# On the phase retrieval problem in optics

T. I. Kuznetsova

*P. N. Lebedev Physics Institute, Academy of Sciences of the USSR, Moscow*
Usp. Fiz. Nauk **154**, 677–690 (April 1988)

Studies on the retrieval of a complex function from data on its modulus and the modulus of its Fourier spectrum are briefly reviewed. An analog optical method of solving this same problem is described. Pictorial explanations are presented concerning the ambiguity of the retrieval of a spectrally bounded one-dimensional function from a single modulus. Some methods of retrieval of time characteristics of radiation are analyzed from the standpoint of the phase retrieval problem that rests on spectral measurements or on measurements of the autocorrelation function.

**CONTENTS**



## 1. INTRODUCTION

The problem of obtaining the phase characteristics of light fields arises in various optical studies. The difficulties of directly measuring the phase in the optical range compel optical physicists to seek roundabout paths: to try to extract phase information from data on the intensity. Of course, attempts to obtain here simple recipes for solving the problem were doomed to failure. However, the past 15 years have marked a significant advance in the problem of retrieval of the phase characteristics of light fields. Studies have appeared in which the retrieval of phase characteristics from the characteristics of the intensity is actually achieved by computational technique. It is important to stress that additional data on the field are introduced here, e.g., in many cases one uses two (rather than one) intensity distributions pertaining to two cross sections of the field.

One could speak to the same extent of progress in processing data of x-ray structure analysis or electron microscopy as of the advances in optics. However, we shall restrict the treatment here to optical problems.

This new line of studies of the phase retrieval problem was founded by Ref. 1. One can become acquainted with the status of the phase retrieval problem in the preceding period, e.g., in Refs. 2–4. We shall proceed below to outline Ref. 1.

## 2. THE GERCHBERG-SAXTON ALGORITHM AND OTHER SIMILAR METHODS OF SOLVING THE PHASE PROBLEM

For retrieving phase information, Gerchberg and Saxton were the first to draw on intensity data pertaining to two planes, rather than one. An iteration method of finding the phase proved especially successful for them,[1] which consists in the following.

Let us assume that the initial monochromatic field bearing information on the object of study has been recorded in two planes: in the image plane and in the Fourier plane. We shall assume (in line with the starting assumptions of Ref. 1) that the field depends only on one space coordinate. Let us denote the field in the image plane by $\mathscr{E}(x) = a(x)\exp(i\Phi(x))$, and that in the Fourier plane as $\mathscr{F}(x) = A(x)\exp(i\Phi(x))$. We shall introduce the following symbol for the Fourier transformation operation:

$$L\mathscr{E}(x) = \frac{1}{\sqrt{2\pi}}\int \mathscr{E}(x')\exp(-ixx')\,dx'.$$

Thus we have

$$F(x) = L\mathscr{E}(x).$$

Assume that in recording the field we obtain information on the functions $a(x)$ and $A(x)$, whereas information on the phase factors is lacking. The problem consists in constructing a complex function from its given modulus and the modulus of its Fourier image. To solve this problem Gerchberg and Saxton proposed an iteration procedure that consists in the following. As the trial function they took a function whose modulus coincided with the given modulus in the image plane $a(x)$, while the phase factor $\exp(i\psi(x))$ was taken arbitrarily; in the study it was constructed by using a random-number generator. The trial function is conveniently designated as $y_0(x)$; henceforth the subscript will coincide with the number of performed iterations, while before starting the iterations we have

 0038-5670/88/040364-08$01.80 © 1988 American Institute of Physics 

$$y_0(x) = a(x)\exp(i\psi(x)).$$

The Fourier transform was constructed for this function:

$$Ly_0 = A_0(x)\exp(i\Phi_0(x)).$$

Then the modulus of the obtained function was replaced by the correct value, i.e., $A(x)$, while keeping the phase. Thus the following function was obtained in the Fourier plane:

$$Y_0(x) = A(x)\exp(i\Phi_0(x)),$$

which can be represented as

$$Y_0 = \frac{ALy_0}{|Ly_0|}.$$

Then an inverse Fourier transformation was performed and the modulus was again corrected. Here they obtained a function in the image plane in the first approximation

$$y_1 = \frac{aL^{-1}Y_0}{|L^{-1}Y_0|}.$$

In the next step the same transformations were carried out. One can represent the entire iteration procedure by the formulas

$$Y_n = \frac{ALy_n}{|Ly_n|}, \quad y_{n+1} = \frac{aL^{-1}Y_n}{|L^{-1}Y_n|}. \tag{1}$$

If the process proves to converge, then we shall denote the function obtained as the limit of iteration as $y$:

$$y = \lim_{n\to\infty} y_n.$$

In many cases the process actually converged, and the function $y$ coincided with the initial complex field $a\exp(i\psi)$. We shall present illustrative material in Sec. 3, where we shall describe our algorithm, which has much in common with the Gerchberg-Saxton algorithm. We note that the Gerchberg-Saxton method guaranteed neither the existence of a solution nor its uniqueness. Nevertheless the numerous successes in numerical experimentation have led to the widespread use of the algorithm by other authors. Moreover, the Gerchberg-Saxton algorithm began to be applied to several different problems, and modifications in it have arisen, due both to a change in formulation of the problems and to striving to accelerate the numerical procedure. We shall take up some of these modifications.

