Intuitive explanation of the phase anomaly of focused light beams

Robert W. Boyd

Institute of Optics, University of Rochester, Rochester, New York 14627

(Received 25 August 1979)

An intuitive argument is presented for the phase anomaly, that is, the 180° phase shift of a light wave in passing through a focus. The treatment is based on the geometrical properties of Gaussian light beams, and suggests a new viewpoint for understanding the origin of the phase shift. Generalizing the argument by including higher-order modes of the light field allows the case of a spherical wave to be treated.

INTRODUCTION

It is well known that a spherical converging light wave undergoes a phase change of 180 degrees in passing through its focus. This *phase anomaly* was first observed by Gouy[1] in 1890, and was explained by him on the basis of Huygens' principle. Gouy[2] also showed that this phase change is predicted for acoustic waves as well, and is, in fact, a general property of any focused wave.

Numerous other authors have also treated this problem. Debye[3] found an exact solution to the wave equation which predicts the phase anomaly at a focus. His solution is valid for all space and thus neatly avoids the problem of applying the proper boundary conditions at the physical aperture that limits the diameter of the converging wave.

Rubinowicz[4] has treated the phase anomaly by use of the theory of the boundary diffraction wave. A light wave in this theory is split into an incident wave and a wave diffracted at the boundary of the aperture. Rubinowicz showed that the phase anomaly is a property of geometrical optics in the sense that it appears in the incident wave and not in the diffracted wave.

Linfoot and Wolf[5] have shown that the 180-degree total phase change results from a rather complicated phase distribution within the focal region. This phase distribution was derived by use of Kirchoff's diffraction theory in terms of the Lommel functions. Linfoot and Wolf also showed that, in large part, the phase anomaly is associated with a lengthening[6] of the wavelength of light in a focal region by the factor $(1 + 1/16f\#^2)$, where $f\#$ is the focal ratio of the converging spherical wave.

These theories provide an adequate theoretical explanation of the phase anomaly. They are all rather mathematical, however. A more intuitive explanation of this phenomenon would serve a useful purpose; it is the goal of this paper to provide such an explanation. The treatment is based on the theory of Gaussian beam propagation. In Sec. I an intuitive explanation is given for the 180-degree phase shift that a Gaussian beam experiences in passing through a focal region. In Sec. II a spherical converging wave is treated as a linear combination of the normal modes of the freely propagating field. In cylindrical coordinates these normal modes have the form of a Gaussian multiplied by a Laguerre polynomial. Since each of these normal modes shows a 180-degree phase shift in passing through the focus, the well-known phase anomaly for the spherical wave is predicted.

I. PHASE ANOMALY FOR A GAUSSIAN BEAM

A. Properties of Gaussian beams

Since many lasers produce output beams in which the transverse intensity distribution is nearly Gaussian, there has been considerable interest in studying the properties of Gaussian beams. Kogelnik and Li[7] have shown that an approximate solution to the scalar wave equation

$$\nabla^2 u + k^2 u = 0 \quad (1)$$

for a beam of Gaussian cross section traveling in the $+z$ direction and centered on the z axis is given by

$$u(r,z) = \frac{w_0}{w(z)} \exp\left[i[kz - \Phi(z)] - r^2\left(\frac{1}{w^2(z)} - \frac{ik}{2R(z)}\right)\right], \quad (2)$$

where $r^2 = x^2 + y^2$, (x,y being Cartesian coordinates in a plane perpendicular to the beam axis) and where $k = 2\pi/\lambda$ with λ the wavelength of light in the medium. The parameter $w(z)$ is loosely called the beam radius because, at fixed z, the modulus of the light amplitude $|u(r,z)|$ falls to $1/e$ of its maximum value at a distance $w(z)$ from the z axis, as depicted in Fig. 1. The beam radius $w(z)$ changes with z according to the formula

$$w^2(z) = w_0^2[1 + (\lambda z/\pi w_0^2)^2]. \quad (3)$$

At $z = 0$, $w(z)$ takes on its minimum value $w(0) \equiv w_0$, and this location is thus called the beam waist. The function $w(z)$, shown in Fig. 2, is a hyperbola with asymptotes inclined to the z axis at the far-field angle

$$\theta_{ff} = \lambda/\pi w_0. \quad (4)$$

The parameter $R(z)$ of Eq. (2) can be interpreted as the radius

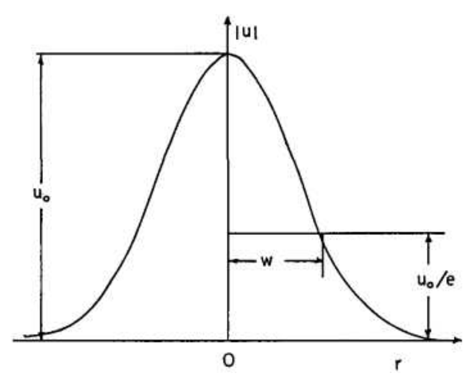

FIG. 1. Amplitude distribution of a Gaussian beam.

