Phase Retrieval via Wirtinger Flow: Theory and Algorithms

Emmanuel J. Candès, Xiaodong Li, and Mahdi Soltanolkotabi

Abstract—We study the problem of recovering the phase from magnitude measurements; specifically, we wish to reconstruct a complex-valued signal $x \in \mathbb{C}^n$ about which we have phaseless samples of the form $y_r = |\langle a_r, x \rangle|^2$, $r = 1, \ldots, m$ (knowledge of the phase of these samples would yield a linear system). This paper develops a nonconvex formulation of the phase retrieval problem as well as a concrete solution algorithm. In a nutshell, this algorithm starts with a careful initialization obtained by means of a spectral method, and then refines this initial estimate by iteratively applying novel update rules, which have low computational complexity, much like in a gradient descent scheme. The main contribution is that this algorithm is shown to rigorously allow the exact retrieval of phase information from a nearly minimal number of random measurements. Indeed, the sequence of successive iterates provably converges to the solution at a geometric rate so that the proposed scheme is efficient both in terms of computational and data resources. In theory, a variation on this scheme leads to a near-linear time algorithm for a physically realizable model based on coded diffraction patterns. We illustrate the effectiveness of our methods with various experiments on image data. Underlying our analysis are insights for the analysis of nonconvex optimization schemes that may have implications for computational problems beyond phase retrieval.

Index Terms—Non-convex optimization, convergence to global optimum, phase retrieval, Wirtinger derivatives.

I. Introduction

WE ARE interested in solving quadratic equations of the form

$$y_r = |\langle a_r, z \rangle|^2, \quad r = 1, 2, \ldots, m, \tag{I.1}$$

where $z \in \mathbb{C}^n$ is the decision variable, $a_r \in \mathbb{C}^n$ are known sampling vectors, and $y_r \in \mathbb{R}$ are observed measurements. This problem is a general instance of a nonconvex quadratic program (QP). Nonconvex QPs have been observed to occur frequently in science and engineering and, consequently, their study is of importance. For example, a class of combinatorial optimization problems with Boolean decision variables may

be cast as QPs [54, Sec. 4.3.1]. Focusing on the literature on physical sciences, the problem (I.1) is generally referred to as the phase retrieval problem. To understand this connection, recall that most detectors can only record the intensity of the light field and not its phase. Thus, when a small object is illuminated by a quasi-monochromatic wave, detectors measure the magnitude of the diffracted light. In the far field, the diffraction pattern happens to be the Fourier transform of the object of interest—this is called Fraunhofer diffraction—so that in discrete space, (I.1) models the data aquisition mechanism in a coherent diffraction imaging setup; one can identify z with the object of interest, a_r with complex sinusoids, and y_r with the recorded data. Hence, we can think of (I.1) as a *generalized phase retrieval* problem. As is well known, the phase retrieval problem arises in many areas of science and engineering such as X-ray crystallography [33], [50], microscopy [47]–[49], astronomy [21], diffraction and array imaging [12], [18], and optics [68]. Other fields of application include acoustics [4], [5], blind channel estimation in wireless communications [3], [60], interferometry [22], quantum mechanics [20], [61] and quantum information [35]. We refer the reader to the tutorial paper [64] for a recent review of the theory and practice of phase retrieval.

Because of the practical significance of the phase retrieval problem in imaging science, the community has developed methods for recovering a signal $x \in \mathbb{C}^n$ from data of the form $y_r = |\langle a_r, x \rangle|^2$ in the special case where one samples the (square) modulus of its Fourier transform. In this setup, the most widely used method is perhaps the error reduction algorithm and its generalizations, which were derived from the pioneering research of Gerchberg and Saxton [29] and Fienup [24], [26]. The Gerchberg-Saxton algorithm starts from a random initial estimate and proceeds by iteratively applying a pair of 'projections': at each iteration, the current guess is projected in data space so that the magnitude of its frequency spectrum matches the observations; the signal is then projected in signal space to conform to some a-priori knowledge about its structure. In a typical instance, our knowledge may be that the signal is real-valued, nonnegative and spatially limited. First, while error reduction methods often work well in practice, the algorithms seem to rely heavily on a priori information about the signals, see [25], [27], [34], [62]. Second, since these algorithms can be cast as alternating projections onto nonconvex sets [8] (the constraint in Fourier space is not convex), fundamental mathematical questions concerning their convergence remain, for the most part, unresolved; we refer to Section III-B for further discussion.

On the theoretical side, several combinatorial optimization problems—optimization programs with discrete design variables which take on integer or Boolean values—can be cast as solving quadratic equations or as minimizing a linear objective subject to quadratic inequalities. In their most general form these problems are known to be notoriously difficult (NP-hard) [54, Sec. 4.3]. Nevertheless, many heuristics have been developed for addressing such problems.[1] One popular heuristic is based on a class of convex relaxations known as Shor's relaxations [54, Sec. 4.3.1] which can be solved using tractable semi-definite programming (SDP). For certain random models, some recent SDP relaxations such as PhaseLift [14] are known to provide exact solutions (up to global phase) to the generalized phase retrieval problem using a near minimal number of sampling vectors [16], [17]. While in principle SDP based relaxations offer tractable solutions, they become computationally prohibitive as the dimension of the signal increases. Indeed, for a large number of unknowns in the tens of thousands, say, the memory requirements are far out of reach of desktop computers so that these SDP relaxations are de facto impractical.

II. ALGORITHM: WIRTINGER FLOW

This paper introduces an approach to phase retrieval based on non-convex optimization as well as a solution algorithm, which has two components: (1) a careful initialization obtained by means of a spectral method, and (2) a series of updates refining this initial estimate by iteratively applying a novel update rule, much like in a gradient descent scheme. We refer to the combination of these two steps, introduced in reverse order below, as the *Wirtinger flow* (WF) algorithm.

A. Minimization of a Non-Convex Objective

Let $\ell(x, y)$ be a loss function measuring the misfit between both its scalar arguments. If the loss function is non-negative and vanishes only when $x = y$, then a solution to the generalized phase retrieval problem (I.1) is any solution to

$$\text{minimize} \quad f(z) := \frac{1}{2m} \sum_{r=1}^{m} \ell\left(y_r, \left|a_r^* z\right|^2\right), \quad z \in \mathbb{C}^n. \quad \text{(II.1)}$$

Although one could study many loss functions, we shall focus in this paper on the simple quadratic loss $\ell(x, y) = (x - y)^2$. Admittedly, the formulation (II.1) does not make the problem any easier since the function f is not convex. Minimizing non-convex objectives, which may have very many stationary points, is known to be NP-hard in general. In fact, even establishing convergence to a local minimum or stationary point can be quite challenging, please see [53] for an example where convergence to a local minimum of a degree-four polynomial is known to be NP-hard.[2] As a side remark,

[1] For a partial review of some of these heuristics as well as some recent theoretical advances in related problems we refer to our companion paper [16, Sec. 1.6] and references therein [7], [13], [30], [31], [36], [55], [56], [65].

[2] Observe that if all the sampling vectors are real valued, our objective is also a degree-four polynomial.

Algorithm 1 Wirtinger Flow: Initialization

Input: Observations $\{y_r\} \in \mathbb{R}^m$.
  Set

$$\lambda^2 = n \frac{\sum_r y_r}{\sum_r \|a_r\|^2}.$$

  Set z_0, normalized to $\|z_0\| = \lambda$, to be the eigenvector corresponding to the largest eigenvalue of

$$Y = \frac{1}{m} \sum_{r=1}^{m} y_r a_r a_r^*.$$

Output: Initial guess z_0.

---

deciding whether a stationary point of a polynomial of degree four is a local minimizer is already known to be NP-hard.

Our approach to (II.1) is simply stated: start with an initialization z_0, and for $\tau = 0, 1, 2, \ldots$, inductively define

$$\begin{aligned} z_{\tau+1} &= z_\tau - \frac{\mu_{\tau+1}}{\|z_0\|^2} \left(\frac{1}{m} \sum_{r=1}^{m} \left(\left|a_r^* z\right|^2 - y_r \right) (a_r a_r^*) z \right) \\ &:= z_\tau - \frac{\mu_{\tau+1}}{\|z_0\|^2} \nabla f(z_\tau). \end{aligned} \quad \text{(II.2)}$$

If the decision variable z and the sampling vectors were all real valued, the term between parentheses would be the gradient of f divided by two, as our notation suggests. However, since $f(z)$ is a mapping from $\mathbb{C}^n$ to $\mathbb{R}$, it is not holomorphic and hence not complex-differentiable. However, this term can still be viewed as a gradient based on Wirtinger derivatives reviewed in Section VI. Hence, (II.2) is a form of steepest descent and the parameter $\mu_{\tau+1}$ can be interpreted as a step size (note nonetheless that the effective step size is also inversely proportional to the magnitude of the initial guess).

B. Initialization via a Spectral Method

Our main result states that for a certain random model, if the initialization z_0 is sufficiently accurate, then the sequence $\{z_\tau\}$ will converge toward a solution to the generalized phase problem (I.1). In this paper, we propose computing the initial guess z_0 via a spectral method, detailed in Algorithm 1. In words, z_0 is the leading eigenvector of the positive semidefinite Hermitian matrix $\sum_r y_r a_r a_r^*$ constructed from the knowledge of the sampling vectors and observations. (As usual, a_r^* is the adjoint of a_r.) Letting A be the $m \times n$ matrix whose rth row is a_r^* so that with obvious notation $y = |Ax|^2$, z_0 is the leading eigenvector of $A^* \text{diag}\{y\} A$ and can be computed via the power method by repeatedly applying A, entrywise multiplication by y and A^*. In the theoretical framework we study below, a constant number of power iterations would give machine accuracy because of an eigenvalue gap between the top two eigenvalues, please see Appendix B for additional information.

C. Wirtinger Flow as a Stochastic Gradient Scheme

We would like to motivate the Wirtinger flow algorithm and provide some insight as to why we expect it to work in

a model where the sampling vectors are random. First, we emphasize that our statements in this section are heuristic in nature; as it will become clear in the proof Section VII, a correct mathematical formalization of these ideas is far more complicated than our heuristic development here may suggest. Second, although our ideas are broadly applicable, it makes sense to begin understanding the algorithm in a setting where everything is real valued, and in which the vectors a_r are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Also without any loss in generality and to simplify exposition in this section we shall assume that the signal has unit Euclidean norm, i. e. $\|\mathbf{x}\| = 1$.

Let $\mathbf{x}$ be a solution to (I.1) so that $y_r = |\langle \mathbf{a}_r, \mathbf{x} \rangle|^2$, and consider the initialization step first. In the Gaussian model, a simple moment calculation gives

$$\mathbb{E}\left[\frac{1}{m}\sum_{r=1}^{m} y_r \mathbf{a}_r \mathbf{a}_r^*\right] = \mathbf{I} + 2\mathbf{x}\mathbf{x}^*.$$

By the strong law of large numbers, the matrix $\mathbf{Y}$ in Algorithm 1 is equal to the right-hand side in the limit of large samples. Since any leading eigenvector of $\mathbf{I} + 2\mathbf{x}\mathbf{x}^*$ is of the form $\lambda \mathbf{x}$ for some scalar $\lambda \in \mathbb{R}$, we see that the intialization step would recover $\mathbf{x}$ perfectly, up to a global sign or phase factor, had we infinitely many samples. Indeed, the chosen normalization would guarantee that the recovered signal is of the form $\pm \mathbf{x}$. As an aside, we would like to note that the top two eigenvalues of $\mathbf{I} + 2\mathbf{x}\mathbf{x}^*$ are well separated unless $\|\mathbf{x}\|$ is very small, and that their ratio is equal to $1 + 2\|\mathbf{x}\|^2$. Now with a finite amount of data, the leading eigenvector of $\mathbf{Y}$ will of course not be perfectly correlated with $\mathbf{x}$ but we hope that it is sufficiently correlated to point us in the right direction.

We now turn our attention to the gradient-update (II.2) and define

$$2\mathbf{z}^*(\mathbf{I} - \mathbf{x}\mathbf{x}^*)\mathbf{z} + \frac{3}{2}\left(\|\mathbf{z}\|^2 - 1\right)^2$$

where here and below, $\mathbf{x}$ is once again our planted solution. The first term ensures that the direction of $\mathbf{z}$ matches the direction of $\mathbf{x}$ and the second term penalizes the deviation of the Euclidean norm of $\mathbf{z}$ from that of $\mathbf{x}$. Obviously, the minimizers of this function are $\pm \mathbf{x}$. Now consider the gradient scheme

$$\mathbf{z}_{\tau+1} = \mathbf{z}_\tau - \frac{\mu_{\tau+1}}{\|\mathbf{z}_0\|^2} \nabla F(\mathbf{z}_\tau) \qquad (\text{II.3})$$

In Section VII-I, we show that if $\min \|\mathbf{z}_0 \pm \mathbf{x}\| \leq 1/8\|\mathbf{x}\|$, then $\{\mathbf{z}_\tau\}$ converges to $\mathbf{x}$ up to a global sign. However, this is all ideal as we would need knowledge of $\mathbf{x}$ itself to compute the gradient of F; we simply cannot run this algorithm.

Consider now the WF update and assume for a moment that $\mathbf{z}_\tau$ is fixed and independent of the sampling vectors. We are well aware that this is a false assumption but nevertheless wish to explore some of its consequences. In the Gaussian model, if $\mathbf{z}$ is independent of the sampling vectors, then a modification of Lemma 7.2 for real-valued $\mathbf{z}$ shows that $\mathbb{E}[\nabla f(\mathbf{z})] = \nabla F(\mathbf{z})$ and, therefore,

$$\mathbb{E}[\mathbf{z}_{\tau+1}] = \mathbb{E}[\mathbf{z}_\tau] - \frac{\mu_{\tau+1}}{\|\mathbf{z}_0\|^2} \mathbb{E}[\nabla f(\mathbf{z}_\tau)]$$

$$\Rightarrow \mathbb{E}[\mathbf{z}_{\tau+1}] = \mathbf{z}_\tau - \frac{\mu_{\tau+1}}{\|\mathbf{z}_0\|^2} \nabla F(\mathbf{z}_\tau).$$

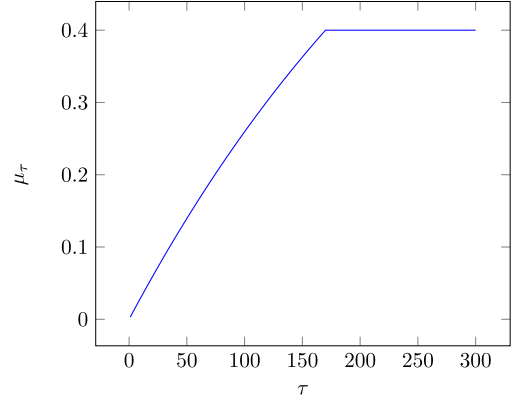Fig. 1. Learning parameter μ_τ from (II.5) as a function of the iteration count τ; here, $\tau_0 \approx 330$ and $\mu_{\max} = 0.4$.

Hence, the average WF update is the same as that in (II.3) so that we can interpret the Wirtinger flow algorithm as a stochastic gradient scheme in which we only get to observe an unbiased estimate $\nabla f(\mathbf{z})$ of the "true" gradient $\nabla F(\mathbf{z})$.

Regarding WF as a stochastic gradient scheme helps us in choosing the learning parameter or step size μ_τ. Lemma 7.7 asserts that

$$\|\nabla f(\mathbf{z}) - \nabla F(\mathbf{z})\|^2 \leq \|\mathbf{x}\|^2 \cdot \min \ \|\mathbf{z} \pm \mathbf{x}\| \qquad (\text{II.4})$$

holds with high probability. Looking at the right-hand side, this says that the uncertainty about the gradient estimate depends on how far we are from the actual solution $\mathbf{x}$. The further away, the larger the uncertainty or the noisier the estimate. This suggests that in the early iterations we should use a small learning parameter as the noise is large since we are not yet close to the solution. However, as the iterations count increases and we make progress, the size of the noise also decreases and we can pick larger values for the learning parameter. This heuristic together with experimentation lead us to consider

$$\mu_\tau = \min(1 - e^{-\tau/\tau_0}, \mu_{\max}) \qquad (\text{II.5})$$

shown in Figure 1. Values of τ_0 around 330 and of $\mu_{\max}$ around 0.4 worked well in our simulations. This makes sure that μ_τ is rather small at the beginning (e.g. $\mu_1 \approx 0.003$ but quickly increases and reaches a maximum value of about 0.4 after 200 iterations or so.

III. MAIN RESULTS

A. Exact Phase Retrieval via Wirtinger Flow

Our main result establishes the correctness of the Wirtinger flow algorithm in the Gaussian model defined below. Later in Section V, we shall also develop exact recovery results for a physically inspired diffraction model.

Definition 3.1: We say that the sampling vectors follow the Gaussian model if $\mathbf{a}_r \in \mathbb{C}^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}/2) + i\mathcal{N}(\mathbf{0}, \mathbf{I}/2)$. In the real-valued case, they are i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

We also need to define the distance to the solution set.

Definition 3.2: Let $\mathbf{x} \in \mathbb{C}^n$ be any solution to the quadratic system (I.1) (the signal we wish to recover). For each $\mathbf{z} \in \mathbb{C}^n$,

define

$$dist(z, x) = \min_{\phi \in [0, 2\pi]} \left\| z - e^{i\phi} x \right\|.$$

Theorem 3.3: Let x be an arbitrary vector in $\mathbb{C}^n$ and $y = |Ax|^2 \in \mathbb{R}^m$ be m quadratic samples with $m \geq c_0 \cdot n \log n$, where c_0 is a sufficiently large numerical constant. Then the Wirtinger flow initial estimate z_0 normalized to have squared Euclidean norm equal to $m^{-1} \sum_r y_r,$[3] obeys

$$dist(z_0, x) \leq \frac{1}{8} \|x\| \tag{III.1}$$

with probability at least $1 - 10e^{-\gamma n} - 8/n^2$ (γ is a fixed positive numerical constant). Further, take a constant learning parameter sequence, $\mu_\tau = \mu$ for all $\tau = 1, 2, \ldots$ and assume $\mu \leq c_1/n$ for some fixed numerical constant c_1. Then there is an event of probability at least $1 - 13e^{-\gamma n} - me^{-1.5m} - 8/n^2$, such that on this event, starting from any initial solution z_0 obeying (III.1), we have

$$dist(z_\tau, x) \leq \frac{1}{8} \left(1 - \frac{\mu}{4} \right)^{\tau/2} \cdot \|x\|.$$

Clearly, one would need $2n$ quadratic measurements to have any hope of recovering $x \in \mathbb{C}^n$. It is also known that in our sampling model, the mapping $z \mapsto |Az|^2$ is injective for $m \geq 4n$ [5] and that this property holds for generic sampling vectors [19].[4] Hence, the Wirtinger flow algorithm loses at most a logarithmic factor in the *sampling complexity*. In comparison, the SDP relaxation only needs a sampling complexity proportional to n (no logarithmic factor) [15], and it is an open question whether Theorem 3.3 holds in this regime.

