Towards Sample-Optimal Compressive Phase Retrieval with Sparse and Generative Priors

Anonymous Author(s)
Affiliation
Address
email

Abstract

Compressive phase retrieval is a popular variant of the standard compressive sensing problem, in which the measurements only contain magnitude information. In this paper, motivated by recent advances in deep generative models, we provide recovery guarantees with order-optimal sample complexity bounds for phase retrieval with generative priors. We first show that when using i.i.d. Gaussian measurements and an L-Lipschitz continuous generative model with bounded k-dimensional inputs, roughly $O(k \log L)$ samples suffice to guarantee that the signal is close to any vector that minimizes an amplitude-based empirical loss function. Attaining this sample complexity with a practical algorithm remains a difficult challenge, and a popular spectral initialization method has been observed to pose a major bottleneck. To partially address this, we further show that roughly $O(k \log L)$ samples ensure sufficient closeness between the signal and any *globally optimal* solution to an optimization problem designed for spectral initialization (though finding such a solution may still be challenging). We adapt this result to sparse phase retrieval, and show that $O(s \log n)$ samples are sufficient for a similar guarantee when the underlying signal is s-sparse and n-dimensional, matching an information-theoretic lower bound. While our guarantees do not directly correspond to a practical algorithm, we propose a practical spectral initialization method motivated by our findings, and experimentally observe significant performance gains over various existing spectral initialization methods of sparse phase retrieval.

1 Introduction

In this paper, we consider the (real-valued) phase retrieval problem, which aims to recover a signal $\mathbf{x} \in \mathbb{R}^n$ from noisy magnitude-only measurements:

$$y_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle| + \eta_i, \quad i = 1, 2, \ldots, m, \tag{1}$$

where $\mathbf{a}_i \in \mathbb{R}^n$ is the i-th sensing vector, and η_i is the additive noise term. This problem arises naturally in areas such as diffraction imaging, X-ray crystallography, microscopy, optics and astronomy [7], where it is often difficult or even impossible to observe the linear measurements directly, and one can only record the magnitudes or intensities (squared magnitudes).

In many real applications, to reduce the required number of measurements, it is of interest to exploit structure in the signal being estimated. In particular, for applications related to signal processing and imaging, it is well-known that the underlying signal typically admits a sparse or approximately sparse representation in some known basis [37]. Motivated by this, and considering the popularity of the standard compressive sensing (CS) problem [15], the sparse phase retrieval problem has attracted significant research interest.

Moreover, inspired by the successful applications of deep generative models in many fields [14], recently, a new perspective of CS has emerged, for which the sparsity assumption is replaced by

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

a generative model assumption. That is, rather than being sparse, the signal is instead assumed to be close to the range of a generative model [4]. In addition to the theoretical developments in [4], the authors also provide impressive numerical results showing that for some imaging applications, using generative priors can significantly reduce the required number of measurements (e.g., by a factor of 5 to 10) for recovering the signal up to a given accuracy. Follow-up works of [4] include [52, 21, 42, 61, 11, 28], just to name a few.

In this paper, we focus on providing recovery guarantees with order-optimal sample complexity bounds for the compressive phase retrieval problem, for which the underlying signal is assumed to be sparse or lie in (or close to) the range of a generative model.

1.1 Related Work

In this subsection, we provide a summary of some relevant works, which can roughly be divided into (i) the general phase retrieval problem with no structural assumptions on the signal, (ii) sparse phase retrieval, and (iii) phase retrieval with generative models.

General phase retrieval: A wide range of approaches have been designed to solve the phase retrieval problem. Error-reduction approaches [17, 13] work well in practice, but they lack provable guarantees. Convex methods such as PhaseLift [8] and PhaseCut [56] lift the phase retrieval problem to a higher dimension and suffer from high computation cost. Convex relaxations that operate in the natural domain of the signal are proposed in [18, 1]. However, these convex relaxation based methods are not empirically competitive against widely-used non-convex methods.

Following the seminal work of Netrapalli *et al.* [39], several works have studied theoretically-guaranteed non-convex optimization algorithms for phase retrieval [7, 9, 57, 66]. These algorithms start from a spectral initialization method, and then use an iterative algorithm (e.g., alternating minimization or gradient descent) to further decrease the approximation error. For the general phase retrieval problem, the optimal sample complexity $O(n)$ can be achieved for some non-convex methods [9, 57, 66]. Interestingly, optimal (up to a logarithmic factor) sample complexity guarantees have also been provided for the case of using a random initialization [49, 10]. It is worth mentioning that some variants of the classic Wirtinger Flow (WF) algorithm [7, 9] and trust-region methods [49] consider squared measurements:

$$y_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|^2 + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, m, \tag{2}$$

and minimize the intensity based empirical loss function:

$$f_I(\mathbf{w}) := \frac{1}{2} \left\| \mathbf{y} - |\mathbf{A}\mathbf{w}|^2 \right\|_2^2, \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the sensing matrix with its i-th row being $\mathbf{a}_i^T$, and $|\cdot|$ is applied element-wise. Nonetheless, numerical results suggest that algorithms minimizing the amplitude based empirical risk (with the measurement model (1))

$$f(\mathbf{w}) := \| \mathbf{y} - |\mathbf{A}\mathbf{w}| \|_2^2 \tag{4}$$

are usually more efficient in computation [57, 66, 48]. Based on this, we focus on the measurement model (1) and the associated loss function (4).

Sparse phase retrieval: Various algorithms have also been devised for sparse phase retrieval [41, 47, 45]. In particular, there exist various non-convex optimization based algorithms, including thresholded/projected WF [6, 48], sparse truncated amplitude flow [58], compressive phase retrieval with alternating minimization [27], and sparse phase retrieval by hard iterative pursuit [5]. All of these approaches are analyzed under the assumption of using a sensing matrix with i.i.d. real Gaussian entries. In addition, similar to non-convex methods for the general phase retrieval problem, all of these approaches start with a spectral initialization step, and then use an iterative algorithm to refine the initial vector. More specifically, in the initialization step, these approaches try to accurately estimate the support of the underlying signal.

