Fast deterministic approach to exit-wave reconstruction

A. J. D'Alfonso,[1] A. J. Morgan,[1] A. V. Martin,[2] H. M. Quiney,[1] and L. J. Allen[1]
[1]*School of Physics, University of Melbourne, Parkville, Victoria 3010, Australia*
[2]*Center for Free-Electron Laser Science, DESY, D-22607 Hamburg, Germany*

(Received 22 November 2011; published 13 January 2012)

We introduce a fast, dependable algorithm to solve for the exit surface wave of a specimen in coherent diffractive imaging for a set of illumination conditions that are not unduly restrictive. It is shown that a direct solution of the phase problem from a diffraction pattern is obtained efficiently and uniquely. The algorithm is deterministic and is known *a priori* to converge to the correct solution in less than a predetermined number of steps. It is based on the conjugate gradient least-squares method implemented with Fourier transforms and offers the possibility of real-time solutions. We also extend the formulation to allow for imaging of extended objects in a manner similar to ptychography.

DOI: 10.1103/PhysRevA.85.013816 PACS number(s): 42.30.Rx, 42.30.Wb

I. INTRODUCTION

The ability to retrieve the phase of a complex function from its modulus is fundamental in the diverse fields of astronomy, biology, materials science, and physics [1–5]. Considerable recent scientific activity has been directed toward the complete reconstruction of the exit surface wave of a specimen from measurements of diffracted intensity. Coherent diffractive imaging (CDI) is one such technique that seeks to restore the exit surface wave from intensity measurements obtained by coherent illumination of an object. CDI is particularly powerful because it has the capacity to reconstruct the exit surface wave from a single intensity measurement under Fraunhofer or Fresnel diffraction conditions.

Numerous strategies have been proposed to perform the restoration of phase information from measurements of intensity data, the most successful of which are elaborations [6–11] of the iterative scheme originally devised by Gerchberg and Saxton [12]. Reconstructions of the exit wave using these iterative approaches often suffer, however, from slow or erratic convergence; in some cases, the iterative sequence simply fails to converge. These approaches are nonsmooth, nonconvex, nonlinear optimization problems and much attention and effort has been directed toward ameliorating issues associated with the stagnation of the iterative sequence and determining whether the images that they generate are unique. One successful approach has been to use multiple measurements, as in ptychography, to overdetermine the problem. Such techniques exclude, however, the possibility of single-shot imaging in CDI, which has recently assumed particular significance in the development of ultrafast imaging strategies using coherent sources of x rays or electrons.

Despite the widespread use of iterative methods, the issue of soundness and decidability is often raised. Elser has remarked [10] that there currently exist "no practical algorithms for phase retrieval" because the procedures that are employed in practice provide no guarantee that a solution will be found at a computational cost that grows modestly with the size of the problem. A truly satisfactory algorithm, in the sense discussed by Elser, should be deterministic, enabling bounds to be placed on the number of steps required to obtain a solution from an arbitrary starting guess [13]. Existing iterative procedures do not fall strictly within any of the commonly accepted definitions of "an algorithm" because they all admit the possibility of stagnation; the procedure may terminate, in the sense that the output of its final step is the same as the input taken from the result of the previous iteration, yet the output need not correspond to the target solution. No bounds may be placed on the number of iterations required to obtain a solution using these procedures, which means that they need never terminate. They all require some level of human intervention to ascertain whether the results that they provide should be accepted or rejected. This intervention usually involves an empirical mixture of *a priori* information about the target image and an examination of the reproducibility of the output of the procedure from randomly chosen starting points.

The procedure proposed by Martin and Allen [15] to retrieve the exit surface wave of an scattering object from its diffraction pattern satisfies the most significant criterion for it to be regarded as a *bona fide* algorithm; determinism (i.e., it is not subject to randomness). Subject to the satisfaction of some illumination conditions that prove to be, in practice, not very restrictive, this formulation leads to the construction of an overdetermined set of linear equations, guaranteeing a solution using standard methods of computational linear algebra. This approach does not, however, satisfy a subsidiary condition identified by Elser for "practical algorithms" because the minimum number of linear equations needed to solve the problem is n_{var}, where n_{var} is the number of unknown variables, and the solution of this set of equations possesses unfavorable n_{var}^3 scaling characteristics. This is illustrated by the diffraction data shown in Fig. 1(a), formed by illuminating a gnat's wing using the circular illumination shown in Fig. 1(b). By construction, the gnat's wing is confined to the region indicated by Fig. 1(c). In a recent paper [16] the exit surface wave in this region was determined by solving an overdetermined set of 311 632 linear equations to obtain the exit surface wave for the 51 026 pixels in the region indicated in Fig. 1(c). The computational resources required for this retrieval were quite demanding, requiring approximately 12 h of computer time in single precision, using 70 GB of RAM on a 12-core AMD Opertron 2.0-GHz computer.

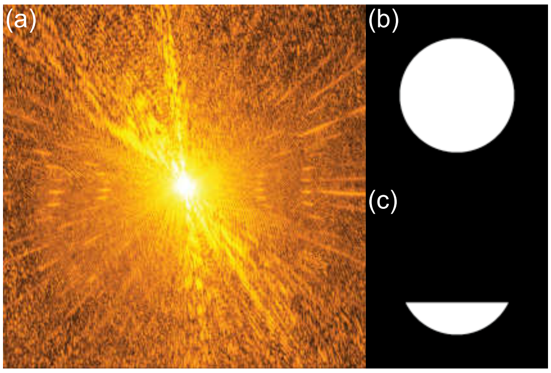

FIG. 1. (Color online) (a) The diffraction pattern formed by illumination of a gnat's wing with a laser, (b) assumed illumination intensity and phase, and (c) the object area used for the iterative linear reconstruction of a gnat's wing transmission function. The data were taken from Ref. [14] with details therein.

In the first section of this communication we present a procedure to determine the exit surface wave in CDI that may be regarded as a true "practical algorithm." The physical model of Martin and Allen [15] is adapted to exploit fully its linear character, producing a solution with a computational scaling of $O(n_{\mathrm{var}} N \ln(N))$ in time. The factor n_{var} indicates that up to n_{var} iterations may be required to solve the linear equations (in practice less). N is the number of pixels that the intensity measurement occupies. Compared to recent demonstrations of the Martin and Allen formulation in Ref. [16], this new algorithm features substantially reduced memory requirements, making its demands for computational resources comparable to those of conventional iterative CDI techniques, such as the hybrid input-output [7] or difference map [10] methods. The new algorithm converges in a number of steps that is fewer than the number of equations being solved and in a time that is reduced by over two orders of magnitude on a standard desktop computer compared with the result in Ref. [16]. The modest memory requirements also make it possible to use a consumer-grade GPU to determine the solution, achieving further efficiency gains of an order of magnitude and reducing solution times to tens of seconds. A reduction in computing time of several orders of magnitude might be achievable using a high-performance computing facility, permitting real-time reconstructions.

