Ultramicroscopy 27 (1989) 413–422
North-Holland, Amsterdam

THE PHASE PROBLEM, MICRODIFFRACTION AND WAVELENGTH-LIMITED RESOLUTION – A DISCUSSION

J.M. RODENBURG

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK

Received 2 January 1989

If it were possible to assign phase to the microdiffraction plane available in a scanning transmission electron microscope (STEM) extremely high spatial resolution would be possible, limited only by the electron wavelength as opposed to the poor electron optics of the objective lens. This paper reviews the phase problem with respect to the microdiffraction plane.

1. Introduction

Gabor holography [1] suggests an elegant solution to the resolution limit imposed by spherical aberration in electron microscopy. This concerns itself with processing the central disc of a microdiffraction pattern available in a scanning transmission electron microscope (STEM) [2]. However, holography requires extremely stable experimental conditions in order to obtain a sufficiently broad angular coherence width across the illuminating beam. Perhaps a more experimentally feasible approach would be to employ a smaller-angle beam convergence and to process also the scattered intensity outside the central disc (the "dark-field" intensity) where correlation data exist, a subject which has been studied recently [3]. Microdiffraction then reduces to a classic phase problem. If it were possible to assign phase to every point in the microdiffraction plane, it would be possible to transform directly back to the wavefield that emanates from the specimen at whatever spatial resolution is required, limited only by the largest angle of diffraction that can be recorded. Assigning phase is not trivial, but there are many instrumental variables which, in principle, will make unique solution possible under certain circumstances.

This paper reviews the phase problem with respect to the microdiffraction plane. It is written under the assumption that it is experimentally possible to digitize the entire microdiffraction plane at TV rates (for example, by using a CCD array coupled to a YAG scintillator [4]), which gives access to very large quantities of diffraction data as a function of probe position. What should be done with all these data? Ideally, an inverse algorithm could solve for super-resolution specimen structure, preferably in real time so that the microscope parameters could be adjusted appropriately. Here, the possibility of unique solution is examined in order to suggest what may or may not be possible.

2. Solving for phase: a brief review

2.1. The problem

If it is possible to record the complex wave disturbance in the far field of a coherent scattering experiment, it is possible to construct a direct linear transform to find the wave distribution at specimen function. Resolution obtainable in real space is inversely proportional to the extent of the wavefield recorded (and, of course, the k-vector of the radiation). Let the object be $\Psi(x)$ in one

0304-3991/89/$03.50

dimension. Then in the Fraunhofer limit we have a complex diffraction pattern $D(k)$, where

$$D(k) = \int_{-\infty}^{\infty} \Psi(x) \exp(\mathrm{i}kx)\, \mathrm{d}x, \tag{1}$$

where it is assumed that k is scaled to x by the wavelength. Given that ultimately we are going to represent $\Psi(x)$ by a discrete set of points (for example, by a display of pixels on a computer screen), it is reasonable to represent $\Psi(x)$ by a set of equally spaced sample values Ψ_j, where $j = 1, 2, \ldots, N$. Hence structural determination reduces to having to find a solution vector Ψ in an N-dimensional space from a diffraction vector D. If we know the complex value of D, this reduces to a simple linear inverse Fourier transform. However, given that we only ever measure $|D(k)|^2$, can we solve for $\Psi(x)$?

Consider, for the sake of simplicity, a bandwidth-limited periodic function which fills half the unit cell. The diffraction pattern is also bandwidth-limited and periodic. Taking the Fourier transform of $|D(k)|^2$ will yield the autocorrelation function (that is, the Patterson function) such that

$$c(j\Delta x) = c_j = \sum_{n=1}^{N} \Psi(x_n)\Psi^*(x_n + j\Delta x), \tag{2}$$

where Δx is the separation of the sample points in $\Psi(x)$, and the asterisk denotes the complex conjugate. If, for example, $N = 6$ and the non-zero elements of $\Psi(x)$ are Ψ_1, Ψ_2 and Ψ_3 (where $\Psi_n = \Psi(x_n)$), eqs. (2) will look like

$$\begin{aligned} c_{-2} &= \Psi_1^* \Psi_3, \\ c_{-1} &= \Psi_1^* \Psi_2 + \Psi_2^* \Psi_3, \\ c_0 &= \Psi_1\Psi_1^* + \Psi_2\Psi_2^* + \Psi_3\Psi_3^*, \\ c_1 &= \Psi_1\Psi_2^* + \Psi_2\Psi_3^*, \\ c_2 &= \Psi_1\Psi_3^*. \end{aligned} \tag{3}$$