Unlike the general phase retrieval setting in which optimal sample complexity guarantees are provided for some practical algorithms, for non-convex methods of sparse phase retrieval, the typical sample complexity requirement is $O(s^2 \log n)$, where s is the sparsity level. This does not match the information theoretic lower bound [44, 51]; the bottleneck comes from the accurate estimation of

the support set [58, 48, 27]. Improved guarantees are also available in the noiseless setting under more restrictive assumptions on the magnitudes of the non-zero entries [62]. In addition, for sparse phase retrieval, recovery guarantees with respect to *globally optimal solutions* of both the intensity and amplitude based loss functions have been provided in [12, 25, 23, 22, 63], achieving the optimal sample complexity $O(s \log n)$. However, designing an *efficient algorithm achieving the optimal sample complexity* for s-sparse vectors remains an important open problem.

Phase retrieval with generative models: Phase retrieval using generative models has been studied in [19, 24, 26, 46, 60]. In particular, the authors of [19] model natural signals as being in the range of a fully-connected ReLU neural network with random weights and no offsets, and study the global geometry for gradient methods. Extensive numerical experiments for phase retrieval with generative models, as well as both Gaussian and coded diffraction pattern measurements, have been presented in [46]. Algorithms proposed in both works [19, 46] minimize the objective function directly over the latent variable $\mathbf{z} \in \mathbb{R}^k$ using gradient descent, where k is the latent dimension. The corresponding objective function is highly non-convex, and performing gradient descent directly over $\mathbf{z}$ limits the explorable solution space, which may lead to getting stuck in local minima.

Recovery guarantees for non-convex optimization algorithms for phase retrieval with pre-trained ReLU neural network or un-trained neural network generative models have been provided in [24, 26], with assumptions of noiseless measurements and the existence of a good initial vector. Recovery guarantees for phase retrieval with generative models have been provided in [60], within a more general setting of using differentiable but unknown link functions. The weaker assumption of an unknown link function comes as a price. In particular, for an unknown nonlinearity, it is an open problem to handle signals with representation error [43]. Moreover, the assumption of differentiability makes it is not directly applicable to the measurement model (1), and the loss function considered in [60] is different from both (3) and (4), which are widely adopted for phase retrieval.

1.2 Contributions

- We show that for compressive phase retrieval using i.i.d. Gaussian measurements, when the signal is close to the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs, roughly $O(k \log L)$ samples suffice for ensuring the closeness between the signal and any vector minimizing the amplitude-based empirical loss function (4).
- To address the sample complexity barrier posed by spectral initialization, we propose a relevant optimization problem, and show that roughly $O(k \log L)$ samples are sufficient to guarantee that any *globally optimal* solution of the optimization problem is close to the true signal. This suggests the plausibility of practical spectral initialization algorithms that are able to find such a global optimum, though we do not attempt to provide one that can provably do so.
- We adapt our analysis to the case of sparse phase retrieval, and show that roughly $O(s \log n)$ samples suffice for a similar recovery guarantee, where s is the sparsity level and n is the ambient dimension. This matches the information-theoretic lower bound.
- Motivated by our theoretical findings for sparse phase retrieval, we propose a spectral initialization method based on sparse principal component analysis. We verify on synthetic experiments that our initialization method significantly outperforms several popular spectral initialization methods of sparse phase retrieval (in the sense of relative error). This further corroborates the optimal sample complexity guarantee in our theory.

1.3 Notation

We use upper and lower case boldface letters to denote matrices and vectors respectively. We write $[N] = \{1, 2, \cdots, N\}$ for a positive integer N. We define the ℓ_2-ball $B_2^k(r) := \{\mathbf{z} \in \mathbb{R}^k : \|\mathbf{z}\|_2 \leq r\}$, and the unit sphere $\mathcal{S}^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = 1\}$. We use G to denote an L-Lipschitz continuous generative model from $B_2^k(r)$ to $\mathbb{R}^n$. For a set $B \subseteq B_2^k(r)$, we write $G(B) = \{G(\mathbf{z}) : \mathbf{z} \in B\}$. The sensing matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is assumed to have i.i.d. standard normal entries, i.e., $a_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. For any $\mathbf{s} \in \mathbb{R}^n$, we use $\bar{\mathbf{s}} = \frac{\mathbf{s}}{\|\mathbf{s}\|_2}$ to denote the corresponding normalized vector. The support (set) of a vector is the index set of its non-zero entries. For any $\mathbf{X} \in \mathbb{R}^{m \times n}$ and any index set $I \subseteq [m]$, $\mathbf{X}_I$ denotes the $|I| \times n$ sub-matrix of $\mathbf{X}$ that only keeps the rows indexed by I. We use $\|\mathbf{X}\|_{2 \to 2}$ to denote the spectral norm of a matrix $\mathbf{X}$. For any positive integer N, we use $\mathbf{I}_N$ to denote the identity

matrix in $\mathbb{R}^{N\times N}$. We use the generic notations C and C' to denote large positive constants, and we use c and c' to denote small positive constants. Their values may differ from line to line.

2 Optimal Solutions of Amplitude-Based Loss Function for Generative Priors

In this section, we provide recovery guarantees with respect to optimal solutions of the amplitude based loss function (4) for phase retrieval with generative models. We let the generative model $G : B_2^k(r) \to \mathbb{R}^n$ be any L-Lipschitz continuous function, and we suppose that the underlying signal $\mathbf{x}$ is close to (but does not need to lie in) the range of G. We are interested in the case $k \ll n$ (i.e., relatively small latent dimension). The sensing matrix $\mathbf{A}$ is assumed to have i.i.d. standard normal entries. Let $\mathbf{q} \in \mathrm{Range}(G)$ minimize (4) to within additive $\tau > 0$ of the optimum, i.e.,

$$\|\mathbf{y} - |\mathbf{A}\mathbf{q}|\|_2^2 \leq \min_{\mathbf{w}\in\mathrm{Range}(G)} \|\mathbf{y} - |\mathbf{A}\mathbf{w}|\|_2^2 + \tau, \tag{5}$$

and define $\mathbf{p} \in \mathrm{Range}(G)$ to be the point that is closest to the signal $\mathbf{x}$, i.e.,

$$\mathbf{p} = \arg\min_{\mathbf{w}\in\mathrm{Range}(G)} \|\mathbf{x} - \mathbf{w}\|_2. \tag{6}$$

With the above settings in place, we present the following theorem, proved in Appendix B.