0030-3941/80/070877-04$00.50

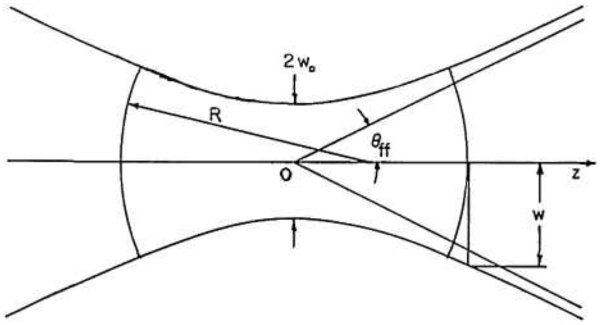

FIG. 2. Profile of a Gaussian beam, showing the change of w with z, the far-field divergence angle θ_{ff}, and the wave-front radius of curvature R.

of curvature of the wave front and is given by

$$R(z) = z[1 + (\pi w_0^2/\lambda z)^2]. \tag{5}$$

The wave fronts are thus plane at the beam waist and in the limit of $z \to \pm\infty$; the wave-front radius of curvature takes on its minimum value of $2\pi w_0^2/\lambda$ at $z = \pi w_0^2/\lambda$. Lastly, the parameter $\Phi(z)$ of Eq. (2) can be interpreted as a phase difference between the Gaussian beam and an infinite plane wave of propagation constant k, also traveling in the $+z$ direction; $\Phi(z)$ is given by

$$\Phi(z) = \arctan(\lambda z/\pi w_0^2). \tag{6}$$

This phase shift increases continuously as the beam passes into and through the beam waist. The total phase shift experienced in passing through the region of the beam waist is the difference in $\Phi(z)$ between the limits $z \to \pm\infty$, and is equal to 180 degrees. This phase shift for a Gaussian beam is the analog of the phase anomaly for a spherical wave.

As can be verified by direct substitution, the Gaussian beam of Eq. (2) is a solution to the wave equation (1) in the limit where

$$e^{ikz}\frac{\partial^2}{\partial z^2}[u(r,z)e^{-ikz}]$$

is much smaller than any other term obtained in evaluating $\nabla^2 u(r,z)$ of Eq. (1). A sufficient condition for this occurrence is $\theta_{ff} \ll 1$, implying that the radiation field is substantially confined to a narrow cone about the z axis.

B. Heuristic explanation of the phase anomaly

The intuitive explanation of the phase anomaly is illustrated by the focused Gaussian beam shown in Fig. 3. The two wave fronts AB and DE are symmetrically located with respect to the beam waist at $z = 0$. According to geometrical optics the optical path length between wave front AB and wave front DE is given by the distance along the straight line BE. In a sense, diffraction causes the light to propagate along the shorter curved path BCD, and thus the optical disturbance at DE is advanced in phase with respect to the value predicted by geometrical optics. The path length BCD is measured along the curve $w(z)$, whose tangent is everywhere parallel to the direction of energy flow. Being curved, $w(z)$ is not a ray path in the sense of geometrical optics, but does possess the property that, for $\theta_{ff} \ll 1$, its tangent is everywhere perpendicular to the surfaces of constant phase of the scalar field $u(r,z)$, and thus distance measured along $w(z)$ is an appropriate measure of the phase of the field for various values of z.

In order to demonstrate that this argument leads to the correct value for the phase anomaly, it is necessary to calculate the difference between path lengths BE and BCD. The path length BCD is given by

$$L = \frac{2\pi w_0^2}{\lambda}\int_0^{z_1/b} dx \left(\frac{1 + (1+\theta_{ff}^2)x^2}{1+x^2}\right)^{1/2}, \tag{7}$$

where $b = \pi w_0^2/\lambda$, and where Eq. (3) for the hyperbola BCD has been used. This integral can be expressed in terms of elliptic integrals as[8]

