Accelerated Wirtinger Flow: A fast algorithm for ptychography

Rui Xu, Mahdi Soltanolkotabi, Justin P. Haldar, Walter Unglaub,
Joshua Zusman, Anthony F. J. Levi, Richard M. Leahy
Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089

Abstract

This paper presents a new algorithm, Accelerated Wirtinger Flow (AWF), for ptychographic image reconstruction from phaseless diffraction pattern measurements. AWF is based on combining Nesterov's acceleration approach with Wirtinger gradient descent. Theoretical results enable prespecification of all AWF algorithm parameters, with no need for computationally-expensive line searches and no need for manual parameter tuning. AWF is evaluated in the context of simulated X-ray ptychography, where we demonstrate fast convergence and low per-iteration computational complexity. We also show examples where AWF reaches higher image quality with less computation than classical algorithms. AWF is also shown to have robustness to noise and probe misalignment.

1 Introduction

The ability to image objects at nm scales is of fundamental importance in a variety of scientific and engineering disciplines. For instance, successful imaging of large protein complexes and biological specimens at very fine scales may enable live imaging of biochemical behavior at the molecular level providing new insights. Similarly, modern multilayered integrated circuits increasingly contain features below 10nm in size. The ability to image such specimens non-destructively can be used to improve quality control during the manufacturing process.

Imaging at finer and finer resolutions necessitates shorter and shorter beam wavelengths. However, lens-like devices and other optical components are difficult to build at very short wavelengths. Phase-less coherent diffraction imaging techniques offer an alternative method for recovery of high resolution images without the need for involved measurment setups that include mirrors and lenses. Invention of new light sources and new experimental setups that enable recording and reconstruction of non-crystalline objects has caused a major revival in the use of phase-less imaging techniques [44, 56, 13, 76, 48, 2, 47, 45, 36, 55, 17, 77, 24]. More recently, successful experiments using ptychography [53, 64, 21, 57, 19, 32], Fourier ptychography [82, 75, 33, 16, 66, 67] and partially coherent Ptychography [12] have further contributed to this surge. There has also been tremendous progress in the development of phase retrieval methods with the introduction of new algorithmic approaches that include maximum-likelihood estimation [65], Ptychographic Iterative Engine (PIE) [51] and extended Ptychographic Iterative Engine (ePIE) [42], Difference Map (DM) [63, 64], Error Reduction (ER) [78], Relaxed Averaged Alternating Reflections (RAAR) [43], semidefinite programming [10, 7, 8, 35, 68], convex relaxation with an anchor vector [27, 4, 20, 26, 54], eigen-based angular

synchronization [3, 34], Wirtinger Flow (WF) [9], proximal algorithms [74, 30, 59], and majorize-minimize methods [73]. See also [14, 61, 58, 80, 5, 81, 6, 71, 40, 50, 18, 62, 28, 15, 70, 22, 11] for many interesting works on first-order methods and/or theoretical analysis with random sensing ensembles.

Despite all the aforementioned progress, major challenges impede the use of phaseless imaging techniques for imaging large specimens at nm scales. One major challenge is computational in nature. For example, imaging a 1 cm × 1 cm integrated circuit specimen at 10 nm resolution results in on the order of 10^{12} pixels that need to be reconstructed, and the image sizes grow dramatically when extended to 3D tomographic imaging applications. Given the computational complexity of processing such large data, it is important that phase retrieval algorithms converge quickly to good solutions.

For practical phaseless imaging modalities such as X-ray ptychography [29, 51, 52, 53], classical phase retrieval algorithms often exhibit slow convergence rates. Another challenge is that the image reconstruction task in phaseless imaging often involves highly non-convex optimization problems with many spurious local optima so that classical approaches can converge to suboptimal solutions or even may not converge at all. Recently, theoretical results have proven that gradient-descent methods, also known as WFs, will converge with high probability to globally optimal solutions in the case of randomized sensing ensembles [9]. Random/psuedo-random sensing ensembles can be realized in the visible light range via spatial light modulators or phase from defocus [38]. Unfortunately, these theoretical results do not apply to the majority of real-world imaging scenarios currently used at nm scales. It is possible to implement randomized models such as low-pass coded diffraction patterns [8], for example by moving a large sand sheet in front of the sample [1]. In the x-ray regime, a similar approach uses a sheet of paper to produce random structure in the illumination beam [60]. When used in practical applications like ptychography, these WF approaches also suffer from slow convergence rates, albeit to a lesser extent.

In this paper, inspired by a seminal acceleration technique by Nesterov from the optimization literature [46] we present a new accelerated algorithm for phase retrieval called Accelerated Wirtinger Flow (AWF). While use of Nesterov's method has previous been proposed in the context of phase retrieval[83], our approach differs in the specification of a fixed step-size based on a Lipschitz-like constant. We derive novel theoretical results that prove the convergence of WF algorithms to stationary points for arbitrary phase-less measurements. These theoretical guidelines allows us to to eliminate computationally-expensive line-search algorithms, which means that WF, and by extension AWF, has low per-iteration computational complexity and has no algorithm parameters to tune. We note that recently a parallel line of research utilizes acceleration in PIE-based algorithms [41]. However, this approach diverges from Nesterov's formulation in several ways and still requires parameter tuning.

Using X-ray ptychography simulations, we observe that our algorithm exhibits much faster convergence when compared to traditional WF and other popular algorithms from the literature. Our results also show that AWF can achieve higher quality image reconstruction while being significantly more efficient in terms of computation cost, and also demonstrates resilience in scenarios with noise and device misalignment.

2   Phase retrieval

2.1   Generic Phase Retrieval

The phase retrieval problem arises in a variety of optical and X-ray imaging scenarios where the detector measures the intensity but not the phase of the diffraction field. The general version of the acquisition model can be written as

$$\mathbf{d} = |\mathbf{A}\mathbf{f}_*|^2 + \mathbf{n}, \tag{1}$$

where $\mathbf{d} \in \mathbb{R}^M$ is the vector of measured intensity values at the detector, $\mathbf{f}_* \in \mathbb{C}^N$ represents samples of the discretized complex-valued object function that we want to reconstruct, the matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ describes the propagation model for the optical system, and $\mathbf{n} \in \mathbb{R}^M$ represents noise perturbations. When applied to a matrix $\mathbf{A}$ or vector $\mathbf{a}$, we use the notation $|\mathbf{A}|$ or $|\mathbf{a}|$ to denote applying the absolute-value function to each entry elementwise, e.g., $[|\mathbf{A}|]_{mn} = |[\mathbf{A}]_{mn}|$ for each $m = 1, 2, \ldots, M$ and each $n = 1, 2, \ldots, N$, where $[\mathbf{A}]_{mn}$ is the entry from the mth row and nth column of $\mathbf{A}$.

2.2   Ptychographic Phase Retrieval

Ptychography [29, 51, 52, 53] is a coherent diffractive imaging method that leads to a special case of the generic phase retrieval problem. In this experiment, a sample is illuminated with several different illumination functions (or "probes") and corresponding diffraction patterns for each probe are measured by a detector in the far field. If $f(\mathbf{r})$ represents the 2D object function (corresponding to $\mathbf{f}$ in Eq. (1)) as a function of the spatial position $\mathbf{r} = (x, y)$ and $p_k(\mathbf{r})$ represents the kth probe function, then the complex field at the detector plane resulting from the kth probe is given by

$$b_k(\mathbf{r}') = \mathcal{F}\{p_k(\mathbf{r})f(\mathbf{r})\}, \tag{2}$$

where $\mathcal{F}\{f(\mathbf{r})\}$ denotes the Fourier transform of $f(\mathbf{r})$ with respect to $\mathbf{r}$. In many cases, the different probe functions $p_k(\mathbf{r})$ are obtained as different spatial shifts of the same basic probe function $p(\mathbf{r})$, i.e., $p_k(\mathbf{r}) = p(\mathbf{r} - \mathbf{r}_k)$ for some set of position vectors $\mathbf{r}_k$, $k = 1, 2, \ldots, K$. A schematic illustration of ptychograhy is shown in Fig. . 1.

If data is collected for K different probe functions, then the $\mathbf{A}$ matrix from Eq. (1) corresponding to a discretization of Eq. (2) can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{F} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{F} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{F} & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{\text{supp}(\mathbf{p}_1)}\text{diag}(\mathbf{p}_1) \\ \mathbf{I}_{\text{supp}(\mathbf{p}_2)}\text{diag}(\mathbf{p}_2) \\ \vdots \\ \mathbf{I}_{\text{supp}(\mathbf{p}_K)}\text{diag}(\mathbf{p}_K) \end{bmatrix}. \tag{3}$$

Here, $\mathbf{p}_k \in \mathbb{C}^N$ represents samples of the discretized complex-valued kth probe function, $\text{diag}(\mathbf{p}_k) \in \mathbb{C}^{N \times N}$ is a diagonal matrix with diagonal entries equal to the entries of $\mathbf{p}_k$. We assume $\mathbf{p}_k$ has contiguous support of size ℓ and $\mathbf{I}_{\text{supp}(\mathbf{p}_k)} \in \mathbb{R}^{\ell \times N}$ is a sub-matrix of the $N \times N$ identity matrix $\mathbf{I}$, indexed by the set of non-zero entries of $\mathbf{p}_k$ or $\text{supp}(\mathbf{p}_k)$. Finally, $\mathbf{F}$ is an $M/K \times \ell$ matrix modeling the 2D Fourier transform and detector sampling operation and it is assumed that each of the K diffraction patterns is measured at M/K spatial positions in the detector plane. Typically, in ptychography $\ell = M/K$ and $\mathbf{F}$ is an $M/K \times M/K$ DFT matrix.
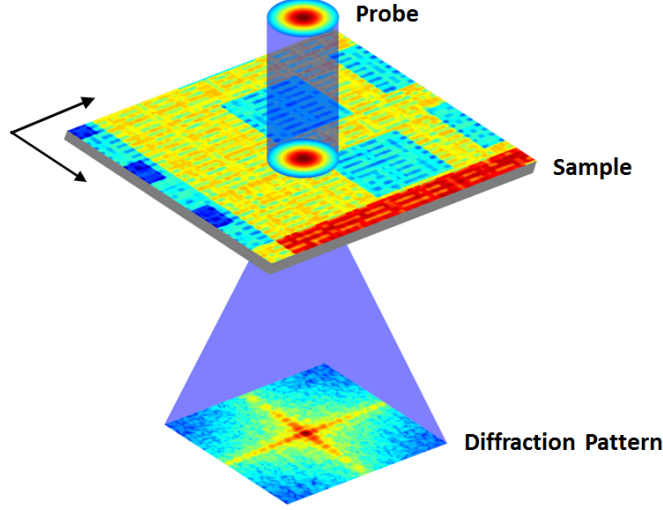
3

Figure 1: A schematic illustration of ptychography. A 2D sample is illuminated by a 2D Gaussian probe function, and a diffraction pattern is measured by a detector in the far field.

3   Reconstruction Algorithm: Accelerated Wirtinger Flow

3.1   Basic Formulation as an Optimization Problem

The AWF algorithm is designed to estimate $\mathbf{f}$ from Eq. (1) by solving an optimization problem of the form

$$\hat{\mathbf{f}} = \arg \min_{\mathbf{f} \in \mathbb{C}^N} \mathcal{L}(\mathbf{f}). \tag{4}$$

In this work, we choose $\mathcal{L} : \mathbb{C}^N \to \mathbb{R}$ to be the cost function

$$\mathcal{L}(\mathbf{f}) = \left\| \mathbf{b} - |\mathbf{A}\mathbf{f}| \right\|_2^2, \tag{5}$$

where $\mathbf{b}$ is obtained by taking the elementwise square-root of $\mathbf{d}$, i.e., $[\mathbf{b}]_m = ([\mathbf{d}]_m)^{1/2}$ for $m = 1, 2, \ldots, M$, and the ℓ_2-norm of the vector $\mathbf{a} \in \mathbb{C}^N$ is defined by $\|\mathbf{a}\|_2 = \left(\sum_{n=1}^N |[\mathbf{a}]_n|^2 \right)^{1/2}$. This choice of cost function, based on the amplitude rather than intensity of the diffraction pattern, can be derived as an approximation of the maximum likelihood (ML) cost function for a Poisson noise model [65], and can therefore be justified from the perspective of statistical estimation theory [39]. In addition, we and many other authors [65, 58, 79, 81, 71, 79] have observed that this choice of cost function leads to faster convergence behavior and better noise robustness compared to other cost functions (e.g., the ML cost function that would be obtained by assuming $\mathbf{n}$ in Eq. (1) follows a Gaussian distribution, or the unapproximated ML cost function for the Poisson distribution). Since this behavior is well established in the literature we focus on the cost (5) in the remainder of the paper and do not provide comparisons using other costs.

