Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Theory of Super-Resolution Electron Microscopy Via Wigner-Distribution Deconvolution

J. M. Rodenburg and R. H. T. Bates

Phil. Trans. R. Soc. Lond. A 1992 **339**, 521-553
doi: 10.1098/rsta.1992.0050

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click **here**

To subscribe to *Phil. Trans. R. Soc. Lond. A* go to:
http://rsta.royalsocietypublishing.org/subscriptions

This journal is © 1992 The Royal Society

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
THE ROYAL SOCIETY A
PHILOSOPHICAL TRANSACTIONS OF

The theory of super-resolution electron microscopy via Wigner-distribution deconvolution

By J. M. Rodenburg[1] and R. H. T. Bates[2]†
[1] *Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, U.K.*
[2] *School of Electrical Engineering, University of Canterbury, Christchurch, New Zealand*

Contents

	PAGE
1. Introduction	522
2. The resolution problem	523
(*a*) Conventional diffraction	523
(*b*) Conventional imaging	524
(*c*) Holography and instability	526
(*d*) Dark-field imaging, reciprocity and microdiffraction	528
(*e*) Experimental considerations of STEM	529
3. Elementary theory and physical interpretation	530
(*a*) Ptychography versus holography	530
(*b*) Wigner deconvolution and phase-retrieval	533
4. Effect of thickness	539
5. Full theory, including effects of partial coherence	542
6. The effect of noise	547
7. Conclusions	550
References	551

The theory of deconvolving the microdiffraction data-set available in a scanning transmission electron microscope or, equivalently, the set of all bright- and dark-field images available in a conventional transmission electron microscope to obtain super-resolution micrographs (which are not limited by the transfer function of the objective lens) is developed and described with reference to holography and other phase-retrieval schemes. By the use of a Wigner distribution, influences of the instrument function can be entirely separated from the information pertaining to the specimen. The final solution yields an unambiguous estimate of the complex value of the specimen function at a resolution which in theory is only limited by the electron wavelength. The faithfulness of the image processing is shown to be not seriously affected by specimen thickness or partial coherence in the illuminating beam. The inversion procedure is remarkably noise insensitive, implying that it should result in a robust and practicable experimental technique, though one that will require very large computing facilities.

† Professor Richard Bates died of cancer before the final draft of this paper had been completed. With characteristic dedication and enthusiasm, he continued to work on this problem until the very end of his life, despite being in considerable pain.

Phil. Trans. R. Soc. Lond. A (1992) **339**, 521–553
© 1992 The Royal Society and the authors
Printed in Great Britain

The Royal Society is collaborating with JSTOR to digitize, preserve, and extend access to *Philosophical Transactions: Physical Sciences and Engineering.*
JSTOR®
www.jstor.org

Downloaded from rsta.royalsocietypublishing.org

1. Introduction

For many years there has been a large disparity between the spatial resolution obtainable in a transmission electron microscope (currently 1.5–2 Å†) and the wavelength of a high-energy electron (typically 0.05 Å or less). We do not discuss here scanning tunnelling microscopy (STM), which is a surface-sensitive technique (Binning *et al.* 1982), but consider the question of retrieving transmitted information from a far-field scattering experiment with the intent of obtaining three-dimensional information on the specimen structure. In these circumstances, the resolution problem has remained remarkably insuperable (for reasons which are reviewed briefly in §2), despite the fact that the present limit is of the same order of magnitude as the atomic bond lengths, so that even a minor improvement can greatly increase the classes of material that can be studied directly by the electron microscope. The most debilitating experimental problems are related to electrical, magnetic and mechanical instabilities. In any scattering experiment where only intensities can be measured, retrieval of the relative phase of the Fraunhofer diffraction plane, which is equivalent to obtaining a full description of the wavefield which has emanated from the specimen, relies upon re-interference of separated parts of the diffraction plane. In the case of electron scattering, this essential requirement is extremely difficult to achieve if the difference in path length between interfering components is large: any instability in the instrument or external interference from stray magnetic fields will alter the relative phase of the beams.

To overcome the resolution problem in a robust and practical way, we are therefore faced with the following problem: Is it possible to construct an electron scattering experiment which does not rely upon interference of beams which possess large differences in their path length, but which nevertheless allows retrieval of the phase of the entire scattered wavefield? We show here that at least in the case of kinematical scattering, there is a complete theoretical solution to this problem. We use a data-set which is equivalent to all possible images which could be collected in a conventional transmission electron microscope (CTEM) as a function of all dark-field (tilted) illumination conditions, though in practice this could be most conveniently recorded in the microdiffraction plane of a scanning transmission electron microscope (STEM). For any one of these images, only a small portion of the diffraction plane can pass through the electron lens, and therefore each image has poor resolution. However, by suitable processing using a Wigner-distribution deconvolution, we can obtain a very high resolution reconstruction of the specimen function. Furthermore, the processing method is shown to be extremely noise-robust, able to cope with partial coherence in the source (in the case of STEM) and is not limited to very thin specimens. The main practical constraint will be that the data-set is somewhat large, but given that computing power is becoming progressively cheaper, particularly in relation to the cost of improving resolution by other means (such as by using higher accelerating voltages), this is a minor difficulty. The important conclusion is that the resolution problem should be tractable within the framework of prevailing experimental apparatus.

In §2 we review the origin of the resolution problem and discuss why many earlier suggestions for overcoming the present limit have failed. In §3 we examine why the microdiffraction plane in STEM offers the possibility of obtaining much higher

† 1 Å = 10^{-10} m.

Downloaded from rsta.royalsocietypublishing.org

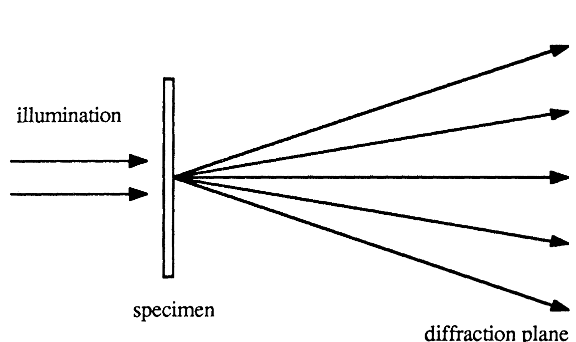


Figure 1. The configuration of a generalized scattering experiment. Parallel illumination is incident upon a thin transmitting specimen. The detector is positioned sufficiently far away for it to lie in the Fraunhofer diffraction plane.

resolution information and we develop an elementary theory for how this can be processed by using a Fourier transform algorithm based on a Wigner-distribution deconvolution. It is shown that the complex transfer function of the poor-quality magnetic lens can be filtered directly from the information pertaining to the specimen, and that obtaining super-resolution reduces to a phase-retrieval problem. The influence of specimen thickness is discussed in §4. In §5 the theory is developed further to account for both thickness effects and incoherence in the source. A one-dimensional calculation is presented in §6 which strongly suggests that the algorithm is not significantly noise sensitive. Conclusions are presented in §7.

2. The resolution problem

Figure 1 shows an idealized electron scattering experiment. Parallel illumination is incident upon a specimen and the detector is sufficiently far away for it to lie in the Fraunhofer diffraction plane. We assume that the waves interact elastically and multiplicatively with a specimen which can be represented as a complex-valued function. Most thin specimens are effectively non-absorbing and so they only introduce a phase change to the incident radiation which is proportional to the projected atomic potential (Cowley 1981*a*). In what follows, we may solve for a strong phase object, which corresponds to a full, dynamical solution to the scattering interaction provided the specimen in infinitesimally thin (Fejes 1977). For scattering from a thick specimen, we assume the first Born approximation (Mott & Massey 1933); that is to say, each part of the specimen is subject to an incident beam amplitude which is unaffected by scattering from other parts of the specimen.

(*a*) *Conventional diffraction*

In figure 1, the detector lies in the Fraunhofer diffraction plane, and so by definition, the Fourier transformation of the complex value of the wave disturbance measured over the whole plane yields a reconstruction of the waves emanating out of the specimen. Unfortunately, being a matter wave, the phase of the electron wavefunction cannot be measured directly. However, being restricted to being able to record only the intensity, which corresponds to the case of conventional electron or X-ray diffraction, is for many classes of specimen not such a severe drawback. The 'phase problem' in diffraction theory (Woolfson 1961) is usually tractable because

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

the experiment is provided with large quantities of *a priori* information; the specimen is crystalline and atomic coordination is subject to chemical constraints. By simply taking the Fourier transform of the diffraction intensity, the resulting Patterson function (Patterson 1934) gives a plot of the autocorrelation function of the specimen, and this clearly shows what information has been lost; diffraction experiments measure the distribution of relative separations within the specimen but are insensitive to absolute position. However, the great advantage of diffraction, which we make use of below, is that it is experimentally robust. Unlike techniques such as imaging and holography, all the pertinent wave interference occurs at, or in the very close vicinity of, the specimen. Once the wave train has set off towards the detector, we do not require it to interfere at some other point in the apparatus and so the source may be relatively incoherent implying that the technique is not very sensitive to physical vibration and other sources of mechanical and electrical interference.

It should be pointed out that there exists a theoretical solution to the Fraunhofer diffraction phase problem, even when the specimen is not crystalline. We may assume that any actual specimen is finite, in which case the set of allowed intensity distributions in the diffraction plane can be shown to be limited (Burge *et al.* 1976). Given a particular intensity distribution, the specimen function is nearly always uniquely defined if it is two dimensional, or of higher dimensionality (Bates & McDonnell 1986). There has recently been rapid progress in the development of iterative algorithms for converging upon a complex object function of known size even if only the intensity of its Fourier transform can be measured (see, for example, Fienup 1987). However, in the case of electrons, these techniques would be difficult to implement. To reconstruct the specimen, it is of course necessary to make a sufficient number of measurements to define its (complex) value at a particular sampling frequency (resolution) in real space over its total spatial extent. The object size defines a corresponding minimum sampling frequency in the diffraction plane, which incidentally must be twice as high as that required for the conventional complex Fourier transform when only intensity is measured (Bates & McDonnell 1986). For example, if it were possible to isolate a specimen of 1000×1000 Å^2 by using an opaque aperture, then for 100 keV electrons, the detector would have to have a resolution of better than 20 μrad. Furthermore, the pattern would have to be recorded with very good counting statistics because iterative algorithms for phase retrieval are not particularly noise insensitive, at least compared with the method developed below, and so this would not seem to be the most practicable solution. It is interesting to note, however, that it is theoretically possible to construct an electron microscope by using a very good quality detector and a large computer; the electrons lens, so often blamed for the resolution problem, is not strictly necessary.

(*b*) *Conventional imaging*

Given that microscopy is equivalent to assigning a phase to the Fraunhofer diffraction plane up to a particular scattering angle (proportional to the final resolution) and at a particular angular resolution (inversely proportional to the total size of the specimen), then the role of a lens can be regarded as simply a convenient method for re-interfering certain sections of the diffraction pattern. This is essentially the Abbé theory of light microscopy.

Figure 2 shows the experimental arrangement for bright-field imaging, which is the most commonly used technique in the electron microscope. Plane wave illumination

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

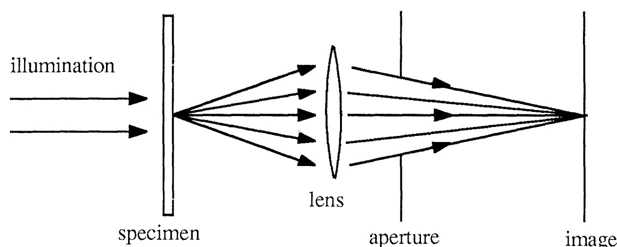


Figure 2. The optical configuration for conventional bright-field imaging. The incident illumination is parallel with the optic axis of the microscope and is transmitted to the image plane, where it acts as a reference wave for those beams scattered by the specimen and refocussed by the lens.

parallel to the optic axis is incident upon the specimen. The lens collects the Fraunhofer diffraction pattern up to a particular cut-off angle which is defined by the objective aperture positioned in the back-focal plane. Though being able to record an image directly is a great improvement in comparison to only having access to the diffraction plane intensity, the technique does not completely overcome the phase problem insofar as the image can still only be recorded in intensity.

To produce a faithful image, we wish the lens to perform a Fourier transform on the complex wavefield lying within the objective aperture. However, because thin, high-resolution specimens are essentially non-absorbing and only introduce a phase change in the incident beam, if the lens is absolutely perfect, the image intensity would be uniformly bright and would therefore, ironically, contain no information at all. Scherzer (1949) has shown that at certain value of defocus, spherical aberration inherent in a magnetic lens (Scherzer 1936) could be usefully employed as a phase plate, as in optical phase microscopy (Bennett *et al.* 1951). Ideally, the (largely unattenuated) incident beam is allowed through to the image to act as a reference wave while all scattered beams, which fall off the optic axis in the back-focal plane, are subject to a phase change. If we assume that the specimen is a weak phase object, it can be shown (see, for example, Spence 1981) that the final image has a contrast which is proportional to the phase change introduced by the specimen (which, as a first approximation, is in turn proportional to the projected atomic potential). But under these circumstances there is also a very clearly defined resolution limit which appears to be insurmountable. The phase change ϕ produced in the back-focal plane as a function of θ, the radial angle from the optic axis, is given to a first approximation by

$$\phi = (\Delta F \pi/\lambda)\,\theta^2 + (C_s\,\pi/2\lambda)\,\theta^4, \tag{1}$$

where ΔF is the defocus of the lens, C_s is the spherical aberration constant of the lens, and λ is the wavelength. For a given spherical aberration constant (which is usually of the order of the focal length) the angular range in which ϕ can be maintained as constant is very limited (of the order of 10 mrad), though in the Scherzer condition it is balanced in part by the defocus term. In the phase contrast image, the strength of the expression of a particular spatial frequency is given by $\sin\phi$, which is referred to as the contrast transfer function (CTF). Clearly, at large values of θ (where the high resolution information resides), the term involving C_s in equation (1) dominates and results in zeros and changes of sign in the CTF which seriously corrupt the fidelity of the final image. It is therefore usual to choose an objective aperture which cuts off the CTF at an angle corresponding to its first zero to avoid uninterpretable (and invariably misleading) artefacts in the image.

Downloaded from rsta.royalsocietypublishing.org

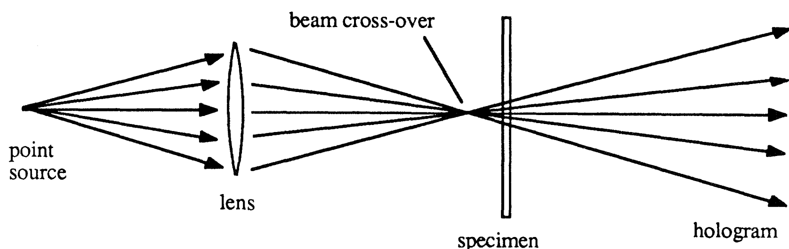


Figure 3. The optical configuration for transmission holography, as first proposed by Gabor. Interference now occurs not in the image plane (as in figure 2) but over a region equivalent to the conventional diffraction plane. The advantage of the technique is that aberrations introduced by the lens can be accounted for during the reconstruction process.

The whole rationale of conventional high-resolution phase-contrast imaging is to obtain as broad a CTF as possible. Unfortunately, any significant improvement in the resolution (i.e. usable values of θ) requires dramatic reduction of the spherical aberration constant, because of the dependence of the power of four in θ (equation (1)). Over the past 20 years or so, great efforts in lens design have managed to reduce C_s to fractions of a millimetre, but it now seems improbable that further gains in resolution by this method will be forthcoming. Recent suggestions for compensation of spherical aberration by higher-order octupoles (Shao *et al.* 1988; Shao & Crewe 1988) may prove hard to implement experimentally. Alternatively, one can reduce the electron wavelength by raising the accelerating voltage but this scheme faces a similar theoretical cut-off, though one which is possibly less severe than the spherical aberration problem. A conventional electron microscope works at an accelerating voltage of between 100–300 keV, at which the electron, having a rest mass of 0.51 MeV, is already relativistic. Gains in resolution are therefore less than proportional to the increase in voltage, and become marginal beyond a few MeV. More seriously, however, above about 100–200 keV many specimens suffer severe knock-on damage: the incident electrons have enough energy to displace atoms right out of the specimen (Makin 1968). Real gains in resolution, say by a factor of ten, can only occur if the total angular range of the diffraction plane contributing to the experiment is increased. Because of the spherical aberration problem, the conventional bright-field image would not appear to hold much promise.