Theorem 1. *Consider the observation model* (1) *with i.i.d. Gaussian measurements, and* $\mathbf{p}, \mathbf{q}$ *satisfying* (5)–(6) *with parameter* τ*. Then, for any* $\delta > 0$*, we have that if* $m = \Omega(k \log \frac{Lr}{\delta})$,[1] *with probability* $1 - e^{-\Omega(m)}$*, it holds that*

$$\min\{\|\mathbf{q} - \mathbf{x}\|_2, \|\mathbf{q} + \mathbf{x}\|_2\} \leq O\left(\|\mathbf{p} - \mathbf{x}\|_2 + \frac{\|\boldsymbol{\eta}\|_2 + \sqrt{\tau}}{\sqrt{m}} + \delta\right). \tag{7}$$

Since we assume that $\mathbf{x}$ is close to the range of G, the representation error $\|\mathbf{p}-\mathbf{x}\|_2$ is small (compared to $\|\mathbf{x}\|_2$). The optimization error τ is also ideally small. In addition, it is common to assume that $\frac{\|\boldsymbol{\eta}\|_2}{\sqrt{m}} \leq c\|\mathbf{x}\|_2$ (see, e.g., [66]), where $c > 0$ is also small. A typical d-layer neural network generative model has $\mathrm{poly}(n)$-bounded weights in each layer, and thus its Lipschitz constant is $L = n^{O(d)}$ [4], where d is the depth. Then, the values of r and δ can be as extreme as $r = n^{O(d)}$ and $\delta = \frac{1}{n^{O(d)}}$ without affecting the final scaling. In this case, Theorem 1 reveals that roughly $O(k \log L)$ samples suffice to guarantee that any optimal solution of the amplitude-based loss function (4) is close to the true signal (up to an unavoidable possible global sign flip). By the analysis of information-theoretic lower bounds in [30, 34], this scaling cannot be improved without further assumptions.

3 Spectral Initializations with Generative Priors

In Theorem 1, we provided recovery guarantees for phase retrieval with generative models with respect to optimal solutions of the amplitude based loss function. However, it is not clear whether such a guarantee is achievable by practical algorithms.

A notable recent work is [24], showing that for a depth-d, width-w ReLU neural network, roughly $O(kd \log w)$ samples[2] suffice for their alternating phase projected gradient descent (APPGD) approach to obtain an accurate estimate of the signal, as long as there exists a proper initial vector $\mathbf{x}^{(0)}$, in the sense that $\min\{\|\mathbf{x} - \mathbf{x}^{(0)}\|_2, \|\mathbf{x} + \mathbf{x}^{(0)}\|_2\} \leq c\|\mathbf{x}\|_2$ for some small positive constant c. However, as we know for sparse phase retrieval, designing a practical algorithm to find a proper initial vector with optimal sample complexity requirements is more difficult than designing the subsequent iterative algorithm that refines the initial guess, and it still remains as an open problem.

Existing initialization methods for sparse phase retrieval typically first estimate the support of the sparse signal, and then perform a power method on a sub-matrix corresponding to the estimated support set to calculate an initial vector. Without extra assumptions on the structure of the sparse

[1] Here and subsequently, all $m = \Omega(\cdot)$ statements are assumed to have a large enough implied constant.

[2] This matches the $O(k \log L)$ scaling upon substituting $L = O(w^d)$ for ReLU networks [4].

signal, these methods have a sub-optimal sample complexity of $O(s^2 \log n)$. For phase retrieval with generative models, the situation is even worse – there is no theoretically-guaranteed initialization method at all. To address this gap, we provide recovery guarantees for a spectral initialization method of phase retrieval with generative models.

We assume that $\mathbf{x}$ is in the range of $G : B_2^k(r) \to \mathbb{R}^n$ (again assumed to be L-Lipschitz continuous), and we assume that the noise terms $\eta_1, \eta_2, \ldots, \eta_m$ are bounded as follows:

$$\frac{\|\boldsymbol{\eta}\|_2}{\sqrt{m}} \leq c_0 \|\mathbf{x}\|_2, \tag{8}$$

where $c_0 > 0$ is a sufficiently small constant. This assumption states that the signal to noise ratio (SNR) is sufficiently large, and has also been made in relevant existing works (see, e.g., [66, Theorem 3]). Moreover, we assume that

$$\|\boldsymbol{\eta}\|_\infty \leq c_1 \|\mathbf{x}\|_2, \tag{9}$$

where $c_1 > 0$ is also a sufficiently small constant. This assumption is satisfied, for example, when $\|\boldsymbol{\eta}\|_\infty = O\big(\frac{\|\boldsymbol{\eta}\|_2}{\sqrt{m}}\big)$ (assuming (8)), i.e., when none of the noise entries are unusually large. Again, similar assumptions have been made in existing works such as [9, Theorem 2], and [65, Theorem 2].[3]

Let λ be defined as

$$\lambda := \sqrt{\frac{\pi}{2}} \cdot \frac{1}{m} \sum_{i=1}^m y_i. \tag{10}$$

Using sub-Gaussian concentration [54, Proposition 5.10], λ is close to $\|\mathbf{x}\|_2$ with high probability given enough samples. In the following, we focus on estimating the normalized signal vector $\bar{\mathbf{x}} := \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. For this purpose, similar to the idea in [32], we consider a normalized generative model $\tilde{G} : \mathcal{D} \to \mathcal{S}^{n-1}$, where $\mathcal{D} := \{\mathbf{z} \in B_2^k(r) : \|G(\mathbf{z})\|_2 > R_{\min}\}$ for some $R_{\min} > 0$,[4] $\mathcal{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$, and $\tilde{G}(\mathbf{z}) = \frac{G(\mathbf{z})}{\|G(\mathbf{z})\|_2}$. Then, $\tilde{G}$ is $\tilde{L}$-Lipschitz continuous with $\tilde{L} = \frac{L}{R_{\min}}$.