3.2 A Review of Wirtinger Flow and some New Results

The AWF algorithm is a generalization of the WF method [9]. As a result, this subsection reviews the WF method, and also presents some useful new theory that can be used to significantly reduce the computational cost associated with each iteration of WF. Later in the paper we will also make use of this computational improvement for AWF.

The optimization problem in Eq. (4) is nonlinear and generally does not have a simple closed-form solution. As a result, it is common to rely on iterative minimization methods. Gradient descent is one of the simplest and most natural iterative minimization algorithms. In this approach, starting from some initial guess $\hat{\mathbf{f}}_0$, the estimate of $\mathbf{f}$ is updated at the τth iteration according to

$$\hat{\mathbf{f}}_{\tau+1} = \hat{\mathbf{f}}_\tau - \mu_\tau \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau), \tag{6}$$

where μ_τ is the step-size for the τth iteration, and $\nabla \mathcal{L}(\hat{\mathbf{f}}_\tau)$ is the gradient of Eq. (5) with respect to $\mathbf{f}$, evaluated at the point $\hat{\mathbf{f}}_\tau$. Strictly speaking, Eq. (5) is not complex-differentiable. However, it is still possible to define a generalized gradient based on the notion of Wirtinger derivatives. Following [9], we shall refer to the iterations defined by Eq. (6) as WF.

The generalized gradient for Eq. (5) takes the form [9]

$$\nabla \mathcal{L}(\mathbf{f}) := \mathbf{A}^H \left(\mathbf{A}\mathbf{f} - \mathbf{b} \odot \mathrm{sgn}\left(\mathbf{A}\mathbf{f} \right) \right). \tag{7}$$

In this expression, we have used $\mathbf{A}^H$ to denote the conjugate transpose of the matrix $\mathbf{A}$, and have used $\odot$ to denote elementwise multiplication of two vectors (i.e., for vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{C}^N$, we have that $\mathbf{a} \odot \mathbf{b}$ is also a length-N vector defined by $[\mathbf{a} \odot \mathbf{b}]_n = [\mathbf{a}]_n [\mathbf{b}]_n$ for $n = 1, 2, \dots, N$). We have also introduced the complex signum function, which is defined for vectors $\mathbf{a} \in \mathbb{C}^N$ as the length-N vector that obeys

$$[\mathrm{sgn}(\mathbf{a})]_n = \begin{cases} 0, & \text{if } [\mathbf{a}]_n = 0 \\ \frac{[\mathbf{a}]_n}{|[\mathbf{a}]_n|}, & \text{otherwise,} \end{cases} \tag{8}$$

for $n = 1, 2, \dots, N$.

Given this expression for the gradient, it remains to choose the step size μ_τ to complete the specification of the WF algorithm. One of the most popular approaches to selecting μ_τ is to solve a simple one-dimensional "line-search" optimization problem [49]

$$\mu_\tau = \arg\min_{\mu \in \mathbb{R}} \mathcal{L}\left(\hat{\mathbf{f}}_\tau - \mu \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau) \right). \tag{9}$$

This choice is useful in the sense that it ensures the maximal possible decrease of the cost function along the gradient direction. On the other hand, solving this optimization problem can also be computationally expensive. Specifically in the case of ptychography with $\mathbf{A}$ defined in Eq. (3), each cost function evaluation requires the computation of K two-dimensional Fourier transform operations in addition to the $2K$ two-dimensional Fourier transform operations that are needed to compute the gradient vector. Consequently, solving Eq. (9) can be computationally expensive if a large number of candidate μ values need to be evaluated.

Here we make the novel observation that the Wirtinger algorithm will always converge if we set μ_τ to be the constant value

$$\mu_\tau = \bar{\mu} := \frac{1}{\lambda_{\max}(\mathbf{A}^H \mathbf{A})} \tag{10}$$

5

for all τ, where $\lambda_{\max}(\mathbf{A}^H\mathbf{A})$ is the largest eigenvalue of the positive semidefinite matrix $\mathbf{A}^H\mathbf{A}$. Importantly, this observation means that WF can be implemented without a line search. The choice of this stepsize is justified by Theorem A.1 in the Appendix. This theorem demonstrates that with our chosen step size the Wirtinger Flow iterates converge to a point where the generalized gradient is zero. This is a non-trivial statement as the loss function is non-smooth and there can be many stationary points where the generalized gradient does not vanish. The main use of this theorem in this paper is to justify our choice of step size in Eq. (10).

Computing $\bar{\mu}$ is particularly straightforward when $\mathbf{A}$ represents the ptychographic propagation model. Specifically, assuming that the $\mathbf{A}$ matrix has the form of Eq. (3), then $\mathbf{A}^H\mathbf{A}$ can be written as

$$\mathbf{A}^H\mathbf{A} = \sum_{k=1}^K \text{diag}(\mathbf{p}_k)^H \mathbf{I}_{\text{supp}(\mathbf{p}_K)}^H \mathbf{F}^H \mathbf{F} \mathbf{I}_{\text{supp}(\mathbf{p}_K)} \text{diag}(\mathbf{p}_k). \tag{11}$$

In practical ptychography contexts, the Fourier transform operator $\mathbf{F}$ is often chosen in such a way that $\mathbf{F}^H\mathbf{F} = \alpha\mathbf{I}$ for some constant α, where $\mathbf{I}$ is the $N \times N$ identity matrix. In such cases, the previous expression simplifies to

$$\mathbf{A}^H\mathbf{A} = \alpha \sum_{k=1}^K \text{diag}(\mathbf{p}_k)^H \mathbf{I}_{\text{supp}(\mathbf{p}_K)}^H \mathbf{I}_{\text{supp}(\mathbf{p}_K)} \text{diag}(\mathbf{p}_k) = \alpha \sum_{k=1}^K \text{diag}(\mathbf{p}_k)^H \text{diag}(\mathbf{p}_k). \tag{12}$$

The right-hand side of Eq. (12) is the sum of K diagonal matrices, and is therefore also diagonal. Since the eigenvalues of a diagonal matrix are equal to its diagonal entries, it is easy to see that $\lambda_{\max}(\mathbf{A}^H\mathbf{A})$ for this case is equal to the maximum entry of the easy-to-compute length-N vector $\alpha \sum_{k=1}^K |\mathbf{p}_k| \odot |\mathbf{p}_k|$.

For other choices of loss function in (4) (in lieu of (5)), it may be possible to solve the exact line search problem (9) in closed form. For instance, the cost function that minimizes the least-squares fit on intensity rather than amplitude values is a fourth order polynomial and hence (9) can be solved by cubic rooting [37, 72]. However, as mentioned previously we and other authors have observed that the amplitude-based loss (5) leads to faster convergence behavior and better noise robustness compared to other loss functions (even when exact line search is used). We are not aware of any closed form solutions to (9) for the cost function (5) that is the focus of this paper.

3.3 From Wirtinger Flow to Accelerated Wirtinger Flow

While the WF method described in the previous subsection often demonstrates faster convergence behavior than other classical phase retrieval schemes, the convergence rate can still be rather slow. To overcome this challenge, we replace the WF iteration from Eq. (6) with the following AWF update equation:

$$\hat{\mathbf{f}}_{\tau+1} = \hat{\mathbf{f}}_\tau + \beta_\tau \left(\hat{\mathbf{f}}_\tau - \hat{\mathbf{f}}_{\tau-1}\right) - \bar{\mu}\nabla\mathcal{L}\left(\hat{\mathbf{f}}_\tau + \beta_\tau \left(\hat{\mathbf{f}}_\tau - \hat{\mathbf{f}}_{\tau-1}\right)\right), \tag{13}$$

with $\bar{\mu}$ defined in Eq. (10), $\nabla\mathcal{L}(\mathbf{f})$ defined in Eq. (7), and β_τ defined by

$$\beta_\tau := \frac{\tau+1}{\tau+3}. \tag{14}$$

It should be noted that, aside from the choice of initialization $\hat{\mathbf{f}}_0$ and the choice of stopping criterion, the AWF algorithm is fully specified in Eq. (13). It has no tuning parameters and requires no line

6

searches. The computational cost of evaluating each iterative update using Eq. (13) is also low, and approximately the same as the computational cost of computing the gradient vector.

This AWF algorithm is inspired by the seminal work of Nesterov [46], who showed that if $\mathcal{L}(\mathbf{f})$ is smooth and convex, then the iterations defined by Eq. (13) converge at an optimal rate. However, the cost function in (5) is neither convex nor differentiable, and we have not proven that AWF will always converge to a minimizer of Eq. (5). Therefore, it may not be immediately clear that our proposed AWF scheme is useful, although many other authors have empirically observed that applying Nesterov-inspired algorithms to nonconvex cost functions can lead to outstanding results [25]. Our numerical evaluations below suggest that in practice, AWF often converges much more rapidly than both WF and classical phase retrieval algorithms.

We note that Zhou, Zhang, and Liang [83] also study accelerated phase retrieval using Nesterov's approach. However this paper differs from ours in terms of the problem setting and conclusions. In particular, this paper does not specify an appropriate step-size for ptychography, uses a different acceleration parameter, develops local convergence guarantees, and is focused on Gaussian designs in lieu of ptychography. We emphasize that utilizing acceleration for nonconvex optimization is of course well established. The novelty of our approach is three-fold: (1) specification of a step-size in terms of the Lipschitz-like constant for this problem, (2) demonstrating that such an approach works empirically for the practically relevant ptychography model without any need for tuning the step size, and (3) demonstrating that this approach can in some cases even escape shallow local minima.

4   Numerical Experiments

In this section, we investigate the performance of AWF in the context of ptychographic phase-less imaging. This work was motivated by x-ray ptychographic imaging of integrated circuits at resolution on the order of 10nm using a synchronton x-ray source[32], and we have designed our simulations in accordance with this application.

4.1   Simulated Ptychography Experimental Setup

We consider complex test images (to be described later) with 500×500 pixels (i.e., $N = 250,000$). To mimic the true illumination in a physical experiment, we use multiple shifts of the basic probe function $p(\mathbf{r})$, and set $p(\mathbf{r})$ to be a 2D Gaussian function with full width at half maximum (FWHM) of 30 pixels as shown in Fig. 2(a). The probe is set to zero outside of the central 78 pixel $\times$ 78 pixel region.

In the ideal noiseless case, this probe is shifted to $K = 313$ positions $\mathbf{r}_k$ based on a hexgonal scan pattern over a 250×250 portion at the center of the sample. The distance between two adjacent scan spots is 15 pixels, 50% of the FWHM. We have depicted the overall illumination pattern in Fig. 2(b) where the pixel values indicate how many times each pixel is illuminated. To prepare a stack of 313 diffraction patterns, we start from the first scan location, extract a 160×160 frame and multiply it by the probe centered on this frame. We then compute a 2D Discrete Fourier Transform (DFT) of the result. The use of 160 pixels in the frame determines the sample interval in Fourier space. The magnitude of the transform is recorded and saved before we move to the next scan location. In such a setting M/K =25,600 corresponding to $M = 8,012,800$.

