On the solution of the phase retrieval problem

Hoenders, B.J.

Published in:
Journal of Mathematical Physics

DOI:
10.1063/1.522769

On the solution of the phase retrieval problem*

B. J. Hoenders†

Department of Physics and Astronomy, University of Rochester, Rochester, New York 14627
(Received 23 August 1974; revised manuscript received 18 February 1975)

It is shown that the intensity in the image plane of a microscope determines *uniquely* the phase of the corresponding image wavefunction up to an over-all phase. This result is obtained using the *a priori* information that both the image wavefunction and the unperturbed wavefunction in the Fraunhofer plane are band-limited and that we have some *a priori* knowledge about the intensity at the rim of the diaphragm in the Fraunhofer plane. If we have no useful *a priori* information about the wavefunction in the Fraunhofer plane, unique phase reconstruction is possible from two exposures, corresponding to two different values of the defocusing.

INTRODUCTION

Since only the modulus of a wavefunction is usually a measurable quantity rather than the wavefunction itself, the phase of a wavefunction is, as a rule, lost during measurement. The problem whether or not the phase of a function can be unambiguously determined from the values of its modulus is known as the phase retrieval problem.

The loss of phase information leads to serious difficulties in the interpretation of measurements and also to ambiguities. It is well known, for example, that the structure of a crystal cannot in general be uniquely determined from intensity measurements on the scattered x rays.

Another example arises in the theory of partial coherence where it is known that in general, the complex degree of coherence cannot be determined from its modulus without additional information, Wolf,[1] Nussenzveig.[2] Another example of this type is provided by the theory of image formation in microscopes, where it is shown that several distributions in the Fraunhofer plane might lead to the same intensity distribution in the Gaussian image plane, Walther.[3]

However, most functions occurring in physics are analytic, which implies that both phase and modulus are connected by the Cauchy–Riemann equations which considerably narrows the class of phase functions which might be assigned to a given modulus. Much work has been done to determine to which degree the phase of an analytic function is determined by its modulus. For an extensive survey see Mandel and Kohler.[4] From the calculations of Wolf,[1] Dialetis and Wolf,[5] Walther,[3] and Nussenzveig[2] it became clear that in several cases of physical interest the phase of an analytic function is not uniquely determined by its modulus. It is the aim of this paper to give an extension of the analysis of the phase problem as developed by Walther[3] and to provide an answer to the following phase problem occurring in the theory of image formation in a microscope: Suppose that an object is imaged through a microscope free of aberrations and the intensity in the image plane is measured. We then investigate whether the intensity distribution determines the phase of the image wavefunction unambiguously. As already shown by Walther,[3] several wavefunctions in the Fraunhofer plane might lead to the same intensity distribution in the image plane. However, it should be stressed that this does *not* prove that the phase cannot be reconstructed. This comes from our *a priori* knowledge that the *unperturbed* wavefunction $g(y)$, viz. the wavefunction just before the diaphragm in the Fraunhofer plane, is band-limited, like the image wavefunction, i.e.,

$$g(y)=\int_{-1}^{+1}\exp(ix_0y)\,\psi(x_0)\,dx_0, \tag{Ia}$$

$$h(x)=\int_{-1}^{+1}\exp(ixy)\,g(y)\,dy, \tag{Ib}$$

where $\psi(x_0)$ is the distribution in the object plane, $g(y)$ the unperturbed distribution in the Fraunhofer plane, and $h(x)$ the distribution in the Gaussian image plane, Born, Wolf.[6] It might so happen that only *one* of the various distributions in the Fraunhofer plane is consistent with our *a priori* knowledge that it must be a *perturbed* band-limited function, viz. a function which, relying on Kirchoff's boundary conditions is part of a band-limited function on the transparent part of the diaphragm and zero on the nontransparent part. This would imply that due to this *a priori* knowledge arg $h(x)$ is uniquely determined by $|h(x)|$. However, unfortunately this conjecture is not true, e.g., the modulus of the band-limited function $h^{(1)}(z)\equiv h^*(z^*)$

$$h^{(1)}(z)=\int_{-1}^{+1}\exp(izy)\,g^*(-y)\,dy,$$

is equal to the modulus of $h(z)$ on the real axis. Therefore the unperturbed wavefunction $g^*(-y)$ yields the same intensity distribution in the image plane as $g(y)$. Moreover, it follows from Eq. (1a) that $g^*(-y)$ is a band-limited function thus obtaining at least two unperturbed bandlimited distributions in the Fraunhofer plane leading to the same intensity in the image plane. In holography $g(y)$ and $g^*(-y)$ would correspond to the reconstructed object and its twin object, "twin images."