The second section of this communication extends this algorithm to allow for the imaging of objects which are larger than the illumination area of the beam. This eases the restrictions outlined in Ref. [15] where the object imaged must be finite and smaller than the illumination. It is shown that a knowledge of the transmission function of the object over a small region is sufficient to seed a ptychographic reconstruction of an extended object using a simple extension of the new algorithm, subject to the usual conditions relating to the extent and position of the unknown portion of the transmission function illuminated by the probe. We also show that the ptychographic reconstruction obtained by this approach exhibits robustness to statistical measurement noise because of the rapid phase variation imprinted on the illuminating wave by its transit through the specimen. In practice, this allows for reconstructions to be obtained beyond the information limit defined by the incident illumination.

II. THEORY

In this section we briefly outline the theory as discussed in Refs. [14–16]. The approach may be regarded as a generalization of Fourier holography. It does not assume a real object, as can be seen by the strong departure from centrosymmetry in Fig. 1(a), nor does it assume positivity. The autocorrelation of an object's exit surface wave can be obtained by taking the inverse Fourier transform of the measured diffraction pattern, which is a consequence of the Wiener-Khinchin theorem. Symbolically, we write the exit-wave autocorrelation as

$$f_e(\mathbf{r}) \equiv \int \psi_e(\mathbf{r}+\mathbf{r}')\psi_e^*(\mathbf{r}')d\mathbf{r}' = \mathcal{F}^{-1}[I(\mathbf{q})], \qquad (1)$$

where $I(\mathbf{q})$ is the measured diffraction pattern, $\psi_e(\mathbf{r})$ is the exit surface wave, and $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operation. This assumes that the whole diffraction pattern is measured, as shown in Fig. 1(a). If the exit surface wave is written as the sum of the illumination, ψ_{illum} (assumed known), and a term expressing the modification due to the illumination passing through the object, ψ_{obj}, then we may write

$$\psi_e(\mathbf{r}) = \psi_{\mathrm{illum}}(\mathbf{r}) + \psi_{\mathrm{obj}}(\mathbf{r}). \qquad (2)$$

The exit-wave autocorrelation can be expressed as

$$f_e(\mathbf{r}) = f_{\mathrm{illum}}(\mathbf{r}) + f_{\mathrm{cross}}(\mathbf{r}) + f_{\mathrm{obj}}(\mathbf{r}), \qquad (3)$$

where $f_{\mathrm{cross}}(\mathbf{r})$ is given by

$$f_{\mathrm{cross}}(\mathbf{r}) = \int \psi_{\mathrm{illum}}(\mathbf{r}+\mathbf{r}')\psi_{\mathrm{obj}}^*(\mathbf{r}')d\mathbf{r}' + \int \psi_{\mathrm{obj}}(\mathbf{r}+\mathbf{r}')\psi_{\mathrm{illum}}^*(\mathbf{r}')d\mathbf{r}'. \qquad (4)$$

Given that $\psi_{\mathrm{illum}}(\mathbf{r})$ is known and satisfies constraints outlined in Ref. [15], then sufficient information exists in a region $\mathcal{D}$ of the autocorrelation $f_e(\mathbf{r})$ to formulate a system of (in general overdetermined) linear equations,

$$f_{\mathrm{cross}}(\mathbf{r})|_{\mathbf{r}\in\mathcal{D}} = \mathcal{F}^{-1}[I(\mathbf{q}) - I_{\mathrm{illum}}(\mathbf{q})]|_{\mathbf{r}\in\mathcal{D}}, \qquad (5)$$

where $I_{\mathrm{illum}}(\mathbf{q})$ is the intensity of the illumination in the diffraction plane. The region $\mathcal{D}$ is shown in pale gray in Fig. 2(b) and is the region of the exit-wave autocorrelation which contains cross terms of the form given by Eq. (4) but which excludes the area to which the autocorrelation of the object wave is assumed to be confined (shown in dark gray).

This procedure leads to a set of linear equations of the form

$$\mathcal{A}\mathbf{x} = \mathbf{b}. \qquad (6)$$

Following Eq. (5), the vector $\mathbf{b}$ is constructed from the pertinent values of $f_e(\mathbf{r})$ in the region $\mathcal{D}$. The matrix $\mathcal{A}$ is constructed from ψ_{illum} and has a Toeplitz plus Hankel-like structure. The unknown values of the object wave in the region to which the object is assumed confined are contained in the vector $\mathbf{x}$. The storage requirements of the matrix $\mathcal{A}$ scale as the square of the number of unknowns in the vector $\mathbf{x}$, n_{var}. This rapidly leads to large storage requirements and to the time required to solve the equations by direct methods of linear algebra.

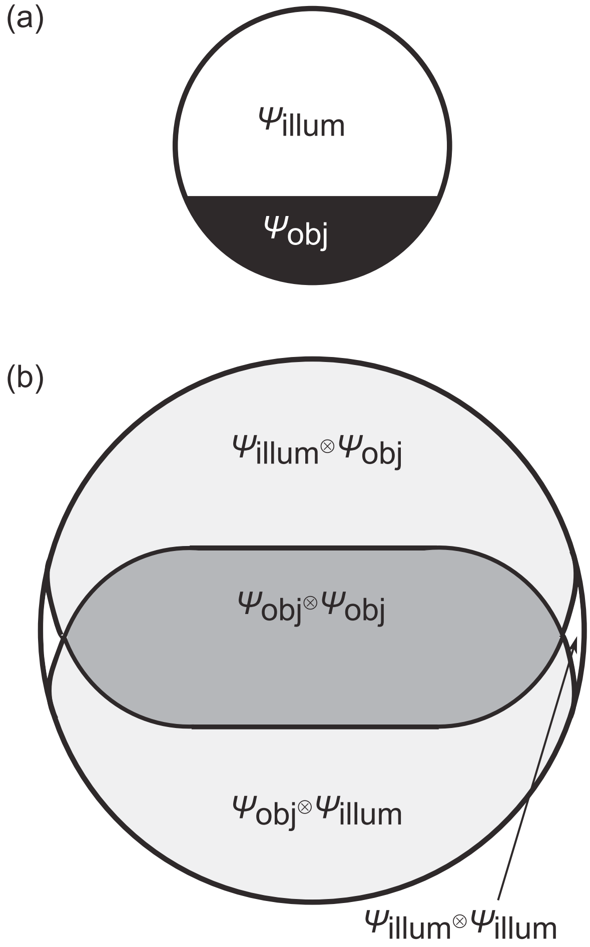
(a)
$\psi_{\rm illum}$
$\psi_{\rm obj}$
(b)
$\psi_{\rm illum}\otimes\psi_{\rm obj}$
$\psi_{\rm obj}\otimes\psi_{\rm obj}$
$\psi_{\rm obj}\otimes\psi_{\rm illum}$
$\psi_{\rm illum}\otimes\psi_{\rm illum}$

FIG. 2. Domains of the various autocorrelation contributions as pertains to Fig. 1 [and Eq. (4)]. We define the cross-correlation function as $P \otimes Q \equiv \int P(\mathbf{r}+\mathbf{r}')Q^*(\mathbf{r}')d\mathbf{r}'$. The region shaded pale gray indicates the cross-correlation region, the region shaded dark gray indicates the object wave autocorrelation region, and the autocorrelation of the illumination wave is defined over the entire region.