There are three independent equations ($c_n = c^*_{-n}$) but they are not in general uniquely soluble. Even if $\Psi(x)$ is known to be real and positive there can still occur unexpected multiple roots apart from the obvious ones such as Ψ', where $\Psi_1' = \Psi_3$, $\Psi_2' = \Psi_2$ and $\Psi_3' = \Psi_1$. Naive substitution from the extremal equations inwards makes the entire solution highly noise-sensitive. Solution by a Newton–Raphson iteration is well-behaved, but is still plagued by multiple roots. In this example we can write

$$\begin{aligned} &\Psi_1^8 + (2c_2 - c_0)\Psi_1^6 + (2c_2^2 + c_1^2 - 2c_0c_2)\Psi_1^4 \\ &\quad + (2c_2^3 - c_0c_2^2)\Psi_1^2 + c_2^4 = 0, \end{aligned} \tag{5}$$

where

$$\Psi_2 = \frac{c_1}{\Psi_1 + \dfrac{c_2}{\Psi_1}}, \quad \Psi_3 = \frac{c_2}{\Psi_1},$$

and where for simplicity it has been assumed that Ψ is real. There are up to eight solutions, though restricting Ψ to be real (which it will not be in the case of electron microscopy) may in special circumstances result in a unique solution (for example, if Ψ is centrosymmetric and positive).

This example illustrates two important points. (i) It is invariably possible to find solutions to the one-dimensional phase problem which fit the recorded data: the problem is that a large number of compatible solutions exist. (ii) The correlation function is twice as wide as the object function, so in order to solve for $\Psi(x)$ without overlap or wrap-around, the unit cell must be at least twice as wide as the object function. This is equivalent to noting that components of intensity have twice the frequency of their underlying complex components, so the diffraction plane must be sampled on a grid of half the spacing of that needed for the direct linear transform.

An alternative view of the question of the multiplicity of solutions can be pictured by considering the continuation of either the object or diffraction functions into the complex plane. For a detailed review of this construction in the one-dimensional case see Saxton [5]. Complete knowledge of $D(k)$ implies complete knowledge of $\Psi(x)$, and vice versa, so the choice of plane is not essential. In taking the Fourier transform of $\Psi(x)$, we could allow k to take all complex values z. The transform then becomes

$$D(z) = \int_a^b \Psi(x) \exp(\mathrm{i}zx)\, \mathrm{d}x, \tag{4}$$

where $z = \alpha + \mathrm{i}\beta$. By stipulating certain conditions on $\Psi(x)$ and the interval (a,b), which in

general will be fulfilled by any physical specimen and which are discussed in detail elsewhere [6], it can be shown that $D(z)$ is an entire function. That is, it can be expanded as a Taylor series at any point in the complex plane, and hence will be defined by its complex zeros. This is similar to expanding a real function as a polynomial, and noting that its roots give a full description of the function. In general, if $D(k)$ (or $\Psi(x)$) require N components (or sample points) to be described fully, then there will be N relevant complex zeros when they are continued into the complex plane. For finite objects (which have infinitely wide transforms), there are an infinite number of zeros, but only N are needed to obtain a good approximation to the original function. Moving zeros around preserves the number of Fourier components in the resulting function, but only a limited set of such movements will give the recorded intensity. In fact, for each zero in $D(z)$ there is one other zero which will maintain the form of $|D(k)|^2$. Therefore, for an N-component function, the are 2^N possible combinations of complex zeros that are compatible with the recorded intensity. This is the origin of the ambiguity of solution in the phase problem.

2.2. *Mathematical conditions for solubility*

Burge et al. [6] have derived a mathematical framework with which to understand the one-dimensional phase problem in terms of logarithmic Hilbert transforms. This is particularly useful for understanding why a hologram yields a unique solution to the phase problem, especially when it occurs only as another form of far-field intensity distribution, as is the case of the central disc of a microdiffraction pattern. If it is known that the object function is finite, certain restrictions are placed on $D(k)$, which can allow direct solution for the phase of the scattered wavefield. These ideas are closely related to causal transforms which arise in many fields, for example, the Kramers–Kronig relations between the real and imaginary parts of the dielectric function [7] and coherence theory [8]. In relation to the phase problem, it is possible to write