(c) *Holography and instability*

In his original invention of holography, Gabor (1948, 1949) proposed that the imitations of phase-contrast electron microscopy could be overcome by using the arrangement shown in figure 3. Here, the lens is not being used to form an image, but is simply a method for providing a reference wave over a wide region of the Fraunhofer diffraction plane. Any type of beam-splitter device may be used. Provided the reference beam is strong relative to the scattered amplitude then the hologram records all amplitude and phase information in the scattered wavefield (which is as much as we could ever hope to obtain), at least to within an ambiguity of two possible reconstructions. Unlike the bright-field image, the technique is no longer subject to the contrast transfer characteristics of the lens, because the interference occurs not in the image but over a region equivalent to the back focal plane. Though the reference beam may have a complicated phase structure (prescribed by equation (1) with a large value of defocus because the specimen does

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

not lie in the focal plane), it may have a large amplitude at all angles. Gabor envisaged that it would be possible to magnify the electron hologram so that it could be later reconstructed by using light optics in such a way that any imperfections in the electron lens could be mimicked accurately. Finally, a wide-angle optical lens would focus the reconstructed wavefield into a high-resolution image. The in-line configuration (as shown in figure 3) suffers from certain difficulties (unlike, for example, side-band holography (Leith & Upatnieks 1962)) in separating the two reconstructions, for reasons which have been well documented (Hanzsen 1982), but it shows that a good quality lens is not a prerequisite for high-resolution imaging if it is used simply as a primitive beam-splitter and one is prepared to do a significant amount of post-experimental processing.

It is somewhat ironic that holography has had so many successful applications, except in the specific field for which it was invented. To record a hologram, the reference beam must be stable with respect to the scattered beam to within a fraction of a wavelength, and whereas this requires care in the case of light, it has proved to be extremely difficult for electrons which have a wavelength five orders of magnitude shorter, and which are also vulnerable to magnetic interference. First successful attempts were hampered by low beam intensities (Haine & Mulvey 1952) which occur once an incoherent source has been demagnified sufficiently to obtain the requisite spatial coherence. During the long exposure times which are then necessary, the specimen inevitably drifts. Recent attempts (Lichte 1986; Lin & Cowley 1986; Tonomura 1987) have been more successful, allowing for the amplitude and phase of the specimen function to be measured directly. Volki & Lichte (1990) have obtained astonishingly fine interference fringes corresponding to a real-space distance of 0.3 Å but have yet to demonstrate that the instrument function can be deconvolved from the image. It may well be that holography will routinely record higher-resolution micrographs than the bright-field technique, but it is unlikely to become experimentally robust and will ultimately remain limited by instability.

The general problem of instrument instability is also encountered in high-resolution bright-field imaging, where it is common experimental practice not to use a physical aperture in the back-focal plane of the objective lens to cut-off unwanted parts of the CTF, but instead rely on instability in the whole apparatus to dampen the contribution of high spatial frequency components in the image.

The instability term is often modelled as a gaussian envelope (Frank 1973), which decays as a function of angle θ. It results from several contributions including chromatic spread in the source (i.e. instability in the accelerating voltage supply), ripple in the objective lens current supply which is equivalent to variation in the defocus term of equation (1), and interference from oscillating magnetic fields which are hard to avoid in the laboratory environment. All of these effects reduce the ability for higher-angle beams to interfere at a fixed phase difference during the integration time of the recording process. Suffice it to say that irrespective of the adverse properties of electron optics, any type of improvement in image resolution must face up to the fundamental problem of instability. Development of the theory of phase-retrieval of electron imaging during the 1970s (for a review, see Saxton 1980) established that apart from holography, there are numerous theoretical methods of processing various types of data to retrieve the phase of the image plane, for example, by iterative convergence algorithms using the intensity distributions in both diffraction and image plane (Gerchberg & Saxton 1972; Gerchberg 1972; Chapman 1975*a*, *b*), by processing several images as a function of defocus (Misell

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

1973), or by using judicious combinations of varying objective aperture shapes (Misell & Greenaway 1974). Once this has been achieved, then it should be possible to deconvolve the result with the complex transfer function of the lens to obtain arbitrarily good resolution; the limitations of the zeros and changes of sign in the contrast transfer function arise from only being able to record image intensity. However, all of these schemes rely on re-interfering beams which have traversed significantly different paths well separated in diffraction space and which are therefore subject to the same experimental difficulties as holography.

(d) *Dark-field imaging, reciprocity and microdiffraction*

From the preceding sections, it is evident that there is certainly no lack of theoretical possibilities for extracting very high resolution information from the electron microscope. However, despite its very considerable failings, the bright-field image has still yet to be routinely improved upon. It has two persuasive advantages: the image can be seen immediately while the experiment is being performed and the instrumental specifications are not nearly as demanding as any of the alternative arrangements discussed so far. Any method which attempts to improve upon the bright-field technique is bound to have to perform some sort of computation akin to a deconvolution or holographic reconstruction, and so the instantaneous nature of the image is likely to be lost. But having conceded that a large computational effort will be required, it would seem logical to concentrate on collecting and processing a data-set which is as insensitive as possible to the known experimental difficulties.

Consider figure 4*a*, which corresponds to a tilted (dark-field) illumination condition in the conventional electron microscope. It is clear that if we no longer restrict ourselves to illumination parallel to the optic axis, there is a large set of images which may be collected from a single specimen. Of course, any one dark-field micrograph can only have the same resolution as the bright-field image, but it is worth asking whether the whole data-set thus acquired may be used to solve for the specimen at super-resolution. The advantage of the optical geometry is that the requirements for the instrument stability are the same as for the bright-field image. Even when the illumination angle is large, the lens only has to re-interfere a similar solid angle of wavefield. The interference processes which allow for very high angle information to contribute to the recorded data are occurring at the specimen itself, as in conventional diffraction. This is the data-set we use in the Wigner-distribution deconvolution method.

Let us consider the scanning transmission electron microscope (STEM). For thin (say of the order of 10 nm thick) elastically scattering specimens, the data-set shown in figure 4*a* can most effectively be measured by placing a two-dimensional detector in the microdiffraction plane of a STEM, as shown in figure 4*b*. By the principle of reciprocity, a STEM can be regarded as a CTEM run in reverse (Cowley 1969), with the source and image planes exchanged. Any particular point in the microdiffraction plane of a STEM corresponds to an angular position of the source in CTEM, while the source in STEM lies at an equivalent position to one point in the image plane of a CTEM. A conventional bright-field image can therefore be obtained in a STEM by scanning the source sequentially across the image plane (which is achieved in practice by deflecting the position of the beam cross-over at the specimen using magnetic deflection coils) while detecting the intensity at the central point in the micro-diffraction plane (defined by the 'collector' aperture) and displaying the resulting signal synchronously on a cathode ray tube.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

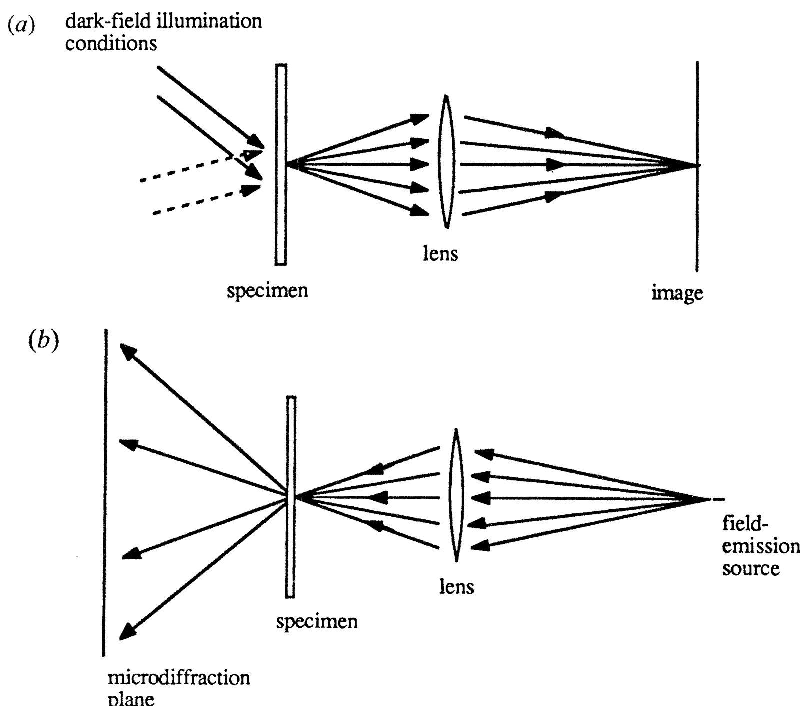


Figure 4. (*a*) Examples of different illumination conditions for dark-field imaging. There is a whole plane of image data available for each possible incident angle. For any one of these images, the lens only has to interfere a limited region of the Fraunhofer diffraction plane. (*b*) The optical configuration for microdiffraction in a scanning transmission electron microscope. This is identical to (*a*), except the direction of all rays have been reversed. By the principle of reciprocity, a microdiffraction pattern can therefore be regarded as a plot of the intensity of one image pixel in the conventional image plane as a function of all angles of illumination in the conventional microscope.

Reciprocity is useful for understanding why features such as Fresnel fringes, thickness fringes and phase contrast effects are virtually identical in the two types of machine despite the fact that they appear to operate by entirely different mechanisms. In the rest of this paper, we will occasionally make use of reciprocity to explain the origin and relationship of various types of information. It should be understood that when we refer to an image we mean either the image recorded in CTEM by using one angle of illumination or, equivalently, the signal collected at one scattering angle in the microdiffraction plane of a STEM as a function of all probe positions. Similarly, a microdiffraction pattern in STEM can be regarded as a plot of the intensity which would be measured at one point in the image plane of a CTEM as a function of all possible angles of illumination.

(e) *Experimental considerations of* STEM

STEM is notoriously poor at producing good quality, real-space images because in order to satisfy the necessary counting statistics for a given spatial resolution, the source must be extremely bright (see, for example, Brown 1981). Having taken into account the demagnification in the objective lens (and earlier lenses if fitted) the source must be smaller than the size of the resolution element of the microscope (as defined by the objective lens aberrations coupled with the chosen objective aperture)

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

otherwise the final image is further compromised by being convolved with the source shape. Crewe *et al.* (1968) overcame the brightness problem by using a cold field-emission gun (FEG), but at the price of requiring the entire microscope to be held at ultra-high vacuum.

This experimental complication, which necessitates indirect recording methods (and not, for example, the use of photographic film which would tend to outgas into the chamber), coupled with the fact that the main success of STEM has been in high-spatial resolution micro-analysis (i.e. recording the X-ray, electron-energy loss and other secondary signals that come from the small illuminated volume of specimen (Brown 1981)) has resulted in the microdiffraction plane being surprisingly under-exploited. It is widely regarded as merely another analytical tool, suitable only for making approximate measurements of the lattice spacings in small crystallites.

More recent instrumental developments (see, for example, Cowley 1980; Rodenburg & McMullan 1985) have indicated that the fine-scale structure of microdiffraction patterns obtained from both crystalline and disordered specimens (Cowley 1979, 1981*b*; Rodenburg 1988) is complicated and heavily dominated by the microscope's transfer function, leading, for example, to a consistent breakdown in Friedel's law (Rodenburg 1988). It has been demonstrated recently (McMullan *et al.* 1990) that it is possible to record energy-filtered microdiffraction patterns through the electron spectrometer at high angular resolution, high efficiency, and good dynamic range via a YAG scintillator coupled optically to a CCD array. This arrangement has the advantage of allowing inelastically scattered electrons, which may otherwise compromise the diffracted elastic signal, to be filtered out from the measured data-set. Given that this is now available, it would appear appropriate to develop the theory for how to process it efficiently, preferably in real time. The main difficulty arises from the enormous size of the data-set. For example, if the probe is moved over a 256×256 grid of points on the specimen, and for each probe position the microdiffraction pattern is recorded over a 256×256 array, the data-set will be about 4 gigabytes in size.

It has already been shown experimentally that super-resolution is possible by tracking the intensity of Patterson components in the microdiffraction plane (Konnert & D'Antonio 1986; Konnert *et al.* 1989), but this method relies on having considerable *a priori* information about the structure being examined. The present work demonstrates that a full set of microdiffraction patterns can be manipulated by using Fourier transforms (which are inherently fast and may be performed by using parallel computing techniques) to obtain a unique estimate of the complex value of the specimen function at higher resolution than the conventional Rayleigh limit prescribed by the lens/aperture configuration without any *a priori* information. It should therefore be useful for imaging all classes of specimen at increased resolution, including amorphous materials.

3. Elementary theory and physical interpretation

(*a*) *Ptychography versus holography*

Earlier work by Hoppe (1969*a*, *b*), Hoppe & Strube (1969), Spence (1977) and Spence & Cowley (1978) has shown that the microdiffraction plane can in principle be used to solve for the phase of high-angle diffracted beams which would normally lie outside the usable region of the objective aperture in conventional microscopy, at least in the case of the specimen being a thin crystal of known unit cell size (for

Downloaded from rsta.royalsocietypublishing.org

a review see Rodenburg 1989). The physical basis of the method relies on being able to move the probe across the specimen (or change the illumination conditions) in such a way that a multiple set of intensity measurements can be made in the microdiffraction plane, allowing for full solution of the phase of every beam. This type of phase-retrieval can be usefully referred to as 'ptychography' (following Hegerl & Hoppe (1972) and Hoppe & Hegerl (1980)) to differentiate it from reference-beam holography.

Ptychography can be most easily understood by considering the ray diagram for STEM shown in figure 5*a*. In the absence of a specimen, the microdiffraction plane consists of a bright disc which corresponds to a shadow image of the objective aperture. Now consider the introduction of a thin crystalline specimen. Because of the range of incident angles in the probe, any given diffraction spot will result in a disc of scattered amplitude. With careful choice of objective aperture size with respect to the unit cell size, it is possible to arrange for a first-order diffracted disc to just overlap with the zero-order disc, as shown in the diagram. Given that the electron source is spatially coherent (i.e. of very small physical extent), the two discs can interfere with one another coherently. The resulting intensity in the region of overlap will depend upon the phase of the underlying reciprocal lattice point, the complex transfer function of the lens and the probe position. Simple consideration of the interference condition in the complex plane (figure 5*b*) shows that this phase is not uniquely defined by measuring the three intensities available (of both discs and of the region of overlap). Unique determination is possible, however, if the probe is shifted laterally and the experiment is repeated. Shifting the probe is equivalent to introducing a phase ramp across the back focal plane of the objective lens, say by the introduction of a very thin prism (figure 5*c*). At the plane of their interference, the two discs also possess the phase ramp, but because opposite edges of the aperture overlap, different phase changes are introduced to the two beams before they interfere, thus allowing for a second estimate of their relative phase which, for a carefully chosen probe movement, can be used to solve for the ambiguity in figure 5*b*.