Here we will discuss an implementation of the conjugate gradient algorithm [17] to solve the linear system of Eq. (6), where the $\mathcal{A}$ matrix is never constructed explicitly. The matrix elements need never be stored, since we require only matrix-vector products of the form $\mathcal{A}\mathbf{x}'$, where $\mathbf{x}'$ is a trial vector. In our implementation, the storage requirements depend linearly on N.

Formally, the full solution of the conjugate gradient method requires $n_{\rm var}$ iterations. For well-conditioned problems, however, the conjugate gradient method is known to converge in a number of steps which is significantly less than $n_{\rm var}$. For the illumination conditions in Fig. 2(a), the system of linear equations is overdetermined so the $\mathcal{A}$ matrix is nonsquare. In this case we use the conjugate gradient least-squares method (CGLS) [17], which is equivalent to solving the system of linear equations

$$\mathcal{A}^T\mathcal{A}\mathbf{x} = \mathcal{A}^T\mathbf{b}. \tag{7}$$

The matrix product, $\mathcal{A}^T\mathcal{A}$ is never formed explicitly, avoiding any ill conditioning associated with the matrix multiplication. The CGLS algorithm begins with the following initialization:

$$\mathbf{x}_0 = 0,\ \ \mathbf{d}_0 = \mathbf{b},\ \ \mathbf{r}_0 = \mathcal{A}^T\mathbf{b},\ \ \mathbf{p}_0 = \mathbf{r}_0,\ \ \mathbf{t}_0 = \mathcal{A}\mathbf{p}_0. \tag{8}$$

Here the initial guess is $\mathbf{x}_0 = 0$, but any starting guess produces the same least-squares solution. We then iterate for $i = 1,2,3,\ldots n_{\rm eq}$, until a stopping criterion is satisfied; this may require significantly fewer than $n_{\rm var}$ steps. For each i, the intermediate quantities that are constructed in the CGLS algorithm are defined by

$$\begin{aligned}\alpha_i &= \|\mathbf{r}_{i-1}\|^2/\|\mathbf{t}_{i-1}\|^2, \quad \mathbf{x}_i = \mathbf{x}_{i-1} + \alpha_i\mathbf{p}_{i-1},\\ \mathbf{d}_i &= \mathbf{d}_{i-1} - \alpha_i\mathbf{t}_{i-1}, \quad \mathbf{r}_i = \mathcal{A}^T\mathbf{d}_i,\\ \beta_i &= \|\mathbf{r}_i\|^2/\|\mathbf{r}_{i-1}\|^2, \quad \mathbf{p}_i = \mathbf{r}_i + \beta_i\mathbf{p}_{i-1},\\ \mathbf{t}_i &= \mathcal{A}\mathbf{p}_i.\end{aligned} \tag{9}$$

Here, $\|\mathbf{v}\|$ denotes the Euclidean norm of a vector $\mathbf{v}$ α, and β are scalars. We define the residual at the ith iteration as $\mathbf{d}_i = \mathbf{b} - \mathcal{A}\mathbf{x}_i$, the residual error estimate as $\mathbf{r}_i = \mathcal{A}^T\mathbf{b} - \mathcal{A}\mathbf{x}_i$ and the estimate for the solution is $\mathbf{x}_i$. As mentioned previously, an advantage of conjugate gradient methods is that the algorithms only reference $\mathcal{A}$ through matrix-vector products, which in sparse matrix representations is an efficient operation. In the CGLS method as outlined in Eq. (9) there are two matrix-vector products which must be computed for each iteration. Assuming that the $\mathcal{A}$ matrix is not sparse and the full matrix must be stored to compute the matrix-vector products, the memory requirements remain the same as usual direct methods; however, computation time can be shortened compared to direct methods if a satisfactory stopping criterion has been reached.

In the context of CDI, however, it is possible to recast CGLS such that the matrix-vector products can be evaluated efficiently without actually constructing $\mathcal{A}$ because of the structure of $\mathcal{A}$. Let us, first, consider $\mathcal{A}\mathbf{p}_i = \mathbf{t}_i$. To evaluate the matrix-vector product of the $\mathcal{A}$ matrix with some arbitrary vector $\mathbf{p}$ is trivial as this is the exact problem that we set out to solve. Writing $\mathbf{p}_i$ explicitly as $\tilde{\psi}_i(\mathbf{r})$ the matrix-vector product is

$$\begin{aligned}\mathcal{A}\mathbf{p}_i \equiv t_i(\mathbf{r}) &= \psi_{\rm illum}(\mathbf{r})\otimes\tilde{\psi}_i(\mathbf{r})^* + \tilde{\psi}_i(\mathbf{r})\otimes\psi^*_{\rm illum}(\mathbf{r})\\ &= \mathcal{F}^{-1}[\Psi_{\rm illum}(\mathbf{q})\tilde{\Psi}_i^*(\mathbf{q}) + \Psi^*_{\rm illum}(\mathbf{q})\tilde{\Psi}_i(\mathbf{q})],\end{aligned} \tag{10}$$

where $\otimes$ is a correlation. As $\psi_{\rm illum}$ is known and the function $\tilde{\psi}$ is the functional form of $\mathbf{p}_i$ at the current iteration i, this quantity may be calculated efficiently—as two Fourier transforms and two vector products. Next, consider $\mathcal{A}^T\mathbf{d}_i = \mathbf{r}_i$. Computing the matrix-vector product for the matrix $\mathcal{A}^T$ with the ith residual vector $\mathbf{d}_i$ can also be done efficiently using Fourier transforms as

$$\begin{aligned}(\mathcal{A}^T\mathbf{d}_i) \equiv r(\mathbf{r}) &= \mathbf{d_i}\otimes\psi^{\rm illum} + \psi^{\rm illum}\star\mathbf{d_i},\\ &= \mathcal{F}^{-1}[\Psi_{\rm illum}(\mathbf{q})D_i^*(\mathbf{q}) + \Psi_{\rm illum}(\mathbf{q})D_i(\mathbf{q})],\end{aligned} \tag{11}$$

where $\star$ is a convolution. It is seen that to perform the matrix-vector product of the matrix $\mathcal{A}^T$ with the $i^{\rm th}$ residual vector $\mathbf{d}$ simply amounts to calculating a handful of Fourier transforms and using the convolution and correlation theorems. We denote the CGLS algorithm of Eq. (9) in conjunction with Eqs. (10)

and (11) as iterative linear retrieval using Fourier transforms (ILRUFT).