First of all we should mention here the Fienup algorithm, which is designed for retrieving a real, non-negative function from the modulus of its Fourier spectrum.[5] We stress that, in contrast to the Gerchberg-Saxton algorithm, here one need not know the modulus of the field in the object plane, i.e., the function $a(x)$. The features of the iteration procedure consist in multiplying the value of the function obtained in a successive step by zero, when recalculating from the Fourier plane to the image plane, if it does not satisfy the condition of non-negativity, but leaving it unchanged in the converse case. The non-negativity condition proves to be so strong that retrieval using it proceeds no more poorly in a number of cases than the retrieval of complex functions from the moduli given in two planes. We should note that the Fienup algorithm has been successfully applied, not only to one-dimensional, but also to two-dimensional functions.[6,7] Although at first glance two-dimensional problems seem more complicated than one-dimensional, and a large bulk of calculations is required at each iteration step in calculations for them, yet in the two-dimensional case the algorithms work more effectively. We shall turn to the peculiarities of two-dimensional problems in Sec. 4. A report appeared relatively recently in the literature on the retrieval of two-dimensional functions from the modulus of the Fourier spectrum, upon which no non-negativity condition was imposed. That is, the functions can be complex both in the image plane and in the Fourier plane.[8]

## 3. AN ALGORITHM FOR RETRIEVAL OF PHASE CHARACTERISTICS MODELING THE ACTION OF AN OPTICAL SYSTEM

Let us ask the question of whether one cannot perform the procedure described in Sec. 2 by using optical transformations. As is known, the operation of Fourier transformation is easily performed in optics by using a lens. One can easily multiply by functions (by $a(x)$ and $A(x)$) if these functions are presented in the form of the transmission function of transparencies. As regards the operation of equalizing the modulus of the complex amplitude, it can be performed only approximately. To do this, one must use a nonlinear element of the type of a darkening filter (e.g., an element with two-photon absorption or with generation of stimulated Raman scattering or an amplifier with saturable amplification). We have described[9,10] the design of an optical instrument for phase retrieval that combined the stated elements and transparencies in a ring resonator. The algorithm corresponding to the action of this optical instrument is given by the formula

$$y_{n+1} = \beta L^{-1}AN(\beta LaN(y_n)). \tag{2}$$

Here $y_n$ and $y_{n+1}$ are the fields in successive iterations, $\beta$ is a constant that denotes the amplification coefficient, and the functions $a$ and $A$ and the operator $L$ are defined in Sec. 2. The symbol $N$ denotes the nonlinear operator that smoothes the intensity distribution, leaving the phase distribution unchanged, and which has the form

$$N(y) = \frac{y}{(1+yy^*)^{1/2}}. \tag{3}$$

This nonlinear transformation corresponds to a nonlinear element in which the losses arise from two-photon absorption or generation of a second optical harmonic.

We should call attention to the fact that we have $N(y) = y/|y|$ at very large fields ($|y|^2 \gg 1$), and the algorithm (2) goes over into algorithm (1). In our calculations we began with small intensities at which the nonlinearity did not alter the field. That is, at the start of the calculations we had $N(y) \approx y$. Depending on the coefficient $\beta$, either the intensity of the field increased during the iterations, which led to "turning on" the nonlinearity, or the intensity declined from step to step, so that the nonlinearity did not operate throughout the entire iteration procedure. In this latter case, which we called the prethreshold case, a stable field distribution along the $x$ coordinate was attained after 8–10 iterations, while the mean value declined from iteration to iteration at a constant rate. In the prethreshold regime the established distribution $y(x)$ and the initial field $a(x)\exp(i\varphi(x))$ had an overlap integral $\eta = 0.75$–$0.90$. The overlap integral is defined by the formula

$$\eta = \frac{|\langle ya\exp(-i\varphi)\rangle|}{(\langle yy^*\rangle\langle a^2\rangle)^{1/2}}.$$

Here the angle brackets denote averaging over the $x$ coordinate. The established distribution did not depend on the trial function with which the calculations began. A defect of the

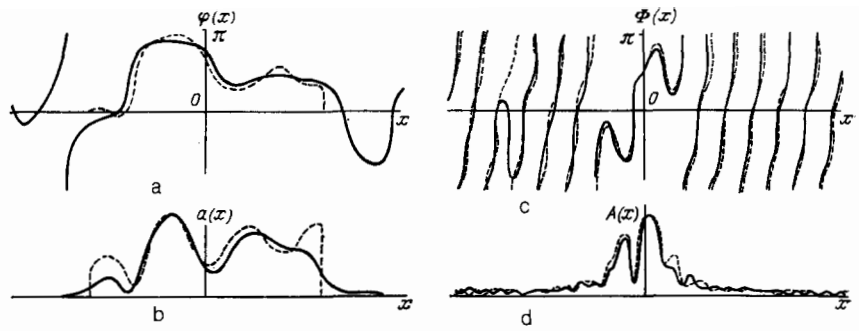


FIG. 1. An example of retrieval of a complex field from two moduli. Dependence of the phase and modulus of the field on the coordinate: a, b—in the image plane, c, d—in the Fourier plane. The dotted lines indicate the starting function, and the solid lines the retrieved function.

system was the low value of the overlap integral. An appreciably better quality of retrieval was achieved in the nonlinear (or superthreshold) regime. In this case, which was achieved at larger values of the coefficient $\beta$, the following occurred. At the very start of the calculations, along with a change in the structure of the field, the mean value of the intensity increased. Then, after several iterations, the nonlinearity was turned on, the increase in intensity decelerated, and finally a stationary value of the intensity and a stable distribution of the field in $x$ were established. Here a considerably higher value of the overlap integral of the functions $y(x)$ and $a(x)\exp(i\varphi(x))$ was attained than in the prethreshold regime. Usually—without any careful choice of the values of $\beta$ and $y_0\,y_0^*$—values of the overlap integral $\eta = 0.997$ were easily attained and were established in 15–20 iterations. In Ref. 1 the necessary number of iterations was of the order of a hundred. Just as in Ref. 1, we retrieved one-dimensional complex functions. The starting fields were constructed by using a random-number generator and smoothing operations over the coordinate. The diagram shows an example of an object field and the result of retrieval (in the superthreshold regime). The high rate of convergence of our algorithm is attained by the linear stage in which an approximate solution is rapidly formed, and is then refined during the nonlinear stage.

## 4. MATHEMATICAL PROBLEMS ASSOCIATED WITH THE PHASE RETRIEVAL PROBLEM

In parallel with the studies on computational methods of solving the phase problem, a large number of studies have been published in the literature devoted to analytical methods of studying the problem.

