Propagation of generally astigmatic Gaussian beams along skew ray paths

Alan W. Greynolds

Breault Research Organization, Inc.
4601 East First Street, Tucson, AZ 85711

Abstract

The propagation of a Gaussian beam through an asymmetric optical system results in a generally or "twisted"" astigmatic beam whose elliptical spot rotates as it propagates in free space. A general mathematical form for this beam is derived by applying the angular spectrum approach to the paraxialized wave equation. This result is compared to those derived by other researchers and it is found that the beam can be equivalently represented by a set of paraxial rays traced about the central beam ray. The application of general Gaussian beam propagation to the analysis of non-Gaussian beam propagation in arbitrary optical systems is also discussed.

Introduction

One would probably assume from the title of this paper that the author was motivated by the need to analyze optical systems that steer, focus, and/or shape only Gaussian beams. Actually, the primary thrust of the research summarized herein is to show that the Gaussian beam can be used more importantly as a general tool in the diffraction analysis of arbitrary wavefronts in any optical system. The Gaussian beam can be considered as a finite extension of the zero-width ray which has been used for centuries as the basic tool for the design and analysis of optical instruments. This is not to say that Gaussian beam tracing will eventually make ray tracing obsolete since ironically the most practical way to propagate Gaussian beams is by tracing a set of closely spaced rays.

System diffraction calculations by the standard ray trace method

The standard method used for the past two decades to calculate the diffraction image in an optical instrument requires optical path data generated by tracing hundreds of rays. The phase of the wavefront must be known on a rectangular grid of points at the exit pupil in order to employ a discrete Fourier transform (DFT) which yields the field at the image plane. This is accomplished either by interpolating the optical path difference (OPD) data into a rectangular grid or by iteratively aiming the rays so that their intersections with the exit pupil form a uniform array of points. Both methods require some computational overhead. If a direct implementation of the DFT is used, then the calculation becomes even more time-consuming. What made this method both practical and popular was the advent of the Fast Fourier Transform (FFT) algorithm which could perform the required two-dimensional transform hundreds of times faster. However, the annoying aliasing problems associated with applying the DFT to non-periodic functions are still present with the FFT.

The standard method is limited to the determination of the diffraction image on a plane close to the focus and nearly perpendicular to the axis of the imaging light cone. Also, the system must have a well-defined exit pupil. These requirements exclude calculating the diffracted field at any point in both imaging and non-imaging systems by this method. It was this goal that prompted the author to seek a more general method for system diffraction calculations.

An alternative approach using decomposition into and tracing of elementary beams

The electromagnetic field is linear, i.e. it obeys the principle of superposition. Therefore, an arbitrary field incident upon a system could be decomposed into some set of elementary fields. As opposed to an infinitesimal ray, these elementary fields would be of non-zero extent so that, in general, they would overlap and interfere with each other. These fields could then be individually propagated (including self-diffraction effects) through the system so that the total field at any point "downstream" is found by recombining the new values for the elementary fields. This concept of decomposition, propagation, and recombination is not new since this is precisely what is done in the angular spectrum method where the elementary field is a plane wave, i.e. the incident field is decomposed into plane waves (Fourier transformed), propagated by a diffraction transfer function, and recombined (inverse Fourier transformed). The angular spectrum method works well with the propagation of fields in free space but cannot be easily applied to general optical systems with bounded non-planar refracting (or reflecting) surfaces. This is mostly due to the infinite extent of the plane waves. If we knew how to propagate some other elementary field that is of

finite extent so that the effects of surface interactions are localized and therefore calculable, then we would have a very general approach to system diffraction. One such field (and the simplest) is the fundamental Gaussian beam mode.

Desirable features of Gaussian beams

There are several features of Gaussian beams that make them the best choice for the proposed technique for system diffraction calculations by beam decomposition. First and foremost, as shall be shown in later sections, Gaussian beams are relatively easy to propagate through any optical system. This is due to the well-known fact that they do not change mathematical form as they propagate, i.e. they are fundamental solutions of the wave equation within the Fresnel approximation region. This in turn can be directly attributable to the fact that the Gaussian function (reference 1) defined in equation (1) is perfectly smooth, i.e. all derivatives are continuous.

$$Gaus(x) = e^{-\pi x^2} \tag{1}$$

Another important feature of Gaussians is their relative compactness. Mathematically the Gaussian never drops to precisely zero (which would imply an infinite width). However, as we can see from Table 1, it drops off so rapidly that for all intent and purposes it can be considered of finite width. For example, at x=3 it has fallen nearly 13 orders of magnitude. If we now square the function as we need to do in order to get the intensity of the field, then this last point has an equivalent intensity value of around 1.E-25. Since this is so close to the lower limit of most single precision floating point number representations on digital computers, it can be considered to be zero and therefore, for all practical purposes our Gaussian has a finite width.

Table 1. Illustration of the Rapid Falloff of a Gaussian function

x	Gaus(x)
0.0	1.0
0.5	0.456
1.0	0.0432
1.5	8.5E-4
2.0	3.5E-6
2.5	3.0E-9
3.0	5.3E-13

A Brief review of the different Gaussian beam forms

The proposed method for system diffraction calculations hinges on our ability to propagate Gaussian beams in any optical system. There has been considerable work done on Gaussian beam propagation since the invention of the laser. Let us briefly review the current state of affairs for the free-space propagation case only.

The simplest Gaussian beam is rotationally symmetric around its propagation axis which we will designate as the z axis in a Cartesian coordinate system. The equation for the field of this beam can be written most compactly in the following form:

$$u = \frac{e^{i2\pi z}}{q(z)} \cdot e^{-\pi\left[\frac{x^2+y^2}{q(z)}\right]}, \quad q(z) = w_o^2 - iz \tag{2}$$

where we have conveniently normalized to unity wavelength. This has exactly the same form as a quadratically approximated spherical wave with a complex axial focal position given by the complex beam parameter q. One can easily separate 1/q into real and imaginary parts and derive the more familiar equations for the width and curvature, respectively, of the beam as a function of axial position z. This type of beam can only exist on the axis of a rotationally symmetric optical system.

The next step up in complexity would be what is normally referred to as an simply astigmatic or orthogonal beam. Due to the separability of the Gaussian function, it can be written essentially as the product of two one-dimensional beams:

$$u = \frac{e^{i2\pi z}}{\sqrt{q_x q_y}} \cdot e^{-\pi\left[\frac{x^2}{q_x} + \frac{y^2}{q_y}\right]} \qquad \begin{cases} q_x = w_x^2 - i(z - z_x) \\ q_y = w_y^2 - i(z - z_y) \end{cases} \tag{3}$$

Now the two complex beam parameters resemble the complex positions of the two perpendicular focal lines of a wave with simple astigmatism. This expression can only be used to describe a beam whose axis lies in a plane of symmetry of the optical system.

In order to completely characterize any optical system, we need to be able to propagate the generally astigmatic Gaussian beams that would occur along skew or nonorthogonal paths. The simplest mathematical form for this beam is:

$$u = e^{i2\pi z} \cdot \sqrt{d} \cdot e^{-\pi[ax^2+2bxy+cy^2]} \tag{4}$$

Here the four complex beam parameters a,b,c,d are functions of z that need to be determined. The d parameter contains the on-axis amplitude and phase shift of the beam. The real parts of the other beam parameters in the Gaussian argument will obviously determine the lateral variation of the amplitude (modulus) of the wave. The contours of amplitude variation will always be elliptical curves if the beam is to be finite. Likewise, the imaginary parts will specify the phase variation of the wavefront, whose contours may be either ellipses of hyberpolas depending on the signs of the principal curvatures of the wavefront.

The general Gaussian argument can also be written in matrix notation as a quadratic form.

$$ax^2+2bxy+cy^2 = (x\ y)\begin{pmatrix} a & b \\ b & c \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \text{constant} \tag{5}$$

Setting this to a constant yields the equation of one contour. The orientation of this contour curve is specified by the following rotation angle (see Figure 1).

$$\theta = \frac{1}{2}\tan^{-1}\left(\frac{2b}{a-c}\right) \tag{6}$$

There are two quantities that are rotationally invariant, i.e. they are independent of the particular orientation of the Cartesian coordinate system (reference 2).

$$a+c \quad \& \quad d = \begin{vmatrix} a & b \\ b & c \end{vmatrix} = ac-b^2 \begin{matrix} = 0 & \text{Parabola} \\ < 0 & \text{Hyperbola} \\ > 0 & \text{Ellipse} \end{matrix} \tag{7}$$

Theses two relationships can be used to derive the equations for the orthogonal coefficients that result when the coordinate system is rotated such that the cross term vanishes.

$$\left.\begin{matrix} a' \\ c' \end{matrix}\right\} = \frac{1}{2}\left[(a+c) \pm \sqrt{(a+c)^2-4(ac-b^2)}\right]$$
$$= \frac{1}{2}\left[(a+c) \pm (a-c)\sqrt{1+\left(\frac{2b}{a-c}\right)^2}\right] \tag{8}$$

Derivation of general Gaussian beam equations

Since the scalar harmonic optical field must be a solution of the Helmholtz wave equation,

$$[\nabla^2 + (2\pi)^2]u = 0 \tag{9}$$

it would seem that the most straight-forward way to to determine the axial dependence of the the general Gaussian beam parameters would be to substitute (4) into (9) and solve the resulting equations. However, we find that the Gaussian is not a rigorous fundamental solution of the exact wave equation. Instead, we will apply the angular spectrum method of solution to an approximate form of the wave equation.