Setting $\mu = c_1/n$ yields ϵ accuracy in a relative sense, namely, $dist(z, x) \leq \epsilon \|x\|$, in $\mathcal{O}(n \log 1/\epsilon)$ iterations. The computational work at each iteration is dominated by two matrix-vector products of the form Az and A^*v, see Appendix B. It follows that the overall computational complexity of the WF algorithm is $\mathcal{O}(mn^2 \log 1/\epsilon)$. Later in the paper, we will exhibit a modification to the WF algorithm of mere theoretical interest, which also yields exact recovery under the same sampling complexity and an $\mathcal{O}(mn \log 1/\epsilon)$ computational complexity; that is to say, the computational workload is now just *linear* in the problem size.

B. Comparison With Other Non-Convex Schemes

We now pause to comment on a few other non-convex schemes in the literature. Other comparisons may be found in our companion paper [16].

Earlier, we discussed the Gerchberg-Saxton and Fienup algorithms. These formulations assume that A is a Fourier transform and can be described as follows: suppose z_τ is the current guess, then one computes the image of z_τ through A and adjust its modulus so that it matches that of the observed

---

[3]The same results holds with the intialization from Algorithm 1 because $\sum_r \|a_r\|^2 \approx m \cdot n$ with a standard deviation of about the square root of this quantity.

[4]It is not within the scope of this paper to explain the meaning of generic vectors and, instead, refer the interested reader to [19].

data vector: with obvious notation,

$$\hat{v}_{\tau+1} = b \odot \frac{Az_\tau}{|Az_\tau|}, \tag{III.2}$$

where $\odot$ is elementwise multiplication, and $b = |Ax|$ so that $b_r^2 = y_r$ for all $r = 1, \ldots, m$. Then

$$v_{\tau+1} = \arg \min_{v \in \mathbb{C}^n} \|\hat{v}_{\tau+1} - Av\|. \tag{III.3}$$

(In the case of Fourier data, the step (III.2)–(III.3) essentially adjusts the modulus of the Fourier transform of the current guess so that it fits the measured data.) Finally, if we know that the solution belongs to a convex set $\mathcal{C}$ (as in the case where the signal is known to be real-valued, possibly non-negative and of finite support), then the next iterate is

$$z_{\tau+1} = P_\mathcal{C}(v_{\tau+1}), \tag{III.4}$$

where $P_\mathcal{C}$ is the projection onto the convex set $\mathcal{C}$. If no such information is available, then $z_{\tau+1} = v_{\tau+1}$. The first step (III.3) is not a projection onto a convex set and, therefore, it is in general completely unclear whether the Gerchberg-Saxton algorithm actually converges. (And if it were to converge, at what speed?) It is also unclear how the procedure should be initialized to yield accurate final estimates. This is in contrast to the Wirtinger flow algorithm, which in the Gaussian model is shown to exhibit geometric convergence to the solution to the phase retrieval problem. Another benefit is that the Wirtinger flow algorithm does not require solving a least-squares problem (III.3) at each iteration; each step enjoys a reduced computational complexity.

A recent contribution related to ours is the interesting paper [56], which proposes an alternating minimization scheme named AltMinPhase for the general phase retrieval problem. AltMinPhase is inspired by the Gerchberg-Saxton update (III.2)–(III.3) as well as other established alternating projection heuristics [26], [42], [43], [45], [46], [69]. We describe the algorithm in the setup of Theorem 3.3 for which [56] gives theoretical guarantees. To begin with, AltMinPhase partitions the sampling vectors a_r (the rows of the matrix A) and corresponding observations y_r into $B + 1$ disjoint blocks $(y^{(0)}, A^{(0)})$, $(y^{(1)}, A^{(1)})$, …, $(y^{(B)}, A^{(B)})$ of roughly equal size. Hence, distinct blocks are stochastically independent from each other. The first block $(y^{(0)}, A^{(0)})$ is used to compute an initial estimate z_0. After initialization, AltMinPhase goes through a series of iterations of the form (III.2)–(III.3), however, with the key difference that each iteration uses a fresh set of sampling vectors and observations: in details,

$$z_{\tau+1} = \arg \min_{z \in \mathbb{C}^n} \|\hat{v}_{\tau+1} - A^{(\tau+1)} z\|,$$
$$\hat{v}_{\tau+1} = b \odot \frac{A^{(\tau+1)} z_\tau}{|A^{(\tau+1)} z_\tau|}. \tag{III.5}$$

As for the Gerchberg-Saxton algorithm, each iteration requires solving a least-squares problem. Now assume a real-valued Gaussian model as well as a real valued solution $x \in \mathbb{R}^n$. The main result in [56] states that if the first block $(y^{(0)}, A^{(0)})$ contains at least $c \cdot n \log^3 n$ samples and each consecutive block contains at least $c \cdot n \log n$ samples—c here denotes a

positive numerical constant whose value may change at each occurence—then it is possible to initialize the algorithm via data from the first block in such a way that each consecutive iterate (III.5) decreases the error $\|z_\tau - x\|$ by 50%; naturally, all of this holds in a probabilistic sense. Hence, one can get ϵ accuracy in the sense introduced earlier from a total of $c \cdot n \log n \cdot (\log^2 n + \log 1/\epsilon)$ samples. Whereas the Wirtinger flow algorithm achieves arbitrary accuracy from just $c \cdot n \log n$ samples, these theoretical results would require an infinite number of samples. This is, however, not the main point.

The main point is that in practice, it is not realistic to imagine (1) that we will divide the samples in distinct blocks (how many blocks should we form a priori? of which sizes?) and (2) that we will use measured data only once. With respect to the latter, observe that the Gerchberg-Saxton procedure (III.2)–(III.3) uses all the samples at each iteration. This is the reason why AltMinPhase is of little practical value, and of theoretical interest only. As a matter of fact, its design and study seem merely to stem from analytical considerations: since one uses an independent set of measurements at each iteration, $A^{(\tau+1)}$ and z_τ are stochastically independent, a fact which considerably simplifies the convergence analysis. In stark contrast, the WF iterate uses all the samples at each iteration and thus introduces some dependencies, which makes for some delicate analysis. Overcoming these difficulties is crucial because the community is preoccupied with convergence properties of algorithms one actually runs, like Gerchberg-Saxton (III.2)–(III.3), or would actually want to run. Interestingly, it may be possible to use some of the ideas developed in this paper to develop a rigorous theory of convergence for algorithms in the style of Gerchberg-Saxton and Fienup, please see [65].

In a recent paper [44], which appeared on the arXiv preprint server as the final version of this paper was under preparation, the authors explore necessary and sufficient conditions for the global convergence of an alternative minimization scheme with generic sampling vectors. The issue is that we do not know when these conditions hold. Further, even when the algorithm converges, it does not come with an explicit convergence rate so that is is not known whether the algorithm converges in polynomial time. As before, some of our methods as well as those from our companion paper [16] may have some bearing upon the analysis of this algorithm. Similarly, another class of nonconvex algorithms that have recently been proposed in the literature are iterative algorithms based on Generalized Approximate Message Passing (GAMP), see [59], [63] as well as [9], [10], [23] for some background literature on AMP. In [63], the authors demonstrate a favorable runtime for an algorithm of this nature. However, this does not come with any theoretical guarantees.

Moving away from the phase retrieval problem, we would like to mention some very interesting work on the matrix completion problem using non-convex schemes by Montanari and coauthors [38]–[40], see also [2], [6], [32], [37], [41], [51], [52]. Although the problems and models are quite different, there are some general similarities between the algorithm named OptSpace in [39] and ours. Indeed, OptSpace operates by computing an initial guess of the solution to a low-rank matrix completion problem by means of a spectral method. It then sets up a nonconvex problem, and proposes an iterative algorithm for solving it. Under suitable assumptions, [39] demonstrates the correctness of this method in the sense that OptSpace will eventually converge to a low-rank solution, although it is not shown to converge in polynomial time.

IV. NUMERICAL EXPERIMENTS

We present some numerical experiments to assess the empirical performance of the Wirtinger flow algorithm. Here, we mostly consider a model of coded diffraction patterns reviewed below.

A. The Coded Diffraction Model

We consider an acquisition model, where we collect data of the form

$$y_r = \left| \sum_{t=0}^{n-1} x[t]\bar{d}_\ell(t)e^{-i2\pi kt/n} \right|^2,$$

$$r = (\ell, k), \quad \begin{array}{l} 0 \le k \le n-1 \\ 1 \le \ell \le L; \end{array} \quad \text{(IV.1)}$$

thus for a fixed ℓ, we collect the magnitude of the diffraction pattern of the signal $\{x(t)\}$ modulated by the waveform/code $\{d_\ell(t)\}$. By varying ℓ and changing the modulation pattern d_ℓ, we generate several views thereby creating a series of *coded diffraction patterns* (CDPs).

In this paper, we are mostly interested in the situation where the modulation patterns are random; in particular, we study a model in which the d_ℓ's are i.i.d. distributed, each having i.i.d. entries sampled from a distribution d. Our theoretical results from Section V assume that d is symmetric, obeys $|d| \le M$ as well as the moment conditions

$$\mathbb{E}\, d = 0, \quad \mathbb{E}\, d^2 = 0, \quad \mathbb{E}\, |d|^4 = 2(\mathbb{E}\, |d|^2)^2. \quad \text{(IV.2)}$$

A random variable obeying these assumptions is said to be *admissible*. Since d is complex valued we can have $\mathbb{E}\, d^2 = 0$ while $d \neq 0$. An example of an admissible random variable is $d = b_1 b_2$, where b_1 and b_2 are independent and distributed as

$$b_1 = \begin{cases} +1 & \text{with prob. } 1/4 \\ -1 & \text{with prob. } 1/4 \\ -i & \text{with prob. } 1/4 \\ +i & \text{with prob. } 1/4, \end{cases} \quad \text{(IV.3)}$$

and

$$b_2 = \begin{cases} \sqrt{2}/2 & \text{with prob. } 4/5 \\ \sqrt{3} & \text{with prob. } 1/5. \end{cases} \quad \text{(IV.4)}$$

We shall refer to this distribution as an *octanary* pattern since d can take on eight distinct values. The condition $\mathbb{E}[d^2] = 0$ is here to avoid unnecessarily complicated calculations in our
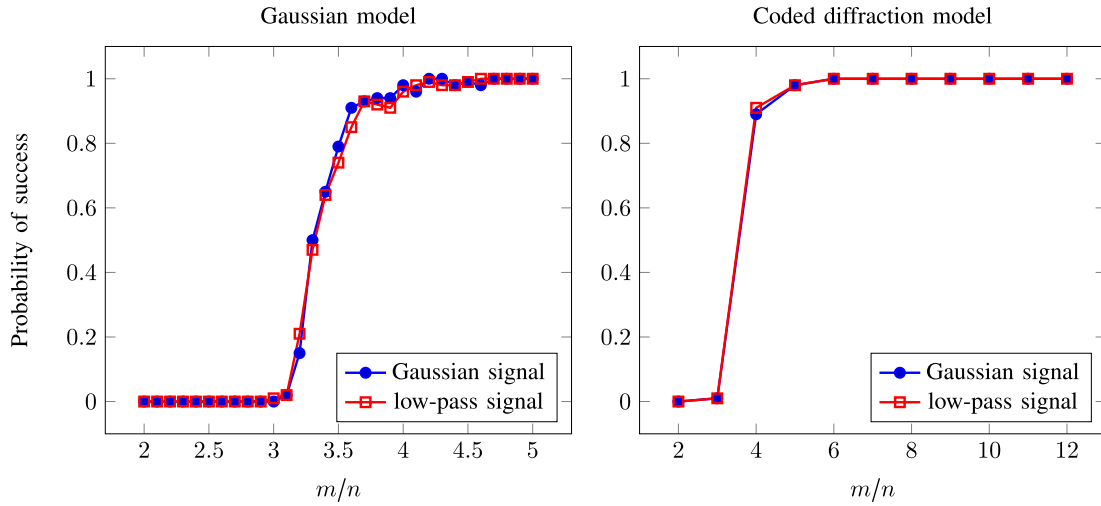
Fig. 2. Empirical probability of success based on 100 random trials for different signal/measurement models and a varied number of measurements. The coded diffraction model uses octanary patterns; the number of patterns $L = m/n$ only takes on integral values.

theoretical analysis. In particular, we can also work with a *ternary* pattern in which d is distributed as

$$d = \begin{cases} +1 & \text{with prob. } 1/4 \\ 0 & \text{with prob. } 1/2 \qquad\qquad \text{(IV.5)} \\ -1 & \text{with prob. } 1/4. \end{cases}$$

We emphasize that the random coded diffraction patterns mentioned above are physically realizable in optical applications specially those that arise in microscopy. However, we should note that phase retrieval has many different applications and in some cases other CDP models may be more suitable for that particular application. We refer to our companion paper [16, Sec. 2.2] for a discussion of other practically relevant models.

B. The Gaussian and Coded Diffraction Models

We begin by examining the performance of the Wirtinger flow algorithm for recovering random signals $x \in \mathbb{C}^n$ under the Gaussian and coded diffraction models. We are interested in signals of two different types:

* *Random Low-Pass Signals:* Here, x is given by

$$x[t] = \sum_{k=-(M/2-1)}^{M/2} (X_k + iY_k)e^{2\pi i(k-1)(t-1)/n},$$

  with $M = n/8$ and X_k and Y_k are i.i.d. $\mathcal{N}(0,1)$.
* *Random Gaussian Signals:* In this model, $x \in \mathbb{C}^n$ is a random complex Gaussian vector with i.i.d. entries of the form $x[t] = X + iY$ with X and Y distributed as $\mathcal{N}(0,1)$; this can be expressed as

$$x[t] = \sum_{k=-(n/2-1)}^{n/2} (X_k + iY_k)e^{2\pi i(k-1)(t-1)/n},$$

  where X_k and Y_k are are i.i.d. $\mathcal{N}(0, 1/8)$ so that the low-pass model is a 'bandlimited' version of this high-pass random model (variances are adjusted so that the expected signal power is the same).

Below, we set $n = 128$, and generate one signal of each type which will be used in all the experiments.

The initialization step of the Wirtinger flow algorithm is run by applying 50 iterations of the power method outlined in Algorithm 3 from Appendix B. In the iteration (II.2), we use the parameter value $\mu_\tau = \min(1 - \exp(-\tau/\tau_0), 0.2)$ where $\tau_0 \approx 330$. We stop after $2,500$ iterations, and report the empirical probability of success for the two different signal models. The empirical probability of succcess is an average over 100 trials, where in each instance, we generate new random sampling vectors according to the Gaussian or CDP models. We declare a trial successful if the relative error of the reconstruction $\mathrm{dist}(\hat{x}, x)/\|x\|$ falls below 10^{-5}.

Figure 2 shows that around $4.5n$ Gaussian phaseless measurements suffice for exact recovery with high probability via the Wirtinger flow algorithm. We also see that about six octanary patterns are sufficient.

C. Performance on Natural Images

We move on to testing the Wirtinger flow algorithm on various images of different sizes; these are photographs of the Naqsh-e Jahan Square in the central Iranian city of Esfahan, the Stanford main quad, and the Milky Way galaxy. Since each photograph is in color, we run the WF algorithm on each of the three RGB images that make up the photograph. Color images are viewed as $n_1 \times n_2 \times 3$ arrays, where the first two indices encode the pixel location, and the last the color band.

We generate $L = 20$ random octanary patterns and gather the coded diffraction patterns for each color band using these 20 samples. As before, we run 50 iterations of the power method as the initialization step. The updates use the sequence $\mu_\tau = \min(1 - \exp(-\tau/\tau_0), 0.4)$ where $\tau_0 \approx 330$ as before. In all cases we run 300 iterations and record the relative recovery error as well as the running time. If x and $\hat{x}$ are the original and recovered images, the relative error is equal to $\|\hat{x} - x\| / \|x\|$, where $\|\cdot\|$ is the Euclidean norm $\|x\|^2 = \sum_{i,j,k} |x(i,j,k)|^2$. The computational time we report

(a)

(b)
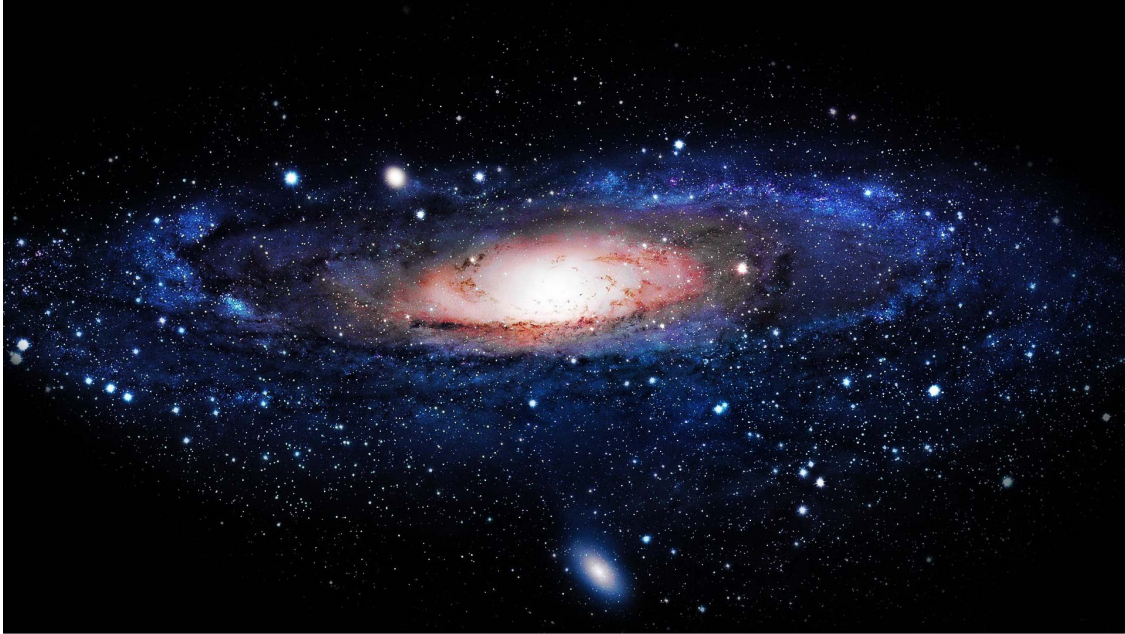
(c)

Fig. 3. Performance of the WF algorithm on three scenic images. Image size, computational time in seconds and in units of FFTs are reported, as well as the relative error after 300 WF iterations. (a) Naqsh-e Jahan Square, Esfahan. Image size is 189×768 pixels; timing is 61.4 sec or about 21,200 FFT units. The relative error is 6.2×10^{-16}. (b) Stanford main quad. Image size is 320×1280 pixels; timing is 181.8120 sec or about 20,700 FFT units. The relative error is 3.5×10^{-14}. (c) Milky way Galaxy. Image size is 1080×1920 pixels; timing is 1318.1 sec or 41,900 FFT units. The relative error is 9.3×10^{-16}.

is the the computational time averaged over the three RGB images. All experiments were carried out on a MacBook Pro with a 2.4 GHz Intel Core i7 Processor and 8 GB 1600 MHz DDR3 memory.