Ptychographical information can also be regarded as residing in image interference fringes available in the conventional electron microscope. In figure 5*d* essentially the same expriment is performed by illuminating the specimen from a point at an angle defined by the region of disc overlap in figure 5*a*. In the back focal plane of the objective lens, two diffraction spots will occur, either side of the optic axis, which will then propagate to the image plane where they will form a set of interference fringes. By reciprocity, this information is identical to a map of the disc-overlap intensity in STEM as a function of all probe positions. The fringes may be thought of as a poor (low-resolution) image of the specimen which maps the periodicity of the unit cell. Information pertaining to the relative phase of the two diffraction spots will lie in the exact position (i.e. the phase) of the interference fringes relative to the optic axis. By choosing many illumination angles, it would be possible in principle (though difficult experimentally) to interfere all pairs of diffraction spots lying in the back-focal plane (even those lying well outside the objective aperture) and thus determine all their relative phases. In none of these experiments would it be necessary to use more than a small region of the object lens close to the optic axis. That is to say, all the measurements can be made within the limits dictated by instrument instability.

Hoppe (1969*a*) originally proposed ptychography as a solution to the conventional diffraction pattern phase problem under circumstances in which it is possible to move a small aperture over the specimen plane. The particular form of the aperture in STEM

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

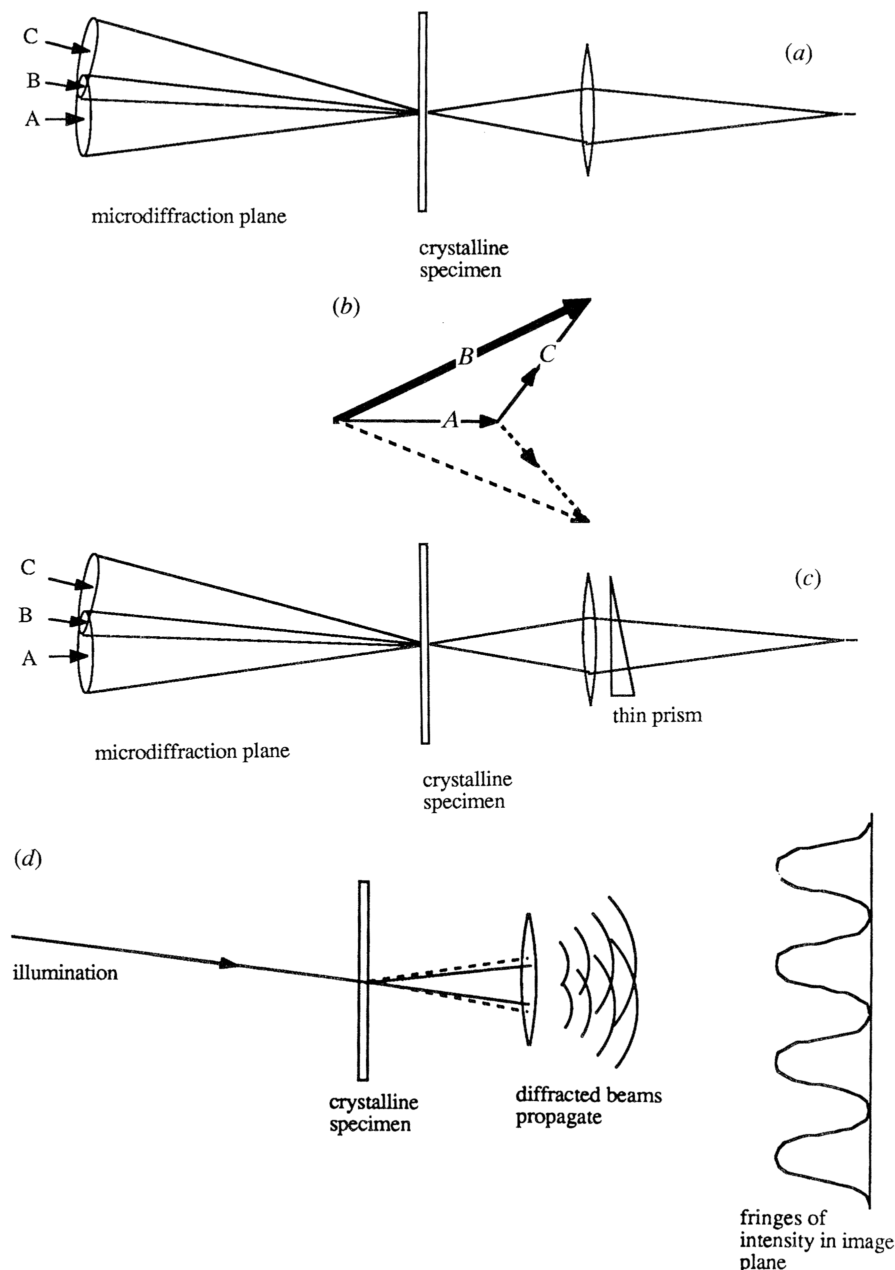


Figure 5. An illustration of the principle of phase determination by ptychography. (*a*) The microdiffraction pattern that will occur when a thin crystalline specimen is illuminated by a STEM probe. For simplicity, only the zero-order and one diffracted order (in the form of a disc) have been shown. The intensity measured at point B in (*a*) is governed by the complex addition shown in (*b*). Complex amplitudes A and C add up to give B. We can measure A^2, B^2 and C^2. This does not determine uniquely the phase relation between A and C (the dotted addition in (*b*) is equally valid, given the available intensity measurements). However, moving the probe, which is equivalent to the introduction of a thin phase wedge shown in (*c*), can resolve the ambiguity by rotating A and C in opposite directions. By reciprocity, the same information resides in the exact position of interference fringes in the conventional two-beam image (*d*).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

(which in this context is not a physical aperture, but corresponds to the probe function at the specimen plane) conveniently results in each diffraction spot being a neat disc which can be arranged to overlap with only one other order at a time. Even so, explicit calculation of the relative phase of each diffraction spot relies on having substantial *a priori* knowledge of the specimen; it must be crystalline and of known unit-cell size. Specimens with very large unit cell size will result in many diffraction orders contributing to each intensity measurement. Furthermore, the exact form of the object lens transfer function as defined by the values of defocus and spherical aberration in equation (1) will radically affect the actual intensity measurements recorded. Indeed, in the worst-case limit of an entirely amorphous, disordered specimen, it is well established that the microdiffraction plane consists of apparently random patches of intensity which can be regarded as a speckle phenomenon principally dominated by the instrumental conditions (Rodenburg 1988). Notwithstanding these difficulties, it is important to note that microdiffraction data collected as a function of all probe positions is rich in phase information and hence super-resolution information; the electron lens is used merely as an imperfect interferometer.

Unlike reference-beam holography, ptychography does not require the entire detector plane to be filled with a reference wave because first-order diffraction discs which interfere with the unscattered beam can themselves be used to phase second-order discs, and so on and so forth. Using diffraction orders themselves as a reference wave should lend the technique much greater experimental stability because, like in conventional diffraction, most of the pertinent interference occurs in the vicinity of the specimen. Furthermore, the Wigner deconvolution described below does not suffer from any ambiguity in the final reconstruction (except in the special case of non-overlapping diffraction discs); there is no need to separate two reconstructions as in Gabor holography. Although this result is evident from the mathematics, some physical insight into where this extra information is coming from may be obtained by considering the Gabor hologram which occurs in the unscattered zero-order beam at the centre of the microdiffraction pattern (note the similarity between the ray diagrams shown in figures 3 and 4*b*). If data from only one probe position is processed, successful reconstruction can only be achieved by severely defocusing the probe so that the two reconstructions (which in the in-line configuration occur on either side of the beam cross-over) can be adequately separated. If the data are also recorded as a function of probe position, however, Lin & Cowley (1989) have suggested that the two reconstructions will appear to move in opposite directions, thus allowing for their unambiguous separation even at small values of defocus. Similar information is available at high scattering angles via ptychography, where no direct reference beam exists, but where probe movement also provides unique solution to the phase problem, as shown in the next section.

(*b*) *Wigner deconvolution and phase-retrieval*

We at first treat the various planes of information in the electron microscope as two dimensional and related to each other by two-dimensional Fourier transforms. We assume, therefore, that all angles involved are small (i.e. $\sin\theta = \theta$) and that the specimen itself is infinitely thin. Although these approximations are commonly used in weak-phase object high-resolution microscopy (Spence 1981), they are much less realistic in the case of the microdiffraction plane, even within the kinematical approximation, for reasons which are discussed and accounted for in the next section.

Downloaded from rsta.royalsocietypublishing.org

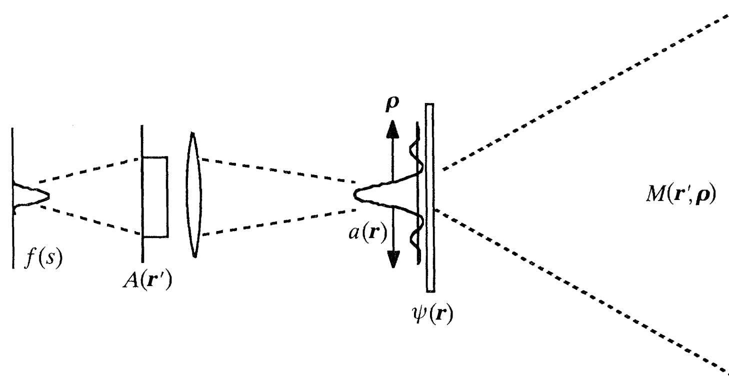


Figure 6. The principle components of the microdiffraction data-set. A source function, $f(\boldsymbol{s})$, illuminates the back focal plane of the objective aperture function, $A(\boldsymbol{r}')$. This distribution is then focused by a lens to a narrow cross-over (the probe function), $a(\boldsymbol{r})$, which lies in the specimen plane. The probe can be moved laterally to various positions $\boldsymbol{\rho}$ across the specimen, $\psi(\boldsymbol{r})$. The microdiffraction plane amplitude $M(\boldsymbol{r}',\boldsymbol{\rho})$ lies in the far-field Fraunhofer diffraction, and its intensity can be measured up to very large scattering angles compared with the convergence angle in the probe (which defines the conventional Rayleigh resolution limit).

It is informative, however, to treat the two-dimensional case first, so that principal characteristics of the Wigner deconvolution method can be illustrated. In what follows, we denote reciprocal space functions by upper case symbols and reciprocal space coordinates by a prime so that, for example, $Q(\boldsymbol{r}')$ is the Fourier transform of $q(\boldsymbol{r})$. The main optical elements of the experiment are shown in figure 6. The probe function, which lies in the specimen plane, is the Fourier transform of the objective aperture function, $A(\boldsymbol{r}')$. In an earlier paper (Bates & Rodenburg 1989) we referred to the probe as $P(\boldsymbol{r})$, but here we will denote it as $a(\boldsymbol{r})$ to emphasize its relationship to the aperture function. Although the amplitude of $A(\boldsymbol{r}')$ is of the form of the top-hat function, its phase depends upon defocus and spherical aberration, as described by equation (1). As a first approximation, the microdiffraction plane is simply the intensity of the Fourier transform of the product of the probe function $a(\boldsymbol{r})$ and the specimen function, which we write as $\psi(\boldsymbol{r})$. We denote the position of the probe with respect to the specimen by the two-dimensional coordinate $\boldsymbol{\rho}$. The complex value of the amplitude at the microdiffraction plane, denoted $M(\boldsymbol{r}',\boldsymbol{\rho})$, as a function of all scattering angles $\boldsymbol{r}'$ and probe positions $\boldsymbol{\rho}$, is then given by

$$M(\boldsymbol{r}',\boldsymbol{\rho}) = \int a(\boldsymbol{r}-\boldsymbol{\rho})\,\psi(\boldsymbol{r})\exp(\mathrm{i}2\pi\boldsymbol{r}\cdot\boldsymbol{r}')\,\mathrm{d}\boldsymbol{r}, \tag{2}$$

where for clarity we have omitted a scaling factor of $1/R\lambda$ in the exponential, where R is the camera length of the microscope (that is, the distance between the specimen plane and the detector plane). It should be noted that in practice R, which is of the order of 10 cm in a typical microscope, is usually a function of the objective lens excitation, because its magnetic field tends to spill over the specimen plane, thus compressing the diffraction pattern. Only the intensity $|M(\boldsymbol{r}',\boldsymbol{\rho})|^2$ can be recorded, which may be written

$$|M(\boldsymbol{r}',\boldsymbol{\rho})|^2 = \iint a(\boldsymbol{b}-\boldsymbol{\rho})\,a^*(\boldsymbol{c}-\boldsymbol{\rho})\,\psi(\boldsymbol{b})\,\psi^*(\boldsymbol{c})\exp[\mathrm{i}2\pi(\boldsymbol{b}\cdot\boldsymbol{r}'-\boldsymbol{c}\boldsymbol{r}')]\,\mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{c}, \tag{3}$$

where $\boldsymbol{b}$ and $\boldsymbol{c}$ have been introduced as dummy variables corresponding to $\boldsymbol{r}$ in

Downloaded from rsta.royalsocietypublishing.org

equation (2), and * denotes the complex conjugate. Alternatively, we may rewrite equation (3) using the convolution theorem in terms of the Fourier transforms of $\psi(\boldsymbol{r})$ and $a(\boldsymbol{r}-\boldsymbol{\rho})$, so that

$$|M(\boldsymbol{r}',\boldsymbol{\rho})|^2 = \iint A(\boldsymbol{b}')\,A^*(\boldsymbol{c}')\,\Psi(\boldsymbol{r}'-\boldsymbol{b}')\,\Psi^*(\boldsymbol{r}'\text{-}\boldsymbol{c}')\exp\left[\mathrm{i}2\pi\boldsymbol{\rho}\cdot(\boldsymbol{b}'-\boldsymbol{c}')\right]\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}', \quad (4)$$

where $\boldsymbol{b}'$ and $\boldsymbol{c}'$ are dummy variables. If the probe is lying on the optic axis, so that $\rho = 0$, the exponential term in equation (4) becomes unity and the expression reduces to the intensity of the convolution of the aperture function $A(\boldsymbol{r}')$ and the reciprocal space of the specimen function $\Psi(\boldsymbol{r}')$, and yields the formulation of microdiffraction derived by Cowley (1978). This is particularly straightforward to visualize in the case when the specimen function is a two-dimensional crystal. $\Psi(\boldsymbol{r}')$ then consists of a series of points each of which lies at the centre of a disc the shape of the shadow image of the objective aperture function, as described in §3*a*. The exponential term can be thought of as the phase ramp across the aperture function, as shown in figure 5*c*, defining the amount of probe shift $\boldsymbol{\rho}$.

We wish to manipulate $|M(\boldsymbol{r}',\boldsymbol{\rho})|^2$ to solve for the complex value of $\psi(\boldsymbol{r})$. We have found that a convenient way of achieving this is to start by taking the forward Fourier transform of $|M(\boldsymbol{r}',\boldsymbol{\rho})|^2$ with respect to $\boldsymbol{\rho}$ and the back Fourier transform with respect to $\boldsymbol{r}'$, to form a quantity we denote as $H(\boldsymbol{r},\boldsymbol{\rho}')$. Rearrangement of the resulting integral shows that each point in $H(\boldsymbol{r},\boldsymbol{\rho}')$ can be expressed as a product of two integrals such that

$$H(\boldsymbol{r},\boldsymbol{\rho}') = \int a^*(\boldsymbol{b})\,a(\boldsymbol{b}+\boldsymbol{r})\exp\left(-\mathrm{i}2\pi\boldsymbol{\rho}'\cdot\boldsymbol{b}\right)\mathrm{d}\boldsymbol{b}\int\psi^*(\boldsymbol{c})\,\psi(\boldsymbol{c}+\boldsymbol{r})\exp\left(\mathrm{i}2\pi\boldsymbol{\rho}'\cdot\boldsymbol{c}\right)\mathrm{d}\boldsymbol{c}. \quad (5)$$

For any general function $q(\boldsymbol{r})$, we can define a quantity

$$\chi_q(\boldsymbol{a},\boldsymbol{b}) = \int q^*(\boldsymbol{c})\,q(\boldsymbol{c}+\boldsymbol{a})\exp\left(\mathrm{i}2\pi\boldsymbol{c}\cdot\boldsymbol{b}\right)\mathrm{d}\boldsymbol{c}, \quad (6)$$

which we refer to as a Wigner distribution, but is sometimes called an offset correlation function or, in signal processing theory, an ambiguity function (for a review of similar functions see Cohen (1989)). This definition allows us to write equation (5) in the compact form

$$H(\boldsymbol{r},\boldsymbol{\rho}') = \chi_a(\boldsymbol{r},-\boldsymbol{\rho}')\,\chi_\psi(\boldsymbol{r},\boldsymbol{\rho}'), \quad (7)$$

and the advantage of forming $H(\boldsymbol{r},\boldsymbol{\rho}')$ now becomes evident since it offers a means of separating out all the influences of the microscope's transfer function which are implicitly contained in $\chi_a(\boldsymbol{r},-\boldsymbol{\rho}')$. Note that the Wigner distribution is highly constrained, being a four-dimensional function derived from only a two-dimensional function.