If we confine ourselves to fields that propagate mostly in one particular direction, i.e. fields that do not differ a great deal from a plane wave, then we can factor the plane wave dependence out to get a new wave function which we will call the reduced field.

$$u(x,y;z) = \Psi(x,y;z)\, e^{i2\pi z} \tag{10}$$

Note that this plane wave factor was common to all the different forms of the Gaussian beam. Substituting this into the our original wave equation, we get a new wave equation.

$$\left[\nabla^2 + i4\pi\frac{\partial}{\partial z}\right]\Psi = 0 \tag{11}$$

Since most of the variation of the original wave field with axial position is contained in factored out plane wave term, we can assume all variations of the reduced field with axial position z are relatively slow, i.e.

$$\left|\frac{\partial^2\Psi}{\partial z^2}\right| << \left|\frac{\partial\Psi}{\partial z}\right| \tag{12}$$

Neglecting the second derivative term, we are left with what is called the paraxialized wave equation (reference 3).

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + i4\pi\frac{\partial}{\partial z}\right]\Psi = 0 \tag{13}$$

This wave equation is "paraxial" in the sense that due to the approximation used, it is limited to making accurate predictions of fields not too far from the axis.

General solution by angular spectrum approach

Let us assume that the reduced field at z=0 can be represented by a two-dimensional Fourier integral, i.e. a continuous angular spectrum of plane waves.

$$\Psi(x,y;z) \overset{\mathcal{F}}{\Leftrightarrow} \Psi(\alpha,\beta;z)$$
$$\begin{cases} \Psi = \iint_{-\infty}^{\infty} \Psi\, e^{i2\pi(\alpha x+\beta y)}\, d\alpha\, d\beta \\ \Psi = \iint_{-\infty}^{\infty} \Psi\, e^{-i2\pi(\alpha x+\beta y)}\, dx\, dy \end{cases} \tag{14}$$

Substituting this expression into the paraxialized wave equation and using the Fourier representation of the two-dimensional Laplacian, we obtain an ordinary differential equation for the spectrum as a function of axial position z.

$$\left[-(4\pi)^2 + i4\pi\frac{d}{dz}\right]\Psi = 0 \tag{15}$$

This equation has the general solution at any axial position given the initial spectrum at z=0:

$$\Psi(\alpha,\beta;z) = \Psi(\alpha,\beta;0)\, e^{-i\pi z(\alpha^2+\beta^2)} \tag{16}$$

Therefore, the spectrum of the reduced field is propagated form one plane to another by multiplication with a "transfer" function. The reduced field itself is then obtained by another Fourier Transform operation.

If the angular spectrum approach had been applied to the exact wave equation for the actual field, the following result would have been obtained (reference 4).

$$U(\alpha,\beta;z) = U(\alpha,\beta;0)\, e^{i2\pi z\sqrt{1-(\alpha^2+\beta^2)}} \tag{17}$$

If the square root in the exponent of the transfer function is expanded in a power series about the axis and only the quadratic terms are kept, then we obtain exactly the same result we did using the paraxialized wave equation.

$$U(\alpha,\beta;z) \approx U(\alpha,\beta;0)\, e^{i2\pi z}\, e^{-i\pi z(\alpha^2+\beta^2)} \tag{18}$$

This transfer function corresponds to that of diffraction within the Fresnel region. In other words, the Fresnel approximation applied to the solution of the exact wave equation is equivalent to the "exact" solution of the paraxialized wave equation.

Particular solution for the generally astigmatic (nonorthogonal) Gaussian beam

Now we can address the problem of interest; the propagation of a general Gaussian beam. The only piece we are missing is an expression for the Fourier Transform of our general Gaussian function. By direct integration, we obtain the following Fourier transform pair:

$$\sqrt{d_o}\, e^{-\pi[a_o x^2 + 2b_o xy + c_o y^2]} \overset{\mathcal{F}}{\Leftrightarrow} e^{-\pi\left[\frac{c_o\alpha^2 - 2b_o\alpha\beta + a_o\beta^2}{d_o}\right]} \tag{19}$$

$$d_o = a_o c_o - b_o^2 \qquad \mathrm{Re}\{a_o, c_o\} > 0$$

The conditions that the real parts of the exponent parameters be positive assures that the volume under the function is finite, a requirement for the existence of this Fourier integral relationship and a confined beam.

We can now write down the spectrum of the Gaussian field at some plane given its spectrum at z=0.

$$\Psi(\alpha,\beta;z) = e^{-\pi[C\alpha^2 - 2B\alpha\beta + A\beta^2]}$$
$$A = \frac{a_o}{d_o} + iz \;,\; B = \frac{b_o}{d_o} \;,\; C = \frac{c_o}{d_o} + iz \;,\; D = AC - B^2 \qquad (20)$$

Applying the Fourier transform to this expression, we obtain the actual field, i.e. explicit expressions for the four beam parameters at any axial position given their complex values at z=0.

$$u(x,y;z) = e^{i2\pi z}\sqrt{d}\; e^{-\pi[ax^2 + 2bxy + cy^2]}$$
$$\begin{cases} a = \frac{A}{D} = \frac{a_o + id_o z}{d_o D} & b = \frac{B}{D} = \frac{b_o}{d_o D} \quad d = ac - b^2 = \frac{1}{D} \\ c = \frac{C}{D} = \frac{c_o + id_o z}{d_o D} & d_o D = (1 + ia_o z)(1 + ic_o z) + b_o^2 z^2 \end{cases} \qquad (21)$$

From the above relationships, it follows that:

$$\frac{b}{a-c} = \frac{b_o}{a_o - c_o} \qquad (22)$$

In other words, this quantity is invariant during propagation and corresponds to a complex rotation of the beam ellipses. Arnaud and Kogelnik (reference 5) arrived at the same result by applying a coordinate rotation to equation (3) for the orthogonal beam, which remains a solution of the paraxialized wave equation even when the angle of rotation is a complex quantity. The fact that this complex rotation is invariant with propagation does not mean that the beam itself does not rotate. In fact, if we calculate this quantity for the real and imaginary parts of the beam parameters separately, we find that the amplitude and phase ellipses, respectively, do in general rotate at different rates as the Gaussian beam propagates. This "twisting" behavior will be demonstrated in a later section.

Although somewhat more complicated than a simple orthogonal Gaussian, the expressions for the complex beam parameters of the general beam are still relatively simple. However, separating these functions into real and imaginary parts is a tedious task if done manually. The author chose an easier route and used the symbolic math program MACSYMA to do this algebra. The resulting expressions are lengthy, involving dozens of terms. Thus, it is difficult to gain any insight into how the amplitude and phase components of the Gaussian vary with propagation because of this and the inherent coupling between them. Simple Gaussian beam propagation can be represented quite easily by a variety graphical techniques, e.g. circle diagram (Smith chart), lateral focii, and the axial projection of a single skew paraxial ray (reference 6). The author attempted to extend these graphical aids to the general Gaussian, but no convenient representation that simplified the understanding of the propagation process could be found. Unfortunately, in this case understanding will have to take a back seat to our primary goal of performing numerical calculations. We will just have to be content to program the equations, letting the computer perform the complex number arithmetic for us.

Refraction of the general Gaussian beam by an optical surface

Ray tracing involves the repeated application of a two-step process of transfer and refraction. The same two operations are also required in the tracing of beams and wavefronts. We have already derived the equations of transfer for the general Gaussian beam, i.e. we can propagate these beams from one refracting surface to another in a homogeneous isotropic medium. We must now determine a method for refracting the beams at an interface surface between two different media.

General characteristics of the refraction process

Only the phase variation of the beam will be altered during refraction, the amplitude variation will remain unchanged. Any terms to first order are automatically taken care of by the refraction of the central beam ray, i.e. the local axis of our Gaussian. The Gaussian component contains only wavefront terms of second order. These terms specify the magnitude and directions of the principal curvatures of the wavefront. It follows that only the local second order properties of the refracting surface will be involved. If the size of the Gaussian beam is so large that higher order terms would be introduced during refraction, then the paraxial approximation is violated.

Assuming that the beam size is small enough so that the paraxial approximation holds in a

localized region around the point of refraction, we need only find a method for determining the magnitude and directions of the principal curvatures of the refracted beam given the same data for the incident wavefront and the refracting surface. Such methods do exist (see reference 7). However, as to be expected, the equations are somewhat more complex than those for straight ray tracing. In addition, for asymmetric non-spherical surfaces, the calculation of the surface curvatures is a non-trivial matter, involving all the second order partial derivatives of the surface function. These complexities lead the author to search for other techniques. In the next section we will look at a more practical approach that uses standard ray tracing techniques to both propagate and refract the general Gaussian beam.