One can find the most complete presentation of the results obtained here in Refs. 11 and 12. We shall give some of them below. We shall present the current views on the ambiguity of the solution of the phase retrieval problem, i.e., on the question of the degree of uncertainty with which the phase distribution is retrieved from a single intensity distribution. The origin of studies on this problem goes back to Refs. 13 and 14. The current status of the studies is found in Refs. 11 and 12. In essence we shall follow Refs. 11 and 12, but shall present the problem in a simplified fashion that does not require recourse to the apparatus of the theory of analytic functions. Simplifications arise because complex fields of less general form are taken than in Refs. 11 and 12. To a certain extent our approach resembles that used in Ref. 15.

We shall treat monochromatic fields that present the information of any type of stationary objects; the information consists in the transverse structure of the fields. First of all we should emphasize that the fields studied in optics belong to a definite class of functions—functions with a bounded spectrum. Whatever the object whose image we are studying, whatever its true spectrum of spatial frequencies—only a finite spectral band always takes part in forming the image, owing to the finite apertures of the optical elements. And even in the absence of aperture restrictions, restrictions arise involving the filtering properties of free space (the higher spatial harmonics are evanescent waves and the information that they bear disappears on going away from the object to a distance greater than the wavelength of the radiation). Let the field of a monochromatic source illuminate some object of finite extent. Let us denote the field in a plane lying in the immediate vicinity of the object as $\mathscr{E}_0(x_0)$, and the Fourier image of the function $\mathscr{E}_0(x_0)$ as $\mathscr{F}_0(\xi)$,

$$\mathscr{F}_0(\xi) = \frac{1}{\sqrt{2\pi}} \int \mathscr{E}_0(x_0) \exp(-i\xi x_0)\, dx_0. \tag{4}$$

As we have already stated above, only part of the true spectrum having a finite support takes part in forming the image. We shall denote this part of the spectrum as $\mathscr{F}(\xi)$. We shall assume that

$$\begin{aligned} \mathscr{F}(\xi) &= \mathscr{F}_0(\xi) \quad \text{when} \quad |\xi| < b, \\ &= 0 \qquad \text{when} \quad |\xi| > b. \end{aligned} \tag{5}$$

We shall denote the image formed by this spectrum as $\mathscr{E}(x)$.

$$\begin{aligned} \mathscr{E}(x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathscr{F}(\xi) \exp(ix\xi)\, d\xi \\ &= \frac{1}{\sqrt{2\pi}} \int_{-b}^{b} \mathscr{F}(\xi) \exp(ix\xi)\, d\xi. \end{aligned} \tag{6}$$

Now we shall make the fundamental simplifying assumption. We shall assume that the spectral function $\mathscr{F}_0(\xi)$ can be represented on the segment $[-b,b]$ by a finite number of Fourier harmonics:

$$\mathscr{F}_0(\xi) = \sum_{m=-n/2}^{n/2} G_m e^{im(\pi/b)\xi}. \tag{7}$$

Here the $G_m$ are arbitrary complex constants. The spectral function of the image, $\mathscr{F}(\xi)$, which is connected to $\mathscr{F}_0(\xi)$ by Eq. (5), is described by the product

$$\mathscr{F}(\xi) = \operatorname{rect}\frac{\xi}{2b} \cdot \sum_{m=-n/2}^{n/2} G_m e^{im(\pi/b)\xi}. \tag{8}$$

Here we are using the step function that is often used in Fourier optics (see, e.g., Ref. 16):

$$\text{rect}\,\xi' = 1 \quad |\xi'| < 1,$$
$$= 0 \quad |\xi'| > 1.$$

In going via Eq. (6) from the spectrum $\mathcal{F}(\xi)$ to the image $\mathcal{E}(x)$, we obtain from (8)

$$\mathcal{E}(x) = \sum_{m=-n/2}^{n/2} G_m \frac{\sin(bx+\pi m)}{bx+\pi m}. \tag{9}$$

We can transform Eq. (9) to the form

$$\mathcal{E}(x) = \sin bx \cdot \sum_{m=-n/2}^{n/2} \frac{G_m(-1)^m}{bx+\pi m} = \sin bx \cdot \frac{Q_n(x)}{\prod_{m=-n/2}^{n/2}(bx+\pi m)}. \tag{10}$$

Here $Q_n$ is a polynomial of degree $n$ with complex coefficients. Upon denoting the roots of the polynomial as $z_j$ ($j = 1,2,\ldots, n$) and the coefficient of the highest power of the polynomial as $Cb^{n+1}$, we can write the following expression instead of (10):

$$\mathcal{E}(x) = C \sin bx \cdot \frac{\prod_{j=1}^{n}(x-z_j)}{x\prod_{m=1}^{n/2}\left(x-\frac{\pi m}{b}\right)\left(x+\frac{\pi m}{b}\right)}. \tag{11}$$

We see from Eq. (11) that—apart from a constant—the image $\mathcal{E}(x)$ is completely determined by the roots $z_j$ of the polynomial and the width $b$ of the spectrum. As is implied by (8) and (10), the number $n$ of roots of the polynomial is determined by the number of harmonics that take part in forming the image. We see from (6) that the number $n$ also determines the number of essentially different elements of the image, which we can call the number of its degrees of freedom. If now we go to the modulus of the field, or more exactly, to the square of the modulus, or intensity, then we can obtain the following representation for the intensity on the basis of (11):

$$I(x) = \mathcal{E}(x)\,\mathcal{E}^*(x)$$
$$= CC^*\left[\frac{\sin bx}{x\prod_{m=1}^{n/2}\left(x^2-\frac{\pi^2m^2}{b^2}\right)}\right]^2 \prod_{j=1}^{n}(x-z_j)(x-z_j^*). \tag{12}$$

We see from (12) that, if we know the function $I(x)$, then we can determine the roots $z_j$, $z_j^*$. Yet if we try to find the field $\mathcal{E}(x)$ from these roots, an ambiguity arises; two possibilities exist for each pair of complex conjugate roots $z_j$ and $z_j^*$; we can assume the root of the function $\mathcal{E}$ to be either $z_j$ or $z_j^*$. By going through all $n$ pairs of complex conjugate roots, we can construct $2^n$ different functions $\mathcal{E}(x)$. Thus the number of admissible solutions amounts to $2^n$. We emphasize that the addition to a given modulus of an arbitrary phase factor might cause the solution to lie outside the required spectral band. The number $2^n$ determines the degree of arbitrariness in choosing the solution. There are $2^n$ different functions compatible with both the given modulus and with the given spectral width.