In addition to the ideal noiseless case, we also considered practical scenarios with either Poisson

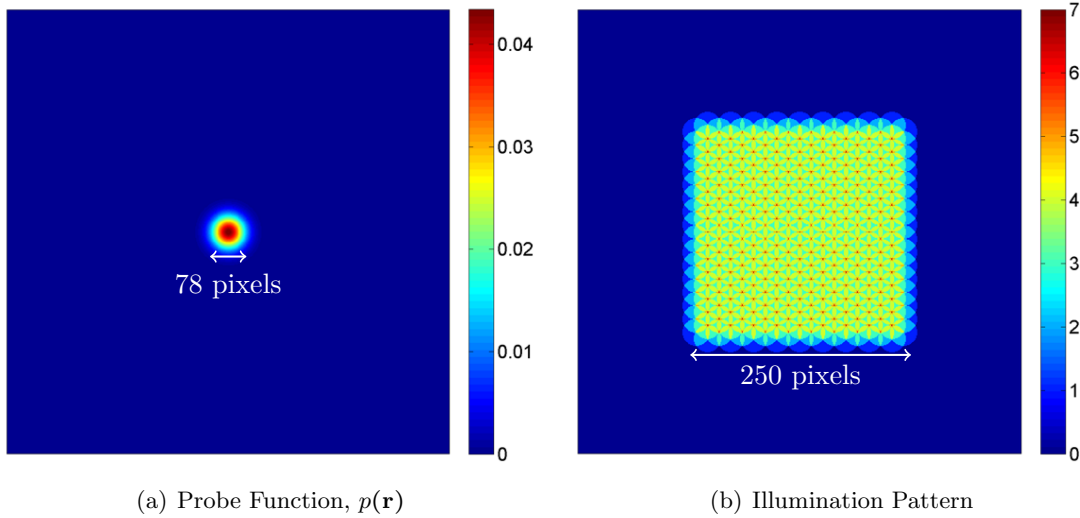(a) Probe Function, $p(\mathbf{r})$ (b) Illumination Pattern

Figure 2: Images depicting (a) the probe function and (b) the illumination pattern used in the nominal simulated ptychography experiments.

noise resulting from a limited number of photons or probe position misalignment introduced by imperfections in either the beam or sample stage positioning hardware. We simulate psuedo-Poisson noise based on the assumption that either 1×10^9 or 1×10^7 photons are delivered to the sample for each diffraction pattern. To model misalignment, we used a random displacement for each $\mathbf{r}_k$ probe position value when generating data (but used the nominal $\mathbf{r}_k$ values in reconstruction). To do this we generated random independent displacement values along x and y by drawing uniformly from the set of possible displacements $\{-1, 0, 1\}$ (in units of voxels) in each direction for each probe position.

4.2   Description of Test Images

Reconstructions were performed using two different test images. one that we numerically simulated based on the structure of an integrated circuit (IC), and the other a popular ptychography test image of gold beads.

The IC test image is generated from the projection of a synthetic IC structure measuring 2.5 μm × 2.5 μm × 4.4 μm with a voxel size of 5 nm, corresponding to 500 × 500 × 880 voxels. This IC structure contains silicon dioxide (SiO_2), silicon (Si), aluminum (Al), tungsten (W), and copper (Cu), and these materials are modeled as complex dielectrics discretized in space, described by complex $n_{index} = 1 - \delta + i\beta$ that depends on the illumination energy of the x-ray probe; δ is the refractive index decrement responsible for shifting the probe phase, while β is the imaginary part that describes attenuation of the probe amplitude by the material. We use a simulated x-ray source with an energy of 6.2 keV ($\lambda_0 = 0.2$ nm) to simulate operating conditions of recent experiments [21, 69, 31]. The test image, Fig. 3, is computed as the complex projection through the chip along the z-axis:

$$T(x,y) = e^{\int_0^{\Delta z} ik_0 n_{index}(x,y,z;k_0)dz},$$ (15)

where $k_0 = 2\pi/\lambda_0$ is the x-ray wavenumber and the integral is calculated over the chip thickness $\Delta z = 4.4$ μm, resulting in a 500×500 complex-valued image. At an x-ray energy of 6.2 keV, Si has values $\delta = 1.285 \times 10^{-5}$ and $\beta = 4.777 \times 10^{-7}$; Al has values $\delt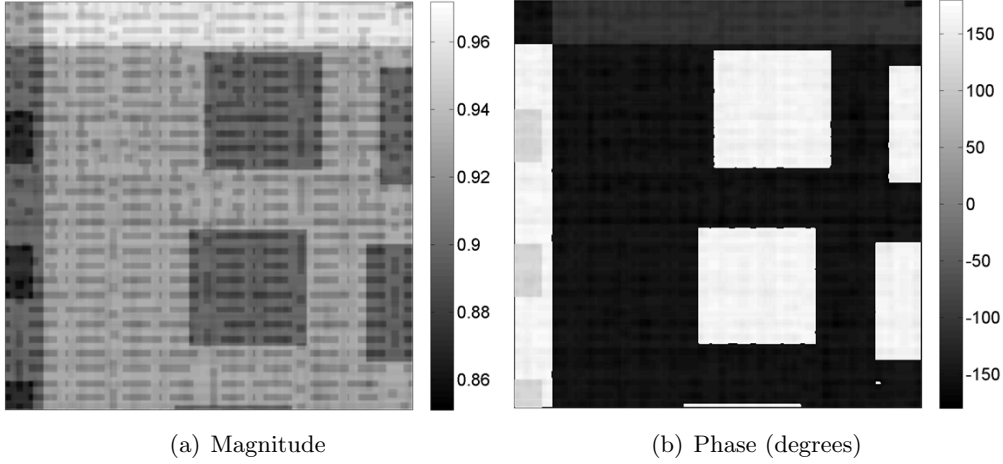a = 1.436 \times 10^{-5}$ and $\beta = 4.305 \times 10^{-7}$; W has values $\delta = 7.971 \times 10^{-5}$ and $\beta = 9.840 \times 10^{-6}$; Cu has values $\delta = 4.264 \times 10^{-5}$ and $\beta = 1.479 \times 10^{-6}$; and SiO$_2$ has values $\delta = 1.206 \times 10^{-5}$ and $\beta = 2.574 \times 10^{-7}$.

(a) Magnitude   (b) Phase (degrees)

Figure 3: Magnitude and phase of the integrated circuit test image.

The second test image is based on the transmission function of a collection of gold beads deposited on a membrane [78]. The magnitude and phase of this test image are depicted in Fig. 4.
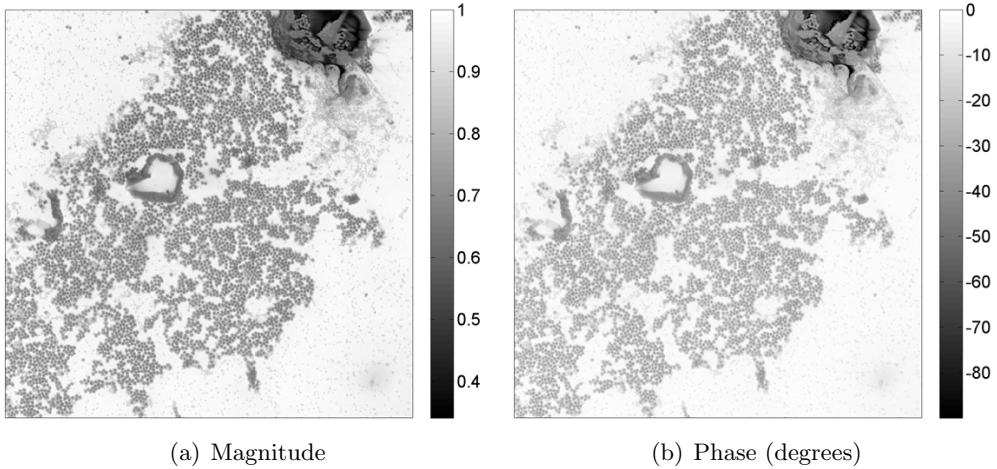

(a) Magnitude   (b) Phase (degrees)

Figure 4: Magnitude and phase of the gold bead test image.

4.3 Comparison of Ptychography Algorithms.

Images were reconstructed using our novel AWF algorithm and several previously-proposed ptychography algorithms for comparison. Specifically, we implemented WF [9], the Polak-Ribiere form

9

of the nonlinear Conjugate Gradient Method (CGM) [65, 78, 49], DM [63, 64], ePIE [42], RAAR [43], and ER [78]. DM, ePIE, RAAR and ER are alternating projections algorithms that do not use the cost function from Eq. (5), while WF, AWF, and CGM are specifically designed to minimize Eq. (5). To ensure a fair comparison, all seven algorithms are initialized in exactly the same way by setting $[\hat{\mathbf{f}}_0]_n = 1$ for all $n = 1, \ldots, N$.

Our WF and CGM implementations incorporate a line-search procedure to pick the step-size μ_τ as in Eq. (9). Our implementation of this line search uses MATLAB's default line search algorithm (golden section search with parabolic interpolation [23]).

Our primary measure of the quality of the reconstructed images uses the normalized root mean-squared error between the ground truth and reconstructed images after compensating for phase ambiguities and masking the central portion of the image:

$$\text{Relative reconstruction error} = \min_{\phi \in [0, 2\pi]} \frac{\left\| \mathbf{M} \left(e^{i\phi} \hat{\mathbf{f}}_\tau - \mathbf{f}_* \right) \right\|_2}{\| \mathbf{M} \mathbf{f}_* \|_2}. \tag{16}$$

The diagonal matrix $\mathbf{M} \in \mathbb{R}^{N/4 \times N}$ has the effect of extracting the central 250×250 pixel region of the reconstructed image and discarding the other regions, and is used because the other image regions received very little illumination and are not expected to be reconstructed accurately. Our error metric has a phase correction factor ϕ because we only have magnitude measurements, which means that $\mathcal{L}(\hat{\mathbf{f}}) = \mathcal{L}(e^{i\phi}\hat{\mathbf{f}})$ for any choice of ϕ. Therefore, the solution to Eq. (4) is almost never unique, and it is only possible to recover the original signal/image up to a global phase factor.

To investigate the convergence behavior of different algorithms, we compute and plot the error metric from Eq. (16) for each iteration up to $\tau = 1000$. This measure ignores variations in per-iteration computation time among algorithms. For example, multiple iterations of a line search can result in significantly higher costs per iteration. True computational cost is highly dependent on the specifics of each implementation and the computing platform. To avoid a potentially unfair comparison, we therefore also plot the error metric as a function of the number of FFTs used. This serves as a surrogate for computation cost since the measure is robust to implementational details and FFTs are the dominant computational burden in ptychographic image formation. We also investigate the convergence of the cost function itself. In this case we only plot the behavior for three of the algorithms since only WF, AWF and CGM explicitly optimize the cost function in Eqs. 4 and 5. As with the image error metric, we plot the change in cost function as a function of number of iterations and number of FFTs.

In most ptychographic applications the probe may not be known exactly (although in many cases an accurate estimate of the probe is available or can be estimated directly from data). To investigate the AWF algorithm when the probe is not known exactly we compare its performance to the e-PIE algorithm. To run AWF when the probe is not known we first initialize the probe (as discussed next) and then in each iteration perform the following two updates: (1) a signal updates using the AWF iterates assuming a known probe and (2) a probe update using e-PIE's probe update strategy. We investigate two initialization scenarios: (1) a *deterministic* initialization where we use the inverse Fourier transform of the average of the diffraction patterns and (2) a *random perturbation* where we add a small amount of Gaussian noise to the original probe and use that as our initialization. In both cases we normalized the initial estimate of the probe so it has total energy one (the sum of squares of the absolute value of entries is set to one). The former method is a standard practical initialization. While the latter initialization is not achievable in practice, the goal of studying this initialization is to demonstrate the performance of the algorithms in
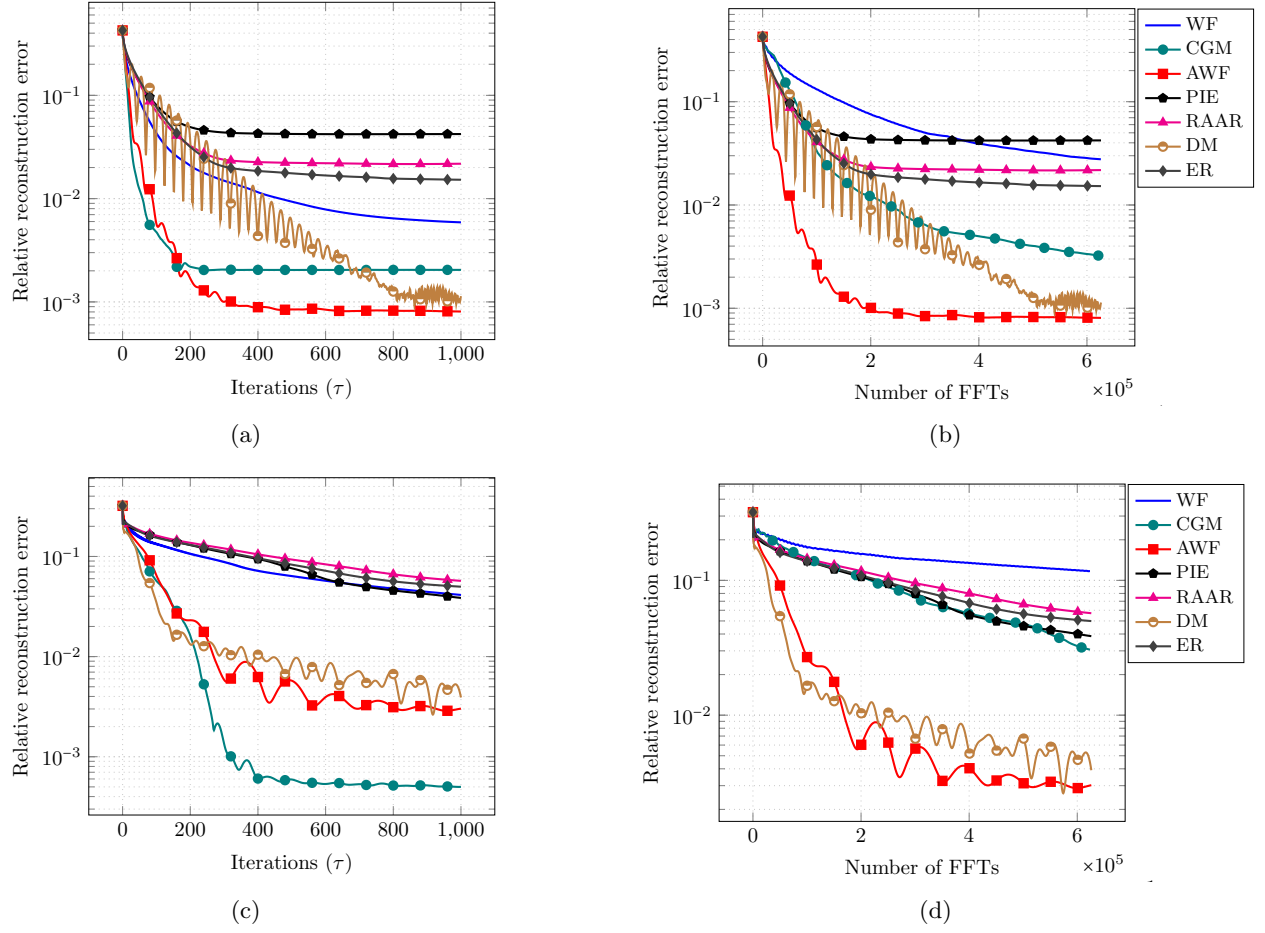
Figure 5: Convergence of the relative reconstruction error (Eq. 16) for different algorithms applied to noiseless data from: (a,b) IC and (c,d) Gold Bead test images, plotted as a function of (a,c) iteration number and (b,d) number of FFT computations.