$$D(z) = |D(z)| \exp[i\phi(z)], \tag{6}$$

and taking the logarithm

$$\ln D(z) = \ln |D(Z)| + i[\phi(z) \pm 2\pi n]. \tag{7}$$

The real part of this function only depends on the modulus of $D(z)$, which is what can be measured along the real axis, and so applying the theory of residues, it is possible to derive a Hilbert transform to determine $\phi(z)$ directly. Here, n denotes the Riemann surface in which $\ln D(z)$ is defined. Zeros in $D(z)$ create branch points in $\ln D(z)$, and there are difficulties if $\lim_{k\to\infty} D(k) = 0$, which will usually be true in practice. The Hilbert transform requires a semicircular path integral of infinite radius in the upper half of the complex plane, which gives the "Hilbert phase". If there are no zeros in $D(k)$ enclosed by this path (i.e. $\ln D(k)$ is analytic throughout the half plane), then the Hilbert phase equals the true phase, provided a means is found to stop $\ln D(z)$ tending to $-\infty$ where $D(z)$ tends to zero, for example by adding a constant to $D(z)$ and accounting for this at a later stage. In general, there will be zeros in $D(z)$, creating branch points in $\ln D(z)$, in which case they have to be found so that their residues can be added as a sum of terms to the Hilbert phase.

These schemes represent two main ideas. (i) If it can be arranged that $D(z)$ has no zeros in one half of the complex plane, then the phase of $D(k)$ can be found directly from its modulus alone. (ii) If there are zeros in both halves of the plane, they have to be located in at least one half, which brings us back to the ambiguities described in the previous section. In physical terms the easiest way to arrange for (i) to be true is to add a large known reference function so that there are no zeros in the upper half plane: this is the principle of holography. In the absence of a reference beam, more than one independent intensity experiment must be performed, so as to resolve the 2^N ambiguity described earlier. The microdiffraction plane is a combination of both these types of information.

2.3. More than one plane or more than one dimension

Conventional transmission electron microscopy (CTEM) is able to access two planes of intensity information, $|\Psi(x)|^2$ and $|D(k)|^2$, but only up to a resolution defined by the objective aperture which truncates $D(k)$. The image plane can nevertheless only be recorded in intensity, and so calculating its phase does qualify as a type of phase problem. It may be postulated that for any general function $\Psi(x)$ which has large variations in phase, the number of complex functions compatible with $|\Psi(x)|^2$ is limited if due account is taken of $|D(k)|^2$. In other words, $|\Psi(x)|^2$ recorded over N pixels has 2^N possible solutions for $\Psi(x)$, but very few of these will give $|D(k)|^2$ where $D(k)$ is the Fourier transform of $\Psi(x)$. This is an example of a one-dimensional phase problem without a holographic reference beam where solution is tractable by virtue of being able to access two independent intensity experiments. Rather than locating all complex zeros and eliminating those which are not compatible with the two sets of measurement, successful solutions to this problem have proceeded by iterative convergence. For example, in the Gerchberg–Saxton algorithm [9–11] a random phase distribution $\phi_r(x)$ is assigned to one plane of data, say $\Psi(x)$, to form $\Psi(x)\exp[i\phi_r(x)]$, where $\Psi(x)$ is real and is the square root of the measured intensity. This is then transformed to the diffraction plane, where it is unlikely to resemble the measured $|D(k)|^2$. The modulus of the resulting transform is discarded, but its phase is used as a first guess for the phase of $D(k)$, which is then transformed back to the image plane. The process is repeated, at each step the modulus being discarded, but the calculated phase is reassigned to the measured modulus in either plane. It turns out there can still occur certain ambiguities in the solution calculated by this technique [12], and in practice there are problems with convergence if either the phase changes in the object function are weak or the data are corrupted by noise [13]. However, this method demonstrates that there can exist practical computational algorithms for converging on solutions when two intensity experiments are available. Two images of the same object taken at different settings of defocus may also suffice [14–16], and in principle more and more sets of data can be collected under different illumination conditions until the final solution is guaranteed to be unique.

In the case of a two-dimensional finite object function, the phase problem becomes much more tractable [17]. Even when only the diffraction plane of intensity information is accessible (which will be referred to as the "strict" phase problem), increasing the dimension of the object function greatly reduces the possible number of ambiguities. Consider a two-dimensional array of $n \times n$ pixels. It is possible to continue either a row or a column of pixels into the complex plane. Each of these functions will have 2^n complex roots. However, the pixel lying on the intersection of the row and column only has a single phase associated with it. There are two sets of possible phases which the pixel may have according to each of the possible root combinations taken along either the row or the column. The intersection of these two sets will normally have far fewer elements than the 2^n ambiguity. This method of elimination can be repeated for the next row or column, or even at diagonals across the plane through the same pixel. Similarly, in terms of correlation equations like those in eq. (3), in 2D there are roughly $2N$ meaningful equations for N unknowns, in comparison to the 1D case of N equations for N unknowns. It is interesting that even as long ago as 1939 Wrinch [18] discussed the existence of degenerate solutions in the case of discrete points scatterers in terms of vector maps, and concluded that the 2D solution is very rarely non-unique.

