Non-iterative phase retrieval by phase modulation through a single parameter

Konijnenberg, Sander; Coene, Wim; Urbach, Paul

DOI
10.1016/j.ultramic.2016.12.017

Publication date
2016

Document Version
Final published version

Published in
Ultramicroscopy

Ultramicroscopy 174 (2017) 70–78

Contents lists available at ScienceDirect

Ultramicroscopy

journal homepage: www.elsevier.com/locate/ultramic

Non-iterative phase retrieval by phase modulation through a single parameter

A.P. Konijnenberg[a,*], W.M.J. Coene[a,b], H.P. Urbach[a]

[a] Optics Research Group, Delft University of Technology, Delft 2628 CH, The Netherlands
[b] ASML Netherlands B.V., De Run 6501, 5504 DR Veldhoven, The Netherlands

A B S T R A C T

We report on a novel non-iterative phase retrieval method with which the complex-valued transmission function of an object can be retrieved with a non-iterative computation, with a limited number of intensity measurements. The measurements are taken in either real space or Fourier space, and for each measurement the phase in its dual space is modulated according to a single optical parameter. The requirement found for the phase modulation function is a general one, which therefore allows for plenty of customization in this method. It is shown that quantitative Zernike phase contrast imaging is one special case of this general method. With simulations we investigate the sampling requirements for a microscopy setup and for a Coherent Diffraction Imaging (CDI) setup.

1. Introduction

There are many applications where one wants to find a complex-valued function $f(x)$, but only its modulus $|f(x)|$ can be measured directly. In the context of Coherent Diffraction Imaging (CDI) this function may represent the transmission function of a sample, but there are many other applications for phase retrieval as well (e.g. quantum state tomography [1–4]). To find the function $f(x)$ itself, one must therefore find a method to retrieve the phase.

In particular, there are phase retrieval problems that involve either measurements or some kind of constraints on a Fourier transform pair, given by $f(x)$ and its transform $\widetilde{f}(x') = \mathcal{F}\{f\}(x')$. An example of such a case is found in CDI. In this case we have a two-dimensional object, with a complex-valued transmission function $O(\mathbf{x})$. Here, $\mathbf{x}$ is a 2D position vector. If we illuminate the object with a plane wave we can measure the intensity of the diffraction pattern in the far field $I(\mathbf{x}') = |\widetilde{O}(\mathbf{x}')|^2$, where $\widetilde{O}$ denotes the Fourier transform of O, and $\mathbf{x}'$ is a 2D vector in Fourier space. Suppose, as in the original proposal by Gerchberg and Saxton [5], that we can only measure the intensity $I(\mathbf{x}')$ directly, and of the function $O(\mathbf{x})$ we only know its support (i.e. our object is an isolated object, of which we know its finite size). In other words, we have an amplitude constraint in Fourier space, and a support constraint in the object space. With projective algorithms such as the Error Reduction algorithm [5] or the Hybrid Input–Output algorithm [6], we alternatively apply the amplitude constraint and the support constraint in the two dual spaces, and that way we can try to reconstruct $O(\mathbf{x})$ and $\widetilde{O}(\mathbf{x}')$. However, these algorithms are known to not always converge to the correct solution. An alternative approach is ptychography, for which algorithms have been developed such as the ptychographic iterative engine (PIE) [7]. In ptychography, an illumination function $P(\mathbf{x})$ is used to illuminate different parts of an object $O(\mathbf{x})$. That is, we shift the illumination function by some vector $\mathbf{X}_j$, and for each $\mathbf{X}_j$ we measure intensity $I_j(\mathbf{x}') = |\mathcal{F}\{O(\mathbf{x})P(\mathbf{x}-\mathbf{X}_j)\}(\mathbf{x}')|^2$. By having $P(\mathbf{x}-\mathbf{X}_j)$ overlap for different $\mathbf{X}_j$, there is redundancy in the intensity measurements $I_j(\mathbf{x}')$, which is used as an extra constraint in the reconstruction, which makes the algorithm more robust. The PIE algorithm has been extended to ePIE [8], and it has been applied to quantum tomography [4]. However, the algorithm is still a black box in the sense that there are no known guarantees for convergence to the correct solution.

The algorithms mentioned so far are all iterative methods. There are also non-iterative methods to retrieve the phase from Fourier pairs. An example of such a method is Zernike phase contrast microscopy [9]. If we have a 2D phase object $O(\mathbf{x}) = e^{i\varphi(\mathbf{x})}$, we can Fourier transform it, shift the phase of the 0th diffraction order by $\pi/2$, and apply an inverse Fourier transform. We then find that the phase information has been converted to amplitude information which can be measured directly. However, the assumption has to be made that the object is a pure phase object, and that the variation of the phase is small (i.e. the weak-phase approximation should hold). A method in which these assumptions do not have to be made is quantitative Zernike phase contrast imaging [10]. In this method, we have an arbitrary 2D complex-valued object

* Corresponding author.
E-mail address: a.p.konijnenberg@tudelft.nl (A.P. Konijnenberg).

http://dx.doi.org/10.1016/j.ultramic.2016.12.017
Received 10 October 2016; Received in revised form 23 December 2016; Accepted 24 December 2016
Available online 26 December 2016

$O(\mathbf{x})$, and we shift the phase of the 0th diffraction order of its Fourier transform $\widetilde{O}(\mathbf{x}')$ by $A_j \in [0, 2\pi)$. We then apply an inverse Fourier transform, and measure the intensity $I_j(\mathbf{x})$. By taking three different measurements for different A_j, the object $O(\mathbf{x})$ can be calculated directly. However, this approach would make it desirable that $|\widetilde{O}(\mathbf{0})|$ is sufficiently large, because otherwise the variations in $I_j(\mathbf{x})$ are too small, which makes the method very sensitive to noise.

A non-iterative phase retrieval method that in a way resembles quantitative Zernike phase contrast imaging is Fourier transform holography [11]. Whereas in the quantitative Zernike phase contrast method a perturbation (i.e. a phase-shifted pixel) is introduced inside the support of the field, in Fourier transform holography a perturbation (i.e. a point source that is coherent with the incident field) is introduced sufficiently far away from the support of the field. This way the autocorrelation of the field (which can be found by inverse Fourier transforming the intensity of the Fourier transform of the field) contains information that is proportional to the original field. The main advantage of this method is that only one intensity measurement is needed. Similar methods that use holography-related techniques with an extended reference are given in [12,13].

Another non-iterative method is the focus-variation method [14,15], for which substantial progress was made during the 1996 Brite–Euram project [16–20]. In this method, we have a 2D object $O(\mathbf{x})$, and we take intensity measurements in different defocus planes $I_j(\mathbf{x}') = |\mathcal{F}\{O(\mathbf{x})e^{iA_j|\mathbf{x}|^2}\}(\mathbf{x}')|^2$. With these intensity measurements we can directly calculate $O(\mathbf{x})$, but only in the approximation that $|\widetilde{O}(\mathbf{0})|$ is sufficiently large. If the distance between two measurement planes is sufficiently small, the Transport of Intensity Equation can be used to solve the field non-iteratively [21,22]. In this method, the difference between the intensities measured in the two planes is described with a differential equation, from which the field can be solved. A related method which uses shifting Gaussian filters is presented in [23].

A method similar to the focus-variation method is the 2D astigmatism variation method [24]. Instead of varying the defocus parameter to get different intensity measurements, two second-order astigmatism parameters are being varied. With this method, the object $O(\mathbf{x})$ can be calculated non-iteratively, and no approximation needs to be made about the magnitude of $|\widetilde{O}(\mathbf{0})|$.

An overview of various non-iterative phase retrieval methods is given in [25].

