Deterministic phase retrieval: a Green's function solution

Michael Reed Teague

The Charles Stark Draper Laboratory, Inc., 555 Technology Square, Cambridge, Massachusetts 02139

Received November 29, 1982; revised manuscript received May 31, 1983

Equations for the propagation of phase and irradiance are derived, and a Green's function solution for the phase in terms of irradiance and perimeter phase values is given. A measurement scheme is discussed, and the results of a numerical simulation are given. Both circular and slit pupils are considered. An appendix discusses the local validity of the parabolic-wave equation based on the factorized Helmholtz equation approach to the Rayleigh–Sommerfeld and Fresnel diffraction theories. Expressions for the diffracted-wave field in the near-field region are given.

1. INTRODUCTION

The optical phase-retrieval problem, which is to deduce optical phase from minimal irradiance measurements by using noninterferometric techniques, has been studied extensively in recent years. Its literature has been repeatedly surveyed and is easily traceable.[1–3]

Recently[4] a scheme was developed that allowed optical phase to be determined from measured irradiance moments in multiple measurement planes (in particular, two planes sufficed in principle) if the pupils were sufficiently "soft," i.e., not sharp. The present paper, which is based on Green's functions, evolved from a study of methods to determine phase from two-plane irradiance measurements without the auxiliary device of introducing irradiance moments with their concomitant mathematical existence problems associated with sharp pupils. (Which of these two schemes would actually perform best in terms of dynamic range, bandwidth, and signal-to-noise ratio in a given measurement situation can be determined only by further analysis, simulation, and experimental studies.) Both methods are deterministic phase-retrieval schemes in the sense that the phase is given directly in terms of the measured irradiance data, and the phase uniqueness question does not arise here as it does in methods based on iterative algorithms[1–3] to retrieve phase. Previous studies based on two-defocus[5,6] measurement methods and analyticity properties[7–9] of the phase have also concluded that additional irradiance measurements (multiple transverse-measurement planes or measurements with different phase filters inserted in one plane) ensure that the phase that is recovered is unique.

In Section 2 the propagation equations for phase and irradiance are derived, and one of the equations furnishes a basis for phase retrieval. A Green's function solution for the phase is given in Section 3. A new feature is that for finite pupils (i.e., non-Gaussian-like beams) a perimeter phase measurement near the actual pupil boundary is also necessary, and two-plane irradiance measurements alone are insufficient to determine the phase across the area of the pupil. In fact, however, for a uniformly illuminated pupil, perimeter phase measurements and irradiance measurements in one plane a distance δz from the pupil plane suffice to determine ϕ. Moreover, for circularly symmetric aberrations, only one-phase irradiance measurements are needed for phase retrieval; the perimeter phase measurements are unnecessary. Irradiance measurements near the pupil plane are generally inconvenient and difficult from the viewpoint of noise considerations. Section 4 indicates how to make the necessary irradiance measurements in a more favorable plane by using a beam compressor. The penultimate section gives results of a numerical simulation of the phase-determination scheme.

The Green's function method of phase retrieval applies locally anywhere and not just at the pupil, as discussed in the last paragraph. In general, the phase inside a region bounded by some perimeter in a transverse plane is determined by the longitudinal gradient of the irradiance in the given plane and by perimeter phase measurements. As is discussed in Section 3, in some cases, e.g., in the image plane, it is permissible to extend the perimeter effectively to infinity. Then image-plane phase is determined only by the longitudinal irradiance gradient at the image plane, and pupil-plane phase is then given by inverse Fresnel transformation.[10]

This paper is based on the local validity on the parabolic-wave equation[10] for describing the diffracted-wave field. As is discussed in an extensive appendix, this is not the same as assuming the validity of Fresnel diffraction theory since the parabolic-wave equation remains approximately valid locally even when the Fresnel diffraction integral gives an inaccurate description of the diffracted wave (e.g., in the near-field region of a sharp pupil). The somewhat unconventional approach to diffraction theory presented in Appendix A is especially suited to near-field-region considerations.

2. PROPAGATION EQUATIONS FOR PHASE AND IRRADIANCE

Suppose that light propagates nominally in the $+z$ direction, and let the time-dependent wave amplitude be written as $\exp(-i2\pi ct/\lambda)u_z(\mathbf{r})$, where $\mathbf{r} = (x, y)$ is a two-dimensional vector in the transverse direction. Then, as is discussed extensively in Appendix A, the amplitude $u_z(\mathbf{r})$ satisfies approximately the parabolic equation[10] [see Eq. (A18) of Appendix A]

0030-3941/83/111434-08$01.00

$$\left(i\frac{\partial}{\partial z}+\frac{\nabla^2}{2k}+k\right)u_z(\mathbf{r})=0, \tag{1}$$

where $\nabla^2 = [(\partial^2/\partial x^2) + (\partial^2/\partial y^2)]$ and $k = 2\pi/\lambda$. The normalization of $u_z(\mathbf{r})$ is such that

$$I_z(\mathbf{r}) \equiv |u_z(\mathbf{r})|^2 \tag{2}$$

is the irradiance at point (x, y, z).

The wave amplitude may be expressed in terms of the irradiance I and the phase ϕ, which are real-valued quantities, i.e.,

$$u_z(\mathbf{r}) = [I_z(\mathbf{r})]^{1/2}\exp[i\phi_z(\mathbf{r})]. \tag{3}$$

Let Eq. (1) be multiplied on the left-hand side by $u_z{}^*$ and the complex conjugate of Eq. (1) be multiplied on the left-hand side by u_z. If the two resulting equations are subtracted, one gets (suppressing the z subscript)

$$\frac{2\pi}{\lambda}\frac{\partial}{\partial z}I=-\nabla\cdot I\nabla\phi, \tag{4}$$

whereas, if they are added and the sum is multiplied by I, one gets

$$\frac{4\pi}{\lambda}I^2\frac{\partial}{\partial z}\phi=\frac{1}{2}I\nabla^2 I-\frac{1}{4}(\nabla I)^2-I^2(\nabla\phi)^2+kI^2. \tag{5}$$

Thus the real quantities ϕ and I satisfy a coupled set of nonlinear equations. If both ϕ and I are unknown, it is obviously easier to solve the single linear Eq. (1) for the complex-valued amplitude u. However, if the irradiance is known (e.g., by measurement), then Eq. (4) becomes a linear equation, which may be solved to obtain the phase ϕ.

On introduction of an auxiliary function ψ, which satisfies

$$\nabla\psi = I\nabla\phi, \tag{6}$$

Eq. (4) becomes

$$\nabla^2\psi=\frac{-2\pi}{\lambda}\frac{\partial}{\partial z}I, \tag{7}$$

i.e., the two-dimensional Poisson equation.[11] Once ψ is known, ϕ is determined (to within an additive constant) by integrating Eq. (6).