But, fortunately, uniqueness can be obtained by using a little more *a priori* information, i.e., if we know *a priori* that either the number of zeros of $h(z)$ is finite in the upper half or the lower half of the complex plane, uniqueness is obtained. The uniqueness is obtained by showing that only one (the band-limited) unperturbed wavefunction in the Fraunhofer plane decays if y tends to plus or minus infinity whereas all the other possible wavefunctions diverge. Moreover it will be shown how to calculate all these possible unperturbed wavefunctions from the intensity. The above mentioned *a priori*

knowledge concerning the distribution of the zeros of $h(z)$ is readily obtained in light—or electron microscopy and, as will be shown in the section "application to microscopy," boils down, to the *a priori* knowledge whether $|g(-1)| > |g(1)|$ or $|g(-1)| < |g(1)|$. This knowledge can be obtained from experiment or *a fortiori* by illuminating the object with a plane wave incident with a certain angle with the optical axis of the microscope. From now on it will be assumed that the number of zeros of $h(z)$ located in the upper half of the complex plane is finite. A theory completely analogous to the one developed in this paper can be derived if a finite number of zeros are located in the lower half of the complex plane. It would be interesting to find whether a similar analysis also holds if the object is imaged into the Fraunhofer plane by a microscope suffering from aberrations. It might so happen that in that case also only *one* wavefunction can be identified as the image in the Fraunhofer plane of an object distribution. Although the author regards this to be highly probable a proof of this conjecture has not been constructed so far. Therefore, let us suppose that in this case all the various wavefunctions $g(y)$ are admissible and hence that the phase cannot be reconstructed unambiguously from the modulus. In order to get rid of this ambiguity more information has to be put in. Recently, Gerchberg and Saxton[7] propose as one possible choice for such additional information the knowledge of the intensity distribution in the Fraunhofer plane.

Another way of obtaining additional information is to measure several intensity distributions in the image plane corresponding to different values of parameters which are at our disposal such as the location of the focus, illumination, or the size of the aperture. Misell[8] proposed a method of phase reconstruction for weak objects from two exposures, obtained by using two diaphragms, one which transmits only positive spatial frequencies, and one which transmits only negative spatial frequencies.

Another idea will be investigated in more detail in this paper. It will be shown that from two exposures, corresponding to two different values of the defocusing, the phases of the two wavefunctions in the image plane can be reconstructed unambiguously, up to a constant. To prove this statement we shall derive in the next section a dispersion relation between the phase and the modulus of a bandlimited function. Moreover it will be shown that all the wavefunctions $g_1(y)$ in the Fraunhofer plane leading to the same intensity in the image plane are related to the true wavefunction in the Fraunhofer plane by a linear Volterra integral equation of the second kind. Comparing the two sets of wavefunctions $\{g_1^{(1)}(y)\}$ and $\{g_1^{(2)}(y)\}$ which correspond to the two exposures it is shown that only one wave function is consistent with both intensity distributions. Hence, using (1) $\arg h(x)$ can be calculated.

A DISPERSION RELATION FOR BAND LIMITED FUNCTIONS

In the introduction we already indicated the need for a dispersion relation between phase and amplitude of a band limited function. Such a dispersion relation may be obtained by means of the following lemma:

Lemma: Let the complex valued function $g(y)$ be defined in the interval $-1 \leq y \leq +1$. Suppose $g'(y) = (d/dy) \times g(y)$ exists everywhere in that interval and is of bounded variation. Suppose that the entire function $h(z)$ is band limited to the interval $-1 \leq y \leq +1$, i.e., it has a representation of the form

$$h(z) = \int_{-1}^{+1} \exp(izy)\, g(y)\, dy. \tag{1}$$

Then, if the number of zeros $a_{n'}$ $(n' = 1, 2, \cdots)$ of $h(z)$ in the upper half of the complex plane is finite, the phase and the modulus of $h(z)$ are related by the dispersion relation

$$\frac{1}{\pi} \oint_{-\infty}^{\infty} \ln|h(x')x'| \, \frac{dx'}{x' - x}$$

$$= -\arg h(x) - \arg\left(x \exp(ix) \prod_{n'} \frac{x - a_{n'}^*}{x - a_{n'}}\right)$$

$$+ \frac{1}{2}\pi + \arg g(-1), \tag{2}$$

the integral on the left being interpreted as the Cauchy principal value and both x and x' are real numbers.

Proof: Repeated integration by parts yields

$$\int_{-1}^{+1} \exp(izy)\, g(y)\, dy = \frac{\exp(izy)}{iz}\, g(y)\Big|_{-1}^{+1} - \frac{\exp(izy)}{(iz)^2}\, g'(y)\Big|_{-1}^{+1}$$

$$+ \frac{1}{(iz)^2} \int_{-1}^{+1} \exp(izy)\, dg'(y). \tag{3}$$

Introducing complex numbers α and β by the relations

$$\exp(i\alpha) = g(1), \quad \exp(-i\beta) = g(-1), \tag{4}$$

we may readily derive from (2) the asymptotic formula

$$h(z) = \frac{2}{z} \exp\left(\frac{i}{2}(\alpha - \beta)\right) \sin[z + \tfrac{1}{2}(\alpha + \beta)] \left(1 + O\left\{\frac{1}{z}\right\}\right),$$

$$0 \leq \arg z \leq 2\pi. \tag{5}$$

Therefore the zeros a_n of $h(z)$ are distributed according to the asymptotic formula (Titchmarsh[9])

$$a_n \sim n\pi - \tfrac{1}{2}(\alpha + \beta), \tag{6a}$$

or Cartwright[10]

$$a_n \sim n\pi + \frac{i}{2} \ln\left(\frac{g(-1)}{g(1)}\right). \tag{6b}$$

Consider the contour integral

$$I(x, R) = \frac{1}{\pi i} \int_C \ln\left(\prod_{n'} \frac{z - a_{n'}^*}{z - a_{n'}}\, h(z)\, z \exp(iz)\right) \frac{dz}{z - x}, \tag{7}$$

for large values of the parameter R defined below.