We note that we might use the expansion

$$\sin bx = bx\prod_{m=1}^{\infty}\left(1-\frac{b^2x^2}{m^2\pi^2}\right)$$

to represent the field $\mathcal{E}(x)$ in the form

$$\mathcal{E}(x) = \text{const}\cdot\prod_{j=1}^{n}(x-z_j)\prod_{m=\frac{n}{2}+1}^{\infty}\left(x-\frac{\pi m}{b}\right)\left(x+\frac{\pi m}{b}\right). \tag{13}$$

If we examine the function $\mathcal{E}(x)$ in the complex plane of the variable $x$, we could say that it vanishes at $x = z_j, j = 1, 2,\ldots, n$ and at $x = \pm(\pi m/b)$, $m = (n/2)+1$, $(n/2)+2,\ldots$. The zeros at $x = \pm(\pi m/b)$ involve only the width of the spectral band and do not depend on the coefficients $G_m$. The zeros at $x = z_j, j = 1, 2,\ldots, n$ bear essential information on the field.

The ideas presented here are valid not only for functions whose Fourier image is given by Eq. (8), but are more general in character. The function is only required to have a bounded spectrum. It is known from the theory of analytic functions that functions having a bounded spectrum can be represented in the form

$$\mathcal{E}(x) = \text{const}\cdot\exp\left(i\frac{a+b}{2}x\right)\prod_{j=-\infty}^{\infty}(x-z_j). \tag{14}$$

Here the support of the spectrum is the segment $[a,b]$. One can find the needed explanations and literature references in Refs. 11 and 12, while a brief presentation of the problem exists in a set of monographs, e.g., in Ref. 17. As is known (see Refs. 11, 12), the form of the function $\mathcal{E}(x)$ depends substantially only on the position of a finite number of zeros. This number, $n$, involves the number of degrees of freedom of the function. However, in the general case this connection does not look so perspicuous as for our special case—for functions of the form of (13). The remaining zeros (except for the stated $n$ zeros), inevitably are found in one half-plane of the complex variable and lie at a constant distance from the real axis. In the asymptotic case, the successive zeros are shifted with respect to one another by a segment of length $2\pi/(b-a)$ parallel to the real axis. For a function of the type of (14) it is known that the retrieval of the function from its modulus gives an ambiguity of $2^n$. Moreover, it is stated (see Ref. 11) that, if we treat the continuation of the spectrum of such a function outside the aperture $[a,b]$, then for only one of the $2^n$ functions will the spectrum decline with increasing modulus of the frequency on the real axis. However, the condition of decline of the spectral function at high frequencies cannot be introduced into iteration algorithms. Therefore it turns out in numerical calculations on the retrieval of complex functions that the data pertaining to one cross section of the field do not suffice, and one must introduce data pertaining to a second cross section.

Unfortunately, the rigorous theory as yet cannot indicate what volume of additional information is necessary for removing the ambiguity. An opinion exists in the literature, based only on the experience of working with algorithms, that singlevaluedness is made possible in the "two-moduli" problem, and also in many cases of retrieval of a real, nonnegative function.[18]

The connection between the phase-retrieval problem and seeking the positions of the zeros of the complex function was elucidated even in the early studies, which did not employ Eq. (14), but relied on the Hilbert transformation (see, e.g., Ref. 2, p. 625). Let us present a formula from the review of Ref. 2 which expresses the connection between the phase and the modulus of the complex function:

$$\varphi(x)=\frac{x}{\pi}P\int_{-\infty}^{\infty}\frac{|\mathscr{E}(x')|\,dx'}{x'(x-x')}+\frac{1}{i}\ln\prod_{j=1}^{n}\frac{x-z_j}{x-z_j^*}. \qquad (15)$$

Here it is assumed that the function $\mathscr{E}(x)$ is regular in the upper half-plane of the complex variable $x$, the $z_j$ are the zeros of the function $\mathscr{E}(x)$ lying in the upper half-plane, and $P$ is the symbol for the principal value of the integral. As we see from (15), this approach also requires solving the problem of reference to the zeros. Here we should bear in mind that, if the positions of the zeros are known, then the function can be obtained by using the Hadamard product (14), so that recourse to the Hilbert transform becomes superfluous.

In the phase retrieval problem the change from one to two or more dimensions is very interesting. The features of two-dimensional phase retrieval problems have been studied by many authors. Among the first studies on this topic we cite Ref. 15, and among the later ones, Refs. 19 and 20 (see also the references in Refs. 18 and 20). The behavior of a function of two variables having a bounded spectrum is also determined by its zeros. However, in contrast to the one-dimensional case, the zeros are no longer isolated points and cannot be moved about individually. Therefore the source of ambiguity is liquidated. Let us present the formulas for the two-dimensional case analogous to (9), (10), and (12); we obtain

$$\mathscr{E}(x,\,y)=\sum_{m=-n/2}^{n/2}\sum_{l=-n/2}^{n/2}G_{ml}\frac{\sin(bx+\pi m)}{bx+\pi m}\frac{\sin(by+\pi l)}{by+\pi l}, \qquad (9a)$$

$$\mathscr{E}(x,\,y)=\frac{C\sin bx\cdot\sin by}{\prod_{m=-n/2}^{n/2}\left(x+m\frac{\pi}{b}\right)\prod_{l=-n/2}^{n/2}\left(y+l\frac{\pi}{b}\right)}Q_{n,n}(x,\,y), \qquad (10a)$$

$$I(x,\,y)=CC^*\left[\frac{\sin bx\cdot\sin by}{\prod_{m=-n/2}^{n/2}\left(x+m\frac{\pi}{b}\right)\prod_{l=-n/2}^{n/2}\left(y+l\frac{\pi}{b}\right)}\right]^2\times Q_{n,n}(x,\,y)\,Q_{n,n}^*(x,\,y). \qquad (12a)$$