Inspired by the initialization method used in [66] for the general phase retrieval problem, we consider the following matrix defined with a truncation operation (see Remark 2 below for a discussion comparing to the non-truncated counterpart):

$$\mathbf{V} := \frac{1}{m} \sum_{i=1}^m y_i \mathbf{a}_i \mathbf{a}_i^T \mathbf{1}_{\{l\lambda < y_i < u\lambda\}}, \tag{11}$$

where l, u are positive constants satisfying $1 < l < u$, and λ is given in (10). Note that for the noiseless case with $y_i = |\langle \mathbf{a}_i, \mathbf{x} \rangle|$, in accordance with the above-mentioned closeness between λ and $\|\mathbf{x}\|_2$, it is useful to consider the following expectation:

$$\mathbf{J} := \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^m y_i \mathbf{a}_i \mathbf{a}_i^T \mathbf{1}_{\{l\|\mathbf{x}\|_2 < y_i < u\|\mathbf{x}\|_2\}}\right] = \|\mathbf{x}\|_2 (\gamma_0 \mathbf{I}_n + \beta_0 \bar{\mathbf{x}}\bar{\mathbf{x}}^T), \tag{12}$$

where for $g \sim \mathcal{N}(0,1)$, we define $\gamma_0 := \mathbb{E}[|g| \mathbf{1}_{\{l<|g|<u\}}]$ and $\beta_0 := \mathbb{E}\big[|g|^3 \mathbf{1}_{\{l<|g|<u\}}\big] - \gamma_0$ (*cf.* Lemma 8 in Appendix C.2). Based on the observation that $\bar{\mathbf{x}}$ is a leading eigenvector of $\mathbf{J}$ and the assumption $\mathbf{x} \in \mathrm{Range}(G)$, we use the following optimization problem to find $\hat{\mathbf{x}}$ approximating $\bar{\mathbf{x}}$:

$$\hat{\mathbf{x}} := \arg \max_{\mathbf{w} \in \tilde{G}(\mathcal{D})} \quad \mathbf{w}^T \mathbf{V} \mathbf{w}. \tag{13}$$

In addition, in order to take the norm of $\mathbf{x}$ into account, we set the the initial vector $\mathbf{x}^{(0)}$ as

$$\mathbf{x}^{(0)} = \lambda \hat{\mathbf{x}}. \tag{14}$$

We show that roughly $O(k \log L)$ samples suffice for ensuring the closeness between $\mathbf{x}^{(0)}$ and $\mathbf{x}$ (up to a global signal flip). Formally, we have the following.

[3] In these papers, the authors consider an intensity based measurement model, and accordingly, the corresponding assumption is for $\frac{\|\boldsymbol{\eta}\|_\infty}{\|\mathbf{x}\|_2^2}$ instead of for $\frac{\|\boldsymbol{\eta}\|_\infty}{\|\mathbf{x}\|_2}$.

[4] As discussed in [32], the dependence on $R_{\min}$ in the sample complexity is very mild. Under the typical scaling $L = n^{O(d)}$ for a d-layer neural network, the scaling laws remain unchanged even with $R_{\min} = \frac{1}{n^{O(d)}}$.

Theorem 2. *Consider the model* (1) *with i.i.d. Gaussian measurements and noise satisfying* (8)–(9)*, and let* $\hat{\mathbf{x}}$ *be as defined in* (13)*. Suppose that*[5] $\bar{\mathbf{x}} \in \tilde{G}(\mathcal{D})$ *and* $u > l > 1 + c_1$*, where* $c_1 \geq \frac{\|\boldsymbol{\eta}\|_\infty}{\|\mathbf{x}\|_2}$ *appears in* (9)*. Given a sufficiently small positive constant* c*, when* $m = \Omega\big(k \log(\tilde{L}nr)\big)$*, we have with probability* $1 - e^{-\Omega(m)}$ *that*

$$\min\{\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\|_2, \|\hat{\mathbf{x}} + \bar{\mathbf{x}}\|_2\} < 0.9c. \tag{15}$$

In addition, we have with probability $1 - e^{-\Omega(m)}$ *that* $\mathbf{x}^{(0)} = \lambda\hat{\mathbf{x}}$ *satisfies*

$$\min\{\|\mathbf{x}^{(0)} - \mathbf{x}\|_2, \|\mathbf{x}^{(0)} + \mathbf{x}\|_2\} < c\|\mathbf{x}\|_2. \tag{16}$$

The proof is given in Appendix C. Using well-established chaining arguments as those in [4, 32, 33], the sample complexity $\Omega(k \log(\tilde{L}nr))$ can be reduced to $\Omega(k \log(\tilde{L}r))$. However, since for a typical d-layer neural network, we have the Lipschitz constant $L = n^{O(d)}$ (and thus $\tilde{L}$ and r can also be of order $n^{O(d)}$) [4], such a reduction is typically of minor importance.

4 Sparse Phase Retrieval

The proof of Theorem 2 relies on the fact that $\bar{\mathbf{x}}$ lies in a subset of the unit sphere with the δ-covering number being upper bounded by $\big(\frac{\tilde{L}r}{\delta}\big)^k$. Letting Σ_s^n is the set of all s-sparse unit vectors in $\mathbb{R}^n$, we know that the δ-covering number of Σ_s^n is upper bounded by $\binom{n}{s}\big(\frac{1}{\delta}\big)^s \leq \big(\frac{en}{\delta s}\big)^s$ [2]. Based on this observation, we can readily adapt the analysis in Theorem 2 and obtain the following theorem.

Theorem 3. *Consider the model* (1) *with i.i.d. Gaussian measurements and noise satisfying* (8)–(9)*, and suppose that* $\mathbf{x} \in \mathbb{R}^n$ *is* s*-sparse. Let* λ *be defined as in* (10)*. Suppose that* $u > l > 1 + c_1$*, where* $c_1 \geq \frac{\|\boldsymbol{\eta}\|_\infty}{\|\mathbf{x}\|_2}$ *appears in* (9)*, and* $\mathbf{V}$ *is defined in* (11)*. Let* $\hat{\mathbf{x}}$ *be defined as*

$$\hat{\mathbf{x}} := \arg\max_{\mathbf{w} \in \Sigma_s^n} \mathbf{w}^T \mathbf{V} \mathbf{w}. \tag{17}$$

Then, when $m = \Omega\big(s \log n\big)$*, we have with probability* $1 - e^{-\Omega(m)}$ *that*

$$\min\{\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\|_2, \|\hat{\mathbf{x}} + \bar{\mathbf{x}}\|_2\} < 0.9c, \tag{18}$$

where $c > 0$ *is a sufficiently small constant. In addition, we have with probability* $1 - e^{-\Omega(m)}$ *that* $\mathbf{x}^{(0)} = \lambda\hat{\mathbf{x}}$ *satisfies*

$$\min\{\|\mathbf{x}^{(0)} - \mathbf{x}\|_2, \|\mathbf{x}^{(0)} + \mathbf{x}\|_2\} < c\|\mathbf{x}\|_2. \tag{19}$$