Arnaud's ray equivalent method for the general Gaussian

In the late 1960's, Arnaud (reference 8) discovered that the propagation of a rotationally symmetric Gaussian beam could be represented by the tracing of a complex ray. The real part of this ray is a paraxial ray with zero height at the Gaussian waist and a ray angle equal to the far field divergence angle of the beam. The imaginary part is a paraxial ray parallel to the axis with a height equal to the width of the beam at the waist. These two rays, called the divergence and waist rays, respectively, bear a close resemblance to the standard chief and marginal rays in first-order optical design and as such can also the thought of as the orthogonal components of a skew ray. If this skew ray is rotated about the axis, it forms the hyperbolic surface corresponding to the width of the beam at any axial location.

Later, Arnuad extended this ray equivalent approach to the generally astigmatic Gaussian beam (reference 9). He started by noting that a general astigmatic geometrical wave is uniquely determined by two rays that are always normal to the wavefront. The equations of these two rays in a two-dimensional transverse coordinate system can be written in the following vector form:

$$\begin{aligned} \vec{h}_1(z) &= \vec{h}_1(0) + \vec{u}_1 z \\ \vec{h}_2(z) &= \vec{h}_2(0) + \vec{u}_2 z \end{aligned} \tag{23}$$

These equations must represent two unique rays of the astigmatic bundle and therefore, must satisfy the the following condition at all axial locations, i.e. their Lagrange invariant must be zero.

$$\vec{h}_1 \cdot \vec{u}_2 - \vec{h}_2 \cdot \vec{u}_1 = 0 \tag{24}$$

This geometrical wave can be written in a form similar to our general Gaussian.

$$u = e^{ikz} \cdot \sqrt{d} \cdot e^{ik/2\,[ax^2 + 2bxy + cy^2]} \tag{25}$$

Here we have reintroduced the axial phase factor and an explicit wavelength (contained in the wave number k). The undetermined coefficients are found by noting that the ray directions are proportional to the derivative of the optical path function. The coefficients are therefore solutions of the following matrix equation.

$$\begin{pmatrix} \vec{h}_1 & \vec{h}_2 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} = \begin{pmatrix} \vec{u}_1 & \vec{u}_2 \end{pmatrix} \tag{26}$$

We will not reproduce the solutions here (see reference 9). It is important to note that the solutions depend on our selection of the transverse coordinate directions. However, there are two quantities that do not depend on the coordinate system, i.e. they can be written entirely in vector form.

$$\begin{aligned} d &= |\vec{h}_1 \times \vec{h}_2|^{-1} \\ a + c &= d\,[\,|\vec{h}_1 \times \vec{u}_2| - |\vec{h}_2 \times \vec{u}_1|\,] \end{aligned} \tag{27}$$

These two quantities are the same rotational invariants in equation (7).

Now one might ask what this geometrical wave has to do with Gaussian beam propagation. First, the wave is a solution of the paraxialized wave equation. Second, this is true even if the positions and directions of the two rays are complex quantities. Therefore, if the two rays are complex and are still solutions of the paraxial ray trace equations such that their Lagrange invariant is zero, then this homogeneous geometrical wave equation is transformed into a ray equivalent expression for the general inhomogeneous Gaussian beam. The beauty of this approach is that the propagation and refraction of the diffracting

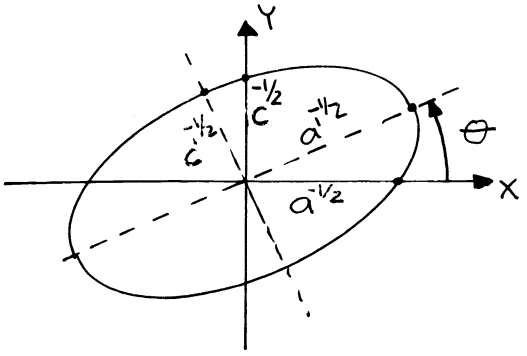

Figure 1. The general elliptical curve.

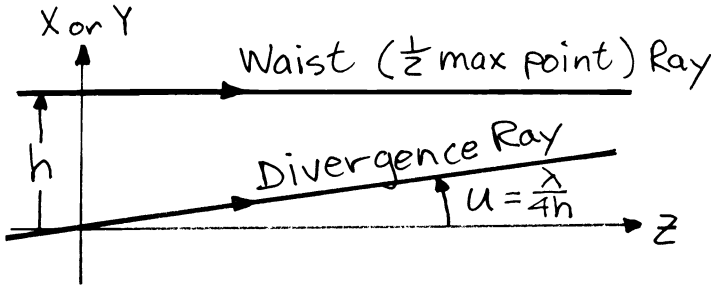

Figure 2. The waist and divergence rays.

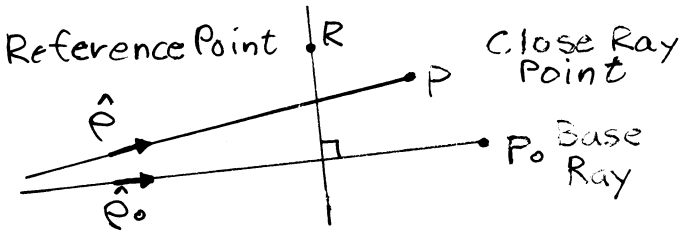

Figure 3. Projection of ray vectors.

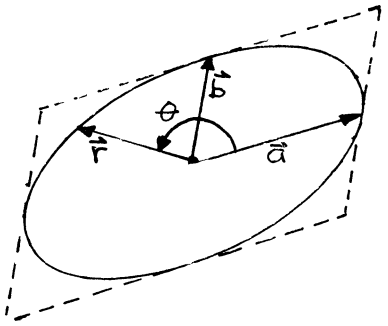

Figure 4. Vector representation of an ellipse.

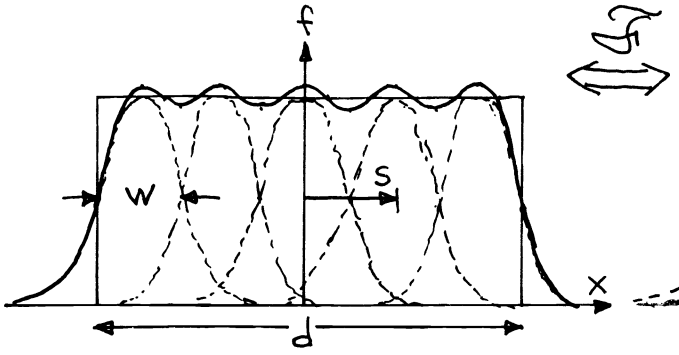

Figure 7. Decomposition of an aperture into Gaussians.

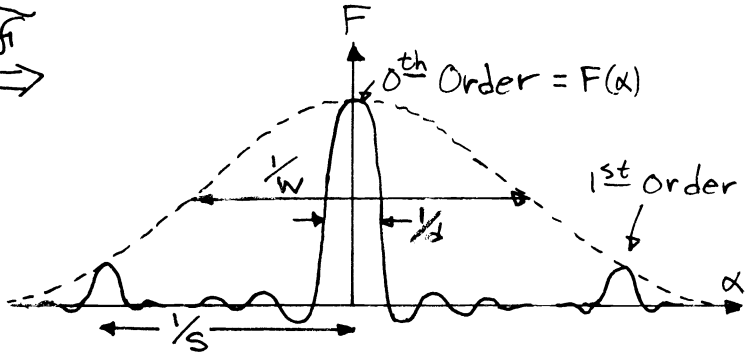

Figure 8. Corresponding focal plane field.

Gaussian beam is automatically taken into account by the propagation and refraction of the geometrical rays that describe it.

Complex ray set for tracing Gaussians

The only requirements placed on the complex rays is that their real and imaginary parts represent actual rays and that they are unique but still have a complex Lagrange invariant that vanishes. A particular convenient choice for these four rays leads to an orthogonal Gaussian beam (reference 10). The real parts are the divergence rays in the two orthogonal directions. The imaginary parts are the corresponding waist rays (see Figure 2). Although these rays initially describe an orthogonal beam, once they are traced along a nonorthogonal or skew path, they describe a generally astigmatic Gaussian.

Transformation of 3D ray data to 2D coordinate system

To implement the ray equivalent approach, we require the ability to trace rays, in particular, the central base ray and four additional neighboring rays. Ray tracing algorithms and computer codes abound. One can choose to either trace the neighboring rays exactly or as first order variations of the base ray if this is more efficient. In the former case, we must transform the three-dimensional ray data into a two-dimensional coordinate system centered on the base ray. This procedure involves first defining a reference plane normal to the base ray that passes through a chosen reference point (see Figure 3). The three-dimensional vector equation of this plane is given by:

$$\hat{\rho}_o \cdot (\vec{r} - \vec{R}) = 0 \tag{28}$$

Each ray is represented by a parametric vector equation.

$$\vec{r} = \vec{p} + \hat{\rho} d \tag{29}$$

Next we find the intersections of the rays with this plane by substituting the ray into the plane equation and solving for the intersection distance.

$$d = - \frac{\hat{\rho}_o \cdot (\vec{p} - \vec{R})}{\hat{\rho}_o \cdot \hat{\rho}} \tag{30}$$

The ray positions in the local coordinate system are found by simply subtracting the base ray position from each one.

$$\vec{h}_i = \vec{r}_i - \vec{r}_o \qquad \vec{r}_i = \vec{p}_i + \hat{\rho}_i d_i \tag{31}$$

The local ray directions are found by projecting them onto the reference plane.

$$\vec{u}_i = \hat{\rho}_i - (\hat{\rho}_i \cdot \hat{\rho}_o) \hat{\rho}_o \tag{32}$$

The only thing left to do is project these vector quantities onto the local coordinate directions (specified by arbitrarily letting the reference point lie on the local x axis).

$$\begin{aligned} \vec{X} &= |X| \hat{X} = \vec{R} - \vec{r}_o \\ \hat{Y} &= \hat{Z} \times \hat{X} \qquad \hat{Z} = \hat{\rho}_o \end{aligned} \tag{33}$$

We now have the local two-dimensional ray data needed to describe the Gaussian beam in the the reference plane.