We can form $\chi_\psi(\boldsymbol{r},\boldsymbol{\rho}')$ by dividing every point in $H(\boldsymbol{r},\boldsymbol{\rho}')$ by $\chi_a(\boldsymbol{r},-\boldsymbol{\rho}')$. This is identical to a conventional deconvolution process, the only difference being that the Fourier transform of the measured data-set has both real and reciprocal coordinates. The usual difficulties associated with a divide-by-zero apply, especially if the original data-set is noisy (which is bound to be the case in practice), and so it is advisable to use a Wiener filter (Bates & McDonnell 1986) such that

$$\chi_\psi(\boldsymbol{r},\boldsymbol{\rho}') = \chi_a^*(\boldsymbol{r},-\boldsymbol{\rho})\,H(\boldsymbol{r},\boldsymbol{\rho}')/(|\chi_a(\boldsymbol{r},-\boldsymbol{\rho}')|^2+\epsilon), \quad (8)$$

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

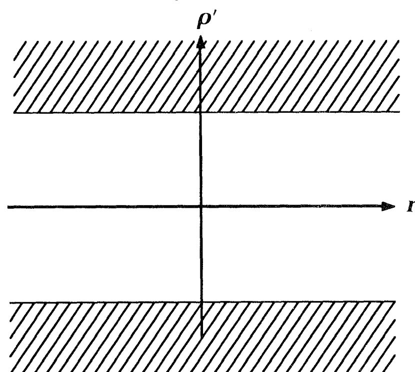


Figure 7. Schematic representation of the data-set $H(\mathbf{r}, \boldsymbol{\rho}')$. Hatched regions lie beyond the $\boldsymbol{\rho}'_{\max}$ cut-off which arises because of the bandwidth-limited nature of the probe. These regions cannot usefully contribute to the deconvolution process.

where ϵ is some small constant. Ideally, ϵ should be a function of both $\mathbf{r}$ and $\boldsymbol{\rho}'$ corresponding to the signal to noise ratio at each point, though experience with conventional deconvolution suggests that a constant term is usually sufficient.

In an earlier paper (Bates & Rodenburg 1989), $\psi(\mathbf{r})$ was obtained from $\chi_{\psi}(\mathbf{r}, \boldsymbol{\rho}')$ by Fourier transforming it with respect to $\mathbf{r}'$, thus obtaining a purely real-space quantity from which $\psi(\mathbf{r})$ could be estimated. It has been found, however, that this leads to certain difficulties associated with the usable region of $H(\mathbf{r}, \boldsymbol{\rho}')$, and can only provide a correct solution if the probe function in real space is finite and sharply truncated. In practice, the probe is bandwidth limited, being composed of a finite range of incident k-vectors. There is therefore only a limited range of points in the $\boldsymbol{\rho}'$ direction, up to a value we call $\boldsymbol{\rho}'_{\max}$, for which $H(\mathbf{r}, \boldsymbol{\rho}')$ has significant value. This is equivalent to stating that any conventional bright- or dark-field image (collected as a function of $\boldsymbol{\rho}$ at a particular value of $\mathbf{r}'$) has a maximum spatial frequency component, which of course corresponds to the conventional resolution limit defined by the size of the objective aperture. Note that here we do not refer to the interpretable resolution limit as in the case of phase-contrast imaging (which depends also upon the exact form of the complex transfer function of the objective lens), but merely the maximum spatial frequency which could exist in the image plane, and which is governed by either the size of the objective aperture or the width of the stability component of the instrument function of the microscope, whichever is narrower.

Consider figure 7 which is a schematic, two-dimensional plot of $H(\mathbf{r}, \boldsymbol{\rho}')$, though in reality it would be a four-dimensional function since both $\mathbf{r}$ and $\boldsymbol{\rho}'$ are two-dimensional vectors. By inspecting equation (6), it is clear that along $\mathbf{r} = 0$, $\chi_a(\mathbf{0}, -\boldsymbol{\rho}')$ will have the form of the Fourier transform of the intensity of probe, or, equivalently, the autocorrelation of the objective aperture function $A(\mathbf{r}')$. Even at other values of $\mathbf{r}$, it is evident that in the shaded regions of the plane lying beyond the total aperture width $\boldsymbol{\rho}'_{\max}$ in the $\boldsymbol{\rho}'$ direction, $\chi_a(\mathbf{r}, -\boldsymbol{\rho}')$ will be zero and so these regions cannot usefully contribute to the deconvolution process. In what follows, however, 'super-resolution' information is derived from the $\mathbf{r}$-coordinate, which has been originally obtained from the high-angle scattering data in the microdiffraction plane. In $H(\mathbf{r}, \boldsymbol{\rho}')$, super-resolution data exist at small values of $\mathbf{r}$, where $\chi_a(\mathbf{r}, -\boldsymbol{\rho}')$ is large, so presenting no problems with the Wiener filtering process.

Having obtained an estimate of $\chi_{\psi}(\mathbf{r}, \boldsymbol{\rho}')$ in the unshaded regions of figure 7,

Downloaded from rsta.royalsocietypublishing.org

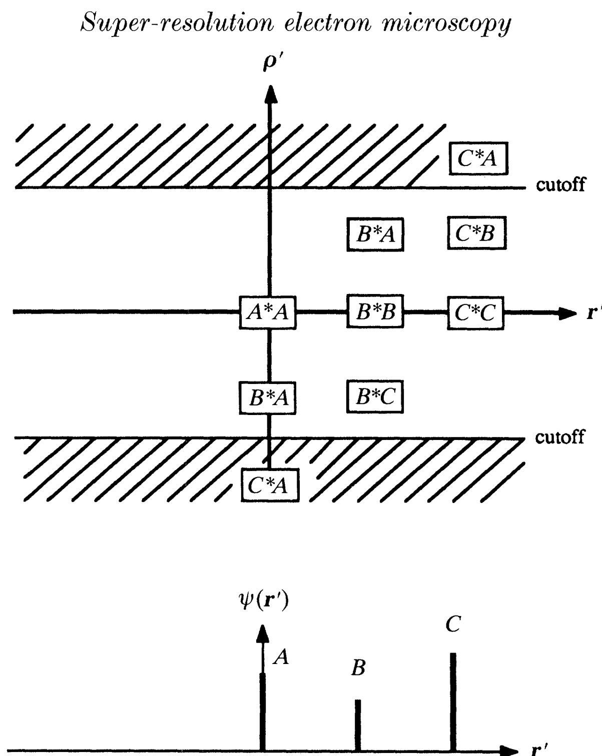


Figure 8. Schematic representation of the data-set $D(\boldsymbol{r}', \boldsymbol{\rho}')$, showing the positions and values of peaks for a simple (and rather artificial) one-dimensional specimen. The diffraction pattern of the specimen is plotted in the lower diagram, where it should be remembered that A, B and C may be complex. $D(\boldsymbol{r}', \boldsymbol{\rho}')$ consists of a series of delta-function spikes, positioned in the centre of the square boxes. Despite the $\boldsymbol{\rho}'_{\max}$ cut-off, all the beams can be phased with respect to one another (see text for details).

solution of the specimen function is best achieved by Fourier transforming this with respect to the $\boldsymbol{r}$ coordinate, to give a function we call $D(\boldsymbol{r}', \boldsymbol{\rho}')$, such that

$$D(\boldsymbol{r}', \boldsymbol{\rho}') = \Psi^*(\boldsymbol{r}' - \boldsymbol{\rho}')\,\Psi(\boldsymbol{r}'), \tag{9}$$

where it should be noted that $\Psi(\boldsymbol{r}')$ is the Fourier transform of the specimen function $\psi(\boldsymbol{r})$. We can envisage $D(\boldsymbol{r}', \boldsymbol{\rho}')$, which is shown schematically in figure 8, as consisting of many estimates of the reciprocal space of the specimen function, each lying in the $\boldsymbol{\rho}'$ direction at fixed $\boldsymbol{r}'$, and each multiplied by the value of $\Psi(\boldsymbol{r}')$ lying at $D(\boldsymbol{r}', \boldsymbol{0})$. Note that there is still a cut-off at $\boldsymbol{\rho}'_{\max}$. The conventional diffraction pattern given by $\Psi^*(\boldsymbol{r}')\,\Psi(\boldsymbol{r}')$ (which is now deconvolved from the broadening effects of the range of incident angles in the objective aperture), lies along $\boldsymbol{\rho}' = 0$. For a full solution of the complex value of $\Psi(\boldsymbol{r}')$, we must use information lying at $\boldsymbol{\rho}' \neq 0$ and hence find relative phases between different parts of the diffraction pattern via equation (9). It is logical that phase information has arrived along the $\boldsymbol{\rho}'$ coordinate, which corresponds to the reciprocal of the probe-position coordinate, or the reciprocal of the conventional image plane. As pointed out in the previous section, probe movement or poor-resolution image structure allows for ptychographical phase-retrieval, though after the deconvolution process, we no longer have to worry about the influences of phase changes introduced by the objective aperture function. It is of course impossible to find the absolute phase of the wavefield, which is anyway a

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

meaningless quantity: multiplying the whole diffraction pattern by a constant phase change will simply alter the phase of the whole image by a similar amount, but all structural information will remain intact. It is necessary, however, to make an arbitrary assignment of phase to one point in the pattern and work from there.

$D(\boldsymbol{r}', \boldsymbol{\rho}')$ is highly redundant: it is a four-dimensional function from which we wish to solve for a two-dimensional image. Given that the original data-set will be noisy and that it is unlikely that the complex transfer function of the microscope will be accurately known, this extra information will be invaluable. Exactly how to proceed from equation (9) will depend upon the type of best-estimate algorithm used. For example, having arrived at $D(\boldsymbol{r}', \boldsymbol{\rho}')$, it may be found that the relative moduli of the estimates lying along the $\boldsymbol{\rho}'$ direction do not correspond to the modulus calculated from the square root of the intensity lying along $\boldsymbol{\rho}' = 0$, implying that the original probe function was incorrectly estimated. There would then be scope for iterative convergence on a better estimate of $a(\boldsymbol{r})$. There is a similarly large number of possible methods for deriving the phase of $\Psi(\boldsymbol{r}')$ from $D(\boldsymbol{r}', \boldsymbol{\rho}')$. The simplest scheme would be to designate $\Psi(0)$ as purely real (so assigning the arbitrary absolute phase of the wavefield) and equal to the square root of $D(0,0)$. Along $\boldsymbol{r}' = 0$, we can then write

$$\Psi(-\boldsymbol{\rho}') = D^*(0, \boldsymbol{\rho}')/\sqrt{D(0,0)}, \tag{10}$$

which gives us the complex value of $\Psi(\boldsymbol{r}')$, but only as far as the cut-off at $\boldsymbol{\rho}'_{max}$. (The variables $\boldsymbol{r}'$ and $\boldsymbol{\rho}'$ are interchangeable as far as Ψ is concerned, being both reciprocal coordinates arising from the redundancy in $D(\boldsymbol{r}', \boldsymbol{\rho}')$.) Having thus determined the phase of a point lying at, say, $\boldsymbol{w}'$, where $|\boldsymbol{w}'| < |\boldsymbol{\rho}'_{max}|$ we can then move to a strip in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ lying along $\boldsymbol{r}' = \boldsymbol{w}'$, and write

$$\Psi(\boldsymbol{w}' - \boldsymbol{\rho}') = D^*(\boldsymbol{w}', \boldsymbol{\rho}')/\Psi^*(\boldsymbol{w}'). \tag{11}$$

Although the cut-off at $\boldsymbol{\rho}'_{max}$ again limits the usefulness of this equation, we now have access to the complex value of a region of reciprocal space displaced by $\boldsymbol{w}'$ and hence we can find measurable estimates of $\Psi(\boldsymbol{w}' - \boldsymbol{\rho}')$ where in some parts $|\boldsymbol{w}' - \boldsymbol{\rho}'| > |\boldsymbol{\rho}'_{max}|$.

Consider, for example, a crystalline specimen with reciprocal lattice peaks of complex value A, B and C, also shown in figure 8. Along $\boldsymbol{\rho}' = 0$ in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ we have the conventional diffraction peaks of height A^*A, B^*B and C^*C. B^*A occurs before the $\boldsymbol{\rho}'_{max}$ cut-off along $\boldsymbol{r}' = 0$ in $D(\boldsymbol{r}', \boldsymbol{\rho}')$, and so assuming we assign arbitrary phase zero to A, we can derive the complex value of B by performing a division similar to equation (10). Along this first strip of data, however, C^*A is obscured by the $\boldsymbol{\rho}'_{max}$ cut-off. Having measured B, we are then able to move to the value of $\boldsymbol{r}'$ in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ where an estimate of C^*B lies within the range of the cut-off, thus allowing us to find the complex value of C relative to B. Repeating this process, estimates of the relative phases of all reciprocal points lying at arbitrarily large reciprocal vectors can be found. A final Fourier transform of $\Psi(\boldsymbol{r}')$ yields the complex, super-resolution image, $\psi(\boldsymbol{r})$.

It is apparent that in the case of a perfectly crystalline specimen of small unit cell size, even the first reciprocal lattice point along $\boldsymbol{r}' = 0$ in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ may be obscured by the cut-off at $\boldsymbol{\rho}'_{max}$. Under these circumstances, the microdiffraction data fail to provide any ptychographical phase information because the reciprocal lattice points are more widely spaced than the total diameter of the objective aperture, so that the discs shown in figure 5*a* do not overlap. This is the price which must be paid for having a limited objective aperture. Because the phasing of higher-order beams relies on interference with lower-order beams, the specimen must either be finite (in which

Downloaded from rsta.royalsocietypublishing.org

case all diffraction orders are blurred out), of large unit cell size (and possibly infinite), or amorphous. In practice, any real specimen of very small unit cell size could be rendered finite or pseudo-amorphous by the addition of some randomly positioned heavy atoms, though this condition is unlikely to occur in practice given that in conventional microscopy it is usually possible to accommodate at least the first-order diffraction spots within the transfer function of the objective lens. Furthermore, it may be argued that since this technique fully accounts for the instrument function, there is really no need for an objective aperture at all. The limit of usable $\boldsymbol{\rho}'$ will then be determined by the stability (or coherence) function of the microscope. In §5, however, it will be shown that it may nevertheless be advantageous to retain the objective aperture in order to overcome complications which arise from the rapid breakdown of the two-dimensional projection approximation at high angles of scatter.

Finally, it is worth mentioning that when $\boldsymbol{r}'$ and $\boldsymbol{\rho}'$ are both two-dimensional vectors, the possibilities for constraining the phase assignment process become much richer. At any value of $\boldsymbol{r}'$, there will be a two-dimensional complex diffraction pattern available in $D(\boldsymbol{r}', \boldsymbol{\rho}')$. We could choose to calculate the phase difference between any point and $\boldsymbol{\rho}' = 0$ by a number of different routes, which take different paths around the plane, and average many such routes to obtain a more noise-robust measure of any particular phase.