the presence of small and random uncertainty in the probe. We evaluate the performance of the algorithms using the following error metric

$$\text{Relative reconstruction error} = \min_{c \in \mathbb{C}} \frac{\left\| \mathbf{M} \left(c\hat{\mathbf{f}}_\tau - \mathbf{f}_* \right) \right\|_2}{\left\| \mathbf{M}\mathbf{f}_* \right\|_2}. \tag{17}$$

This is analogous to (17). The only difference is that we use a complex scalar $c \in \mathbb{C}$ instead of the global phase factor $e^{i\phi}$ to account for the inherent ambiguity in the scale of the image when the probe is not known since the measurements are formed from the product of these two unknown quantities.

4.4   Numerical Results

Fig. 5 shows convergence behavior of the relative reconstruction error (RRE) (Eq. 16) in the noiseless case for the IC and Gold Bead test images. The plots on the left show the change in RRE vs number of iterations. Those on the right are individually rescaled in the horizontal axis to reflect number of FFT computations. Note that the curves on the right show behavior for different maximum numbers of iterations for some of the methods, depending on the number of FFTs required per iteration. There are two interesting aspects to these plots: (a) the convergence behavior varies markedly across methods and (b) there are substantial differences across methods in the RRE even after 1,000 iterations. This latter observation reflects, in part, differences in convergence rate, but also the fact that not all are converging towards the same solution. Since ER, DM, RAAR and PIE are not explicitly optimizing a cost function it is not suprising that they would not converge to the same solution. However, there are also significant differences in RRE for WF, AWF and CGM, all of which optimize the same cost function. In that case, the differences may be largely attributable to different convergence rates, although it is also possible that, since the cost function is not convex, they may be converging towards different solutions.

The AWF method shows substantially faster convergence in RRE vs. iteration than simple WF, even though the former uses a fixed step size while the latter uses a line search. For this reason, the differences also appear even larger when RRE is plotted against number of FFTs. These differences reflect the well known accelaration associated with Nesterov-like iterations compared to steepest descent. Conjugate gradient also exhibits fast convergence relative to most other methods, in fact even faster than AWF in the case of the gold bead test image, Fig. 5(c). However, when we account for the additional cost of the line search by plotting convergence vs. number of FFTs, CGM becomes relatively slow leaving AWF as the most rapidly converging algorithm. Note also that DM exhibits rapid convergence in these noiseless data. However, as we see below, DM becomes unstable in the presense of noisy or inconsistent data.

Fig. 6 shows the behavior of WF, AWF and CGM in terms of the cost function rather than the image error. Similar observations can be made with respect to their relative behavior. Although convergence rate for CGM can be fast as a function of iteration number, Fig. 6(a,c), once the additional costs of the line search is taken into account, AWF is substantially faster Fig. 6(b,d).

Fig. 7 shows convergence behavior in the noisy case (10^9 photons per probe position) for the IC and Gold Bead test images while Fig. 10 shows behavior when the data are misaligned. Figs. 8 and 9 show the 256 x 256 pixel central portion of the magnitude and phase of the ground truth images as well as those of the reconstructed images along with the absolute value of the magnitude and phase difference images (phase adjusted reconstructed images minus ground truth image)
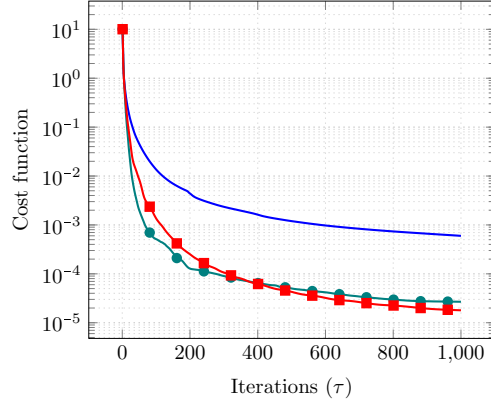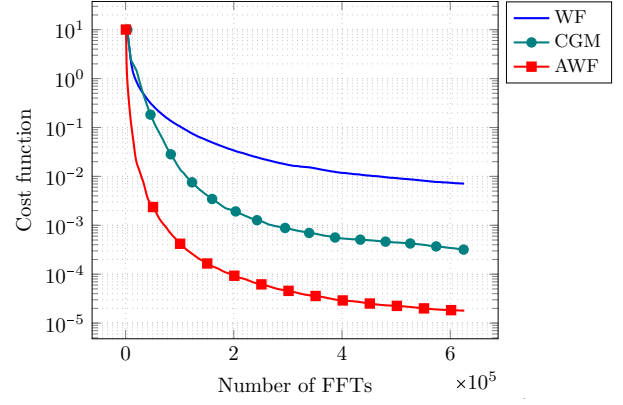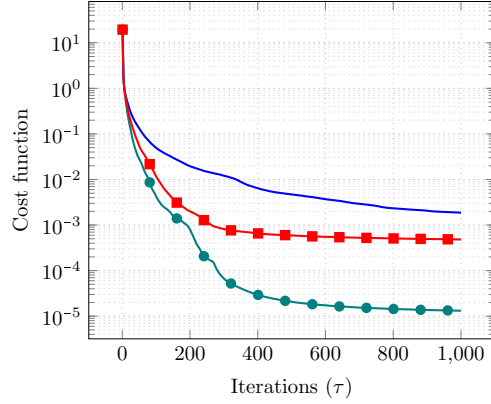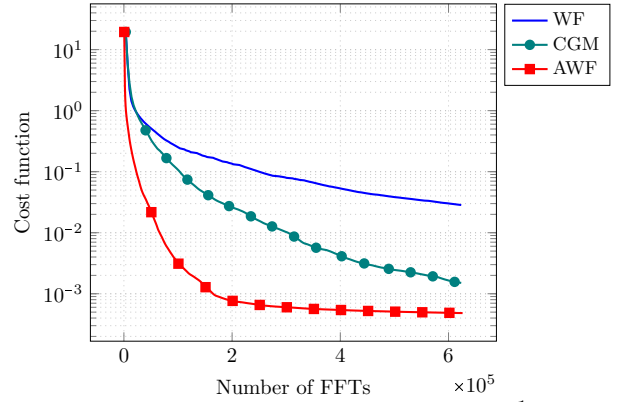
Figure 6: Convergence of cost function in Eqs. 4and 5 for CGM, WF and AWF algorithms for noiseless data from (a,b) IC and (c,d) Gold Beads test images as a function of (a,c) iterations and (b,d) FFT computations.

for the case of 10^7 photons per probe position. We observe that most algorithms demonstrate similar characteristics to those that were seen in the noiseless case. The major exception is the DM algorithm which exhibits quite unstable behavior in both cases. Consistently in these results, AWF and CGM exhibit the fastest convergence relative to iteration number. With the benefit of fewer FFTs required per iteration for AWF relative to CGM, the former shows substantially faster convergence when plotted against number of FFTs. Ther reconstructed images are visually similar for most cases, although errors are significant for DM. Overall, errors are larger in the magnitude than the phase component. Errors are consistently smaller for the optimization based methods (WF, CGM and AWF) than for PIE, RAAR, DM and ER. Interestingly, in some instances the errors in Fig. 7 are smaller than for the noiseless case, Fig. 5. We note that because the cost functions is non-convex these algorithms can become trapped in local optima and it is possible that a small amount of additional noise in the data may result in convergence to a different minimum that represents a better solution.

Fig. 11 shows convergence behavior of AWF and e-PIE with probe updates at each iteration when they are both initialized via the random perturbation model for the noisy case (10^9 photons per probe position) for the IC and Gold Bead test images. Fig. 12 shows this convergence behavior when both algorithms are initialized via the deterministic initialization described in the previous section. We observe that in both cases, while errors are larger than in the case when the probe is known exactly, AWF converges faster to a smaller error compared to e-PIE.

5   Discusion and Conclusions

We have described the novel AWF algorithm for phase retrieval and demonstrated it on simulated ptychography data. AWF combines the Wirtinger flow approach to phase retrieval with Nesterov's momentum method for accelerated gradient descent. The primary advantage of AWF over alternative algorithms is that it explicitly optimizes a cost function without the need for a line search. We show that use of an easily computed Lipschitz-like constant ensures convergence of WF with a fixed step-size and also results in effective convergence in practical cases for AWF. Furthermore, computational results show favorable performance of AWF in terms of convergence rate, not only relative to WF and CGM (both of which optimize the same cost function) but also in comparison to other widely used ptychography methods.

Interestingly, although the visible diffferences in the final images were relatively small, we do see significant differences in terms of residual errors, even after 1000 iterations of each algorithm, with either CGM or AWF consistently achieving the smallest error. Both of these fast algorithms are optimizing the same cost function so the similarity in performance when compared as a function of iteration is perhaps not surprising. When also accounting for the higher complexity in CGM as a result of the line-search, AWF exhibits the best combination of residual error and computation cost. WF also optimizes the same cost function but convergence is far slower. The other methods tested converge slower than AWF and use a range of alternation schemes which do not explicitly optimize a cost function and can result in differences in the final images with residual errors larger than AWF.

