Misspecified Phase Retrieval with Generative Priors

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we study phase retrieval under model misspecification and generative priors. In particular, we aim to estimate an n-dimensional signal $\mathbf{x}$ from m i.i.d. realizations of the single index model $y = f(\mathbf{a}^T\mathbf{x})$, where f is an unknown and possibly random nonlinear link function and $\mathbf{a} \in \mathbb{R}^n$ is a standard Gaussian vector. We make the assumption $\mathrm{Cov}[y, (\mathbf{a}^T\mathbf{x})^2] \neq 0$, which corresponds to the misspecified phase retrieval problem. In addition, the underlying signal $\mathbf{x}$ is assumed to lie in the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs. We propose a two-step approach, for which the first step plays the role of spectral initialization and the second step refines the estimated vector produced by the first step iteratively. We show that both steps enjoy a statistical rate of order $\sqrt{(k \log L) \cdot (\log m)/m}$ under suitable conditions. Experiments on image datasets are performed to demonstrate that our approach performs on par with or even significantly outperforms several competing methods.

1 Introduction

Compressed sensing (CS) is perhaps the most popular instance of high-dimensional inverse problems, for which one has the *linear* measurement model

$$y = \mathbf{a}^T\mathbf{x} + \eta, \tag{1}$$

where $\mathbf{a} \in \mathbb{R}^n$ is the sensing vector, $\mathbf{x} \in \mathbb{R}^n$ is the sparse signal to estimate, and η represents additive noise. It has been well-known for CS that roughly $m = O(s \log(n/s))$ i.i.d. Gaussian measurements are sufficient to ensure the accurate recovery of a signal with s non-zero entries [79, 1, 22, 10, 69].

Phase retrieval (PR) arises in numerous scientific areas including X-ray crystallography, acoustics, astronomy, microscopy, optics, wireless communications, and quantum information [12], where one cannot measure $\mathbf{a}^T\mathbf{x}$ directly, and can only record its magnitude. For example, the following *noisy magnitude-only* measurement model has been adopted in various prior works on sparse (real-valued) PR [83, 38, 24, 35, 8, 6]:

$$y = |\mathbf{a}^T\mathbf{x}| + \eta, \tag{2}$$

where the signal $\mathbf{x}$ is assumed to be sparse, and the sensing vector $\mathbf{a}$ is assumed to be a standard Gaussian vector.

However, both the linear and magnitude-only measure models in (1) and (2) are idealized views of the data generating process. In practice, it is more desirable to utilize the following semi-parametric *single index model (SIM)* for general nonlinear models:

$$y = f(\mathbf{a}^T\mathbf{x}), \tag{3}$$

where $f : \mathbb{R} \to \mathbb{R}$ is an *unknown* (possibly random) nonlinear link function, and $\mathbf{a}$ is typically assumed to be Gaussian. In addition, since the norm of $\mathbf{x} \in \mathbb{R}^n$ is absorbed into the SIM, for brevity, it is common to assume that $\mathbf{x}$ is a unit vector, i.e., $\|\mathbf{x}\|_2 = 1$. SIMs have long been studied in the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

conventional setting where the number of measurements $m > n$ [29, 72, 44]. In recent years, they have also been analyzed in the high-dimensional setting where $m \ll n$, mainly under the assumption that the underlying signal is sparse. Relevant works include but are not limited to [64, 55, 25, 65, 60]. For all these works, it is crucial to impose the following assumption on the SIM:

$$\mathrm{Cov}[y, \mathbf{a}^T\mathbf{x}] \neq 0. \tag{4}$$

The pivotal condition in (4) is fairly generic and encompasses notable special examples such as noisy 1-bit measurements and general (non-binary) quantization schemes. However, it is not satisfied by PR models including (2) and

$$y = |\mathbf{a}^T\mathbf{x} + \eta|, \quad y = (\mathbf{a}^T\mathbf{x})^2 + \eta, \tag{5}$$

where η refers to zero-mean random noise that is independent of $\mathbf{a}$. In order to formalize a class of SIMs that encompass the above-mentioned PR models (and more general models, see the discussion in Section 2.2) as special cases, misspecified phase retrieval (MPR) has been studied in [56, 89], with the condition (4) being replaced by the assumption

$$\mathrm{Cov}[y, (\mathbf{a}^T\mathbf{x})^2] \neq 0. \tag{6}$$

In both works [56, 89], the signal $\mathbf{x}$ is assumed to be sparse. Recently, motivated by tremendous successful applications of deep generative models and following the seminal work [5] on generative model based linear CS, it has been popular to study high-dimensional inverse problems under generative priors [68, 30, 31, 33, 87, 40, 2, 59, 86, 53, 39, 57]. In particular, instead of being sparse, the underlying signal is assumed to be contained in (or lie near) the range of a generative model. It has been empirically demonstrated in [5] and its various follow-up works (e.g., see [18, 76, 46]) that compared to sparsity based methods, corresponding generative model based algorithms require significantly fewer samples to recover the signal up to a given accuracy.