In this paper, we present another non-iterative phase retrieval method based on parameter variation. Just like in the case of focus variation and 2D astigmatism variation, we modulate the phase in one space (real space or Fourier space) according to a parameter A_j, and we measure intensities I_j in the dual space. However, as opposed to focus variation, our method does not require the approximation of $|\widetilde{O}(\mathbf{0})|$ being large, and as opposed to 2D astigmatism variation, we only need to vary one parameter. Our method gives a general form of the phase modulation function we need to apply, and we will demonstrate that in a special case this method reduces to quantitative Zernike phase contrast. Thus, in a way our general method provides a framework which connects focus variation and astigmatism variation with quantitative Zernike phase contrast, while providing an entire class of alternatives as well.

2. Method

The novel non-iterative phase retrieval method that we explain in this section can be applied in a microscopy setup (see Fig. 1a), or in a focused probe or CDI setup (see Fig. 1b). Let us for the sake of notation decide that we are treating the case for the CDI setup, but the same derivation holds for the microscopy setup if we interchange the roles of object space and Fourier space (if we assume there are no incoherent effects). It should be noted though that from a practical point of view the microscopy setup would be easier to implement than the CDI setup: in the microscopy setup one could with a Spatial Light Modulator (SLM) directly alter the phase of the field in the Fourier plane, while in the CDI setup it may not be so straightforward to shape the phase of the probe.

$O(\mathbf{x})$ can be reconstructed non-iteratively from intensity measurements as follows:

1. We have a complex-valued transmission function $O(\mathbf{x})$ of an object. We illuminate it with an illumination function $P_A(\mathbf{x}) = e^{2\pi i A f(\mathbf{x})}$.
2. For N different A, spaced by some interval Δ_A, we measure the intensity in the diffraction plane $I_A(\mathbf{x}') = |\mathcal{F}\{O \cdot P_A\}(\mathbf{x}')|^2$.
3. We reconstruct the object in $\mathbf{x} \neq \mathbf{0}$ using

$$O^*(\mathbf{0})O(\mathbf{x}) = \sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})}, \tag{1}$$

where $H(A)$ is a sampling function which we can choose to be e.g. Gaussian multiplied with a series of delta peaks that determine for which A we sample.
4. To reconstruct the object in $\mathbf{x} = \mathbf{0}$, we need to find $|O(\mathbf{0})|^2$. This is done by solving a quadratic equation.

In the following paragraphs we will demonstrate that this method works if $f(\mathbf{x})$ is chosen appropriately.

First, we will rewrite Eq. (1) so that it becomes more apparent why we can reconstruct $O(\mathbf{x})$ with this expression. Note that if $H(A)$ consists of multiple delta functions which indicate for which A we sample $I_A(\mathbf{x}')$, we can rewrite the right side of Eq. (1) as an integral over A

$$\sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})} \propto \int \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})}\mathrm{d}A. \tag{2}$$

We can rewrite $\mathcal{F}^{-1}\{I_A\}(\mathbf{x})$ as an autocorrelation function

$$\mathcal{F}^{-1}\{I_A\}(\mathbf{x}) = \int O(\mathbf{y})^* O(\mathbf{x}+\mathbf{y})e^{2\pi i A(f(\mathbf{x}+\mathbf{y})-f(\mathbf{y}))}\mathrm{d}\mathbf{y}. \tag{3}$$

Plugging this into Eq. (2) and defining $\widetilde{H}(A')$ as the Fourier transform of $H(A)$ we get

$$\begin{aligned}&\sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})}\\ &= \iint O(\mathbf{y})^* O(\mathbf{x}+\mathbf{y})e^{2\pi i A(f(\mathbf{x}+\mathbf{y})-f(\mathbf{y})-f(\mathbf{x}))}H(A)\mathrm{d}\mathbf{y}\mathrm{d}A\\ &= \int O(\mathbf{y})^* O(\mathbf{x}+\mathbf{y})\widetilde{H}(f(\mathbf{x}+\mathbf{y})-f(\mathbf{y})-f(\mathbf{x}))\mathrm{d}\mathbf{y}.\end{aligned} \tag{4}$$

Let us for now assume the ideal case that $H(A) = 1$ so that $\widetilde{H}(A') = \delta(A')$, i.e. we assume that we can sample $I_A(\mathbf{x}')$ for all A continuously. In that case Eq. (4) reduces to

$$\sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})} = \int O(\mathbf{y})^* O(\mathbf{x}+\mathbf{y})\delta(f(\mathbf{x}+\mathbf{y})-f(\mathbf{y}) - f(\mathbf{x}))\mathrm{d}\mathbf{y}. \tag{5}$$

Let us have a closer look at the argument of the delta function, which we define as

$$g(\mathbf{x},\mathbf{y}) = f(\mathbf{x}+\mathbf{y}) - f(\mathbf{y}) - f(\mathbf{x}). \tag{6}$$

Note that if $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, then $g(\mathbf{x},\mathbf{y}) = 0$ (where we have assumed without loss of generality that $f(\mathbf{0}) = 0$). For now we will assume that $\mathbf{x} \neq \mathbf{0}$. Suppose that $g(\mathbf{x},\mathbf{y}) = 0$ only if $\mathbf{y} = \mathbf{0}$. In that case Eq. (5) will reduce to

$$\sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x})H(A)e^{-2\pi i A f(\mathbf{x})} \propto O(\mathbf{0})^* O(\mathbf{x}), \tag{7}$$

which is what we want (the expressions are in this case proportional to each other, not equal, because the determinant of the Jacobian is omitted. However, in case that we pixelate $I_A(\mathbf{x}')$ and $O(\mathbf{x})$, i.e. we discretize $\mathbf{x}$, as will always be the case in practice, this factor is irrelevant). Although the preceding derivation was not very rigorous in using delta functions, it can be made mathematically rigorous by

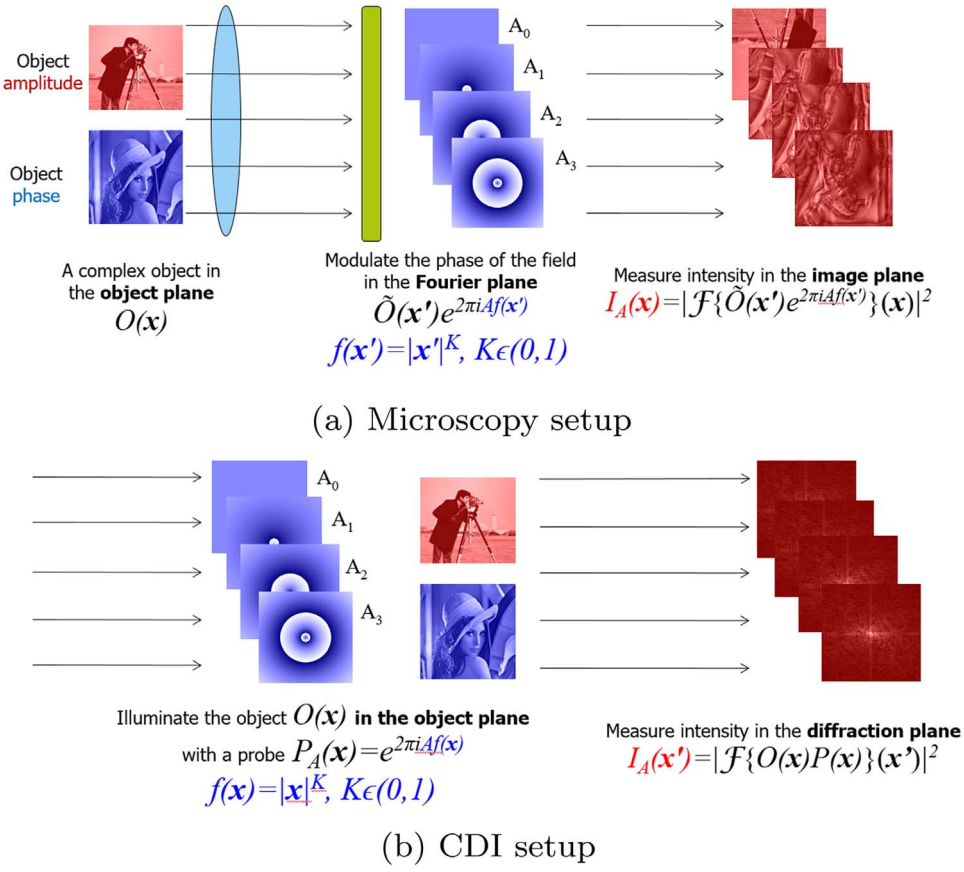


Fig. 1. Illustrations of the microscopy setup and the CDI setup in which the proposed non-iterative phase retrieval method may be used.

approximating the delta functions with narrow continuous functions. The question we need to answer now is the following: how should we choose $f(\mathbf{x})$, such that $g(\mathbf{x}, \mathbf{y}) = 0$ *only if* $\mathbf{y} = \mathbf{0}$ or $\mathbf{x} = \mathbf{0}$?