Gaussian beam amplitude pattern as the convolution of two geometrical patterns

Before proceeding onto some numerical examples, let us look at an alternative mathematical form for the Gaussian that not only simplifies the drawing of the elliptical beam cross sections but also leads to a different way of thinking about general Gaussian beam propagation.

We already pointed out that the argument of our general Gaussian beam expression is a scalar equation of a general ellipse in two dimensions. An ellipse can also be represented by a parametric vector equation of the form:

$$\vec{r} = \vec{a} \cos\theta + \vec{b} \sin\theta \tag{34}$$

The two vectors specifying the ellipse can be thought of as the generating vectors of a parallelogram that contains the ellipse (see Figure 4). The parametric parameter is an angle measured from the first vector. This is the most convenient form of the ellipse for plotting purposes. We simply vary the angle form 0 to 360 degrees to move our "pen" all the

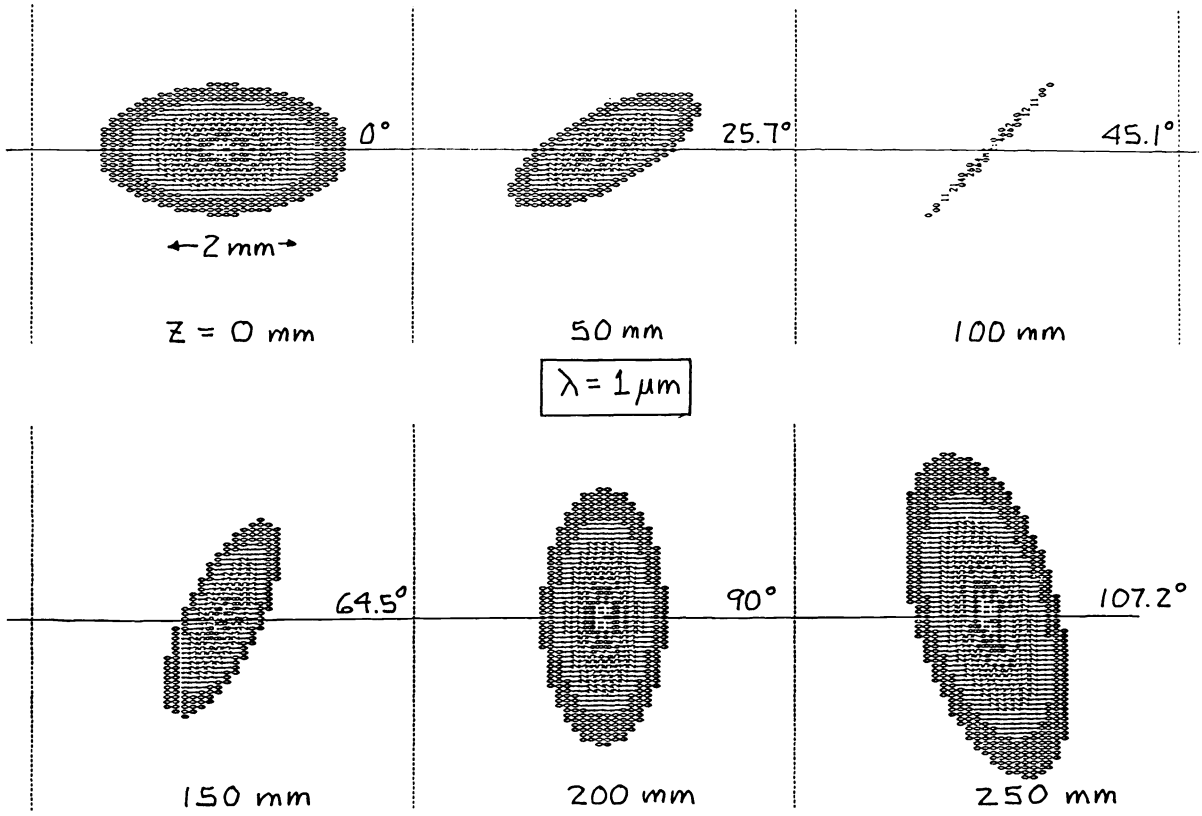

Figure 5. Propagation of a twisted astigmatic Gaussian beam at a wavelength of one micron.

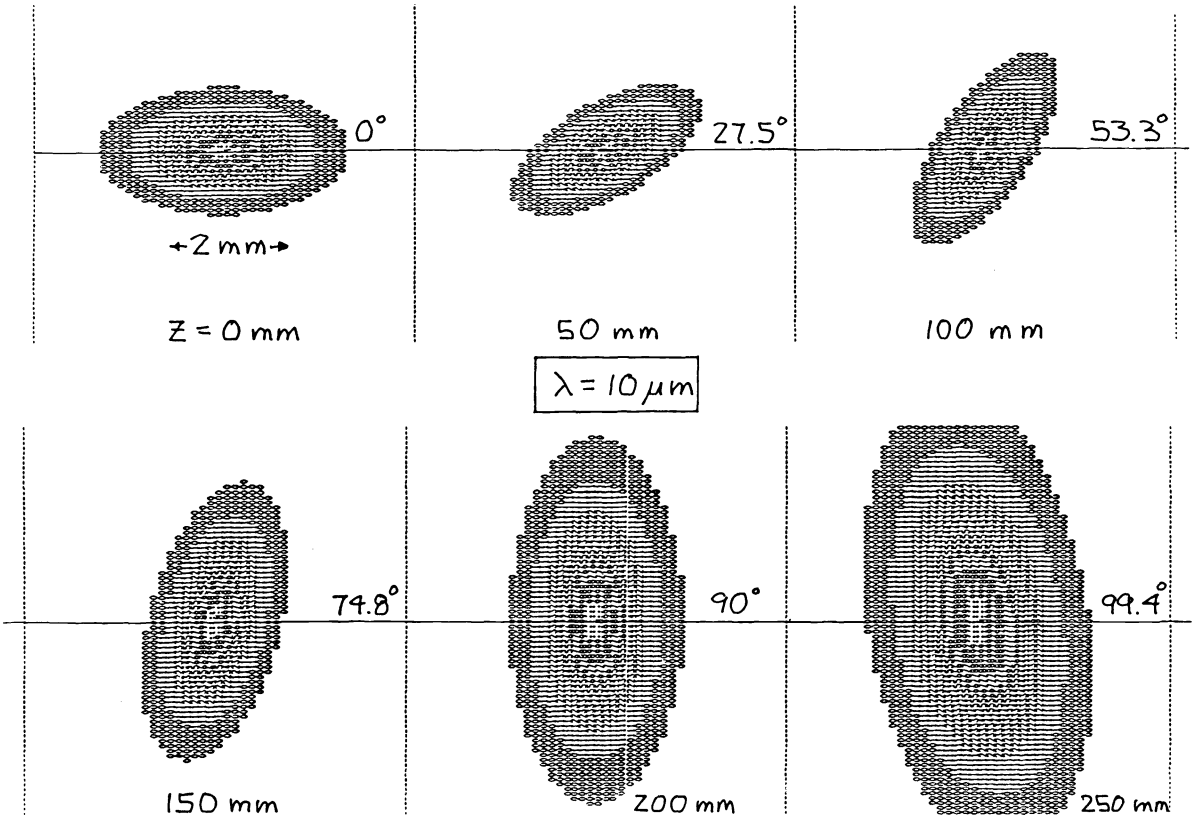

Figure 6. Propagation of twisted astigmatic Gaussian at a wavelength of ten microns.

way around the elliptical curve. Since the two two-dimensional vectors are specified by four scalar components and a general ellipse requires only three scalar quantities, we see that there are an infinite number of vector pairs that could be used to represent the same ellipse.

We can eliminate the angle parameter by taking the vector cross product of equation (34) with the first vector then the second, squaring the two resulting equations, and adding them. Therefore, a general ellipse can also be represented by the following expression:

$$(\vec{a}\times\vec{r})^2+(\vec{b}\times\vec{r})^2=(\vec{a}\times\vec{b})^2 \tag{35}$$

It should be noted that pi times the cross product on the right is equal to the area of the ellipse so that the vectors can not be parallel.

Besides being independent of the coordinate system, the real advantage to using the vector expressions for the ellipse comes when we associate the generating vectors with the transverse positions of two neighboring rays with respect to a base ray. We find that we can express the propagation of a geometrical Gaussian beam (i.e. no diffraction) in this form given two nominally othrogonal generating rays. But more importantly, it can be shown that the amplitude component of a general Gaussian beam in the presence of diffraction can be represented as the convolution of two geometrical Gaussians, one propagated by the waist ray set and the other by the divergence rays. In others words, the diffraction process (in the case of Gaussians only) can be rigorously expressed as a convolution at any axial position of the near field geometrical wave with the far field geometrical wave.