A. Preconditioning

The rate of convergence of congugate gradient methods depends on the condition number of the $\mathcal{A}$ matrix. Furthermore, the condition number indicates how robust the solution is to inconsistencies in measured data, such as measurement errors in the diffraction pattern. Preconditioning was used in Ref. [14] to stabilize a solution using experimental data. Preconditioning was not used for the reconstructions in this paper, but we outline the mathematical details here for completeness. In ILRUFT we seek to precondition the set of linear equations so the retrieval is insensitive to the effects of spurious high-frequency Fourier components. To do so, we seek to retrieve a lower resolution $\psi_{\text{obj}}(\mathbf{r})$ which could also be produced with a convolution of ψ_{obj} with a Gaussian, $p(\mathbf{r}) \star \psi_{\text{obj}}$. This is represented in matrix notation as the operation $\mathcal{P}\mathbf{x} = \mathbf{x}'$. Starting from Eq. (6) and inserting $\mathcal{P}^{-1}\mathcal{P}$ we obtain

$$\mathcal{A}\mathcal{P}^{-1}\mathcal{P}\mathbf{x} = \mathbf{b}, \quad \mathcal{A}\mathcal{P}^{-1}\mathbf{x}' = \mathbf{b}, \quad \mathcal{A}'\mathbf{x}' = \mathbf{b}, \tag{12}$$

where $\mathcal{A} \rightarrow \mathcal{A}'$. It can be shown that the preconditioned matrix vector products in ILRUFT are then

$$\mathcal{A}'\mathbf{p}_i \equiv t_i(\mathbf{r}) = \mathcal{F}^{-1}[\Psi_{\text{illum}}(\mathbf{q})H^*(\mathbf{q})\tilde{\Psi}_i^*(\mathbf{q}) + \Psi^*_{\text{illum}}(\mathbf{q})H(\mathbf{q})\tilde{\Psi}_i(\mathbf{q})] \tag{13}$$

and

$$\mathcal{A}'^T\mathbf{d}_i \equiv r_i(\mathbf{r}) = \mathcal{F}^{-1}[H^*(\mathbf{q})\Psi_{\text{illum}}(\mathbf{q})D_i^*(\mathbf{q}) + H(\mathbf{q})\Psi_{\text{illum}}(\mathbf{q})D_i(\mathbf{q})], \tag{14}$$

where $H(\mathbf{q})$ is the reciprocal of the Fourier transform of the Gaussian that is used to reduce the resolution of $\psi_{\text{obj}}(\mathbf{r})$.

B. Partial coherence

The effects of partial coherence are readily incorporated within the ILRUFT algorithm. This is achieved by formulating the procedure in terms of the object transmission function, $T(\mathbf{r})$, rather than an exit surface wave, which cannot be used to characterize the scattering from a partially coherent source. The mutual optical intensity of the illumination, $J(\mathbf{r}_1,\mathbf{r}_2)$, describes the statistical properties of the illumination in the entrance surface of the sample in a quasimonochromatic approximation, while $T^*(\mathbf{r}_1)J(\mathbf{r}_1,\mathbf{r}_2)T(\mathbf{r}_2)$ describes the exit surface properties of the optical field.

The coherent mode expansion of $J(\mathbf{r}_1,\mathbf{r}_2)$ in any plane perpendicular to the optical axis takes the form [13]

$$J(\mathbf{r}_1,\mathbf{r}_2) = \sum_k \eta_k \psi^*_{e,k}(\mathbf{r}_1)\psi_{e,k}(\mathbf{r}_2), \tag{15}$$

where $k = 1,2,\ldots$, and the real, non-negative parameter, η_k, represents the occupancy of two-dimensional mode $\psi_k(\mathbf{r})$; typically, only a few terms in the sum over k are required to describe a partially coherent field. Following the conventions established in Fig. 2(a), each of the modes in the exit surface of the object, $\psi_{e,k}(\mathbf{r})$, may be partitioned according to

$$\psi_{e,k}(\mathbf{r}) = \psi^{(1)}_{\text{illum},k}(\mathbf{r}) + \psi^{(2)}_{\text{illum},k}(\mathbf{r})T_{\text{obj}}(\mathbf{r}), \tag{16}$$

where $\psi^{(1)}_{\text{illum},k}(\mathbf{r})$ is the illumination outside the region assumed to contain the object, shown in Fig. 1(c), $\psi^{(1)}_{\text{illum},k}(\mathbf{r})$ is the illumination inside that region and $T_{\text{obj}}(\mathbf{r})$ represents the transmission function in the region assumed to contain the object. This multiplicative representation of the effect of the object on the illumination differs from that expressed in Eq. (2) but is convenient for the inclusion of partial coherence in the linear formalism.

For partially coherent illumination, the function $f_e(\mathbf{r})|_{\mathbf{r}\in\mathcal{D}}$ appearing in Eq. (5) then assumes the generalized form

$$f_e(\mathbf{r})|_{\mathbf{r}\in\mathcal{D}} = \int\left[\sum_k \eta_k \psi^{(1)}_{\text{illum},k}(\mathbf{r}+\mathbf{r}')\psi^{*(2)}_{\text{illum},k}(\mathbf{r}')\right]T^*(\mathbf{r}')d\mathbf{r} + \int\left[\sum_k \eta_k \psi^{(2)}_{\text{illum},k}(\mathbf{r}')\psi^{(1)}_{\text{illum},k}(\mathbf{r}'-\mathbf{r})\right]T(\mathbf{r}')d\mathbf{r}'. \tag{17}$$

This reduces to the formulation obtained for fully coherent illumination, Eq. (4), when restricted to $\mathbf{r} \in \mathcal{D}$, in the limit $\eta_k = \delta_{1,k}$ and $\psi_{\text{obj}}(\mathbf{r}) = \psi^{(2)}_{\text{illum},1}(\mathbf{r})T(\mathbf{r})$. The discretization of $f_e(\mathbf{r})|_{\mathbf{r}\in\mathcal{D}}$ in Eq. (17) generates a set of linear equations of the form $\bar{\mathbf{A}}\mathbf{t} = \mathbf{b}$, where $\mathbf{t}$ represents a vector of sample values of $T(\mathbf{r})$ and $\mathbf{b}$ is defined in Eq. (5), in which the measured intensities are now generated by partially coherent sources.