In Eq. (7) the contour C consists of the part of the real axis between $-R$ and $+R$, indented at $z = x$ and at the possible zeros of $h(z)$ at the real axis with semicircles with radii ϵ in the upper half of the complex plane, and a semicircle in the upper half of the complex plane of radius R and centre at the origin. Using Eq. (3)

$$\exp(iz)\int_{-1}^{1}\exp(izy)\,g(y)\,dy=\frac{ig(-1)}{z}+O\left\{\frac{1}{z^2}\right\},$$

$$\theta<\arg z<\pi. \tag{8}$$

On using the asymptotic formula (8) we derive the formula

$$\frac{1}{\pi i}\int_{\text{semicircle}}\ln\left(\prod_{n'}\frac{z-a_{n'}^*}{z-a_n}\,h(z)z\exp(iz)\right)\frac{dz}{z-x}$$

$$=\frac{1}{\pi i}\int_{\text{semicircle}}\left(\ln\left(ig(-1)\right)+O\left\{\frac{1}{z}\right\}\right)\frac{dz}{z-x}$$

$$=\ln\,ig(-1)\;+O\left\{\frac{1}{R}\right\}. \tag{9}$$

We inserted the asymptotic expansion (8) into the lhs of (9) although (8) is not valid for real values of z. However, Eq. (3) shows that $\exp(iz)zh(z)$ is bounded for all values of $\arg z$ in the closed interval $[0,\pi]$. Hence, recalling that the number of zeros of $h(z)$ in the upper half of the complex plane is finite, which means using (6b) that $|g(-1)/g(1)|<1$, there exists a positive number M, independent of R, such that for sufficiently large values of R

$$\left|\ln\left(\prod_{n'}\frac{z-a_{n'}^*}{z-a_{n'}}\,h(z)z\exp(iz)\right)\right|<M,\quad z=R\exp(i\phi),$$

$$0\leq\phi\leq\pi. \tag{10}$$

Using (10), we observe that the contributions to (9) from those parts of the semicircle corresponding to values of $\arg z$ lying in the intervals

$$0\leq\arg z\leq\delta,\quad \pi-\delta\leq\arg z\leq\pi,$$

can be made arbitrarily small by choosing the positive number δ to be small enough. This proves that

$$\lim_{R\to\infty}\frac{1}{\pi i}\int_{\text{semicircle}}\ln\left(\prod_{n'}\frac{z-a_{n'}^*}{z-a_{n'}}\,h(z)z\exp(iz)\right)\frac{dz}{z-x}$$

$$=\ln(ig(-1)). \tag{11}$$

The argument of the logarithm of the integrand of (7) is an analytic function which by construction has no zeros in the upper half of the complex z-plane but possibly on the real axis. Hence the logarithm is an analytic function within the domain with boundary C and Cauchy's theorem applied to (7), yields the result

$$\lim_{R\to\infty}I(x,R)=0. \tag{12}$$

Letting the radii ϵ of the semicircles tend to zero and using the property that the possible zeros of $h(z)$ at the real axis gives no contribution to the integral (7), Eqs. (7), (11), and (12) leads to

$$\lim_{R\to\infty}I(x,R)=0=\ln(ig(-1))+\frac{1}{\pi i}\oint_{-\infty}^{+\infty}\ln\left(\prod_{n'}\frac{x-a_n^{*\prime}}{x'-a_{n'}}\right.$$

$$\left.\times h(x')x'\exp(ix')\right)\frac{dx'}{x'-x}-\ln\left(\prod_{n'}\frac{x-a_{n'}^*}{x-a_{n'}}\,h(x)x\exp(ix)\right). \tag{13}$$

Equating the imaginary parts of Eq. (13) leads to the desired dispersion relation

$$\frac{1}{\pi}\oint_{-\infty}^{+\infty}\ln|h(x')x'|\,\frac{dx'}{x'-x}=-\arg h(x)-\arg\left(xe^{ix}\prod_{n'}\frac{x-a_{n'}^*}{x-a_{n'}}\right)$$

$$+\frac{\pi}{2}+\arg(g(-1)). \tag{14}$$

RELATIONS BETWEEN POSSIBLE SOLUTIONS OF THE PHASE RETRIEVAL PROBLEM

Equation (14) reveals that, knowing the location of the zeros a_n of $h(z)$ in the upper half of the complex z-plane, the phase of $h(z)$ can be calculated from the values of its modulus along the real axis up to an overall phase constant. This result is similar to the well-known relation between the modulus and phase of the complex degree of coherence (Wolf[1], Nussenzveig[2]). Hence the question now arises whether or not the zeros a_n are determined from the knowledge of $|h(x)|$ along the real axis.