In practice, of course, there will be no formal solution to all the possible correlation equations that can be constructed in the two-dimensional problem because data are bound to be corrupted by noise. However, Fienup (see, for example, ref. [19]) has developed convergent algorithms similar to the Gerchberg–Saxton scheme for the strict 2D phase problem which are relatively noise-robust, yet for which only knowledge of the support of the object function is required (as opposed to its intensity distribution). (The support of a function is the region over which it is non-zero.) This is quite remarkable when compared to the intracta-

bility of the one-dimensional problem, especially when the object is allowed to be complex [20].

3. Microscopy and microdiffraction

In this section it will be assumed that (i) the electron wave interacts multiplicatively with the specimen – i.e. if the probe profile is given by $P(x)$, then immediately beyond the specimen the wave distribution is $P(x)\Psi(x)$; (ii) the Ewald sphere is adequately represented as flat – that is, that the specimen appears as a projection at one level of defocus. This approximation is often used in high resolution imaging, but it must be borne in mind it is a poor approximation in microdiffraction, where very large angles of scatter are accessible. In reality, multiple scattering effects are strong in all but the very thinnest specimens. However, it is assumed that it should normally be possible to find a thin edge of a specimen, and in what follows solution of atomic structure will anyway only be practical in the very thinnest specimens.

In terms of electron microscopy, confining ourselves to the strict phase problem is equivalent to having to calculate the complex specimen function using only the diffracted intensity in the back focal plane of the objective lens. Surprisingly, the previous section suggests that, being a 2D problem, this is in principle tractable as long as the specimen is finite. However, conventional (selected-area) diffraction is performed over relatively large areas of specimen (of the order of microns) which gives diffraction patterns with extremely fine structure. In practice, this structure cannot be measured, partly because the detector is unlikely to have a fine enough element size, and partly because the illumination is unlikely to be sufficiently coherent (i.e. the diffraction pattern will be convolved with a large source size). Solution of phase relies upon extracting all information from the coherent diffracted intensity – that is, it is necessary to record the pattern on a grid corresponding to the Nyquist sampling frequency (in reciprocal space) of the most rapidly varying intensity component, which is proportional to the size of the specimen. It should be emphasised that it is not possible to solve for phase from a "cross-grating" diffraction pattern from an infinite crystalline object. The solution techniques described above rely on finite object function support, so that diffraction orders are blurred and they have the opportunity to interfere with one another. A very small crystallite will scatter significant intensity outside the reciprocal lattice points, and all this information must be collected in order to solve for phase. Clearly, this scheme is both impractical and wasteful of the properties of the objective lens, which is capable of re-interfering large sections of the diffraction pattern in a phase-conserving way.

Let us consider the information available in a microscope which is being run with an objective aperture chosen for optimum resolution at Scherzer defocus. It is usual to think only of the bright-field image intensity. However, there is also the possibility of tilting the illuminating beam to form many distinct dark-field images, thus exploring diffraction orders which exist well outside the objective aperture when the illumination is paraxial. By reciprocity [21], a microdiffraction pattern is a plot of the intensity at one image pixel as a function of all angles of illumination (see fig. 1). This represents a far greater body of information than the conventional image. In order to consider how to process it, it is easiest to study the microdiffraction plane as a phase problem. We

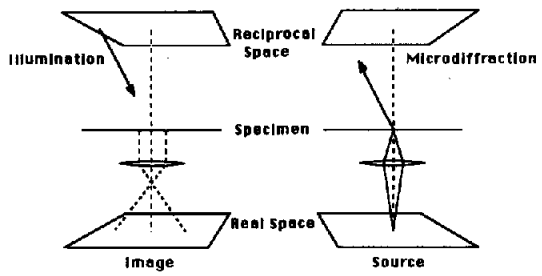

Fig. 1. Comparison of the information available in a CTEM (left) and STEM (right). In both machines there are 4 dimensions of intensity measurements available. In CTEM, a 2D image plane for every incident angle of illumination; in STEM, a 2D microdiffraction plane for every image pixel. The latter reduces to a far-field phase problem with the specimen modulated by the probe profile, which is an aberrated Airy disc function.

can regard the richer structural information available in a STEM as resulting from the probe-profile delineating such a small region of specimen that the far-field diffraction plane (the microdiffraction plane) possesses large enough intensity variations for a practical, multi-element detector to be able to extract all the available intensity diffraction information from the single image pixel being illuminated. If, for example, the probe is incident upon an amorphous material, intensity variations due to the finite size of the illuminated volume are clearly visible [22]. These are conditions in which solution of phase should become possible. A relatively small objective aperture can be used, yet very high angles of scatter are available. Furthermore, we have the ability to move the probe relative to the specimen, so performing essentially independent intensity experiments.