2.1. Choosing $f(\mathbf{x})$

First, let us look at the method of focus variation and see why it fails to meet our requirements. In case of focus variation, we have $f(\mathbf{x}) = |\mathbf{x}|^2$, in which case we get

$$g(\mathbf{x}, \mathbf{y}) = (\mathbf{x} + \mathbf{y})\cdot(\mathbf{x} + \mathbf{y}) - \mathbf{y}\cdot\mathbf{y} - \mathbf{x}\cdot\mathbf{x} = 2\mathbf{x}\cdot\mathbf{y}. \tag{8}$$

Obviously, this fails our requirement because $\mathbf{x}\cdot\mathbf{y} = 0$ whenever $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, not only if $\mathbf{x} = 0$ or $\mathbf{y} = 0$. We demonstrate that a function of the following form will satisfy the requirement:

$$f(\mathbf{x}) = h(n(\mathbf{x})). \tag{9}$$

Here, $n(\mathbf{x})$ is a vector norm (e.g. the Euclidean norm $n(\mathbf{x}) = \sqrt{\mathbf{x}\cdot\mathbf{x}}$), and $h(a)$ is a monotonically increasing subadditive function, i.e.

$$h(a + b) \leq h(a) + h(b), \tag{10}$$

where equality holds only when a=0 or b=0. An example would be $h(a) = a^K$, with $K \in (0, 1)$. To see why a function $f(\mathbf{x})$ of the form $h(n(\mathbf{x}))$ works, consider the inequality

$$\begin{aligned} f(\mathbf{x} + \mathbf{y}) &= h(n(\mathbf{x} + \mathbf{y})) \leq h(n(\mathbf{x}) + n(\mathbf{y})) \leq h(n(\mathbf{x})) + h(n(\mathbf{y})) \\ &= f(\mathbf{x}) + f(\mathbf{y}). \end{aligned} \tag{11}$$

The first inequality holds because of the triangle inequality (which holds by definition of a vector norm) and because $h(a)$ is a monotonically increasing function. The second inequality holds because $h(a)$ is a subadditive function. Note that equality only holds when $n(\mathbf{x}) = 0$ or $n(\mathbf{y}) = 0$, which by definition of a vector norm holds only when $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$. Thus, $g(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} + \mathbf{y}) - (f(\mathbf{x}) + f(\mathbf{y}))$ only vanishes when $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$, which is what we required.

2.2. The sampling function H(A)

In Eq. (5) we have assumed that $H(A) = 1$, i.e. that we can sample $I_A(\mathbf{x}')$ continuously over an infinite range. Now we will have a look at what happens when we sample A in a discrete number of N points spaced by intervals of Δ_A over a limited range $N\Delta_A$. By the properties of the Fourier transform, $\widetilde{H}(A')$ will consist of aliases separated by intervals of $1/\Delta_A$. If we choose the envelope of $H(A)$ to be a Gaussian (to prevent sidelobes in $\widetilde{H}(A')$), then the width of $\widetilde{H}(A')$ is inversely proportional to the width of $H(A)$. This is illustrated in Fig. 2.

Ideally, $\widetilde{H}(A')$ would be a delta function as is assumed in Eq. (5). In practice, we can only make it a narrow peak with a finite width that is inversely proportional to the sampling range $N\Delta_A$. We can make the following remarks about the required sampling range $N\Delta_A$ and how it is affected by our choice of $f(\mathbf{x})$:

- For practical reasons, we want to make as few measurements as possible. Thus, we desire the sampling range $N\Delta_A$ to be small, which means $\widetilde{H}(A')$ would have to be broad.
- At the same time, from Eq. (4) we see we want $\widetilde{H}(g(\mathbf{x}, \mathbf{y}))$ to have large values for a small range of $\mathbf{y}$ around $\mathbf{y} = \mathbf{0}$. This could be achieved by making $\widetilde{H}(A')$ narrower, but this would be in conflict with the previous point.
- Alternatively one could make sure that $g(\mathbf{x}, \mathbf{y})$ is small for a small range of $\mathbf{y}$ around $\mathbf{y} = \mathbf{0}$, so that even if $\widetilde{H}(A')$ is broad, $\widetilde{H}(g(\mathbf{x}, \mathbf{y}))$ has a large value for only a small range of $\mathbf{y}$ around $\mathbf{y} = \mathbf{0}$.
- If we choose $f(\mathbf{x}) = |\mathbf{x}|^K$, $K \in (0, 1)$, then the region of $\mathbf{y}$ for which $g(\mathbf{x}, \mathbf{y})$ is small decreases as K decreases. As we will show in Section 2.4, for $K \to 0$ our method is equivalent to quantitative Zernike phase contrast.

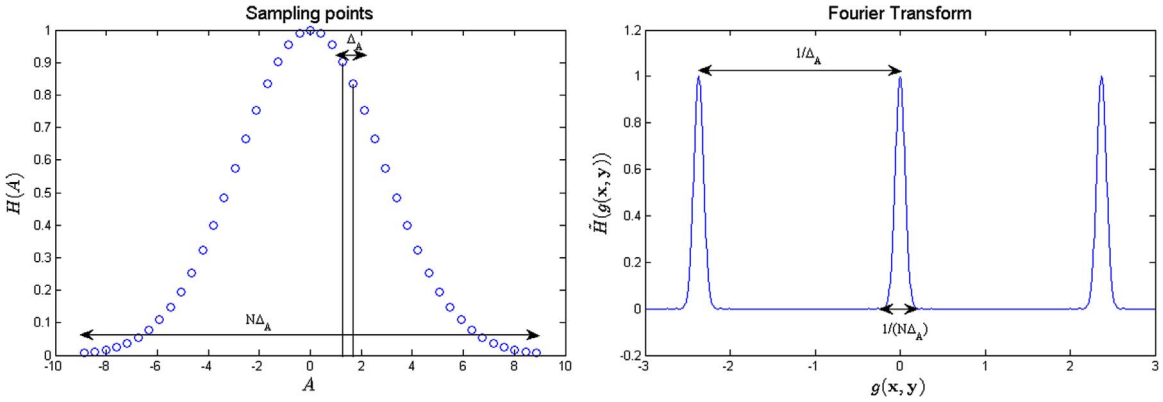
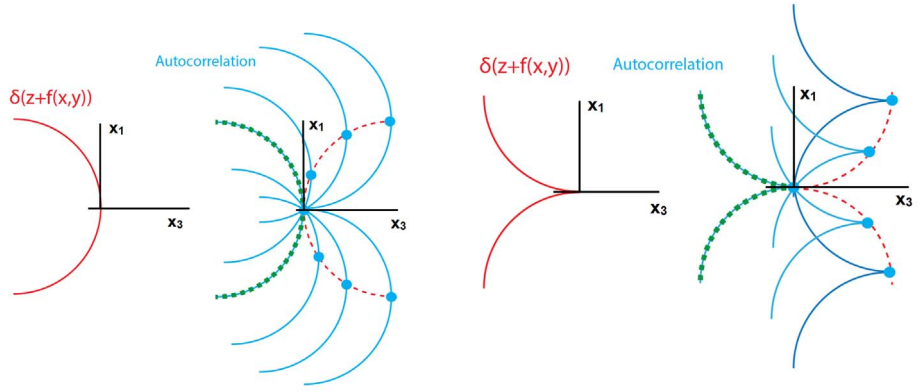
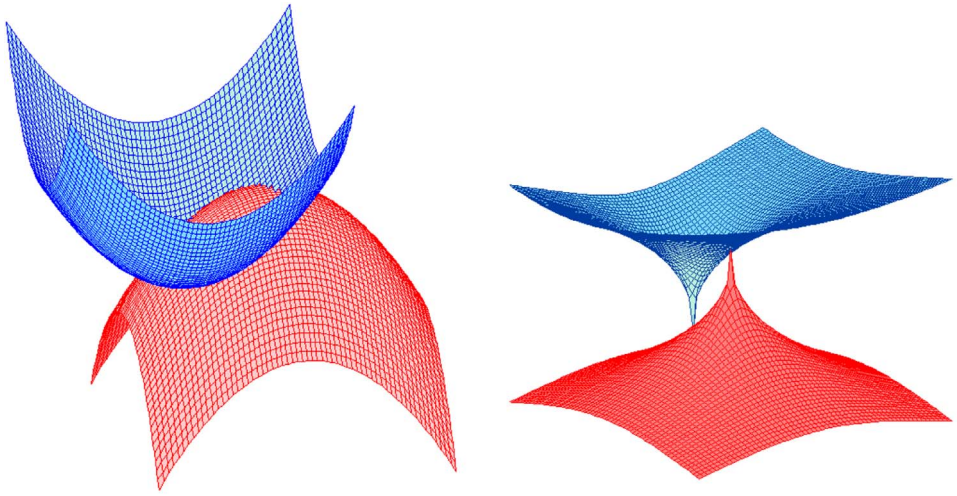