3. GREEN'S FUNCTION SOLUTION FOR THE PHASE

The solution of Eq. (7) inside a two-dimensional region R, which is perpendicular to the z axis and bounded by perimeter P, is given by[11]

$$\psi_z(\mathbf{r})=\frac{-2\pi}{\lambda}\iint_R \mathrm{d}\mathbf{r}'G(\mathbf{r},\mathbf{r}')\frac{\partial}{\partial z}I_z(\mathbf{r}') - \int_P \mathrm{d}s'\left[G(\mathbf{r},\mathbf{r}')\frac{\partial\psi_z(\mathbf{r}')}{\partial n'}-\psi_z(\mathbf{r}')\frac{\partial}{\partial n'}G(\mathbf{r},\mathbf{r}')\right], \tag{8}$$

where $\partial/\partial n' = \hat{n}'\cdot\nabla'$, $\hat{n}'$ is the outward-pointing unit vector normal to the perimeter at $\mathbf{r}'$, and $G(\mathbf{r}, \mathbf{r}')$ is a Green's function that depends on the shape of the region R.

For now we consider the special case of a circular region R of radius a, with xy-coordinate origin at the center of the circle [e.g., round pupil centered on the optical axis. However, for reasons discussed below, the perimeter to be used in Eq. (8) is slightly inside the actual physical perimeter of the round pupil]. We choose, for convenience, to have G satisfy Dirichlet boundary conditions; i.e.,

$$G(\mathbf{r},\mathbf{r}')=0, \tag{9}$$

for $\mathbf{r}'$ on the perimeter.

In addition,

$$\nabla^2 G(\mathbf{r},\mathbf{r}')=\delta(\mathbf{r}-\mathbf{r}') \tag{10}$$

for $\mathbf{r}, \mathbf{r}' \in R$. By using elementary considerations (and the method of images[11]), it is straightforward to construct G satisfying Eqs. (9) and (10); it is

$$G(\mathbf{r},\mathbf{r}')=\frac{1}{4\pi}\ln\left(\frac{\mathbf{r}^2+\mathbf{r}'^2-2\mathbf{r}\cdot\mathbf{r}'}{\dfrac{r^2r'^2}{a^2}+a^2-2\mathbf{r}\cdot\mathbf{r}'}\right). \tag{11}$$

For convenience, we also quote

$$\frac{\partial G(\mathbf{r},\mathbf{r}')}{\partial n'}=\frac{1}{2\pi a}\left(\frac{a^2-r^2}{a^2+r^2-2\mathbf{r}\cdot\mathbf{r}'}\right), \tag{12}$$

where in Eq. (12), but not in Eq. (11), $\mathbf{r}' = a\hat{r}'$ and $\hat{r}'$ is a unit vector.]

Finally, we now specialize to the case that the irradiance is constant ($I = I_0$) over the pupil plane ($z = 0$). Then Eq. (6) implies that

$$\psi = I_0(\phi + \text{constant}). \tag{13}$$

Using Eqs. (8)–(13), we may express the phase (over the interior of a constant-irradiance, round region) as

$$\phi_0(\mathbf{r})=\iint_R \mathrm{d}\mathbf{r}'G(\mathbf{r},\mathbf{r}')\left[\frac{-2\pi}{\lambda I_0}\frac{\partial I_0(\mathbf{r}')}{\partial z}\right] + \int_P \mathrm{d}s'\phi_0(\mathbf{r}')\frac{\partial G}{\partial n'}(\mathbf{r},\mathbf{r}'). \tag{14}$$

[The constant in Eq. (13) drops out of Eq. (14) because of the identity

$$\int_P \mathrm{d}s'\frac{\partial}{\partial n'}G(\mathbf{r},\mathbf{r}')=1 \tag{15}$$

for any $\mathbf{r}$, as is straightforward[12] to verify from Eq. (12).]

Equation (14) indicates that the phase $\phi_0(\mathbf{r})$ at the pupil plane is determined by the value of $\phi_0(\mathbf{r})$ on the perimeter P and by the value of the longitudinal gradient of the irradiance over the area of the pupil. In fact, since the gradient must be approximated by measured values

$$\frac{\partial I_0(\mathbf{r}')}{\partial z}\approx\frac{I_{\delta z}(\mathbf{r}')-I_0}{\delta z},$$

and since I_0 is a (known) constant, irradiance measurements need to be made only in the single plane $z = \delta z$. In subsequent sections we discuss how to make this measurement and how to choose δz.

The perimeter phase measurements to be used in Eq. (14) determine only the relative phase of points on the perimeter, and the relative phase between perimeter points and interior points can be indeterminate to within an overall unknown constant. Because of Eq. (15) this affects ϕ as determined by

Eq. (14) only by the same unknown constant (and hence is a trivial piston phase error).

In some cases it is permissible to extend the radius of the perimeter to infinity (e.g., with an untruncated Gaussian beam). Then the Green's function becomes $G(\mathbf{r}, \mathbf{r}') = (1/2\pi)\ln(|\mathbf{r} - \mathbf{r}'|/\Lambda)$, where Λ is a constant length and $\nabla\phi$ is independent of the particular choice of Λ. Moreover, the perimeter integrals in Eq. (8) do not appear; thus two-plane irradiance measurements alone suffice to determine ϕ.

(For reference purposes we quote here the analogous results for an infinite slit aperture over which the irradiance is uniform. Again $+z$ is the propagation direction, y is a transverse coordinate parallel to the two parallel slit edges, which are separated by distance $2a$, and x is a transverse coordinate perpendicular to the slit edges. The coordinates' origin is at the center of the slit, and it is assumed that the wave amplitude is independent of y.) Then, analogous to Eq. (14), we have

$$\phi_0(x) = \int_{-a}^{a} \mathrm{d}x' G(x, x') \left[\frac{-2\pi}{\lambda I_0}\frac{\partial}{\partial z} I_0(x')\right] + \left(\frac{x+a}{2a}\right)\phi_0(a) - \left(\frac{x-a}{2a}\right)\phi_0(-a), \quad (16)$$

where the Green's function is

$$G(x, x') = \frac{1}{2}|x - x'| - \frac{1}{2}\left|\frac{xx'}{a} - a\right| \quad (17)$$

and satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} G(x, x') = \delta(x - x'), \quad (18)$$

$$G(x, a) = G(x, -a) = 0. \quad (19)$$

4. MEASUREMENT SCHEME

In this paper we assume that there is no basic problem in making the necessary perimeter phase measurements (e.g., by using a set of Hartmann[13] sensors). We also assume that the light is nominally collimated at the pupil. Then $\phi_0(\mathbf{r})$ is the pupil-plane phase aberration. In general it is difficult to measure $\partial I_0/\partial z$ by subtracting irradiance values at points near the pupil plane because [see (Eq. (4)] $\delta I_0/I_0 \approx (\lambda\delta z/a^2) \times$ (number of waves of aberration). For example, for visible light and large pupils ($a \approx 1$ m), $\delta I_0/I_0 \approx 10^{-6}$ (δz/m). The longitudinal irradiance variation can be greatly increased by using a beam compressor (see Fig. 1) with, e.g., two converging lenses with focal lengths F and f, respectively. The magnification is $m = f/F \ll 1$ for our case. (In fact, the first lens will usually already be part of the optical system, and ϕ_0 is the phase aberration just before the lens plus the aberration introduced by the lens.)