In this paper, we study the MPR problem under generative modeling assumptions.

1.1 Related Work

In this subsection, we summarize some relevant works on PR and SIM, for both cases with or without generative priors.

PR and SIM without generative priors: There is a large amount of literature providing practical algorithms for PR, including convex methods [58, 13, 11, 81, 4, 28] and empirically more competitive non-convex methods [26, 20, 54, 62]. In particular, in the seminar work [54] and a variety of its follow-up works [12, 9, 14, 83, 82, 90, 38, 73, 7], no matter it is for general PR (with no priors on the signal) or sparse PR, two-step approaches have been proposed with provable guarantees. The first step consists of a spectral initialization method, and the second step is typically an iterative (such as alternating minimization and gradient descent) algorithm that further refines the initial guess of the first step. For general PR, the spectral initialization step turns out to be not necessary, and optimal sample complexity guarantees can be established even when using random initialization [74, 15]. However, to the best of our knowledge, it is not the case for sparse PR. More specifically, all theoretically-guaranteed non-convex algorithms for sparse PR require spectral initialization, and this typically results in a sub-optimal sample complexity of $O\big(s^2 \log n\big)$ (instead of $O(s \log n)$), where s refers to the number of non-zero entries of the signal to estimate.

MPR is also closely related to SIMs, which have been extensively studied in the conventional setting where $m > n$, see, e.g., [29, 34, 52]. High-dimensional SIMs have received a lot of attention in recent years, with theoretical guarantees for signal estimation and support recovery [21, 23, 66, 75, 51, 64, 65, 16, 88, 27, 84, 61, 19]. In particular, motivated by the idea that under the assumption (4), an SIM can be converted into a scaled linear measurement model with unconventional noise, the authors of [64] consider minimizing the linear least-squares objective function over a convex set. They show that a reliable estimation of the signal can be obtained by such a simple method despite the unknown nonlinear link function.

As mentioned earlier, MPR with sparse priors has been studied in [56, 89]. The work [56] implements a two-step procedure based on convex optimization, with the first step being a spectral initialization method. More specifically, a semidefinite program is solved to produce an initial vector, and then an ℓ_1 regularized least square is solved to obtain a refined estimator. Such a procedure suffers from high

computational costs. A more efficient two-step approach, which is a simple variant of the thresholded Wirtinger flow method [9], is proposed in [89]. In the first step, identical to that in [9], the initial vector is calculated by a thresholded spectral method that first estimates the support of the sparse signal by thresholding and then performs the classic power method over the submatrix corresponding to the estimated support. In the second step, a thresholded gradient descent algorithm is employed. Both approaches in [56, 89] can attain the optimal statistical rate of order $\sqrt{(s \log n)/m}$, provided that the sensing vector is standard Gaussian and the number of samples $m = \Omega\big(s^2 \log n\big)$. In addition, the second step of the approach proposed in [89] is shown to achieve a linear convergence rate.

PR and SIM with generative priors: PR with generative priors has been studied in [30, 36, 37, 71, 3, 45]. More specifically, an approximate message passing algorithm is proposed in [3]. The authors of [30, 71] minimize the objective over the latent space in $\mathbb{R}^k$ using gradient descent, where k is the latent dimension of the generative prior. Corresponding algorithms may easily get stuck in local minima since the explorable solution space is limited. Recovery guarantees for projected gradient descent algorithms over the ambient space in $\mathbb{R}^n$ for noiseless PR with pre-trained or untrained neural network priors have been proposed in [36, 37]. No initialization methods have been proposed in these works, making the assumption on the initial vector therein a bit stringent. On the other hand, the authors of [45] propose a spectral initialization method for PR with generative priors and provide recovery guarantees with respect to globally optimal solutions of a corresponding optimization problem. The optimization problem is non-convex, and a projected power method is proposed in [47] to approximately find an optimal solution.