$$L = 2w_0\sqrt{1+\theta_{ff}^{-2}}\left(\frac{1}{1+\theta_{ff}^2}F(\phi,\kappa) - E(\phi,\kappa) + \frac{z_1\theta_{ff}}{w_0}\sqrt{\frac{(1+\theta_{ff}^{-2})(w_0^2\theta_{ff}^{-2}+z_1^2)}{w_0^2\theta_{ff}^{-4}+z_1^2(1+\theta_{ff}^{-2})}}\right), \tag{8}$$

where $F(\phi,\kappa)$ and $E(\phi,\kappa)$ are the elliptic integrals of the first and second kinds, respectively, and where the parameters ϕ and κ are given by

$$\phi = \sin^{-1}\sqrt{z_1^2/[z_1^2 + w_0^2(\theta_{ff}^2 + \theta_{ff}^4)^{-1}]} \tag{9}$$

and

$$\kappa = \sqrt{\theta_{ff}^2/(1+\theta_{ff}^2)}\,. \tag{10}$$

Similarly, the straight-line distance between points B and E is given by

$$L' = 2z_1\sqrt{1+\theta_{ff}^2 + w_0^2 z_1^{-2}}\,. \tag{11}$$

The phase anomaly $\Delta\Phi$ is then given by

$$\Delta\Phi = \lim_{z_1\to\infty}\frac{2\pi}{\lambda}(L' - L), \tag{12}$$

where the limit expresses the physical requirement that the accumulated phase shift over the entire focal region be evaluated. In the limit $z_1 \to \infty$, the last term of Eq. (8) for L cancels the contribution from L', and the phase anomaly is given by

$$\Delta\Phi = \frac{-4}{\theta_{ff}^2}\sqrt{1+\theta_{ff}^2}\left(\frac{1}{1+\theta_{ff}^2}F(\pi/2,\kappa) - E(\pi/2,\kappa)\right). \tag{13}$$

Since by Eq. (10) the parameter κ depends only on θ, it is apparent that the phase anomaly $\Delta\Phi$ depends only on θ_{ff}. This expression for $\Delta\Phi$ is plotted in Fig. 4 as a function of $(2\theta_{ff})^{-1}$, which can be interpreted roughly as the focal ratio of the

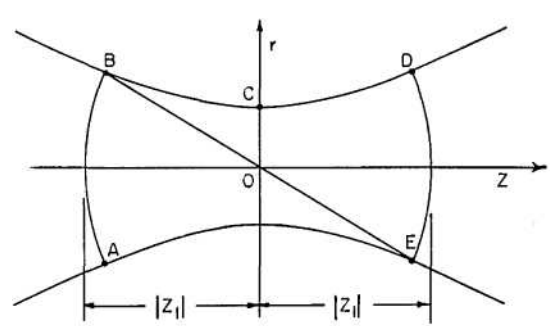

FIG. 3. The wave fronts AB and DE of a Gaussian beam are separated by the optical path length BCD. The difference between the path length BCD and the geometrical separation BE gives rise to the phase anomaly.

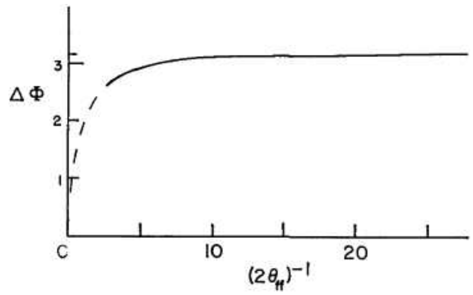

FIG. 4. The phase anomaly $\Delta\Phi$ predicted by the heuristic argument described here is shown as a function of $(2\theta_{ff})^{-1} \equiv f^{\#}$. For large values of $f^{\#}$, the predicted phase anomaly agrees with the actual value of π radians. The broken portion of the curve corresponds to $f^{\#} \lesssim 1$, where the approximations used in the derivation are invalid.

Gaussian beam. The predicted phase shift is nearly π radians for normal beams. In the limit of small θ_{ff}, Eq. (13) approaches the value π, as can be verified by expressing θ_{ff} in terms of k by Eq. (10) and introducing the power-series expansions[9] for $F(\pi/2,k)$ and $E(\pi/2,k)$. For very tightly focused beams, the value of the phase shift predicted by this model differs significantly from π. It is exactly in this limit, however, that the Gaussian-beam solution (2) fails to satisfy the wave equation.