We treat the beam compressor as a linear filter, assume paraxial optics and thin, aberration-free lenses, and ignore finite-aperture effects. If the output plane z_2 is chosen so that

$$z_2 = z_1 + (1 + m)f \approx z_1 + f,$$

then one can show, by using standard techniques,[14] that

$$u_{z_2}(\mathbf{r}) = -\frac{1}{m} u_0(-\mathbf{r}/m). \quad (20)$$

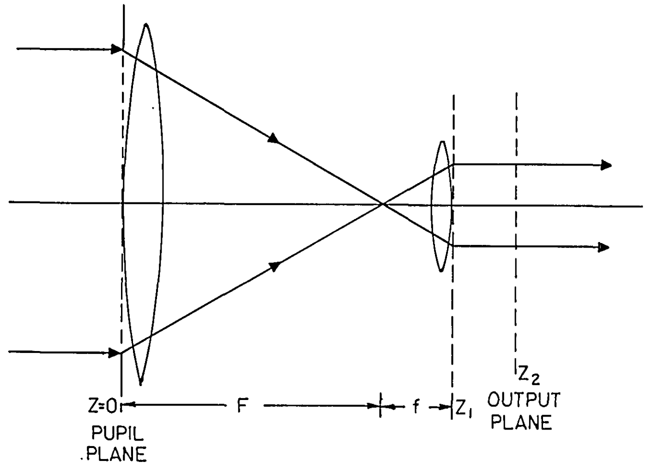

Fig. 1. Beam compressor. Output plane is located at $z_2 = z_1 + (1 + f/F)f$. The output-plane wave amplitude is $u_{z_2}(\mathbf{r}) = (-1/m)u_0(-\mathbf{r}/m)$, where $m = f/F$.

Hereafter, let the xy-coordinate system at z_2 be rotated 180° about the z axis. Then

$$I_{z_2}(\mathbf{r}) = \frac{1}{m^2} I_0(\mathbf{r}/m) \quad (21)$$

and

$$\phi_{z_2}(\mathbf{r}) = \phi_0(\mathbf{r}/m) + \pi, \quad (22)$$

and, using Eq. (4), which applies both to planes $z = 0$ and $z = z_2$, we have

$$\frac{1}{I_{z_2}}\frac{\partial I_{z_2}(\mathbf{r})}{\partial z} = \frac{1}{m^2}\frac{1}{I_0}\frac{\partial I_0(\mathbf{r}/m)}{\partial z}. \quad (23)$$

Thus $\delta I_{z_2}/I_{z_2} \approx (\lambda\delta z/m^2a^2) \times$ (number of waves of aberration) and, for a high-ratio compressor (e.g., $m \approx 10^{-3}$), $\delta I_{z_2}/I_{z_2} \approx$ (δz/1 meter), which is readily measurable. Equation (23) is then used to obtain, for use in Eq. (14), $(1/I_0)\partial I_0/\partial z$ from measured data. (Incidentally, the beam-compressor technique makes feasible the two-plane irradiance-moment method[4] of pupil-plane phase retrieval.)

5. NOISE CONSIDERATIONS AND CHOICE OF δz

Near the plane $z = z_2$, the irradiance may be written as

$$I_{z_2+\delta z}(\mathbf{r}) = I_{z_2}[1 + b_1\delta z + b_2(\delta z)^2] + \Delta I_{z_2}, \quad (24)$$

where

$$b_1 = -2\left(\frac{\lambda}{4\pi}\right)\nabla^2\phi_{z_2} \quad (25)$$

and

$$b_2 = 4\left(\frac{\lambda}{4\pi}\right)^2[(\phi_{xx})^2 + \phi_x\phi_{xxx} + \phi_{xx}\phi_{yy} + \phi_y\phi_{yyy} + (\phi_{yy})^2]_{z=z_2}. \quad (26)$$

In Eq. (26) $\phi_{xx} = (\partial^2/\partial x^2)\phi$, etc. Equations (24)–(26) are derived by using Eq. (A32) of Appendix A to second order in δz. ΔI_{z_2} is a noise term. Assuming the most favorable case of photon-noise-limited detection (Poisson-detection statistics) and unity quantum efficiency, one can show[15] that ΔI_{z_2}

has zero mean and standard deviation:

$$(\Delta I_{z2})_{\rm rms} = I_{z2}\left(\frac{hc/\lambda}{A_d \Delta t I_{z2}}\right)^{1/2} = I_{z2}/\sqrt{N_{z2}}, \tag{27}$$

where A_d is the detector element area and N_{z2} is the total number of photons counted in area A_d in time Δt. [Since $I_{z2+\delta z}$ differs from I_{z2} by only a few per cent, one can use z_2 rather than $z_2 + \delta z$ for estimation purposes in Eq. (27).] The measured quantity is

$$\frac{I_{z2+\delta z} - I_{z2}}{\delta z I_{z2}} = b_1 + b_2 \delta z + \frac{\Delta I_{z2}}{\delta z I_{z2}}. \tag{28}$$

δz must be small enough that the first term dominates the second:

$$\delta z \ll |b_1/b_2|,$$

and large enough that the noise term is small:

$$\frac{1}{|b_1|\sqrt{N_{z2}}} \ll \delta z. \tag{29}$$

[One can show that, for fixed b_2 and N_{z2}, the optimal, in the sense of minimizing the last two terms on the right-hand side of Eq. (28), δz is $|b_2|^{-1/2} N_{z2}^{-1/4}$.]