In practice, the matrix products $\bar{\mathbf{A}}\mathbf{t}$ and $\bar{\mathbf{A}}^T\mathbf{t}'$ are constructed by modal extensions of Eqs. (10) and (11). An input trial vector is multiplied by the mode $\psi^{(2)}_{\text{illum},k}(\mathbf{r})$ to form $\tilde{\Psi}_{k,i}(\mathbf{r})$ [Eq. (10)] or $D_{k,i}(\mathbf{r})$ [Eq. (11)]. The required matrix-vector products $\bar{\mathbf{A}}\mathbf{t}$ and $\bar{\mathbf{A}}^T\mathbf{t}'$ are formed from the weighted sum of terms of the form Eqs. (10) and (11), respectively, with weighting coefficients η_k. The increase in cost, relative to the fully coherent case, scales linearly with the number of terms that are included in the modal expansion. Typically, this factor is less than 10 and depends only on the coherence properties of the source rather than the dimensionality of the problem. The favorable scaling characteristics of the ILRUFT algorithm are, therefore, preserved within a partially coherent formulation.

III. EXPERIMENTAL TEST CASE

In this section we demonstrate and evaluate the ILRUFT algorithm using experimental data. The experimental data is the gnat's wing data in Ref. [14]. The diffraction pattern, illumination, and object area used in ILRUFT are shown in Fig. 1. At high resolution (1200 × 1200 pixels) and using single precision this data set requires 70 Gb of memory to construct and store the $\mathcal{A}$ matrix explicitly. To solve the problem using QR decomposition required approximately 12 h of computing time, more if SVD is used. Using ILRUFT at double precision the program requires less than 0.1 Gb of memory and each iteration takes approximately 0.25 s.

The intensity and phase retrieved from the ILRUFT algorithm, as a function of iteration, are shown in Fig. 3. Before display the single regularization step described in Ref. [16] was implemented but the unregularized result was carried forward in the ILRUFT procedure. The regularization procedure is simply the propagation of the computed exit surface wave to the detector plane, the replacement of the modulus of this

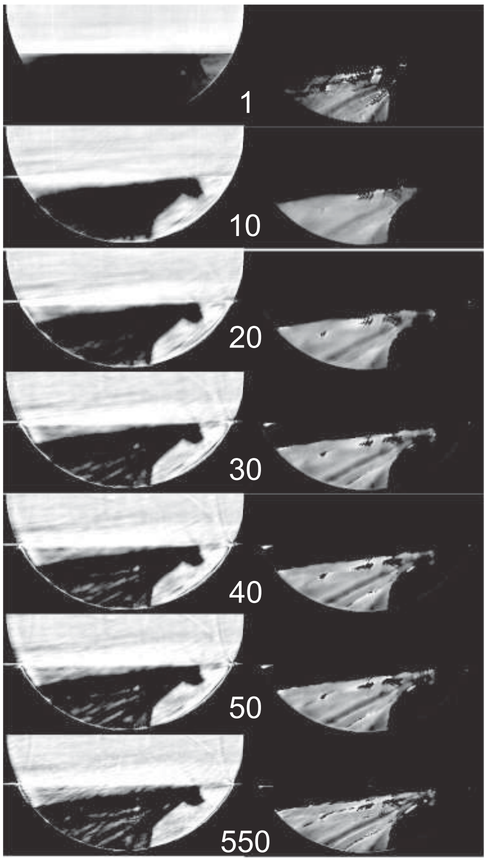

FIG. 3. The retrieved intensity (left) and corresponding phase (right) of the exit surface wave from a gnat's wing as a function of ILRUFT iterate. The iterate number is inlaid on the figure. Before display regularization is performed but the unregularized result was carried forward in the ILRUFT procedure.

wave using the measured intensity in the diffraction plane, and then the propagation back to the exit surface. It is a striking feature of the linear method that an excellent estimate of the solution is found in a handful of iterations; further iterations effectively fit measurement noise and systematic errors. An example of a systematic error that may be prominent in the retrieved intensity is the misalignment of the illumination, which is discussed in Ref. [16]. Figure 4 shows the residual produced by ILRUFT (without regularization) as a function of iterate number. The residual decays quite rapidly and asymptotes to the value 3.3×10^{-3} after a handful of iterations. Examining the regularized result, one could justifiably claim to have obtained the full deterministic solution after less than 200 iterations. This implementation of ILRUFT is based on the CGLS method and, as such, brings with it a well-established set of analytic tools regarding rates of convergence, solution stability, and matrix preconditioning. These methods are typically not available with conventional CDI techniques because of their intrinsic nonconvex character. Furthermore, the efficient evaluation of the matrix-vector product using Fourier methods for the overdetermined problem can be applied to more sophisticated iterative linear equation solvers. This is not to suggest, however, that CGLS is poor or that significant improvements in performance are required; on the contrary, ILRUFT provides an excellent solution in under a minute of CPU time. Implementing ILRUFT using state-of-the-art linear equation solvers and massively parallel graphics chips may, however, reduce the wall-clock solution time to tenths of a second, making possible real-time deterministic phase retrieval.

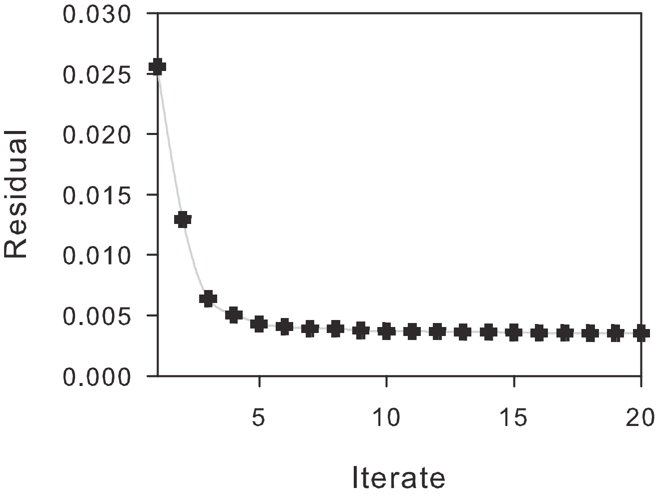

FIG. 4. The residual as a function of ILRUFT iterate for the experimental gnat's wing data. The residual was calculated without regularization.