Most of the simulation studies presented here assume that the probe is known exactly. However, we also show examples where we update the probe estimate at each iteration using an e-PIE like procedure. These results shows that AWF retains its faster convergence behavior relative to e-PIE. These comparisons are preliminary in nature and are meant as a baseline comparison of AWF with
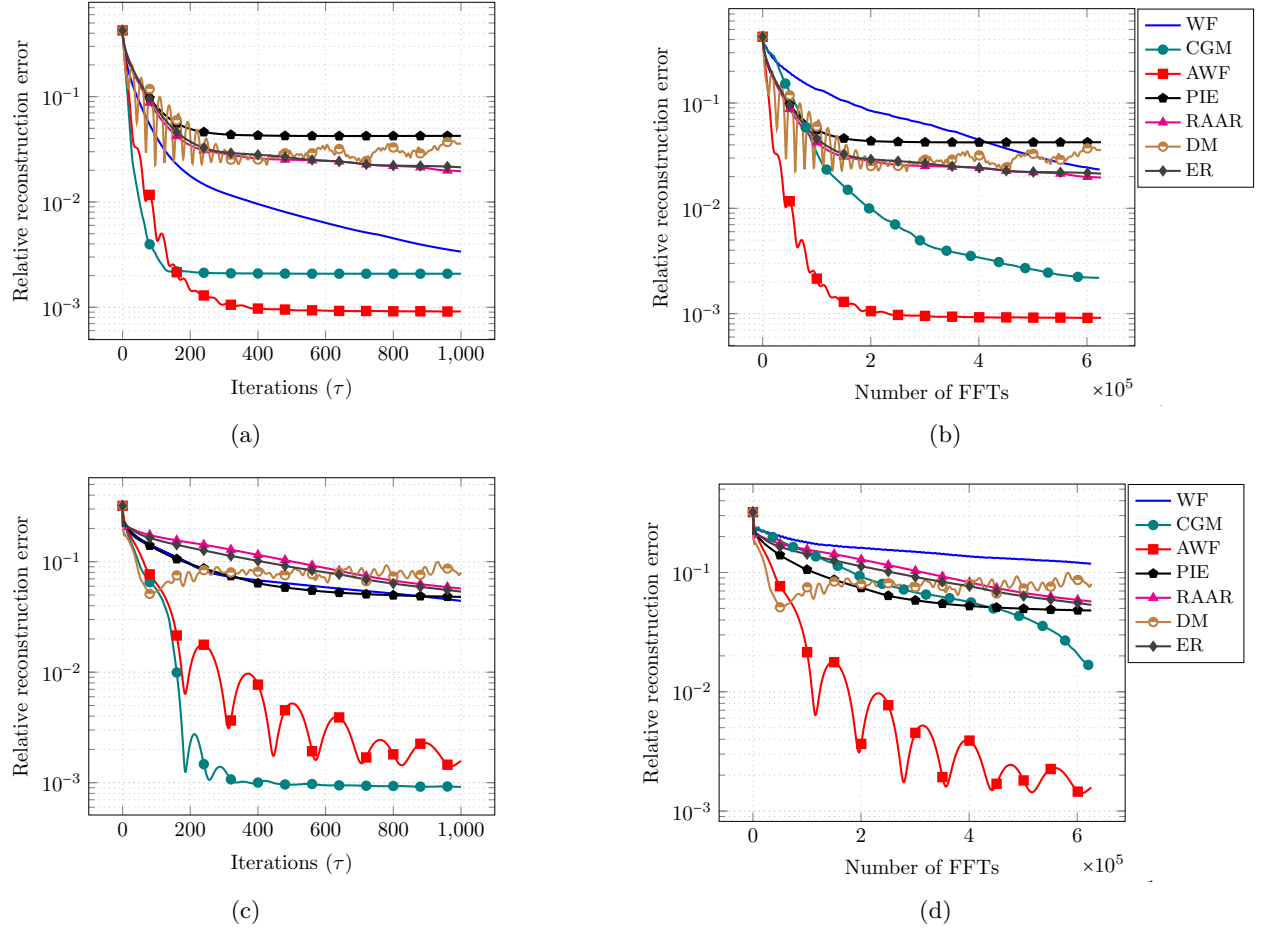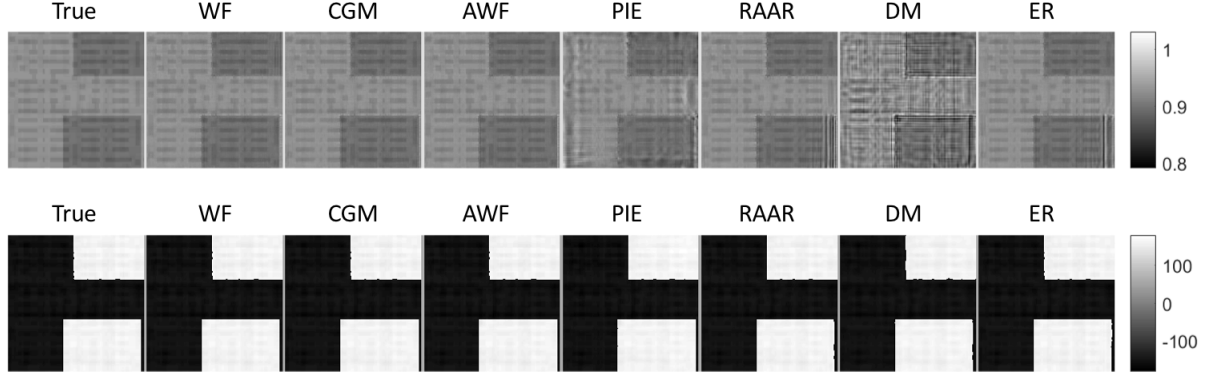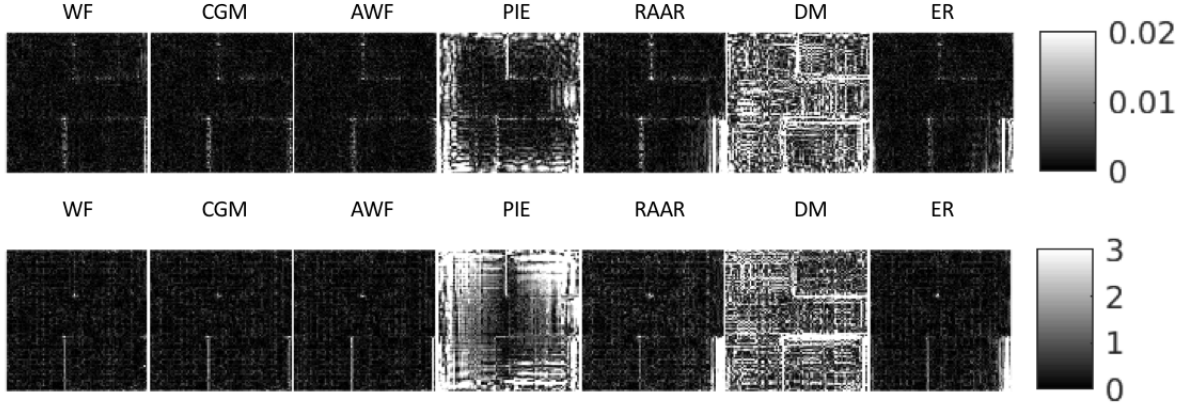
Figure 7: Convergence of the relative reconstruction error (Eq. 16) for different algorithms applied to data with Poisson noise from: (a,b) IC and (c,d) Gold Bead test images, plotted as a function of (a,c) iteration number and (b,d) number of FFT computations.
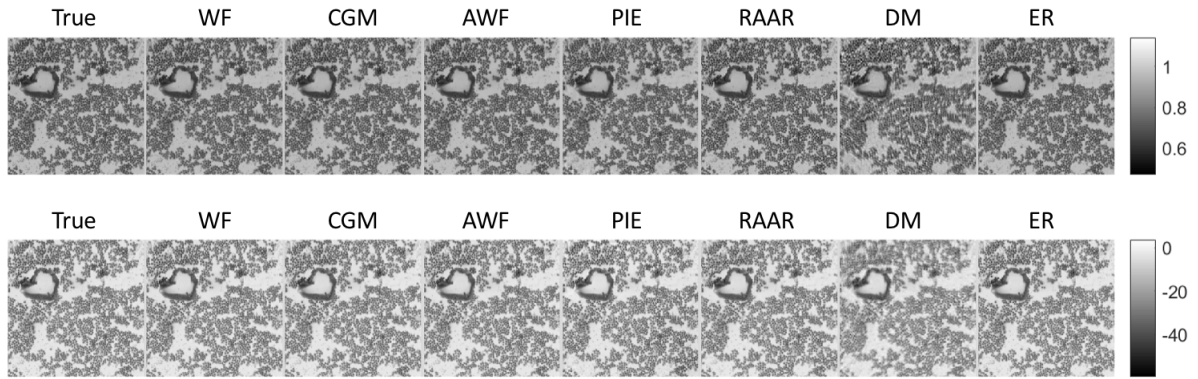
Figure 8: Magnitude (upper) and phase (lower) of (a) the ground truth and reconstructed images from different algorithms (b) the absolulute value of the magnitude and phase difference images (phase adjusted reconstructed image minus ground truth) for IC sample from Poisson data, 10^7 photons per probe position.
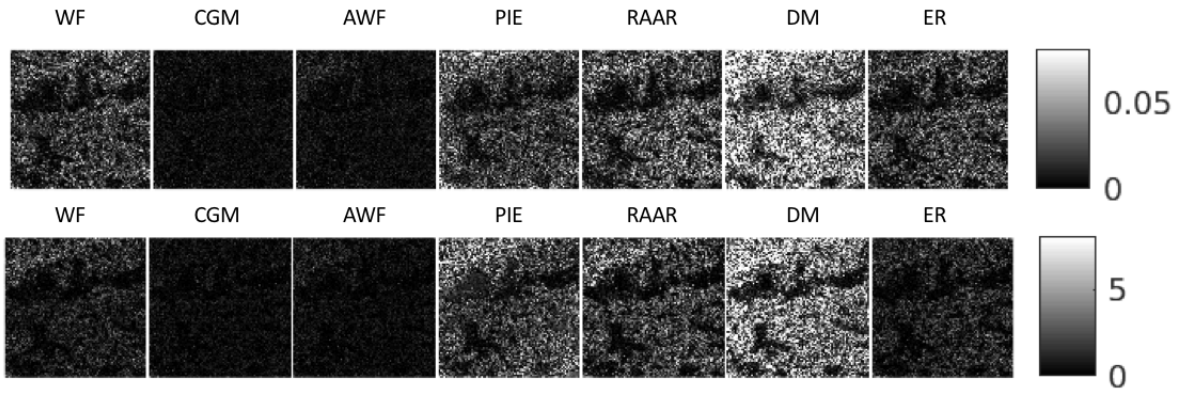
16

Figure 9: Magnitude (upper) and phase (lower) of (a) the ground truth and reconstructed images from different algorithms (b) the absolulute value of the magnitude and phase difference images (phase adjusted reconstructed image minus ground truth) for the golden bead sample from Poisson data, 10^7 photons per probe position.
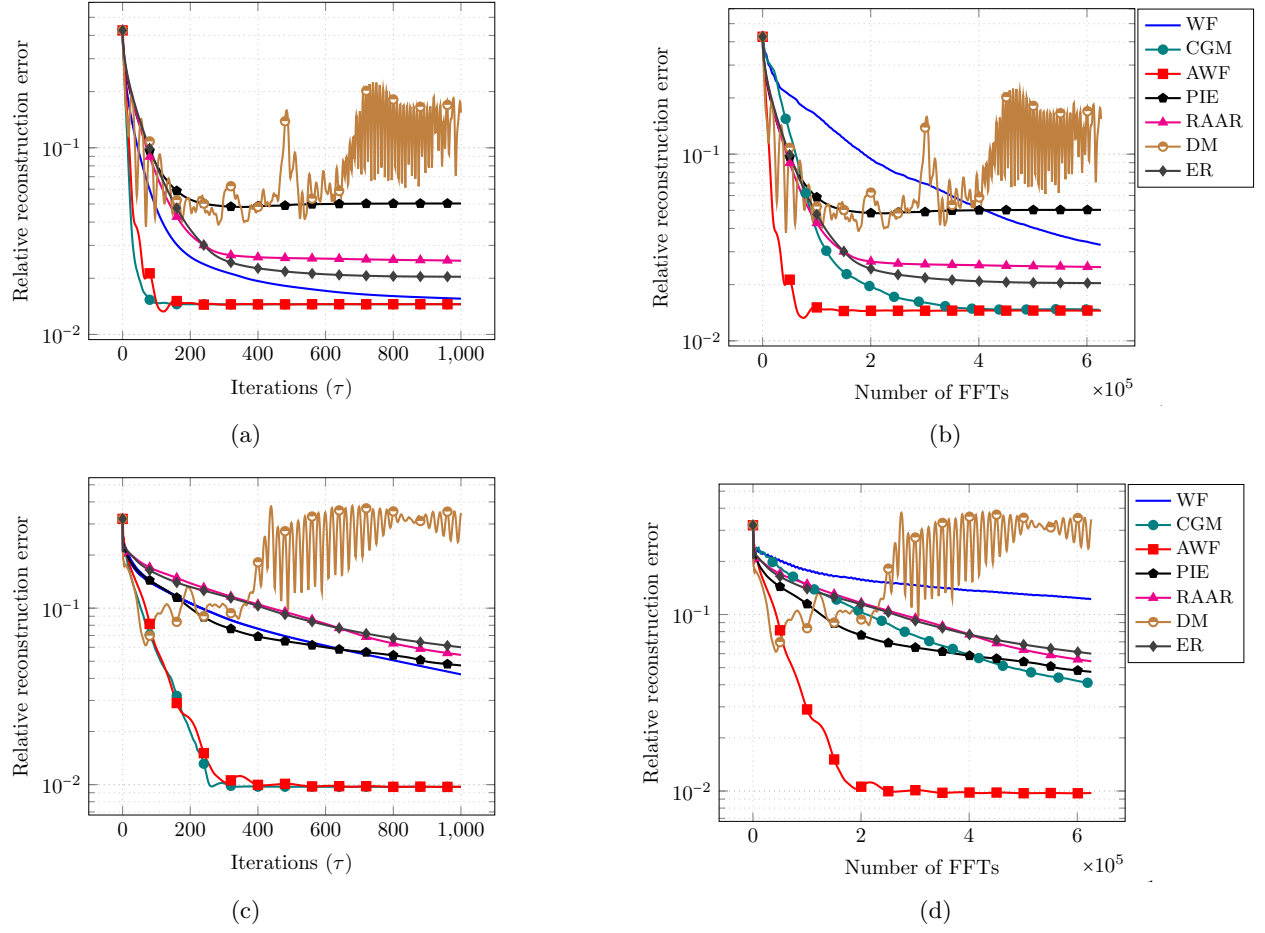
Figure 10: Convergence of the relative reconstruction error (Eq. 16) for different algorithms applied to misaligned data from: (a,b) IC and (c,d) Gold Bead test images, plotted as a function of (a,c) iteration number and (b,d) number of FFT computations.