Implications regarding sparse PCA and support recovery: We note that (17) is essentially a problem of sparse principal component analysis (SPCA) [67, 35]. Existing spectral initialization methods for sparse phase retrieval typically start from estimating the support of $\mathbf{x}$, and an accurate estimate of the support requires a sub-optimal $O(s^2 \log n)$ sample complexity [6, 58, 27], which is the major bottleneck of the sample complexity upper bound of the whole procedure (i.e., the initialization step and the subsequent iterative step). It is unknown whether there is a practical and sample-optimal approach for estimating the support. In fact, even the following question has not been answered previously:

Key question: *Does there exist any spectral-based method (computationally efficient or otherwise) attaining an accurate initial vector with order-optimal sample complexity?*

We provide a positive answer to this question, showing that $O(s \log n)$ samples suffice to ensure any optimal solution of the SPCA problem (17) is close to the underlying (normalized) signal.

Motivated by the strong sample complexity guarantee provided in Theorem 3, we propose a spectral initialization method for sparse phase retrieval as described in Algorithm 1. In particular, after calculating $\mathbf{V}$ from $\mathbf{A}$ and $\mathbf{y}$, we use algorithms for SPCA to solve (17). We observe via numerical experiments in Section 5 that our initialization method significantly outperforms several popular spectral initialization methods used for sparse phase retrieval.

[5] This is weaker than the assumption that $\|\mathbf{x}\|_2 > R_{\min}$ made in [32]. Moreover, even if $\|\mathbf{x}\|_2 \leq R_{\min}$, the condition $\bar{\mathbf{x}} \in \tilde{G}(\mathcal{D})$ may be satisfied (e.g., when G is a ReLU neural network generative model with no offsets, if $0 < \|\mathbf{x}\|_2 \leq R_{\min}$, we can choose a sufficiently large C such that $C\|\mathbf{x}\|_2 > R_{\min}$, with $C\mathbf{x} \in \mathrm{Range}(G)$).

Algorithm 1 A spectral initialization method for sparse phase retrieval based on SPCA (`PRI-SPCA`)

Input: $\mathbf{A}$, $\mathbf{y}$, l, u
Output: An initial vector $\mathbf{x}^{(0)}$
1) Calculate λ from (10) and then calculate $\mathbf{V}$ from (11).
2) Use an SPCA algorithm to compute $\hat{\mathbf{x}}$ according to (17).
3) Let $\mathbf{x}^{(0)} = \lambda\hat{\mathbf{x}}$.

Remark 1. *In various works on SPCA, it has been shown that there exists a statistical-computational gap for achieving consistency and optimal convergence rates uniformly over a parameter space (e.g., see [3, 59, 6]). However, the models therein are all different from ours (e.g., the spiked covariance model [16]), so it is unclear to what extent we should expect similar computational difficulties.*

Remark 2. *The truncation function* $\mathbf{1}_{\{l\lambda<y_i<u\lambda\}}$ *in* (11) *plays an important role in reducing the required sample complexity in our analysis. Without the truncation function, if we use the same proof technique to derive a sample complexity bound, we need to make use of concentration inequalities for random variables that have heavier tail than sub-exponential random variables, e.g., sub-Weibull random variables of order* $\alpha < 1$ *(see, e.g., [20, Theorem 3.1]). This leads to a larger sample complexity bound. In Section 5, we present numerical results for applying SPCA to the non-truncated weighted empirical matrix*

$$\tilde{\mathbf{V}} = \frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i \mathbf{a}_i^T. \tag{20}$$

We will see in our experiments that the method corresponding to the truncated empirical matrix $\mathbf{V}$ *significantly outperforms the method corresponding to the non-truncated empirical matrix* $\tilde{\mathbf{V}}$*, although the expectations of these two matrices have a very similar structure. Thus, we expect that truncation has the potential of overcoming the computational challenges discussed above.*

We note that [40, 50] consider solving sparse phase retrieval via practical SPCA methods, in a more general setting considering an unknown link function. They consider a weighted empirical covariance matrix with no truncation, and the derived sample complexity has a quadratic dependence on s.

5 Numerical Experiments

In this section, we perform synthetic experiments to complement our theoretical results. We focus only on the sparse setting, since numerous existing algorithms are known for sparse priors but not for generative priors. Specifically, we compare our method (`PRI-SPCA`; see Algorithm 1) with spectral initialization methods used in several popular algorithms for sparse phase retrieval: Thresholded Wirtinger flow (`ThWF`) [6], sparse truncated amplitude flow (`SPARTA`) [58], and compressive phase retrieval with alternating minimization (`CoPRAM`) [27].[6] We also compare with the approach of solving an SPCA problem with respect to $\tilde{\mathbf{V}}$ (*cf.* (20)), where $\tilde{\mathbf{V}}$ is a non-truncated version of $\mathbf{V}$. The corresponding method is termed `PRI-SPCA-NT`, with 'NT' standing for no truncation. For `PRI-SPCA`, we set $l = 1$ and $u = 5$, as previously suggested in [66].

Various algorithms have been developed to tackle the SPCA problem, see, e.g., [38, 29, 36]. In our experiments, we present the numerical results using the generalized Rayleigh quotient iteration (`GRQI`) method [31], but we observed that other popular SPCA methods such as the truncated power method (`TPower`) [64] give similar averaged relative errors.[7] For `GRQI`, we set the deflation parameter to be 0.2, and the total number of iterations to be 100. Since the sparsity level s is typically assumed to be known a priori for `SPARTA` and `CoPRAM` (`ThWF` assumes the knowledge of a parameter that plays the similar role as s), for `GRQI`, we set the parameter about the maximum number of non-zero indices as s. Note that the time complexity of each iteration of `GRQI` is $O(ns + s^3)$, which scales well with the subsequent $O(s^2 n \log n)$ time complexity of the iterative methods used in the three approaches (for `ThWF`, the corresponding time complexity is $O(n^2 \log n)$. See [27, Table I]). The number of power method steps used in all the methods is fixed to be 100.

[6] We use the MATLAB package shared by the authors of [27] at `https://github.com/GauriJagatap/model-copram`. The package not only contains the codes for `CoPRAM`, but also for `ThWF` and `SPARTA`.

