } \rho_\mathrm{p} >\\
&\frac{\rho_0(z_\mathrm{q} - z_\mathrm{p}) - \rho_\mathrm{q}z_\mathrm{p}}{z_\mathrm{q}}\left[1 - 2\frac{\rho_\mathrm{q}z_\mathrm{p}}{\rho_0(z_\mathrm{p} - z_\mathrm{q})}\cos^2\frac{\varphi_\mathrm{p} - \varphi_\mathrm{q}}{2}\right]\\
&\text{with } z_\mathrm{p} > 0,\\
&\sqrt{\Delta_\mathrm{s}} = -|\sqrt{\Delta_\mathrm{s}}| \text{ elsewhere.}
\end{aligned} \tag{158}$$

All exhausting functions based on the far solutions can be obtained from these formulas inverting the sign of ρ_0. The errors are $O((\varphi_\mathrm{p} - \varphi_\mathrm{q})^4)$ or $O((\varphi_\mathrm{p} - \varphi_\mathrm{q} - \pi)^4)$, respectively.

Generally, one may solve equation (152) numerically. The effort for this is orders of magnitudes less than for a direct numerical solution of Maxwell's equations.

14. Outlook

What might be novel in this article?

The representation theorem of electrodynamics was never stated before in such generality as here, see Section 2. It was never anticipated that the boundary conditions on metallic surfaces are as simple for the representatives as deduced in Section 3. These discoveries are synthesized with Sommerfeld's criticisms and proposals in the Sections 4 and 5. The thus derived solution of electromagnetic diffraction, polarization inclusive, described in Section 5, is a new finding. Compared to this, the discussion of the triangle function in Section 6 appears at first sight of minor importance. But the root of the triangle function and a consistent assignment of its signs are of prime importance for applications.

However, the big thing is the principle of utter exhaust declared in Section 8. Utter exhaust shows its power in the universal formula of diffraction, see Section 9, which is new. Its application to diffraction by an edge in Section 10 yields a comprehensive formula which encloses, as a limiting case, Sommerfeld's stringent solution, see Section 11. The great wonder is the perfect agreement though utter exhaust is just an asymptotic evaluation for short wavelengths.

For diffraction by a slit, as discussed in Section 12, utter exhaust copes with Fresnel diffraction and Fraunhofer interference in a unified manner. The gradual transition between these fundamentally different patterns can be calculated now for the first time.

The formulas in Section 13 describing diffraction by a round stop, derived via utter exhaust, are new.

At last we have a straightforward theory of electrodynamic diffraction derived from Maxwell's equations and resulting in observables.

Also part of acoustics will change. The methods developed from Section 3 on apply to the velocity field of sound if the representatives $a_k(\mathbf{r}_\mathrm{p})$ and $b_k(\mathbf{r}_\mathrm{p})$ are construed as scalar potentials. The boundary conditions derived in Section 3 are just the most often used ones in acoustics, namely on the soft or the hard wall, i. e. acoustic impedance zero or infinite. Almost everything remains the same, in particular the universal formula of diffraction (105). Only the dispersion relation of sound is different from (42) since compressible fluids refuse the telegraph equation [14].

The physical discussion of the formulas derived here has only begun. Important systems, e. g. diffraction by a hook or a cusp, were not yet analyzed though they are now accessible. The circular aperture is an especially rich system which deserves further studies. The foundation of its theory is the integral (58). Here it was evaluated for short wavelengths only, but more asymptotics must be considered. Evaluation for large distances from the screen will yield Fourier integrals and thus comfort all those who understand imaging as a sequence of Fourier and inverse Fourier transforms. The asymptotics for small angles are scarcely understood though they are most important for applications in optical systems.

Quantitative comparisons with experiments are surprisingly rare. The best measurements are probably those with microwaves [33].

[1] G. Kirchhoff, Berlin. Ber. **XX**, 641 (1882).

[2] G. Kirchhoff, Ann. Phys. **18**, 663 (1883).

[3] G. Kirchhoff, Mathematische Optik, Teubner, Leipzig 1891.

[4] E. Hecht, Optics, Addison Wesley, Upper Saddle River 2002.

[5] D. Meschede, Optics, Light and Lasers, Wiley VCH, Weinheim 2004.

[6] K.K. Sharma, Optics, Academic Press, Amsterdam 2006.

[7] J. A. Kong, Electromagnetic Wave Theory, EMW Publishing, Cambrigde Mass. 2008.

[8] K. Hönl, A. W. Maue, and K. Westphal, Theorie der Beugung in Handbuch der Physik Bd. XXV/1 (ed. S. Flügge), Springer, Berlin 1961.

[9] G. Mie, Ann. Physik **25**, 377 (1908).

[10] A. Sommerfeld, Math. Ann. **47**, 317 (1896).

[11] A. Sommerfeld, Optik, Geest & Portig, Leipzig 1964.

[12] A. Sommerfeld, Optics, Academic Press, New York 1964.

[13] U. Brosa, Zur Lösung von Randwertproblemen mit Vektorfeldern, Habilitationsschrift, Universität Marburg 1985.

[14] U. Brosa, Z. Naturforsch. **41a**, 1141 (1986).

[15] S. Grossmann, Mathematischer Einführungskurs für die Physik, 7th edition, B. G. Teubner, Stuttgart 1993.

[16] J. D. Jackson, Classical Electrodynamics, 2nd edition, John Wiley & Sons, New York 1975.

[17] M. von Laue, Röntgenstrahl-Interferenzen, 2. Kapitel, Geest & Portig, Leipzig 1964.

[18] P. Debye, Ann. Physik **30**, 57, 1909.

[19] J. Meixner and F. W. Schäfke, Mathieusche Funktionen und Sphäroidfunktionen, Springer, Berlin 1954, pp. 357 – 358.

[20] L. A. Stratton and L. J. Chu, Phys. Rev. **56**, 99 (1939).

[21] W. R. Smythe, Phys. Rev. **72**, 1066 (1947).

[22] W. R. Smythe, Static and Dynamic Electricity, Benjamin, New York 1969.

[23] O. Theimer, G. D. Wassermann and E. Wolf, Proc. Roy. Soc. **A212**, 426 (1952).

[24] L. Boberg and U. Brosa, Z. Naturforsch. **43a**, 697 (1988).

[25] U. Brosa, J. Stat. Phys. **55**, 1303 (1989).

[26] B. Hof, J. Westweel, T. M. Schneider, and B. Eckardt, Nature **443**, 59 (2006).

[27] A. Erdélyi, Asymptotic Expansions, Dover, New York 1956.

[28] M. Born and E. Wolf, Principles of Optics, Cambridge University Press, Cambridge 1997.

[29] M. Abramowitz and I. Stegun, Handbook of Mathematical Functions, Dover, New York 1970.

[30] E. Jahnke, F. Emde, and F. Lösch, Tables of Higher Functions, Teubner, Stuttgart 1966.

[31] A. Fresnel, Œuvres complètes, Imprimerie impériale, Paris 1866.

[32] H. M. Macdonald, Proc. Lond. Math. Soc. **14**, 410 (1915).

[33] R. W. P. King and T. T. Wu, The Scattering and Diffraction of Waves, Havard University Press, Cambridge Mass. 1959.