Figure 3 shows the images recovered via the Wirtinger flow algorithm. In all cases, WF gets 12 or 13 digits of precision in a matter of minutes. To convey an idea of timing that is platform-independent, we also report time in units of FFTs; one FFT unit is the amount of time it takes to perform a single FFT on an image of the same size. Now all the workload is dominated by matrix vector products of the form Az and A^*v. In details, each iteration of the power method in
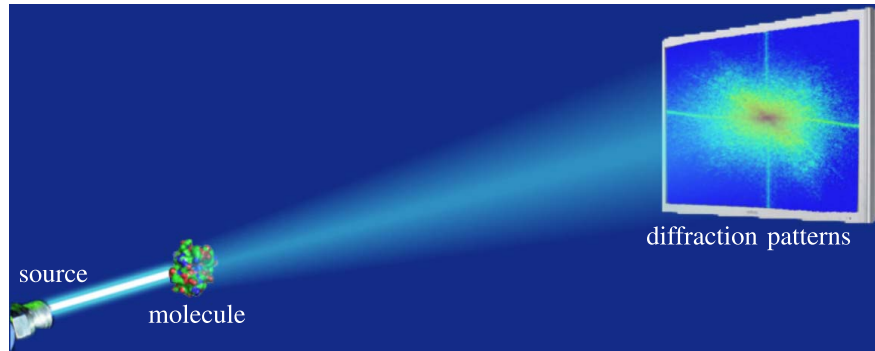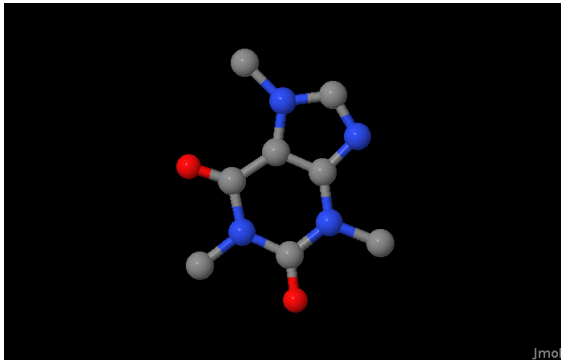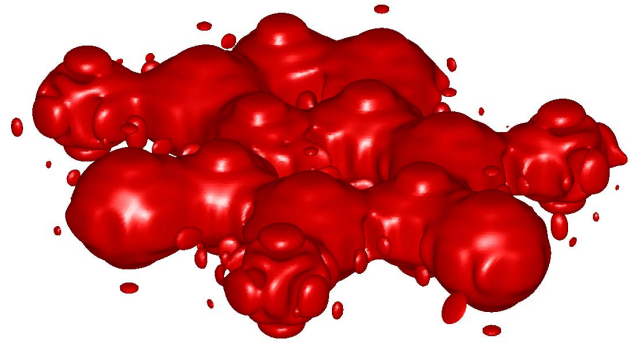
Fig. 4. An illustrative setup of diffraction patterns.

(a)

(b)

Fig. 5. Schematic representation and electron density map of the Caffeine molecule. (a) Schematic representation (b) Electron density map.

the initialization step, or each update (II.2) requires 40 FFTs; the factor of 40 comes from the fact that we have 20 patterns and that each iteration involves one FFT and one adjoint or inverse FFT. Hence, the total number of FFTs is equal to

20 patterns $\times$ 2 (one FFT and one IFFT)

$\times$(300 gradient steps $+$ 50 power iterations) $= 14,000$.

Another way to state this is that the total workload of our algorithm is roughly equal to 350 applications of the sensing matrix A and its adjoint A^*. For about 13 digits of accuracy (relative error of about 10^{-13}), Figure 3 shows that we need between 21,000 and 42,000 FFT units. This is within a factor between 1.5 and 3 of the optimal number computed above. This increase has to do with the fact that in our implementation, certain variables are copied into other temporary variables and these types of operations cause some overhead. This overhead is non-linear and becomes more prominent as the size of the signal increases.

For comparison, SDP based solutions such as PhaseLift [14], [17] and PhaseCut [67] would be prohibitive on a laptop computer as the lifted signal would not fit into memory. In the SDP approach an n pixel image become an $n^2/2$ array, which in the first example already takes storing the lifted signal even for the smallest image requires $(189 \times 768)^2 \times 1/2 \times 8$ Bytes, which is approximately 85 GB of space. (For the image of the Milky Way, storage would be about 17 TB.) These large memory requirements

prevent the application of full-blown SDP solvers on desktop computers.

D. 3D Molecules

Understanding molecular structure is a great contemporary scientific challenge, and several techniques are currently employed to produce 3D images of molecules; these include electron microscopy and X-ray imaging. In X-ray imaging, for instance, the experimentalist illuminates an object of interest, e.g. a molecule, and then collects the intensity of the diffracted rays, please see Figure 4 for an illustrative setup. Figures 5 and 6 show the schematic representation and the corresponding electron density maps for the Caffeine and Nicotine molecules: the density map $\rho(x_1, x_2, x_3)$ is the 3D object we seek to infer. In this paper, we do not go as far 3D reconstruction but demonstrate the performance of the Wirtinger flow algorithm for recovering projections of 3D molecule density maps from simulated data. For related simulations using convex schemes we refer the reader to [28].

To derive signal equations, consider an experimental apparatus as in Figure 4. If we imagine that light propagates in the direction of the x_3-axis, an approximate model for the collected data reads

$$I(f_1, f_2) = \left| \int \left(\int \rho(x_1, x_2, x_3) \mathrm{d}x_3 \right) e^{-2i\pi(f_1 x_1 + f_2 x_2)} \mathrm{d}x_1 \mathrm{d}x_2 \right|^2.$$

In other words, we collect the intensity of the diffraction pattern of the projection $\int \rho(x_1, x_2, x_3) \mathrm{d}x_3$. The 2D image
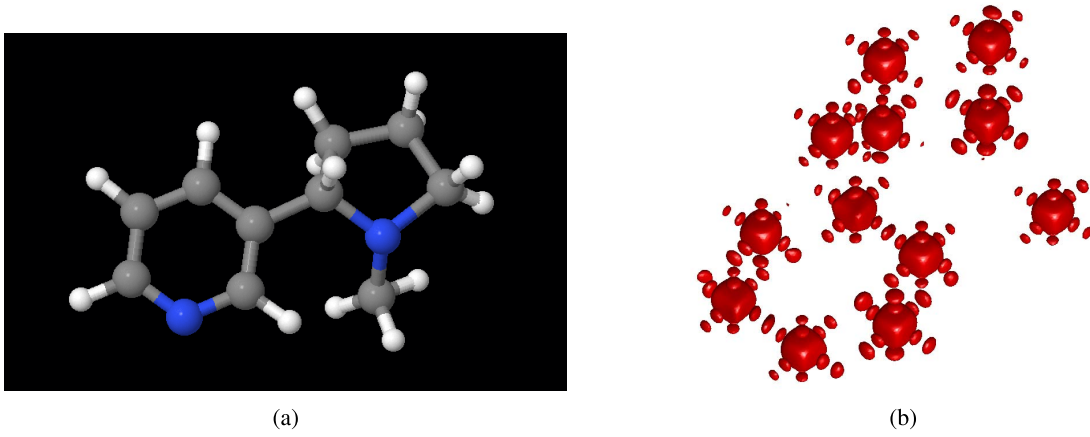
Fig. 6.   Schematic representation and electron density map of the Nicotine molecule. (a) Schematic representation (b) Electron density map.
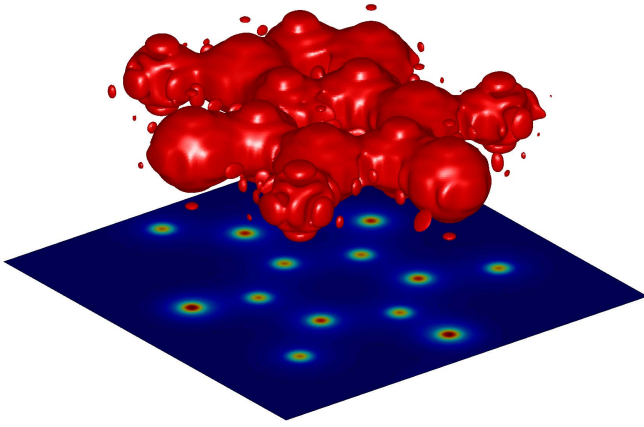

Fig. 7.   Electron density $\rho(x_1, x_2, x_3)$ of the Caffeine molecule along with its projection onto the $x_1 x_2$-plane.

we wish to recover is thus the line integral of the density map along a given direction. As an example, the Caffeine molecule along with its projection on the $x_1 x_2$-plane (line integral in the x_3 direction) is shown in Figure 7. Now, if we let R be the Fourier transform of the density ρ, one can re-express the identity above as

$$I(f_1, f_2) = |R(f_1, f_2, 0)|^2.$$

Therefore, by imputing the missing phase using phase retrieval algorithms, one can recover a slice of the 3D Fourier transfom of the electron density map, i.e. $R(f_1, f_2, 0)$. Viewing the object from different angles or directions gives us different slices. In a second step we do not perform in this paper, one can presumably recover the 3D Fourier transform of the electron density map from all these slices (this is the tomography or blind tomography problem depending upon whether or not the projection angles are known) and, in turn, the 3D electron density map.

We now generate 51 observation planes by rotating the $x_1 x_2$-plane around the x_1-axis by equally spaced angles in the interval $[0, 2\pi]$. Each of these planes is associated with a 2D projection of size 1024×1024, giving us

20 coded diffraction octanary patterns (we use the same patterns for all 51 projections). We run the Wirtinger flow algorithm with exactly the same parameters as in the previous section, and stop after 150 gradient iterations. Figure 8 reports the average relative error over the 51 projections and the total computational time required for reconstructing all 51 images.

V. THEORY FOR THE CODED DIFFRACTION MODEL

We complement our study with theoretical results applying to the model of coded diffraction patterns. These results concern a variation of the Wirtinger flow algorithm: whereas the iterations are exactly the same as (II.2), the initialization applies an iterative scheme which uses fresh sets of sample at each iteration. This is described in Algorithm 2. In the CDP model, the partitioning assigns to the same group all the observations and sampling vectors corresponding to a given realization of the random code. This is equivalent to partitioning the random patterns into $B + 1$ groups. As a result, sampling vectors in distinct groups are stochastically independent.

Theorem 5.1: Let $\boldsymbol{x}$ be an arbitrary vector in $\mathbb{C}^n$ and assume we collect L admissible coded diffraction patterns with $L \geq c_0 \cdot (\log n)^4$, where c_0 is a sufficiently large numerical constant. Then the initial solution $\boldsymbol{z}_0$ of Algorithm 2[5] obeys

$$dist(\boldsymbol{z}_0, \boldsymbol{x}) \leq \frac{1}{8\sqrt{n}} \|\boldsymbol{x}\| \tag{V.1}$$

with probability at least $1 - (4L + 2)/n^3$. Further, take a constant learning parameter sequence, $\mu_\tau = \mu$ for all $\tau = 1, 2, \dots$ and assume $\mu \leq c_1$ for some fixed numerical constant c_1. Then there is an event of probability at least $1 - (2L+1)/n^3 - 1/n^2$, such that on this event, starting from any initial solution $\boldsymbol{z}_0$ obeying (V.1), we have

$$dist(\boldsymbol{z}_\tau, \boldsymbol{x}) \leq \frac{1}{8\sqrt{n}} \left(1 - \frac{\mu}{3}\right)^{\tau/2} \cdot \|\boldsymbol{x}\|. \tag{V.2}$$

[5]We choose the number of partitions B in Algorithm 2 to obey $B \geq c_1 \log n$ for c_1 a sufficiently large numerical constant.
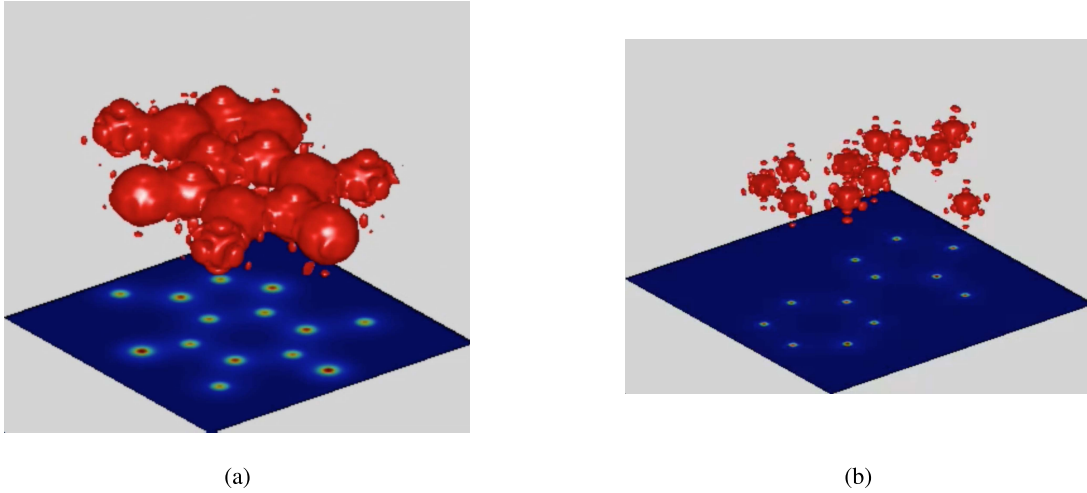
Fig. 8. Reconstruction sequence of the projection of the Caffeine and Nicotine molecules along different directions. To see the videos please download and open the PDF file using Acrobat Reader. (a) Caffeine molecule Mean rel. error is 9.6×10^{-6} Total time is 5.4 hours. (b) Nicotine molecule Mean rel. error is 1.7×10^{-5} Total time is 5.4 hours

---

Algorithm 2 Initialization via Resampled Wirtinger Flow

Input: Observations $\{y_r\} \in \mathbb{R}^m$ and number of blocks B.

Partition the observations and sampling vectors $\{y_r\}_{r=1}^m$ and $\{a_r\}_{r=1}^m$ into $B + 1$ groups of size $m' = \lfloor m/(B+1) \rfloor$. For each group $b = 0, 1, \ldots, B$, set

$$f(z; b) = \frac{1}{2m'} \sum_{r=1}^{m'} \left(y_r^{(b)} - \left| \langle a_r^{(b)}, z \rangle \right|^2 \right)^2,$$

where $\{a_r^{(b)}\}$ are those sampling vectors belonging to the bth group (and likewise for $\{y_r^{(b)}\}$).

Initialize $\tilde{u}_0$ to be eigenvector corresponding to the largest eigenvalue of

$$Y = \frac{1}{m'} \sum_{r=1}^{m'} y_r^{(0)} a_r^{(0)} a_r^{(0)*}$$

normalized as in Algorithm (1).

Loop:

for $b = 0$ **to** $B - 1$ **do**

$$u_{b+1} = u_b - \frac{\tilde{\mu}}{\|u_0\|^2} \nabla f(u_b; b)$$

end for

Output: $z_0 = u_B$.

In the Gaussian model, both statements also hold with high probability provided that $m \geq c_0 \cdot n (\log n)^2$, where c_0 is a sufficiently large numerical constant.

Hence, we achieve perfect recovery from on the order of $n(\log n)^4$ samples arising from a coded diffraction experiment. Our companion paper [16] established that PhaseLift—the SDP relaxation—is also exact with a sampling complexity on the order of $n(\log n)^4$ (this has recently been improved to $n(\log n)^2$ [31]). We believe that the sampling complexity of both approaches (WF and SDP) can be further reduced to

$n \log n$ (or even n for certain kind of patterns). We leave this to future research.

Setting $\mu = c_1$ yields ϵ accuracy in $\mathcal{O}(\log 1/\epsilon)$ iterations. As the computational work at each iteration is dominated by two matrix-vector products of the form Az and A^*v, it follows that the overall computational is at most $\mathcal{O}(nL \log n \log 1/\epsilon)$. In particular, this approach yields a near-linear time algorithm in the CDP model (linear in the dimension of the signal n). In the Gaussian model, the complexity scales like $\mathcal{O}(mn \log 1/\epsilon)$.

VI. WIRTINGER DERIVATIVES

Our gradient step (II.2) uses a notion of derivative, which can be interpreted as a Wirtinger derivative. The purpose of this section is thus to gather some results concerning Wirtinger derivatives of real valued functions over complex variables. Here and below, M^T is the transpose of the matrix M, and $\bar{c}$ denotes the complex conjugate of a scalar $c \in \mathbb{C}$. Similarly, the matrix $\bar{M}$ is obtained by taking complex conjugates of the elements of M.

Any complex-or real-valued function

$$f(z) = f(x, y) = u(x, y) + iv(x, y)$$

of several complex variables can be written in the form $f(z, \bar{z})$, where f is holomorphic in $z = x + iy$ for fixed $\bar{z}$ and holomorphic in $\bar{z} = x - iy$ for fixed z. This holds as long as a the real-valued functions u and v are differentiable as functions of the real variables x and y. As an example, consider

$$f(z) = \left(y - |a^*z|^2 \right)^2 = (y - \bar{z}^T aa^* z)^2 = f(z, \bar{z}).$$

with $z, a \in \mathbb{C}^n$ and $y \in \mathbb{R}$. While $f(z)$ is not holomorphic in z, $f(z, \bar{z})$ is holomorphic in z for a fixed $\bar{z}$, and vice versa.