Fig. 2. Illustrations of the sampling function $H(A)$ and its Fourier transform $\widetilde{H}(A')$ which in Eq. (4) is evaluated in $A' = g(\mathbf{x}, \mathbf{y})$.



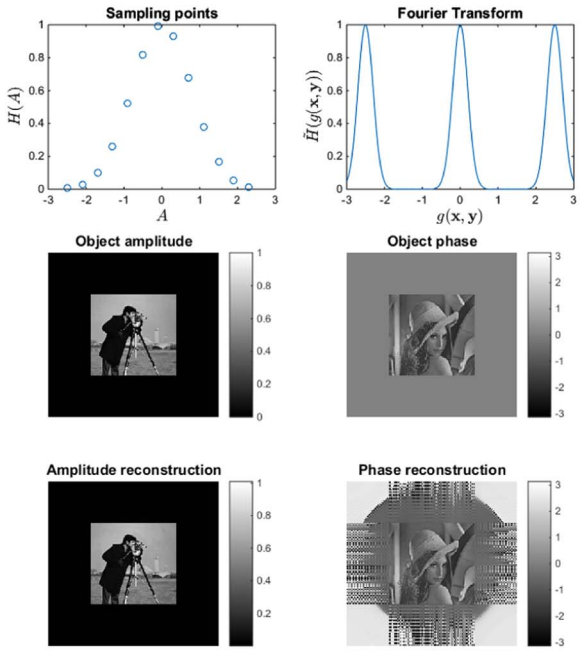
(a) $f(x_1, x_2) \propto x_1^2 + x_2^2$, focus variation (b) $f(x_1, x_2) \propto \sqrt{x_1^2 + x_2^2}^K$, $K \in (0, 1)$, proposed method

Fig. 3. Plots showing cross-sections of autocorrelations for different choices of $f(x_1, x_2)$. The autocorrelation of $O_{3D}(x_1, x_2, x_3)$ consists of a sum of copies of $O_{3D}(x_1, x_2, x_3)$ (blue curves) that are shifted by $(-y_1, -y_2, -y_3)$ for those (y_1, y_2, y_3) for which $O_{3D}(y_1, y_2, y_3)$ is nonzero (dotted red curve). The surface $\delta(x_3 + f(x_1, x_2))$ is obtained by rotating the illustrated cross-section around its symmetry axis (see Fig. 4), we see that in the case of $f(x, y) \propto x^2 + y^2$ the autocorrelation evaluated in a point on the red dotted surface ($x_3 = f(-x_1, -x_2)$) or green dotted surface ($x_3 = f(x_1, x_2)$) contains the contributions of multiple blue copies. However, in the case of $f(x_1, x_2) \propto \sqrt{x_1^2 + x_2^2}^K$, $K \in (0, 1)$ the autocorrelation evaluated in a point on the red dotted surface or the green dotted surface contains the contribution of just one blue copy, so the values on the surfaces are directly proportional to the values of $O(x_1, x_2)$ or $O(-x_1, -x_2)^*$. The only exception is at the cusp of the surface. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

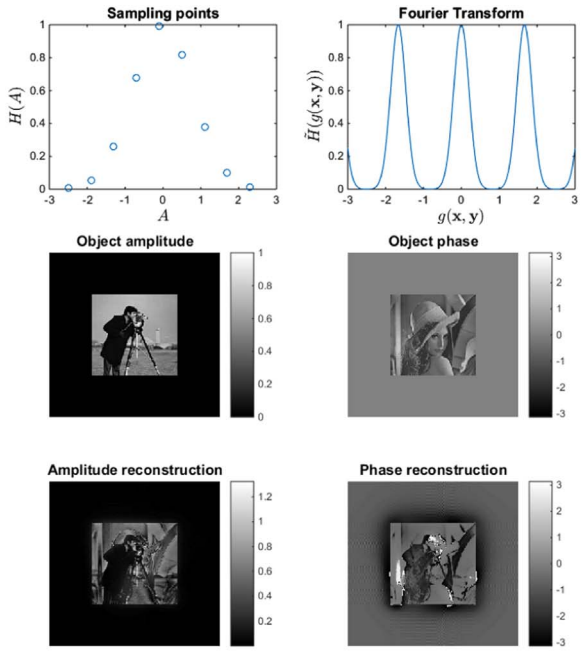


(a) $f(x_1, x_2) \propto x_1^2 + x_2^2$, focus variation (b) $f(x_1, x_2) \propto \sqrt{x_1^2 + x_2^2}^K$, $K \in (0, 1)$, proposed method

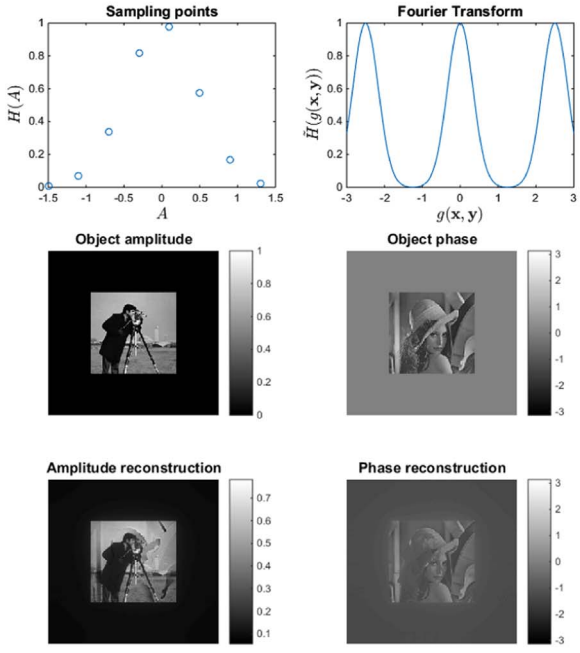
Fig. 4. Illustration of how the surfaces of Fig. 3 intersect. The red surface here corresponds to the red dotted surface in Fig. 3, and the blue surface here corresponds to the blue copies in Fig. 3. In this 3D plot it becomes apparent that the paraboloids we get with focus variation intersect the red dotted surface of Fig. 3 in many points, while for the proposed method the surfaces intersect in only two distinct points.