There is no computational advantage in this method over the methods we have already presented. Therefore, the mathematical details of this convolution process will not be covered here. The convolution of two general Gaussians expressed in vector form can be derived by working with the products of their Fourier transforms in the frequency domain and making use of the following property of two-dimensional transforms (see Gaskill page 315 for the scalar form).

$$f(\vec{a}\times\vec{r},\vec{b}\times\vec{r})\overset{\mathcal{F}}{\Leftrightarrow}\frac{1}{|\vec{a}\times\vec{b}|}F\left(\frac{\vec{b}\cdot\vec{\rho}}{\vec{a}\times\vec{b}},\frac{-\vec{a}\cdot\vec{\rho}}{\vec{a}\times\vec{b}}\right)$$
$$\vec{a}\times\vec{b}=a_xb_y-a_yb_x=\begin{vmatrix}a_x & b_x\\ a_y & b_y\end{vmatrix} \tag{36}$$

Examples of single beam propagation

The ray equivalent method for general Gaussian beam propagation was easily incorporated into an already existing ray trace computer code. The following examples of single beam propagation were then generated by passing a 1 by 2 millimeter elliptical Gaussian through a cylindrical focusing mirror (100 millimeter focal length) oriented at 45 degrees relative to the axes of the ellipse. This produces a "twisted" beam whose amplitude principal axes are not parallel to the phase axes. The cross section of the beam amplitude was then viewed at six different axial locations ranging from 0 to 250 millimeter from the mirror. Figure 5 shows the results for a wavelength of 1 micron. The angles of rotation shown on the figure were calculated separately using the angular spectrum solution. We can see that there is very good agreement in this respect between the the ray equivalent and angular spectrum approaches. Also, note that the angle of rotation at z=100 millimeter which corresponds to the focus of the cylindrical element is not exactly 45 degrees as would be the case for purely geometrical propagation. Figure 6 shows the same results but for a wavelength 10 times longer where we would expect diffraction effects to be more apparent. In addition to the more noticeable spread of the beam as it propagates, we see that the rotation angle at the astigmatic focus is now quite different from 45 degrees.

Decomposition of truncated wavefronts into Gaussian beams

Now that we can propagate generally astigmatic Gaussian beams in any optical system, we can turn our attention back to the original goal of modeling the propagation of arbitrary wavefronts by decomposing them into Gaussians. In this paper, we will restrict ourselves to finite clear apertures illuminated by a uniform plane wave and followed by an optical system that may or may not aberrate the wavefront. This is a simple case but one that is routinely encountered in practice.

Simple one-dimensional analysis

Let us first start by looking at the one-dimensional problem of decomposing a slit pattern into a set of evenly spaced identical Gaussians. This simple mathematical analysis will give us a feel for the kinds of problems and tradeoffs associated with Gaussian sampling of general apertures.

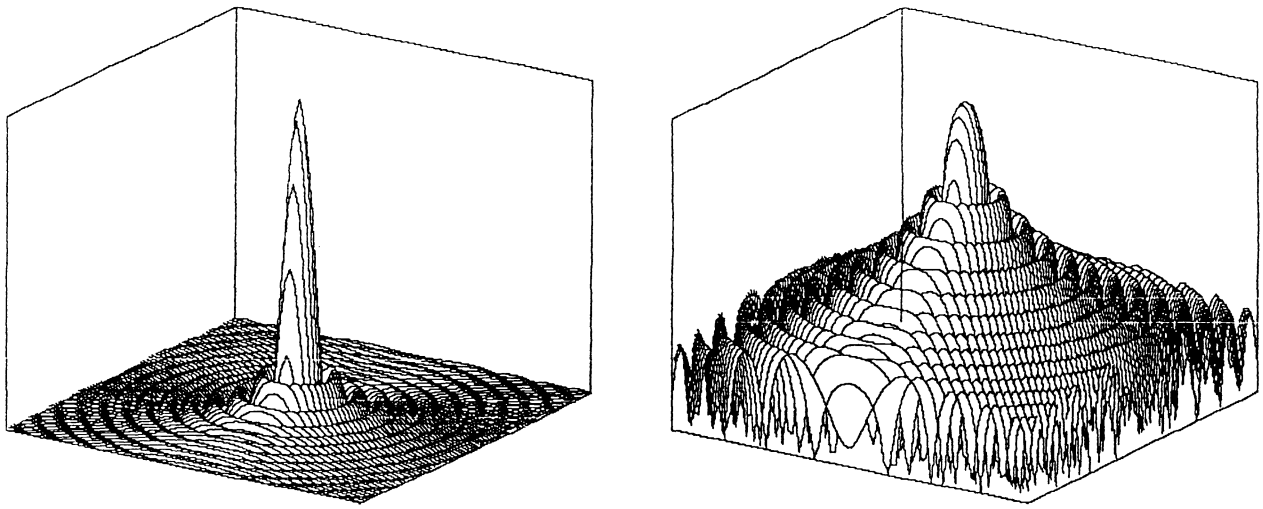

Figure 9. Amplitude and log intensity distributions calculated from closed-form analytical solution of perfect circular aperture diffraction.

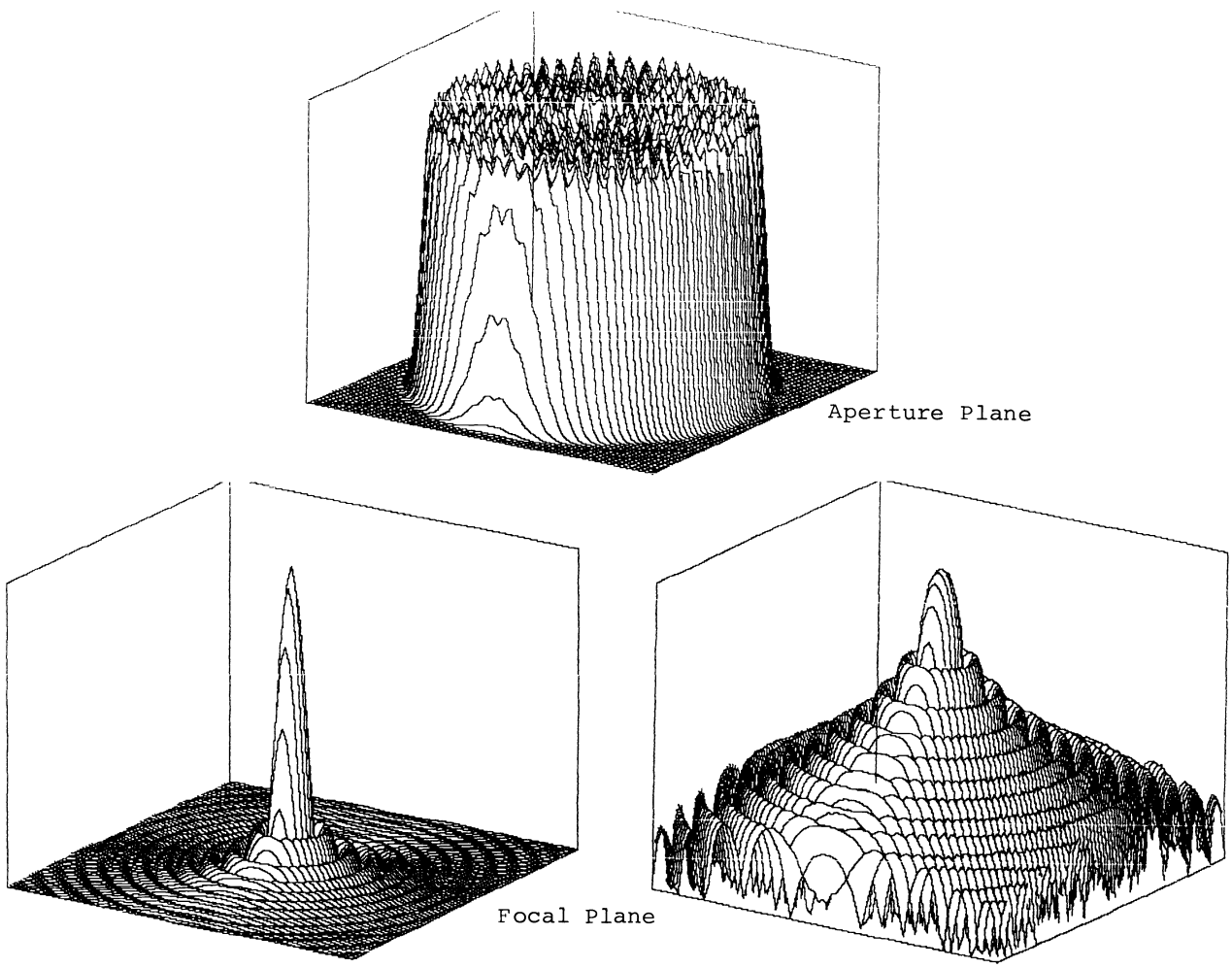

Figure 10. Corresponding distributions calculated from decomposition of aperture into 331 uniformly spaced Gaussian beams.