This fact underlies the development of the *Wirtinger calculus*. In essence, the *conjugate coordinates*

$$\begin{bmatrix} z \\ \bar{z} \end{bmatrix} \in \mathbb{C}^n \times \mathbb{C}^n, \quad z = x + iy \text{ and } \bar{z} = x - iy,$$

can serve as a formal substitute for the representation $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{2n}$. This leads to the following derivatives

$$
\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{z}} &:= \frac{\partial f(\boldsymbol{z}, \bar{\boldsymbol{z}})}{\partial \boldsymbol{z}}\big|_{\bar{\boldsymbol{z}}=\text{constant}} \\
&= \left[\frac{\partial f}{\partial z_1}, \frac{\partial f}{\partial z_2}, \ldots, \frac{\partial f}{\partial z_n}\right]_{\bar{\boldsymbol{z}}=\text{constant}}, \\
\frac{\partial f}{\partial \bar{\boldsymbol{z}}} &:= \frac{\partial f(\boldsymbol{z}, \bar{\boldsymbol{z}})}{\partial \bar{\boldsymbol{z}}}\big|_{\boldsymbol{z}=\text{constant}} \\
&= \left[\frac{\partial f}{\partial \bar{z}_1}, \frac{\partial f}{\partial \bar{z}_2}, \ldots, \frac{\partial f}{\partial \bar{z}_n}\right]_{\boldsymbol{z}=\text{constant}}.
\end{aligned}
$$

Our definitions follow standard notation from multivariate calculus so that derivatives are row vectors and gradients are column vectors. In this new coordinate system the complex gradient is given by

$$
\nabla_c f = \left[\frac{\partial f}{\partial \boldsymbol{z}}, \frac{\partial f}{\partial \bar{\boldsymbol{z}}}\right]^*.
$$

Similarly, we define

$$
\begin{aligned}
\mathcal{H}_{\boldsymbol{z}\boldsymbol{z}} &:= \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial f}{\partial \boldsymbol{z}}\right)^*, \quad \mathcal{H}_{\bar{\boldsymbol{z}}\boldsymbol{z}} := \frac{\partial}{\partial \bar{\boldsymbol{z}}}\left(\frac{\partial f}{\partial \boldsymbol{z}}\right)^*, \\
\mathcal{H}_{\boldsymbol{z}\bar{\boldsymbol{z}}} &:= \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{\partial f}{\partial \bar{\boldsymbol{z}}}\right)^*, \quad \mathcal{H}_{\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}} := \frac{\partial}{\partial \bar{\boldsymbol{z}}}\left(\frac{\partial f}{\partial \bar{\boldsymbol{z}}}\right)^*.
\end{aligned}
$$

In this coordinate system the complex Hessian is given by

$$
\nabla^2 f := \begin{bmatrix} \mathcal{H}_{\boldsymbol{z}\boldsymbol{z}} & \mathcal{H}_{\bar{\boldsymbol{z}}\boldsymbol{z}} \\ \mathcal{H}_{\boldsymbol{z}\bar{\boldsymbol{z}}} & \mathcal{H}_{\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}} \end{bmatrix}.
$$

Given vectors $\boldsymbol{z}$ and $\Delta \boldsymbol{z} \in \mathbb{C}^n$, we have defined the gradient and Hessian in a manner such that Taylor's approximation takes the form

$$
\begin{aligned}
f(\boldsymbol{z} + \Delta \boldsymbol{z}) &\approx f(\boldsymbol{z}) + (\nabla_c f(\boldsymbol{z}))^* \begin{bmatrix} \Delta \boldsymbol{z} \\ \overline{\Delta \boldsymbol{z}} \end{bmatrix} \\
&+ \frac{1}{2} \begin{bmatrix} \Delta \boldsymbol{z} \\ \overline{\Delta \boldsymbol{z}} \end{bmatrix}^* \nabla^2 f(\boldsymbol{z}) \begin{bmatrix} \Delta \boldsymbol{z} \\ \overline{\Delta \boldsymbol{z}} \end{bmatrix}.
\end{aligned}
$$

If we were to run gradient descent in this new coordinate system, the iterates would be

$$
\begin{bmatrix} \boldsymbol{z}_{\tau+1} \\ \bar{\boldsymbol{z}}_{\tau+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{z}_{\tau} \\ \bar{\boldsymbol{z}}_{\tau} \end{bmatrix} - \mu \begin{bmatrix} (\partial f/\partial \boldsymbol{z})^*|_{\boldsymbol{z}=\boldsymbol{z}_\tau} \\ (\partial f/\partial \boldsymbol{z})^*|_{\boldsymbol{z}=\boldsymbol{z}_\tau} \end{bmatrix} \quad \text{(VI.1)}
$$

Note that when f is a real-valued function (as in this paper) we have

$$
\overline{\frac{\partial f}{\partial \boldsymbol{z}}} = \frac{\partial f}{\partial \bar{\boldsymbol{z}}}.
$$

Therefore, the second set of updates in (VI.1) is just the conjugate of the first. Thus, it is sufficient to keep track of the first update, namely,

$$
\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \mu \, (\partial f/\partial \boldsymbol{z})^*.
$$

For real valued functions of complex variables, setting

$$
\nabla f(\boldsymbol{z}) = \left(\frac{\partial f}{\partial \boldsymbol{z}}\right)^*
$$

gives the gradient update

$$
\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \mu \nabla f(\boldsymbol{z}_\tau).
$$

The reader may wonder why we choose to work with conjugate coordinates as there are alternatives: in particular, we could view the complex variable $\boldsymbol{z} = \boldsymbol{x} + i\boldsymbol{y} \in \mathbb{C}^n$ as a vector in $\mathbb{R}^{2n}$ and just run gradient descent in the $\boldsymbol{x}, \boldsymbol{y}$ coordinate system. The main reason why conjugate coordinates are particularly attractive is that expressions for derivatives become significantly simpler and resemble those we obtain in the real case, where $f : \mathbb{R}^n \to \mathbb{R}$ is a function of real variables.

VII. PROOFS

A. Preliminaries

We first note that in the CDP model with admissible CDPs $\|\boldsymbol{a}_r\| \leq \sqrt{6n}$ for all $r = 1, 2, \ldots, m$, as the entries of the CDPs obey $|d| \leq \sqrt{3} < \sqrt{6}$. In the Gaussian model the measurements vectors also obey $\|\boldsymbol{a}_r\| \leq \sqrt{6n}$ for all $r = 1, 2, \ldots, m$ with probability at least $1 - me^{-1.5n}$. Throughout the proofs, we assume we are on this event without explicitly mentioning it each time.

Before we begin with the proofs we should mention that we will prove our result using the update

$$
\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \frac{\mu}{\|\boldsymbol{x}\|^2} \nabla f(\boldsymbol{z}_\tau), \quad \text{(VII.1)}
$$

in lieu of the WF update

$$
\boldsymbol{z}_{\tau+1} = \boldsymbol{z}_\tau - \frac{\mu_{\text{WF}}}{\|\boldsymbol{z}_0\|^2} \nabla f(\boldsymbol{z}_\tau). \quad \text{(VII.2)}
$$

Since $\left|\|\boldsymbol{z}_0\|^2 - \|\boldsymbol{x}\|^2\right| \leq \frac{1}{64}\|\boldsymbol{x}\|^2$ holds with high probability as proven in Section VII-H, we have

$$
\|\boldsymbol{z}_0\|^2 \geq \frac{63}{64}\|\boldsymbol{x}\|^2. \quad \text{(VII.3)}
$$

Therefore, the results for the update (VII.1) automatically carry over to the WF update with a simple rescaling of the upper bound on the learning parameter. More precisely, if we prove that the update (VII.1) converges to a global optimum as long as $\mu \leq \mu_0$, then the convergence of the WF update to a global optimum is guaranteed as long as $\mu_{\text{WF}} \leq \frac{63}{64}\mu_0$. Also, the update in (VII.1) is invariant to the Euclidean norm of $\boldsymbol{x}$. Therefore, without loss of generality we will assume throughout the proofs that $\|\boldsymbol{x}\| = 1$.

We remind the reader that throughout $\boldsymbol{x}$ is a solution to our quadratic equations, i.e. obeys $\boldsymbol{y} = |\boldsymbol{Ax}|^2$ and that the sampling vectors are independent from $\boldsymbol{x}$. Define

$$
P := \{\boldsymbol{x}e^{i\phi} : \phi \in [0, 2\pi]\}.
$$

to be the set of all vectors that differ from the planted solution $\boldsymbol{x}$ only by a global phase factor. We also introduce the set of all points that are close to P,

$$
E(\epsilon) := \{\boldsymbol{z} \in \mathbb{C}^n : \text{dist}(\boldsymbol{z}, P) \leq \epsilon\}, \quad \text{(VII.4)}
$$

Finally for any vector $\boldsymbol{z} \in \mathbb{C}^n$ we define the phase $\phi(\boldsymbol{z})$ as

$$
\phi(\boldsymbol{z}) := \underset{\phi \in [0, 2\pi]}{\arg\min} \, \left\|\boldsymbol{z} - e^{i\phi}\boldsymbol{x}\right\|,
$$

so that

$$
\text{dist}(\boldsymbol{z}, \boldsymbol{x}) = \left\|\boldsymbol{z} - e^{i\phi(\boldsymbol{z})}\boldsymbol{x}\right\|.
$$

B. Formulas for the Complex Gradient and Hessian

We gather some useful gradient and Hessian calculations that will be used repeatedly. Starting with

$$f(z) = \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - \bar{z}^T (a_r a_r^*) z \right)^2$$

$$= \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - z^T (a_r a_r^*)^T \bar{z} \right)^2,$$

we establish

$$\left(\frac{\partial}{\partial z} f(z) \right)^T = \frac{1}{m} \sum_{r=1}^{m} \left(z^T (a_r a_r^*)^T \bar{z} - y_r \right) (a_r a_r^*)^T \bar{z}.$$

This gives

$$\nabla f(z) = \left(\frac{\partial}{\partial z} f(z) \right)^* = \frac{1}{m} \sum_{r=1}^{m} \left(\bar{z}^T (a_r a_r^*) z - y_r \right) (a_r a_r^*) z.$$

For the second derivative, we write

$$\mathcal{H}_{zz} = \frac{\partial}{\partial z} \left(\frac{\partial}{\partial z} f(z) \right)^* = \frac{1}{m} \sum_{r=1}^{m} \left(2 |a_r^* z|^2 - y_r \right) a_r a_r^*$$

and

$$\mathcal{H}_{\bar{z} z} = \frac{\partial}{\partial \bar{z}} \left(\frac{\partial}{\partial z} f(z) \right)^* = \frac{1}{m} \sum_{r=1}^{m} (a_r^* z)^2 a_r a_r^T.$$

Therefore,

$$\nabla^2 f(z)$$
$$= \frac{1}{m} \sum_{r=1}^{m} \begin{bmatrix} (2 |a_r^* z|^2 - y_r) a_r a_r^* & (a_r^* z)^2 a_r a_r^T \\ (\overline{a_r^* z})^2 \bar{a}_r a_r^* & (2 |a_r^* z|^2 - y_r) \bar{a}_r a_r^T \end{bmatrix}.$$

C. Expectation and Concentration

This section gathers some useful intermediate results whose proofs are deferred to Appendix A. The first two lemmas establish the expectation of the Hessian, gradient and a related random variable in both the Gaussian and admissible CDP models.[6]

Lemma 7.1: Recall that x is a solution obeying $\|x\| = 1$, which is independent from the sampling vectors. Furthermore, assume the sampling vectors a_r are distributed according to either the Gaussian or admissible CDP model. Then

$$\mathbb{E}[\nabla^2 f(x)] = I_{2n} + \frac{3}{2} \begin{bmatrix} x \\ \bar{x} \end{bmatrix} [x^*, x^T] - \frac{1}{2} \begin{bmatrix} x \\ -\bar{x} \end{bmatrix} [x^*, -x^T].$$

Lemma 7.2: In the setup of Lemma 7.1, let $z \in \mathbb{C}^n$ be a fixed vector independent of the sampling vectors. We have

$$\mathbb{E}[\nabla f(z)] = (I - x x^*) z + 2 \left(\|z\|^2 - 1 \right) z.$$

The next lemma gathers some useful identities in the Gaussian model.

Lemma 7.3: Assume $u, v \in \mathbb{C}^n$ are fixed vectors obeying $\|u\| = \|v\| = 1$ which are independent of the sampling

---

[6]In the CDP model the expectation is with respect to the random modulation pattern.

vectors. Furthermore, assume the measurement vectors a_r are distributed according to the Gaussian model. Then

$$\mathbb{E}\left[\left(\mathrm{Re}(u^* a_r a_r^* v) \right)^2 \right] = \frac{1}{2} + \frac{3}{2} (\mathrm{Re}(u^* v))^2 - \frac{1}{2} (\mathrm{Im}(u^* v))^2 \tag{VII.5}$$

$$\mathbb{E}\left[\mathrm{Re}(u^* a_r a_r^* v) \left| a_r^* v \right|^2 \right] = 2 \mathrm{Re}(u^* v) \tag{VII.6}$$

$$\mathbb{E}\left[\left| a_r^* v \right|^{2k} \right] = k!. \tag{VII.7}$$

The next lemma establishes the concentration of the Hessian around its mean for both the Gaussian and the CDP model.

Lemma 7.4: In the setup of Lemma 7.1, assume the vectors a_r are distributed according to either the Gaussian or admissible CDP model with a sufficiently large number of measurements. This means that the number of samples obeys $m \geq c(\delta) \cdot n \log n$ in the Gaussian model and the number of patterns obeys $L \geq c(\delta) \cdot \log^3 n$ in the CDP model. Then

$$\left\| \nabla^2 f(x) - \mathbb{E}[\nabla^2 f(x)] \right\| \leq \delta, \tag{VII.8}$$

holds with probability at least $1 - 10 e^{-\gamma n} - 8/n^2$ and $1 - (2L+1)/n^3$ for the Gaussian and CDP models, respectively.

We will also make use of the two results below, which are corollaries of the three lemmas above. These corollaries are also proven in Appendix A.

Corollary 7.5: Suppose $\left\| \nabla^2 f(x) - \mathbb{E}[\nabla^2 f(x)] \right\| \leq \delta$. Then for all $h \in \mathbb{C}^n$ obeying $\|h\| = 1$, we have

$$\frac{1}{m} \sum_{r=1}^{m} \mathrm{Re} \left(h^* a_r a_r^* x \right)^2 = \frac{1}{4} \sum_{r=1}^{m} \begin{bmatrix} h \\ \bar{h} \end{bmatrix}^* \nabla^2 f(x) \begin{bmatrix} h \\ \bar{h} \end{bmatrix}$$

$$\leq \frac{1}{2} \|h\|^2 + \frac{3}{2} \mathrm{Re}(x^* h)^2$$

$$- \frac{1}{2} \mathrm{Im}(x^* h)^2 + \frac{\delta}{2}.$$

In the other direction,

$$\frac{1}{m} \sum_{r=1}^{m} \mathrm{Re} \left(h^* a_r a_r^* x \right)^2 \geq \frac{1}{2} \|h\|^2 + \frac{3}{2} \mathrm{Re}(x^* h)^2$$

$$- \frac{1}{2} \mathrm{Im}(x^* h)^2 - \frac{\delta}{2}.$$

Corollary 7.6: Suppose $\left\| \nabla^2 f(x) - \mathbb{E}[\nabla^2 f(x)] \right\| \leq \delta$. Then for all $h \in \mathbb{C}^n$ obeying $\|h\| = 1$, we have

$$\frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 |a_r^* h|^2 = h^* \left(\frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 a_r a_r^* \right) h$$

$$\geq (1-\delta) \|h\|^2 + |h^* x|^2$$

$$\geq (1-\delta) \|h\|^2,$$

and

$$\frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 |a_r^* h|^2 = h^* \left(\frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 a_r a_r^* \right) h$$

$$\leq (1+\delta) \|h\|^2 + |h^* x|^2$$

$$\leq (2+\delta) \|h\|^2.$$

The next lemma establishes the concentration of the gradient around its mean for both Gaussian and admissible CDP models.

Lemma 7.7: In the setup of Lemma 7.4, let $z \in \mathbb{C}^n$ be a fixed vector independent of the sampling vectors obeying $dist(z, x) \leq \frac{1}{2}$. Then

$$\|\nabla f(z) - \mathbb{E}[\nabla f(z)]\| \leq \delta \cdot dist(z, x).$$

holds with probability at least $1 - 20e^{-\gamma m} - 4m/n^4$ in the Gaussian model and $1 - (4L + 2)/n^3$ in the CDP model.

We finish with a result concerning the concentration of the sample covariance matrix.

Lemma 7.8: In the setup of Lemma 7.4,

$$\left\| I_n - m^{-1} \sum_{r=1}^{m} a_r a_r^* \right\| \leq \delta,$$

holds with probability at least $1 - 2e^{-\gamma m}$ for the Gaussian model and $1 - 1/n^2$ in the CDP model. On this event,

$$(1 - \delta) \|h\|^2 \leq \frac{1}{m} \sum_{r=1}^{m} |a_r^* h|^2 \leq (1 + \delta) \|h\|^2, \quad \forall h \in \mathbb{C}^n.$$
(VII.9)

D. General Convergence Analysis

We will assume that the function f satisfies a regularity condition on $E(\epsilon)$, which essentially states that the gradient of the function is well behaved. We remind the reader that $E(\epsilon)$, as defined in (VII.4), is the set of points that are close to the path of global minimizers.

Condition 7.9 (Regularity Condition): We say that the function f satisfies the regularity condition or $RC(\alpha, \beta, \epsilon)$ if for all vectors $z \in E(\epsilon)$ we have

$$\text{Re} \left(\langle \nabla f(z), z - x e^{i\phi(z)} \rangle \right) \geq \frac{1}{\alpha} dist^2(z, x) + \frac{1}{\beta} \|\nabla f(z)\|^2.$$
(VII.10)

In the lemma below we show that as long as the regularity condition holds on $E(\epsilon)$ then Wirtinger Flow starting from an initial solution in $E(\epsilon)$ converges to a global optimizer at a geometric rate. Subsequent sections shall establish that this property holds.

Lemma 7.10: Assume that f obeys $RC(\alpha, \beta, \epsilon)$ for all $z \in E(\epsilon)$. Furthermore, suppose $z_0 \in E$, and assume $0 < \mu \leq 2/\beta$. Consider the following update

$$z_{\tau+1} = z_\tau - \mu \nabla f(z_\tau).$$

Then for all τ we have $z_\tau \in E(\epsilon)$ and

$$dist^2(z_\tau, x) \leq \left(1 - \frac{2\mu}{\alpha} \right)^\tau dist^2(z_0, x).$$

We note that for $\alpha\beta < 4$, (VII.10) holds with the direction of the inequality reversed.[7] Thus, if $RC(\alpha, \beta, \epsilon)$ holds, α and β must obey $\alpha\beta \geq 4$. As a result, under the stated assumptions of Lemma 7.10 above, the factor $1 - 2\mu/\alpha \geq 1 - 4/(\alpha\beta)$ is non-negative.

Proof: The proof follows a structure similar to related results in the convex optimization literature see [55, Th. 2.1.15]. However, unlike these classical

[7]One can see this by applying Cauchy-Schwarz and calculating the determinant of the resulting quadratic form.

results where the goal is often to prove convergence to a unique global optimum, the objective function f does not have a unique global optimum. Indeed, in our problem, if x is solution, then $e^{i\phi} x$ is also solution. Hence, proper modification is required to prove convergence results.

We prove that if $z \in E(\epsilon)$ then for all $0 < \mu \leq 2/\beta$

$$z_+ = z - \mu \nabla f(z)$$

obeys

$$dist^2(z_+, x) \leq \left(1 - \frac{2\mu}{\alpha} \right) dist^2(z, x). \quad \text{(VII.11)}$$

Therefore, if $z \in E(\epsilon)$ then we also have $z_+ \in E(\epsilon)$. The lemma follows by inductively applying (VII.11). Now simple algebraic manipulations together with the regularity condition (VII.10) give

$$\begin{aligned}
\left\| z_+ - x e^{i\phi(z)} \right\|^2 &= \left\| z - x e^{i\phi(z)} - \mu \nabla f(z) \right\|^2 \\
&= \left\| z - x e^{i\phi(z)} \right\|^2 \\
&\quad - 2\mu \, \text{Re} \left(\langle \nabla f(z), \left(z - x e^{i\phi(z)} \right) \rangle \right) \\
&\quad + \mu^2 \|\nabla f(z)\|^2 \\
&\leq \left\| z - x e^{i\phi(z)} \right\|^2 \\
&\quad - 2\mu \left(\frac{1}{\alpha} \left\| z - x e^{i\phi(z)} \right\|^2 + \frac{1}{\beta} \|\nabla f(z)\|^2 \right) \\
&\quad + \mu^2 \|\nabla f(z)\|^2 \\
&= \left(1 - \frac{2\mu}{\alpha} \right) \left\| z - x e^{i\phi(z)} \right\|^2 \\
&\quad + \mu \left(\mu - \frac{2}{\beta} \right) \|\nabla f(z)\|^2 \\
&\leq \left(1 - \frac{2\mu}{\alpha} \right) \left\| z - x e^{i\phi(z)} \right\|^2,
\end{aligned}$$

where the last line follows from $\mu \leq 2/\beta$. The definition of $\phi(z_+)$ gives

$$\left\| z_+ - x e^{i\phi(z_+)} \right\|^2 \leq \left\| z_+ - x e^{i\phi(z)} \right\|^2,$$

which concludes the proof. ∎

E. Proof of the Regularity Condition

For any $z \in E(\epsilon)$, we need to show that

$$\text{Re} \left(\langle \nabla f(z), z - x e^{i\phi(z)} \rangle \right) \geq \frac{1}{\alpha} dist^2(z, x) + \frac{1}{\beta} \|\nabla f(z)\|^2.$$
(VII.12)

We prove this with $\delta = 0.01$ by establishing that our gradient satisfies the local smoothness and local curvature conditions defined below. Combining both these two properties gives (VII.12).