The development until now has been in terms of the properties of the function $w(z)$ defined by arbitrary convention to be the radial distance to the $1/e$ points of $|u(,z)|$. The function $w(z)$ could equally well be defined to be the radial distance to the point where $|u(r,z)|$ falls to any fraction f of its radial maximum $|u(0,z)|$. For any such definition, the argument presented here predicts a 180-degree phase shift, provided that $\theta_{ff} \ll 1$. It is furthermore true that for $\theta_{ff} \ll 1$ the distance BCD between the wave fronts AB and DE is independent of the fraction f used to define $w(z)$, indicating that this distance does correspond to the optical path between these wave fronts.

II. HEURISTIC TREATMENT OF THE PHASE ANOMALY FOR A SPHERICAL CONVERGING BEAM

It will now be shown that the heuristic procedure introduced in Sec. I predicts the phase anomaly not only for Gaussian beams but also for weakly converging beams of arbitrary amplitude distribution, and in particular for a weakly converging spherical wave of circular cross section. For simplicity, only beams with axial symmetry are explicitly considered.

The Gaussian-beam solution given by Eq. (2) can be regarded as the first member of a set of approximate solutions to the wave equation, valid in the limit of $\theta_{ff} \ll 1$, given by[7]

$$u_p(r,z) = L_p\left(\frac{2r^2}{w^2(z)}\right)\frac{w_0}{w(z)}\exp\left[i[kz - \Phi_p(z)] - r^2\left(\frac{1}{w^2(z)} - \frac{ik}{2R(z)}\right)\right], \quad (14)$$

where $L_p(x)$ is the Laguerre polynomial of order p, where $w(z)$ and $R(z)$ are still given by Eqs. (2) and (4), respectively, and where the phase change $\Phi_p(z)$ is now given by

$$\Phi_p(z) = (2p+1)\arctan(\lambda z/\pi w_0^2). \quad (15)$$

The total phase shift for propagation from $-\infty$ to $+\infty$ is thus $(2p+1)\pi$ rad. The contribution of π rad, independent of p, can be understood by the heuristic argument presented in Sec. I, since each solution (14) has a beam profile characterized by $w(z)$ of Eq. (2). The additional phase shift of $2p\pi$ rad, which is unimportant in determining the phase anomaly, can be understood intuitively as a π phase shift occurring each time the geometrical ray BE of Fig. 3 passes through one of the zeros of $u_p(r,z)$.

The set of functions $u_p(r,z)$ constitutes the normal modes of the freely propagating light field in the sense that any monochromatic, axially symmetric, scalar field distribution of small far-field diffraction angle can be expanded in terms of these functions. This follows from the fact that the Laguerre polynomials form a complete set[10] such that any function $f(x)$ defined on the interval $0 \le x < \infty$ can be expressed in the form

$$f(x) = \sum_{p=0}^{\infty} C_p L_p(x). \quad (16)$$

It is, therefore, possible to expand a spherical wave in terms of the normal modes defined by Eq. (14).

A spherical wave front of half-angle θ converging toward the origin from $z < 0$ can be expressed as

$$u_s(r,z) = \begin{cases} \dfrac{A}{\sqrt{r^2+z^2}} e^{-ik\sqrt{r^2+z^2}} & \dfrac{r}{|z|} \le \theta \\ 0 & \dfrac{r}{|z|} > \theta \end{cases} \quad \text{for } z \ll 0, \quad (17)$$

where the positive square root is to be taken here and below. This spherical wave $u_s(r,z)$ can be expanded in terms of the functions $u_p(r,z)$ in the plane z_1 = constant for $\lambda z_1/\pi w_0^2 \ll -1$; $u_s(r,z)$ then takes the form

$$u_s(r,z) = \sum_p C_p u_p(r,z), \quad (18)$$

where explicit expressions for the expansion parameters C_p are not needed. If Eq. (18) is now evaluated at large positive z, each of the $u_p(r,z)$ will have undergone a phase shift of 180° for the reason previously given, and the spherical wave $u_s(r,z)$ becomes

$$u_x(r,z) = \begin{cases} \dfrac{-A}{\sqrt{r^2+z^2}} e^{ik\sqrt{r^2+z^2}} & \dfrac{r}{|z|} \le \theta \\ 0 & \dfrac{r}{|z|} > \theta \end{cases} \quad \text{for } z \gg 0, \quad (19)$$

which differs by a minus sign from the result predicted from geometrical considerations. The anomalous phase shift for a spherical wave of circular cross section is thus seen to be understandable in terms of the simple physical picture presented here.

ACKNOWLEDGMENTS

Discussion of these matters with N. Erickson, C. R. Stroud, Jr., and E. Wolf are gratefully acknowledged. This work was supported in part by NSF Grants SER77-06914 and ENG79-08038.