4. Effect of thickness

Consider a three-dimensional specimen, which we denote $\psi(\boldsymbol{r}, z)$, where as before, $\boldsymbol{r}$ is a two-dimensional vector lying in the plane normal to the optic axis. Let Z be a scalar distance measured along the optic axis. We define the Fourier transform of $\psi(\boldsymbol{r}, z)$ as

$$\Psi(\boldsymbol{r}', z') = \int\int\int \psi(\boldsymbol{r}, z) \exp[\mathrm{i}2\pi(\boldsymbol{r}' \cdot \boldsymbol{r} + z'z)]\, \mathrm{d}z\, \mathrm{d}\boldsymbol{r}, \tag{12}$$

where the three integral signs have been written out explicitly to exphasize the number of dimensions over which we must perform the transform. In figure 9, we show $\Psi(\boldsymbol{r}', z')$ together with the Ewald sphere construction for scattering in the case of microdiffraction. Instead of the single incident vector usually encountered in scattering theory, in microdiffraction we have a range of incident beams, defined by the width of the objective aperture. The strength of scattering in a particular direction can be thought of as an integral in reciprocal space over a small, slightly curved disc represented by the surface A' (Hoppe & Hegerl 1980; Rodenburg 1988), a slice through which is shown in the diagram. To account for aberrations and defocus, A' must also introduce the appropriate phase changes defined by equation (1), with a value of defocus taken relative to the origin in z. When both this surface and the curvature of the Ewald sphere are approximated as being flat, the microdiffraction intensity is adequately represented by equation (4), namely a convolution which depends entirely on $\boldsymbol{r}'$ and has no z' dependence. Under these circumstances we are effectively treating the specimen as a projection, such that

$$\psi(\boldsymbol{r}) = \int \psi(\boldsymbol{r}, z)\, \mathrm{d}z, \tag{13}$$

where the Fourier transform of $\psi(\boldsymbol{r})$ resides along $z' = 0$ in $\Psi(\boldsymbol{r}', z')$. We may derive

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

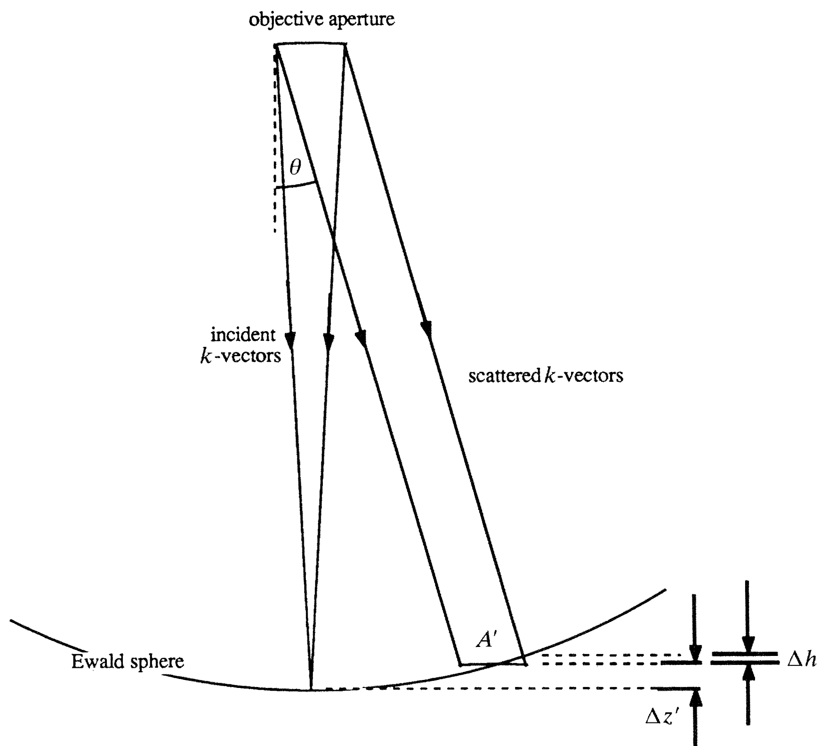


Figure 9. The Ewald sphere construction for microdiffraction. A range of incident k-vectors extend from the objective aperture to the origin of reciprocal space. Scattering to a particular angle θ is governed by an integration over the slightly curved surface A'. The departure of this surface from the Ewald sphere by $\Delta h'$ is less significant than the departure of the Ewald sphere by $\Delta z'$ from the plane of the projection approximation.

an approximate value of the maximum allowable thickness for this approximation as follows. For a given scattering angle θ, let the distance in reciprocal space by which the Ewald sphere has departed from $z' = 0$ be $\Delta z'$ (see figure 9), so that

$$\Delta z' = |k|(1-\cos\theta) \approx \tfrac{1}{2}|k|\,\theta^2, \tag{14}$$

where $|k|$ is the magnitude of the incident k-vector given by $2\pi/\lambda$. We expect the typical size of a feature in reciprocal space in the z' direction to be about $2\pi/T$, where T is the specimen thickness. Furthermore, for a given resolution Δr, we need to process scattering data up to an angle of $\theta \approx \lambda/\Delta r$. Assuming that the largest allowable value of $\Delta z'$ is $2\pi/T$, we can then write

$$T \approx 2(\Delta r)^2/\lambda. \tag{15}$$

If we require Δr to be, say, 1.8 Å, as in a conventional high-resolution electron microscope, then the maximum specimen thickness is about $T = 175$ Å for an accelerating voltage of 100 keV ($\lambda = 0.037$ Å). This is quite possible to achieve in practice (self-supporting carbon films, for example, can be manufactured to be as little as 20 Å thick). Indeed, in many classes of crystalline specimen, dynamical scattering effects would already be significant at this value of thickness, and so for genuine phase-contrast imaging, specimens should ideally be thinner. Suffice it to say

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

that in bright-field imaging, breakdown of the projection approximation is not a serious problem. Equivalently, we may observe that the depth of focus of the electron optics is much larger than the typical specimen thickness. In the microdiffraction plane, however, there is no real restriction on the size of detector plane and hence the total angle θ contributing to the experiment. Of course, this is where the super-resolution information resides, but only at the price of greatly restricting the useful values of T. It would be quite practicable to collect all scattering data up to, say, ± 37 mrad (about five times the width of the objective aperture) with the intention of solving for the specimen function at 0.5 Å resolution, but the projection approximation would fail for any specimen thicker than only 13.5 Å, which in practice would be virtually impossible to manufacture. The difficulty arises from the squared Δr term in equation (15), which can be understood intuitively by considering apparent changes in the projection of the specimen in real space. As it is viewed from different angles, atoms residing at the top and bottom surfaces will appear to move with respect to one another. If the resolution is increased by a certain factor, the ability to detect changes in orientation will increase by the same amount, and thus restrict the values of θ for which the projection approximation holds. However, to achieve that gain in resolution, the range of angles processed by the microscope must also increase by the same factor, leading to a minimum thickness proportional to the square of the desired resolution.

In the next section, we show that blind application of the Wigner distribution deconvolution will yield an estimate of the complex value of the reciprocal space $\Psi(\boldsymbol{r}', z')$ over the surface of the Ewald sphere, at least to within a limit dictated by the size of the objective aperture, which turns out to be much less severe than the breakdown of the projection approximation outlined above. Having obtained this quantity, which is as much as we could wish to measure from a single scattering experiment, we may choose to pick out a particular level of defocus within the bulk of the specimen by emulating a conventional, unaberrated lens which re-interferes the complex wavefield. After all, in normal light optics, the lens processes a wide region of reciprocal space which reconstructs a particular level of the specimen at high resolution. In the final image, parts of the specimen not lying in the focal plane appear as a blurred background. Of course, it is impossible to reconstruct a three-dimensional object from a two-dimensional data-set, but the curvature of the Ewald sphere does contain at least some information about the z coordinate which should not be ignored. In the case of a specimen consisting of heavy atoms distributed at low density in a weakly scattering matrix, it may be possible to weigh the final reconstruction in favour of one particular level in the specimen, in the same way as a biological scientist may focus an optical microscope on one cell, provided the scattering from material above and below the region of interest does not heavily obscure the image.

We use the parabolic approximation for the curvature of the Ewald sphere represented by equation (14). For all practical purposes, this will be sufficiently accurate given that the conventional resolution limit could be surpassed by a factor of 4 by processing only up to about $\theta = \pm 5°$ of the microdiffraction plane. Accounting for the very small curvature in the disc A' (figure 9) is possible but in the present context will be ignored. It is equivalent to the fact that in real space, the probe is focused through a varying beam waist; in other words it is z-dependent. If a conventional objection aperture is used (with a semi-angle of 8 mrad), at the specimen plane the probe is diffraction-limited, relatively broad and has such a

Downloaded from rsta.royalsocietypublishing.org

shallow angle of convergence that its profile does not vary appreciably over a depth (Δz) of several hundred ångströms. The argument is similar to that described above; for this range of convergence angles the projection approximation is valid. In the next section, therefore, we treat the probe as a two-dimensional function which does not vary significantly through the specimen depth, but the specimen is treated as three dimensional to account for the much larger curvature of the Ewald sphere at high angles of θ.

5. Full theory, including effects of partial coherence

For completeness, in this section we derive the full kinematical theory including finite specimen thickness and the effects of partial coherence in the illuminating beam (i.e. finite source size). The main mathematical tool we use for manipulation is collapsing multidimensional Fourier integrals by means of the Dirac delta function. In practice, however, it will not be possible to perform integrals over all space because our data-set will necessarily be limited by both the microdiffraction plane detector size (in reciprocal space) and the region of specimen over which we can move the probe (in real space). These limitations can be expressed by a window function

$$w\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{q}' \end{pmatrix},$$

dependent upon either a real-space coordinate $\boldsymbol{q}$ or a reciprocal space coordinate $\boldsymbol{q}'$. It follows that we may then define an effective delta function given by

$$\tilde{\delta}\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{q}' \end{pmatrix} = \int\!\!\int w\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{q}' \end{pmatrix} \exp\left(\pm \mathrm{i}2\pi \begin{pmatrix} \boldsymbol{q}'\cdot\boldsymbol{q} \\ \boldsymbol{q}\cdot\boldsymbol{q}' \end{pmatrix}\right) \mathrm{d}\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{q}' \end{pmatrix} \tag{16}$$

so that we may write that for some general function Q,

$$\int\!\!\int Q\begin{pmatrix} \boldsymbol{q}' \\ \boldsymbol{q} \end{pmatrix} \tilde{\delta}\begin{pmatrix} \boldsymbol{q}'-\boldsymbol{p}' \\ \boldsymbol{q}-\boldsymbol{p} \end{pmatrix} \mathrm{d}\begin{pmatrix} \boldsymbol{q}' \\ \boldsymbol{q} \end{pmatrix} = Q\begin{pmatrix} \boldsymbol{p}' \\ \boldsymbol{p} \end{pmatrix} \tag{17}$$

on the understanding that

$$Q\begin{pmatrix} \boldsymbol{p}' \\ \boldsymbol{p} \end{pmatrix}$$

is resolved only to the limit inherent in the extent of

$$w\begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{q}' \end{pmatrix}.$$

Unlike in §3*b*, where one integral sign was used for sums over vector quantities, in what follows, each sign for each dimension will be written out explicitly.

With reference to figure 6, we allow the electron source to have an instantaneous spatial and temporal complex amplitude denoted $f(\boldsymbol{s}, t)$, where $\boldsymbol{s}$ is a two-dimensional vector in the source plane and t is time. In practice, the objective lens demagnifies by a factor of between 10 and 20 and there may be a gun or condenser lens mounted before it. However, for theoretical purposes, we can regard the source as being focused directly onto the specimen via a unity-magnification objective, it being understood that the shape chosen for $f(\boldsymbol{s}, t)$ accommodates these other effects. All time-varying quantities are understood to be analytical.

Downloaded from rsta.royalsocietypublishing.org

Each point in the source is incoherent with respect to any other point, so that

$$\langle f(\boldsymbol{s},t)f^*(\boldsymbol{s}+\boldsymbol{\sigma},t)\rangle_t = |f(\boldsymbol{s})|^2\delta(\boldsymbol{\sigma}), \tag{18}$$

where angle brackets represent an average over time, and δ is the conventional Dirac delta function.

The back focal plane of the objective lens lies in the Fraunhofer diffraction plane of the source. As a function of reciprocal coordinate $\boldsymbol{s}'$, the complex amplitude of the field in the back focal plane is then given by $F(\boldsymbol{s}',t)\,A(\boldsymbol{s}')$ where

$$F(\boldsymbol{s}',t) = \iint f(\boldsymbol{s},t)\exp\left[\mathrm{i}2\pi(\boldsymbol{s}'\cdot\boldsymbol{s})\right]\mathrm{d}\boldsymbol{s} \tag{19}$$

(the time taken for propagation of the wavefield has been ignored), and $A(\boldsymbol{s}')$ is the objective lens transfer function which will have unity amplitude within the objective aperture, zero amplitude outside it, and a phase governed by equation (1), with $\theta = |\boldsymbol{s}'|\,\lambda$. The field illuminating the specimen when the probe is centred on a position $\boldsymbol{\rho}$ is then given by

$$g(\boldsymbol{r},z,\boldsymbol{\rho},t) = [f(\boldsymbol{r},t)\odot a(\boldsymbol{r}-\boldsymbol{\rho})]\exp(-\mathrm{i}kz), \tag{20}$$

where $\odot$ is the two-dimensional convolution operator, $a(\boldsymbol{r})$ is the Fourier transform of the aperture function $A(\boldsymbol{r}')$, k is the wavenumber of the source, and z is the distance measured along the optic axis. We have implicitly used the approximation that the probe does not vary throughout the thickness of the specimen, as discussed in the previous section. Bearing in mind that we are working under the first Born approximation, the scattering amplitude in the vicinity of the specimen is now given by

$$m(\boldsymbol{r},z,\boldsymbol{\rho},t) = \psi(\boldsymbol{r},z)\,g(\boldsymbol{r},z,\boldsymbol{\rho},t), \tag{21}$$

where we adopt the three-dimensional specimen function $\psi(\boldsymbol{r},z)$ as discussed earlier. The far-field microdiffraction plane as a function of scattering vector $\boldsymbol{r}'$, probe position $\boldsymbol{\rho}$, and time t is then

$$M(\boldsymbol{r}',\boldsymbol{\rho},t) = \iiint m(\boldsymbol{r},z,\boldsymbol{\rho},t)\exp\left[\mathrm{i}2\pi(\boldsymbol{r}'\cdot\boldsymbol{r}+\tfrac{1}{2}z\lambda|\boldsymbol{r}'|^2)\right]\mathrm{d}\boldsymbol{r}\,\mathrm{d}z, \tag{22}$$

where the second term in the exponential is a parabolic approximation for the curvature of the Ewald sphere, also discussed in the previous section. We record intensity over many cycles of the electron wavefield, and so the measured quantity in the microdiffraction plane is the time average

$$I(\boldsymbol{r}',\boldsymbol{\rho}) = \langle M^*(\boldsymbol{r}',\boldsymbol{\rho},t)M(\boldsymbol{r}',\boldsymbol{\rho},t)\rangle_t. \tag{23}$$

It should be noted that any instability in the specimen or the source (in STEM the largest interference tends to be stray magnetic field reaching the field-emission gun, adding ripple to the effective probe position) may be accommodated by choosing an integration time in the detector plane comparable with the period of the oscillation (e.g. the mains supply period) and by using a time-averaged source distribution for $|f(\boldsymbol{s})|^2$.