IV. PTYCHOGRAPHIC EXTENSION

As has been discussed, the primary advantage of ILRUFT is its deterministic solution of a well-defined least-squares problem, as well as its speed and computational efficiency. However, the linear retrieval as introduced in Ref. [15] requires that certain experimental and illumination conditions are met in order to perform the inversion. The ILRUFT algorithm divides the exit surface wave in the probe region into two parts. The first part contains the scattering object, which must satisfy certain conditions about its spatial extent relative to the dimensions of the probe and its position within the probe. The second part is assumed, following Ref. [15], to be defined by the incident illumination, which passes through the probe region with unit transmission. While our previous studies have typically assumed that this illumination is uniform and structureless in both amplitude and phase, this configuration is not essential. Consequently, the exit surface wave in the second part could be the result of a highly structured incident illumination or, alternatively, of the modification of a plane wave illumination by a highly structured transmission function. The only requirements for the solution of the transmission function in the first part of the probe region are that the exit surface wave be well characterized in the second part and that the illumination be well characterized over the entire probe region. Ptychographic reconstructions of extended regions may, as a consequence, be achieved if the transmission function is known over a region that is sufficient to seed an ILRUFT solution in a region over which it is unknown, subject to the usual conditions relating to the extent and position of the unknown part in the probe region. In this section we present a modification of the linear method using ptychographic principles to allow the imaging of

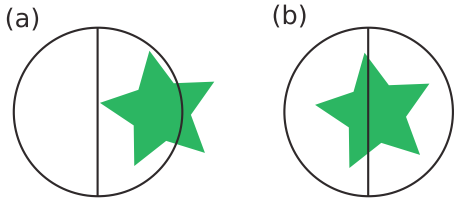

FIG. 5. (Color online) Geometry of the experiment showing the different probe positions at which each diffraction pattern measurement is made.

extended objects. We label this ptychographic extension of the ILRUFT algorithm PILRUFT. Consider the object shown in Fig. 5. It is possible to solve for the exit surface wave from the autocorrelation of the exit wave using an illumination as shown in Fig. 5(a). Conversely, for centrosymmetric illumination, it is not possible to solve for the exit wave for the geometry shown in Fig. 5(b) as this violates the conditions outlined in Ref. [15]. Suppose, however, that the probe position indicated in Fig. 5(a) is solved first and that the object exit surface wave is known *in that region*. It is then also possible to solve for Fig. 5(b) using a modified formalism derived from that described in Sec. II. Starting from Eq. (2) the exit surface wave may be written as

$$\psi_e(\mathbf{r}) = \psi_{\text{illum}}(\mathbf{r}) + \psi_{\text{ko}}(\mathbf{r}) + \psi_{\text{uo}}(\mathbf{r}), \tag{18}$$

where $\psi_{\text{ko}}(\mathbf{r})$ is the known portion of the object wave obtained from the probe position indicated in Fig. 5(a), translated appropriately, and $\psi_{\text{uo}}(\mathbf{r})$ is the unknown portion of the object wave. The corresponding autocorrelation can be written as

$$\begin{aligned} f_e(\mathbf{r}) = {} & f_{\text{illum}}(\mathbf{r}) + f_{\text{cross}}(\mathbf{r}) + f_{\text{ko}}(\mathbf{r}) + f_{\text{uo}}(\mathbf{r}) \\ & + f_{\text{illum,ko}}(\mathbf{r}) + f_{\text{ko,illum}}(\mathbf{r}), \end{aligned} \tag{19}$$

where f_{uo} indicates the autocorrelation of the unknown object wave, f_{ko} is the autocorrelation of the known object wave, and $f_{\text{illum,ko}}$ and $f_{\text{ko,illum}}$ are the cross-correlations of the known object and illumination. The sum of cross-correlations, $f_{\text{cross}}(\mathbf{r})$, is now given by

$$\begin{aligned} f_{\text{cross}}(\mathbf{r}) = {} & \int \psi_{\text{illum}}(\mathbf{r}+\mathbf{r}')\psi^*_{\text{uo}}(\mathbf{r}')d\mathbf{r}' \\ & + \int \psi_{\text{uo}}(\mathbf{r}+\mathbf{r}')\psi^*_{\text{illum}}(\mathbf{r}')d\mathbf{r}' \\ & + \int \psi_{\text{ko}}(\mathbf{r}+\mathbf{r}')\psi^*_{\text{uo}}(\mathbf{r}')d\mathbf{r}' \\ & + \int \psi_{\text{uo}}(\mathbf{r}+\mathbf{r}')\psi^*_{\text{ko}}(\mathbf{r}')d\mathbf{r}'. \end{aligned} \tag{20}$$

As a consequence, Eq. (20) can be recast as a set of linear equations, as was achieved previously for the probe position indicated in Fig. 5(a). The $\mathcal{A}$ matrix is then constructed; now, however, from the sum of the known illumination and the known portion of the previously calculated exit surface wave. Conceptually, this process may be regarded most simply as involving the incorporation of the previously retrieved exit surface wave as a part of the illumination that interferes with the unknown exit surface wave when propagated to the far field. The extension to multiple probe positions is obvious and allows for the imaging of arbitrary sized objects.

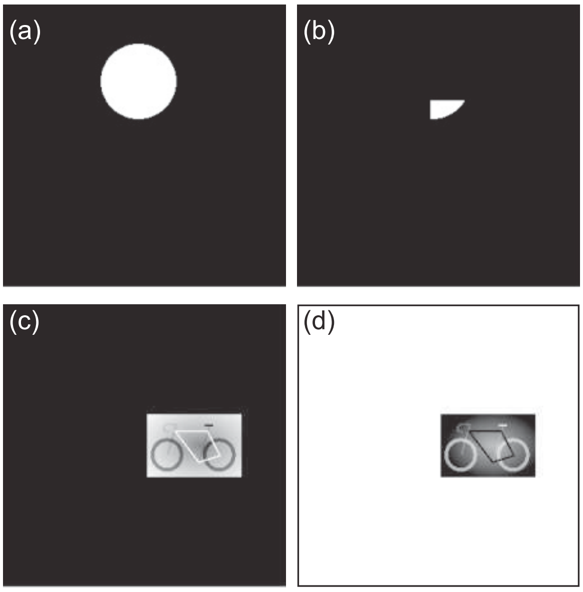

FIG. 6. Numerical inputs to the PILRUFT algorithm, (a) illumination amplitude and phase, (b) assumed object area, (c) object intensity, and (d) object phase.

The assumption that the action of the illumination on the object is multiplicative is critical, because it enables the object to be freely translated within the beam. Furthermore, we also assume that the probe step size is chosen to be sufficiently small that the number of equations that need to be solved in the augmented system is greater than the number of variables. Supposing the above conditions are satisfied and that the first probe position either satisfies the conditions of Martin and Allen [15] or was obtained by other means (such as by an independent iterative phase retrieval of a finite segment of the object), the PILFUFT procedure then is well defined for any sized object.

As a demonstration of PILRUFT, Fig. 6 shows the inputs used for a PILRUFT reconstruction. The numerical grid contained 300 × 300 pixels. The illumination had a diameter of 80 pixels and the object was confined to a rectangle of side length 100 × 66 pixels. The reconstruction used 15 horizontal probe positions and 10 vertical probe positions. During the reconstruction, the probe was raster scanned horizontally, 7 pixels at a time. On the completion of the horizontal scan the probe was then translated vertically 7 pixels and a new horizontal scan was performed. At each probe position the ILRUFT algorithm was allowed to run until either the residual changed by less than 1 part in 10^7 or the residual was smaller than 10^{-19}. At each step of the reconstruction, the object wave in the area of the illumination defined by Fig. 6(b) was obtained. This was subsequently averaged with previous overlapping calculations, each of which were also averaged. The full exit wave was then regularized as described in Ref. [16]. Shown in Fig. 7 is a PILRUFT reconstruction for various probe positions using perfect inputs, including no noise or measurement errors. Clearly evident is a retrieved object wave devoid of numerical artifacts which has been obtained at the precision specified for each ILRUFT iteration.