6. NUMERICAL SIMULATION

For simulation purposes the following system parameters were chosen arbitrarily: $F = 3$ m, $a = 1$ m, $f = 0.93$ mm (thus $m = 3.1 \times 10^{-4}$), $\lambda = 10^{-6}$ m, $A_d = (20\ \mu\text{m})^2$. Then the detector array at z_2 has 31 × 31 elements. The demagnified round pupil, with a diameter of 610 μm, then just fits as a circle inscribed inside the 610-μm × 610-μm square detector array. This array size was chosen empirically as this mesh size guaranteed that the phase calculated by Eq. (14) agreed with the known phase aberration to within 1/20 wave when there was no added noise. The area integrals were performed by a 31- × 31-point 2-*d* Simpson's rule. For the perimeter integrals a 93-point Simpson's rule was used. The Green's function and its normal derivative have integrable singularities that slow the convergence of the numerical integration. An *ad hoc* procedure of drawing a circle of radius 0.001 (relative to a pupil radius of 1.0) around the singularities and setting the integrand to zero inside the small circle was adopted. A larger keep-out radius speeds convergence but to an inaccurate answer.

Gaussian random noise of zero mean and 1/20-wave standard deviation was added to the perimeter phase values. Gaussian random noise of zero mean and standard deviation corresponding to $\Delta I/I = 1/10$ (i.e., 100 photons/detector cell) was added to the area integrations.

Figures 2 and 3 show results of the calculations for one wave of coma and one wave of spherical aberration, respectively. The quantity plotted is

$$\Delta\phi(x, 0) = \phi(x, 0)_{\rm actual} - \phi(x, 0)_{\rm calculated}.$$

For these conditions, $\delta z \approx 2$ mm allowed inequalities (25) and (26) to be met.

The purpose of the numerical simulation was to verify the correctness of Eqs. (11), (12), and (14) and their relative noise immunity. No attempt was intended to optimize a system design, e.g., in terms of minimum array size or maximum tolerable noise.

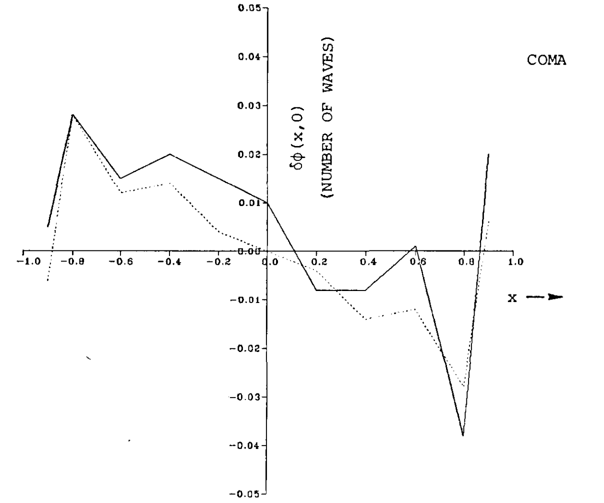

Fig. 2. Phase error for one wave of coma. Both here and in Fig. 3 a plot of $\delta\phi(x, 0) = \phi_{\rm actual}(x, 0) - \phi_{\rm calculated}(x, 0)$ is given. The solid curve has simulated noise added to the calculation. The dotted curve is the noise-free calculation and is nonzero because the number of points used in the numerical integrations was finite. The pupil radius is 1.0.

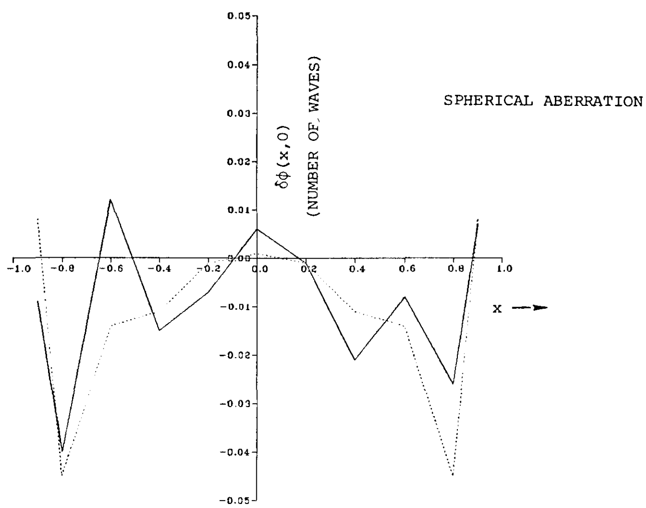

Fig. 3. Phase error for one wave of spherical aberration.

In all cases for which numerical calculations were made, the area integral and the perimeter integral in Eq. (14) produced results of the same order of magnitude; thus neither can be neglected for a finite pupil {unless it is known that the aberration is circularly symmetric, in which case the phase is constant on the line integral and only the area integral need be calculated [see Eq. (15)]}.

7. CONCLUDING COMMENTS

The approach to phase retrieval presented in this paper allows optical phase to be determined uniquely from longitudinal irradiance gradients and possibly also perimeter phase measurements. The solution presented involves Green's functions, and a numerical simulation showed that adequate accuracy is attained with a relatively small detector array. Moreover, the algorithm is not particularly sensitive to noise. If pupil-plane phase is desired, then irradiance measurements must be made at a demagnified version of the pupil by using a beam compressor. This technique would not be necessary if measurements were made at the image plane, but then the scheme of this paper will give the optical phase at the image plane.

APPENDIX A. EXACT DIFFRACTION THEORY, APPROXIMATE VALIDITY OF THE PARABOLIC-WAVE EQUATION, AND NEAR-FIELD CONSIDERATIONS

In this Appendix it is shown in Section A.2 that, while the Fresnel diffraction integral and the parabolic-wave equation imply each other, the parabolic-wave equation is, locally, approximately valid even when the global Fresnel diffraction integral produces absurd results (e.g., in the near-field diffraction region of a sharp pupil). In Section A.3, expressions describing the diffracted-wave field of aberrated pupils in the near-field region are given. The results of Sections A.2 and A.3 follow from scalar diffraction theory as summarized in Section A.1 starting with the factorized wave equation, which has long been known in the physics literature[16–18] and has been used increasingly for treating problems in optics[19–21] and acoustics.[22,23] The reader is reminded that the methods of the factorized wave equation (which leads to the diffraction operator approach[19–21]), the angular spectrum,[14,24] and the Rayleigh–Sommerfeld diffraction theory[11,14,25] are all equivalent formulations of diffraction theory.