Equation (10a) (in contrast to (10)) contains a polynomial in $x$ and $y$. The probability that the polynomial $Q_{n,n}$ will prove reducible for arbitrary coefficients $G_{ml}$ is small. Therefore most often a given intensity distribution will correspond to only two functions for the field, which can be transformed into one another by taking the complex conjugate. This evident ambiguity usually does not cause difficulties, and as a rule, is not even mentioned in the literature. We note that, in studies on image retrieval, one commonly speaks of the "form of the image"[8] and considers functions equivalent on this level that can be obtained from one another by taking the complex conjugate, shifting, or inverting. Thus in the two-dimensional case the phase is associated with the intensity "almost always" unambiguously; the exceptions amount to functions for which the dependence on the two coordinates can be separated, and other factorable functions.[20] These exceptions are discussed, e.g., in Ref. 21, where ambiguity is discussed that is not removed even by the condition of non-negativity.

We note that, in the presented features of the two-dimensional phase retrieval problem, an analogy is traced with the problem of dislocations of a wave front.[22–25] In the dislocation problem the change from one to two transverse coordinates also leads to qualitative special features due to change in the properties of the zeros of the field (in real, rather than complex, space). The studies existing up to now enable us as yet to speak only of a superficial analogy.

In closing the presentation of this section, let us mention another mathematical problem—the dependence of solutions of the phase problem on small deviations in the starting data. Since noise always exists in an experiment, the recorded field intensity distribution differs from the true value, and to some extent these differences will affect the retrieval of the phase distribution. The problem of the role of noise has been posed in a number of studies on phase retrieval.[18,26–29] In other optical problems the studies of stability and methods of regularization are being ever more widely applied (see Ref. 30). In the phase problem the effect of noise yet requires quantitative study.

## 5. SOME APPLICATIONS OF THE PHASE RETRIEVAL PROBLEM. THE KINOFORM

The problem of retrieving the dependence on the coordinate of the phase of the field is associated with various applied problems. Among them the problems most fully dealt with in the literature are those that arise in the construction of images in astronomy. Here in a number of cases signals are recorded in such a way that the data processing is reduced to retrieving a non-negative function from the modulus of its Fourier spectrum. Here it proves expedient to use the Fienup algorithm.[5,6] Studies along this line have been reflected in many collected volumes and monographs (see, e.g., Refs. 31–33). Therefore we shall not treat them here.

We shall call attention to a problem that has not been described in the literature so thoroughly. This is the problem of influencing the spatial characteristics of radiation. It consists in calculating and constructing phase elements that could focus the radiation of a laser source into a given line, figure, or some previously prescribed spatial pattern. Such phase elements are commonly called kinoforms.[34] If we restrict the treatment to the mathematical side of the problem, we can show that the synthesis of a kinoform solves the problem of selecting the phase function from two intensity distributions. In the near zone one takes a uniform distribution (usually piecewise-constant), or sometimes a Gaussian profile $\exp(-r^2/r_0^2)$, and a given complicated spatial pattern in the far zone. Formally we have here the same starting data as in the problem of retrieving the phase of a field whose intensity is recorded in two planes. Therefore one applies algorithms of the Gerchberg-Saxton type in constructing kinoforms. One can find references to the use of the algorithms in Ref. 35. Moreover, there is another approach to calculating kinoforms (see, e.g., Ref. 36).

However, there is an important distinction between the phase problem and the problem of synthesizing a kinoform. In the case of phase retrieval one always knows that a solution exists. Actually, in the phase problem we have two intensity distributions obtained for an object field that has really existed: these two distributions can be called a coordinate pair. In the case of a kinoform it is not evident beforehand whether a complex field exists that contains both given intensity distributions.

Numerical calculations on kinoforms have shown that, if one requires to reproduce a spatial distribution in the far

zone with low accuracy, then usually one can construct a solution. However, as a rule, the problem of increasing the accuracy proves to be unworkable (the given input and output distributions are not a coordinate pair). This situation is described in Ref. 35, and a pathway of improving the characteristics of kinoforms is proposed: high accuracy in the central region of the output distribution is attained by admitting a certain arbitrariness in the peripheral region. Here an increase in the area of the peripheral region leads to a more exact result and to increased rate of convergence of the algorithm.

## 6. RETRIEVAL OF THE TIME CHARACTERISTICS OF RADIATION

The problems that we have discussed in the previous sections pertained to the spatial structure of light fields. Similar problems arise with regard to the time structure of radiation.

The development of lasers and nonlinear optics has posed the problem of studying radiation having unusual time characteristics, e.g., light pulses of very short duration, and pulses with strong phase modulation. The need has arisen of measuring the time characteristics with high accuracy. We recall that in 1967 indirect methods of recording radiation were proposed[37,38] and began to be applied that were based on measuring the autocorrelation functions of the intensity. In those years the electronic technology did not allow making direct measurements with the needed time resolution. Hence the indirect methods rapidly won many proponents. In subsequent years the direct methods strode far ahead. The most important advances in this field are presented, e.g., in Ref. 39. However, even now electronic technology lags behind the demands of laser physics, which now has shifted from the picosecond to the femtosecond range. In this regard, as before, the indirect methods of measuring the time characteristics are widely applied in laboratories. In using these methods, i.e., in processing autocorrelation functions in order to obtain the time characteristics of radiation, the same problems arise as in processing the spatial characteristics of fields directed toward retrieving the spatial distribution of the phase. However, these analogies and the methods applied for processing the spatial characteristics are not employed in the time studies. Moreover, in studying a number of problems bearing on time measurements, researchers have relied on false concepts, although analogous problems in the language of spatial functions have already been elucidated and presented in the literature. As an example, let us point out a reference to the principle of "maximum entropy" in Ref. 40, and the presentation of the potentialities of correlation methods in the monographs, e.g., Ref. 41. On this level it would be very useful to treat time measurements as one of the aspects of the phase retrieval problem.