Figure 7 depicts the modeling of a slit of width d by a set of Gaussians of width w and spaced a center-to-center distance s apart. Mathematically, we can write this approximation to the slit function as:

$$f(x) \approx \left[\mathrm{rect}\left(\frac{x}{d}\right) \cdot \frac{1}{s}\,\mathrm{comb}\left(\frac{x}{s}\right)\right] * \frac{1}{w}\,\mathrm{Gaus}\left(\frac{x}{w}\right) \tag{37}$$

We have adopted the notation of Gaskill here. A finite set of evenly spaced delta functions represented by the product of the Rect and Comb functions is convolved with the Gaussian sampling function to form the approximate slit function. We can immediately notice two defects in this approximation. The top of the function is not flat but instead is rippled. In addition the sides of the function have a smooth gradual falloff instead of the perfectly sharp behaviour of the slit.

The question is how does this approximate sampling affect the diffracted field as it propagates. It is easiest to look at the far field pattern, which is equivalent to assuming that a plane wave incident on this aperture is focused by a perfect optical system. The resulting field at the focal plane will be proportional to the Fourier transform of the aperture function.

$$F(\alpha) \approx \left[d\,\mathrm{sinc}(d\alpha) * \mathrm{comb}(s\alpha)\right] \cdot \mathrm{Gaus}(w\alpha) \tag{38}$$

This field distribution is plotted in Figure 8. We see that the ripple at the slit has lead to diffracted orders and that the lack of sharpness of the edges results in multiplying the whole pattern by a Gaussian envelope. The zeroth order would accurately represent the field from a perfect slit if not for the additional falloff caused by the envelope. Therefore, in order to improve the sampling approximation, we must somehow reduce the higher order effects without adversely affecting the desired zeroth order pattern. If we use the ratio of first to zeroth order peak values as a measure of the error in the approximation,

$$\mathrm{Gaus}\left(\frac{w}{s}\right) = \mathrm{Gaus}\left(N\frac{w}{d}\right) \tag{39}$$

we find that these two goals are mutually exclusive when using uniformly spaced samples. The only way to beat down the higher orders is to reduce the ripple at the slit. This is done by increasing the width to spacing ratio of the Gaussians. However, the width of the envelope at the focal plane is inversely proportional to the width of the Gaussians at the slit so that the falloff of the zeroth order increases.

Another obvious way to reduce the sampling error would be to increase the number of samples N given by the ratio of the Gaussian spacing to the full slit width. This has the effect of moving the higher orders further away from the zeroth order. If we are only interested in modeling diffraction effects close to the central core, then this is a good approach. However, there is a practical limitation placed on the number of samples. Since the Gaussians get narrower as the number of samples increase, they will spread more rapidly as they propagate and the chances of them being large enough to violate the paraxial approximation at some surface increases. A Gaussian must be roughly at least 100 waves wide in order to have a far field divergence angle of less than a degree. If we set the minimum number of samples across an aperture at around ten (which puts the first order "noise" out beyond the 10th diffraction ring), then we are limited to apertures at least 1000 waves wide. For a one micron wavelength, this lower limit corresponds to one millimeter and therefore, does not restrict us from modeling most optical systems. Even apertures much smaller than this can be modeled by carrying out the decomposition in the angular frequency domain, i.e. a set of Gaussians many times wider than the aperture and incident upon it over a range of angles will interfere to form a good approximation to the narrow aperture.

So far we have confined our attention to Gaussian samples of constant width and even spacing across the aperture. If we allow both the widths and spacings of the Gaussians to be larger at the center of the aperture than at the edge, we can break up the periodicity of the slit ripple (which will wash out the higher order peaks) and at the same time increase the sharpness of the edges of the slit. Therefore, as we shall see in the next section, non-uniform sampling can produce more accurate results using less samples.

Results of computations for a circular aperture

The simplest and most often encountered two-dimensional problem that we can use to test the Gaussian decomposition technique is the circular aperture. In particular, let the aperture be 20,000 waves in diameter (e.g. a 20 centimeter diameter aperture at a 10 micron wavelength) and located at the front focal plane of an f/1 parabolic mirror. The closed form exact diffraction pattern at the back focus is shown in Figure 9 for both amplitude and log intensity over six orders of magnitude. This should be compared with the results in

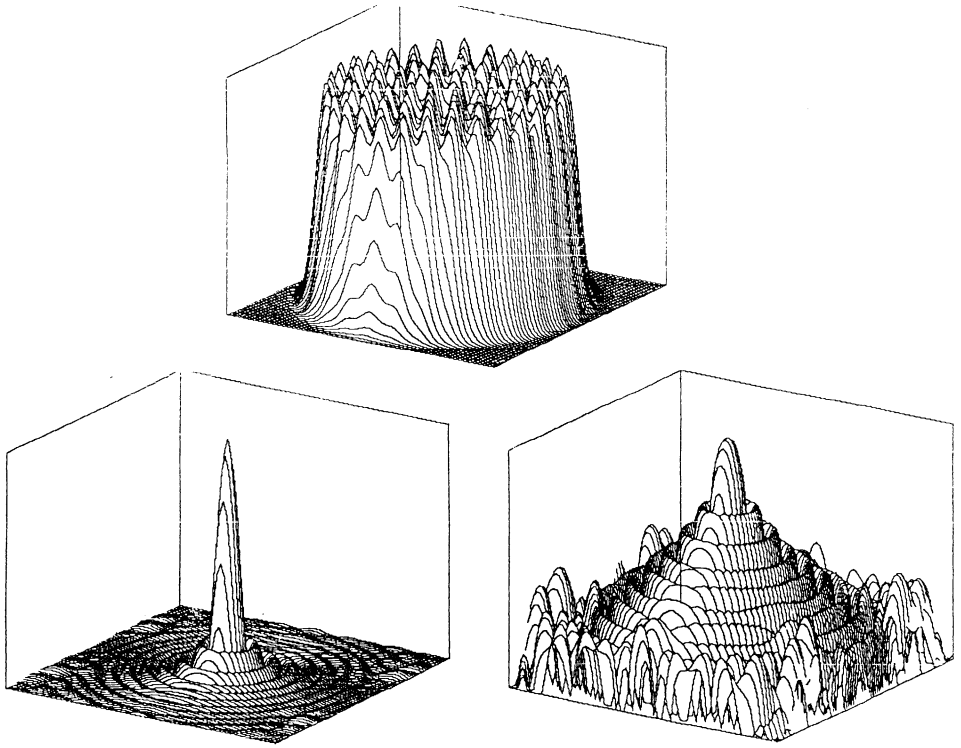

Figure 11. Aperture and focal plane distributions using 91 uniformly spaced Gaussians.

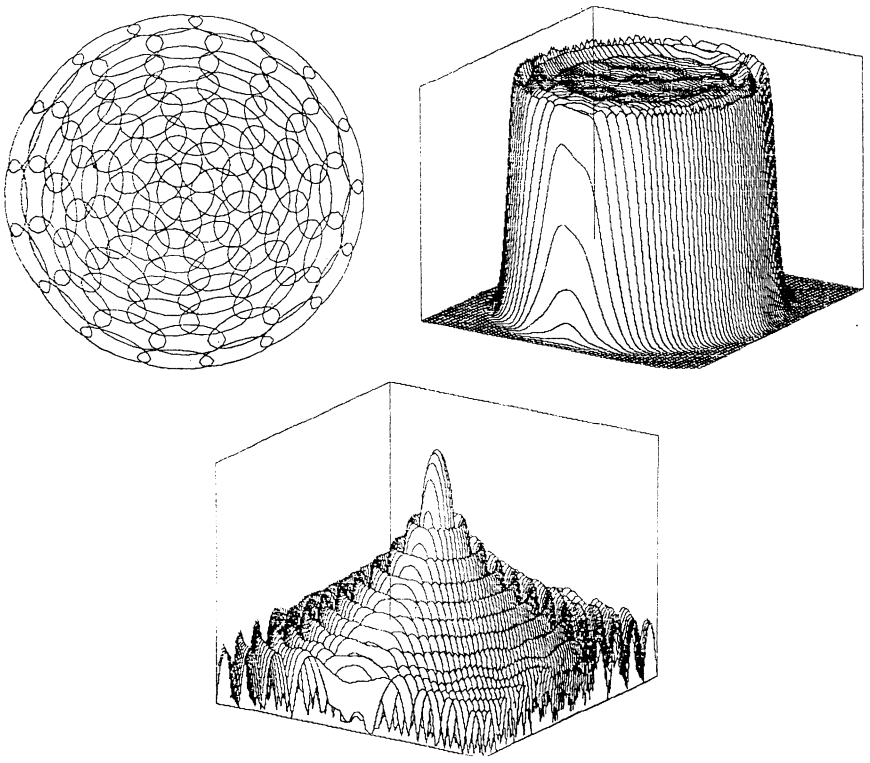

Figure 14. Corresponding distributions using the non-uniform sampling shown.