3.1. As linear and quadratic equations

Working, for simplicity, in one dimension, let the probe-profile in a STEM be $P(x)$, a complex quantity which can be represented by a discrete set of values P_j, where $j = 1, 2 \ldots N$. $P(x)$ is the Fourier transform of the objective aperture function, which will contain the usual phase changes due to spherical aberration and defocus. Taking the Fourier transform of the intensity of a single microdiffraction pattern, we can rewrite the quadratic correlation equations (3) as

$$\begin{aligned}
&P_1P_1^*\Psi_1\Psi_1^* + P_2P_2^*\Psi_2\Psi_2^* + \ldots P_NP_N^*\Psi_N\Psi_N^* \\
&\quad = H_0^0, \\
&P_1P_2^*\Psi_1\Psi_2^* + P_2P_3^*\Psi_2\Psi_3^* + \ldots P_{N-1}P_N^*\Psi_{N-1}\Psi_N^* \\
&\quad = H_1^0, \\
&\vdots \\
&P_1P_N^*\Psi_1\Psi_N^* = H_N^0,
\end{aligned} \tag{8}$$

where H_g^f is the gth Patterson component from the fth probe-position. In STEM we can shift the probe relative to the specimen to produce yet more simultaneous equations of the form (for $f = 5$)

$$\begin{aligned}
&P_1P_1^*\Psi_5\Psi_5^* + P_2P_2^*\Psi_6\Psi_6^* + \ldots P_NP_N^*\Psi_{N+5}\Psi_{N+5}^* \\
&\quad = H_0^5, \\
&P_1P_2^*\Psi_5\Psi_6^* + P_2P_3^*\Psi_6\Psi_7^* \\
&\quad + \ldots P_{N-1}P_N^*\Psi_{N+4}\Psi_{N+5}^* = H_1^5, \\
&\vdots \\
&P_1P_N^*\Psi_5\Psi_{N+1}^* = H_N^5.
\end{aligned} \tag{9}$$

At first it may appear as if infinitely many such equations are available, but they do not all contain independent information. To show this, we could construct a matrix P of the form

$$P = \begin{pmatrix}
P_1P_1^* & P_2P_2^* & P_3P_3^* & \ldots & P_NP_N^* \\
P_NP_N^* & P_1P_1^* & P_2P_2^* & \ldots & P_{N-1}P_{N-1}^* \\
P_{N-1}P_{N-1}^* & P_NP_N^* & P_1P_1^* & & P_{N-2}P_{N-2}^* \\
\vdots & & & & \\
\ldots & \ldots & \ldots & & P_1P_1^*
\end{pmatrix},$$

such that

$$P\Psi = H_0, \tag{10}$$

where

$$\Psi = \begin{pmatrix}
\Psi_1\Psi_1^* \\
\Psi_2\Psi_2^* \\
\Psi_3\Psi_3^* \\
\vdots
\end{pmatrix},$$

where H_0 is a vector of elements H_0^f, where $f = 1, 2, \ldots, M$. Here, we have simply taken the first equation of (8), and first equation of all such sets of equations such as (9). We could write similar sets of linear equations by using the jth equation from each quadratic set, so as to solve for a vector composed of terms like $\Psi_n\Psi_{n+j}^*$. As written here, P is a finite Teoplitz matrix, very similar to those encountered in signal processing theory [23]. Indeed, microdiffraction is closely akin to processing theory, in that it represents a space-resolved spatial-frequency spectrum, whereas in signal processing such as speech recognition, data are often represented as a time-resolved frequency