The accuracy of the linear reconstruction that is actually achieved depends critically on the quality of the

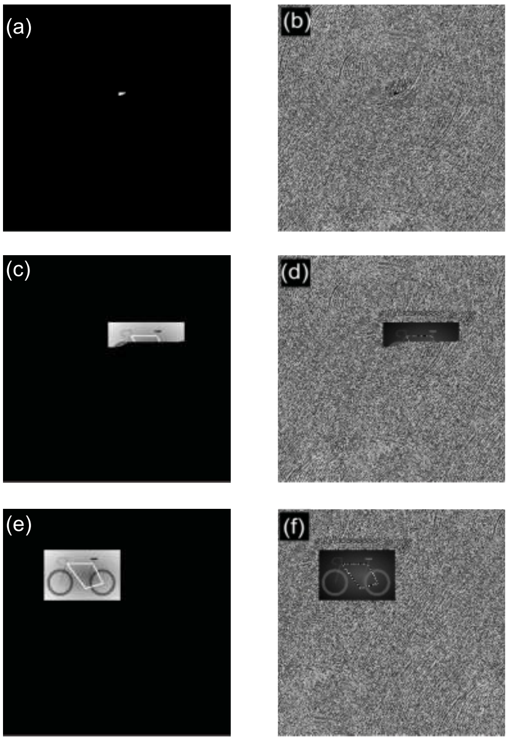

FIG. 7. Results from the full PILRUFT reconstruction with scanning in two dimensions. Images (a), (c), and (e) are the intensity and (b), (d), and (f) are the phase of the object transmission function for probe positions $(x,y) = (1,1),(5,7)$, and $(15,10)$.

characterization of the illumination. For the augmented system of linear equations, where some region of the object wave that has previously been solved for is now used to redefine the illumination, this dependency apparently places limits on the robustness of the PILRUFT algorithm. Inconsistencies in the data, such as measurement noise or inadequate convergence for the ILRUFT reconstruction, have the potential to impact negatively on the PILRUFT reconstruction, because errors propagate to subsequent probe positions. To investigate error propagation we incorporated measurement noise in the calculated diffraction pattern for each probe position by including statistical errors of 0.5 and 1.0% on the brightest pixel. The results of the PILRUFT reconstruction of the transmission function with these noise levels are shown in Fig. 8 and appear surprisingly robust. For a noise level corresponding to 0.5% [Figs. 8(a)–8(h)] the object transmission function displays some artifacts but, nonetheless, it provides a reasonable reconstruction of both the amplitude and phase. Interestingly, doubling the noise level to 1.0% [Figs. 8(i)–8(q)] still allows the PILRUFT reconstruction to produce a qualitatively accurate reconstruction of the object transmission function. A significantly worse reconstruction was obtained (results not shown) for a single probe position which used an expanded beam to illuminate the object.

A. Robustness

The excellent phase reconstruction for the noisy simulation shown in Fig. 8(q) demonstrates a striking robustness of the PILRUFT algorithm to statistical measurement noise. This robustness is most easily attributed to the overdetermined nature of the ptychographic problem. In the reconstruction for each probe position where overlap exists between previously reconstructed parts of the transmission function and the current reconstruction *in the region defined to be "unknown,"* the data sets for each transmission function were averaged [see Fig. 9]. This, in effect, averages out the uncorrelated noise that is present on the reconstructed transmission function for each probe position. The characterization of the exit wave is then improved, which is then used to redefine the illumination for subsequent probe positions. More importantly, however, the rapid phase variation present in the exit surface wave provides strong interference with the unknown portion of the exit wave. This rapid variation in the phase provides a more

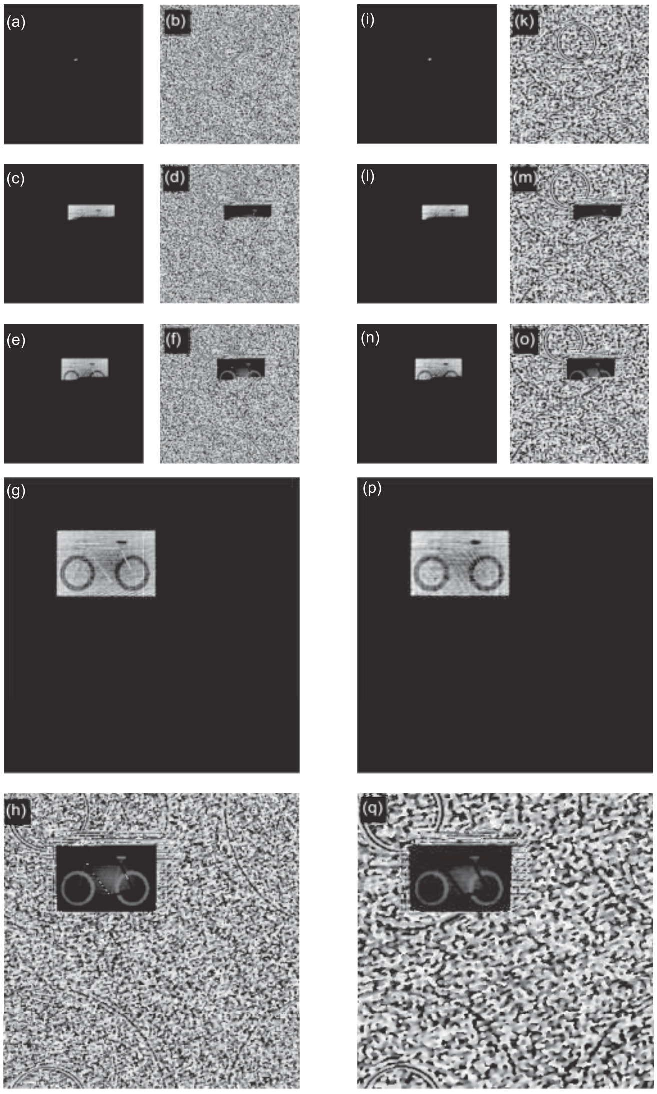

FIG. 8. Results from the full PILRUFT simulation with scanning in two dimensions. Results to the left [(a)–(h)] are for the 0.5% noise levels, whereas those to the right [(i)–(q)] correspond to 1.0% noise levels. The intensity (black background) and corresponding phase (speckled background) of the computed exit wave are shown as a function of probe position. Panels (g) and (h) show the full transmission function for 0.5% noise and panels (p) and (q) show the full transmission function for 1.0% noise. In (h) the bright speckle on the bicycle is a result of phase wrapping.