Figure 11: Convergence of the relative reconstruction error (Eq. 17) for e-PIE and AWF (using e-PIE's probe update applied to (a) IC and (b) Gold Bead test images, ploted as a function of iteration number. In this plot, the initial probe is set to a random Gaussian perturbation of the probe for both algorithms.
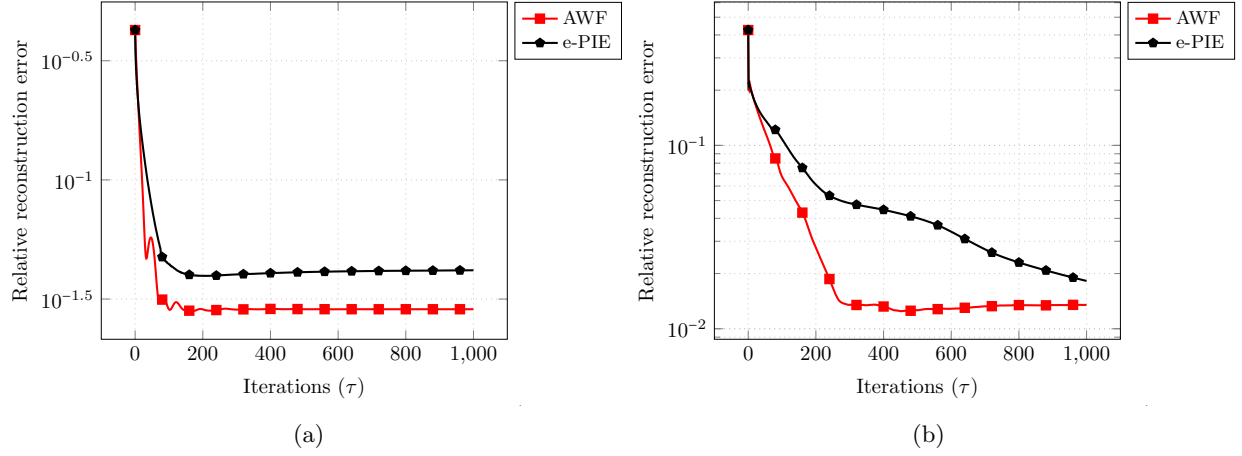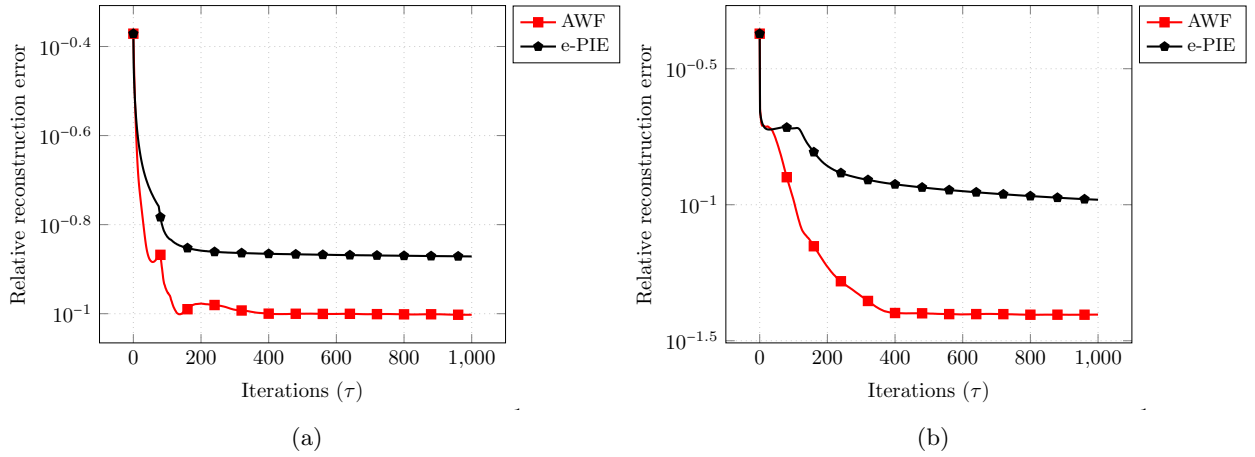

Figure 12: Convergence of the relative reconstruction error (Eq. 17) for e-PIE and AWF (using e-PIE's probe update applied to (a) IC and (b) Gold Bead test images, ploted as a function of iteration number. In this plot, the initial probe is set to the inverse Fourier transform of the average of the diffraction patterns for both algorithms.

19

a well known approach. We believe that eventually one may be able to develop an algorithm where the probe and signal are updated simultaneously using accelerated gradient updates. It may also be possible to develop rigorous theory for convergence to stationary points with a carefully developed step-size in this case. Such investigations are an interesting direction for future research

Acknowledgements

References

[1] *Personal communication Stefano Marchesini and Andreas Menzel.* August 2017.

[2] B. Abbey, K. A. Nugent, G. J. Williams, J. N. Clark, A. G. Peele, M. A. Pfeifer, M. De Jonge, and I. McNulty. Keyhole coherent diffractive imaging. *Nature Physics*, 4(5):394, 2008.

[3] B. Alexeev, A. S. Bandeira, M. Fickus, and D. G. Mixon. Phase retrieval with polarization. *SIAM Journal on Imaging Sciences*, 7(1):35–66, 2014.

[4] S. Bahmani and J. Romberg. Phase retrieval meets statistical learning theory: A flexible convex relaxation. *arXiv preprint arXiv:1610.04210*, 2016.

[5] T. Bendory, Y. C. Eldar, and N. Boumal. Non-convex phase retrieval from STFT measurements. *IEEE Transactions on Information Theory*, 64(1):467–484, 2018.

[6] T. T. Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *arXiv preprint arXiv:1506.03382*, 2015.

[7] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski. Phase retrieval via matrix completion. *SIAM Journal on Imaging Sciences*, 6(1):199–225, 2013.

[8] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 2014.

[9] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[10] E. J. Candes, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[11] R. Chandra, C. Studer, and T. Goldstein. Phasepack: A phase retrieval library. *arXiv preprint arXiv:1711.10175*, 2017.

[12] H. Chang, P. Enfedaque, Y. Lou, and S. Marchesini. Partially coherent ptychography by gradient decomposition of the probe. *Acta Crystallographica Section A: Foundations and Advances*, 2018.

[13] H. N. Chapman, A. Barty, S. Marchesini, A. Noy, S. P. Hau-Riege, C. Cui, M. R. Howells, R. Rosen, H. He, J. C. H. Spence, et al. High-resolution ab initio three-dimensional x-ray diffraction microscopy. *JOSA A*, 23(5):1179–1200, 2006.

[14] Y. Chen and E. Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pages 739–747, 2015.

[15] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *arXiv preprint arXiv:1803.07726*, 2018.

[16] J. Chung, X. Ou, R. P. Kulkarni, and C. Yang. Counting white blood cells from a blood smear using Fourier ptychographic microscopy. *PloS ONE*, 10(7):e0133489, 2015.

[17] J. N. Clark, L. Beitra, G. Xiong, A. Higginbotham, D. M. Fritz, H. T. Lemke, D. Zhu, M. Chollet, G. J. Williams, M. Messerschmidt, et al. Ultrafast three-dimensional imaging of lattice dynamics in individual gold nanocrystals. *Science*, 341(6141):56–59, 2013.

[18] D. Damek, D. Dmitriy, and P. Courtney. The nonsmooth landscape of phase retrieval. *arxiv:1711.03247*, 2017.

[19] J. Deng, D. J. Vine, S. Chen, Y. S. G. Nashed, Q. Jin, N. W. Phillips, T. Peterka, R. Ross, S. Vogt, and C. J. Jacobsen. Simultaneous cryo x-ray ptychographic and fluorescence microscopy of green algae. *Proceedings of the National Academy of Sciences*, 112(8):2314–2319, 2015.

[20] O. Dhifallah, C. Thrampoulidis, and Y. M. Lu. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *arXiv preprint arXiv:1805.09555*, 2018.

[21] M. Dierolf, A. Menzel, P. Thibault, P. Schneider, C. M. Kewish, R. Wepf, O. Bunk, and F. Pfeiffer. Ptychographic X-ray computed tomography at the nanoscale. *Nature*, 467(7314):436, 2010.

[22] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv preprint arXiv:1705.02356*, 2017.

[23] G. E. Forsythe, C. B. Moler, and M. A. Malcolm. *Computer Methods for Mathematical Computations.* Prentice-Hall, 1977.

[24] Marcus Gallagher-Jones, Yoshitaka Bessho, Sunam Kim, Jaehyun Park, Sangsoo Kim, Daewoong Nam, Chan Kim, Yoonhee Kim, Osamu Miyashita, Florence Tama, et al. Macromolecular structures probed by combining single-shot free-electron laser diffraction with synchrotron coherent x-ray imaging. *Nature Communications*, 5:3798, 2014.

[25] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 2016.

[26] R. Ghods, A. S. Lan, T. Goldstein, and C. Studer. Phaselin: Linear phase retrieval. In *Information Sciences and Systems (CISS), 2018 52nd Annual Conference on*, pages 1–6. IEEE, 2018.

[27] T. Goldstein and C. Studer. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 2018.

[28] J. Halyun and G. C. Sinan. Convergence of the randomized kaczmarz method for phase retrieval. *arXiv preprint arXiv:1706.10291*, 2017.

[29] R. Hegerl and W. Hoppe. Dynamische theorie der kristallstrukturanalyse durch elektronenbeugung im inhomogenen primarstrahlwellenfeld. *Berichte der Bunsengesellschaft für physikalische Chemie*, 74(11):1148–1154, 1970.

[30] R. Hesse, D. R. Luke, S. Sabach, and M. K. Tam. Proximal heterogeneous block implicit-explicit method and application to blind ptychographic diffraction imaging. *SIAM J. Imaging Sciences*, 8(1):426–457, 2015.

[31] M. Holler, A. Diaz, M. Guizar-Sicairos, P. Karvinen, E. Farm, E. Harkonen, M. Ritala, A. Menzel, J. Raabe, and O. Bunk. X-ray ptychographic computed tomography at 16 nm isotropic 3D resolution. *Scientific Reports*, 4:3857, 2014.

[32] M. Holler, M. Guizar-Sicairos, E. H. R. Tsai, R. Dinapoli, E. Müller, O. Bunk, J. Raabe, and G. Aeppli. High-resolution non-destructive three-dimensional imaging of integrated circuits. *Nature*, 543(7645):402–406, 2017.

[33] R. Horstmeyer, X. Ou, G. Zheng, P. Willems, and C. Yang. Digital pathology with Fourier ptychography. *Computerized Medical Imaging and Graphics*, 42:38–43, 2015.