Generative model based SIMs have been studied in [48, 85]. The authors of [48] provide optimal sample complexity upper bounds under the assumption of Gaussian sensing vectors. But their results rely on the assumption (4), which is not satisfied by widely adopted PR models. The SIM studied in [85] encompasses certain PR models as special cases, and the sensing vector can be non-Gaussian. However, the nonlinear link function f is assumed to be differentiable, making it not directly applicable to the PR model in (2). Moreover, it is worth mentioning that in both works [85, 48], the recovery guarantees are with respect to globally optimal solutions of typically highly non-convex optimization problems. Attaining these optimal solutions is practically difficult.

1.2 Contributions

Throughout this paper, we assume that the signal $\mathbf{x}$ lies in the range of an L-Lipschitz continuous generative model with bounded k-dimensional inputs. Our main contributions are as follows:

- We propose a two-step approach for MPR with generative priors. In particular, in the first step, we make use of the projected power method proposed in [47] to obtain a good initial vector for the iterative algorithm used in the second step.

- We show that under appropriate initialization, both steps attain a statistical rate of order $\sqrt{(k \log L) \cdot (\log m)/m}$, and the second step achieves a linear convergence rate. The initialization condition for the first step is mild in the sense that it allows the inner product between the starting point and the signal $\mathbf{x}$ to be a sufficiently small positive constant. In contrary, the initialization condition for the second step is more restrictive, making the first step necessary. Notably, unlike for the works on MPR with sparse priors [56, 89] that require the sub-optimal $O\big(s^2 \log n\big)$ sample complexity, the required number of samples for our recovery guarantees is $O((k \log L) \cdot (\log m))$, which is naturally conjectured to be near-optimal [49, 41].

- Numerical experiments are performed on image datasets to corroborate our theoretical results. In particular, for the noisy magnitude-only measurement model (2), we observe that our approach gives reconstructions that are competitive with those of the alternating phase projected gradient descent (APPGD) algorithm proposed in [36], which is the corresponding state-of-the-art method, though we do not utilize the knowledge of the link function and allow for model misspecification. Moreover, for several closely related measurement models that satisfy the condition (6), our approach leads to superior reconstruction performance compared to all other competing methods, including APPGD.

2 Problem Formulation

In this section, we overview some important assumptions that we adopt. Before proceeding, we present the notation used in this paper.

2.1 Notation

We use upper and lower case boldface letters to denote matrices and vectors respectively. For any positive integer N, we use the shorthand notation $[N] = \{1, 2, \cdots, N\}$, and $\mathbf{I}_N$ represents the identity matrix in $\mathbb{R}^{N\times N}$. The support (set) of a vector is the index set of its non-zero entries. We use $\|\mathbf{X}\|_{2\to 2}$ to represent the spectral norm of $\mathbf{X}$. We use $B^k(r)$ to denote the radius-r in $\mathbb{R}^k$, i.e., $B^k(r) := \{\mathbf{z} \in \mathbb{R}^k : \|\mathbf{z}\|_2 \le r\}$, and $\mathcal{S}^{n-1}$ represents the unit sphere in $\mathbb{R}^n$, i.e., $\mathcal{S}^{n-1} := \{\mathbf{s} \in \mathbb{R}^n : \|\mathbf{s}\|_2 = 1\}$. We use G to denote a pre-trained L-Lipschitz continuous generative model from $B^k(r)$ to $\mathbb{R}^n$. We focus on the setting where $k \ll n$. For any set $S \subseteq B^k(r)$, we write $G(S) = \{G(\mathbf{z}) : \mathbf{z} \in S\}$. Our goal is to estimate the signal $\mathbf{x} \in \mathrm{Range}(G) = G(B^k(r))$ from realizations of the sensing vector $\mathbf{a}$ and the observation y (generated according to the SIM in (3)). For any two sequences of real values $\{a_j\}$ and $\{b_j\}$, we write $a_j = O(b_j)$ if there exist an absolute constant C_1 and a positive integer j_1 such that for any $j > j_1$, $|a_j| \le C_1 b_j$, $a_j = \Omega(b_j)$ if there exist an absolute constant C_2 and a positive integer j_2 such that for any $j > j_2$, $|a_j| \ge C_2 b_j$. We use the generic notations C and C' to denote large positive constants, and we use c to denote a small positive constant; their values may differ from line to line.

2.2 Setup

First, the following are the standard definitions of a sub-exponential random variable and the associated sub-exponential norm.