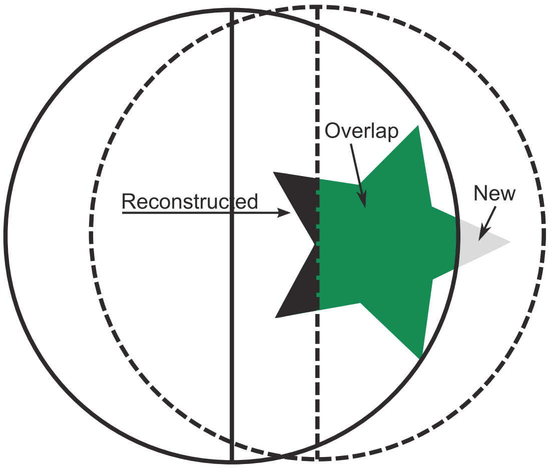

FIG. 9. (Color online) Schematic of the overlap region describing the averaging used in the reconstruction. Here, it is assumed that we solve for the entire area defined by the right-hand side semicircle of the illumination. The initial probe position reconstructs the black region. The green (or dark gray in grayscale) region indicates the overlap between the first and second probe positions and the light gray region indicates new information about the specimen.

robust reconstruction, aiding convergence of conventional iterative techniques [18,19], including in cases in which statistical measurement noise is present in the data [20]. In the present case, the rapid phase variation coupled with the overdetermined solution within the overlapping region improves the stability and the robustness of the PILRUFT reconstruction.

B. Resolution limits

The resolution limits of holographic techniques critically depends on the Fourier components contained in the illumination. In the present case, with circular illumination, this is defined by the sharpness of the circle's edge. For focused coherent illumination, like that found in scanning transmission electron microscopy, this would be the aperture cutoff used to form the probe. Using regularization incorporates nonlinear data within the reconstruction. This, in effect, improves the resolution of the reconstructed object beyond that allowed by holography alone. Using PILRUFT includes this nonlinear data in a linear fashion. The augmented illumination, which contains the characterized illumination and part of the transmission function, is used to interfere with the unknown portion of the exit wave and contains Fourier components of higher spatial frequency than those contained by the illumination alone. The result of further ILRUFT reconstructions are, without subsequent regularizations, at a higher resolution than would be obtained if the illumination was expanded to cover the entire specimen.

V. CONCLUSION

We have presented a deterministic phase retrieval algorithm that scales modestly with the size of the problem provided that certain illumination conditions are satisfied (which in practice are not very restrictive). The approach here ensures sufficient linear (in addition to nonlinear) information is available in the inverse Fourier transform of the diffraction intensity to solve the phase problem. In that sense it is simply a special case of the most general CDI phase problem. ILRUFT is the only algorithm that can claim to be a CDI algorithm in the strictest sense; it is both deterministic and requires resources that scale favorably with the size of the problem. Furthermore, ILRUFT is amenable to all forms of illumination, provided that the Fresnel imaging geometry is adopted for unstructured forms of illumination. Problems of stagnation that plague conventional techniques are not an issue in ILRUFT. In cases for which stagnation is not a problem, ILRUFT remains computationally competitive and obtains the solution in the same number or less iterations as conventional techniques. It has also been demonstrated that ILRUFT is numerically stable and practical when used with noisy experimental data.

We have shown that ILRUFT can be modified to allow for the imaging of extended objects in a pytchographic extension we have designated PILRUFT. This modification improves the robustness of the reconstruction with respect to unavoidable experimental issues such as measurement noise. Finally, unlike conventional techniques which provide a least-squares solution for the measured intensity problem, ILRUFT and PILRUFT obtain the least-squares solution for the unknown object at the numerical precision of the machine (in the absence of noise). The deterministic nature of PILRUFT ensures that it can also be used to check the uniqueness of a conventional ptychographic reconstruction.

ACKNOWLEDGMENTS

This research was supported under Australian Research Council's Discovery Projects funding scheme (project number DP1096025).

[1] J. R. Fienup, J. C. Marron, T. J. Schulz, and J. H. Seldin, Appl. Opt. **32**, 1747 (1993).

[2] J. Miao, K. O. Hodgson, T. Ishikawa, C. A. Larabell, M. A. LeGros, and Y. Nishino, Proc. Natl. Acad. Sci. USA **100**, 110 (2003).

[3] D. Shapiro, P. Thibault, T. Beetz, V. Elser, M. Howells, C. Jacobsen, J. Kirz, E. Lima, H. Miao, A. M. Neiman *et al.*, Proc. Natl. Acad. Sci. USA **102**, 15343 (2005).

[4] H. N. Chapman, P. Fromme, A. Barty, T. A. White, R. A. Kirian, A. Aquila, M. S. Hunter, J. Schulz, D. P. DePonte, U. Weierstall *et al.*, Nature **470**, 73 (2011).

[5] M. M. Seibert, T. Ekeberg, F. R. N. C. Maia, M. Svenda, J. Andreasson, O. Jonsson, D. Odic, B. Iwan, A. Rocker, D. Westphal *et al.*, Nature **470**, 78 (2011).

[6] J. R. Fienup, Opt. Lett. **3**, 27 (1978).

[7] J. R. Fienup, Appl. Opt. **21**, 2758 (1982).

[8] J. R. Fienup and C. C. Wackerman, J. Opt. Soc. Am. A **3**, 1897 (1986).

[9] J. R. Fienup, J. Opt. Soc. Am. A **4**, 118 (1987).

[10] V. Elser, J. Opt. Soc. Am. A **20**, 40 (2003).

[11] D. R. Luke, Inverse Probl. **21**, 37 (2005).

[12] R. W. Gerchberg and W. O. Saxton, Optik **35**, 237 (1972).

[13] H. M. Quiney, J. Mod. Opt. **57**, 1109 (2010).

[14] A. V. Martin, A. I. Bishop, D. M. Paganin, and L. J. Allen, Ultramicroscopy **111**, 777 (2011).

[15] A. V. Martin and L. J. Allen, Opt. Commun. **281**, 5114 (2008).

[16] A. J. Morgan, A. J. D'Alfonso, A. V. Martin, A. I. Bishop, H. M. Quiney, and L. J. Allen, Phys. Rev. B **84**, 144122(6) (2011).

[17] M. R. Hestenes and E. Stiefel, J. Res. Natl. Bur. Stand. **49**, 409 (1952).

[18] H. M. Quiney, K. A. Nugent, and A. G. Peele, Opt. Lett. **30**, 1638 (2005).

[19] G. J. Williams, H. M. Quiney, B. B. Dhal, C. Q. Tran, K. A. Nugent, A. G. Peele, D. Paterson, and M. D. de Jonge, Phys. Rev. Lett. **97**, 025506 (2006).

[20] Y. J. Liu, B. Chen, E. R. Li, J. Y. Wang, A. Marcelli, S. W. Wilkins, H. Ming, Y. C. Tian, K. A. Nugent, P. P. Zhu *et al.*, Phys. Rev. A **78**, 023817 (2008).