We recall that one performs the following measurements in the correlation methods of studying radiation. The light beam to be studied is split into two beams with a controllable path difference, both beams are directed into a nonlinear medium in which an effect quadratic in the light intensity is excited, and on this basis one records the function

$$g(\tau)=\frac{1}{T}\int_0^T I(t)\,I(t+\tau)\,dt. \qquad (16)$$

Here $I(t)$ is the time-varying intensity of the radiation, $T$ is the total time of recording, and $\tau$ is the path difference of the two beams; the correlation function $g(\tau)$ is measured in relative (rather than absolute) units. A feature of the problem is the presence of two substantially different time scales: the value of $T$—the total duration (which we can associate with the period determined by the laser resonator, or in a system with a shutter that isolates one pulse from the train, we can simply relate it to the period), which amounts to $\approx 10^{-8}$ s, and the value of $\tau_{corr}$, which characterizes the peak width of the correlation function and is of the order of $10^{-12}$ s. In order not to complicate the treatment, we shall not deal with the fact here that one most often records experimentally the sum of the function $g(\tau)$ and a certain constant "background". A detailed treatment of this situation and the calculation of the proportionality constant relating the "background" to the quantity $g(0)$ for the case of two-photon luminescence can be found in Refs. 42–44. Thus in the simplest variant the problem consists in determining the function $I(t)$ from a measured function of the form of (16).

If we know in advance that the radiation amounts to a solitary pulse, then we can find from the peak width of the function $g(\tau)$ the duration of this pulse. Yet if additional information on the properties of the radiation is lacking, then it is not at all easy to prove from such measurements that the pulse was single. The difficulties involve the fact that substantially different signals (e.g., single-pulse and multipulse, or single-pulse with zero background and single-pulse on a broad pedestal) have correlation functions that differ little, while the differences can be lost owing to noise in recording. Here one requires a high accuracy of measurement of the function $g(\tau)$ and an appropriate processing of the measurements.

For example, let us examine, following Refs. 42–44, the problem of what accuracy is needed to estimate from the autocorrelation function the instantaneous power attained in the radiation being studied. In the general case the width of the peak of the autocorrelation function (which is usually all that is measured) yields only the characteristic scale of the intensity variation: if the peak width amounts to $\tau_{corr}$, then if we neglect coefficients $\sim\sqrt{2}$, we can assume that excursions of duration $\tau_{corr}$ exist in the time pattern of the radiation. The maximum instantaneous power would be attained if only one such excursion existed on a zero background; here the instantaneous power would exceed the mean by a factor of $T/\tau_{corr}$;

$$I_{inst}\approx\frac{I_{av}\,T}{\tau_{corr}}.$$

In lasers with self-synchronization of modes this increase can amount to $10^3$–$10^4$ ($T/\tau_{corr}\approx 10^3$–$10^4$). The maximum value corresponds to the case of complete mode synchronization. With incomplete synchronization the value of $I_{inst}/I_{av}$ can lie in the range from 1–2 to $10^3$–$10^4$.

Let us introduce a quantity with which we can characterize the value of the instantaneous power:

$$t_{eff}=\left(\int_0^T I(t)\,dt\right)^2\left(\int_0^{T_i} I^2(t)\,dt\right)^{-1}\equiv T\langle I\rangle^2\langle I^2\rangle^{-1}. \qquad (17)$$

Evidently we have $I_{inst}/I_{av}\sim T/t_{eff}$. We call the quantity $t_{eff}$ the effective duration; it gives the characteristic time inter-

val in which the greater part of the energy of the radiation is concentrated. Let us express the quantity $t_{\text{eff}}$ in terms of the correlation function $g(\tau)$. We obtain from Eq. (16)

$$\int_{-T}^{T} g(\tau)\,d\tau = \frac{1}{T}\left(\int_{0}^{T} I(t)\,dt\right)^2,$$

$$g(0) = \frac{1}{T}\int_{0}^{T} I^2(t)\,dt.$$

From the definition (17), this implies that

$$t_{\text{eff}} = \frac{1}{g(0)}\int_{-T}^{T} g(\tau)\,d\tau.$$

Now it is easy to associate the error in measuring the effective duration with the accuracy of measuring the correlation function:

$$\delta(t_{\text{eff}}) = \frac{T\delta g}{g(0)}. \tag{18}$$

If we must prove that $t_{\text{eff}} \approx t_{\text{corr}}$, then of course this error must not exceed $t_{\text{corr}}$: $\delta(t_{\text{eff}}) \leqslant t_{\text{corr}}$, or $T\delta g/g(0) \leqslant t_{\text{corr}}$. This means that the permissible error in measuring the correlation function amounts to the very small quantity $\delta g/g(0) \leqslant t_{\text{corr}}/T \approx 10^{-3}$. Besides, the demands on accuracy can be somewhat lowered if one makes direct measurements of the time course of the radiation intensity simultaneously with measuring the correlation function. Let the direct methods have a resolution $t_{\text{res}}$ and indicate the presence of a single pulse of duration $t_{\text{res}}$ in the period $T$. Then we can replace the quantity $T$ in the estimate (18) with $t_{\text{res}}$. Accordingly the demands on accuracy of measurement of the autocorrelation function acquire the form

$$\frac{\delta g}{g(0)} \leqslant \frac{t_{\text{corr}}}{t_{\text{res}}}.$$

Let us note that the problem of retrieving the time course of the intensity from the autocorrelation function can be treated as an example of the phase retrieval problem. In fact, let us introduce the spectrum of the intensity, i.e., the function $\tilde{I}(\Omega)$ with which the sought function $I(t)$ is associated by the transformation

$$I(t) = \frac{1}{\sqrt{2\pi}}\int \tilde{I}(\Omega)\,e^{i\Omega t}\,d\Omega,$$

Then by using the Wiener-Khinchin theorem or directly performing a Fourier transformation of (16), we can obtain

$$|\tilde{I}(\Omega)|^2 = \frac{1}{\sqrt{2\pi}}\int g(\tau)\,e^{-i\Omega\tau}\,d\tau.$$