spectrum. For simplicity, P has been allowed to wrap around itself, implying that the probe is repeated periodically. Repetitions make no difference provided, once again, that the specimen only fills half the unit cell, or the probe function is effectively zero throughout half the unit cell. In practice, the probe is an Airy disc function which decays rapidly away from the central lobe. These outer regions have very serious consequences when the specimen is crystalline and of small unit cell (see next section). Choosing a higher sampling frequency in both Ψ and P increases the dimension of eq. (10) but nevertheless a solution vector using this set of equations is not available up to arbitrarily good resolution. This is because the $P(x)$ is bandwidth-limited, being the transform of a finite aperture. The convolution of $P(x)$ and Ψ (i.e. the scanned image) is also bandwidth-limited, and so increasing the sampling rate above the Nyquist frequency of the highest component in the image renders the matrix P singular. This is most easily seen by attempting to solve (10) by employing the deconvolution theorem. Because the transform of $P(x)$, the aperture function, is finite, a divide by zero will occur at frequency components higher than those present in the probe. This is equivalent to saying that Fourier components of intensity in the microdiffraction plane do not vary significantly when the probe is moved laterally by a distance less than the conventional incoherent image resolution of the microscope.

However, it the unit cell is larger than the image resolution, it is possible to generate at least more than one set of quadratic equations pertaining to the same specimen elements. This ability to perform multiple intensity experiments (and hence effectively be able to locate the complex zeros of the specimen function without requiring the central disc holographic reference beam where conventional bright-field image information resides) renders the microdiffraction phase problem much more tractable than the strict phase problem. Furthermore, the microdiffraction plane is inherently two-dimensional, which will further enhance its solubility. Also, for a very thin specimen, it would be reasonable to assume the phase-grating approximation, so that $|\Psi(x)| = 1$, which would greatly limit compatible solutions. Unfortunately, though, certain problems are created by the bandwidth-limited nature of the probe. These are easy to recognize in the case of a perfect crystal (see below), but may also arise in other specimen functions which are of infinite extent.

3.2. Solving for phase in reciprocal space

The extra information available by being able to move the probe in STEM can be conveniently thought of in terms of a technique first suggested by Hoppe [24–26], later referred to as "ptychography" [27], and discussed by Spence with respect to the microdiffraction plane [28]. Ptychography was originally viewed as a diffraction phenomenon employing a small aperture at the specimen plane which could be moved laterally by a small amount. The effect was to change the interference conditions in the diffraction pattern so that simultaneous equations could be derived to solve for the phase of each diffracted beam. Of course, shifting an aperture and observing the transmitted intensity is a primitive form of imaging capability. However, if the entire diffraction pattern is recorded and given phase, the object can be observed at a resolution corresponding to the highest angle of diffraction available. The aperture size may be large, provided the spacing of the detector elements in the far field is appropriately fine. This is the problem posed by microdiffraction. In a later review, Hoppe and Hegerl [29] eulogized the possibility of using a STEM as a sophisticated diffractometer, but they did not comment on the possible difficulties associated with having a bandwidth-limited probe function instead of a well defined aperture in the specimen plane.

Consider a specimen of moderately large unit cell illuminated by a STEM probe. In the far field, each reciprocal lattice point is convolved in amplitude with the circular objective aperture function as in fig. 2. If we assign an arbitrary phase of zero to the zero-order beam, the phase of the first-order diffracted beam can be determined to within two possible solutions by measuring the intensity in that disc and in the region of overlap where the beams interfere. Higher-order beams can be phased similarly with respect to lower-order beams. The two-fold ambiguity can be removed by shifting the

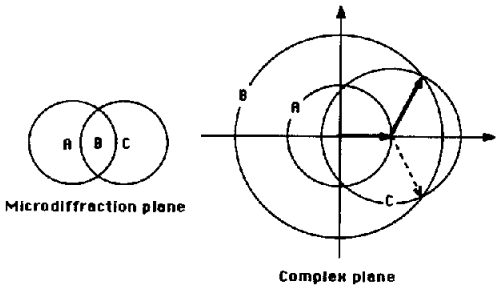

Fig. 2. Two overlapping discs in the microdiffraction plane allow for the possibility of measuring three intensities at A, B and C. If the phase of one of these diffraction orders is put to zero or is known, the other beam can be phased to two possible alternatives. In the complex plane, the square root of the intensities at A, B and C will give rise to circles of the radii shown. Ptychography (probe movement) resolves the ambiguity.

probe by a small amount in the specimen plane. Shifting is equivalent to adding a phase ramp across the aperture function which then adds a different phase change to both beams before they interfere, thus allowing the construction of simultaneous equations. In the two-dimensional case, ambiguity is further reduced (as one would expect) because there is an increased ratio in the number of overlaps to the number of discs.