A.1. Rayleigh-Sommerfeld versus Fresnel Diffraction Theories

The time-independent wave equation in empty space

$$\left[\frac{\partial^2}{\partial z^2} + \nabla^2 + \left(\frac{2\pi}{\lambda}\right)^2\right] \Psi_z(\mathbf{r}) = 0 \tag{A1}$$

[recall that $\nabla^2 = \partial_x{}^2 + \partial_y{}^2$ and $\mathbf{r} = (x, y)$] that describes waves traveling in all directions may be written as

$$L_+L_-\Psi_z(\mathbf{r}) = 0, \tag{A2}$$

where

$$L_\pm = \frac{\partial}{\partial z} \mp \frac{i2\pi}{\lambda}\left[1 + \left(\frac{\lambda\nabla}{2\pi}\right)^2\right]^{1/2}. \tag{A3}$$

The solutions of Eq. (A1) thus separate into two classes:

$$L_+u_z(\mathbf{r}) = 0 \tag{A4}$$

and

$$L_-v_z(\mathbf{r}) = 0. \tag{A5}$$

(Choose $z = 0$ to be the initial plane.) The solutions $\{u_z\}$ describe either oscillatory waves with a positive z component of the wave vector $2\pi[\boldsymbol{\rho}, (1/\lambda)(1 - \lambda^2\rho^2)^{1/2}]$ or evanescent waves if $\lambda^2\rho^2 > 1$ and $z > 0$, whereas the solutions $\{v_z\}$ describe either oscillatory waves with a negative z component of the wave vector $2\pi[\boldsymbol{\rho}, -(1/\lambda)(1 - \lambda^2\rho^2)^{1/2}]$ or evanescent waves if $\lambda^2\rho^2 > 1$ and $z < 0$. In the absence of scattering by charged matter, the two solution classes do not mix. Hereafter, we consider only the solutions $\{u_z\}$.

A formal solution of Eq. (A4) may be written as

$$u_z(\mathbf{r}) = \exp\left[ikz\left(1 + \frac{\nabla^2}{k^2}\right)^{1/2}\right] u_0(\mathbf{r}), \tag{A6}$$

where $k = 2\pi/\lambda$, and the square-root operator is defined ultimately (as discussed in detail in Section A.3) in terms of the Fourier transform of Eq. (A6), i.e.,

$$U_z(\boldsymbol{\rho}) = \exp[ikz(1 - \lambda^2\rho^2)^{1/2}]U_0(\boldsymbol{\rho}), \tag{A7}$$

where the two-dimensional Fourier transform is generally defined by

$$U_z(\boldsymbol{\rho}) = \int \mathrm{d}\mathbf{r} \exp(-i2\pi\boldsymbol{\rho}\cdot\mathbf{r})u_z(\mathbf{r}) \equiv \mathrm{FT}[u_z(\mathbf{r})]_{\boldsymbol{\rho}}. \tag{A8}$$

It is, however, known[26,27] that the inverse Fourier-transform relationship

$$\mathrm{FT}^{-1}\{\exp[ikz(1 - \lambda^2\rho^2)^{1/2}]\} = -\frac{1}{2\pi}\frac{\partial}{\partial z}\frac{e^{ikR}}{R} \tag{A9}$$

holds, where

$$R \equiv (z^2 + r^2)^{1/2}.$$

Therefore, from Eqs. (A7) and (A9), we have

$$u_z(r) = -\frac{1}{2\pi}\frac{\partial}{\partial z}\left[u_0(\mathbf{r}) ** \frac{e^{ikR}}{R}\right], \tag{A10}$$

which is just the usual Rayleigh–Sommerfeld diffraction theory result,[11,14,25] which gives the wave amplitude in a transverse plane $z \geq 0$ in terms of the wave amplitude in an earlier plane $z = 0$. Notice that $**$ denotes two-dimensional convolution, i.e., for two functions f and g:

$$f(\mathbf{r}) ** g(\mathbf{r}) \equiv \int \mathrm{d}\mathbf{r}' f(\mathbf{r}')g(\mathbf{r} - \mathbf{r}'). \tag{A11}$$

The conventional Fresnel diffraction theory results when the square roots in Eqs. (A6) and (A7) are expanded to lowest order to get

$$u_{z,F}(\mathbf{r}) = \exp(ikz)\exp\left[\frac{(i\lambda z\nabla^2)}{4\pi}\right] u_0(\mathbf{r}) \tag{A12}$$

and

$$U_{z,F}(\boldsymbol{\rho}) = \exp(ikz)\exp(-i\pi\lambda z\rho^2)U_0(\boldsymbol{\rho}), \tag{A13}$$

and, using the elementary (inverse) Fourier-transform relationship,

$$\mathrm{FT}^{-1}[\exp(-i\pi\lambda z\rho^2)] = \frac{\exp(i\pi r^2/\lambda z)}{i\lambda z}, \tag{A14}$$

we have from Eq. (A12)

$$u_{z,F}(\mathbf{r}) = \exp(ikz)\left[u_0(\mathbf{r}) ** \frac{\exp(i\pi r^2/\lambda z)}{i\lambda z}\right], \tag{A15}$$

which is the usual Fresnel diffraction theory result[14] (with the obliquity factor taken equal to unity).

One can see immediately that there are cases in which the Rayleigh–Sommerfeld and Fresnel diffraction theories predict essentially the same diffracted-wave field for any $z \geq 0$. For example, if the spectrum $U_0(\boldsymbol{\rho})$ of the initial plane-wave amplitude is essentially zero for $|\boldsymbol{\rho}| > \epsilon/\lambda$ (where ϵ is a small number, say, $1/10$), then Eq. (A7) is effectively the same as Eq. (A13), and both theories predict essentially the same diffracted wave. This situation occurs characteristically when the initial plane has a smooth wave-amplitude distribution with a transverse width $\gg\lambda$ and explains why the Fresnel diffraction theory accurately describes the propagation of a Gaussian beam for any $z \geq 0$.

On the other hand, it is characteristic of the two theories to predict significant differences in the near-field region when the initial plane has a sharp pupil. This can be seen immediately for a round pupil of radius a, normally illuminated by collimated, monochromatic light. For *axial* observation points ($\mathbf{r} = 0$), it is straightforward to evaluate analytically both Eqs. (A10) and (A15) to get

$$U_z(0) = \exp(ikz) - \frac{z \exp[ik(z^2 + a^2)^{1/2}]}{(z^2 + a^2)^{1/2}} \qquad \text{(A16)}$$

and

$$U_{z,F}(0) = \exp(ikz) - \exp\left[ik\left(z + \frac{1}{2}\frac{a^2}{z}\right)\right]. \qquad \text{(A17)}$$

[In arriving at Eq. (A16) it is best to use $R' = (z^2 + r'^2)^{1/2}$ rather than r' as an integration variable in Eq. (A10). Moreover, it appears generally true in evaluating Eq. (A10) analytically that it is better to differentiate with respect to z *after* the convolution is performed.] For $z \gg a$, both theories predict the same result. [A more precise condition for the agreement of Eqs. (A16) and (A17) is that $z/a \gg (a/\lambda)^{1/3}$ be satisfied, as discussed on p. 59 of Ref. 14.] Significant differences occur when $z \sim a$, and for $z \ll a$ Eq. (A17) predicts longitudinal oscillations in the wave field over distances even smaller than λ, which is an unphysical result. Equation (A16), however, correctly reproduces the initial field as $z \rightarrow 0$.