To answer this question we consider the function $h(z)h^*(z^*)$. First we observe that if $h(z)$ is an entire function so is $h^*(z^*)$, as follows immediately from the Cauchy–Riemann equations. Hence $h(z)\,h^*(z^*)$ is the analytical continuation of $h(x)h^*(x)$ into the whole complex plane and can be calculated with any desired tolerance from the values of $|h(x)|$ at a sufficiently large number of points.[11] Let $\{a_n\}$ denote the set of zeros of $g(z)$ and $\{b_n\}$ denote the set of zeros of $g^*(z^*)$. Therefore the set of zeros of $h(z)\,h^*(z^*)$ is the union $\{a_n\}U\{b_n\}$, which can in principle, be determined from the knowledge of $|h(x)|$ at the real axis.[11,12]

In order to apply the dispersion relation (2) we should be able to decide whether a particular zero in the upper half of the complex plane belongs to the set $\{a_n\}$ of zeros of $h(z)$ or to the set $\{b_n\}$ of zeros of $h^*(z^*)$. Unfortunately, as already observed by Walther,[3] it is impossible to make such a distinction. This can be seen by "flipping" one of the zeros of $h(z)$ about the real axis, which is equivalent to multiplying it by a so-called Blaschke factor $(z-a_n^*)/(z-a_n)$. We then obtain a new entire function, one zero of which has been replaced by its complex conjugate and the value of its modulus along the real axis has not changed because

$$\left|\frac{x-a_n^*}{x-a_n}\right|=1,\quad x\text{ real}.$$

Furthermore, it can be shown that multiplying a band limited function $h(z)$ by a Blaschke factor transforms it into another band limited function[15] (Walther[3]). (A proof different from Walther's will be given in Theorem 1.)

Observing that the set $\{b_n\}$ of zeros of $g(z)$ are the complex conjugates of the set $\{a_n\}$ of zeros of $g^*(z^*)$, it is clear that multiplication of the original function $h(z)$ with a suitable product of Blaschke factors yields a new band limited function, the zeros of which may be any finite number of combinations of all those elements of

the union $\{a_n\}U\{b_n\}$ lying in the upper half of the complex plane. The reason that we allow only a finite number of combinations is due to our *a priori* knowledge that only a *finite* number of zeros of $h(z)$ are located in the upper half of the complex plane as is discussed in the introduction. Moreover, the moduli of all these functions are the same along the real axis. Therefore any combination of Blaschke factors, generated by all possible combinations of those elements of the union $\{a_n\}U\{b_n\}$ which lie in the lower half of the complex plane, yields, after insertion of these factors into the dispersion relation (2), a possible solution for the phase problem. All the possible band limited functions $h_1(z)$, labeled by the index 1, having the same modulus along the real axis, and a finite number of zeros in the upper half of the complex plane, can be represented by the formula

$$h_1(z)=\int_{-1}^{+1}\exp(izy)\,g_1(y)\,dy, \tag{15}$$

where

$$h_1(z)=\prod_{\{n_1\}}\frac{z-a_n^*}{z-a_n}\,h(z) \tag{16}$$

and

$$h(z)=\int_{-1}^{+1}\exp(izy)\,g(y)\,dy. \tag{17}$$

The product in Eq. (16) is taken over any subset $\{n_1\}$ containing a finite number of elements of the union $\{a_n\}U\{b_n\}$ in the lower half of the complex plane.

The following theorem will show that all the functions $g_1(y)$ are related to $g(y)$ by a linear Volterra integral equation of the second kind.

Theorem 1: Let the function $g(y)$ be defined on the interval $-1\leq y\leq+1$ and suppose that $g'(y)$ exists everywhere in this closed interval and is of bounded variation. Suppose the complex numbers a_n denote the zeros of the function

$$h(z)=\int_{-1}^{+1}\exp(izy)\,g(y)\,dy, \tag{18}$$

and let

$$h_1(z)=\prod_{\{n_1\}}\frac{z-a_n^*}{z-a_n}\,h(z), \tag{19}$$

where the index n_1 labels all the finite number of possible combinations of Blaschke factors. Then the functions $h_1(z)$ are band limited,[15] i. e., there exist functions $g_1(y)\in L^2$ such that

$$h_1(z)=\int_{-1}^{+1}\exp(izy)\,g_1(y)\,dy, \tag{20}$$

and $g(y)$ and $g_1(y)$ are related by the following Volterra integral equation of the second kind:

$$g_1(y)=g(y)-i\sum_{n''}(a_{n''}-a_{n''}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n''}}\frac{a_{n''}-a_n^*}{a_{n''}-a_n}\times\int_{-1}^{y}e^{ia_{n''}(y-y')}g(y')\,dy'+i\sum_{n}(a_n-a_n^*)\prod_{\substack{\{n_1\}\\ n_1\neq n'}}\frac{a_{n'}-a_n^*}{a_{n'}-a_n}\times\int_{y}^{1}e^{ia_{n'}(y-y')}g(y')\,dy'. \tag{21}$$

In Eq. (21) the index n'' labels all the poles in the lower half of the complex plane of the Blaschke products occurring in (19). Similarly the index n' labels all the poles a_n in the upper half of the complex plane.