Now let us examine the effects of specimen size and unit cell size in ptychography. Towards the limit of an amorphous specimen, that is an extremely large unit cell, each point in the diffraction plane will have contributions from many overlapping discs. However, there is a correspondingly large number of independent experiments that can be performed by shifting the probe to many distinct places within the unit cell. For an isolated, finite specimen, reciprocal space does not need to be sampled on an infinitely fine grid. The sampling theorem implies all the pertinent information is encapsulated in intensity measurements in the diffraction plane at points corresponding to a real-space unit cell twice the size of the object function. In practice, finite detector-element size in the microdiffraction plane will define a region of specimen over which it is possible to solve for specimen structure at any one time. By reciprocity, such a region is equivalent to the coherence width in the image plane of a CTEM due to the finite source size.

The opposite extreme of very small unit cell raises an interesting problem: what happens if the diffracted discs do not overlap? Solution of the phase becomes definitely impossible. This is because structural determination in the general formulation of the phase problem (and without a reference beam) relies either on finite support of the object function or multiple independent intensity experiments. Ptychography only works on an infinite crystal if there exist distinctly different positions to which the probe can be moved within the unit cell. However, if an infinite specimen of small unit cell were rendered finite, simply by the addition of a single very heavy atom lying within it, then the diffracted amplitude from that atom could, in principle, be used to phase the rest of the microdiffraction pattern as the probe is moved relative to the specimen.

Now let us consider the zero-order diffraction disc. For a thin or mostly transparent specimen, this beam is very strong relative to the specimen diffracted amplitude and so can be employed for in-line holography as originally suggested by Gabor [1,30]. The only information lost is the sign of the wavefield amplitude, which results in the existence of two reconstructed images. These can be well separated in side-band holography, but in Gabor holography they must be separated by severely defocussing the probe, so that the reconstructed images exist either side of the beam cross-over. The microscopical advantage of the technique is that the electron lens used to form the hologram may possess large aberrations, provided that this is accounted for in the reconstruction. Under these conditions, one reconstruction is true to the original specimen, while the other is doubly aberrated.

Holography has been investigated more recently by Lin and Cowley [31], especially with application to STEM. They point out that the ability to shift the probe leads to further possibilities in separating the two reconstructed images. We can understand this further reduction of ambiguity by considering the hologram as a zero-order ptychography disc. Solving for the two-fold ambiguity of ptychography relies on shifting the

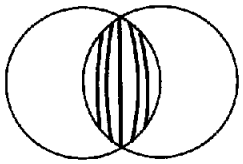

Fig. 3. Overlapping discs with severe defocus will cause fine fringes in the region of overlap. This information resolves the phase ambiguity in fig. 2, and is similar to the need to defocus in Gabor holography.

probe. Alternatively, greatly defocussing the probe creates much finer interference fringes within the regions of disc overlap (see fig. 3). This allows unambiguous determination of phase because many phase combinations can be sampled according to the relative position of the interfering beams with respect to the aperture function which will contain concentric variations in phase due to defocus. Of course, this requires a very fine detector element size because the region of specimen illuminated is much larger when the probe is defocussed. However, the two-fold ambiguity of ptychography can be viewed as analogous to the need to separate reconstructed images in holography: they both result from the ambiguity of addition in the complex plane. The Gabor holographic solution to this problem is to defocus, the ptychographic solution is to shift the probe. However, whereas holography can only be performed within the central disc which is strong relative to diffracted amplitude, ptychography can be used between beams of comparable strength and therefore can be used to phase very high angles of diffraction where super-resolution information resides.

4. Conclusions

There are three main ways of resolving the ambiguities which arise in the phase problem: (i) by the addition of a reference beam (holography); (ii) by increasing the dimension of the problem, given that the object function is finite; (iii) by performing more than one independent intensity experiment. The information in the central disc of the microdiffraction plane falls squarely into category (i). However, the "dark-field" scattered intensity falls rather awkwardly between categories (ii) and (iii). Although probe movement (i.e. the conventional imaging capability of a microscope) allows multiple intensity experiments to be performed, the object function is never finite because the probe function decays away from a central maximum. This is particularly debilitating in the case of an infinite crystal of small unit cell, where phasing the higher-order beams becomes definitely impossible. For objects of large or infinite unit cell, it should be possible to develop an algorithm which employs all a-priori knowledge of the specimen, the two-dimensional constraints of the strict phase problem and the ability to perform independent intensity experiments by moving the probe. The aim would be to solve for specimen structure at spatial resolution corresponding to the largest angle of diffracted intensity that can be recorded.