Condition 7.11 (Local Curvature Condition): We say that the function f satisfies the local curvature condition or

$LCC(\alpha, \epsilon, \delta)$ if for all vectors $z \in E(\epsilon)$,

$$\mathrm{Re}\left(\langle \nabla f(z), z - x e^{i\phi(z)}\rangle\right) \geq \left(\frac{1}{\alpha} + \frac{(1-\delta)}{4}\right) dist^2(z, x)$$
$$+ \frac{1}{10m}\sum_{r=1}^{m}\left|a_r^*(z - e^{i\phi(z)}x)\right|^4.$$
(VII.13)

This condition essentially states that the function curves sufficiently upwards (along most directions) near the curve of global optimizers.

Condition 7.12 (Local Smoothness Condition): We say that the function f satisfies the local smoothness condition or $LSC(\beta, \epsilon, \delta)$ if for all vectors $z \in E(\epsilon)$ we have

$$\|\nabla f(z)\|^2$$
$$\leq \beta\left(\frac{(1-\delta)}{4} dist^2(z, x) + \frac{1}{10m}\sum_{r=1}^{m}\left|a_r^*(z - e^{i\phi(z)}x)\right|^4\right).$$
(VII.14)

This condition essentially states that the gradient of the function is well behaved (the function does not vary too much) near the curve of global optimizers.

F. Proof of the Local Curvature Condition

For any $z \in E(\epsilon)$, we want to prove the local curvature condition (VII.13). Recall that

$$\nabla f(z) = \frac{1}{m}\sum_{r=1}^{m}\left(|\langle a_r, z\rangle|^2 - y_r\right)(a_r a_r^*)z,$$

and define $h := e^{-i\phi(z)}z - x$. To establish (VII.13) it suffices to prove that

$$\frac{1}{m}\sum_{r=1}^{m}\left(2\,\mathrm{Re}(h^*a_r a_r^* x)^2 + 3\,\mathrm{Re}(h^*a_r a_r^* x)|a_r^*h|^2 + |a_r^*h|^4\right)$$
$$\geq \left(\frac{1}{10m}\sum_{r=1}^{m}|a_r^*h|^4\right) + \left(\frac{1}{\alpha} + \frac{(1-\delta)}{4}\right)\|h\|^2, \quad \text{(VII.15)}$$

holds for all h satisfying $\mathrm{Im}(h^*x) = 0$, $\|h\|_2 \leq \epsilon$. Equivalently, we only need to prove that for all h satisfying $\mathrm{Im}(h^*x) = 0$, $\|h\|_2 = 1$ and for all s with $0 \leq s \leq \epsilon$,

$$\frac{1}{m}\sum_{r=1}^{m}\left(2\,\mathrm{Re}(h^*a_r a_r^* x)^2 + 3s\,\mathrm{Re}(h^*a_r a_r^* x)|a_r^*h|^2\right.$$
$$\left. + \frac{9}{10}s^2\left|a_r^*h\right|^4\right) \geq \frac{1}{\alpha} + \frac{(1-\delta)}{4}. \quad \text{(VII.16)}$$

By Corollary 7.5, with high probability,

$$\frac{1}{m}\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2 \leq \frac{1+\delta}{2} + \frac{3}{2}\mathrm{Re}(x^*h)^2,$$

holds for all h obeying $\|h\| = 1$. Therefore, to establish the local curvature condition (VII.13) it suffices to show that

$$\frac{1}{m}\sum_{r=1}^{m}\left(\frac{5}{2}\mathrm{Re}(h^*a_r a_r^* x)^2 + 3s\,\mathrm{Re}(h^*a_r a_r^* x)|a_r^*h|^2\right.$$
$$\left. + \frac{9}{10}s^2\left|a_r^*h\right|^4\right) \geq \left(\frac{1}{\alpha} + \frac{1}{2}\right) + \frac{3}{4}\mathrm{Re}(x^*h)^2. \quad \text{(VII.17)}$$

We will establish (VII.17) for different measurement models and different values of ϵ. Below, it shall be convenient to use the shorthand

$$Y_r(h, s) := \frac{5}{2}\mathrm{Re}(h^*a_r a_r^* x)^2 + 3s\,\mathrm{Re}(h^*a_r a_r^* x)|a_r^*h|^2$$
$$+ \frac{9}{10}s^2|a_r^*h|^4,$$

$$\langle Y_r(h, s)\rangle := \frac{1}{m}\sum_{r=1}^{m}Y_r(h, s).$$

1) Proof of (VII.17) With $\epsilon = 1/8\sqrt{n}$ in the Gaussian and CDP Models: Set $\epsilon = 1/8\sqrt{n}$. We show that with high probability, (VII.17) holds for all h satisfying $\mathrm{Im}(h^*x) = 0$, $\|h\|_2 = 1$, $0 \leq s \leq \epsilon$, $\delta \leq 0.01$, and $\alpha \geq 30$. First, note that by Cauchy-Schwarz inequality,

$$\langle Y_r(h, s)\rangle$$
$$\geq \frac{5}{2m}\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2$$
$$- \frac{3s}{m}\sqrt{\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2}\sqrt{\sum_{r=1}^{m}|a_r^*h|^4}$$
$$+ \frac{9}{10}\frac{s^2}{m}\sum_{r=1}^{m}|a_r^*h|^4$$
$$= \left(\sqrt{\frac{5}{2m}\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2} - s\sqrt{\frac{9}{10m}\sum_{r=1}^{m}|a_r^*h|^4}\right)^2$$
$$\geq \frac{5}{4m}\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2 - \frac{9s^2}{10m}\sum_{r=1}^{m}|a_r^*h|^4. \quad \text{(VII.18)}$$

The last inequality follows from $(a - b)^2 \geq \frac{a^2}{2} - b^2$. By Corollary 7.5,

$$\frac{1}{m}\sum_{r=1}^{m}\mathrm{Re}(h^*a_r a_r^* x)^2 \geq \frac{1-\delta}{2} + \frac{3}{2}\mathrm{Re}(x^*h)^2 \quad \text{(VII.19)}$$

holds with high probability for all h obeying $\|h\| = 1$. Furthermore, by applying Lemma 7.8,

$$\frac{1}{m}\sum_{r=1}^{m}|a_r^*h|^4 \leq (\max_r \|a_r\|^2)\left(\frac{1}{m}\sum_{r=1}^{m}|a_r^*h|^2\right) \leq 6(1+\delta)n \quad \text{(VII.20)}$$

holds with high probability. Plugging (VII.19) and (VII.20) in (VII.18) yields

$$\langle Y_r(h, s)\rangle \geq \frac{15}{8}\mathrm{Re}(x^*h)^2 + \frac{5}{8}(1-\delta) - \frac{27}{5}s^2(1+\delta)n.$$

(VII.17) follows by using $\alpha \geq 30$, $\epsilon = \frac{1}{8\sqrt{n}}$ and $\delta = 0.01$.

2) Proof of (VII.17) With $\epsilon = 1/8$ in the Gaussian Model: Set $\epsilon = 1/8$. We show that with high probability, (VII.17) holds for all h satisfying $\mathrm{Im}(h^*x) = 0$, $\|h\|_2 = 1$, $0 \leq s \leq \epsilon$, $\delta \leq 2$, and $\alpha \geq 8$. To this end, we first state a result about the tail of a sum of i.i.d. random variables. Below, Φ is the cumulative distribution function of a standard normal variable.

Lemma 7.13 [11]: Suppose $X_1, X_2, \ldots, X_m$ are i.i.d. real-valued random variables obeying $X_r \leq b$ for some nonrandom $b > 0$, $\mathbb{E} X_r = 0$, and $\mathbb{E} X_r^2 = v^2$. Setting $\sigma^2 = m \max(b^2, v^2)$,

$$P(X_1 + \ldots + X_m \geq y)$$
$$\leq \min\left(\exp\left(-\frac{y^2}{2\sigma^2}\right), c_0 \left(1 - \Phi(y/\sigma)\right)\right),$$

where one can take $c_0 = 25$.

To establish (VII.17) we first prove it for a fixed $\boldsymbol{h}$, and then use a covering argument. Observe that

$$Y_r := Y_r(\boldsymbol{h}, s) = \left(\sqrt{\frac{5}{2}} \operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}) + \sqrt{\frac{9}{10}} s \left|\boldsymbol{a}_r^* \boldsymbol{h}\right|^2\right)^2.$$

By Lemma 7.3,

$$\mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})^2] = \frac{1}{2} + \frac{3}{2}(\operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h}))^2$$

and

$$\mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})|\boldsymbol{a}_r^* \boldsymbol{h}|^2] = 2 \operatorname{Re}(\boldsymbol{u}^* \boldsymbol{v}),$$

compare (VII.5) and (VII.6). Therefore, using $s \leq \frac{1}{8}$,

$$\mu_r = \mathbb{E} Y_r = \frac{5}{4}(1 + 3 \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2) + 6s \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h}) + \frac{27}{10}s^2 < 6.$$

Now define $X_r = \mu_r - Y_r$. First, since $Y_r \geq 0$, $X_r \leq \mu_r < 6$. Second, we bound $\mathbb{E} X_r^2$ using Lemma 7.3 and Holder's inequality with $s \leq 1/8$:

$$\mathbb{E} X_r^2 \leq \mathbb{E} Y_r^2 = \frac{25}{4} \mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})^4] + \frac{81}{100}s^4 \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^8]$$
$$+ \frac{27}{2}s^2 \mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})^2 |\boldsymbol{a}_r^* \boldsymbol{h}|^4]$$
$$+ 15s \mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})^3 |\boldsymbol{a}_r^* \boldsymbol{h}|^2]$$
$$+ \frac{27}{5}s^3 \mathbb{E}[\operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}) |\boldsymbol{a}_r^* \boldsymbol{h}|^6]$$
$$\leq \frac{25}{4}\sqrt{\mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^8] \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{x}|^8]}$$
$$+ \frac{81}{100}s^4 \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^8]$$
$$+ \frac{27}{2}s^2\sqrt{\mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^{12}] \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{x}|^4]}$$
$$+ 15s\sqrt{\mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^{10}] \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{x}|^6]}$$
$$+ \frac{27}{5}s^3\sqrt{\mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{h}|^{14}] \mathbb{E}[|\boldsymbol{a}_r^* \boldsymbol{x}|^2]}$$
$$< 20s^4 + 543s^3 + 513s^2 + 403s + 150$$
$$< 210.$$

We have all the elements to apply Lemma 7.13 with $\sigma^2 = m \max(9^2, 210) = 210m$ and $y = m/4$:

$$\mathbb{P}\left(m\mu - \sum_{r=1}^{m} Y_r \geq \frac{m}{4}\right) \leq e^{-2\gamma m}$$

with $\gamma = 1/840$. Therefore, with probability at least $1 - e^{-2\gamma m}$, we have

$$\langle \boldsymbol{Y}_r(\boldsymbol{h}, s) \rangle \geq \frac{5}{4}(1 + 3 \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2) + 6s \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h}) + 2.7s^2 - \frac{1}{4}$$
$$\geq \frac{3}{4} + \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2 + 3 \left(\operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h}) + s\right)^2$$
$$+ \left(\frac{1}{4} - \frac{3}{10}s^2\right)$$
$$\geq \frac{3}{4} + \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2. \qquad \text{(VII.21)}$$

provided that $s \leq \sqrt{5/6}$. The inequality above holds for a fixed vector $\boldsymbol{h}$. To prove (VII.17) for all $\boldsymbol{h} \in \mathbb{C}^n$ with $\|\boldsymbol{h}\| = 1$, define

$$p_r(\boldsymbol{h}) := \sqrt{\frac{5}{2}} \operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}) + \sqrt{\frac{9}{10}} s \left|\boldsymbol{a}_r^* \boldsymbol{h}\right|^2.$$

Using the fact that $\max_r \|\boldsymbol{a}_r\| \leq \sqrt{6n}$ and $s \leq 1/8$, we have $|p_r(\boldsymbol{h})| \leq 2 |\boldsymbol{a}_r^* \boldsymbol{h}| |\boldsymbol{a}_r^* \boldsymbol{x}| + s |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \leq 13n$. Moreover, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$ obeying $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$,

$$|p_r(\boldsymbol{u}) - p_r(\boldsymbol{v})| \leq \left|\sqrt{\frac{5}{2}} \operatorname{Re}\left((\boldsymbol{u} - \boldsymbol{v})^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}\right)\right|$$
$$+ \sqrt{\frac{9}{10}} s \left|\boldsymbol{a}_r^* (\boldsymbol{u} + \boldsymbol{v})\right| \left|\boldsymbol{a}_r^* (\boldsymbol{u} - \boldsymbol{v})\right|$$
$$\leq \frac{27}{2}n \|\boldsymbol{u} - \boldsymbol{v}\|.$$

Introduce

$$q(\boldsymbol{h}) := \frac{1}{m} \sum_{r=1}^{m} p_r(\boldsymbol{h})^2 - \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2$$
$$= \langle Y_r(\boldsymbol{h}, s) \rangle - \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2.$$

For any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$ obeying $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$,

$$|q(\boldsymbol{u}) - q(\boldsymbol{v})| = \left|\frac{1}{m} \sum_{r=1}^{m} (p_r(\boldsymbol{u}) - p_r(\boldsymbol{v}))(p_r(\boldsymbol{u}) + p_r(\boldsymbol{v}))\right.$$
$$\left. - \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^*(\boldsymbol{u} - \boldsymbol{v})) Re(\boldsymbol{x}^*(\boldsymbol{u} + \boldsymbol{v}))\right|$$
$$\leq \frac{27n}{2} \times 2 \times 13n \|\boldsymbol{u} - \boldsymbol{v}\| + \frac{3}{2} \|\boldsymbol{u} - \boldsymbol{v}\|$$
$$= \left(351n^2 + \frac{3}{2}\right) \|\boldsymbol{u} - \boldsymbol{v}\|. \qquad \text{(VII.22)}$$

Therefore, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^n$ obeying $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = 1$ and $\|\boldsymbol{u} - \boldsymbol{v}\| \leq \eta := \frac{1}{6000n^2}$, we have

$$q(\boldsymbol{v}) \geq q(\boldsymbol{u}) - \frac{1}{16}. \qquad \text{(VII.23)}$$

Let $\mathcal{N}_\eta$ be an η-net for the unit sphere of $\mathbb{C}^n$ with cardinality obeying $|\mathcal{N}_\eta| \leq (1 + \frac{2}{\eta})^{2n}$. Applying (VII.21) together with the union bound we conclude that for all $\boldsymbol{u} \in \mathcal{N}_\eta$

$$\mathbb{P}\left(q(\boldsymbol{u}) \geq \frac{3}{4}\right) \geq 1 - |\mathcal{N}_\eta| e^{-2\gamma m}$$
$$\geq 1 - (1 + 12000n^2)^n e^{-2\gamma m}$$
$$\geq 1 - e^{-\gamma m}. \qquad \text{(VII.24)}$$

The last line follows by choosing m such that $m \geq c \cdot n \log n$, where c is a sufficiently large constant. Now for any $\boldsymbol{h}$ on the unit sphere of $\mathbb{C}^n$, there exists a vector $\boldsymbol{u} \in \mathcal{N}_\eta$ such that $\|\boldsymbol{h} - \boldsymbol{u}\| \leq \eta$. By combining (VII.23) and (VII.24),

$$q(\boldsymbol{h}) \geq \frac{3}{4} - \frac{1}{16} > \frac{5}{8}$$

$$\Rightarrow \langle Y_r(\boldsymbol{h}, s) \rangle \geq \left(\frac{1}{8} + \frac{1}{2} \right) + \frac{3}{4} \operatorname{Re}(\boldsymbol{x}^* \boldsymbol{h})^2,$$

holds with probability at least $1 - e^{-\gamma m}$. This concludes the proof of (VII.17) with $\alpha \geq 8$.