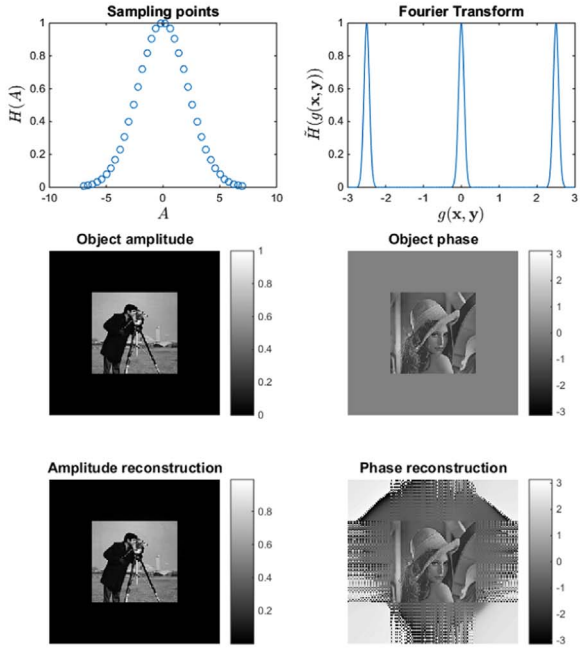
(a) Reconstruction, $K = 0.1, \Delta_A = 0.4, N = 13$



(b) $K = 0.1, \Delta_A = 0.4, N = 8$, too short sampling range $N\Delta_A$



(c) $K = 0.1, \Delta_A = 0.6, N = 9$, too large sampling interval Δ_A



(d) Reconstruction, $K = 0.3, \Delta_A = 0.4, N = 36$

Fig. 5. Simulation results showing how the sampling interval Δ_A and the sampling range $N\Delta_A$ affect the reconstruction quality for different K in case we choose $f(\mathbf{x}) = |\mathbf{x}|^K$. We assume the microscopy setup as in Fig. 1(a).

The sampling interval Δ_A determines how far the aliases lie apart. In Eq. (4), we want the integrand to contribute to the integral only when $g(\mathbf{x}, \mathbf{y})$ is small. Thus, in order to prevent the aliases from contributing to the integral, we require that $O(\mathbf{x}) \approx 0$ for those $g(\mathbf{x}, \mathbf{y})$. Thus, the required sampling interval Δ_A is determined by the extent of the object $O(\mathbf{x})$, while the required sampling range $N\Delta_A$ is determined by the resolution with which we want to reconstruct $O(\mathbf{x})$.

2.3. Reconstructing $O(\mathbf{x})$ in $\mathbf{x} = \mathbf{0}$

Looking at Eq. (5) we see that if $\mathbf{x} \neq \mathbf{0}$, then the integrand only contributes to the integral when $\mathbf{y} = \mathbf{0}$, giving a direct reconstruction of $O(\mathbf{0})^* O(\mathbf{x})$. However, for $\mathbf{x} = \mathbf{0}$ we want to reconstruct $|O(\mathbf{0})|^2$, but, if we discretize $\mathbf{x}$ (since the image is pixellated), the integral of Eq. (5) gives

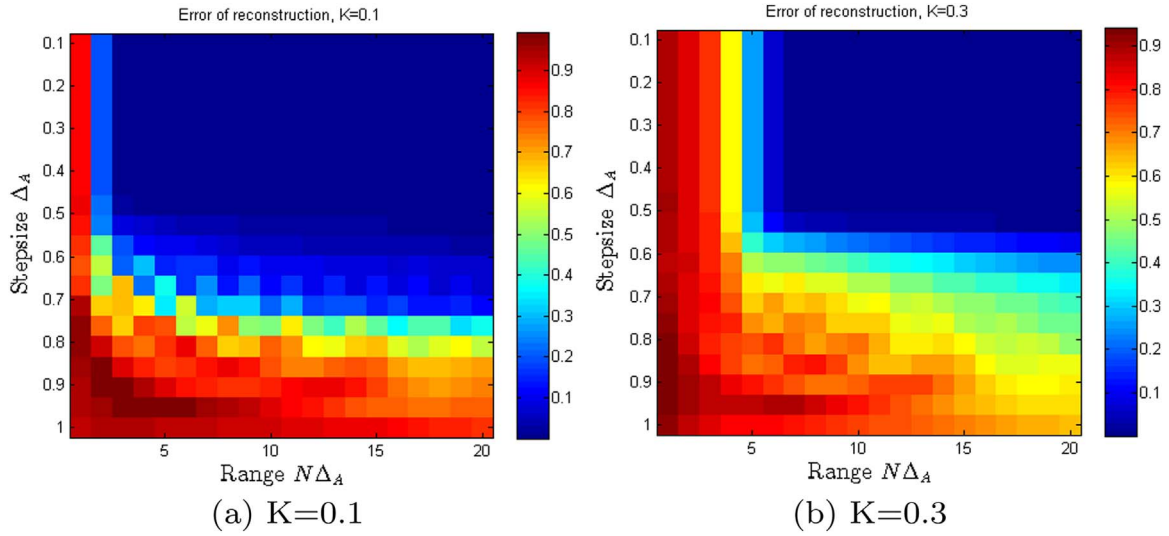


Fig. 6. Plots of the reconstruction error for the microscopy setup in case K=0.1 and K=0.3. It is seen that once the sampling range $N\Delta_A$ and the sampling interval Δ_A exceed certain thresholds, the reconstruction is successful.

$$O_{\text{recon}}(\mathbf{x}) = \begin{cases} O(\mathbf{0})^* O(\mathbf{x}) & \text{if } \mathbf{x} \neq \mathbf{0}, \\ \sum_{\mathbf{y}} |O(\mathbf{y})|^2 & \text{if } \mathbf{x} = \mathbf{0}. \end{cases} \tag{12}$$

From this we can derive the equation

$$\sum_{\mathbf{y}\neq\mathbf{0}} |O_{\text{recon}}(\mathbf{y})|^2 = |O(\mathbf{0})|^2 O_{\text{recon}}(\mathbf{0}) - |O(\mathbf{0})|^4. \tag{13}$$

This quadratic equation in $|O(\mathbf{0})|^2$ can be solved as

$$|O(\mathbf{0})|^2 = \frac{O_{\text{recon}}(\mathbf{0}) \pm \sqrt{O_{\text{recon}}(\mathbf{0})^2 - 4\sum_{\mathbf{y}\neq\mathbf{0}} |O_{\text{recon}}(\mathbf{y})|^2}}{2}. \tag{14}$$

To determine which sign gives the correct answer, one can see for which of the two possible values of $|O(\mathbf{0})|^2$ the calculated intensity patterns $I(\mathbf{x}'; A) = |\mathcal{F}\{O(\mathbf{x})P(\mathbf{x})\}|^2$ best match the measured intensity patterns. With this, the reconstruction of $O(\mathbf{0})^* O(\mathbf{x})$ is completed.

2.4. Quantitative Zernike phase contrast as a special case

It was argued before that if we choose $f(\mathbf{x}) = |\mathbf{x}|^K$, then for $K \to 0$ few measurements are needed for a good reconstruction. In this section we demonstrate that for $K \to 0$ the method reduces to quantitative Zernike phase contrast. When $K \to 0$ we get

$$f(\mathbf{x}) = \begin{cases} 0 & \text{if } |\mathbf{x}| = 0, \\ 1 & \text{if } |\mathbf{x}| \neq 0. \end{cases} \tag{15}$$

Thus, we are changing the phase in all but one pixel, which is equivalent to changing the phase in only one pixel. With this choice of $f(\mathbf{x})$, it follows that if $\mathbf{x} \neq \mathbf{0}$

$$g(\mathbf{x}, \mathbf{y}) = \begin{cases} 0 & \text{if } |\mathbf{y}| = 0, \\ -2 & \text{if } \mathbf{y} = -\mathbf{x}, \\ -1 & \text{otherwise.} \end{cases} \tag{16}$$

Thus, according to Eq. (4), we must choose $H(A)$ such that

$$\widehat{H}(0) = 1 \quad \widehat{H}(-1) = 0 \quad \widehat{H}(-2) = 0. \tag{17}$$

An option would be

$$H(A) \propto \delta(A) + \delta\left(A - \frac{1}{3}\right) + \delta\left(A - \frac{2}{3}\right). \tag{18}$$

That is, we need to take three measurements, namely with A=0, $A = \frac{1}{3}$, and $A = \frac{2}{3}$. This procedure is the same as in quantitative Zernike phase contrast. However, if one changes the phase in only one pixel, the variation in the measured intensity patterns will be very small (unless the amplitude in that one pixel is very large), and thus the method can be very sensitive to noise. By choosing K small but finite, one can still obtain a non-iterative reconstruction, while having a larger diversity in the intensity measurements. However, one would then need to take more than three measurements.