Proof: Consider the two integrals

$$I_1(y,c)=\frac{1}{2\pi}\int_{-c}^{+c}\exp(-iyx)\left(\prod_{\{n_1\}}\frac{x-a_n^*}{x-a_n}\times\int_{-1}^{y}\exp(ix\tau)\,g(\tau)\,d\tau\right)dx\quad\text{if } -1<y<+1, \tag{22}$$

$$I_2(y,c)=\frac{1}{2\pi}\int_{-c}^{+c}\exp(-iyx)\left(\prod_{\{n_1\}}\frac{x-a_n^*}{x-a_n}\times\int_{y}^{1}\exp(ix\tau)\,g(\tau)\,d\tau\right)dx\quad\text{if } -1<y<+1, \tag{23}$$

for large values of the real parameter c. On adding Eqs. (22) and (23) we obtain at once

$$I_1(y,c)+I_2(y,c)=\frac{1}{2\pi}\int_{-c}^{+c}\exp(-iyx)\,h_1(x)\,dx. \tag{24}$$

Let us close the contour of I_1 by a semicircle C^- in the lower half of the complex plane, with radius c, and centered at the origin, and let us close the contour of I_2 by a semicircle C^+ in the upper half of the complex plane, with radius c and centered at the origin.

Using the asymptotic expansions

$$\int_{-1}^{y}\exp[iz(-y+\tau)]\,g(\tau)\,d\tau=\frac{g(y)}{iz}+O\left\{\frac{1}{z^2}\right\},\quad \pi<\arg z<2\pi,\quad -1<y<+1, \tag{25}$$

$$\int_{y}^{+1}\exp[iz(-y+\tau)]\,g(\tau)\,d\tau=-\frac{g(y)}{iz}+O\left\{\frac{1}{z^2}\right\},\quad 0<\arg z<\pi,\quad -1<y<+1, \tag{26}$$

that follows from (3), we obtain from (22) and (25), with the help of the residue theorem,[13]

$$I_1(y,c)=-i\sum_{n''}(a_{n''}-a_{n''}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n''}}\frac{a_{n''}-a_n^*}{a_{n''}-a_n}\times\int_{-1}^{y}\exp[ia_{n''}(-y+\tau)]\,g(\tau)\,d\tau+\frac{1}{2\pi}\int_{\pi}^{2\pi}\left(\frac{g(y)}{ice^{i\phi}}+O\left\{\frac{1}{c^2}\right\}\right)\times ce^{i\phi}\,id\phi,\quad -1<y<+1, \tag{27}$$

and

$$I_2(y,c)=i\sum_{n'}(a_{n'}-a_{n'}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n'}}\frac{a_{n'}-a_n^*}{a_{n'}-a_n}\times\int_y^1\exp[ia_{n'}(-y+\tau)]\,g(\tau)\,d\tau+\int_\pi^0\left(\frac{g(y)}{-ice^{i\phi}}+O\left\{\frac{1}{c^2}\right\}\right)cie^{i\phi}d\phi,\quad -1<y<+1, \tag{28}$$

Hence if c tends to infinity we derive from (27) and (28)

$$\begin{aligned}\lim_{c\to\infty}\{I_1(y,c)+I_2(y,c)\}&=g(y)-i\sum_{n''}(a_{n''}-a_{n''}^*)\times\prod_{\substack{\{n_1\}\\ n_1\neq n''}}\frac{a_{n''}-a_n^*}{a_{n''}-a_n}\int_{-1}^y\exp[ia_{n''}(-y+\tau)]\,g(\tau)\,d\tau\\&+i\sum_{n'}(a_{n'}-a_{n'}^*)\times\prod_{\substack{\{n_1\}\\ n_1\neq n'}}\frac{a_{n'}-a_n^*}{a_{n'}-a_n}\int_y^1\exp[ia_{n'}(-y+\tau)]\,g(\tau)\,d\tau,\\&\qquad -1<y<+1.\end{aligned} \tag{29}$$

The relation (24) shows that the left-hand side of (29) is the Fourier transform of the entire function $h_1(z)$. Moreover, defining $g(\tau)$ to have the value zero if $|\tau|>1$ we derive in a similar way the result

$$\begin{aligned}\lim_{c\to\infty}\{I_1(y,c)+I_2(y,c)\}&=-i\sum_{n''}(a_{n''}-a_{n''}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n''}}\frac{a_{n''}-a^*}{a_{n''}-a_n}\times\int_{-1}^{+1}\exp[ia_{n''}(-y+\tau)]\,g(\tau)\,d\tau=0\quad\text{if } y>1,\\&=i\sum_{n'}(a_{n'}-a_{n'}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n'}}\frac{a_{n'}-a_n^*}{a_{n'}-a_n}\int_{-1}^{+1}\exp[ia_{n'}(-y+\tau)]\times g(\tau)d\tau=0\quad\text{if } y<-1.\end{aligned} \tag{30}$$