G. Proof of the Local Smoothness Condition

For any $\boldsymbol{z} \in E(\epsilon)$, we want to prove (VII.14), which is equivalent to proving that for all $\boldsymbol{u} \in \mathbb{C}^n$ obeying $\|\boldsymbol{u}\| = 1$, we have

$$\left| (\nabla f(\boldsymbol{z}))^* \boldsymbol{u} \right|^2$$
$$\leq \beta \left(\frac{(1-\delta)}{4} \operatorname{dist}^2(\boldsymbol{z}, \boldsymbol{x}) + \frac{1}{10m} \sum_{r=1}^m \left| \boldsymbol{a}_r^*(\boldsymbol{z} - e^{i\phi(\boldsymbol{z})} \boldsymbol{x}) \right|^4 \right).$$

Recall that

$$\nabla f(\boldsymbol{z}) = \frac{1}{m} \sum_{r=1}^m \left(|\langle \boldsymbol{a}_r, \boldsymbol{z} \rangle|^2 - y_r \right) (\boldsymbol{a}_r \boldsymbol{a}_r^*) \boldsymbol{z}$$

and define

$$g(\boldsymbol{h}, \boldsymbol{w}, s) = \frac{1}{m} \sum_{r=1}^m 2 \operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}) \operatorname{Re}(\boldsymbol{w}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})$$
$$+ \frac{1}{m} \sum_{r=1}^m s |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \operatorname{Re}(\boldsymbol{w}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x})$$
$$+ \frac{1}{m} \sum_{r=1}^m 2s \operatorname{Re}(\boldsymbol{h}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{x}) \operatorname{Re}(\boldsymbol{w}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{h})$$
$$+ \frac{1}{m} \sum_{r=1}^m s^2 |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \operatorname{Re}(\boldsymbol{w}^* \boldsymbol{a}_r \boldsymbol{a}_r^* \boldsymbol{h}).$$

Define $\boldsymbol{h} := e^{-\phi(\boldsymbol{z})} \boldsymbol{z} - \boldsymbol{x}$ and $\boldsymbol{w} := e^{-i\phi(\boldsymbol{z})} \boldsymbol{u}$, to establish (VII.14) it suffices to prove that

$$|g(\boldsymbol{h}, \boldsymbol{w}, 1)|^2 \leq \beta \left(\frac{1-\delta}{4} \|\boldsymbol{h}\|^2 + \frac{1}{10m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4 \right). \quad \text{(VII.25)}$$

holds for all $\boldsymbol{h}$ and $\boldsymbol{w}$ satisfying $\operatorname{Im}(\boldsymbol{h}^* \boldsymbol{x}) = 0$, $\|\boldsymbol{h}\| \leq \epsilon$ and $\|\boldsymbol{w}\| = 1$. Equivalently, we only need to prove for all $\boldsymbol{h}$ and $\boldsymbol{w}$ satisfying $\operatorname{Im}(\boldsymbol{h}^* \boldsymbol{x}) = 0$, $\|\boldsymbol{h}\| = \|\boldsymbol{w}\| = 1$ and $\forall s : 0 \leq s \leq \epsilon$,

$$|g(\boldsymbol{h}, \boldsymbol{w}, s)|^2 \leq \beta \left(\frac{1-\delta}{4} + \frac{s^2}{10m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4 \right). \quad \text{(VII.26)}$$

Note that since $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$

$$|g(\boldsymbol{h}, \boldsymbol{w}, s)|^2 \leq \left| \frac{1}{m} \sum_{r=1}^m 2 |\boldsymbol{h}^* \boldsymbol{a}_r| |\boldsymbol{w}^* \boldsymbol{a}_r| |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \right.$$
$$+ \frac{1}{m} \sum_{r=1}^m 3s |\boldsymbol{h}^* \boldsymbol{a}_r|^2 |\boldsymbol{a}_r^* \boldsymbol{x}| |\boldsymbol{w}^* \boldsymbol{a}_r|$$
$$\left. + \frac{1}{m} \sum_{r=1}^m s^2 |\boldsymbol{a}_r^* \boldsymbol{h}|^3 |\boldsymbol{w}^* \boldsymbol{a}_r| \right|^2$$
$$\leq 3 \left| \frac{2}{m} \sum_{r=1}^m |\boldsymbol{h}^* \boldsymbol{a}_r| |\boldsymbol{w}^* \boldsymbol{a}_r| |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \right|^2$$
$$+ 3 \left| \frac{3s}{m} \sum_{r=1}^m |\boldsymbol{h}^* \boldsymbol{a}_r|^2 |\boldsymbol{a}_r^* \boldsymbol{x}| |\boldsymbol{w}^* \boldsymbol{a}_r| \right|^2$$
$$+ 3 \left| \frac{s^2}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^3 |\boldsymbol{w}^* \boldsymbol{a}_r| \right|^2$$
$$:= 3(I_1 + I_2 + I_3). \quad \text{(VII.27)}$$

We now bound each of the terms on the right-hand side. For the first term we use Cauchy-Schwarz and Corollary 7.6, which give

$$I_1 \leq \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{x}|^2 |\boldsymbol{a}_r^* \boldsymbol{w}|^2 \right) \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{x}|^2 |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \right)$$
$$\leq (2 + \delta)^2. \quad \text{(VII.28)}$$

Similarly, for the second term, we have

$$I_2 \leq \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4 \right) \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{w}|^2 |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \right)$$
$$\leq \frac{2 + \delta}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4. \quad \text{(VII.29)}$$

Finally, for the third term we use the Cauchy-Schwarz inequality together with Lemma 7.8 (inequality) (VII.9) to derive

$$I_3 \leq \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^3 \max_r \|\boldsymbol{a}_r\| \right)^2$$
$$\leq 6n \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^3 \right)^2$$
$$\leq 6n \left(\frac{1}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4 \right) \left(\sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \right)$$
$$\leq \frac{6n(1 + \delta)}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4. \quad \text{(VII.30)}$$

We now plug these inequalities into (VII.27) and get

$$|g(\boldsymbol{h}, \boldsymbol{w}, s)|^2 \leq 12(2 + \delta)^2 + \frac{27s^2(2 + \delta)}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4$$
$$+ \frac{18s^4 n(1 + \delta)}{m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4$$
$$\leq \beta \left(\frac{1-\delta}{4} + \frac{s^2}{10m} \sum_{r=1}^m |\boldsymbol{a}_r^* \boldsymbol{h}|^4 \right), \quad \text{(VII.31)}$$

which completes the proof of (VII.26) and, in turn, establishes the local smoothness condition in (VII.14). However, the last line of (VII.31) holds as long as

$$\beta \geq \max \left(48 \frac{(2+\delta)^2}{1-\delta}, 270(2+\delta) + 180\epsilon^2 n(1+\delta) \right).$$
(VII.32)

In our theorems we use two different values of $\epsilon = \frac{1}{8\sqrt{n}}$ and $\epsilon = \frac{1}{8}$. Using $\delta \leq 0.01$ in (VII.32) we conclude that the local smoothness condition (VII.31) holds as long as

$$\begin{aligned} \beta &\geq 550 && \text{for } \epsilon = 1/(8\sqrt{n}), \\ \beta &\geq 3n + 550 && \text{for } \epsilon = 1/8. \end{aligned}$$

H. Wirtinger Flow Initialization

In this section, we prove that the WF initialization obeys (III.1) from Theorem 3.3. Recall that

$$Y := \frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 a_r a_r^*.$$

and that Lemma 7.4 gives

$$\|Y - (xx^* + \|x\|^2 I)\| \leq \delta := 0.001.$$

Let $\tilde{z}_0$ be the eigenvector corresponding to the top eigenvalue λ_0 of Y obeying $\|\tilde{z}_0\| = 1$. It is easy to see that

$$\begin{aligned} \left| \lambda_0 - (|\tilde{z}_0^* x|^2 + 1) \right| &= \left| \tilde{z}_0^* Y \tilde{z}_0 - \tilde{z}_0^* (xx^* + I) \tilde{z}_0 \right| \\ &= \left| \tilde{z}_0^* \left(Y - (xx^* + I) \right) \tilde{z}_0 \right| \\ &\leq \|Y - (xx^* + I)\| \\ &\leq \delta. \end{aligned}$$

Therefore,

$$|\tilde{z}_0^* x|^2 \geq \lambda_0 - 1 - \delta.$$

Also, since λ_0 is the top eigenvalue of Y, and $\|x\| = 1$, we have

$$\lambda_0 \geq x^* Y x = x^* \left(Y - (I + xx^*) \right) x + 2 \geq 2 - \delta.$$

Combining the above two inequalities together, we have

$$\begin{aligned} |\tilde{z}_0^* x|^2 &\geq 1 - 2\delta \\ &\Rightarrow \text{dist}^2(\tilde{z}_0, x) \leq 2 - 2\sqrt{1-2\delta} < \frac{1}{256} \\ &\Rightarrow \text{dist}(\tilde{z}_0, x) \leq \frac{1}{16}. \end{aligned}$$

Recall that $z_0 = \left(\sqrt{\frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2} \right) \tilde{z}_0$. By Lemma 7.8, equation (VII.9), with high probability we have

$$\left| \|z_0\|^2 - 1 \right| = \left| \frac{1}{m} \sum_{r=1}^{m} |a_r^* x|^2 - 1 \right| \leq \frac{31}{256}$$

$$\Rightarrow |\|z_0\| - 1| \leq \frac{1}{16}.$$

Therefore, we have

$$\begin{aligned} \text{dist}(z_0, x) &\leq \|z_0 - \tilde{z}_0\| + \text{dist}(\tilde{z}_0, x), \\ &= |\|z_0\| - 1| + \text{dist}(\tilde{z}_0, x), \\ &\leq \frac{1}{8}. \end{aligned}$$

I. Initialization via Resampled Wirtinger Flow

In this section, we prove that that the output of Algorithm 2 obeys (V.1) from Theorem 5.1. Introduce

$$F(z) = \frac{1}{2} z^* (I - xx^*) z + \left(\|z\|^2 - 1 \right)^2.$$

By Lemma 7.2, if $z \in \mathbb{C}^n$ is a vector independent from the measurements, then

$$\mathbb{E} \nabla f(z; b) = \nabla F(z).$$

We prove that F obeys a regularization condition in $E(1/8)$, namely,

$$\begin{aligned} &\text{Re} \left(\langle \nabla F(z), z - x e^{i\phi(z)} \rangle \right) \\ &\qquad \geq \frac{1}{\alpha'} \text{dist}^2(z, x) + \frac{1}{\beta'} \|\nabla F(z)\|^2. \end{aligned} \quad \text{(VII.33)}$$

Lemma 7.7 implies that for a fixed vector z,

$$\begin{aligned} &\text{Re} \left(\langle \nabla f(z; b), z - x e^{i\phi(z)} \rangle \right) \\ &\quad = \text{Re} \left(\langle \nabla F(z), z - x e^{i\phi(z)} \rangle \right) \\ &\qquad + \text{Re} \left(\langle \nabla f(z; b) - \nabla F(z), z - x e^{i\phi(z)} \rangle \right) \\ &\quad \geq \text{Re} \left(\langle \nabla F(z), z - x e^{i\phi(z)} \rangle \right) \\ &\qquad - \|\nabla f(z; b) - \nabla F(z)\| \, \text{dist}(z, x) \\ &\quad \geq \text{Re} \left(\langle \nabla F(z), z - x e^{i\phi(z)} \rangle \right) - \delta \text{dist}(z, x)^2 \\ &\quad \geq \left(\frac{1}{\alpha'} - \delta \right) \text{dist}(z, x)^2 + \frac{1}{\beta'} \|\nabla F(z)\|^2, \quad \text{(VII.34)} \end{aligned}$$

holds with high probability. The last inequality follows from (VII.33). Applying Lemma 7.7, we also have

$$\begin{aligned} \|\nabla F(z)\|^2 &\geq \frac{1}{2} \|\nabla f(z; b)\|^2 - \|\nabla f(z; b) - \nabla F(z)\|^2 \\ &\geq \frac{1}{2} \|\nabla f(z; b)\|^2 - \delta^2 \text{dist}(z, x)^2. \end{aligned}$$

Plugging the latter into (VII.34) yields

$$\begin{aligned} &\text{Re} \left(\langle \nabla f(z; b), z - x e^{i\phi(z)} \rangle \right) \\ &\quad \geq \left(\frac{1}{\alpha'} - \frac{\delta^2}{\beta'} - \delta \right) \text{dist}^2(z, x) + \frac{1}{2\beta'} \|\nabla f(z; b)\| \\ &\quad := \frac{1}{\tilde{\alpha}} \text{dist}^2(z, x) + \frac{1}{\tilde{\beta}} \|\nabla f(z; b)\|. \end{aligned}$$

Therefore, using the general convergence analysis of gradient descent discussed in Section VII-D we conclude that for all $\tilde{\mu} \leq \tilde{\mu}_0 := 2/\tilde{\beta}$,

$$\text{dist}^2(u_{b+1}, x) \leq \left(1 - \frac{2\tilde{\mu}}{\tilde{\alpha}} \right) \text{dist}^2(u_b, x).$$

Finally,

$$B \geq -\frac{\log n}{\log\left(1 - \frac{2\tilde{\mu}}{\tilde{\alpha}}\right)}$$

$$\Rightarrow \operatorname{dist}(\boldsymbol{u}_B, \boldsymbol{x}) \leq \left(1 - \frac{2\tilde{\mu}}{\tilde{\alpha}}\right)^{\frac{B}{2}} \operatorname{dist}(\boldsymbol{u}_0, \boldsymbol{x})$$

$$\leq \left(1 - \frac{2\tilde{\mu}}{\tilde{\alpha}}\right)^{\frac{B}{2}} \frac{1}{8}$$

$$\leq \frac{1}{8\sqrt{n}}.$$

It only remains to establish (VII.33). First, without loss of generality, we can assume that $\phi(\boldsymbol{z}) = 0$, which implies $\operatorname{Re}(\boldsymbol{z}^*\boldsymbol{x}) = |\boldsymbol{z}^*\boldsymbol{x}|$ and use $\|\boldsymbol{z} - \boldsymbol{x}\|$ in lieu of $\operatorname{dist}(\boldsymbol{z}, \boldsymbol{x})$. Set $\boldsymbol{h} := \boldsymbol{z} - \boldsymbol{x}$ so that $\operatorname{Im}(\boldsymbol{x}^*\boldsymbol{h}) = 0$. This implies

$$\nabla F(\boldsymbol{z}) = (\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z} + 2\left(\|\boldsymbol{z}\|^2 - 1\right)\boldsymbol{z}$$

$$= (\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*)(\boldsymbol{x} + \boldsymbol{h}) + 2\left(\|\boldsymbol{x} + \boldsymbol{h}\|^2 - 1\right)(\boldsymbol{x} + \boldsymbol{h})$$

$$= (\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{h} + 2\left(2\operatorname{Re}(\boldsymbol{x}^*\boldsymbol{h}) + \|\boldsymbol{h}\|^2\right)(\boldsymbol{x} + \boldsymbol{h})$$

$$= (1 + 4(\boldsymbol{x}^*\boldsymbol{h}) + 2\|\boldsymbol{h}\|^2)\boldsymbol{h}$$
$$\quad + (3(\boldsymbol{x}^*\boldsymbol{h}) + 2\|\boldsymbol{h}\|^2)\boldsymbol{x}.$$

Therefore,

$$\|\nabla F(\boldsymbol{z})\| \leq 4\|\boldsymbol{h}\| + 6\|\boldsymbol{h}\|^2 + 2\|\boldsymbol{h}\|^3 \leq 5\|\boldsymbol{h}\|, \quad \text{(VII.35)}$$

where the last inequality is due to $\|\boldsymbol{h}\| \leq \epsilon \leq 1/8$. Furthermore,

$$\operatorname{Re}\left(\langle \nabla F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}\rangle\right)$$
$$= \operatorname{Re}\left(\left\langle (1 + 4(\boldsymbol{x}^*\boldsymbol{h}) + 2\|\boldsymbol{h}\|^2)\boldsymbol{h} + (3(\boldsymbol{x}^*\boldsymbol{h}) + 2\|\boldsymbol{h}\|^2)\boldsymbol{x}, \boldsymbol{h}\right\rangle\right)$$
$$= \|\boldsymbol{h}\|^2 + 2\|\boldsymbol{h}\|^4 + 6\|\boldsymbol{h}\|^2(\boldsymbol{x}^*\boldsymbol{h}) + 3|\boldsymbol{x}^*\boldsymbol{h}|^2$$
$$\geq \frac{1}{4}\|\boldsymbol{h}\|^2, \quad\quad\quad\quad\quad\quad\quad\quad\quad \text{(VII.36)}$$

where the last inequality also holds because $\|\boldsymbol{h}\| \leq \epsilon \leq 1/8$. Finally, (VII.35) and (VII.36) imply

$$\operatorname{Re}\left(\langle \nabla F(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{x}\rangle\right) \geq \frac{1}{4}\|\boldsymbol{h}\|^2$$

$$\geq \frac{1}{8}\|\boldsymbol{h}\|^2 + \frac{1}{200}\|\nabla F(\boldsymbol{z})\|^2$$

$$:= \frac{1}{\alpha'}\|\boldsymbol{h}\|^2 + \frac{1}{\beta'}\|\nabla F(\boldsymbol{z})\|^2,$$

where $\alpha' = 8$ and $\beta' = 200$.

APPENDIX A
EXPECTATIONS AND DEVIATIONS

We provide here the proofs of our intermediate results. Throughout this section we use $\boldsymbol{D} \in \mathbb{C}^{n \times n}$ to denote a diagonal random matrix with diagonal elements being i.i.d. samples from an admissible distribution d (recall the definition (IV.2)

of an admissible random variable). For ease of exposition, we shall rewrite (IV.1) in the form

$$y_r = \left|\sum_{t=0}^{n-1} x[t]\bar{d}_\ell(t)e^{-i2\pi kt/n}\right|^2 = \left|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}\right|^2,$$

$$r = (\ell, k), \quad \begin{array}{l} 0 \leq k \leq n - 1 \\ 1 \leq \ell \leq L, \end{array}$$

where $\boldsymbol{f}_k^*$ is the kth row of the $n \times n$ DFT matrix and $\boldsymbol{D}_\ell$ is a diagonal matrix with the diagonal entries given by $d_\ell(0), d_\ell(1), \ldots, d_\ell(n-1)$. In our model, the matricces $\boldsymbol{D}_\ell$ are i.i.d. copies of $\boldsymbol{D}$.

A. Proof of Lemma 7.1

The proof for admissible coded diffraction patterns follows from [16, Lemmas 3.1 and 3.2]. For the Gaussian model, it is a consequence of the two lemmas below, whose proofs are ommitted.

Lemma A.1: Suppose the sequence $\{\boldsymbol{a}_r\}$ follows the Gaussiam model. Then for any fixed vector $\boldsymbol{x} \in \mathbb{C}^n$,

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m |\boldsymbol{a}_r^*\boldsymbol{x}|^2 \boldsymbol{a}_r\boldsymbol{a}_r^*\right) = \boldsymbol{x}\boldsymbol{x}^* + \|\boldsymbol{x}\|^2 \boldsymbol{I}.$$

Lemma A.2: Suppose the sequence $\{\boldsymbol{a}_r\}$ follows the Gaussiam model. Then for any fixed vector $\boldsymbol{x} \in \mathbb{C}^n$,

$$\mathbb{E}\left(\frac{1}{m}\sum_{r=1}^m (\boldsymbol{a}_r^*\boldsymbol{x})^2 \bar{\boldsymbol{a}}_r\boldsymbol{a}_r^*\right) = 2\boldsymbol{x}\boldsymbol{x}^T.$$

B. Proof of Lemma 7.2

Recall that

$$\nabla f(\boldsymbol{z}) = \frac{1}{m}\sum_{r=1}^m \left(|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2 - y_r\right)(\boldsymbol{a}_r\boldsymbol{a}_r^*)\boldsymbol{z}$$

$$= \frac{1}{m}\sum_{r=1}^m \left(|\langle \boldsymbol{a}_r, \boldsymbol{z}\rangle|^2 - |\langle \boldsymbol{a}_r, \boldsymbol{x}\rangle|^2\right)(\boldsymbol{a}_r\boldsymbol{a}_r^*)\boldsymbol{z}.$$

Thus by applying [16, Lemma 3.1] (for the CDP model) and Lemma VII-A above (for the Gaussian model) we have

$$\mathbb{E}[\nabla f(\boldsymbol{z})] = \frac{1}{m}\mathbb{E}\left[\sum_{r=1}^m (|\boldsymbol{a}_r^*\boldsymbol{z}|^2 \boldsymbol{a}_r\boldsymbol{a}_r^*\boldsymbol{z} - |\boldsymbol{a}_r^*\boldsymbol{x}|^2 \boldsymbol{a}_r\boldsymbol{a}_r^*\boldsymbol{z})\right]$$

$$= (\boldsymbol{z}\boldsymbol{z}^* + \|\boldsymbol{z}\|^2 \boldsymbol{I})\boldsymbol{z} - (\boldsymbol{x}\boldsymbol{x}^* + \boldsymbol{I})\boldsymbol{z}$$

$$= 2(\|\boldsymbol{z}\|^2 - 1)\boldsymbol{z} + (\boldsymbol{I} - \boldsymbol{x}\boldsymbol{x}^*)\boldsymbol{z}.$$

C. Proof of Lemma 7.3

Suppose $\boldsymbol{a} \in \mathbb{C}^n \sim \mathcal{N}(0, \boldsymbol{I}/2) + i\mathcal{N}(0, \boldsymbol{I}/2)$. Since the law of $\boldsymbol{a}$ is invariant by unitary transformation, we may just as well take $\boldsymbol{v} = \boldsymbol{e}_1$ and $\boldsymbol{u} = s_1 e^{i\phi_1}\boldsymbol{e}_1 + s_2 e^{i\phi_2}\boldsymbol{e}_2$, where

s_1, s_2 are positive real numbers obeying $s_1^2 + s_2^2 = 1$. We have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\operatorname{Re}(\boldsymbol{u}^*\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{v})\right)^2\right] \\
&= \mathbb{E}[\left(\operatorname{Re}(s_1 e^{i\phi_1}|a_1|^2 + s_2 e^{i\phi_2}a_1\bar{a}_2)\right)^2] \\
&= 2s_1^2\cos^2(\phi_1) + \frac{1}{2}s_2^2 \\
&= \frac{1}{2} + \frac{3}{2}s_1^2\cos^2(\phi_1) - \frac{1}{2}s_1^2\sin^2(\phi_1) \\
&= \frac{1}{2} + \frac{3}{2}\left(\operatorname{Re}(\boldsymbol{u}^*\boldsymbol{v})\right)^2 - \frac{1}{2}\left(\operatorname{Im}(\boldsymbol{u}^*\boldsymbol{v})\right)^2.
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\left(\operatorname{Re}(\boldsymbol{u}^*\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{v})\right)|\boldsymbol{a}^*\boldsymbol{v}|^2\right] \\
&= \mathbb{E}[\left(\operatorname{Re}(s_1 e^{-i\phi_1}|a_1|^2 + s_2 e^{-i\phi_2}\bar{a}_1 a_2)\right)|a_1|^2] \\
&= 2s_1\cos(\phi_1) \\
&= 2\operatorname{Re}(\boldsymbol{u}^*\boldsymbol{v}).
\end{aligned}
$$

The identity (VII.7) follows from standard normal moment calculations.