[7] Although using `TPower` also results in small relative error for our `PRI-SPCA` method, in our experiments, as we want a fair comparison and do not carefully design the starting point of `TPower`, it is not as stable as `GRQI`.

The authors of [31] suggest using the column of maximal norm as a starting point of GRQI. Because this policy has a time complexity of $O(n^2)$, we instead choose the column corresponding to the largest diagonal entry as the starting point. Considering the structure of $\mathbf{V}$ (*cf.* (12)), this modified strategy is in fact identical to the suggested policy, but with a smaller time complexity requirement. For a fair comparison, all power methods used in ThWF, SPARTA, and CoPRAM are similarly initialized. For the initialization method used in ThWF, we set the tuning parameter $\alpha = 0.1$, as suggested in [6]. Similarly, we set $|\bar{\mathcal{I}}| = \lceil \frac{1}{6}m \rceil$ for the initialization method of SPARTA.

The support of the signal vector $\mathbf{x}$ is uniformly random, and the non-zero entries are generated from i.i.d. standard normal distributions. The noise vector $\boldsymbol{\eta}$ is generated from the distribution $\mathcal{N}(\mathbf{0}, \sigma^2 \|\mathbf{x}\|_2^2 \mathbf{I}_m)$, where $\sigma > 0$ is a parameter indicating the noise level. The initial relative error is defined as $\frac{\min\{\|\mathbf{x}-\mathbf{x}^{(0)}\|_2, \|\mathbf{x}+\mathbf{x}^{(0)}\|_2\}}{\|\mathbf{x}\|_2}$, where $\mathbf{x}^{(0)}$ is an initial vector produced by each of the above mentioned initialization methods. The reported relative error is averaged over 50 random trials.[8]

1. **Sample size effect:** In our first experiment, similarly to [58], we consider noiseless measurements, and we fix the signal dimension as $n = 1000$. We consider s equaling both 10 and 20. The number of measurements m increases in $\{100, 200, \ldots, 3000\}$. From Figure 1, we observe that our PRI-SPCA method almost always achieves the smallest initial relative error, thus outperforming all other methods, including the PRI-SPCA-NT method, which uses the non-truncated empirical matrix $\tilde{\mathbf{V}}$.

2. **Sparsity effect**: We consider noiseless measurements and fix $n = 1000$, and set $m = 1000$ or 2000. We vary s in $\{5, 10, \ldots, 50\}$. We observe from Figure 2 that for this experiment, our PRI-SPCA method also gives the best numerical performance.

3. **Noise effect**: To demonstrate that PRI-SPCA also performs well in the noisy case, we vary σ in $\{0.1, 0.2, \ldots, 1\}$. In addition, similarly to [58], we fix $n = 1000$, $m = 3000$, and consider s equaling both 10 and 20. For this experiment, similarly to that in [27], we do not compare with ThWF, as it uses quadratic measurements. From Figure 3, we observe that PRI-SPCA outperforms SPARTA and CoPRAM, and it is better than PRI-SPCA-NT at low noise levels.

In Appendix E, we additionally compare with other initialization methods in terms of the effectiveness when combining with the subsequent iterative algorithm of CoPRAM. In particular, we compare the relative error and empirical success rate for the noiseless case, and for the noisy case, we compare the relative error when using approximately the same running time. We observe that for most cases, our PRI-SPCA method is superior to all other methods we compare with. This again suggests that PRI-SPCA can serve as a good initialization method in practice.

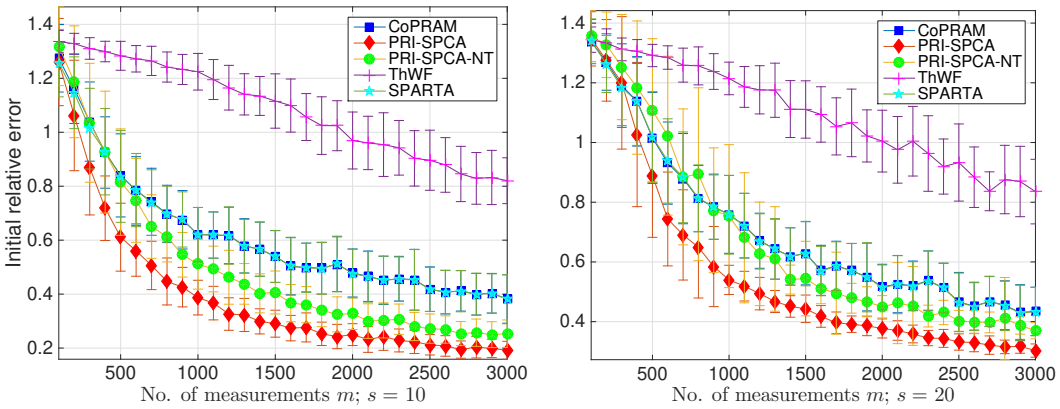

Figure 1: The relation between the average relative error and the number of measurements m.

[8] All numerical experiments are conducted using MATLAB R2014b on a computer with an Intel Core i5 CPU at 1.8 GHz and 8GB RAM.

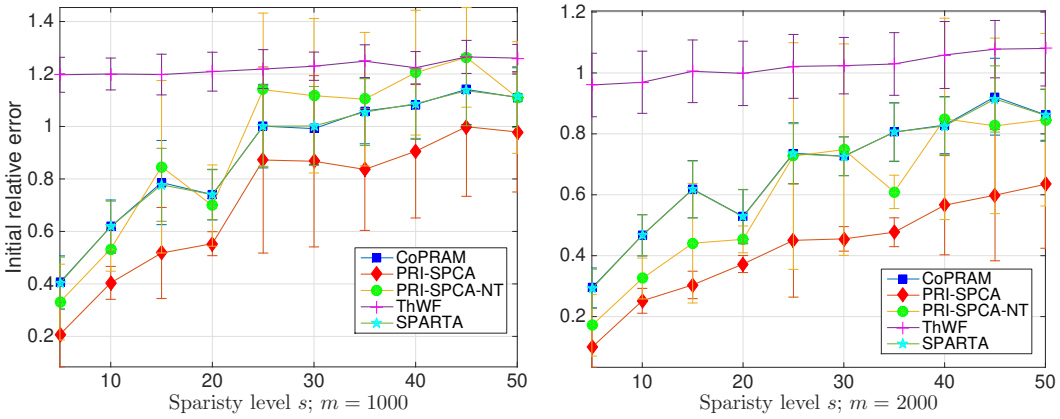

Figure 2: The relation between the average relative error and the sparsity level s.