Hence the Fourier transform of the function $h_1(z)$ vanishes outside the interval $|x|\leq+1$. Therefore, the entire function $h_1(z)$ can be represented on the real axis and by, analytical continuation everywhere in the complex plane, by the formula

$$h_1(z)=\int_{-1}^{+1}\exp(izy)\,g_1(y)\,dy, \tag{31}$$

where

$$g_1(y)=\int_{-\infty}^{+\infty}\exp(-ixy)\,h_1(y)\,dy. \tag{32}$$

Combination of (24), (29), and (32) yields

$$\begin{aligned}g_1(y)=g(y)&-i\sum_{n''}(a_{n''}-a_{n''}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n''}}\frac{a_n-a_n^*}{a_{n''}-a_n}\times\int_{-1}^y\exp[ia_{n''}(-y+\tau)]\,g(\tau)\,d\tau\\&+i\sum_{n'}(a_{n'}-a_{n'}^*)\prod_{\substack{\{n_1\}\\ n_1\neq n'}}\frac{a_{n'}-a_n^*}{a_{n'}-a_n}\times\int_y^1\exp[ia_{n'}(-y+\tau)]\,g(\tau)\,d\tau.\end{aligned} \tag{33}$$

APPLICATION TO MICROSCOPY

The preceding calculations have shown that the phase of a band limited function is not uniquely determined by its modulus along the real axis. Hence *a priori* information has to be used in order to obtain an unambiguous relation between the modulus and phase. We will now show that if we apply the preceding theorems to microscopy we have such a required *a priori* information, especially about the function $g(y)$ of Eq. (1).

Let us consider image formation by a microscope, free of aberrations, of a monochromatically illuminated object.

We know from the preceding analysis that several field distributions $g(y)$ in the Fraunhofer plane lead to the same intensity distribution $|h(x)|$ in the Gaussian image plane. However, we have the *a priori* information that, according to (1a), $g(y)$ is *band limited* and that the number of zeros of $h(z)$ located in the upper half of the complex plane is finite. The band limitation of the unperturbed wavefunction $g(y)$ is due to the imaging properties of the microscope [Eqs. (1a) and (1b)] and is therefore valid for any object imaged by a microscope whereas in general the number of zeros of $h(z)$ may be finite or infinite in the upper half of the complex plane [e.g. consider the example discussed in the introduction which shows that if a finite number of the zeros of $h(z)$ are located in the upper half of the complex plane, the band limited function $h^*(z^*)$ having the same modulus as $h(z)$ on the real axis, has an infinite number of zeros located in the upper half of the complex plane]. Recalling Eq. (6b),

$$a_n\sim n\pi+\frac{i}{2}\ln\left(\frac{g(-1)}{g(1)}\right), \tag{6b}$$

we deduce that the number of zeros of $h(z)$ located in the upper half of the complex plane is finite if $|g(-1)|<|g(1)|$. Hence, in microscopy, just by measuring the intensity in the end points of the Fraunhofer plane we can decide whether the condition $|g(-1)|<|g(1)|$ is valid or not. If $|g(-1)|>|g(1)|$, we can still apply the theory of this paper considering the function $h(-x)$ instead of $h(x)$, because using (1b)

$$h(-x)=\int_{-1}^{+1}\exp(ixy)\,g(-y)\,dy.$$

It is even possible to ensure that the condition $|g(-1)|<|g(1)|$ is valid *a fortiori* if, as usual in electron microscopy, we are dealing with *weak* objects, i.e., objects which only slightly perturbs the illuminating plane wave. Because in this case we can choose the angle of

incidence of the illuminating wave in such a way that the maximum of the diffraction spot coincides with the rim of the diaphragm. It will now be shown that only one function out of set of all possible distributions in the Fraunhofer plane is consistent with our *a priori* knowledge and, moreover, can be easily determined.

According to (34), $g(y)$ is an entire function. Therefore the rhs of (33) is an entire function, which is equal to the entire function $g_1(y)$ in the interval $-1 < y < +1$. Hence, by the principle of analytical continuation, Eq. (33) holds everywhere in the complex plane. Recalling that the numbers $a_{n'}$ and $a_{n''}$ have nonzero imaginary parts, it is apparent from (33) that in general $g_1(y)$ diverges if y tends to either $+\infty$ or $-\infty$. However, one and only one of all the possible field distributions [viz. $g(y)$] is bounded, as required by (34) and the asymptotic expansion (3), if y tends to either $+\infty$ or $-\infty$. Inspection of (33) shows that this will be the case if $g_1(y) = g(y)$.

Hence just by inspection of the asymptotic behavior of all the possible field distributions in the Fraunhofer plane and by using the *a priori* information of band limitation and hence boundedness, we can uniquely determine the field distribution in the Fraunhofer plane. Inserting this uniquely determined function in (35) yields the unique solution to the phase retrieval problem.