D. Proof of Lemma 7.4

1) The CDP Model: Write the Hessian as

$$
\boldsymbol{Y} := \nabla^2 f(\boldsymbol{x}) = \frac{1}{nL}\sum_{\ell=1}^{L}\sum_{k=1}^{n}W_k(\boldsymbol{D}_\ell)
$$

where

$$
W_k(\boldsymbol{D}) := \begin{bmatrix} \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^* \end{bmatrix}\begin{bmatrix} A_k(\boldsymbol{D}) & B_k(\boldsymbol{D}) \\ \overline{B_k(\boldsymbol{D})} & \overline{A_k(\boldsymbol{D})} \end{bmatrix}\begin{bmatrix} \boldsymbol{D}^* & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix}
$$

and

$$
A_k(\boldsymbol{D}) = |\boldsymbol{f}_k^*\boldsymbol{D}^*\boldsymbol{x}|^2\,\boldsymbol{f}_k\boldsymbol{f}_k^*, \quad B_k(\boldsymbol{D}) = (\boldsymbol{f}_k^*\boldsymbol{D}^*\boldsymbol{x})^2\boldsymbol{f}_k\boldsymbol{f}_k^T.
$$

It is useful to recall that

$$
\mathbb{E}\,\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{I} + \boldsymbol{x}\boldsymbol{x}^* & 2\boldsymbol{x}\boldsymbol{x}^T \\ 2\bar{\boldsymbol{x}}\boldsymbol{x}^* & \boldsymbol{I} + \bar{\boldsymbol{x}}\boldsymbol{x}^T \end{bmatrix}.
$$

Now set

$$
\widetilde{\boldsymbol{Y}} = \frac{1}{nL}\sum_{\ell=1}^{L}\sum_{k=1}^{n}W_k(\boldsymbol{D}_\ell)\mathbb{1}_{\{|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}|\leq\sqrt{2R\log n}\}},
$$

where R is a positive scalar whose value will be determined shortly, and define the events

$$
\begin{aligned}
E_1(R) &= \{\|\widetilde{\boldsymbol{Y}} - \mathbb{E}\,\boldsymbol{Y}\| \leq \epsilon\}, \\
E_2(R) &= \{\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}\}, \\
E_3(R) &= \bigcap_{k,\ell}\{|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}| \leq \sqrt{2R\log n}\}, \\
E &= \{\|\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y}\| \leq \epsilon\}.
\end{aligned}
$$

Note that $E_1 \cap E_2 \subset E$. Also, if $|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}| \leq \sqrt{2R\log n}$ for all pairs (k, ℓ), then $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}$ and thus $E_3 \subset E_2$. Putting all of this together gives

$$
\begin{aligned}
\mathbb{P}(E^c) &\leq \mathbb{P}(E_1^c \cup E_2^c) \\
&\leq \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c) \\
&\leq \mathbb{P}(E_1^c) + \mathbb{P}(E_3^c) \\
&\leq \mathbb{P}(E_1^c) + \sum_{\ell=1}^{L_0}\sum_{k=1}^{n}\mathbb{P}(|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}| > \sqrt{2R\log n}). \quad \text{(A.1)}
\end{aligned}
$$

A slight modification to [16, Lemma 3.9] gives $\mathbb{P}(E_1^c) \leq 1/n^3$ provided $L \geq c(R)\log^3 n$ for a sufficiently large numerical constant $c(R)$. Since Hoeffding's inequality yields $\mathbb{P}(|\boldsymbol{f}_k^*\boldsymbol{D}_\ell^*\boldsymbol{x}| > \sqrt{2R\log n}) \leq 2n^{-R}$, we have

$$
\mathbb{P}(E^c) \leq n^{-3} + 2(nL)n^{-R}.
$$

Setting $R = 4$ completes the proof.

2) The Gaussian Model: By unitary invariance, we may take $\boldsymbol{x} = \boldsymbol{e}_1$. Letting $\boldsymbol{z}(1)$, be the first coordinate of a vector $\boldsymbol{z}$, to prove Lemma 7.4 for the Gaussian model it suffices to prove the two inequalities,

$$
\left\|\frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^2\,\boldsymbol{a}_r\boldsymbol{a}_r^* - \left(\boldsymbol{I} + \boldsymbol{e}_1\boldsymbol{e}_1^T\right)\right\| \leq \frac{\delta}{4} \quad \text{(A.2)}
$$

and

$$
\left\|\frac{1}{m}\sum_{r=1}^{m}\overline{\boldsymbol{a}_r(1)}^2\,\boldsymbol{a}_r\boldsymbol{a}_r^T - 2\boldsymbol{e}_1\boldsymbol{e}_1^T\right\| \leq \frac{\delta}{4}. \quad \text{(A.3)}
$$

For any $\epsilon > 0$, there is a constant $C > 0$ with the property that $m \geq C \cdot n$ implies

$$
\frac{1}{m}\sum_{r=1}^{m}(|\boldsymbol{a}_r(1)|^2 - 1) \leq \epsilon, \quad \frac{1}{m}\sum_{r=1}^{m}(|\boldsymbol{a}_r(1)|^4 - 2) < \epsilon,
$$

$$
\frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^6 < 10.
$$

with probability at least $1 - 3n^{-2}$; this is a consquence of Chebyshev's inequality. Moreover a union bound gives

$$
\max_{1\leq r\leq m}|\boldsymbol{a}_r(1)| \leq \sqrt{10\log m}
$$

with probability at least $1 - n^{-2}$. Denote by E_0 the event on which the above inequalities hold. We show that there is another event E_1 of high probability such that (A.2) and (A.3) hold on $E_0 \cap E_1$. Our proof strategy is similar to that of [64, Th. 39]. To prove (A.2), we will show that with high probability, for any $\boldsymbol{y} \in \mathbb{C}^n$ obeying $\|\boldsymbol{y}\| = 1$, we have

$$
\begin{aligned}
I_0(\boldsymbol{y}) &:= \left|\boldsymbol{y}^*\left(\frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^2\,\boldsymbol{a}_r\boldsymbol{a}_r^* - \left(\boldsymbol{I} + \boldsymbol{e}_1\boldsymbol{e}_1^T\right)\right)\boldsymbol{y}\right| \\
&= \left|\frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^2\,|\boldsymbol{a}_r^*\boldsymbol{y}|^2 - (1+|\boldsymbol{y}(1)|^2)\right| \leq \frac{\delta}{4}. \quad \text{(A.4)}
\end{aligned}
$$

For this purpose, partition $\boldsymbol{y}$ in the form $\boldsymbol{y} = (\boldsymbol{y}(1), \tilde{\boldsymbol{y}})$ with $\tilde{\boldsymbol{y}} \in \mathbb{C}^{n-1}$, and decompose the inner product as

$$
|\boldsymbol{a}_r^*\boldsymbol{y}|^2 = \left(|\boldsymbol{a}_r(1)|^2\,|\boldsymbol{y}(1)|^2 + 2\operatorname{Re}\left(\tilde{\boldsymbol{a}}_r^*\tilde{\boldsymbol{y}}\boldsymbol{a}_r(1)\overline{\boldsymbol{y}(1)}\right) + |\tilde{\boldsymbol{a}}_r^*\tilde{\boldsymbol{y}}|^2\right).
$$

This gives

$$
\begin{aligned}
I_0(\boldsymbol{y}) = &\left|\frac{1}{m}\sum_{r=1}^{m}(|\boldsymbol{a}_r(1)|^4 - 2)\,|\boldsymbol{y}(1)|^2\right. \\
&+ 2\operatorname{Re}\left(\frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^2\,\boldsymbol{a}_r(1)\overline{\boldsymbol{y}(1)}\,\tilde{\boldsymbol{a}}_r^*\tilde{\boldsymbol{y}}\right) \\
&\left.+ \frac{1}{m}\sum_{r=1}^{m}|\boldsymbol{a}_r(1)|^2\,|\tilde{\boldsymbol{a}}_r^*\tilde{\boldsymbol{y}}|^2 - \|\tilde{\boldsymbol{y}}\|^2\right|,
\end{aligned}
$$

which follows from $|\mathbf{y}(1)|^2 + \|\tilde{\mathbf{y}}\|^2 = 1$ since $\mathbf{y}$ has unit norm. This gives

$$
\begin{aligned}
I_0(\mathbf{y}) \\
\leq \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 - 1 \right| \|\tilde{\mathbf{y}}\|^2 + \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^4 - 2 \right| |\mathbf{y}(1)|^2 \\
+ 2 \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 \, \mathbf{a}_r(1) \overline{\mathbf{y}(1)} \, \tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}} \right| \\
+ \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 \left(|\tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}}|^2 - \|\tilde{\mathbf{y}}\|^2 \right) \right| \\
\leq 2\epsilon + 2 \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 \, \mathbf{a}_r(1) \overline{\mathbf{y}(1)} \, \tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}} \right| \\
+ \left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 \left(|\tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}}|^2 - \|\tilde{\mathbf{y}}\|^2 \right) \right|. \quad \text{(A.5)}
\end{aligned}
$$

We now turn our attention to the last two terms of (A.5). For the second term, the ordinary Hoeffding's inequality ([64], Proposition 10) gives that for any constants δ_0 and γ, there exists a constant $C(\delta_0, \gamma)$, such that for $m \geq C(\delta_0, \gamma) \sqrt{n \left(\sum_{r=1}^m |\mathbf{a}_r(1)|^6 \right)}$,

$$
\left| \frac{1}{m} \sum_{r=1}^m (|\mathbf{a}_r(1)|^2 \, \mathbf{a}_r(1) \overline{\mathbf{y}(1)}) \tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}} \right| \leq \delta_0 |\mathbf{y}(1)| \, \|\tilde{\mathbf{y}}\| \leq \delta_0
$$

holds with probability at least $1 - 3e^{-2\gamma n}$. To control the final term, we apply the Bernstein-type inequality ([64], Proposition 16) to assert the following: for any positive constants δ_0 and γ, there exists a constant $C(\delta_0, \gamma)$, such that for $m \geq C(\delta_0, \gamma)(\sqrt{n(\sum_{r=1}^m |\mathbf{a}_r(1)|^4)} + n \max_{r=1}^m |\mathbf{a}_r(1)|^2)$,

$$
\left| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r(1)|^2 \left(|\tilde{\mathbf{a}}_r^* \tilde{\mathbf{y}}|^2 - \|\tilde{\mathbf{y}}\|^2 \right) \right| \leq \delta_0 \|\tilde{\mathbf{y}}\|^2 \leq \delta_0
$$

holds with probability at least $1 - 2e^{-2\gamma n}$.

Therefore, for any unit norm vector $\mathbf{y}$,

$$
I_0(\mathbf{y}) \leq 2\epsilon + 2\delta_0 \quad \text{(A.6)}
$$

holds with probability at least $1 - 5e^{-2\gamma n}$. By [64, Lemma 5.4], we can bound the operator norm via an ϵ-net argument:

$$
\max_{\mathbf{y} \in \mathbb{C}^n} I_0(\mathbf{y}) \leq 2 \max_{\mathbf{y} \in \mathcal{N}} I_0(\mathbf{y}) \leq 4\epsilon + 4\delta_0,
$$

where $\mathcal{N}$ is an $1/4$-net of the unit sphere in $\mathbb{C}^n$. By applying the union bound and choosing appropriate δ_0, ϵ and γ, (A.2) holds with probability at least $1 - 5e^{-\gamma n}$, as long as $m \geq C'(\sqrt{n \sum_{r=1}^m |\mathbf{a}_r(1)|^6} + \sqrt{n \sum_{r=1}^m |\mathbf{a}_r(1)|^4} + n \max_{1 \leq r \leq m} |\mathbf{a}_r(1)|^2)$. On E_0 this inequality follows from $m \geq C \cdot n \log n$ provided C is sufficiently large. In conclusion, (A.2) holds with probability at least $1 - 5e^{-\gamma n} - 4n^{-2}$.

The proof of (A.3) is similar. The only difference is that the random matrix is not Hermitian, so we work with

$$
I_0(\mathbf{u}, \mathbf{v}) = \left| \mathbf{u}^* \left(\frac{1}{m} \sum_{r=1}^m \overline{\mathbf{a}_r(1)}^2 \, \mathbf{a}_r \mathbf{a}_r^T - 2 \mathbf{e}_1 \mathbf{e}_1^T \right) \mathbf{v} \right|,
$$

where $\mathbf{u}$ and $\mathbf{v}$ are unit vectors.

E. Proof of Corollary 7.5

It follows from $\|\nabla^2 f(\mathbf{x}) - \mathbb{E}[\nabla^2 f(\mathbf{x})]\| \leq \delta$ that $\nabla^2 f(\mathbf{x}) \preceq \mathbb{E}[\nabla^2 f(\mathbf{x})] + \delta \mathbf{I}$. Therefore, using the fact that for any complex scalar c, $\mathrm{Re}(c)^2 = \frac{1}{2} |c|^2 + \frac{1}{2} \mathrm{Re}(c^2)$, we have

$$
\begin{aligned}
\frac{1}{m} \sum_{r=1}^m \mathrm{Re}(\mathbf{h}^* \mathbf{a}_r \mathbf{a}_r^* \mathbf{x})^2 \\
= \frac{1}{4} \sum_{r=1}^m \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix}^* \begin{bmatrix} |\mathbf{a}_r^* \mathbf{x}|^2 \mathbf{a}_r \mathbf{a}_r^* & (\mathbf{a}_r^* \mathbf{x})^2 \mathbf{a}_r \mathbf{a}_r^T \\ (\overline{\mathbf{a}_r^* \mathbf{x}})^2 \bar{\mathbf{a}}_r \mathbf{a}_r^* & |\mathbf{a}_r^* \mathbf{x}|^2 \bar{\mathbf{a}}_r \mathbf{a}_r^T \end{bmatrix} \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix} \\
\preceq \frac{1}{4} \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix}^* \left(\mathbf{I}_{2n} + \frac{3}{2} \begin{bmatrix} \mathbf{x} \\ \bar{\mathbf{x}} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \bar{\mathbf{x}} \end{bmatrix}^* - \frac{1}{2} \begin{bmatrix} \mathbf{x} \\ -\bar{\mathbf{x}} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ -\bar{\mathbf{x}} \end{bmatrix}^* \right) \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix} \\
+ \frac{\delta}{4} \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix}^* \begin{bmatrix} \mathbf{h} \\ \bar{\mathbf{h}} \end{bmatrix} \\
\preceq \left(\frac{1}{2} \|\mathbf{h}\|^2 + \frac{3}{2} \mathrm{Re}(\mathbf{x}^* \mathbf{h})^2 - \frac{1}{2} Im(\mathbf{x}^* \mathbf{h})^2 \right) + \frac{\delta}{2}.
\end{aligned}
$$

The other inequality is established in a similar fashion.

F. Proof of Corollary 7.6

In the proof of Lemma 7.4, we established that with high probability,

$$
\left\| \frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r^* \mathbf{x}|^2 \, \mathbf{a}_r \mathbf{a}_r^* - \left(\mathbf{x} \mathbf{x}^* + \|\mathbf{x}\|^2 \mathbf{I} \right) \right\| \leq \delta.
$$

Therefore,

$$
\frac{1}{m} \sum_{r=1}^m |\mathbf{a}_r^* \mathbf{x}|^2 \, \mathbf{a}_r \mathbf{a}_r^* \succeq \left(\mathbf{x} \mathbf{x}^* + \|\mathbf{x}\|^2 \mathbf{I} \right) - \delta \mathbf{I}.
$$

This concludes the proof of one side. The other side is similar.