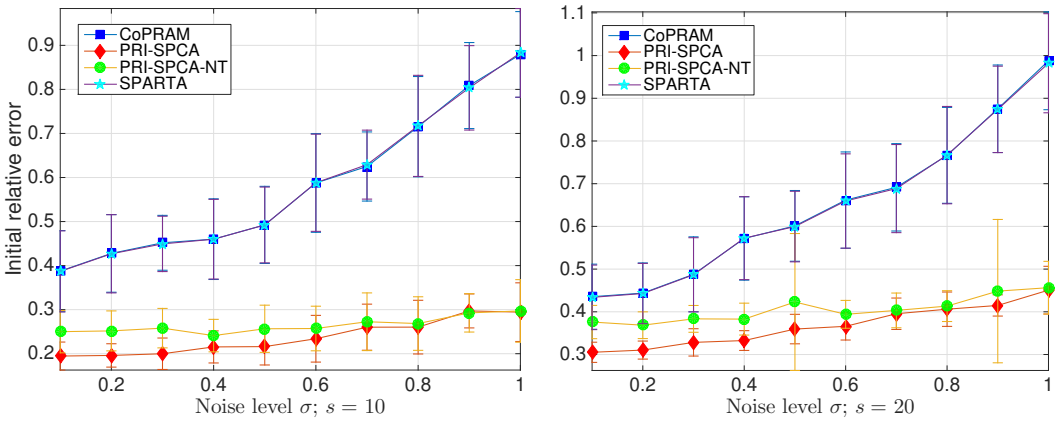

Figure 3: The relation between the average relative error and the noise level σ.

6 Conclusion and Future Work

We have provided an order-optimal recovery guarantee for amplitude-based loss minimization for phase retrieval with generative priors. In addition, motivated by the bottleneck of the spectral initialization for compressive phase retrieval, we have provided order-optimal recovery guarantees with respect to the optimal solutions of optimization problems designed for spectral initialization. An immediate future research direction is to design provably sample-optimal *and computationally efficient* spectral initialization methods that build on our theoretical results. In addition, extensions to complex-valued phase retrieval would be of significant interest.

References

[1] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: A flexible convex relaxation," in *Int. Conf. Artif. Intell. Stat. (AISTATS)*. PMLR, 2017, pp. 252–260.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[3] Q. Berthet and P. Rigollet, "Complexity theoretic lower bounds for sparse principal component detection," in *Conf. Learn. Theory (COLT)*. PMLR, 2013, pp. 1046–1066.

[4] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 537–546.

[5] J.-F. Cai, J. Li, X. Lu, and J. You, "Sparse signal recovery from phaseless measurements via hard thresholding pursuit," *https://arxiv.org/abs/2005.08777*, 2020.

[6] T. T. Cai, X. Li, Z. Ma *et al.*, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *Ann. Stat.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[7] E. J. Candès, X. Li, and M. Soltanolkotabi, "Phase retrieval via Wirtinger flow: Theory and algorithms," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 1985–2007, 2015.

[8] E. J. Candès, T. Strohmer, and V. Voroninski, "PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming," *Comm. Pure Appl. Math.*, vol. 66, no. 8, pp. 1241–1274, 2013.

[9] Y. Chen and E. J. Candès, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," *Comm. Pure Appl. Math.*, vol. 70, no. 5, pp. 822–883, 2017.

[10] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Math. Program.*, vol. 176, no. 1, pp. 5–37, 2019.

[11] J. Cocola, P. Hand, and V. Voroninski, "Nonasymptotic guarantees for spiked matrix recovery with generative priors," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 33, 2020.

[12] Y. C. Eldar and S. Mendelson, "Phase retrieval: Stability and recovery guarantees," *Appl. Comput. Harmon. Anal.*, vol. 36, no. 3, pp. 473–494, 2014.

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

[14] D. Foster, *Generative Deep Learning : Teaching Machines to Paint, Write, Compose and Play*. O'Reilly Media, Inc, USA, 2019.

[15] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

[16] C. Gao, Z. Ma, H. H. Zhou *et al.*, "Sparse CCA: Adaptive estimation and computational barriers," *Ann. Stat.*, vol. 45, no. 5, pp. 2074–2101, 2017.

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

[18] T. Goldstein and C. Studer, "Phasemax: Convex phase retrieval via basis pursuit," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2675–2689, 2018.

[19] P. Hand, O. Leong, and V. Voroninski, "Phase retrieval under a generative prior," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2018, pp. 9154–9164.

[20] B. Hao, Y. Yadkori, Z. Wen, and G. Cheng, "Bootstrapping upper confidence bound," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2019.

[21] R. Heckel and P. Hand, "Deep decoder: Concise image representations from untrained non-convolutional networks," in *Int. Conf. Learn. Repr. (ICLR)*, 2019.

[22] M. Huang and Z. Xu, "The estimation performance of nonlinear least squares for phase retrieval," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7967–7977, 2020.

[23] M. Huang and Z. Xu, "Performance bound of the intensity-based model for noisy phase retrieval," *https://arxiv.org/abs/2004.08764*, 2020.

[24] R. Hyder, V. Shah, C. Hegde, and M. S. Asif, "Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval," in *Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*. IEEE, 2019, pp. 7705–7709.

[25] M. Iwen, A. Viswanathan, and Y. Wang, "Robust sparse phase retrieval made easy," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 42, pp. 135–142, 2017.

[26] G. Jagatap and C. Hegde, "Algorithmic guarantees for inverse imaging with untrained network priors," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 32, 2019.

[27] G. Jagatap and C. Hegde, "Sample-efficient algorithms for recovering structured signals from magnitude-only measurements," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4434–4456, 2019.

[28] A. Jalal, L. Liu, A. G. Dimakis, and C. Caramanis, "Robust compressed sensing using generative models," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2020.

[29] M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre, "Generalized power method for sparse principal component analysis." *J. Mach. Learn. Res.*, vol. 11, no. 2, 2010.

[30] A. Kamath, S. Karmalkar, and E. Price, "Lower bounds for compressed sensing with generative models," *https://arxiv.org/abs/1912.02938*, 2019.