If *a priori* information of $g(y)$ is not available or at least not in a form which can be treated analytically, additional information could be obtained by making a second exposure, for different defocusing. We will now show that these two measurements are sufficient to allow us to determine $g(y)$ up to an overall phase. In one dimension the relation between the image wavefunction $h(x)$ and the wavefunction $g(y)$ in the Fraunhofer plane is

$$h_k(x) = \int_{-1}^{+1} \exp\left(ixy + \frac{\Delta z_k y^2}{f^2}\right) g(y)\, dy, \quad k = 1, 2, \tag{34}$$

where Δz_k is the distance between the defocused observation plane and the Gaussian image plane.

Inserting all possible combinations of Blaschke factors into the dispersion relation (3) we obtain from both measurements ($k = 1, 2$) two sets of functions $\{h_1^{(1)}(y)\}$ and $\{h_1^{(2)}(y)\}$ and by Fourier inversion two sets of functions $\{g_1^{(1)}(x)\}$ and $\{g_1^{(2)}(x)\}$. We know from (34) that only those functions for which

$$\frac{g_1^{(1)}(y)}{g_1^{(2)}(y)} = \exp\frac{i}{f^2}(\Delta z_1 - \Delta z_2)y^2, \tag{35}$$

are consistent with our *a priori* knowledge that both exposures are taken with two different values of the defocusing.

Using (33), condition (35) yields

$$g(y)\exp\left(\frac{i\Delta z_1}{f^2} y^2\right) - i\sum_{n''}(a_{n''} - a_{n''}^*) \prod_{\substack{\{n_1\} \\ n_1 \neq n''}} \frac{a_{n''} - a_n^*}{a_{n''} - a_n}$$

$$\times \int_{-1}^{y} \exp\left(ia_{n''}(-y + \tau) + \frac{i\Delta z_1}{f^2}\tau^2\right) g(\tau)\, d\tau$$

$$+ i\sum_{n'}(a_{n'} - a_{n'}^*)$$

$$\times \prod_{\substack{\{n_1\} \\ n_1 \neq n'}} \frac{a_{n'} - a_n^*}{a_{n'} - a_n} \int_{y}^{1} \exp\left(ia_{n'}(-y + \tau) + i\frac{\Delta z_1}{f^2}\tau^2\right) g(\tau)\, d\tau$$

$$= \left(\exp\frac{i}{f^2}(\Delta z_1 - \Delta z_2)y^2\right)\left[g(y)\exp\left(i\frac{\Delta z_2}{f^2} y^2\right)\right.$$

$$- i\sum_{n''}(b_{n''} - b_{n''}^*) \prod_{\substack{\{n_1\} \\ n_1 \neq n''}} \frac{b_{n''} - b_n^*}{b_{n''} - b_n} \int_{-1}^{y} \exp\left(ib_{n''}(-y + \tau)\right.$$

$$\left. + i\frac{\Delta z_2}{f^2}\tau^2\right) g(\tau)\, d\tau + i\sum_{n'}(b_{n'} - b_{n'}^*) \prod_{\substack{\{n_1\} \\ n_1 \neq n'}} \frac{b_{n'} - b_n^*}{b_{n'} - b_n}$$

$$\left. \times \int_{y}^{1} \exp\left(ib_{n'}(-y + \tau) + i\frac{\Delta z_2}{f^2}\tau^2\right) g(\tau)\, d\tau\right], \tag{36}$$

where the numbers $a_{n'}$ and $b_{n'}$ denote the zeros of the functions $h_1(z)$ and $h_2(z)$ in the upper half of the complex plane, and the numbers $a_{n''}$ and $b_{n''}$ the zeros of the functions $h_1(z)$ and $h_2(z)$ in the lower half of the complex plane.

Equation (36) is identically satisfied if none of the summations appear, i.e., if $g_1(y) \equiv g(y)$. Otherwise, as in the discussion following Eq. (33), we derive from Eq. (3) that if y tends to infinity the left-hand side of Eq. (36) is $O\{\exp(+c_1 y)\}$ if $y \to +\infty$, or $O\{\exp(+c_2 y)\}$ if $y \to -\infty$, where $c_1 = \max\{\mathrm{Im}\, a_{n'}\}$ and $c_2 = \max\{\mathrm{Im}\, a_{n''}\}$, whereas the right-hand side of Eq. (36) is $O\{\exp i(\Delta z_1 - \Delta z_2)y^2 + c_3 y\}$ if $y \to \infty$ or $O\{\exp(i(\Delta z_1 - \Delta z_2)y^2 + c_4 y)\}$ if $y \to -\infty$, where $c_3 = \max\{\mathrm{Im}(b_{n'})\}$ and $c_4 = \max\{\mathrm{Im}(b_{n''})\}$. Hence, Eq. (36) only can be satisfied if none of the summations appear. Therefore, only $g(y)$ satisfies condition (35), and can be determined up to a constant by the following procedure. Divide each function belonging to the set of functions $\{g_1^{(1)}(y)\}$ by every function belonging to the set of functions $\{g_1^{(2)}(y)\}$ and test if condition (35) is satisfied. Then one and only one pair of functions will be found satisfying (37), and these are the functions which determine the unknown function $g(y)$. Having determined $g(y)$, which is the main goal of image reconstruction, $h(y)$ and $\arg h(y)$ can be calculated. This provides the required solution to the phase reconstruction problem.