Definition 1. *A random variable X is said to be sub-exponential if there exists a positive constant C such that $(\mathbb{E}[|X|^p])^{1/p} \le Cp$ for all $p \ge 1$, or equivalently, if there exists a positive constant C' such that $\mathbb{P}(|X| > u) \le \exp(1 - u/C')$ for all $u \ge 0$. The sub-exponential norm of X is defined as $\|X\|_{\psi_1} := \sup_{p\ge 1} p^{-1} (\mathbb{E}[|X|^p])^{1/p}$.*

We will focus on the following settings except where stated otherwise:

- The observations are independently generated according to the SIM (3), where f is the link function that is *unknown* and possibly random.
- We have an L-Lipschitz continuous generative model $G : B^k(r) \to \mathbb{R}^n$. For convenience, similarly to [46, 47], we assume that the generative model is normalized such that $\mathrm{Range}(G) \subseteq \mathcal{S}^{n-1}$. For a general (unnormalized) generative model, we may essentially consider its normalized version. See, e.g., [47, Remark 1].
- The signal $\mathbf{x}$ is contained in the range of G, i.e., $\mathbf{x} \in \mathrm{Range}(G) \subseteq \mathcal{S}^{n-1}$.
- The sensing vector $\mathbf{a} \in \mathbb{R}^n$ is standard Gaussian, i.e., $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.
- The random variable $y = f(\mathbf{a}^T\mathbf{x})$ is sub-exponential with the sub-exponential norm being denoted by K_y, i.e., $K_y := \|y\|_{\psi_1}$. In addition, we use M_y to denote the expectation of y, i.e., $M_y := \mathbb{E}[y]$.
- We consider MPR and assume that[1]
$$\nu := \mathrm{Cov}\left[y, (\mathbf{a}^T\mathbf{x})^2\right] > 0, \tag{7}$$
or equivalently,
$$\nu := \mathrm{Cov}\left[f(g), g^2\right] > 0, \tag{8}$$
where $g \sim \mathcal{N}(0, 1)$ is a standard normal random variable. Note that the assumption (8) is only with respect to the nonlinear link function f. The condition in (7) is satisfied by PR models described in (2) and (5). It is also satisfied by relevant models such as [89]
$$y = |\mathbf{a}^T\mathbf{x}| + 2\tanh(|\mathbf{a}^T\mathbf{x}|) + \eta, \quad y = 2(\mathbf{a}^T\mathbf{x})^2 + 3\sin(|\mathbf{a}^T\mathbf{x}|) + \eta. \tag{9}$$
See [56, Proposition 3 and Remark 4] for more general examples.

[1] The case that $\nu < 0$ can be similarly handled by considering $-y$.

3 Algorithm

In this section, we describe our two-step algorithm devised for MPR with generative priors. Suppose that we have m i.i.d. realizations of $\mathbf{a}$ and y, namely $\mathbf{a}_1, \ldots, \mathbf{a}_m$ and $y_1, \ldots, y_m$. To estimate the signal $\mathbf{x}$, we consider the following two-step approach:

1. We perform T_1 iterations in the first step. In particular, let

$$\mathbf{V} = \frac{1}{m} \sum_{i=1}^{m} y_i \left(\mathbf{a}_i \mathbf{a}_i^T - \mathbf{I}_n \right). \tag{10}$$

   We perform the projected power method proposed in [47]: For $t = 0, 1, \ldots, T_1 - 1$, let

$$\mathbf{w}^{(t+1)} = \mathcal{P}_G \left(\mathbf{V} \mathbf{w}^{(t)} \right), \tag{11}$$

   where $\mathcal{P}_G(\cdot)$ denotes the projection function onto $\mathrm{Range}(G)$,[2] and we obtain $\mathbf{x}^{(0)} := \mathbf{w}^{(T_1)}$. Similarly to [42, 45], we set the starting point $\mathbf{w}^{(0)}$ as the column of $\frac{1}{m} \sum_{i=1}^{m} y_i \mathbf{a}_i \mathbf{a}_i^T$ (i.e., a shifted version of $\mathbf{V}$) that corresponds to the largest diagonal entry.