In what follows, $\Gamma(\boldsymbol{s}')$ is defined as the two-dimensional Fourier transform of the intensity of the source function $|f(\boldsymbol{s})|^2$. It is well known from light optics (the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

Van-Cittert–Zernike theorem (Born & Wolf 1964)) that $\Gamma(\boldsymbol{s}')$ corresponds to the coherence function of the source, namely the propensity for two regions of the wavefield lying in the back-focal plane to be in phase with one another. From equations (20)–(23), where the number beneath the integral indicates the number of integrals,

$$\begin{aligned} I(\boldsymbol{r}',\boldsymbol{\rho}) &= \int_6 m(\boldsymbol{b},z,\boldsymbol{\rho},t)\, m^*(\boldsymbol{c},z,\boldsymbol{\rho},t) \exp\{\mathrm{i}2\pi[\boldsymbol{r}'\cdot(\boldsymbol{b}-\boldsymbol{c})]\} \exp\{\mathrm{i}\pi[(z-\zeta)\,\lambda|\boldsymbol{r}'|^2]\} \\ &\quad \times \mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{c}\,\mathrm{d}z\,\mathrm{d}\zeta \\ &= \int_{10} \langle f(\boldsymbol{p},t) f^*(\boldsymbol{q},t)\rangle_t\, a(\boldsymbol{b}-\boldsymbol{\rho}-\boldsymbol{q})\, a^*(\boldsymbol{c}-\boldsymbol{\rho}-\boldsymbol{q})\, \psi(\boldsymbol{b},z)\, \psi^*(\boldsymbol{c},\zeta) \\ &\quad \times \exp\{\mathrm{i}2\pi[\boldsymbol{r}'\cdot(\boldsymbol{b}-\boldsymbol{c})]\} \exp\{\mathrm{i}\pi[(z-\zeta)\,\lambda|\boldsymbol{r}'|^2]\}\,\mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{c}\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}z\,\mathrm{d}\zeta \\ &= \int_8 \langle |f(\boldsymbol{q})|^2\rangle_t\, a(\boldsymbol{b}-\boldsymbol{\rho}-\boldsymbol{q})\, a^*(\boldsymbol{c}-\boldsymbol{\rho}-\boldsymbol{q})\, \psi(\boldsymbol{b},z)\, \psi^*(\boldsymbol{c},\zeta) \\ &\quad \times \exp\{\mathrm{i}2\pi[\boldsymbol{r}'\cdot(\boldsymbol{b}-\boldsymbol{c})]\} \exp\{\mathrm{i}\pi[(z-\zeta)\,\lambda|\boldsymbol{r}'|^2]\}\,\mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{c}\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}z\,\mathrm{d}\zeta. \end{aligned} \tag{24}$$

Note that, as in §3*b*, $\boldsymbol{r}'$ and the total amplitude of this integral must be scaled appropriately in view of the microscope dimensions, and that the exponential factor in equation (20) can be ignored, given that in the Ewald sphere construction $\boldsymbol{r}'$ is obtained by subtracting the incident k-vector from the scattered k-vector.

It is now convenient to express all quantities in terms of Fourier transforms, such that

$$\begin{aligned} I(\boldsymbol{r}',\boldsymbol{\rho}) &= \int_{20} \Gamma(\boldsymbol{q}')\, A(\boldsymbol{b}')\, A^*(\boldsymbol{c}')\, \Psi(\boldsymbol{m}',z')\, \Psi^*(\boldsymbol{n}',z) \\ &\quad \times \exp\{\mathrm{i}2\pi[\boldsymbol{c}'\cdot(\boldsymbol{c}-\boldsymbol{\rho}-\boldsymbol{q}) - \boldsymbol{b}'\cdot(\boldsymbol{b}-\boldsymbol{\rho}-\boldsymbol{q}) - \boldsymbol{q}'\cdot\boldsymbol{q} + \boldsymbol{n}'\cdot\boldsymbol{c} - \boldsymbol{m}'\cdot\boldsymbol{b} - z'\cdot z \\ &\qquad + z'\cdot z + \boldsymbol{r}'\cdot(\boldsymbol{b}-\boldsymbol{c}) + \tfrac{1}{2}(z-\zeta)\,\lambda|\boldsymbol{r}'|^2]\} \\ &\quad \times \mathrm{d}\boldsymbol{b}\,\mathrm{d}\boldsymbol{c}\,\mathrm{d}\boldsymbol{q}\,\mathrm{d}\boldsymbol{q}'\,\mathrm{d}z\,\mathrm{d}\zeta\,\mathrm{d}z'\,\mathrm{d}\zeta'\,\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}'\,\mathrm{d}\boldsymbol{m}'\,\mathrm{d}\boldsymbol{n}' \\ &= \int_{12} \Gamma(\boldsymbol{q}')\, A(\boldsymbol{b}')\, A^*(\boldsymbol{c}')\, \Psi(\boldsymbol{m}',z')\, \Psi(\boldsymbol{n},z) \\ &\quad \times \delta(\boldsymbol{r}'-\boldsymbol{m}'-\boldsymbol{b}')\, \delta(\boldsymbol{c}'+\boldsymbol{n}'-\boldsymbol{r}')\, \delta(\boldsymbol{b}'-\boldsymbol{c}'-\boldsymbol{q}')\, \delta(z'-\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\, \delta(\zeta-\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2) \\ &\qquad \times \exp[\mathrm{i}2\pi(\boldsymbol{b}'-\boldsymbol{c}')\cdot\boldsymbol{\rho}]\,\mathrm{d}\boldsymbol{q}'\,\mathrm{d}z'\,\mathrm{d}\zeta'\,\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}'\,\mathrm{d}\boldsymbol{m}'\,\mathrm{d}\boldsymbol{n}' \\ &= \int_4 \Gamma(\boldsymbol{b}'-\boldsymbol{c}')\, A(\boldsymbol{b}')\, A^*(\boldsymbol{c}')\, \Psi(\boldsymbol{r}'-\boldsymbol{b}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\, \Psi^*(\boldsymbol{r}'-\boldsymbol{c}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2) \\ &\quad \times \exp[\mathrm{i}2\pi(\boldsymbol{b}'-\boldsymbol{c}')\cdot\boldsymbol{\rho}]\,\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}', \end{aligned} \tag{25}$$

which may be compared with equation (4). Being only able to scan a limited region of real space with the probe, Ψ will only be resolved to a particular resolution in

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

reciprocal space. We may express this by the use of a window function $w(\boldsymbol{\rho})$, such that the quantity $H(\boldsymbol{r},\boldsymbol{\rho}')$ used in the Wigner deconvolution is the Fourier transform of $I(\boldsymbol{r}',\boldsymbol{\rho})\,w(\boldsymbol{\rho})$. Hence

$$\begin{aligned} H(\boldsymbol{r},\boldsymbol{\rho}') &= \int_8 \Gamma(\boldsymbol{b}'-\boldsymbol{c})\,A(\boldsymbol{b}')\,A^*(\boldsymbol{c}')\,\Psi(\boldsymbol{r}'-\boldsymbol{b}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\,\Psi^*(\boldsymbol{r}'-\boldsymbol{c}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\,w(\boldsymbol{\rho}) \\ &\qquad \times \exp\left[\mathrm{i}2\pi(\boldsymbol{b}'\cdot\boldsymbol{\rho}-\boldsymbol{c}'\cdot\boldsymbol{\rho}+\boldsymbol{\rho}'\cdot\boldsymbol{\rho}-\boldsymbol{r}'\cdot\boldsymbol{r})\right]\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}'\,\mathrm{d}\boldsymbol{r}'\,\mathrm{d}\boldsymbol{\rho} \\ &= \int_6 \Gamma(\boldsymbol{b}'-\boldsymbol{c}')\,A(\boldsymbol{b}')\,A^*(\boldsymbol{c}')\,\Psi(\boldsymbol{r}'-\boldsymbol{b}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\,\Psi^*(\boldsymbol{r}'-\boldsymbol{c}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2) \\ &\qquad \times \tilde{\delta}(\boldsymbol{\rho}'+\boldsymbol{b}'-\boldsymbol{c}')\exp\left[-\mathrm{i}2\pi\boldsymbol{r}'\cdot\boldsymbol{r}\right]\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{c}'\,\mathrm{d}\boldsymbol{r}' \\ &= \Gamma(-\boldsymbol{\rho}')\int_4 A(\boldsymbol{b}')\,A^*(\boldsymbol{b}'+\boldsymbol{\rho}')\,\Psi(\boldsymbol{r}'-\boldsymbol{b}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2)\,\Psi^*(\boldsymbol{r}'-\boldsymbol{b}'-\boldsymbol{\rho}',\tfrac{1}{2}\lambda|\boldsymbol{r}'|^2) \\ &\qquad \times \exp\left[-\mathrm{i}2\pi\boldsymbol{r}'\cdot\boldsymbol{r}\right]\mathrm{d}\boldsymbol{b}'\,\mathrm{d}\boldsymbol{r}' \\ &= \Gamma(-\boldsymbol{\rho}')\iint A(\boldsymbol{b}')\,A^*(\boldsymbol{b}'+\boldsymbol{\rho}')\exp\left[-\mathrm{i}2\pi\boldsymbol{b}'\cdot\boldsymbol{r}\right]\tilde{\chi}^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r},\boldsymbol{b}')\,\mathrm{d}\boldsymbol{b}', \end{aligned} \tag{26}$$

where

$$\tilde{\chi}^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r},\boldsymbol{b}') = \iint \Psi(\boldsymbol{c}',\tfrac{1}{2}\lambda|\boldsymbol{c}'+\boldsymbol{b}'|^2)\,\Psi^*(\boldsymbol{c}'+\boldsymbol{\rho}',\tfrac{1}{2}\lambda|\boldsymbol{c}'+\boldsymbol{b}'|^2)\exp\left[-\mathrm{i}2\pi\boldsymbol{c}'\cdot\boldsymbol{r}\right]\mathrm{d}\boldsymbol{c}', \tag{27}$$

the tilda indicating that this can only be resolved to a 'reciprocal resolution' dictated by the size of the probe-movement window function.

Equation (26) has a similar form to equation (5), though now that the thickness dependence has been included, the integrals are no longer separable. In the limit of an infinitesimally thin specimen, then

$$\tilde{\chi}^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r},\boldsymbol{b}') = \tilde{\chi}^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r},\boldsymbol{0}) = \chi^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r}), \tag{28}$$

where $\chi^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r})$ is defined as before by equation (6). Note that under this definition, the relationship between Wigner distributions written in terms of some general real-space function q and its Fourier transform Q is given by

$$\chi_q(\boldsymbol{a},\boldsymbol{b}) = \chi^*_Q(-\boldsymbol{b},\boldsymbol{a}), \tag{29}$$

and thus the thin-specimen relationship equivalent to equation (7), but now including partial coherence effects, would be

$$H(\boldsymbol{r},\boldsymbol{\rho}') = \Gamma(-\boldsymbol{\rho}')\,\chi^*_A(\boldsymbol{\rho}',\boldsymbol{r})\,\chi^*_{\Psi}(-\boldsymbol{\rho}',\boldsymbol{r}). \tag{30}$$

From this point, deconvolution and phase retrieval may proceed as described in §3*b*. It is interesting to note that the coherence function $\Gamma(-\boldsymbol{\rho}')$ is entirely separable from $H(\boldsymbol{r},\boldsymbol{\rho}')$ and is only a function of $\boldsymbol{\rho}'$. This means that it could in theory be accounted for in the deconvolution process (equation (8)). Indeed, after transforming $H(\boldsymbol{r},\boldsymbol{\rho}')$ with respect to $\boldsymbol{r}$ to obtain the final data-set $D(\boldsymbol{r}',\boldsymbol{\rho}')$, it should be straightforward to measure $\Gamma(-\boldsymbol{\rho}')$ by comparing the modulus of $D(\boldsymbol{r}',\boldsymbol{\rho}')$ measured along $\boldsymbol{\rho}'=0$ with that in the $\boldsymbol{\rho}'$ direction at various other values of $\boldsymbol{r}'$. Although it may well possess zeros (or, worse, decay rapidly to zero), this need not affect the resolution of the final reconstruction. As with the method described in §3*b* which may be used to avoid the objective aperture cut-off, it should be possible to arrange to take only small steps

MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
THE ROYAL SOCIETY A
PHILOSOPHICAL TRANSACTIONS OF

Downloaded from rsta.royalsocietypublishing.org

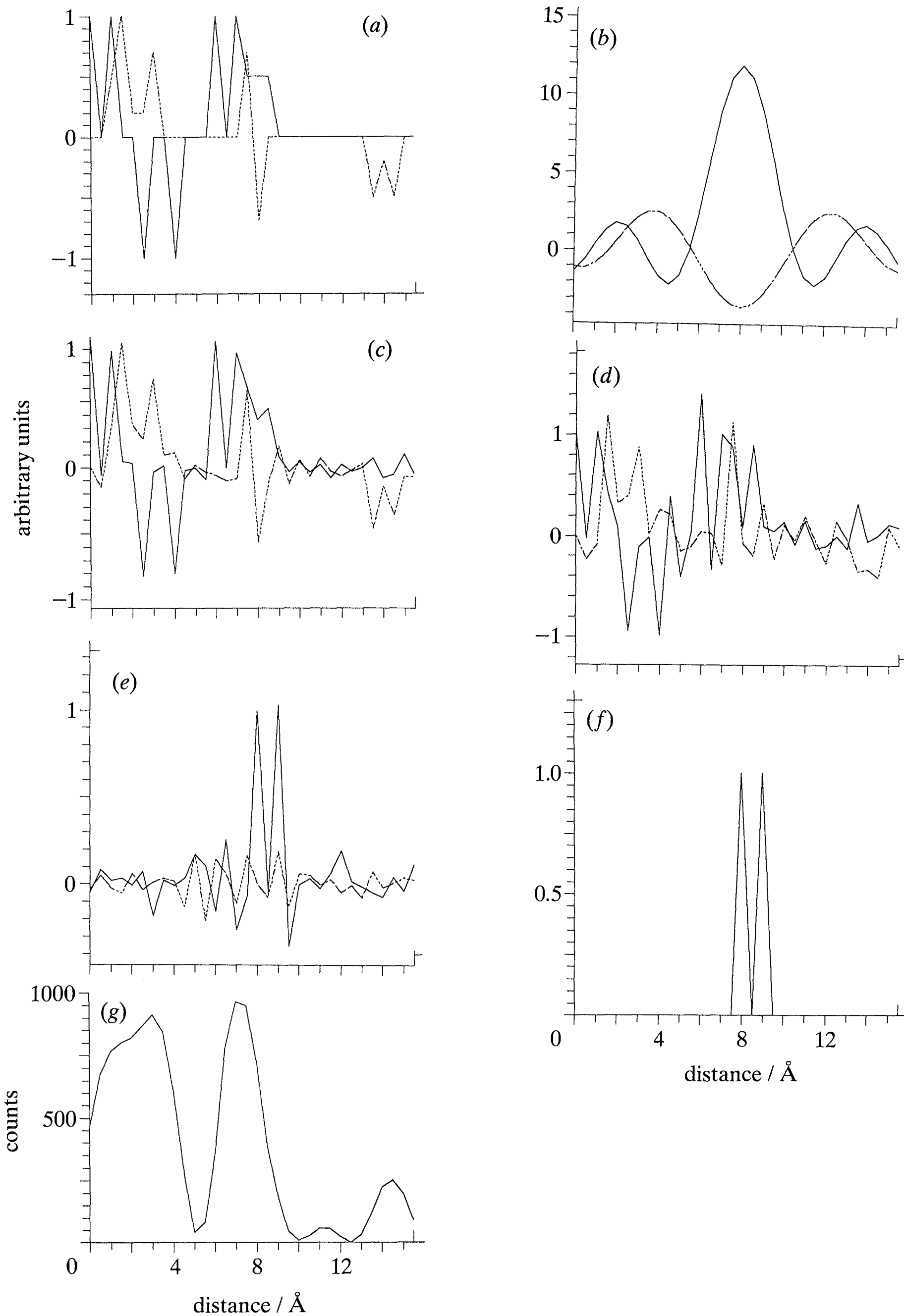


Figure 10. A model calculation of the reconstruction algorithm for a one-dimensional complex specimen function. Real parts of functions are denoted by solid lines, imaginary parts by discontinuous lines. The calculation was performed on a 64-element grid of sample points separated by 0.5 Å, of which only the central 32 elements have been displayed. (*a*) The specimen function. (*b*) The band-width limited, aberrated probe function. (*c*) The reconstruction with $n = 10000$ (see text), corresponding to a noise level in the recorded data of 1 %. Note that only a minimal fraction of $D(\boldsymbol{r}', \boldsymbol{\rho}')$ has been processed to obtain this result. (*d*) The reconstruction with $n = 100$, corresponding to 10 % noise. The main features are seen to remain intact. (*e*) The reconstruction

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

out into the $\boldsymbol{\rho}'$ direction at any given $\boldsymbol{r}'$, provided, as before, that the specimen possesses at least one reciprocal lattice point within the width of the coherence function. It is worth mentioning that we may obtain some intuitive feel for why Γ is purely a function of $\boldsymbol{\rho}'$ by considering reciprocity. A large source in STEM is equivalent to a large detector element size in the image plane of CTEM. Despite the usual limitations of the objective lens transfer function, this further reduces the final resolution of any particular image (which remember is collected as a function of $\boldsymbol{\rho}$), and hence creates a cut-off in $\boldsymbol{\rho}'$ at a value less than that dictated by the electron optics. However, like the transfer function, Γ can be fully accounted for by the Wigner devonvolution method. This is perhaps one of the most important consequences of the analysis, in view of the fact that the most limiting experimental factor in obtaining higher resolution by other means (such as holography) is the finite degree of coherence that can normally be obtained in the laboratory environment, whereas in this method it does not appear to present such a serious problem.