[34] M. A. Iwen, B. Preskitt, R. Saab, and A. Viswanathan. Phase retrieval from local measurements: Improved robustness via eigenvector-based angular synchronization. *arXiv preprint arXiv:1612.01182*, 2016.

[35] K. Jaganathan, Y. C. Eldar, and B. Hassibi. STFT phase retrieval: Uniqueness guarantees and recovery algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):770–781, 2016.

[36] H. Jiang, C. Song, C. C. Chen, R. Xu, K. S. Raines, B. P. Fahimian, C. H. Lu, T. K. Lee, A. Nakashima, J. Urano, et al. Quantitative 3D imaging of whole, unstained cells by using x-ray diffraction microscopy. *Proceedings of the National Academy of Sciences*, 107(25):11234–11239, 2010.

[37] X. Jiang, S. Rajan, and X. Liu. Wirtinger flow method with optimal stepsize for phase retrieval. *IEEE Signal Processing Letters*, 23(11):1627–1631, 2016.

[38] Z. Jingshan, R. A. Claus, J. Dauwels, L. Tian, and L. Waller. Transport of intensity phase imaging by intensity spectrum fitting of exponentially spaced defocus planes. *Optics Express*, 22(9):10661–10674, 2014.

[39] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, Upper Saddle River, 1993.

[40] J. Ma, J. Xu, and A. Maleki. Optimization-based AMP for phase retrieval: The impact of initialization and ℓ_2-regularization. *arXiv preprint arXiv:1801.01170*, 2018.

[41] A. Maiden, D. Johnson, and P. Li. Further improvements to the ptychographical iterative engine. *Optica*, 4(7):736–745, 2017.

[42] A. M. Maiden and J. M. Rodenburg. An improved ptychographical phase retrieval algorithm for diffractive imaging. *Ultramicroscopy*, 109(10):1256–1262, 2009.

[43] S. Marchesini, H. Krishnan, B. J. Daurer, D. A. Shapiro, T. Perciano, J. A. Sethian, and F. R. N. C. Maia. SHARP: a distributed GPU-based ptychographic solver. *Journal of Applied Crystallography*, 49(4):1245–1252, 2016.

[44] J. Miao, P. Charalambous, J. Kirz, and D. Sayre. Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. *Nature*, 400:342–344, 1999.

[45] J. Nelson, X. Huang, J. Steinbrener, D. Shapiro, J. Kirz, S. Marchesini, A. M. Neiman, J. J. Turner, and C. Jacobsen. High-resolution x-ray diffraction microscopy of specifically labeled yeast cells. *Proceedings of the National Academy of Sciences*, 107(16):7235–7239, 2010.

[46] Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[47] Y. Nishino, Y. Takahashi, N. Imamoto, T. Ishikawa, and K. Maeshima. Three-dimensional visualization of a human chromosome using coherent x-ray diffraction. *Physical Review Letters*, 102(1):018101, 2009.

[48] M. A. Pfeifer, G. J. Williams, I. A. Vartanyants, R. Harder, and I. K. Robinson. Three-dimensional mapping of a deformation field inside a nanocrystal. *Nature*, 442(7098):63, 2006.

[49] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C*. Cambridge University Press, Cambridge, second edition, 1992.

[50] Q. Qu, Y. Zhang, Y. C. Eldar, and J. Wright. Convolutional phase retrieval via gradient descent. *arXiv preprint arXiv:1712.00716*, 2017.

[51] J. M. Rodenburg and H. M. L. Faulkner. A phase retrieval algorithm for shifting illumination. *Applied physics letters*, 85(20):4795–4797, 2004.

[52] J. M. Rodenburg, A. C. Hurst, and A. G. Cullis. Transmission microscopy without lenses for objects of unlimited size. *Ultramicroscopy*, 107(2):227–231, 2007.

[53] J. M. Rodenburg, A. C. Hurst, A. G. Cullis, B. R. Dobson, F. Pfeiffer, O. Bunk, C. David, K. Jefimovs, and I. Johnson. Hard-X-ray lensless imaging of extended objects. *Physical Review Letters*, 98(3):034801, 2007.

[54] F. Salehi, E. Abbasi, and B. Hassibi. A precise analysis of phasemax in phase retrieval. *arXiv preprint arXiv:1801.06609*, 2018.

[55] 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. Single mimivirus particles intercepted and imaged with an x-ray laser. *Nature*, 470(7332):78, 2011.

[56] D. Shapiro, P. Thibault, T. Beetz, V. Elser, M. Howells, C. Jacobsen, J. Kirz, E. Lima, H. Miao, A. M. Neiman, et al. Biological imaging by soft x-ray diffraction microscopy. *Proceedings of the National Academy of Sciences*, 102(43):15343–15346, 2005.

[57] D. A. Shapiro, Y. S. Yu, T. Tyliszczak, J. Cabana, R. Celestre, W. Chao, K. Kaznatcheev, A. L. D. Kilcoyne, F. Maia, S. Marchesini, et al. Chemical composition mapping with nanometre resolution by soft x-ray microscopy. *Nature Photonics*, 8(10):765–769, 2014.

[58] M. Soltanolkotabi. Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization. *arXiv preprint arXiv:1702.06175*, 2017.

[59] F. Soulez, E. Thiebaut, A. Schutz, A. Ferrari, F. Courbin, and M. Unser. Proximity operators for phase retrieval. *Applied Optics*, 55:7412–7421, 2016.

[60] M. Stockmar, P. Cloetens, I. Zanette, B. Enders, M. Dierolf, F. Pfeiffer, and P. Thibault. Near-field ptychography: Phase retrieval for inline holography using a structured illumination. *Sci. Rep. 3*, (1927), 2013.

[61] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *arXiv preprint arXiv:1602.06664*, 2016.

[62] Y. S. Tan and R. Vershynin. Phase retrieval via randomized kaczmarz: Theoretical guarantees. *arxiv:1706.09993*, 2017.

[63] P. Thibault, M. Dierolf, O. Bunk, A. Menzel, and F. Pfeiffer. Probe retrieval in ptychographic coherent diffractive imaging. *Ultramicroscopy*, 109(4):338–343, 2009.

[64] P. Thibault, M. Dierolf, A. Menzel, O. Bunk, C. David, and F. Pfeiffer. High-resolution scanning x-ray diffraction microscopy. *Science*, 321(5887):379–382, 2008.

[65] P. Thibault and M. Guizar-Sicairos. Maximum-likelihood refinement for coherent diffractive imaging. *New J. Phys.*, 14:063004, 2012.

[66] L. Tian, Z. Liu, L. H. Yeh, M. Chen, J. Zhong, and L. Waller. Computational illumination for high-speed in vitro Fourier ptychographic microscopy. *Optica*, 2(10):904–911, 2015.

[67] Lei Tian and Laura Waller. 3d intensity and phase imaging from light field measurements in an led array microscope. *Optica*, 2(2):104–111, 2015.

[68] M. Udell, A. Yurtsever, V. Cevher, and J. Tropp. Sketchy decisions: Convex low-rank matrix optimization with optimal storage.

[69] J. Vila-Comamala, A. Diaz, M. Guizar-Sicairos, A. Mantion, Cameron M. Kewish, A. Menzel, O. Bunk, and C. David. Characterization of high-resolution diffractive X-ray optics by ptychographic coherent diffractive imaging. *Optics Express*, 19(22):21333–21344, 2011.

[70] I. Waldspurger. Phase retrieval with random Gaussian sensing vectors by alternating projections. *arXiv preprint arXiv:1609.03088*, 2016.

[71] G. Wang and G. B. Giannakis. Solving random systems of quadratic equations via truncated generalized gradient flow. *arXiv preprint arXiv:1605.08285*, 2016.

[72] Z. Wei, W. Chen, C. W. Qiu, and X. Chen. Conjugate gradient method for phase retrieval based on the Wirtinger derivative. *JOSA A*, 34(5):708–712, 2017.

[73] D. S. Weller, A. Pnueli, G. Divon, O. Radzyner, Y. C. Eldar, and J. A. Fessler. Undersampled phase retrieval with outliers. *IEEE Trans. Comput. Imaging*, 1:247–258, 2015.

[74] Z. Wen, C. Yang, X. Liu, and S. Marchesini. Alternating direction methods for classical and ptychographic phase retrieval. *Inverse Problems*, 28:115010, 2012.

[75] A. Williams, J. Chung, X. Ou, G. Zheng, S. Rawal, Z. Ao, R. Datar, C. Yang, and R. Cote. Fourier ptychographic microscopy for filtration-based circulating tumor cell enumeration and analysis. *Journal of Biomedical Optics*, 19(6):066007–066007, 2014.

[76] 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. Fresnel coherent diffractive imaging. *Physical Review Letters*, 97(2):025506, 2006.

[77] R. Xu, H. Jiang, C. Song, J. A. Rodriguez, Z. Huang, C. C. Chen, D. Nam, J. Park, M. Gallagher-Jones, S. Kim, et al. Single-shot 3D structure determination of nanocrystals with femtosecond x-ray free electron laser pulses. *arXiv preprint arXiv:1310.8594*, 2013.

[78] C. Yang, J. Qian, A. Schirotzek, F. Maia, and S. Marchesini. Iterative algorithms for ptychographic phase retrieval. *arXiv preprint arXiv:1105.5628*, 2011.

[79] L. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller. Experimental robustness of Fourier ptychography phase retrieval algorithms. *Optics Express*, 23(26):33214–33240, 2015.

[80] H. Zhang, Y. Chi, and Y. Liang. Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow. *arXiv preprint arXiv:1603.03805*, 2016.

[81] H. Zhang and Y. Liang. Reshaped Wirtinger flow for solving quadratic systems of equations. *arXiv preprint arXiv:1605.07719*, 2016.

[82] G. Zheng, R. Horstmeyer, and C. Yang. Wide-field, high-resolution Fourier ptychographic microscopy. *Nature Photonics*, 7(9):739–745, 2013.

[83] Y. Zhou, H. Zhang, and Y. Liang. Geometrical properties and accelerated gradient solvers of non-convex phase retrieval. In *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 331–335. IEEE, 2016.

APPENDIX

A   Theory for convergence to stationary points

He we present a theorem that provides theoretical justification for our choice of the step sizes (10) and (14) for Wirtinger Flow and Accelerated Wirtinger Flow.

Theorem A.1 *Let $\mathbf{f} \in \mathbb{C}^N$ denote the desired signal and assume we have M arbitrary noisy measurements of the form $\mathbf{b} = \left(|\mathbf{A}\mathbf{f}_\star|^2 + \mathbf{n}\right)^{1/2} \in \mathbb{R}^M$, where the square root is applied element-wise. Consider the cost function from Eq. (5). We run the Wirtinger Flow updates from Eq. (6) using the generalized gradient from Eq. (7) and the step size from Eq. (10). Also let $\hat{\mathbf{f}} \in \arg\min_{\mathbf{f} \in \mathbb{C}^n} \mathcal{L}(\mathbf{f})$ be a global optima. Then, the following identities hold*

$$\lim_{\tau \to \infty} \left\|\nabla \mathcal{L}(\hat{\mathbf{f}}_\tau)\right\|_2 \to 0 \quad and \quad \min_{\tau \in \{0,1,2,\dots,T\}} \left\|\nabla \mathcal{L}(\hat{\mathbf{f}}_\tau)\right\|_2^2 \le \frac{\left(\mathcal{L}(\hat{\mathbf{f}}_0) - \mathcal{L}(\hat{\mathbf{f}})\right)}{\bar{\mu}(T+1)}. \tag{18}$$

B   Proofs

We begin by studying the convergence of Wirtinger Flow iterates on a smoothed version of Eq. (5) defined by

$$\mathcal{L}_\epsilon(\mathbf{f}) := \sum_{m=1}^{M} \left(\left(\left|\mathbf{a}_m^H \mathbf{f}\right|^2 + \epsilon\right)^{1/2} - [\mathbf{b}]_m\right)^2, \tag{19}$$

where $\mathbf{a}_m^H \in \mathbb{C}^{1 \times N}$ is the mth row of $\mathbf{A}$ and $\epsilon \in \mathbb{R}$ is a nonnegative scalar. This cost function equals the cost function from Eq. (5) when $\epsilon = 0$, and is smooth when $\epsilon > 0$.