2.5. Reconstructing the twin image $O(-\mathbf{x})^*$

The reconstruction formula in Eq. (1) was chosen such that in the autocorrelation integral of Eq. (3), $O(\mathbf{0})^* O(\mathbf{x})$ is sifted out. This means that the argument of the delta function in Eq. (5), which we defined to be $g(\mathbf{x}, \mathbf{y})$, should vanish only in $\mathbf{x} = \mathbf{0}$ or $\mathbf{y} = \mathbf{0}$. However, one may ask if it is also possible to sift out $O(-\mathbf{x})^* O(\mathbf{0})$ instead. This would mean $g(\mathbf{x}, \mathbf{y})$ would have to vanish in $\mathbf{x} = -\mathbf{y}$. We can achieve this by choosing as our reconstruction function

$$O(-\mathbf{x})^* O(\mathbf{0}) = \sum_A \mathcal{F}^{-1}\{I_A\}(\mathbf{x}) H(A) e^{2\pi i A f(-\mathbf{x})}. \tag{19}$$

In this case, we get for the argument of the delta function

$$g(\mathbf{x}, \mathbf{y}) = f(\mathbf{x} + \mathbf{y}) - f(\mathbf{y}) + f(-\mathbf{x}). \tag{20}$$

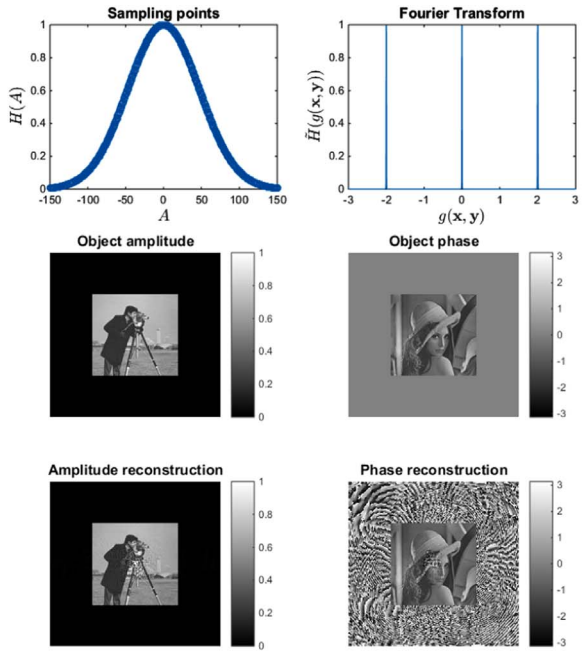
Indeed one can simply verify that this function vanishes if $\mathbf{x} = \mathbf{0}$ or $\mathbf{x} = -\mathbf{y}$ (assuming $f(\mathbf{0}) = 0$ as before). Now we need to make sure that $f(\mathbf{x})$ is chosen such that $g(\mathbf{x}, \mathbf{y})$ vanishes *only* in these points. One can substitute $\mathbf{a} = \mathbf{x} + \mathbf{y}$, $\mathbf{b} = -\mathbf{x}$ to get

$$g(\mathbf{a}, \mathbf{b}) = f(\mathbf{a}) - f(\mathbf{a} + \mathbf{b}) + f(\mathbf{b}). \tag{21}$$

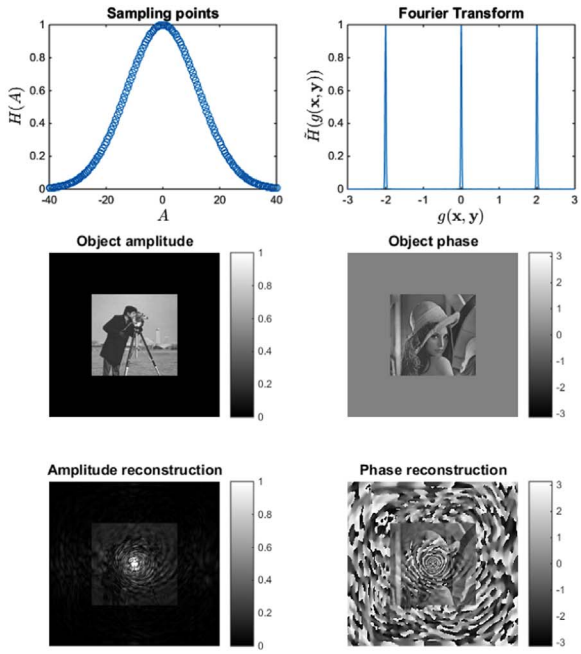
This equation has, aside from a minus sign, the same form as Eq. (6). Also, we impose the same condition on $f(\mathbf{a})$: it should be such that $g(\mathbf{a}, \mathbf{b})$ vanishes only for $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$. Thus, the function $f(\mathbf{x}) = |\mathbf{x}|^K$ is also in this case a valid solution. Because $f(-\mathbf{x}) = f(\mathbf{x})$, the difference between the reconstruction formulas of Eqs. (1) and (19) is merely the sign in the complex exponential.

2.6. Geometric interpretation using 3D autocorrelation functions

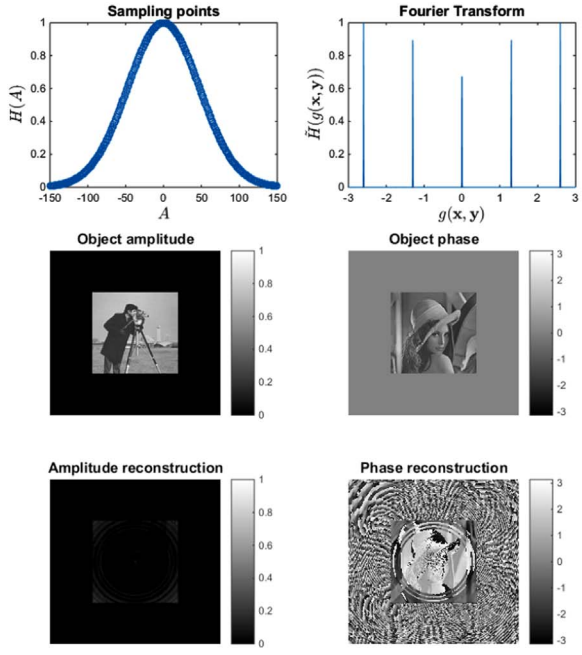
We can interpret the results found previously in a geometric way using autocorrelation functions. To do this, we interpret the set of



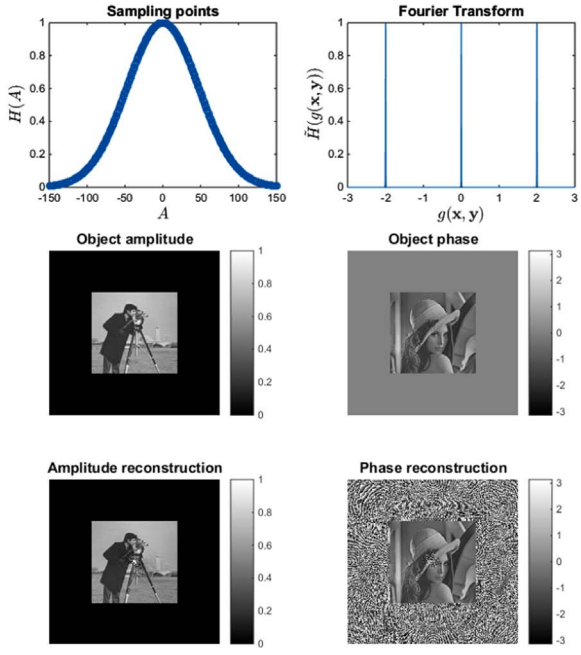
(a) Reconstruction, $K = 0.1, \Delta_A = 0.5, N = 601$ (amplitude scale cut off at 1)