G. Proof of Lemma 7.7

Note that

$$
\|\nabla f(\mathbf{z}) - \mathbb{E} \nabla f(\mathbf{z})\| = \max_{\mathbf{u} \in \mathbb{C}^n, \|\mathbf{u}\|=1} \langle \mathbf{u}, \nabla f(\mathbf{z}) - \mathbb{E} \nabla f(\mathbf{z}) \rangle
$$

Therefore, to establish the concentration of $\nabla f(\mathbf{z})$ around its mean we proceed by bounding $|\langle \mathbf{u}, \nabla f(\mathbf{z}) - \mathbb{E} \nabla f(\mathbf{z}) \rangle|$. From Section VII-B,

$$
\nabla f(\mathbf{z}) = \frac{1}{m} \sum_{r=1}^m \left(|\langle \mathbf{a}_r, \mathbf{z} \rangle|^2 - y_r \right) (\mathbf{a}_r \mathbf{a}_r^*) \mathbf{z}.
$$

Define $\mathbf{h} := e^{-\phi(\mathbf{z})} \mathbf{z} - \mathbf{x}$ and $\mathbf{w} := e^{-i\phi(\mathbf{z})} \mathbf{u}$, we have

$$
\begin{aligned}
\langle \mathbf{u}, \nabla f(\mathbf{z}) \rangle \\
= \frac{1}{m} \sum_{r=1}^m \mathbf{w}^* \left((\mathbf{a}_r^* \mathbf{x})^2 \mathbf{a}_r \mathbf{a}_r^T \right) \bar{\mathbf{h}} + \mathbf{w}^* \left(|\mathbf{a}_r^* \mathbf{x}|^2 \mathbf{a}_r \mathbf{a}_r^* \right) \mathbf{h} \\
+ 2 \mathbf{w}^* \left(|\mathbf{a}_r^* \mathbf{h}|^2 \mathbf{a}_r \mathbf{a}_r^* \mathbf{h} \right) \mathbf{x} + \mathbf{w}^* \left((\mathbf{a}_r^* \mathbf{h})^2 \mathbf{a}_r \mathbf{a}_r^T \right) \bar{\mathbf{x}} \\
+ \mathbf{w}^* \left(|\mathbf{a}_r^* \mathbf{h}|^2 \mathbf{a}_r \mathbf{a}_r^* \right) \mathbf{h}. \quad \text{(A.7)}
\end{aligned}
$$

By Lemma 7.2 we also have

$$
\begin{aligned}
\mathbf{w}^* \mathbb{E}[\nabla f(\mathbf{z})] = \mathbf{w}^* \left(2\mathbf{x} \mathbf{x}^T \right) \bar{\mathbf{h}} + \mathbf{w}^* \left(\mathbf{x} \mathbf{x}^* + \|\mathbf{x}\|^2 \mathbf{I} \right) \mathbf{h} \\
+ 2\mathbf{w}^* \left(\mathbf{h} \mathbf{h}^* + \|\mathbf{h}\|^2 \mathbf{I} \right) \mathbf{x} + \mathbf{w}^* \left(2\mathbf{h} \bar{\mathbf{h}}^T \right) \bar{\mathbf{x}} \\
+ \mathbf{w}^* \left(\mathbf{h} \mathbf{h}^* + \|\mathbf{h}\|^2 \mathbf{I} \right) \mathbf{h}. \quad \text{(A.8)}
\end{aligned}
$$

Combining (A.7) and (A.8) together with the triangular inequality and Lemma 7.4 give

$$\left| \langle \boldsymbol{u}, \nabla f(\boldsymbol{z}) - \mathbb{E}[\nabla f(\boldsymbol{z})] \rangle \right|$$

$$\leq \left| \boldsymbol{w}^* \left(\frac{1}{m} \sum_{r=1}^{m} (\boldsymbol{a}_r^* \boldsymbol{x})^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^T - 2\boldsymbol{x}\boldsymbol{x}^T \right) \bar{\boldsymbol{h}} \right|$$

$$+ \left| \boldsymbol{w}^* \left(\frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - (\boldsymbol{x}\boldsymbol{x}^* + \|\boldsymbol{x}\|^2 \, \boldsymbol{I}) \right) \boldsymbol{h} \right|$$

$$+ 2 \left| \boldsymbol{w}^* \left(\frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - \left(\boldsymbol{h}\boldsymbol{h}^* + \|\boldsymbol{h}\|^2 \, \boldsymbol{I} \right) \right) \boldsymbol{x} \right|$$

$$+ \left| \boldsymbol{w}^* \left(\frac{1}{m} \sum_{r=1}^{m} (\boldsymbol{a}_r^* \boldsymbol{h})^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^T - 2\boldsymbol{h}\bar{\boldsymbol{h}}^T \right) \bar{\boldsymbol{x}} \right|$$

$$+ \left| \boldsymbol{w}^* \left(\frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - \left(\boldsymbol{h}\boldsymbol{h}^* + \|\boldsymbol{h}\|^2 \, \boldsymbol{I} \right) \right) \boldsymbol{h} \right|.$$

$$\leq \left\| \frac{1}{m} \sum_{r=1}^{m} (\boldsymbol{a}_r^* \boldsymbol{x})^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^T - 2\boldsymbol{x}\boldsymbol{x}^T \right\| \|\boldsymbol{h}\|$$

$$+ \left\| \frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{x}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - (\boldsymbol{x}\boldsymbol{x}^* + \|\boldsymbol{x}\|^2 \, \boldsymbol{I}) \right\| \|\boldsymbol{h}\|$$

$$+ 2 \left\| \frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - (\boldsymbol{h}\boldsymbol{h}^* + \|\boldsymbol{h}\|^2 \, \boldsymbol{I}) \right\|$$

$$+ \left\| \frac{1}{m} \sum_{r=1}^{m} (\boldsymbol{a}_r^* \boldsymbol{h})^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^T - 2\boldsymbol{h}\bar{\boldsymbol{h}}^T \right\|$$

$$+ \left\| \frac{1}{m} \sum_{r=1}^{m} |\boldsymbol{a}_r^* \boldsymbol{h}|^2 \, \boldsymbol{a}_r \boldsymbol{a}_r^* - \left(\boldsymbol{h}\boldsymbol{h}^* + \|\boldsymbol{h}\|^2 \, \boldsymbol{I} \right) \right\| \|\boldsymbol{h}\|$$

$$\leq 3\delta \|\boldsymbol{h}\| (1 + \|\boldsymbol{h}\|)$$

$$\leq \frac{9}{2} \delta \|\boldsymbol{h}\|.$$

H. Proof of Lemma 7.8

The result for the CDP model follows from [16, Lemma 3.3]. For the Gaussian model, it is a consequence of standard results, see [64, Th. 5.39], concerning the deviation of the sample covariance matrix from its mean.

APPENDIX B
THE POWER METHOD

We use the power method (Algorithm 3) with a random initialization to compute the first eigenvector of $\boldsymbol{Y} = \boldsymbol{A} \operatorname{diag}\{\boldsymbol{y}\} \boldsymbol{A}^*$. Since, each iteration of the power method asks to compute the matrix-vector product

$$\boldsymbol{Y}\boldsymbol{z} = \boldsymbol{A} \operatorname{diag}\{\boldsymbol{y}\} \boldsymbol{A}^* \boldsymbol{z},$$

we simply need to apply $\boldsymbol{A}$ and $\boldsymbol{A}^*$ to an arbitrary vector. In the Gaussian model, this costs $2mn$ multiplications while in the CDP model the cost is that of $2L$ n-point FFTs. We now turn our attention to the number of iterations required to achieve a sufficiently accurate initialization.

Standard results from numerical linear algebra show that after k iterations of the power method, the accuracy of the eigenvector is $\mathcal{O}(\tan \theta_0 (\lambda_2/\lambda_1)^k)$, where λ_1 and λ_2 are the top two eigenvalues of the positive semidefinite matrix $\boldsymbol{Y}$, and

---

Algorithm 3 Power Method
Input: Matrix $\boldsymbol{Y}$
  $\boldsymbol{v}_0$ is a random vector on the unit sphere of $\mathbb{C}^n$
  for $\tau = 1$ **to** T **do**
    $\boldsymbol{v}_\tau = \frac{\boldsymbol{Y}\boldsymbol{v}_{\tau-1}}{\|\boldsymbol{Y}\boldsymbol{v}_{\tau-1}\|}$
  end for
Output: $\tilde{\boldsymbol{z}}_0 = \boldsymbol{v}_T$

θ_0 is the angle between the initial guess and the top eigenvector. Hence, we would need on the order of $\log(n/\epsilon) / \log(\lambda_1/\lambda_2)$ for ϵ accuracy. Under the stated assumptions, Lemma 7.4 bounds below the eigenvalue gap by a numerical constant so that we can see that few iterations of the power method would yield accurate estimates.

REFERENCES

[1] *EE 378B—Inference, Estimation, and Information Processing*. [Online]. Available: http://web.stanford.edu/class/ee378b/

[2] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Ann. Statist.*, vol. 40, no. 5, pp. 2452–2482, 2012.

[3] A. Ahmed, B. Recht, and J. Romberg. (2012). "Blind deconvolution using convex programming." [Online]. Available: http://arxiv.org/abs/1211.5608

[4] R. Balan, "On signal reconstruction from its spectrogram," in *Proc. 44th Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2010, pp. 1–4.

[5] R. Balan, P. Casazza, and D. Edidin, "On signal reconstruction without phase," *Appl. Comput. Harmonic Anal.*, vol. 20, no. 3, pp. 345–356, 2006.

[6] L. Balzano and S. J. Wright. (2013). "Local convergence of an algorithm for subspace identification from partial data." [Online]. Available: http://arxiv.org/abs/1306.3391

[7] A. S. Bandeira, J. Cahill, D. G. Mixon, and A. A. Nelson. (2013). "Saving phase: Injectivity and stability for phase retrieval." [Online]. Available: http://arxiv.org/abs/1302.4618

[8] H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Hybrid projection–reflection method for phase retrieval," *J. Opt. Soc. Amer. A*, vol. 20, no. 6, pp. 1025–1034, 2003.

[9] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[10] M. Bayati and A. Montanari, "The LASSO risk for Gaussian matrices," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 1997–2017, Apr. 2012.

[11] V. Bentkus, "An inequality for tail probabilities of martingales with differences bounded from one side," *J. Theoretical Probab.*, vol. 16, no. 1, pp. 161–173, 2003.

[12] O. Bunk *et al.*, "Diffractive imaging for periodic samples: Retrieving one-dimensional concentration profiles across microfluidic channels," *Acta Crystallograph. A, Found. Crystallogr.*, vol. 63, pp. 306–314, Jul. 2007.

[13] J. Cahill, P. G. Casazza, J. Peterson, and L. Woodland. (2013). "Phase retrieval by projections." [Online]. Available: http://arxiv.org/abs/1305.6226

[14] E. J. Candès, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM J. Imag. Sci.*, vol. 6, no. 1, pp. 199–225, 2013.

[15] E. J. Candès and X. Li, "Solving quadratic equations via PhaseLift when there are about as many equations as unknowns," *Found. Comput. Math.*, vol. 66, pp. 1241–1274.

[16] E. J. Candès, X. Li, and M. Soltanolkotabi. (2013). "Phase retrieval from coded diffraction patterns." [Online]. Available: http://arxiv.org/1310.3240

[17] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Commun. Pure Appl. Math.*, vol. 38, no. 2, pp. 346–356, Mar. 2015.

[18] A. Chai, M. Moscoso, and G. Papanicolaou, "Array imaging using intensity-only measurements," *Inverse Problems*, vol. 27, no. 1, p. 015005, 2011.

[19] A. Conca, D. Edidin, M. Hering, and C. Vinzant. (2013). "An algebraic characterization of injectivity in phase retrieval." [Online]. Available: http://arxiv.org/abs/1312.0158

[20] J. V. Corbett, "The Pauli problem, state reconstruction and quantum-real numbers," *Rep. Math. Phys.*, vol. 57, no. 1, pp. 53–68, 2006.

[21] J. C. Dainty and J. R. Fienup, "Phase retrieval and image reconstruction for astronomy," in *Image Recovery: Theory and Application*, H. Stark, Ed. San Diego, CA, USA: Academic, 1987, pp. 231–275.

[22] L. Demanet and V. Jugnon. (2013). "Convex recovery from interferometric measurements." [Online]. Available: http://arxiv.org/abs/1307.6864

[23] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, 2009.

[24] J. R. Fienup, "Reconstruction of an object from the modulus of its Fourier transform," *Opt. Lett.*, vol. 3, no. 1, pp. 27–29, 1978.

[25] J. R. Fienup, "Fine resolution imaging of space objects," Radar Opt. Division, Environ. Res. Inst. Michigan, Ann Arbor, MI, USA, Final Sci. Rep. 01/1982-1, 1982.

[26] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, 1982.

[27] J. R. Fienup, "Comments on 'The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform,'" *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 31, no. 3, pp. 738–739, Jun. 1983.

[28] F. Fogel, I. Waldspurger, and A. d'Aspremont. (2013). "Phase retrieval for imaging problems." [Online]. Available: http://arxiv.org/abs/1304.7735

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

[30] D. Gross, F. Krahmer, and R. Kueng. (2013). "A partial derandomization of PhaseLift using spherical designs." [Online]. Available: http://arxiv.org/abs/1310.2267

[31] D. Gross, F. Krahmer, and R. Kueng. (2014). "Improved recovery guarantees for phase retrieval from coded diffraction patterns." [Online]. Available: http://arxiv.org/abs/1402.6286

[32] M. Hardt. (2013). "On the provable convergence of alternating minimization for matrix completion." [Online]. Available: http://arxivweb3.library.cornell.edu/abs/1312.0925v1

[33] R. W. Harrison, "Phase problem in crystallography," *J. Opt. Soc. Amer. A*, vol. 10, no. 5, pp. 1046–1055, 1993.

[34] M. H. Hayes, "The reconstruction of a multidimensional sequence from the phase or magnitude of its Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 30, no. 2, pp. 140–154, Apr. 1982.

[35] T. Heinosaari, L. Mazzarella, and M. M. Wolf, "Quantum tomography under prior information," *Commun. Math. Phys.*, vol. 318, no. 2, pp. 355–374, 2013.

[36] K. Jaganathan, S. Oymak, and B. Hassibi, "On robust phase retrieval for sparse signals," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2012, pp. 794–799.

[37] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. 45th Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[38] R. H. Keshavan, "Efficient algorithms for collaborative filtering," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2012.

[39] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[40] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, Mar. 2010.

[41] K. Lee, Y. Wu, and Y. Bresler. (2013). "Near optimal compressed sensing of sparse rank-one matrices via sparse power factorization." [Online]. Available: http://arxiv.org/abs/1312.0525

[42] S. Marchesini, "Invited article: A unified evaluation of iterative projection algorithms for phase retrieval," *Rev. Sci. Instrum.*, vol. 78, no. 1, p. 011301, 2007.

[43] S. Marchesini, "Phase retrieval and saddle-point optimization," *J. Opt. Soc. Amer. A*, vol. 24, no. 10, pp. 3289–3296, 2007.

[44] S. Marchesini, Y.-C. Tu, and H.-T. Wu. (2014). "Alternating projection, ptychographic imaging and phase synchronization." [Online]. Available: http://arxiv.org/abs/1402.0550

[45] 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*, vol. 400, no. 6742, pp. 342–344, 1999.

[46] J. Miao, T. Ishikawa, B. Johnson, E. H. Anderson, B. Lai, and K. O. Hodgson, "High resolution 3D X-ray diffraction microscopy," *Phys. Rev. Lett.*, vol. 89, no. 8, p. 088303, Aug. 2002.

[47] J. Miao, T. Ishikawa, Q. Shen, and T. Earnest, "Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes," *Annu. Rev. Phys. Chem.*, vol. 59, pp. 387–410, May 2008.

[48] L. Waller, L. Tian, and G. Barbastathis, "Transport of intensity phaseamplitude imaging with higher order intensity derivatives," *Opt. Exp.*, vol. 18, no. 12, pp. 12552–12561, 2010.

[49] L. Tian, X. Li, K. Ramchandran, and L. Waller, "Multiplexed coded illumination for Fourier ptychography with an LED array microscope," *Biomed. Opt. Exp.*, vol. 5, pp. 2376–2389, 2014.

[50] R. P. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Amer. A*, vol. 7, no. 3, pp. 394–411, 1990.

[51] Y. Mroueh. (2014). "Robust phase retrieval and super-resolution from one bit coded diffraction patterns." [Online]. Available: http://arxiv.org/abs/1402.2255

[52] Y. Mroueh and L. Rosasco. (2013). "Quantization and greed are good: One bit phase retrieval, robustness and greedy refinements." [Online]. Available: http://arxiv.org/abs/1312.1830

[53] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Program.*, vol. 39, no. 2, pp. 117–129, 1987.

[54] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. Philadelphia, PA, USA: SIAM, 2001.

[55] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course* (Applied Optimization), vol. 87. Boston, MA, USA: Kluwer, 2004,

[56] P. Netrapalli, P. Jain, and S. Sanghavi. (2013). "Phase retrieval using alternating minimization." [Online]. Available: http://arxiv.org/abs/1306.0160

[57] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry. (2011). "Compressive phase retrieval from squared output measurements via semidefinite programming." [Online]. Available: http://arxiv.org/abs/1111.6323

[58] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. (2012). "Simultaneously structured models with application to sparse and low-rank matrices." [Online]. Available: http://arxiv.org/abs/1212.3753

[59] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul./Aug. 2011, pp. 2168–2172.

[60] J. Ranieri, A. Chebira, Y. M. Lu, and M. Vetterli. (2013). "Phase retrieval for sparse signals: Uniqueness conditions." [Online]. Available: http://arxiv.org/abs/1308.3058

[61] H. Reichenbach, *Philosophic Foundations of Quantum Mechanics*. Berkeley, CA, USA: Univ. California Press, 1965.

[62] J. L. C. Sanz, T. S. Huang, and T. Wu, "A note on iterative Fourier transform phase reconstruction from magnitude," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1251–1254, Dec. 1984.

[63] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2012, pp. 815–822.

[64] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. (2014). "Phase retrieval with application to optical imaging." [Online]. Available: http://arxiv.org/abs/1402.7350

[65] M. Soltanolkotabi, "Algorithms and theory for clustering and non-convex quadratic programming," Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, USA, 2014.

[66] R. Vershynin, "Compressed sensing: Theory and applications," in *Introduction to the Non-Asymptotic Analysis of Random Matrices*, Y. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[67] I. Waldspurger, A. d'Aspremont, and S. Mallat. (2012). "Phase recovery, maxcut and complex semidefinite programming." [Online]. Available: http://arxiv.org/abs/1206.0102

[68] A. Walther, "The question of phase retrieval in optics," *Opt. Acta, Int. J. Opt.*, vol. 10, no. 1, pp. 41–49, 1963.

[69] G.-Z. Yang, B.-Z. Dong, B.-Y. Gu, J.-Y. Zhuang, and O. K. Ersoy, "Gerchberg–Saxton and Yang–Gu algorithms for phase retrieval in a nonunitary transform system: A comparison," *Appl. Opt.*, vol. 33, no. 2, pp. 209–218, 1994.

Emmanuel J. Candès is the Barnum-Simons Chair in Mathematics and Statistics, and professor of electrical engineering (by courtesy) at Stanford University. Up until 2009, he was the Ronald and Maxine Linde Professor of Applied and Computational Mathematics at the California Institute of Technology. His research interests are in applied mathematics, statistics, information theory, signal processing and mathematical optimization with applications to the imaging sciences, scientific computing and inverse problems. Cands graduated from the Ecole Polytechnique in 1993 with a degree in science and engineering, and received his Ph.D. in statistics from Stanford University in 1998. Emmanuel received the 2006 Alan T. Waterman Award from NSF, which recognizes the achievements of early-career scientists. Other honors include the 2013 Dannie Heineman Prize presented by the Academy of Sciences at Gttingen, the 2010 George Polya Prize awarded by the Society of Industrial and Applied Mathematics (SIAM), and the 2015 AMS-SIAM George David Birkhoff Prize in Applied Mathematics. He is a member of the National Academy of Sciences and the American Academy of Arts and Sciences.

Xiaodong Li is a postdoctoral research associate in the Wharton Statistics Department, University of Pennsylvania. He received his B.S. in mathematics from Peking University, Beijing, China in 2008. He obtained his Ph.D in mathematics from Stanford University in 2013. His research interests are in statistics, mathematical signal processing, machine learning and optimization.

Mahdi Soltanolkotabi obtained his B.S. in electrical engineering at Sharif University of Technology, Tehran, Iran in 2009. He completed his M.S. and Ph.D. in electrical engineering at Stanford University in 2011 and 2014. He was a postdoctoral researcher in the Electrical Engineering and Computer Science and Statistics departments at the University of California, Berkeley during the 2014-2015 academic year. He is currently an assistant professor in the Ming Hsieh Department of Electrical Engineering at the University of Southern California. His research interests include mathematical optimization, machine learning, signal processing, high-dimensional statistics, and geometry with emphasis on applications in information and physical sciences.