A.2. Local Validity of the Parabolic Equation

From Eq. (A15) it may be verified that the Fresnel result satisfies exactly the parabolic equation[10]

$$\left(i\frac{\partial}{\partial z} + \frac{\nabla^2}{2k} + k\right) u_{z,F}(\mathbf{r}) = 0 \qquad \text{(A18)}$$

[which is also Eq. (A4) with the square root expanded to lowest nontrivial order]. In fact, Eq. (A18) has a greater region of validity than the integral result of Eq. (A15) in the sense that locally, away from sharp pupil edges, the correct (Rayleigh–Sommerfeld) wave field satisfies approximately the parabolic equation. To see this, we write [using Eqs. (A7) and (A8)]

$$u_z(\mathbf{r}) = \int \mathrm{d}\boldsymbol{\rho} \exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r})\exp[ikz(1 - \lambda^2\rho^2)^{1/2}]U_0(\boldsymbol{\rho}) \qquad \text{(A19)}$$

and apply the operator $[i(\partial/\partial z) + (\nabla^2/\partial k) + k]$ to both sides of this equation to get

$$\begin{aligned}\left(i\frac{\partial}{\partial z} + \frac{\nabla^2}{\partial k} + k\right) u_z(\mathbf{r}) &= \int \mathrm{d}\boldsymbol{\rho} \exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r})\exp[ikz(1 - \lambda^2\rho^2)^{1/2}] \\ &\times U_0(\boldsymbol{\rho})[-k(1 - \lambda^2\rho^2)^{1/2} - \pi\lambda\rho^2 + k]. \end{aligned} \qquad \text{(A20)}$$

In general, it is permissible to expand, in powers of ρ^2, the second expression inside square brackets in Eq.(A20), and moreover this is not equivalent to neglecting evanescent waves since the square root in the exponential $\exp[ikz(1 - \lambda^2\rho^2)^{1/2}]$, to which the integral in Eq. (A20) is much more sensitive, is not expanded. For example, for a round pupil of radius a, normally illuminated, the initial plane spectrum is proportional to

$$\frac{J_1(2\pi a\rho)}{(2\pi a\rho)},$$

and, for ρ as large as $1/3\lambda$, $U_0(\rho)$ has already fallen to the small value

$$U_0\left(\frac{1}{3\lambda}\right) \approx \left(\frac{\lambda}{a}\right)^{3/2} U_0(0), \qquad \text{(A21)}$$

assuming $a \gg \lambda$. To lowest nonvanishing order, a term proportional to ρ^4 remains inside the second set of square brackets in Eq. (A20), and this may be brought outside the integral as an operator proportional to ∇^4 if the point $(\mathbf{r}, z)$ is not located at a pupil edge so that Eq. (A20) becomes

$$\left(i\frac{\partial}{\partial z} + \frac{\nabla^2}{2k} + k - \frac{1}{8}\frac{\nabla^4}{k^3}\right) u_z(\mathbf{r}) \approx 0. \qquad \text{(A22)}$$

[We note that Eq. (A22) is formally the same as Eq. (A4) with the square-root operator expanded to terms of order ∇^4; however, a straightforward expansion of Eq. (A4) gives no indication of the connection between the accuracy of a truncated expansion and the initial plane spectrum $U_0(\boldsymbol{\rho})$.]

We now consider, for several cases, the order of magnitude of the ratio of the term involving ∇^4 in Eq. (A22) to the term involving ∇^2, i.e., we look at

$$\eta_{42} \equiv \frac{1}{4k^2}\frac{\nabla^4 u_z(\mathbf{r})}{\nabla^2 u_z(\mathbf{r})}$$

as follows:

(1) Near the pupil, but away from pupil edges, for nominally collimated light

$$\eta_{42} \approx 0\left[\frac{\lambda^2}{a^2} \times (\text{number of waves of aberration})^2\right],$$

where a is the pupil radius.

(2) Near the pupil, away from pupil edges, for light nominally focused at $z = f$,

$$\eta_{42} \approx 0\left[\frac{1}{(f\#)^2}\right],$$

where $f\# = f/2a$.

(3) In a focused system, near the image plane,

$$\eta_{42} \approx 0\left[\frac{r^2/\lambda^2}{(f\#)^4}\right].$$

In these cases we see that η_{42} is small and hence the parabolic equation is approximately valid locally, away from pupil edges, if $a \gg \lambda$ and, for focused system, if $f\# \gg 1$. These conclusions are consistent with results recently reported,[27] which were based on direct numerical calculations comparing the Fresnel and Rayleigh–Sommerfeld theories.

A.3. Diffraction Operators and Near-Field Considerations

Equations (A6) and (A12) give, respectively, for the Rayleigh–Sommerfeld and Fresnel diffraction theories, a diffraction-operator solution of the wave equation. In this subsection we consider how these operators are defined and use the results to investigate the near-field behavior of the diffracted-wave field.

It is convenient to consider first the mathematically simpler Fresnel theory. It is natural to attempt to define the exponential operator appearing in Eq. (A12) by its power-series expansion, i.e.,

$$\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)u_0(\mathbf{r}) \overset{?}{=} \sum_{N=0}^{\infty}\frac{1}{N!}\left(\frac{i\lambda z\nabla^2}{4\pi}\right)^N u_0(\mathbf{r}). \quad \text{(A23)}$$

Obviously this will work only if $u_0(\mathbf{r})$ is infinitely differentiable at $\mathbf{r}$ and the series in Eq. (A23) converges. If $u_0(\mathbf{r})$ describes a round sharp pupil of radius a, normally illuminated, Eq. (A23) gives, for any $z > 0$, zero if $|\mathbf{r}| < a$ or $|\mathbf{r}| > a$ and an indeterminate (and infinite) result $|\mathbf{r}| = a$. So Eq. (A23) produces an absurd result if the initial plane contains sharp pupils. The fact that a power-series representation of Eq. (A12) generally fails to be valid is not surprising since any operator involving effectively an infinite number of derivatives is actually a nonlocal operator and, in fact, one correct representation of Eq. (A12) is just the familiar Fresnel-diffraction result of Eq. (A15).