To calculate Wirtinger derivatives we have to first rewrite $\mathcal{L}_\epsilon(\mathbf{f})$ as a holomorphic function of $\mathbf{f}$ and its conjugate $\bar{\mathbf{f}}$. For the loss above this takes the form

$$\mathcal{L}_\epsilon(\mathbf{f}, \bar{\mathbf{f}}) := \mathcal{L}_\epsilon(\mathbf{f}) = \sum_{m=1}^{M} \left(\left(\mathbf{f}^T \left(\mathbf{a}_m \mathbf{a}_m^H\right)^T \bar{\mathbf{f}} + \epsilon\right)^{\frac{1}{2}} - [\mathbf{b}]_m\right)^2$$

Then the Wirtinger gradient can be calculated via the transpose of the partial derivative of $\mathcal{L}_\epsilon(\mathbf{f}, \bar{\mathbf{f}})$ with respect to $\mathbf{f}$ with $\bar{\mathbf{f}}$ fixed (see [9, Section 6] for background and further detail on Wirtinger gradient calculations). As a result for the loss considered in this paper the Wirtinger gradient is equal to

$$\nabla \mathcal{L}_\epsilon = \left(\frac{\partial}{\partial \mathbf{f}} \mathcal{L}_\epsilon(\mathbf{f}, \bar{\mathbf{f}})\right)^T = \sum_{m=1}^{M} \frac{\left(\left(\mathbf{f}^T \left(\mathbf{a}_m \mathbf{a}_m^H\right)^T \bar{\mathbf{f}} + \epsilon\right)^{\frac{1}{2}} - [\mathbf{b}]_m\right)}{\left(\mathbf{f}^T \left(\mathbf{a}_m \mathbf{a}_m^H\right)^T \bar{\mathbf{f}} + \epsilon\right)^{\frac{1}{2}}} \left(\mathbf{a}_m \mathbf{a}_m^H\right)^T \bar{\mathbf{f}}.$$

Also

$$\mathcal{H}_{\mathbf{ff}} = \frac{\partial}{\partial \mathbf{f}} \left(\frac{\partial \mathcal{L}_\epsilon}{\partial \mathbf{f}} \right)^H$$

$$= \frac{\partial}{\partial \mathbf{f}} \left(\sum_{m=1}^{M} \frac{\left(\left(\mathbf{f}^T \left(\mathbf{a}_m \mathbf{a}_m^H \right)^T \bar{\mathbf{f}} + \epsilon \right)^{\frac{1}{2}} - [\mathbf{b}]_m \right)}{\left(\mathbf{f}^T \left(\mathbf{a}_m \mathbf{a}_m^H \right)^T \bar{\mathbf{f}} + \epsilon \right)^{\frac{1}{2}}} \left(\mathbf{a}_m \mathbf{a}_m^H \right) \mathbf{f} \right)$$

$$= \sum_{m=1}^{M} \frac{\left(\sqrt{|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon} - [\mathbf{b}]_m \right)}{\sqrt{|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon}} \left(\mathbf{a}_m \mathbf{a}_m^H \right) + \frac{1}{2} \sum_{m=1}^{M} [\mathbf{b}]_m \frac{|\mathbf{a}_m^H \mathbf{f}|^2}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \left(\mathbf{a}_m \mathbf{a}_m^H \right)$$

$$= \sum_{m=1}^{M} \left(1 - \frac{1}{2} \frac{[\mathbf{b}]_m}{\sqrt{|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon}} - \frac{\epsilon}{2} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \right) \left(\mathbf{a}_m \mathbf{a}_m^H \right).$$

$$\mathcal{H}_{\bar{\mathbf{f}}\mathbf{f}} = \frac{\partial}{\partial \bar{\mathbf{f}}} \left(\frac{\partial \mathcal{L}_\epsilon}{\partial \mathbf{f}} \right)^H$$

$$= \frac{\partial}{\partial \bar{\mathbf{f}}} \left(\sum_{m=1}^{m} \frac{\left(\sqrt{\bar{\mathbf{f}}^T \left(\mathbf{a}_m \mathbf{a}_m^H \right) \mathbf{f} + \epsilon} - [\mathbf{b}]_m \right)}{\sqrt{\bar{\mathbf{f}}^T \left(\mathbf{a}_m \mathbf{a}_m^H \right) \mathbf{f} + \epsilon}} \left(\mathbf{a}_m \mathbf{a}_m^H \right) \mathbf{f} \right)$$

$$= \sum_{m=1}^{M} \frac{1}{2} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \left(\mathbf{a}_m^H \mathbf{f} \right)^2 \mathbf{a}_m \mathbf{a}_m^T$$

The Hessian takes the form

$$\nabla^2 \mathcal{L}_\epsilon(\mathbf{f}) = \begin{bmatrix} \mathcal{H}_{\mathbf{ff}} & \mathcal{H}_{\bar{\mathbf{f}}\mathbf{f}} \\ \mathcal{H}_{\mathbf{f}\bar{\mathbf{f}}} & \mathcal{H}_{\bar{\mathbf{f}}\bar{\mathbf{f}}} \end{bmatrix}.$$

We thus have

$$
\begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix}^H \nabla^2 \mathcal{L}_\epsilon(\mathbf{f}) \begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix} = \mathbf{u}^H \mathcal{H}_{\mathbf{ff}} \mathbf{u} + \mathbf{u}^H \mathcal{H}_{\bar{\mathbf{f}}\mathbf{f}} \bar{\mathbf{u}} + \mathbf{u}^T \mathcal{H}_{\mathbf{f}\bar{\mathbf{f}}} \mathbf{u} + \mathbf{u}^T \mathcal{H}_{\bar{\mathbf{f}}\bar{\mathbf{f}}} \mathbf{u}
$$

$$
= 2 \sum_{m=1}^{M} \left(1 - \frac{1}{2} \frac{[\mathbf{b}]_m}{\sqrt{|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon}} - \frac{\epsilon}{2} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \right) |\mathbf{a}_m^H \mathbf{u}|^2
$$

$$
+ \sum_{m=1}^{M} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \Re \left(\left(\mathbf{a}_m^H \mathbf{f} \right)^2 (\mathbf{u}^H \mathbf{a}_m)^2 \right)
$$

$$
= 2 \sum_{m=1}^{M} \left(1 - \frac{1}{2} \frac{[\mathbf{b}]_m |\mathbf{a}_m^H \mathbf{f}|^2}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} - \epsilon \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \right) |\mathbf{a}_m^H \mathbf{u}|^2
$$

$$
+ \sum_{m=1}^{M} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \Re \left(\left(\mathbf{a}_m^H \mathbf{f} \right)^2 (\mathbf{u}^H \mathbf{a}_m)^2 \right)
$$

$$
= 2 \sum_{m=1}^{M} \left(1 - \epsilon \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \right) |\mathbf{a}_m^H \mathbf{u}|^2
$$

$$
+ \sum_{m=1}^{M} \frac{[\mathbf{b}]_m}{\left(|\mathbf{a}_m^H \mathbf{f}|^2 + \epsilon \right)^{\frac{3}{2}}} \left(\Re \left(\left(\mathbf{a}_m^H \mathbf{f} \right)^2 (\mathbf{u}^H \mathbf{a}_m)^2 \right) - |\mathbf{a}_m^H \mathbf{f}|^2 |\mathbf{a}_m^H \mathbf{u}|^2 \right)
$$

$$
\leq 2 \sum_{m=1}^{M} |\mathbf{a}_m^H \mathbf{u}|^2
$$

$$
\leq 2 \| \mathbf{A} \|^2 \| \mathbf{u} \|_2^2
$$

$$
= \| \mathbf{A} \|^2 \left\| \begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix} \right\|_2^2
$$

$$
= \| \mathbf{A}^H \mathbf{A} \| \left\| \begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix} \right\|_2^2 .
$$

In short for any $\mathbf{u}, \mathbf{f} \in \mathbb{C}^n$ we have

$$
\begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix}^H \nabla^2 \mathcal{L}_\epsilon(\mathbf{f}) \begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix} \leq L \left\| \begin{bmatrix} \mathbf{u} \\ \bar{\mathbf{u}} \end{bmatrix} \right\|_2^2 , \tag{20}
$$

with $L = \| \mathbf{A}^H \mathbf{A} \|$. Combining the latter identity with a Wirtinger derivate version of Taylor's

approximation theorem (e.g. see [9, Section 6]) we have that for any $\mathbf{f}, \mathbf{f}^+ \in \mathbb{C}^n$

$$\mathcal{L}_\epsilon(\mathbf{f}^+) = \mathcal{L}_\epsilon(\mathbf{f}) + \begin{bmatrix} \nabla \mathcal{L}_\epsilon(\mathbf{f}) \\ \overline{\nabla \mathcal{L}_\epsilon(\mathbf{f})} \end{bmatrix}^H \begin{bmatrix} \mathbf{f}^+ - \mathbf{f} \\ \overline{\mathbf{f}^+ - \mathbf{f}} \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \mathbf{f}^+ - \mathbf{f} \\ \overline{\mathbf{f}^+ - \mathbf{f}} \end{bmatrix}^H \left(\int_0^1 \nabla^2 \mathcal{L}_\epsilon \left(\mathbf{f} + t(\mathbf{f}^+ - \mathbf{f}) \right) \right) \begin{bmatrix} \mathbf{f}^+ - \mathbf{f} \\ \overline{\mathbf{f}^+ - \mathbf{f}} \end{bmatrix},$$

$$\leq \mathcal{L}_\epsilon(\mathbf{f}) + \begin{bmatrix} \nabla \mathcal{L}_\epsilon(\mathbf{f}) \\ \overline{\nabla \mathcal{L}_\epsilon(\mathbf{f})} \end{bmatrix}^H \begin{bmatrix} \mathbf{f}^+ - \mathbf{f} \\ \overline{\mathbf{f}^+ - \mathbf{f}} \end{bmatrix} + \frac{L}{2} \left\| \begin{bmatrix} \mathbf{f}^+ - \mathbf{f} \\ \overline{\mathbf{f}^+ - \mathbf{f}} \end{bmatrix} \right\|_2^2.$$

Now plugging $\mathbf{f}^+ = \hat{\mathbf{f}}_{\tau+1} = \hat{\mathbf{f}}_\tau - \mu \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau)$ with $\mu = 1/L$ and $\mathbf{f} = \hat{\mathbf{f}}_\tau$ in the above identity we conclude that

$$\mathcal{L}_\epsilon(\hat{\mathbf{f}}_{\tau+1}) \leq \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) - \mu \left(1 - \frac{\mu L}{2} \right) \left\| \begin{bmatrix} \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) \\ \overline{\nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau)} \end{bmatrix} \right\|_2,$$

$$= \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) - \frac{1}{2L} \left\| \begin{bmatrix} \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) \\ \overline{\nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau)} \end{bmatrix} \right\|_2^2,$$

$$= \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) - \frac{1}{L} \left\| \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) \right\|_2^2.$$

Rearranging the above inequality we conclude that

$$\left\| \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) \right\|_2^2 \leq L \left(\mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) - \mathcal{L}_\epsilon(\hat{\mathbf{f}}_{\tau+1}) \right).$$

Now summing over $\tau = 1, 2, \ldots, T$ we conclude that

$$\sum_{\tau=0}^T \left\| \nabla \mathcal{L}_\epsilon(\hat{\mathbf{f}}_\tau) \right\|_2^2 \leq L \left(\mathcal{L}_\epsilon(\hat{\mathbf{f}}_0) - \mathcal{L}_\epsilon(\hat{\mathbf{f}}_{T+1}) \right) \leq L \left(\mathcal{L}_\epsilon(\hat{\mathbf{f}}_0) - \mathcal{L}_\epsilon(\hat{\mathbf{f}}) \right).$$

Note that the above expression holds for any ϵ. Thus taking the limit of both sides as $\epsilon \to 0$ we conclude that

$$\sum_{\tau=0}^T \left\| \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau) \right\|_2^2 \leq L \left(\mathcal{L}(\hat{\mathbf{f}}_0) - \mathcal{L}(\hat{\mathbf{f}}) \right).$$

Since the series above converges we must have $\lim_{\tau \to \infty} \left\| \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau) \right\|_F \to 0$. Furthermore, we have

$$\min_{\tau \in \{0,1,2,\ldots,T\}} \left\| \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau) \right\|_F^2 \leq \frac{1}{T+1} \sum_{\tau=0}^T \left\| \nabla \mathcal{L}(\hat{\mathbf{f}}_\tau) \right\|_2^2 \leq \frac{L \left(\mathcal{L}(\hat{\mathbf{f}}_0) - \mathcal{L}(\hat{\mathbf{f}}) \right)}{T+1},$$

proving the identities (18) and concluding the proof.