The effect of specimen thickness is clearly less straightforward, because the dependence of $\tilde{\chi}^*_{\Psi}(-\boldsymbol{\rho}', \boldsymbol{r}, \boldsymbol{b}')$ on $\boldsymbol{b}'$ renders the integrals in equation (26) inseparable. It should be borne in mind, however, that the value of $A(\boldsymbol{b}')A^*(\boldsymbol{b}'+\boldsymbol{\rho}')$ will be zero for values of $\boldsymbol{b}'$ beyond the objective aperture width $\boldsymbol{b}' = \boldsymbol{\rho}'_{\max}$. For moderately thin specimens, for which $\Psi(\boldsymbol{r}', z')$ will be slowly varying in the z'-direction, the complex value of $\Psi(\boldsymbol{r}', z')$ over the region of the scattered objective aperture disc will be close to the complex value measured over the Ewald sphere. The maximum departure of the aperture from the surface of the sphere is denoted $\Delta h'$ (see figure 9) and is given by $\alpha|k| \sin\theta \approx \alpha|k|\,\theta$, where α is the semi-angle of the objective aperture. It follows that the maximum thickness of specimen allowable for a notional resolution of Δr (using a similar argument to that in §4) is $T = \Delta r/\alpha$. To solve at a resolution of 0.5 Å, the maximum allowable specimen thickness is about 65 Å (compared with 13.5 Å for the strict projection approximation), which should be quite practicable to achieve experimentally. Of course, Fourier transforming the complex distribution lying over the sphere as if it were a two-dimensional function will not yield a clean image, but one reconstructed as if focused at a particular plane through the specimen. Needless to say, the only way to overcome completely the projection difficulty would be by performing many experiments at different specimen tilt angles. In this way, a whole section of $\Psi(\boldsymbol{r}', z')$ could be reconstructed in complex amplitude, though care would still have to be taken to avoid the difficulties inherent in equation (26), perhaps by employing an iterative algorithm which makes progressively better-informed guesses of $\Psi(\boldsymbol{r}', z')$, and then feeds back this information, allowing better separation of the instrument and specimen contributions.

6. The effect of noise

In this section we demonstrate that even when using a very simple (and intrinsically wasteful) scheme for retrieving phase from the final data-set $D(\boldsymbol{r}', \boldsymbol{\rho}')$, the Wigner-distribution deconvolution method is remarkably noise robust. Previous experience has shown that other phase-retrieval methods designed to process more limited data-sets (for example, iterative algorithms that process only the two-

of (f) using the same probe function and an average of only one count per detector pixel. (f) A very simple Young's slits specimen function. (g) The intensity of the conventional bright-field image for the same specimen and probe functions as in (a) and (b).

Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

dimensional Fraunhofer intensity distribution or the pair of image and diffraction planes available in a conventional electron microscope (see the review by Saxton 1980) may work well in the double-precision digital computer but may either fail catastrophically when faced with real experimental data, or must be implemented with considerable care (Chapman 1975*b*). Furthermore, many specimens damage with over-exposure to the beam, and so meaningful images must be collected with relatively poor counting statistics. Other more mundane experimental constraints, such as specimen drift due to differential thermal expansion of the specimen stage, make it highly favourable to reduce the integration time necessary for data collection, in which case the reconstruction process must be as noise robust as possible.

We have performed some illustrative calculations in a one-dimensional system. The specimen function, probe function and some sample reconstructions are shown in figure 10. Some examples of the two-dimensional data-sets such as the microdiffraction plane intensity distribution, the quantity $H(\boldsymbol{r}, \boldsymbol{\rho}')$, the amplitude of the probe Wigner distribution $\chi_a(\boldsymbol{r}, -\boldsymbol{\rho}')$, and the amplitude of $D(\boldsymbol{r}', \boldsymbol{\rho}')$ are shown in figure 11. Choosing to do these calculations in one-dimension has the advantage that each stage of the deconvolution process can be illustrated graphically. It must be remembered that when the image is two dimensional, data-sets such as $H(\boldsymbol{r}, \boldsymbol{\rho}')$ are four-dimensional. Furthermore, since it is well established that for reasons to do with the geometrical constraints of the Fourier transform, phase-retrieval methods are always much more robust in more than one dimension (see, for example, Bates & McDonnell 1986), demonstrating that the method works in the one-dimensional case can be regarded as a more stringent test of the new technique. The calculation ignores the complications of thickness effects and finite source size, but we introduce shot-noise in the microdiffraction plane as follows. The average intensity over all pixels in $|M(\boldsymbol{r}', \boldsymbol{\rho})|^2$ is scaled to unity. The particular value at each detector pixel is then multiplied by a factor n to give a measured intensity I. To this is added random gaussian noise with a standard deviation of $\sqrt{I}$, and the resulting value is then rounded to the closest integer value. Negative values of intensity, which occasionally occur at points of very low signal (because of the tail of the random-noise gaussian) are assigned to zero. n can therefore be regarded as a measure of the average number of electrons per microdiffraction pixel, and the final microdiffraction measurement as being an integral number of electrons corrupted by shot-noise statistics.

We assume here that $A(\boldsymbol{r}')$ (and hence the probe function $a(\boldsymbol{r})$) is known accurately. In practice, this could be measured to a first approximation by using standard techniques employed in high-resolution microscopy; namely by measuring the aberration constant of the lens, and by processing a through-focal series of images of a thin amorphous film to determine the level of defocus. Because of the highly redundant nature of the final data-set, there will be opportunity to refine the exact form of $\chi_a(\boldsymbol{r}, -\boldsymbol{\rho}')$ (see §3*b*), but for the time being we merely establish that the deconvolution represented in equation (9) is stable. The constant ϵ used in the Wiener filter is varied according to the level noise present until the final reconstruction is as clean as possible. Figure 11*e* shows a plot of the phase difference in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ between perfect data (no noise) and noisy data with $n = 100$ (i.e. an average noise level of about 10%). We plot a zero phase difference as white, with grey-levels proportional the modulus of the phase difference up to a maximum of π, represented as black. As one would expect, large phase differences occur in the noisy data where the amplitude of $D(\boldsymbol{r}', \boldsymbol{\rho}')$ is low (compare with figure 11*d*), but there are

Downloaded from rsta.royalsocietypublishing.org

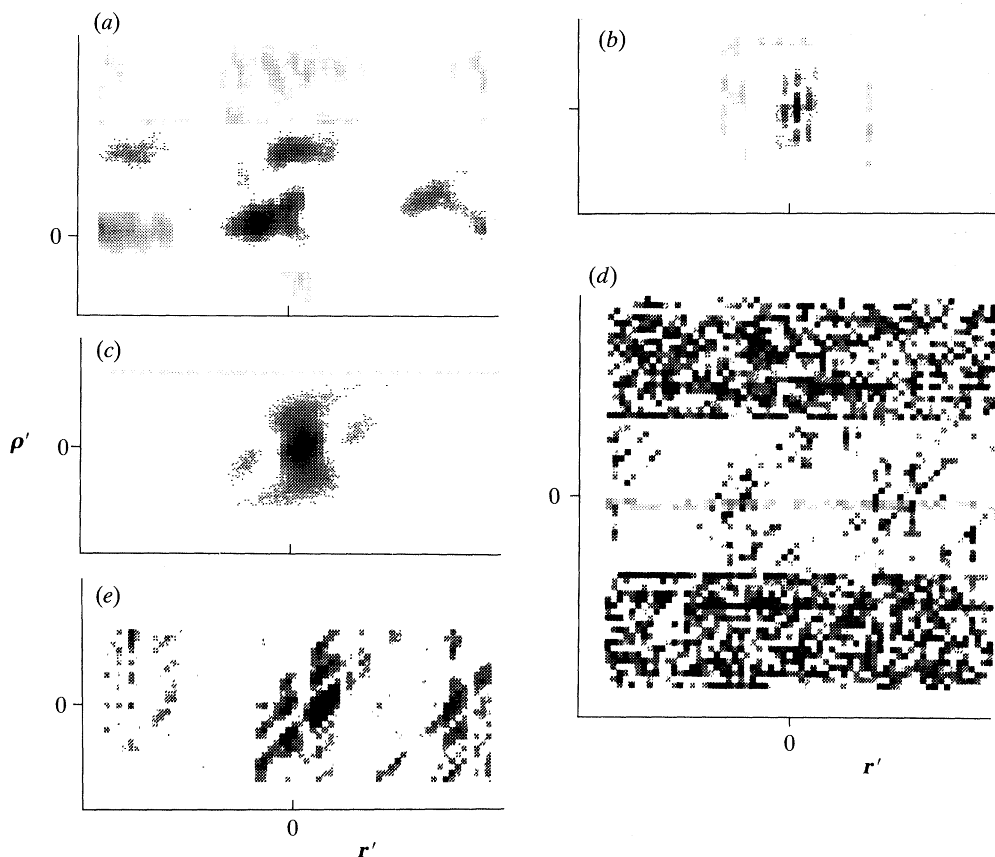


Figure 11. Some sample plots of the two-dimensional data-sets which occur during the reconstruction process. The specimen and probe were the same as those used in figure 10*a* and *b*. All the functions (apart from the recorded data in (*a*)) are complex quantities. Unless otherwise stated, the modulus of the function has been displayed. (*a*) The recorded data-set $|M(\boldsymbol{r}',\boldsymbol{\rho})|^2$. The data is asymmetric because the specimen (figure 10*a*) lies mostly at positive values in real space. (*b*) The function $H(\boldsymbol{r},\boldsymbol{\rho}')$, where the filter processing occurs. (*c*) The function $\chi_a(\boldsymbol{r}, -\boldsymbol{\rho}')$. Note the cut-off in the $\boldsymbol{\rho}$ direction. (*d*) The $D(\boldsymbol{r}',\boldsymbol{\rho}')$ data-set. It is in this plane that the phase-retrieval occurs. (*e*) Plot of the phase difference in $D(\boldsymbol{r}',\boldsymbol{\rho}')$ between perfect data and noisy data (with $n = 100$). See main text for details.

still large regions which have remained relatively uncorrupted, even with these low counting statistics. The characteristic cross-hatched form of $D(\boldsymbol{r}',\boldsymbol{\rho}')$ arises from the nature of equation (9). We derive $\Psi(\boldsymbol{r}')$ from $D(\boldsymbol{r}',\boldsymbol{\rho}')$ by using equation (10) as far as the $\boldsymbol{\rho}'_{\max}$ cut-off. To improve resolution further, we determine offset values for $\boldsymbol{w}'$ in equation (11) by seeking values of $D(\boldsymbol{r}',\boldsymbol{\rho}')$ that have maximum magnitude for any given value of $\boldsymbol{r}'$, but which lie at points which are more than half way towards $\boldsymbol{\rho}'_{\max}$. This elementary algorithm only uses a minimum number of values in $D(\boldsymbol{r}',\boldsymbol{\rho}')$ to reconstruct $\Psi(\boldsymbol{r}')$, but it does reduce the accumulation of phase error by avoiding the dark regions of figure 11, while at the same time reducing the total number of steps required to reach high-resolution information.

It has been found that the accuracy of the final reconstruction depends not only on n, as defined above, but also upon certain attributes of the specimen such as its total transparency and crystallinity. The reason for this is that the dynamic range

Downloaded from rsta.royalsocietypublishing.org

of a given microdiffraction pattern becomes larger if the central disc is largely unattenuated. Similarly, if the specimen is more ordered, scattered intensity arrives in localized regions (diffracted discs) such that pertinent phase information lying outside the discs, or lying in small regions of disc overlap, have relatively low intensity. This information then tends to be lost in the scaling process described above. For example, if we postulate a simple Young's slit experiment illustrated in figure 10*f* then it is possible to get a good reconstruction with an average of only one electron count per detector pixel (figure 10*e*). However, for more transparent specimens, where the central, unscattered disc of intensity becomes very bright, higher average count values are necessary in order not to lose the weakly diffracted super-resolution information. Experimentally, care will have to be taken to avoid the loss of low-intensity regions of the microdiffraction plane, perhaps by having a detector of variable response as a function of scattering angle. Alternatively, a more elaborate filtering and phase-retrieval algorithm could be used to process the data collected so as to avoid high-intensity regions in the microdiffraction plane. This will clearly have to be the subject of further work. However, suffice it to say that with the simple phase-assignment algorithm we use here, the reconstruction is well conditioned, especially in view of the fact that the specimen function is one dimensional and that we have only used a minimal fraction of the $D(\boldsymbol{r}', \boldsymbol{\rho}')$ data-set.

7. Conclusions

Although the microdiffraction data-set collected as a function of all probe positions is a complicated mixture of diffraction, image, holographic and ptychographic information, it is possible to deconvolve it by using a Wigner distribution to obtain the complex value of the specimen function at a resolution dictated by the largest diffraction angle which can be measured in the microdiffraction plane. Because this is much larger than the width of the objective lens transfer function of the microscope, the technique offers a practical means of overcoming the long-standing resolution problem in electron microscopy. An important difference between the Wigner deconvolution technique and conventional holography is that it can cope with partial coherence within the source because it does not require beams which have been well separated within the experimental apparatus to re-interfere with one another coherently. The degree of coherence ordinarily available in the conventional electron microscope (of the order of the width of the objective aperture) is sufficient for the reconstruction process. This is because the technique retrieves the phase of high-angle diffraction beams by interfering them with lower-order diffraction beams, thus using a 'reference' wave which is generated within the sample. Although the existence of this type of information has been known for some years, up until the present work there has been no obvious way of processing all the data available simultaneously, while at the same time accounting for the effects of the instrument function.

That the method reduces to Fourier transforms suggests that the technique may eventually be run in real time. The main practical difficulty will be dealing with such large quantities of information. With an efficient recording medium, enough intensity arrives in the microdiffraction plane for TV-rate exposure times (25–50 ms), even when the specimen is thin (see, for example, results from the film-recording method of Rodenburg (1988)). A data-set using 100×100 probe positions could therefore be recorded in about 10 minutes, which should be short enough to avoid the

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

adverse effects of specimen drift and contamination. The problem is to record, shift into memory and process the diffraction patterns from all such probe positions. However, given the rapid development of CCD technology and the current rate of reduction in the cost of raw computing power, this should not be regarded as a significant obstacle.