Hereafter we shall always consider only initial wave distributions $u_0(\mathbf{r})$ that have a Fourier transform [see Eq. (A8)] $U_0(\boldsymbol{\rho})$. Then the Fresnel-diffraction operator can always be defined by the first equality of the equation

$$\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)u_0(\mathbf{r}) \equiv \int d\boldsymbol{\rho}\exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r}) \times \exp(-i\pi\lambda z\rho^2)U_0(\boldsymbol{\rho}) \quad \text{(A24)}$$

$$\overset{?}{=} \sum_{N=0}^{\infty}\frac{1}{N!}\int d\boldsymbol{\rho}\exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r}) \times (-i\pi\lambda z\rho^2)^N U_0(\boldsymbol{\rho}). \quad \text{(A24')}$$

Whereas Eq. (A24) is always valid for physically realizable cases, e.g., sharp pupils or smooth pupils, Eq. (A24′), if valid, requires each term in the series to be finite and also the series to converge. This will occur if, e.g., as $\rho \to \infty$, $U_0(\boldsymbol{\rho}) \sim 0[\exp(-\text{const. } \rho^n)]$ for $n = 1, 2, 3, \ldots$, and this implies that $u_0(\mathbf{r})$ is a smooth, e.g., Gaussian-like, pupil. We do not wish to limit our considerations to such smooth pupils, however, and in the following paragraphs extend the results to more-general pupils.

We now consider the near-field behavior of the Fresnel diffracted wave. We express the initial wave as

$$u_0(\mathbf{r}) = p_0(\mathbf{r})f_0(\mathbf{r}), \quad \text{(A25)}$$

where $p_0(\mathbf{r})$ describes a sharp pupil and necessitates the use of Eq. (A24) while $f_0(\mathbf{r})$ describes a smooth function and ultimately follows Eq. (A23) to be used operating on $f_0(\mathbf{r})$. For example, we might have

$$p_0(\mathbf{r}) = \begin{cases} 1, & r \le a \\ 0, & \text{otherwise} \end{cases}$$

while

$$f_0(\mathbf{r}) = \exp[ikw(\mathbf{r})]\exp(-\epsilon r^2),$$

where $w(\mathbf{r})$ is the pupil-plane wave-front aberration and ϵ is a very small (ultimately set equal to zero) convergence parameter used to ensure that the Fourier transform of f_0 exists. Then from Eq. (A24) and (A25) we have

$$\begin{aligned}\exp\left(i\frac{\lambda z\nabla^2}{4\pi}\right)u_0(\mathbf{r}) &= \int d\boldsymbol{\rho}\exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r})\exp(-i\pi\lambda z\rho^2) \\ &\quad\times\int d\boldsymbol{\rho}'P_0(\boldsymbol{\rho}')F_0(\boldsymbol{\rho}-\boldsymbol{\rho}') \\ &= \int d\boldsymbol{\rho}'P_0(\boldsymbol{\rho}')\int d\boldsymbol{\rho}\exp(i2\pi\boldsymbol{\rho}\cdot\mathbf{r}) \\ &\quad\times\exp(-i\pi\lambda z\rho^2)F_0(\boldsymbol{\rho}-\boldsymbol{\rho}') \\ &= \int d\boldsymbol{\rho}'P_0(\boldsymbol{\rho}')\exp(i2\pi\boldsymbol{\rho}'\cdot\mathbf{r}) \\ &\quad\times\exp(-i\pi\lambda z\rho'^2) \\ &\quad\times\left\{\int d\boldsymbol{\rho}''\exp[i2\pi\boldsymbol{\rho}''\cdot(\mathbf{r}-\lambda z\boldsymbol{\rho}')]\right. \\ &\quad\left.\times\exp(-i\pi\lambda z\rho''^2)F_0(\boldsymbol{\rho}'')\right\}, \end{aligned} \quad \text{(A26)}$$

where the third equality of Eq. (A26) results from the second equality on introduction of dummy variable $\boldsymbol{\rho}'' = \boldsymbol{\rho} - \boldsymbol{\rho}'$. The expression inside the braces may be written as

$$\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)f_0(\mathbf{r}-\lambda z\boldsymbol{\rho}') = \exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)\sum_{M=0}^{\infty}\sum_{N=0}^{\infty} \times f_0^{(m,n)}(\mathbf{r})\frac{(-\lambda z\xi)^m(-\lambda z\eta)^n}{m!n!}, \quad \text{(A27)}$$

where $\boldsymbol{\rho} = (\xi, \eta)$ and $f_0^{(m,n)}(\mathbf{r}) = \partial^m/\partial x^m\partial^n/\partial y^n f_0(\mathbf{r})$. Thus, finally,

$$\begin{aligned}u_z(\mathbf{r}) = \exp(ikz)\sum_{m-0}^{\infty}\sum_{n=0}^{\infty}\frac{1}{m!n!}&\left[\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)f_0^{m,n}(\mathbf{r})\right] \\ \times&\left\{\left(-\frac{\lambda z}{2\pi i}\frac{\partial}{\partial x}\right)^m\left(-\frac{\lambda z}{2\pi i}\frac{\partial}{\partial y}\right)^n\int d\boldsymbol{\rho}'P_0(\boldsymbol{\rho}')\right. \\ \times&\left.\exp(i2\pi\boldsymbol{\rho}'\cdot\mathbf{r})\exp(-i\pi\lambda z\rho'^2)\right\}. \end{aligned} \quad \text{(A28)}$$

Equation (A28) is an exact result of Fresnel diffraction theory, valid for all $z \ge 0$. The expression in square brackets describes the effect of aberrations, and the expression in braces involves only the aberration-free diffracted-wave field. In the near field (lowest order in z), we have

$$u_z(\mathbf{r}) \approx p_{z,F}(\mathbf{r})\left[\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right)f_0(\mathbf{r})\right], \quad \text{(A29)}$$

where $P_{z,F}(\mathbf{r})$ is the Fresnel diffracted-wave field of the aberration-free pupil.