2. We perform T_2 iterations in the second step. In particular, let

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i \tag{12}$$

   be the empirical mean of the observations. We perform the following iterative procedure: For $t = 0, 1, 2, \ldots, T_2 - 1$, let

$$\hat{\nu}^{(t)} = \frac{1}{m} \sum_{i=1}^{m} (y_i - \bar{y}) \cdot \left(\mathbf{a}_i^T \mathbf{x}^{(t)} \right)^2 \tag{13}$$

$$\tilde{y}_i^{(t)} = (y_i - \bar{y}) \cdot \left(\mathbf{a}_i^T \mathbf{x}^{(t)} \right), i = 1, 2, \ldots, m. \tag{14}$$

$$\tilde{\mathbf{x}}^{(t+1)} = \mathbf{x}^{(t)} - \frac{\zeta}{m} \cdot \sum_{i=1}^{m} \left(\hat{\nu}^{(t)} \cdot \left(\mathbf{a}_i^T \mathbf{x}^{(t)} \right) - \tilde{y}_i^{(t)} \right) \mathbf{a}_i, \tag{15}$$

$$\mathbf{x}^{(t+1)} = \mathcal{P}_G \left(\tilde{\mathbf{x}}^{(t+1)} \right), \tag{16}$$

   where $\zeta > 0$ is a tuning parameter, and $\hat{\nu}^{(t)}$ can be thought of as an approximation of ν defined in (7). The idea behind calculating $\tilde{y}_i^{(t)}$ is that by comparing (4) and (6), we observe that to transform the MPR model into a conventional SIM, we may use $(y - \mathbb{E}[y])(\mathbf{a}^T \mathbf{x})$ to replace y, see also [56]. Moreover, (15) can be considered as a gradient descent step with respect to a linear measurement model with the scale factor $\hat{\nu}^{(t)}$ and observations $\tilde{y}_i^{(t)}$. Then, in (16), we project the calculated vector $\tilde{\mathbf{x}}^{(t+1)}$ onto $\mathrm{Range}(G)$. Finally, we use $\hat{\mathbf{x}} := \mathbf{x}^{(T_2)}$ to represent the estimated vector obtained after T_2 iterations.

For convenience, we summarize the details in Algorithm 1.

4 Theoretical Results

The following theorem establishes recovery guarantees for the first step of Algorithm 1. The proof is given in the supplementary material. Note that $K_y := \|y\|_{\psi_1}$ (*cf.* Section 2.2) is considered a fixed constant and is omitted in the $O(\cdot)$ notation.

Theorem 1. *Assume that there exists a positive integer t_0 such that $\mathbf{x}^T \mathbf{w}^{(t_0)} \geq c_0$, where c_0 is a sufficiently small positive constant. Suppose that $m = \Omega((k \log(nLr)) \cdot (\log m))$ with a large enough implied constant. Then, we have that with probability $1 - O(1/m)$, it holds for all $t > t_0$ that*

$$\left\| \mathbf{w}^{(t)} - \mathbf{x} \right\|_2 \leq \frac{C K_y}{c_0} \sqrt{\frac{(k \log(nLr)) \cdot (\log m)}{m}} = O\left(\sqrt{\frac{(k \log(nLr)) \cdot (\log m)}{m}} \right). \tag{17}$$

[2] That is, for any $\mathbf{s} \in \mathbb{R}^n$, $\mathcal{P}_G(\mathbf{s}) := G\big(\arg\min_{\mathbf{z} \in B^k(r)} \|G(\mathbf{z}) - \mathbf{s}\|_2 \big)$. Similarly to [70, 36, 37, 63, 47], we implicitly assume the exact projection in our analysis. In practice, approximate methods such as gradient- and GAN-based projections [70, 67] have been shown to work well.

Algorithm 1 A two-step approach for misspecified phase retrieval with generative priors

Input: $\{(\mathbf{a}_i, y_i)\}_{i=1}^m$, step size $\zeta > 0$, number of iterations T_1 for the first step, number of iterations T_2 for the second step, pre-trained generative model G, initial vector $\mathbf{w}^{(0)}$

First step:

for $t = 0, 1, \ldots, T_1 - 1$ **do**
  $\mathbf{w}^{(t+1)} = \mathcal{P}_G\big(\mathbf{V}\mathbf{w}^{(t)}\big)$
end for

Second step:

Let $\mathbf{x}^{(0)} := \mathbf{w}^{(T_1)}$

for $t = 0, 1, \ldots, T_2 - 1$ **do**
  Calculate $\hat{\nu}^{(t)}$, $\tilde{y}_i^{(t)}$, $\tilde{\mathbf{x}}^{(t+1)}$ and $\mathbf{x}^{(t+1)}$ according to (13), (14), (15), and (16), respectively
end for