The advantage of microdiffraction is that it is relatively straightforward experimentally. Unlike high-angle holography, it does not require such extremely stable apparatus, even though the inverse calculation is much more complicated. However, there are very severe problems which should not be underestimated. Here we assume that the electron wave interacts linearly with a two-dimensional projection of the specimen. In practice, at high angles of scatter, there will be significant interference between waves scattered from the top and bottom surfaces of the specimen. This may ultimately allow for 3D solution of specimen structure, but would complicate the types of iterative solutions already developed for the 2D problem. More serious difficulties (which incidentally also apply to the holographic technique) will arise from multiple, inelastic and thermal-diffuse scattering, all of which will mask the simple Fourier-optic signal discussed above. Any practical, useful algorithm for the inverse calculation of the object function will have to account for all these effects, or else only operate within well defined limits (e.g. only for very thin specimens). Furthermore, the technique will still have to face all the usual experimental frustrations like specimen damage and contamination. We can conclude, therefore, that though the phase problem in microdiffraction is much more tractable than the strict

phase problem, and though there may be great gains in spatial resolution possible by processing the whole plane as a function of probe position, achieving this experimentally will not be a trivial exercise.

Acknowledgements

The author would like to thank Dr W.O. Saxton for helpful discussions, and is grateful for financial support from The Royal Society.

References

[1] D. Gabor, Nature 161 (1948) 777.
[2] K.J. Hanzsen, Advan. Electron. Electron Phys 59 (1982) 1.
[3] J. Konnert and P. D'Antonio, Ultramicroscopy 19 (1986) 267.
[4] J.C.H. Spence and J.M. Zuo, Rev. Sci. Instr. 59 (1988) 2102.
[5] W.O. Saxton, in: Computer Processing of Electron Microscope Images, Topics in Current Physics, Vol. 13, Ed. P.W. Hawkes (Springer, Berlin, 1980) ch.2.
[6] R.E. Burge, M.A. Fiddy, A.H. Greenaway and G. Ross, Proc. Roy. Soc. (London) A350 (1976) 191.
[7] J.S. Toll, Phys. Rev. 104 (1956) 1760.
[8] E. Wolf, Proc. Phys. Soc. (London) 80 (1962) 1269.
[9] R.W. Gerchberg and W.O. Saxton, Optik 35 (1972) 237.
[10] R.W. Gerchberg and W.O. Saxton, J. Phys. D6 (1973) L31.
[11] R.W. Gerchberg, Nature 240 (1972) 404.
[12] P. Schiske, Optik 40 (1974) 261.
[13] J.N. Chapman, Phil. Mag. 32 (1975) 527, 541.
[14] A.J.J. Drenth, A.M.J. Huiser and H.A. Ferwerda, Opt. Acta 22 (1976) 615.
[15] B.J. Hoenders and H.A. Ferwerda, Opt. Acta 23 (1976) 445.
[16] P. van Toorn and H.A. Ferwerda, Opt. Acta (1978) 457 and 469.
[17] R.H.T. Bates, Optik 61 (1982) 247.
[18] D.M. Wrinch, Phil. Mag. 27 (1939) 98.
[19] G.B. Feldkamp and J.R. Fienup, in: SPIE, Vol. 231, 1980 Intern. Optical Computing Conf., Ed. W.T. Rhodes (1980) p. 84.
[20] J.R. Fienup, J. Opt. Soc. Am. A4 (1987) 118.
[21] J.M. Cowley, Appl. Phys. Letters 15 (1969) 58.
[22] J.M. Rodenburg, Ultramicroscopy 25 (1988) 329.
[23] C. Gueguen, in: Signal Processing, Proc. Les Houches Summer School, Session XLV, Eds. J.L. Lacoume, T.S. Durrani and R. Stora (North-Holland, Amsterdam, 1987) p. 707.
[24] W. Hoppe, Acta Cryst. A25 (1969) 495.
[25] W. Hoppe, Acta Cryst. A25 (1969) 502.
[26] W. Hoppe, Acta Cryst. A25 (1969) 508.
[27] R. Hegerl and W. Hoppe, in: Proc. 5th European Congr. on Electron Microscopy, Manchester, 1972, Inst. Phys. Conf. Ser. 14, Ed. A.M. Glauert (Inst. Phys., London–Bristol, 1972) p. 628.
[28] J.H.C. Spence, Optik 49 (1977) 117.
[29] W. Hoppe and R. Hegerl, in: Computer Processing of Electron Microscope Images, Topics in Current Physics, Vol. 13, Ed. P.W. Hawkes (Springer, Berlin, 1980) ch. 4.
[30] D. Gabor, Proc. Roy. Soc. (London) A197 (1949) 454.
[31] J.A. Lin and J.M. Cowley, Ultramicroscopy 19 (1986) 179.