In the case of the exact Rayleigh–Sommerfeld diffraction theory, the effects of the sharp pupil and the smoothly varying aberrations can be separated only approximately. If we start with Eqs. (A6) and (A25) and follow steps analogous to Eq. (A26) through Eq. (A29), problems arise in the third equality of the equation analogous to Eq. (A26). One can proceed with the derivation only if the spectra $P_0(\boldsymbol{\rho})$ and $F_0(\boldsymbol{\rho})$ are such that the approximation

$$[1-\lambda^2(\boldsymbol{\rho}'+\boldsymbol{\rho}'')^2]^{1/2} \approx (1-\lambda^2\rho'^2)^{1/2} - \lambda^2\boldsymbol{\rho}'\cdot\boldsymbol{\rho}'' + (1-\lambda^2\rho''^2)^{1/2} - 1 \quad \text{(A30)}$$

is valid. For small aberrations, $F_0(\boldsymbol{\rho})$ is sharply peaked about $|\boldsymbol{\rho}| = 0$ while $P_0(\boldsymbol{\rho})$ has properties typical of Eq. (A21). Thus Eq. (A30) is a reasonable approximation and will not lead to the type of physical absurdity in the near-field region characteristic of Eq. (A17). With Eq. (A30) one gets as an approximation to the Rayleigh–Sommerfeld diffracted wave:

$$u_z(\mathbf{r}) \approx \sum_{m-0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{m!n!} \{\exp[ikz(1 + \nabla^2/k^2)^{1/2}] f_0^{(m,n)}(\mathbf{r})\}$$
$$\times \left\{ \left(-\frac{\lambda z}{2\pi i}\frac{\partial}{\partial x}\right)^m \left(-\frac{\lambda z}{2\pi i}\frac{\partial}{\partial y}\right)^n \int \mathrm{d}\boldsymbol{\rho}' P_0(\boldsymbol{\rho}')\exp(i2\pi\boldsymbol{\rho}'\cdot\mathbf{r})\right.$$
$$\left.\times \exp[ikz(1-\lambda^2\rho^2)^{1/2}]\right\} \exp(ikz). \qquad \text{(A31)}$$

The first expression in braces describes the effects of aberrations, and here the diffraction operator may be expanded in a power series. The second factor in braces now involves the Rayleigh–Sommerfeld diffracted wave from an aberration-free pupil. For the near-field region,

$$u_z(\mathbf{r}) \approx p_z(\mathbf{r}) \left[\exp\left(i\frac{\lambda z}{4\pi}\nabla^2\right) f_0(\mathbf{r})\right], \qquad \text{(A32)}$$

where $P_z(r)$ is now the wave predicted by Rayleigh–Sommerfeld theory for an aberration-free pupil.

ACKNOWLEDGMENT

The author expresses his gratitude to V. N. Mahajan for helpful discussions regarding this paper.

A preliminary version of this paper was presented as a postdeadline paper at the Topical Meeting on Signal Recovery and Synthesis with Incomplete Information and Partial Constraints, January 12–14, 1983, Incline Village, Nevada.

REFERENCES

1. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
2. R. H. T. Bates and W. R. Fright, "Reconstructing images from their Fourier intensities," in *Signal and Image Reconstruction from Incomplete Data. Theory & Experiment*, Vol. I of Advances in Computer Vision and Image Processing, T. S. Huang, ed. (JAI, Greenwich, Conn., 1983), Chap. 5.
3. H. P. Baltes, ed., *Inverse Source Problems in Optics* (Springer-Verlag, New York, 1978); *Inverse Scattering Problems in Optics* (Springer-Verlag, New York, 1980).
4. M. R. Teague, "Irradiance moments: their propagation and use for unique retrieval of phase," J. Opt. Soc. Am. **72**, 1199–1209 (1982).
5. B. J. Hoenders, "On the solution of the phase retrieval problem," J. Math. Phys. **16**, 1719–1725 (1975).
6. A. M. Huiser and H. A. Ferweda, "The problem of phase retrieval in light and electron microscopy of strong objects II. On the uniqueness and stability of object reconstruction procedures using two defocused images," Opt. Acta **23**, 445–456 (1976).
7. J. G. Walker, "The phase retrieval problem: a solution based on zero location by exponential apodization," Opt. Acta **28**, 735–738 (1981).
8. J. M. Wood, M. A. Fiddy, and R. E. Burge, "Phase retrieval using two intensity measurements in the complex plane," Opt. Lett. **6**, 514–516 (1981).
9. P. Kiedron, "Phase retrieval methods using differential filters," Proc. Soc. Photo-Opt. Instrum. Eng. **413** (to be published).
10. A. Papoulis, *Systems and Transforms with Applications in Optics* (McGraw-Hill, New York, 1918). The parabolic equation is treated in problem 9.9.
11. J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1975), Chaps. I, II, IX.
12. I. S. Gradsteyn and I. M. Ryzhik, *Tables of Integrals, Series and Products* (Academic, New York, 1965), formula (2.553), no. 3.
13. J. W. Hardy, "Active optics: a new technology for the control of light," Proc. IEEE **66**, 651–697 (1978).
14. J. W. Goodman, *Introduction to Fourier Optics* (McGraw-Hill, New York, 1968).
15. R. H. Kingston, *Detection of Optical and Infrared Radiation* (Springer-Verlag, Berlin, 1978).
16. P. A. M. Dirac, "The quantum theory of the electron," Proc. R. Soc. London Ser. A **117**, 610–624 (1928).
17. N. N. Bogoliubov and D. V. Shirkov, *Introduction to the Theory of Quantized Fields* (Wiley-Interscience, New York, 1959), Sec. 6-1.
18. J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics* (McGraw-Hill, New York, 1964), p. 5.
19. M. D. Feit and J. A. Fleck, "Light propagation in graded index optical fibers," Appl. Opt. **17**, 3990–3998 (1978).
20. M. J. Bastiaans, "Transport equations for the Wigner distribution functions," Opt. Acta **26**, 1265–1272 (1979).
21. A. W. Lohmann, J. Ojeda-Castaneda, and N. Streibl, "Symmetries in coherent and partially coherent fields," Opt. Acta **30**, 399–402 (1983).
22. F. D. Tappert, "The parabolic approximation method," in *Wave Propagation and Underwater Acoustics*, J. Ehlers *et al.*, eds. (Springer-Verlag, Berlin, 1977), p. 272.
23. L. Fishman and J. J. McCoy, "Direct and inverse wave propagation theories and the factorized Helmholtz equation. Path integral representations," Proc. Soc. Photo-Opt. Instrum. Eng. (to be published).
24. M. Born and E. Wolf, *Principles of Optics*, 5th ed. (Pergamon, Oxford, 1975).
25. R. K. Luneburg, *Mathematical Theory of Optics* (U. California Press, Berkeley, Calif., 1964), Sec. 45.
26. E. Lalor, "Conditions for the validity of the angular spectrum of plane waves," J. Opt. Soc. Am. **58**, 1235–1237 (1968).
27. W. H. Southwell, "Validity of the Fresnel approximation in the near field," J. Opt. Soc. Am. **71**, 7–14 (1981).