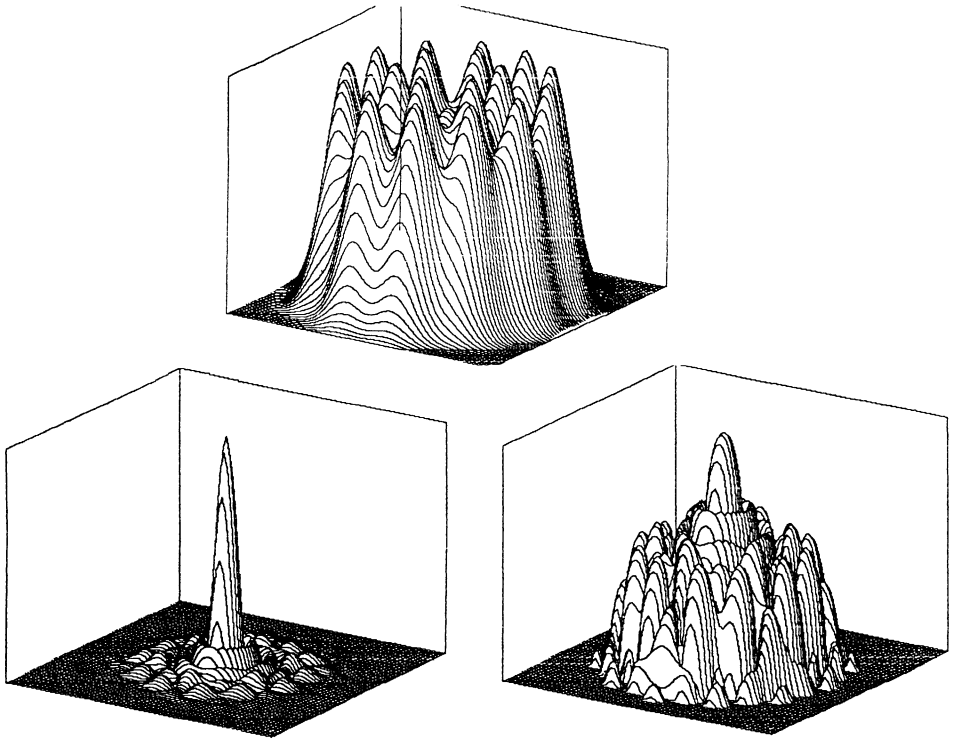

Figure 12. Aperture and focal plane distributions for 19 beam sampling with Gaussians having a width-to-spacing ratio of 1.

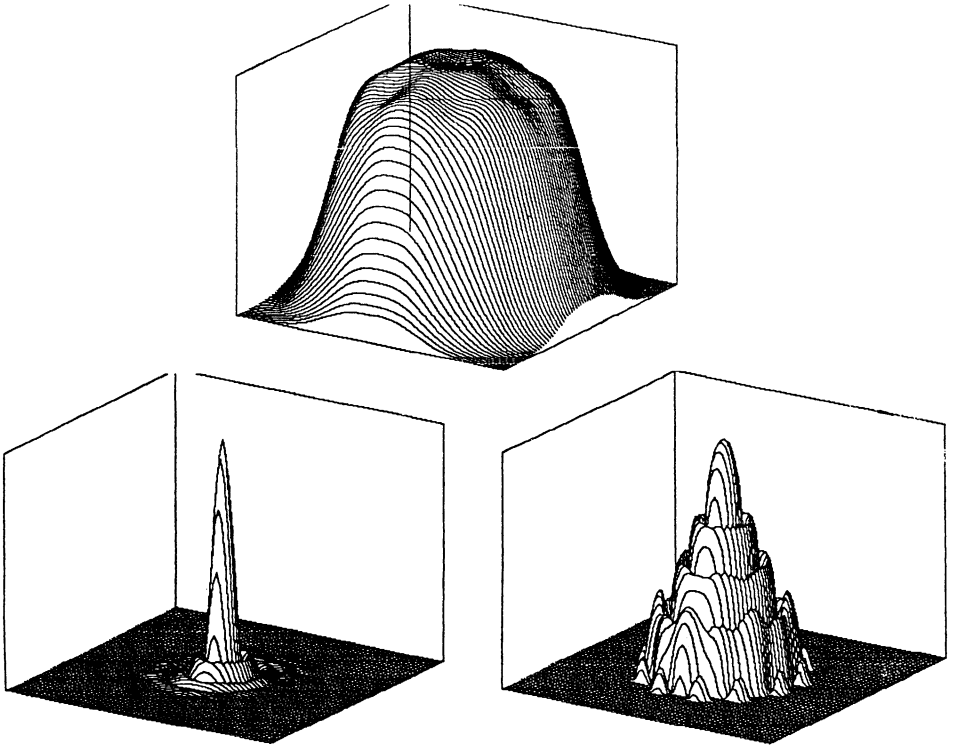

Figure 13. Corresponding distributions for a width-to-spacing ratio of 1.5.

Figure 10 obtained after tracing 331 Gaussian beams (arranged in a uniform hexapolar grid in the aperture) to the focal plane. The diameter of each beam was approximately equal to the spacing between beams and there were 10 radial zones. One has to look very hard at the two figures in order to see any difference because the data area was cleverly chosen such that the first order diffraction peaks are just outside it. However, if we go to 91 larger beams corresponding to 5 radial zones, we can clearly see these defects in Figure 11.

A more extreme case is shown in Figure 12 where only 19 beams in two radial zones were used to sample the aperture. Here the aperture function closely resembles that of a car's distributor cap. At the focal plane, the higher order diffraction peaks are virtually invisible along with the most of the normal diffraction rings because of the narrowness of the envelope Gaussian. Increasing the beam width to spacing ratio to 1.5 improves the aperture function in Figure 13 but eliminates almost all the rings in the focal plane.

If we drop the limitations associated with uniform sampling, it is possible to vastly improve the results for the 91 beam case to the point where the diffraction pattern is nearly the same for the 331 beam case (see Figure 14).

Circular aperture diffraction in the presence of aberrations

The diffraction disk calculation in the absence wavefront aberrations was mostly an academic exercise and does not show the true power of the Gaussian decomposition technique in real world applications where aberrations are invariably present. In the presence of aberrations, the optical path length, directions, and positions of each central beam ray will deviate from the ideal. In addition, the wavefront principal curvatures of each Gaussian will differ according to the the aberrations introduced at surfaces along their separate paths. Amplitude variations and piston error phase shifts due to passages through multilayer interfaces could also be taken into account.

Figure 15 shows the resulting amplitude and log intensity diffraction patterns when a wave of astigmatism is introduced by making the parabolic mirror slightly toric. The aperture was sampled with 91 uniformly spaced Gaussians and the diffraction pattern generated on the medial focal plane. These results compare favorably with analytical results shown on page 480 of Born and Wolf's classic text on optics.

In a similar manner, a wave of spherical aberration can be introduced into the system by slightly altering the conic constant of the mirror so that it is not exactly a paraboloid. We again could produce a diffraction pattern on a plane normal to the axis at focus. However, the through focus characteristics of spherical aberration are of much more interest. Since the Gaussian beam decomposition technique can calculate the diffracted field at any point in the system, we can just as easily determine the field distribution in a plane containing the optical axis as one that is normal to it. Figure 16 is a through focus amplitude and log intensity distribution for an aberration free f/1 beam. Again, this Gaussian beam decomposition calculation closely agrees with analytical results (see Born and Wolf page 440). As to be expected, it is symmetric about the paraxial focal plane (vertical line at center). With a wave of spherical aberration, the diffraction patterns are no longer symmetric about the paraxial focus (see Figure 17). The maximum intensity is shifted to a point halfway between the marginal and paraxial focii. This is exactly what is predicted from theory and illustrates the power and accuracy of the Gaussian beam decomposition method for diffraction calculations.

Extension of decomposition technique to partially coherent vector fields

If we could propagate vector Gaussian fields by taking into account their polarization, then the beam decomposition technique could be extended to the propagation of general electromagnetic wavefronts. To first order, this is a simple matter because the slow divergence (small departure from a plane wave) of an individual Gaussian means that the polarization can not vary much over the extent of the beam. Therefore, we need only concern ourselves with the polarization of the region near the central beam ray. This region can be considered a small area of a plane wave. Then we can use Fresnel's equations to modify the complex polarization vectors at each refractive and/or multilayer interface. When we sum the beams at any particular point in space, we must make sure that any intensity reductions due to cross polarizations are taken into account. In addition, it is an easy matter to also take into account any incoherence between beams by using the following equation to calculate the resulting energy density from the complex electric field vectors of all the beams.

$$I = \sum_{i}^{\text{Beams}} \sum_{j} \gamma_{ij} \left(\vec{E}_i \cdot \vec{E}_j^{*}\right), \quad 0 < \gamma_{ij} < 1 \quad \begin{cases} \gamma_{ii} = 1 \\ \gamma_{ij} = \gamma_{ji} \end{cases} \tag{40}$$

$$= \sum_{i} \left(\vec{E}_i \cdot \vec{E}_i^{*}\right) + 2 \sum_{i} \sum_{j<i} \gamma_{ij} \left(\vec{E}_i \cdot \vec{E}_j^{*}\right)$$