Thus the problem contains information on the quantity $|\tilde{I}(\Omega)|$, i.e., on the modulus of the Fourier spectrum of the sought function, while the phase of the Fourier spectrum is unknown; Moreover, the intensity $I(t)$ is a non-negative function having the finite support $[0,T]$. This means that we have the problem of retrieving a non-negative function from the modulus of its Fourier spectrum. Thus the experience accumulated in solving similar problems can be used to find the time characteristics of radiation. Here we must bear in mind both the advances involving the application of the algorithms,[5,6] and the difficulties pertaining to the cases in which the solution is not unique.[21] Of course, application of the algorithms does not remove the demands on accuracy of measurement of the correlation function.

The methods of measuring time characteristics based on obtaining correlation functions deal in all their variants, not with the field $\mathscr{E}(t)$, but with the intensity of the field $I(t) = \mathscr{E}(t)\mathscr{E}^*(t)$. In their traditional information these methods are directed toward retrieving the time course of the modulus of the field $|\mathscr{E}(t)|$. Information on the modulus of the field is very important, and it often turns out to be all that is needed in experimental studies. However, in a number of cases in the passage of light pulses through a medium having nonlinear refraction, a strong frequency self-modulation arises.[45–48] Detailed information on the parameters of the modulation can prove essential for studying sequential processes in which light pulses take part.

Let us assume that the question arises of studying frequency self-modulation, i.e., of retrieving the time course of the phase of the field. The retrieval of the time course of the phase can be reduced to the problem discussed above of retrieving the phase from two moduli. Actually, let us obtain data on the time course of the radiation intensity and on the intensity of the spectrum of the field. Here we shall assume that the spectrum is recorded with an instrument having a high spectral resolution, as, e.g., in Ref. 49. Such data could be processed by using the Gerchberg-Saxton algorithm or the algorithm presented in Sec. 3. Of course, a necessary condition for applying an algorithm is a high accuracy of measurements and good resolution, both in the time pattern and in the spectral pattern. The problem of the admissible error of measurements is not clear at present. It is only clear that this problem is one of a number of other incorrectly posed problems that arise in optics.[30]

## 7. CONCLUSION

The major aim of this review consisted in drawing attention to the advances in studies of recent years in the field of the phase retrieval problem. The studies on the phase retrieval problem have been published mainly in specialized collected volumes devoted to the processing of optical images, and are insufficiently known to physicists of other specialities. A considerable fraction of the studies performed in this field is empirical in character and is directed toward developing and refining numerical algorithms for solving some particular variant of the phase retrieval problem. In these studies the results are of great interest that bear on the ambiguity of solutions in the one-dimensional case and the features that arise in going to the two-dimensional case. The mathematical studies on the phase retrieval problem as yet have yielded no final conclusions on the existence and singleness of solutions for a number of the algorithms that work in practice. The studies also lack a detailed analysis of the stability of solutions with respect to small changes in the starting data.

Nevertheless, despite the lack of a finished theoretical basis, the results of the studies on the phase retrieval problem find widespread application for solving many applied problems (restoration of images after passing through a turbulent atmosphere, control of the spatial characteristics of laser radiation). It is important that the advances that have been made here should be adopted in other fields of physical studies. It would be especially important to transfer the ideas formed in solving phase retrieval problems into the field of study of the time characteristics of radiation.

I express sincere gratitude to B. Ya. Zel'dovich and N. G. Preobrazhenskiĭ for discussing the problems presented here.