[1]Gouy, C. R. Acad. Sci. Paris **110,** 1251–1253 (1890).
[2]Gouy, Ann. Chim. Phys. **6, XXIV,** 145–213 (1891).
[3]P. Debye, Ann. Phys. **30,** 755 (1909). A discussion of Debye's work can be found in J. Picht, *Optische Abbildung* (Vieweg & Sohn, Braunschweig, 1931). References to much of the early work in this field, including that of F. Reiche and K. Schwarzschild, can also be found in this book.
[4]A. Rubinowicz, "On the anomalous propagation of phase in the focus," Phys. Rev. **54,** 931–936 (1938).
[5]E. H. Linfoot and E. Wolf, "Phase Distribution near Focus in an Aberration-free Diffraction Image," Proc. Phys. Soc. London **69,** 823–832 (1956).
[6]Due to an algebraic error, the original paper states that the wavelength is decreased by this factor (E. Wolf, private communication).
[7]H. Kogelnik and T. Li, "Laser Beams and Resonators," Appl. Opt. **5,** 1550–1567 (1966).
[8]P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists* (Springer-Verlag, Berlin, 1954), p. 3.
[9]M. Abramowitz and I. A. Segen, *Handbook of Mathematical Functions* (Dover, New York, 1965), Eqs. (17.3.11) and (17.3.12).
[10]N. N. Lebedev, *Special Functions and their Applications* (Dover, New York, 1972), p. 88.

Inversion of the nonlinear equations of reflection ellipsometry for uniaxial crystals in symmetrical orientations

M. Elshazly-Zaghloul
Electrical Engineering Department, Faculty of Engineering, Cairo University, Cairo, Egypt

R. M. A. Azzam
Department of Electrical Engineering, School of Engineering, University of New Orleans, Lakefront, New Orleans, Louisiana 70122

(Received 15 June 1979; revised 7 January 1980)

The complex ordinary (N_o) and extraordinary (N_e) refractive indices of an absorbing uniaxial crystal can be determined using reflection ellipsometry. The measurements are taken with the optic axis parallel and perpendicular to the crystal's surface. The equations obtained are solved without resort to iterative methods; N_o and N_e are determined separately. Sixteen solution sets (N_o, N_e) are obtained and the correct solution can be easily identified. We present an optimum angle of incidence that minimizes the relative errors in N_o and N_e.

INTRODUCTION

With the increased interest in uniaxial crystals and their applications, e.g., in electro-optics, the need arises for accurate determination of their optical properties. One of the powerful tools for this purpose is ellipsometry. By studying the change of the polarization state of light upon reflection from the surface of a crystal, the optical properties of the crystal can be determined.

Until now the equations relating the optical properties of a uniaxial crystal and the quantities measured by a conventional ellipsometer were solved using iterative numerical methods. In this paper we present a new inversion method for determining separately the ordinary (N_o) and extraordinary (N_e) refractive indices. An eighth-degree polynomial in N_o is first obtained whose coefficients are determined from the measured data. This polynomial is solved using a new numerical method that is not iterative in nature.[1] Knowing the solution for N_o, we obtain N_e by direct substitution.[2] A detailed example in which this method is applied to a GaSe crystal is presented. Also, using a simulated error analysis, we determine the range of angles of incidence ϕ with the least percent error in N_o and N_e caused by an error of ϕ for ice, calcite, and GaSe crystals.

I. METHOD

For a uniaxially symmetric crystal, two relatively simple cases can be distinguished depending on whether the optic axis (axis of symmetry) is in the plane of incidence or perpendicular to it.[3–5] With respect to the first case, consider only the configuration where the optic axis is perpendicular to the reflecting crystal surface. We take the optic axis parallel to the z axis of a Cartesian xyz coordinate system. The optical constants $N_o(N_x)$ and $N_e(N_z)$ of the crystal may then be written as

$$N_o = N_x = N_y = n_o - jk_o, \tag{1a}$$

$$N_e = N_z = n_e - jk_e. \tag{1b}$$

When the z axis is perpendicular to the crystal surface, the Fresnel reflection coefficients take the form

$$r_p = (N_e\, N_o \cos\phi - n\, X_e)/(N_e\, N_o \cos\phi + n\, X_e), \tag{2a}$$

$$r_s = (n \cos\phi - X_o)/(n \cos\phi + X_o), \tag{2b}$$

where ϕ is the angle of incidence, n is the refractive index of the ambient, and

$$X_e = (N_e^2 - n^2 \sin^2\phi)^{1/2}, \tag{3a}$$

 0030-3941/80/070880-04$00.50