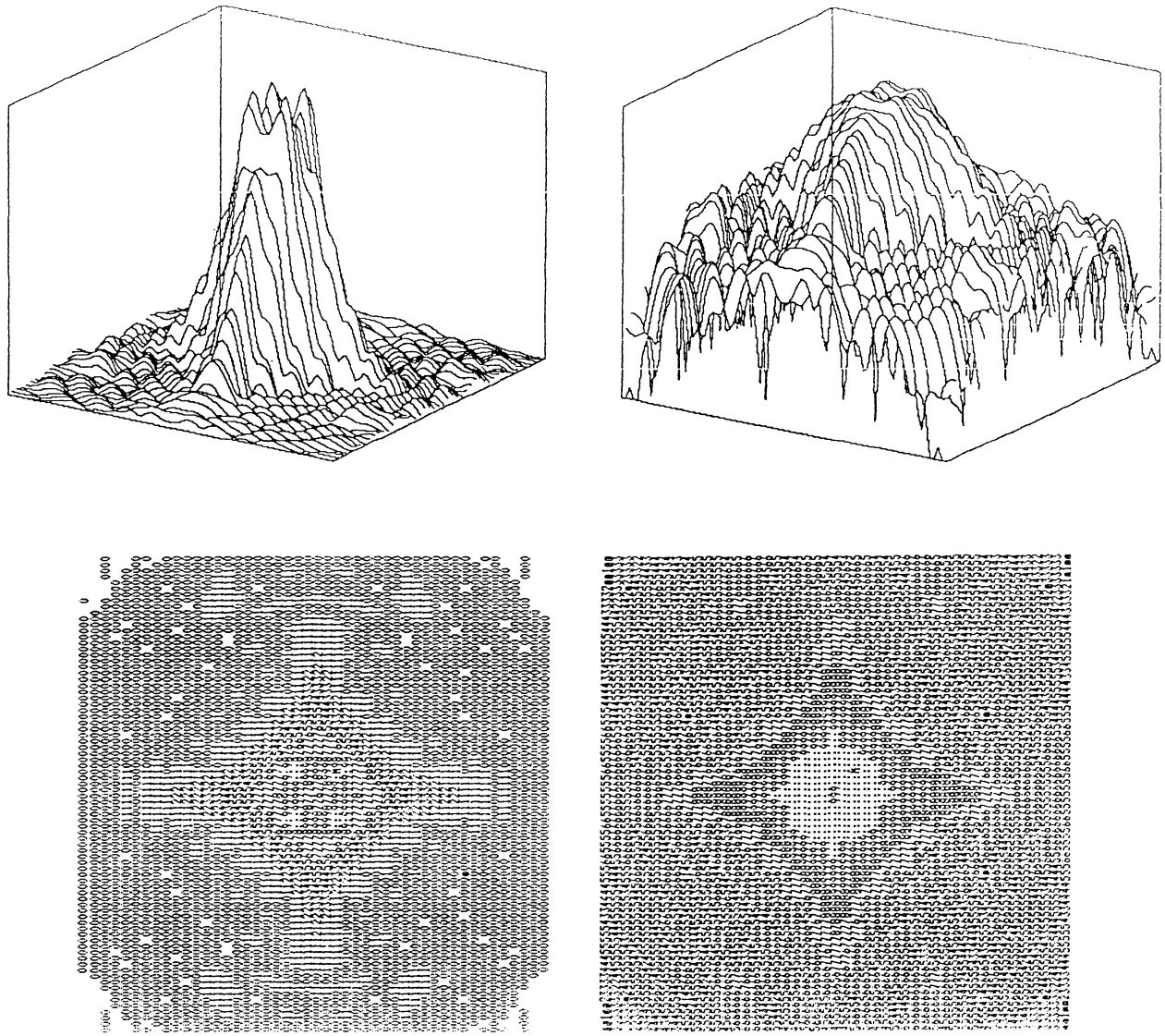

Figure 15. Amplitude and log intensity distributions at medial focus for one wave of astigmatism.

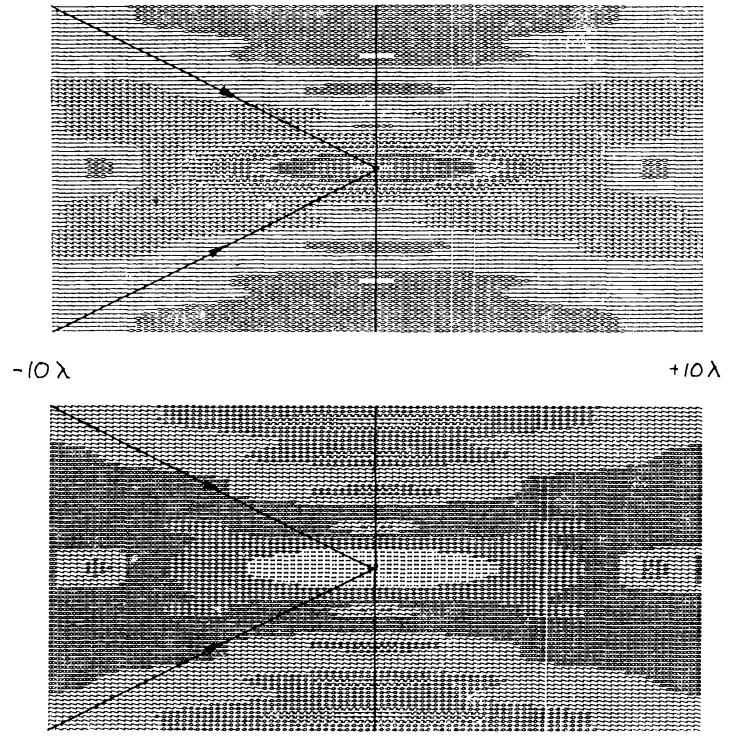

Figure 16. Through focus amplitude and log intensity meridional plane distributions for a perfect F/1 beam.

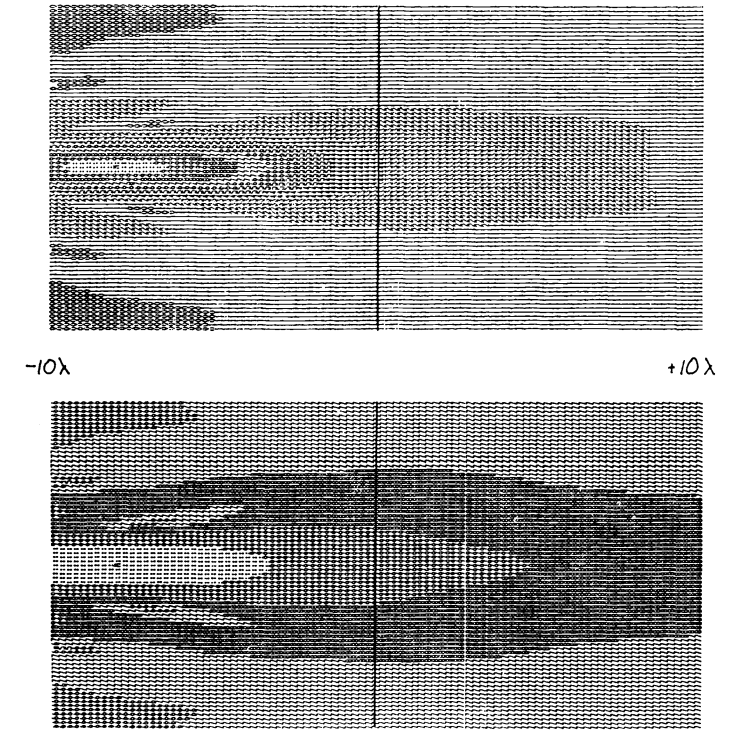

Figure 17. Through focus distributions in the presence of one wave of spherical aberration.

The coherence factor can range from zero (total incoherence) to one (perfect coherence). It can be a function of the difference in optical path between the beams and spectral characteristics of the source (temporal coherence) or the difference in beam positions in the aperture plane and the source size (spatial coherence).

Summary

An alternative method for doing system diffraction calculations was proposed. It is based on the decomposition of the incident wavefront into Gaussian beams. These generally astigmatic beams can then be individually propagated through any optical system using a straight-forward ray equivalent approach. The partially coherent vector field at any point in the system is simply found by properly summing the beam contributions at that point. This technique has been incorporated into a very powerful and flexible ray trace computer code that allows one to easily analyze diffraction effects in not only standard imaging systems, but also non-imaging concentrators, multimode fibers, interferometers, and synthetic aperture systems.

References

1. Gaskill, J. D. , Linear Systems, Fourier Transforms, and Optics, Wiley, 1978.

2. Thomas, G. B., Calculus and Analytical Geometry, Addison-Wesley, 1968, pg. 352.

3. Yariv, A., Introduction to Optical Electronics, Holt-Rinehart-Winston, 1971, pg. 31.

4. Goodman, J. W., Introduction to Fourier Optics, McGraw-Hill, 1968, pg. 48.

5. Arnaud, J. and Kogelnik, H., "Gaussian Light Beams with General Astigmatism", Applied Optics, Vol. 8, No. 8, Aug. 1969, pg. 1687.

6. Shack, R. V., private communication, University of Arizona, Tucson, 1985.

7. Stavroudis, O. N., The Optics of Rays, Wavefronts, and Caustics, Academic Press, 1972, pg. 161.

8. Arnaud, J. A., "Representation of Gaussian beams by complex rays", Applied Optics, Vol. 24, No. 4, Feb. 1985, pg. 538.

9. Arnaud, J. A., "Nonorthogonal Optical Waveguides and Resonators", Bell System Technical Journal, Nov. 1970, pg. 2311.

10. Herloski, Marshall, and Antos, "Gaussian beam ray-equivalent modeling and optical design", Applied Optics, Vol. 22, No. 8, April 1983, Pg. 1168.

11. Born, M. and Wolf, E., Principles of Optics, Pergamon Press, 1975.