[1]R. W. Gerchberg and W. O. Saxton, Optik **35**, 237 (1972).
[2]L. Mandel and E. Wolf, Rev. Mod. Phys. **37**, 231 (1965) [Russ. transl. Usp. Fiz. Nauk **88**, 619 (1966)].
[3]H. M. Nussenzweig, *Causality and Dispersion Relations*, Academic Press, New York (1972) (Russ. transl., Mir, M., 1976).
[4]J. R. Klauder and E. C. G. Sudarshan, *Fundamentals of Quantum Optics*, Benjamin, New York (1968) (Russ. transl., Mir, M., 1970).
[5]J. R. Fienup, Opt. Lett. **3**, 27 (1978).
[6]J. R. Fienup, Appl. Opt. **21**, 2758 (1982).
[7]R. H. T. Bates and W. R. Fright, J. Opt. Soc. Am. **73**, 358 (1983).
[8]R. H. T. Bates and D. G. H. Tan, *ibid.*, Ser. A **2**, 2013 (1985).
[9]T. I. Kuznetsova and D. Yu. Kuznetsov, Kvantovaya Elektron. (Moscow) **12**, 2507 (1985) [Sov. J. Quantum Electron. **15**, 1661 (1985)].
[10]T. I. Kuznetsova and D. Yu. Kuznetsov, Kratk. Soobshch. Fiz., No. 2, 12 (1986) [Sov. Phys. Lebedev Inst. Rep. No. 2, 13 (1986)]; Opt. Commun. **61**, 374 (1987).
[11]H. A. Ferwerda, in *Inverse Source Problems in Optics*, ed. H. P. Baltes, Springer-Verlag, Berlin, 1978 (Topics in Current Physics, Vol. 9) (Russ. transl., Mashinostroenie, M., 1984).
[12]G. Ross, M. A. Fiddy, and M. Nieto-Vesperinas, in *Inverse Scattering Problems in Optics*, ed. H. P. Baltes, Springer-Verlag, Berlin, 1980 (Topics in Current Physics, Vol. 20).
[13]E. L. O'Neill and A. Wahlter, Opt. Acta **10**, 33 (1963).
[14]A. Wahlter, *ibid.*, p. 41.
[15]Yu. M. Bruck and L. G. Sodin, Opt. Commun. **30**, 304 (1979).
[16]J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, San Francisco, 1968 (Russ. transl., Mir, M., 1970).
[17]G. I. Vasilenko and A. M. Taratorin, Retrieval of Images (in Russian), Radio i svyaz', M., 1986.
[18]R. Barakat and G. Newsam, J. Opt. Soc. Am. Ser. A **2**, 2027 (1985).
[19]M. H. Hayes and J. H. McClellan, Proc. IEEE **70**, 197 (1982).
[20]J. L. C. Sanz and T. S. Huang, J. Opt. Soc. Am. **73**, 1446 (1983).
[21]J. R. Fienup, *ibid.*, Ser. A **3**, 284 (1986).
[22]N. B. Baranova and B. Ya. Zel'dovich, Zh. Eksp. Teor. Fiz. **80**, 1789 (1981) [Sov. Phys. JETP **53**, 925 (1981)].
[23]N. B. Baranova, B. Ya. Zel'dovich, A. V. Mamaev, N. F. Pilipetskiĭ, and V. V. Shkunov, Pis'ma Zh. Eksp. Teor. Fiz. **33**, 206 (1981) [JETP Lett. **33**, 195 (1981)].
[24]N. B. Baranova, B. Ya. Zel'dovich, A. V. Mamaev, N. F. Pilipetskiĭ, and V. V. Shkunov, Zh. Eksp. Teor. Fiz. **83**, 1702 (1982) [Sov. Phys. JETP **56**, 983 (1982)].
[25]N. B. Baranova, A. V. Mamaev, N. F. Pilipetsky, V. V. Shkunov, and B. Ya. Zel'dovich, J. Opt. Soc. Am. **73**, 525 (1983).
[26]J. R. Fienup, *ibid.*, p. 1421.
[27]J. L. C. Sanz, T. S. Huang, and F. Cukierman, *ibid.*, p. 1442.
[28]V. V. Aristov, A. I. Erko, Ch. V. Kopetskiĭ, S. M. Kuznetsov, and N. G. Ushakov, Opt. Spektrosk. **62**, 1105 (1987) [Opt. Spectrosc. (USSR) **62**, 653 (1987)].
[29]P. A. Bakut, A. A. Pakhomov, A. D. Ryakhin, K. N. Sviridov, and N. D. Ustinov, Dokl. Akad. Nauk SSSR **290**, 89 (1986) [Sov. Phys. Dokl. **31**, 710 (1986)].
[30]M. Bertero, C. DeMol., and G. A. Viano, see Ref. 12.
[31]C. van Schooneveld, ed., Image Formation from Coherence Functions in Astronomy, Reidel, Dordrecht, 1979 [Russ. transl., Mir, M., 1982].
[32]B. R. Frieden, ed., *The Computer in Optical Research*, Springer-Verlag, Berlin, N. Y., 1980 (Russ. Transl., Mir, M., 1983).
[33]N. D. Ustinov, I. N. Matveev, and V. V. Protopopov, Methods of Processing Optical Fields in Laser Location (in Russian), Nauka, M., 1983.
[34]R. J. Collier, C. B. Burckhardt, and L. H. Lin, *Optical Holography*, Academic Press, New York (1971) (Russ. transl., Mir, M., 1973).
[35]Hiroshi Akahori, Appl. Opt. **25**, 802 (1986).
[36]A. V. Goncharskiĭ, V. A. Danilov, V. V. Popov, A. M. Prokhorov, I. N. Sisakyan, V. A. Soĭfer, and V. V. Stepanov, Dokl. Akad. Nauk SSSR **273**, 605 (1983) [Sov. Phys. Dokl. **28**, 955 (1983)].
[37]J. A. Armstrong, Appl. Phys. Lett. **10**, 16 (1967).
[38]J. A. Giordmaine, P. M. Rentzepis, S. L. Shapiro, and K. W. Wecht, *ibid.* **11**, 216 (1967).
[39]M. Ya. Shchelev, Tr. Fiz. Inst. Akad. Nauk SSSR **155**, 3 (1985) [Proc. (Tr.) P. N. Lebedev Phys. Inst. Acad. Sci. USSR **155**, (1985)].
[40]T. Anderson and S. T. Eng, Opt. Commun. **47**, 288 (1983).
[41]S. Shapiro, ed., Ultrashort Light Pulses, Springer-Verlag, N. Y. (1977) [Russ. transl., Mir, M., 1981].
[42]T. I. Kuznetsova, On the Features of Some Methods of Recording the Duration of Ultrashort Light Pulses (in Russian), Preprint No. 47 of the P. N. Lebedev Physics Institute, Academy of Sciences of the USSR, Moscow, 1968.
[43]T. I. Kuznetsova, Zh. Eksp. Teor. Fiz. **5**, 2453 (1969) [Sov. Phys. JETP **8**, 1303 (1969)].
[44]T. I. Kuznetsova, Tr. Fiz. Inst. Akad. Nauk SSSR **84**, 62 (1975) [Proc. (Tr.) P. N. Lebedev Phys. Inst. Acad. Sci. USSR **84** (1975)].
[45]L. A. Ostrovskiĭ, Pis'ma Zh. Eksp. Teor. Fiz. **6**, 807 (1967) [JETP Lett. **6**, 260 (1967)].
[46]V. V. Korobkin, A. A. Malyutin, and A. M. Prokhorov, *ibid.* **12**, 216 (1970) [JETP Lett. **12**, 150 (1970)].
[47]R. H. Stolen and C. Lin, Phys. Rev. A **17**, 1448 (1978).
[48]E. M. Dianov, A. Ya. Karasik, A. M. Prokhorov, and V. N. Serkin, Izv. Akad. Nauk SSSR Ser. Fiz. **50**, 1042 (1986) [Bull. Acad. Sci. USSR Phys. Ser. **50**(6), 1 (1986)].
[49]V. I. Malyshev, A. V. Masalov, and A. A. Sychev, Zh. Eksp. Teor. Fiz. **59**, 48 (1970) [Sov. Phys. JETP **32**, 27 (1971)].

Translated by M. V. King