(b) $K = 0.1, \Delta_A = 0.5, N = 161$, too short sampling range $N\Delta_A$



(c) $K = 0.1, \Delta_A = 0.77, N = 390$, Too large sampling interval Δ_A (amplitude scale cut off at 1)



(d) Reconstruction, $K = 0.3, \Delta_A = 0.5, N = 601$ (amplitude scale cut off at 1)

Fig. 7. Simulation results showing how the sampling interval Δ_A and the sampling range $N\Delta_A$ affect the reconstruction quality for different K in case we choose $f(\mathbf{x}) = |\mathbf{x}|^K$. We assume the CDI setup as in Fig. 1(b).

intensity measurements as one 3D object $I_{3D}(x'_1, x'_2, A)$, rather than a sequence of 2D objects $I(x'_1, x'_2)$. We can transform the two-dimensional object $O(x_1, x_2)$ into a 3D object $O_{3D}(x_1, x_2, x_3)$ by defining

$$O_{3D}(x_1, x_2, x_3) = O(x_1, x_2)\delta(x_3 + f(x_1, x_2)), \tag{22}$$

which means we stretch out $O(x_1, x_2)$ over a surface $x_3 = -f(x_1, x_2)$. For example, in the case of paraxial focus variation we have $f(x_1, x_2) \propto x_1^2 + x_2^2$, which means $O(x_1, x_2)$ is stretched out onto a paraboloid. With this definition of $O_{3D}(x_1, x_2, x_3)$ we find that

$$I_{3D}(x'_1, y'_2, A) := I_A(x'_1, x'_2) = \left| \iint O(x_1, x_2) e^{-2\pi i (x_1 x'_1 + x_2 x'_2)} e^{2\pi i A f(x_1, x_2)} \mathrm{d}x_1 \mathrm{d}x_2 \right|^2 = |\mathcal{F}_{3D}\{O_{3D}\}(x'_1, y'_2, A)|^2. \quad (23)$$

Since $I_{3D}(x'_1, x'_2, A) = |\mathcal{F}\{O_{3D}\}(x'_1, x'_2, A)|^2$, $\mathcal{F}^{-1}\{I_{3D}\}(x_1, x_2, x_3)$ gives the 3D autocorrelation function of $O_{3D}(x_1, x_2, x_3)$. In Eq. (5) we evaluate this autocorrelation function in $x_3 = -f(x_1, x_2)$ (and in Eq. (19) we evaluate it in the surface $x_3 = f(-x_1, -x_2)$). In the derivation that followed, we essentially demonstrated that this region of the autocorrelation function is directly proportional to the original object $O(x_1, x_2)$ (or in the case of Eq. (19) to its twin image $O(-x_1, -x_2)^*$).

There are two reasons why this interpretation may be valuable:

1. It allows for a visual interpretation (see Fig. 3) of why certain choices of $f(x_1, x_2)$ may or may not work, and, using the theory of sampling in the Fourier domain, one may better understand how the discrete sampling of A may affect the object reconstruction.
2. It makes more obvious the link to other reconstruction methods that also obtain a direct reconstruction of the object from a certain region of the autocorrelation function. These methods would include digital holography, and, when speaking in the context of CDI, Fourier transform holography [11] in particular. In all these methods, there is a region that corresponds to the original object $O(x_1, x_2)$, and a region that corresponds to its twin image $O(-x_1, -x_2)^*$.

3. Simulations

In Section 2.2 we discussed the importance of a correct sampling interval Δ_A and sampling range $N\Delta_A$. In Figs. 5 and 7 simulations are shown for respectively the microscopy setup (Fig. 1a) and the CDI setup (Fig. 1b). The difference is that in the microscopy setup we reconstruct the Fourier transform $\widetilde{O}(\mathbf{x}')$ from which we find the complex-valued object $O(\mathbf{x})$, whereas in the CDI setup we reconstruct $O(\mathbf{x})$ directly.

If we choose $f(\mathbf{x}) = |\mathbf{x}|^K$, then for K=0.1 it is shown in Figs. 5a–c for the microscopy setup how the sampling interval and the sampling range affect the quality of the reconstruction. Indeed, if K is increased to K=0.3, it is shown in Fig. 5d that a larger sampling interval is required by increasing N. In Fig. 6 it is shown that there is indeed a relatively sharp threshold for Δ_A and $N\Delta_A$ for the reconstruction to be successful. The functional we have used to characterize the reconstruction error is

$$E[O_{\mathrm{recon}}(\mathbf{x})] = \frac{\int |O(\mathbf{x}) - cO_{\mathrm{recon}}(\mathbf{x})|^2 \mathrm{d}\mathbf{x}}{\int |O(\mathbf{x})|^2 \mathrm{d}\mathbf{x}}. \quad (24)$$

Here, c is a complex constant that minimizes E. This assures that if $O_{\mathrm{recon}}(\mathbf{x}) = e^{i\theta} O(\mathbf{x})$, the error is 0 as it should be. c is found by solving $\mathrm{d}E/\mathrm{d}c = 0$, which gives

$$c^* = \frac{\int O^*(\mathbf{x}) O_{\mathrm{recon}}(\mathbf{x}) \mathrm{d}\mathbf{x}}{\int |O_{\mathrm{recon}}(\mathbf{x})|^2 \mathrm{d}\mathbf{x}}. \quad (25)$$

In Fig. 7 we see that for the CDI setup we need significantly more measurements to reconstruct the object correctly. This is because in the CDI setup we reconstruct $O(\mathbf{x})$ directly, while in the microscopy setup we reconstruct $\widetilde{O}(\mathbf{x}')$ from which we can find $O(\mathbf{x})$. It should be noted that $\widetilde{O}(\mathbf{x}')$ peaks sharply at $\mathbf{x}' = \mathbf{0}$, whereas $O(\mathbf{x})$ does not peak sharply anywhere. The reconstruction error that comes from integral of Eq. (4) is (for $\mathbf{x} \neq \mathbf{0}$)

$$\int_{\mathbf{y} \neq \mathbf{0}} O(\mathbf{y})^* O(\mathbf{x} + \mathbf{y}) \widetilde{H}(f(\mathbf{x} + \mathbf{y}) - f(\mathbf{y}) - f(\mathbf{x})) \mathrm{d}\mathbf{y}. \quad (26)$$

In case we reconstruct $O(\mathbf{x})$ directly (as in the CDI setup), the value of $O(\mathbf{y})^* O(\mathbf{x} + \mathbf{y})$ has the same order of magnitude for all $\mathbf{y}$, so the reconstruction error is mainly determined by how sharply $\widetilde{H}(g(\mathbf{x}, \mathbf{y}))$ peaks at $\mathbf{y} = \mathbf{0}$ for all $\mathbf{x} \neq \mathbf{0}$, which is determined by the sampling range $N\Delta_A$. When reconstructing $\widetilde{O}(\mathbf{x}')$ however (as in the microscopy setup), the error term is

$$\int_{\mathbf{y}' \neq \mathbf{0}} \widetilde{O}(\mathbf{y}')^* \widetilde{O}(\mathbf{x}' + \mathbf{y}') \widetilde{H}(f(\mathbf{x}' + \mathbf{y}') - f(\mathbf{y}') - f(\mathbf{x}')) \mathrm{d}\mathbf{y}'. \quad (27)$$