The method may also prove to be useful for radiations other than electrons, for which it is difficult to manufacture a good quality lens, but for which an array detector is easy to build. Indeed, the lens is entirely redundant provided some sort of aperture (corresponding to the probe $a(r)$) can be reliably moved across the specimen. An obvious candidate is X-ray microscopy. At present, the best resolution available in a scanning zone-plate X-ray microscope is of the order of about 150 nm (Kenny *et al.* 1989), which again is much larger than the associated wavelength (1–10 Å). Furthermore, experiments with X-ray holography (Howells 1984) suggest that the degree of coherence for a synchrotron source is not insignificant. Placing a CCD array in the far-field would not be experimentally difficult, and by using the Wigner deconvolution all the effects of the poor quality zone-plate could be accounted for as well as allowing full use of the high-angle diffraction data. It could be argued that since X-rays produce much less radiation damage than electrons, such a scheme may ultimately surpass the usefulness of the electron microscope, allowing biological samples to be imaged at atomic resolution. However, there will be some significant difficulties, the worst of which will probably be specimen drift (as in the early days of holography) due to the long integration times necessary for a low flux source and the fact that X-rays have a very much smaller cross section of interaction with matter compared with electrons.

J. M. R. is grateful to the Royal Society and the SERC for financial support. This collaboration was made possible by a Royal Society travel scholarship.

References

Bates, R. H. T. & McDonnell, M. J. 1986 *Image restoration and reconstruction*. Oxford: Clarendon Press.

Bates, R. H. T. & Rodenburg, J. M. 1989 Sub-ångström transmission microscopy: a Fourier transform algorithm for microdiffraction plane intensity information. *Ultramicroscopy* **31**, 303–308.

Bennett, A. H., Jupnik, H., Osterberg, H. & Richards, O. W. 1951 *Phase microscopy, principles and applications*. New York: Wiley; London: Chapman and Hall.

Binning, G., Rohrer, H., Gerber, Ch. & Weibel, E. 1982 Surface studies by scanning tunneling microscopy. *Phys. Rev. Lett.* **49**, 57–61.

Born, M. & Wolf, E. 1964 *Principles of applied optics*, 2nd edn, p. 508. Oxford: Pergamon.

Brown, L. M. 1981 Scanning transmission electron microscopy: microanalysis for the micro-electronic age. *J. Phys.* F **11**, 1–26.

Burge, R. E., Fiddy, M. A., Greenaway, A. H. & Ross, G. 1976 The phase problem. *Proc. R. Soc. Lond.* A **350**, 191–212.

Chapman, J. N. 1975*a* The application of iterative techniques to the investigation of strong phase objects in the electron microscope. I. Test calculations. *Phil. Mag.* **32**, 527–540.

Chapman, J. N. 1975*b* The application of iterative techniques to the investigation of strong phase objects in the electron microscope. II. Experimental results. *Phil. Mag.* **32**, 541–552.

Cohen, L. 1989 Time-frequency distributions – a review. *Proc. IEEE* **77**, 941–981.

Cowley, J. M. 1969 Image contrast in a transmission scanning electron microscope. *Appl. Phys. Lett.* **15**, 58–59.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Downloaded from rsta.royalsocietypublishing.org

Cowley, J. M. 1978 Electron microdiffraction. *Adv. Elect. elect. Phys.* **46**, 1–53.

Cowley, J. M. 1979 Coherent interference in convergent-beam electron diffraction and shadow imaging. *Ultramicroscopy* **4**, 435–450.

Cowley, J. M. 1980 Optical processing of diffraction information in STEM. In *Scanning electron microscopy*, vol. 1 (ed. O. Johari), pp. 61–65. Scanning Electron Microscopy, Inc., AMF O'Hare.

Cowley, J. M. 1981*a* *Diffraction physics*, 2nd edn. Amsterdam, New York and Oxford: North-Holland.

Cowley, J. M. 1981*b* Electron microdiffraction and microscopy of amorphous solids. In *Diffraction studies in non-crystalline substances* (ed. I. Hargiltan & W. J. Orville Thomas), pp. 849–891. Hungarian Academy of Sciences.

Crewe, A. V., Wall, J. & Welter, L. M. 1968 A high-resolution scanning transmission electron microscope. *J. Appl. Phys.* **39**, 5861–5868.

Fejes, P. L. 1977 Approximations for the calculations of high-resolution electron microscope images of thin films. *Acta. crystallogr.* A **33**, 109–113.

Fienup, J. R. 1987 Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. *J. opt. Soc. Am.* A **4**, 118–123.

Frank, J. 1973 The envelope of electron microscopic transfer functions for partially coherent illumination. *Optik* **38**, 519–536.

Gabor, D. 1948 A new microscopic principle. *Nature, Lond.* **161**, 777–778.

Gabor, D. 1949 Microscopy by reconstructed wave-fronts. *Proc. R. Soc. Lond.* A **197**, 454–488.

Gerchberg, R. W. 1972 Holography without fringes in the electron microscope. *Nature, Lond.* **240**, 404–406.

Gerchberg, R. W. & Saxton, W. O. 1972 A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246.

Haine, M. E. & Mulvey, T. 1952 The formation of the diffraction image with electrons in the Gabor diffraction microscope. *J. opt. Soc. Am.* **42**, 763–773.

Hanzsen, K. J. 1982 Holography in electron microscopy. *Adv. Elect. elect. Phys.* **59**, 1–77.

Hegerl, R. & Hoppe, W. 1972 Phase evaluation in generalized diffraction (ptychography). In *Proc. Fifth Congress in Electron Microscopy* (*Manchester*). Inst. Phys. Conf. Ser. vol. 14 (ed. A. M. Glauert), pp. 628–629. London and Bristol: Institute of Physics.

Hoppe, W. 1969*a* Beugung im Inhomogenen Primärstrahlwellenfeld. I. Prinzip einer Phasenessung von Elecktronenbeugungsinterferenzen. *Acta. crystallogr.* A **25**, 495–501.

Hoppe, W. 1969*b* Beugung im Inhomogenen Primärstrahlwellenfeld. III. Amplituden-und Phasenbestimmung bei imperiodischen Objekten. *Acta. crystallogr.* A **25**, 508–514.

Hoppe, W. & Hegerl, R. 1980 Three dimensional structure determination by electron microscopy (non-periodic specimens). In *Computer processing of electron microscope images* (ed. P. W. Hawkes), vol. 13, pp. 127–185 (Topics in Current Physics). Berlin: Springer.

Hoppe, W. & Strube, G. 1969 Beugung im Inhomogenen Primärstrahlwellenfeld. II. Lichtoptishe Analogieversuche zur Phasenmessung von Gitterinterferenzen. *Acta. crystallogr.* A **25**, 502–507.

Howells, M. R. 1984 Possibilities for X-ray holography using synchrotron radiation. In *X-ray microscopy* (ed. G. Schmahl & D. Rudolph), pp. 318–335. Berlin: Springer-Verlag.

Kenny, J. M., Morrison, G. R., Browne, M. T., Buckley, C. J., Burge, R. E., Cave, R. C., Charalambous, P. S., Duke, P. J., Hare, A. R., Hills, C. P. B., Michette, A. G., Ogawa, K. & Rogoyski, A. M. 1989 Evaluation of a scanning transmission X-ray microscope using undulator radiation at the SERC Daresbury Laboratory. *J. Phys.* E **22**, 234–238.

Konnert, J. & D'Antonio, P. 1986 Image reconstruction using electron microdiffraction patterns from overlapping regions. *Ultramicroscopy* **19**, 267–278.

Konnert, J., D'Antonio, P., Cowley, J. M., Higgs, A. & Ou, H.-J. 1989 Determination of atomic positions using electron nanodiffraction patterns from overlapping regions: Si[110]. *Ultramicroscopy* **30**, 371–384.

Leith, E. N. & Upatnieks, J. 1962 Reconstructed wavefronts and communication theory. *J. opt. Soc. Am.* **52**, 1123–1130.

Lichte, H. 1986 Electron holography approaching atomic resolution. *Ultramicroscopy* **20**, 293–304.

MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
THE ROYAL SOCIETY A
PHILOSOPHICAL TRANSACTIONS OF
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES
THE ROYAL SOCIETY A
PHILOSOPHICAL TRANSACTIONS OF

Downloaded from rsta.royalsocietypublishing.org

Lin, J. A. & Cowley, J. M. 1986 Reconstruction from in-line electron holograms by digital processing. *Ultramicroscopy* **19**, 179–190.

Makin, M. J. 1968 Electron displacement damage in copper and aluminium in a high voltage electron microscope. *Phil. Mag.* **13**, 637–653.

McMullan, D., Rodenburg, J. M. & Pike, W. T. 1990 Post-spectrometer instrumentation for STEM. In *Proc. XII International Congress for Electron Microscopy, Seattle*, vol. 2, pp. 104–105.

Misell, D. L. 1973 An examination of an iterative method for the solution of the phase problem in optics and electron optics. I. Test calculations. *J. Phys.* D **6**, 2200–2216.

Misell, D. L. & Greenaway, A. H. 1974 An application of the Hilbert transform in electron microscopy. I. Bright-field microscopy. *J. Phys.* D **7**, 832–855.

Mott, N. F. & Massey, H. S. W. 1933 *The theory of atomic collisions*, ch. V. Oxford: Clarendon Press.

Patterson, A. L. 1934 A Fourier series method for the determination of the components of the interatomic distances in crystals. *Phys. Rev.* **46**, 372–376.

Rodenburg, J. M. & McMullan, D. 1985 The recording of microdiffraction patterns in scanning transmission electron microscopy. *J. Phys.* E **18**, 949–953.

Rodenburg, J. M. 1988 Properties of electron microdiffraction patterns from amorphous materials. *Ultramicroscopy* **25**, 329–344.

Rodenburg, J. M. 1989 The phase problem, microdiffraction and wavelength-limited resolution – a discussion. *Ultramicroscopy* **27**, 413–422.

Saxton, W. O. 1980 Recovery of specimen information for strongly scattering objects. In *Computer processing of electron microscope images* (ed. P. W. Hawkes), vol. 13 (Topics in Current Physics), pp. 35–87. Berlin: Springer.

Scherzer, O. 1936 Uber einige Fehler von Elektronenlinsen. *Zeit. Phys.* **101**, 593–602.

Scherzer, O. 1949 The theoretical resolution limit of the electron microscope. *J. appl. Phys.* **20**, 20–29.

Shao, Z., Beck, V. & Crewe, A. V. 1988 A study of octupoles as correctors. *J. appl. Phys.* **64**, 1646–1651.

Shao, Z. & Crewe, A. V. 1988 Ideal lenses and the Scherzer theorem: a supplement. *Ultramicroscopy* **26**, 385–388.

Spence, J. H. C. 1977 Phase determination of multiply scattered beams in STEM. *Optik* **49**, 117–120.

Spence, J. H. C. & Cowley, J. M. 1978 Lattice imaging in STEM. *Optik* **50**, 129–142.

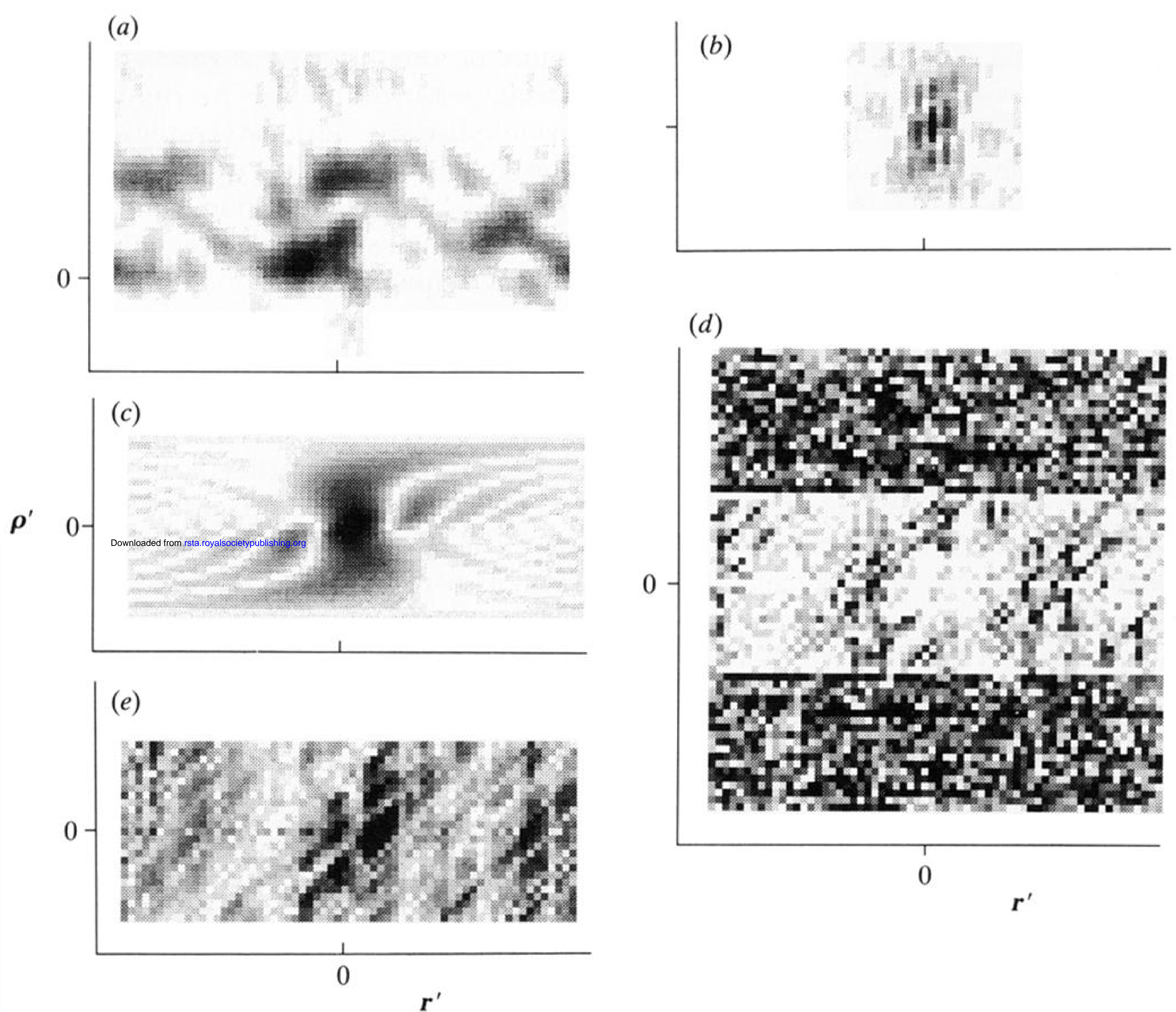
Spence, J. C. H. 1981 *Experimental high-resolution electron microscopy*. Oxford: Clarendon Press.

Tonomura, A. 1987 Applications of electron holography. *Rev. mod. Phys.* **59**, 639–669.

Volki, E. & Lichte, H. 1990 Electron holograms for subangstrom point resolution. *Ultramicroscopy* **32**, 177–180.

Woolfson, M. M. 1961 *Direct methods in crystallography*. Oxford.

Received 29 April 1991; accepted 29 November 1991



Downloaded from rsta.royalsocietypublishing.org

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Figure 11. Some sample plots of the two-dimensional data-sets which occur during the reconstruction process. The specimen and probe were the same as those used in figure 10*a* and *b*. All the functions (apart from the recorded data in (*a*)) are complex quantities. Unless otherwise stated, the modulus of the function has been displayed. (*a*) The recorded data-set $|M(\boldsymbol{r}', \boldsymbol{\rho})|^2$. The data is asymmetric because the specimen (figure 10*a*) lies mostly at positive values in real space. (*b*) The function $H(\boldsymbol{r}, \boldsymbol{\rho}')$, where the filter processing occurs. (*c*) The function $\chi_a(\boldsymbol{r}, -\boldsymbol{\rho}')$. Note the cut-off in the $\boldsymbol{\rho}$ direction. (*d*) The $D(\boldsymbol{r}', \boldsymbol{\rho}')$ data-set. It is in this plane that the phase-retrieval occurs. (*e*) Plot of the phase difference in $D(\boldsymbol{r}', \boldsymbol{\rho}')$ between perfect data and noisy data (with $n = 100$). See main text for details.