[31] V. Kuleshov, "Fast algorithms for sparse principal component analysis based on Rayleigh quotient iteration," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2013, pp. 1418–1425.

[32] Z. Liu, S. Gomes, A. Tiwari, and J. Scarlett, "Sample complexity bounds for 1-bit compressive sensing and binary stable embeddings with generative priors," in *Int. Conf. Mach. Learn. (ICML)*, 2020.

[33] Z. Liu and J. Scarlett, "The generalized Lasso with nonlinear observations and generative priors," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 33, 2020.

[34] Z. Liu and J. Scarlett, "Information-theoretic lower bounds for compressive sensing with generative models," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 292–303, 2020.

[35] T. Ma and A. Wigderson, "Sum-of-squares lower bounds for sparse PCA," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 28, pp. 1612–1620, 2015.

[36] Z. Ma *et al.*, "Sparse principal component analysis and iterative thresholding," *Ann. Stat.*, vol. 41, no. 2, pp. 772–801, 2013.

[37] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.

[38] B. Moghaddam, Y. Weiss, and S. Avidan, "Spectral bounds for sparse PCA: Exact and greedy algorithms," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 18, p. 915, 2006.

[39] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," *IEEE Trans. Sig. Proc.*, vol. 63, no. 18, pp. 4814–4826, 2015.

[40] M. Neykov, Z. Wang, and H. Liu, "Agnostic estimation for misspecified phase retrieval models," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2016.

[41] H. Ohlsson, A. Yang, R. Dong, and S. Sastry, "CPRL–An extension of compressive sensing to the phase retrieval problem," *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, vol. 25, pp. 1367–1375, 2012.

[42] G. Ongie, A. Jalal, C. A. Metzler, R. G. Baraniuk, A. G. Dimakis, and R. Willett, "Deep learning techniques for inverse problems in imaging," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 39–56, 2020.

[43] Y. Plan and R. Vershynin, "The generalized Lasso with non-linear observations," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1528–1537, 2016.

[44] J. Scarlett and V. Cevher, "An introductory guide to Fano's inequality with applications in statistical estimation," *https://arxiv.org/1901.00555*, 2019.

[45] P. Schniter and S. Rangan, "Compressive phase retrieval via generalized approximate message passing," *IEEE Trans. Sig. Proc.*, vol. 63, no. 4, pp. 1043–1055, 2014.

[46] F. Shamshad and A. Ahmed, "Compressed sensing-based robust phase retrieval via deep generative priors," *IEEE Sens. J.*, vol. 21, no. 2, pp. 2286–2298, 2020.

[47] Y. Shechtman, A. Beck, and Y. C. Eldar, "GESPAR: Efficient phase retrieval of sparse signals," *IEEE Trans. Sig. Proc.*, vol. 62, no. 4, pp. 928–938, 2014.

[48] M. Soltanolkotabi, "Structured signal recovery from quadratic measurements: Breaking sample complexity barriers via nonconvex optimization," *IEEE Trans. Inf. Theory*, vol. 65, no. 4, pp. 2374–2400, 2019.

[49] J. Sun, Q. Qu, and J. Wright, "A geometric analysis of phase retrieval," *Found. Comput. Math.*, vol. 18, no. 5, pp. 1131–1198, 2018.

[50] Y. S. Tan, "Sparse phase retrieval via sparse PCA despite model misspecification: A simplified and extended analysis," *https://arxiv.org/1712.04106*, 2017.

[51] L. V. Truong and J. Scarlett, "Support recovery in the phase retrieval model: Information-theoretic fundamental limits," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7887–7910, 2020.

[52] D. Van Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *https://arxiv.org/1806.06438*, 2018.

[53] S. S. Vempala, *The random projection method*. American Mathematical Soc., 2005, vol. 65.

[54] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," *https://arxiv.org/abs/1011.3027*, 2010.

[55] V. Voroninski and Z. Xu, "A strong restricted isometry property, with an application to phaseless compressed sensing," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 386–395, 2016.

[56] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Math. Program.*, vol. 149, no. 1, pp. 47–81, 2015.

[57] G. Wang, G. B. Giannakis, and Y. C. Eldar, "Solving systems of random quadratic equations via truncated amplitude flow," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 773–794, 2017.

[58] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen, "Sparse phase retrieval via truncated amplitude flow," *IEEE Trans. Sig. Proc.*, vol. 66, no. 2, pp. 479–491, 2017.

[59] T. Wang, Q. Berthet, R. J. Samworth *et al.*, "Statistical and computational trade-offs in estimation of sparse principal components," *Ann. Stat.*, vol. 44, no. 5, pp. 1896–1930, 2016.

[60] X. Wei, Z. Yang, and Z. Wang, "On the statistical rate of nonlinear recovery in generative models with heavy-tailed data," in *Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6697–6706.

[61] J. Whang, Q. Lei, and A. G. Dimakis, "Compressed sensing with invertible generative models and dependent noise," *https://arxiv.org/2003.08089*, 2020.

[62] F. Wu and P. Rebeschini, "Hadamard Wirtinger flow for sparse phase retrieval," in *Int. Conf. Artif. Intell. Stat. (AISTATS)*. PMLR, 2021, pp. 982–990.

[63] Y. Xia and Z. Xu, "The recovery of complex sparse signals from few phaseless measurements," *Appl. Comput. Harmon. Anal.*, vol. 50, pp. 1–15, 2021.

[64] X.-T. Yuan and T. Zhang, "Truncated power method for sparse eigenvalue problems," *J. Mach. Learn. Res.*, vol. 14, no. 4, 2013.

[65] H. Zhang, Y. Chi, and Y. Liang, "Provable non-convex phase retrieval with outliers: Median truncated Wirtinger flow," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2016, pp. 1022–1031.

[66] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi, "A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms," *J. Mach. Learn. Res.*, vol. 18, 2017.

[67] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat.*, vol. 15, no. 2, pp. 265–286, 2006.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The limitations motivate our future work, and are described in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [No] Some proofs are not included to avoid repetition. In particular, in Appendix B.1, we do not provide the full proof of Lemma 6 because it can be directly derived from Lemma 5 and a standard (but lengthy) chaining argument used in [4, 32]. We do not provide a proof of Theorem 3 because the proof is identical to the proof of Theorem 2, except that the covering number of the corresponding set is changed.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The codes are included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]