In this case, the value of $\widetilde{O}(\mathbf{y}')^* \widetilde{O}(\mathbf{x}' + \mathbf{y}')$ peaks sharply at $\mathbf{y}' = \mathbf{0}$ and $\mathbf{y}' = -\mathbf{x}'$, because $\widetilde{O}(\mathbf{x}')$ peaks at $\mathbf{x}' = \mathbf{0}$. The reconstruction will therefore be approximately proportional to

$$\widetilde{O}(\mathbf{0})^* \widetilde{O}(\mathbf{x}') \widetilde{H}(0) + \widetilde{O}(-\mathbf{x}')^* \widetilde{O}(\mathbf{0}) \widetilde{H}(-2f(\mathbf{x}')). \quad (28)$$

Thus, whereas when reconstructing $O(\mathbf{x})$ we have to make sure that $\widetilde{H}(g(\mathbf{x}, \mathbf{y}))$ peaks very sharply at $\mathbf{y} = \mathbf{0}$, when reconstructing $\widetilde{O}(\mathbf{x})$ it suffices to make sure that $\widetilde{H}(g(\mathbf{x}', -\mathbf{x}')) = \widetilde{H}(-2f(\mathbf{x}'))$ is small for $\mathbf{x}' \neq \mathbf{0}$. This is a much less strict requirement, meaning the required sampling range for reconstructing $\widetilde{O}(\mathbf{x}')$ is much smaller than the required sampling range for reconstructing $O(\mathbf{x})$. If $\widetilde{H}(-2f(\mathbf{x}'))$ does not decrease quickly enough with increasing $|\mathbf{x}'|$, we will get an error in reconstructing the lower spatial frequencies of $O(\mathbf{x})$. The mixing of amplitude and phase information which is observed in Fig. 5b confirms this.

We have noted before that a high 0th diffraction order is beneficial for phase contrast methods such as Zernike phase contrast imaging [9], quantitative Zernike phase contrast imaging [10], and the focus-variation method [14,16]. Indeed, seeing how these methods are very much related to our proposed phase retrieval method, it is not surprising that a high 0th diffraction order is also beneficial for our method.

4. Conclusion

We have derived a non-iterative phase retrieval method where by modulating the phase in one plane (real space or Fourier space) by $e^{i2\pi A f(\mathbf{x})}$ and measuring the intensity patterns $I_A(\mathbf{x}')$ in the dual space, we can reconstruct the object transmission function. For the phase modulation function $f(\mathbf{x})$ we found a general requirement: it has to be a composition of a vector norm and a monotonically increasing subadditive function. A particular set of functions that satisfy this requirement is $f(\mathbf{x}) = |\mathbf{x}|^K$, $K \in (0, 1)$, and in case we choose $K \to 0$ the method reduces to quantitative Zernike phase contrast as in [10]. Moreover, we have shown how this method can be interpreted as obtaining an object reconstruction directly from a part of an autocorrelation function, as is also the case in Fourier transform holography [11]. We have discussed how the sampling function $H(A)$ affects the reconstruction, and illustrated this statement with simulations. The method can be applied in either a microscopy setup or a CDI setup, though we have shown that the number of intensity measurements required for successful object reconstruction is significantly larger for a CDI setup. Given the general formulation of the phase retrieval method which allows for plenty of customization, the applications may be diverse.

References

[1] Z. Bialynicka-Birula, I. Bialynicki-Birula, Reconstruction of the wavefunction from the photon number and quantum phase distributions, J. Mod. Opt. 41 (November (11)) (1994) 2203–2208.

[2] N. Nakajima, Reconstruction of a wave function from the q function using a phase-retrieval method in quantum-state measurements of light, Phys. Rev. A 59 (June (6)) (1999) 4164–4171.

[3] K.A. Nugent, D. Paganin, Matter-wave phase measurement: a noninterferometric approach, Phys. Rev. A 61 (May (6)) (2000).

[4] P. Thibault, A. Menzel, Reconstructing state mixtures from diffraction measurements, Nature 494 (February (7435)) (2013) 68–71.

[5] R.W. Gerchberg, W.O. Saxton, A practical algorithm for the determination of phase from image and diffraction plane pictures, Optik 35 (1972) 237.

[6] J.R. Fienup, Reconstruction of an object from the modulus of its fourier transform, Opt. Lett. 3 (July (1)) (1978) 27.

[7] J.M. Rodenburg, H.M.L. Faulkner, A phase retrieval algorithm for shifting illumination, Appl. Phys. Lett. 85 (20) (2004) 4795.

[8] Andrew M. Maiden, John M. Rodenburg, An improved ptychographical phase

retrieval algorithm for diffractive imaging, Ultramicroscopy 109 (September (10)) (2009) 1256–1262.
[9] F. Zernike, Phase contrast, a new method for the microscopic observation of transparent objects, Physica 9 (July (7)) (1942) 686–698.
[10] P. Gao, B. Yao, I. Harder, N. Lindlein, F.J. Torcal-Milla, Phase-shifting zernike phase contrast microscopy for quantitative phase measurement, Opt. Lett. 36 (November (21)) (2011) 4305.
[11] S. Eisebitt, J. Lüning, W.F. Schlotter, M. Lörgen, O. Hellwig, W. Eberhardt, J. Stöhr, Lensless imaging of magnetic nanostructures by x-ray spectro-holography, Nature 432 (December (7019)) (2004) 885–888.
[12] M. Guizar-Sicairos, J.R. Fienup, Direct image reconstruction from a fourier intensity pattern using HERALDO, Opt. Lett. 33 (November (22)) (2008) 2668.
[13] A.V. Martin, L.J. Allen, Direct retrieval of a complex wave from its diffraction pattern, Opt. Commun. 281 (October (20)) (2008) 5114–5121.
[14] W.O. Saxton, Advances in Electronics and Electron Physics: Computer Techniques for Image Processing in Electron Microscopy, Academic Press, New York, 1978 (Chapter 9.7).
[15] D. Van Dyck, W. Coene, A new procedure for wave function restoration in high resolution electron microscopy, Optik 77 (3) (1987) 125–128.
[16] M. op de Beeck, D. van Dyck, W. Coene, Wave function reconstruction in HRTEM: the parabola method, Ultramicroscopy 64 (August (1–4)) (1996) 167–183.
[17] A.H. Buist, A. van den Bos, M.A.O. Miedema, Optimal experimental design for exit wave reconstruction from focal series in TEM, Ultramicroscopy 64 (August (1–4)) (1996) 137–152.
[18] A. Thust, W.M.J. Coene, M. Op de Beeck, D. Van Dyck, Focal-series reconstruction in HRTEM: simulation studies on non-periodic objects, Ultramicroscopy 64 (August (1–4)) (1996) 211–230.
[19] H.W. Zandbergen, D. Tang, J. Jansen, R.J. Cava, The use of through focus exit wave reconstruction in the structure determination of several intermetallic superconductors, Ultramicroscopy 64 (August (1–4)) (1996) 231–247.
[20] W.M.J. Coene, A. Thust, M. Op de Beeck, D. Van Dyck, Maximum-likelihood method for focus-variation image reconstruction in high resolution transmission electron microscopy, Ultramicroscopy 64 (August (1–4)) (1996) 109–135.
[21] M.R. Teague, Deterministic phase retrieval: a Green's function solution, J. Opt. Soc. Am. 73 (November (11)) (1983) 1434.
[22] T.E. Gureyev, A. Roberts, K.A. Nugent, Phase retrieval with the transport-of-intensity equation: matrix solution with use of Zernike polynomials, J. Opt. Soc. Am. A 12 (September (9)) (1995) 1932.
[23] N. Nakajima, Phase retrieval from Fresnel zone intensity measurements by use of Gaussian filtering, Appl. Opt. 37 (September (26)) (1998) 6219.
[24] W.M.J. Coene, A.J.E.M. Janssen, Method of reconstructing an image in a particle-optical apparatus, US Patent US5753913 A, 1996.
[25] L.J. Allen, A.J. DAlfonso, A.V. Martin, A.J. Morgan, H.M. Quiney, Deterministic approaches to coherent diffractive imaging, J. Opt. 18 (1) (2016) 014002.