DISCUSSION

The preceding calculations not only prove the uniqueness of phase reconstruction of optical images but also provides an explicit procedure for calculating the phase. However, the question of stability is not considered and the procedure might be very sensitive to noise and errors in measurements. Hence computer simulated calculations should perhaps be employed to indicate the feasibility of the procedure or the need of another algorithm.

Uniqueness was obtained by using *a priori* information about the unperturbed wavefunction $g(y)$ in the Fraunhofer plane, namely that this wavefunction is band

limited and that we know that either $|g(-1)| < |g(1)$ or $|g(-1)| > |g(1)|$. Another *a priori* bit of information which could be used would be the knowledge of the quotient of two functions calculated from two images corresponding to two different settings of the defocusing. Gerchberg and Saxton[7] suggested that knowledge of the modulus of the wavefunction in the Fraunhofer plane determines uniquely the phases of both this wavefunction and of the image wavefunction. Computer simulated calculations sustained their claim. The results of this paper give additional support to their hypothesis.

Our theory is one dimensional. However, since micrographs are essentially two dimensional, a two dimensional extension of our theory is required. Clearly such a theory will be more complicated than the one presented here. For example flipping of zeros expressed by the use of Blaschke factors will then in general not apply, as is obvious from Weierstrass's preparation theorem, Osgood.[14]

These points will be discussed in a future publication.

ACKNOWLEDGMENT

The author wishes to thank Professor E. Wolf for the many stimulating discussions and critical comments and Dr. H. A. Ferwerda for his aid in revising this paper.

*This research was carried out during the tenure of a fellowship from the Netherlands Organization for the Advancement of Pure Research, (Z.W.O.) and was also supported in part by the Army Research Office (Durham).

†Present address: Technical Physical Laboratories, State University at Groningen, Nijenborgh 18, Groningen 8002, The Netherlands.

[1]E. Wolf, Proc. Phys. Soc. 80, 1269 (1962).
[2]H.M. Nussenzveig, J. Math. Phys. 8, 561 (1967).
[3]A. Walther, Opt. Acta 10, 41 (1963).
[4]D. Kohler and L. Mandel, J. Opt. Soc. Am. 63, 126 (1973).
[5]D. Dialetis and E. Wolf, Nuovo Cimento 47, 113 (1967).
[6]M. Born and E. Wolf, *Principles of Optics* (Pergamon, New York, 1965), especially Sec. 8.6.3.
[7]R.W. Gerchberg and W.O. Saxton, Optik 34, 275 (1971).
[8]D.L. Misell, R.E. Burge, and A.H. Greenaway, J. Phys. D 7, 27 (1974). D.L. Misell, J. Phys. D 7, 69 (1974), 7, 832 (1974).
[9]E.C. Titchmarsh, Proc. Lond. Math. Soc. 25, 283 (1926).
[10]M.L. Cartwright, Q. J. Math. Oxford, Ser. (1) 1, 38 (1930) and 2, 113 (1931).
[11]H.A. Ferwerda and B.J. Hoenders, Optik 37, 542 (1973).
[12]For an explicit procedure see H.A. Ferwerda and B.J. Hoenders, Optik 40, 14 (1974).
[13]We substituted the asymptotic expansions (25) and (26) into (27) and (28) although (25) is not valid in the sectors $\pi \leq \arg z \leq \pi + \delta$ and $2\pi - \delta \leq \arg z \leq 2\pi$ and (26) is not valid in the sectors $0 \leq \arg z \leq \delta$ and $\pi - \delta \leq \arg z \leq \pi$, where δ is an arbitrarily small positive number. However, it is an immediate consequence of (3) that the left-hand sides of (25) and (26) are $O\{1/z\}$ if $|z|$ tends to infinity in the *closed* sectors $\pi \leq \arg z \leq 2\pi$, resp. $0 \leq \arg z \leq \pi$, and that therefore the error made by substituting (25) and (26) into (27) and (28) as if they were valid for *closed* sectors can be made arbitrarily small.
[14]W.F. Osgood, *Lehrbuch der Funktionentheorie*, Vol. II, 1, Chap. 2, Sec. 2.
[15]This follows immediately from the following theorem which can be found in R.A.C. Paley and N. Wiener, *Fourier Transforms in the Complex Domain* (Am. Math. Soc., New York, 1934), especially Chap. 3, Sec. 6: Suppose that an entire function $f(z) \in L^2$ along the real axis and suppose there exists a real number A such that $f(z) = 0$ $(\exp A|z|)$. Then there exists a function $F(x) \in L^2$ such that

$$f(z) = \int_{-A}^{+A} \exp(izx) F(x)\, dx.$$