Output: $\hat{\mathbf{x}} := \mathbf{x}^{(T_2)}$

Since a d-layer feedforward neural network generative model from $B^k(r)$ to $\mathbb{R}^n$ is typically L-Lipschitz continuous with $L = n^{\Theta(d)}$ [5] and we may set $r = n^{\Theta(d)}$, the upper bound in (17) is of order $\sqrt{(k \log L) \cdot (\log m)/m}$. Such a statistical rate is naturally conjectured to be near-optimal according to information-theoretic lower bounds established for MPR with sparse priors [56] and generative model based principal component analysis [47]. Therefore, Theorem 1 reveals that the first step of Algorithm 1 attains the near-optimal statistical rate under appropriate initialization and exact projections. The accurate projection assumption is perhaps the major caveat to Theorem 1. However, it is a standard assumption in relevant works including [70, 36, 37, 63, 47]. In practice, both gradient- and GAN-based projection methods [70, 67] have been shown to be highly effective in approximating the projection step.

Remark 1. *Spectral initialization steps in relevant works on* sparsity *based PR [83, 38, 24, 35, 8, 6] or MPR [56, 89] require the* sub-optimal *sample complexity $O(s^2 \log n)$, where s refers to the number of non-zero entries. In contrary, according to Theorem 1, our spectral initialization step only requires the* near-optimal $O((k \log L) \cdot (\log m))$ *sample complexity (with a linear rather than a quadratic dependence on k). However, we note that such an advantage of our spectral initialization step comes at a price. In particular, we require the initialization condition $\mathbf{x}^T \mathbf{w}^{(t_0)} \geq c_0$, which is not required by spectral initialization steps in the above-mentioned works on sparse PR/MPR.*

Remark 2. *We may assume that the dataset contains all non-negative vectors, which is a natural assumption for image datasets. We may also assume that the range of the generative model is contained in the non-negative orthant. This is a mild assumption since we can easily set the activation function of the last layer of the pre-trained neural network generative model to be a certain non-negative function such as ReLU or sigmoid during pre-training. Therefore, the assumption that $\mathbf{x}^T \mathbf{w}^{(t_0)} \geq c_0$ for a sufficiently small positive constant c_0 is also mild. Similar assumptions have been made in relevant works including [17, 47] where it is not appropriate to assume that $-\mathbf{x}$ is also contained in the structured set (such as a closed convex cone or the range of a deep generative model). As a result, we provide an upper bound on $\|\mathbf{w}^{(t)} - \mathbf{x}\|_2$, instead of $\min\{\|\mathbf{w}^{(t)} - \mathbf{x}\|_2, \|\mathbf{w}^{(t)} + \mathbf{x}\|_2\}$, which is a commonly adopted distance measure in relevant literature on real-valued PR.*

Moreover, although the projected power iterations in the first step of Algorithm 1 can attain the near-optimal statistical rate under appropriate conditions, it is evident in a large body of literature on PR (see, e.g., [54, 12, 9, 90, 89, 56]) that such a spectral method better serves as the initialization step of a subsequent iterative approach. This motivates us to propose the second step of Algorithm 1, and in our numerical experiments, we clearly observe the benefit of the second step. More specifically, compared to simply performing the projected power method, performing both steps of Algorithm 1 leads to significantly better reconstructed images when the total number of iterations is fixed to be the same, namely $T_1 + T_2$.

Next, we present the following theorem, which is proved in the supplementary material. This theorem shows that under appropriate initialization and the assumption of exact projections, the iterative algorithm in the second step of Algorithm 1 converges linearly to a point achieving the near-optimal statistical rate of order $\sqrt{(k \log L) \cdot (\log m)/m}$.

Theorem 2. *Assume that the step size ζ and $\mathbf{x}^{(0)}$, which is the initial vector for the second step of Algorithm 1, satisfy*

$$2 \cdot |1 - \zeta\nu| + 5\zeta\nu \cdot \left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|_2 + \beta_1 = 1 - \beta_2, \tag{18}$$

where both β_1 and β_2 are positive constants. Suppose that $m = \Omega((k\log(nLr)) \cdot (\log m))$ with a large enough implied constant. Then, we have with probability $1 - O(1/m)$, the following event occurs: There exists a positive integer $T_0 = O\big(\log\big(\frac{m}{(k\log(nLr))\cdot(\log m)}\big)\big)$, such that the sequence $\big\{\|\mathbf{x}^{(t)} - \mathbf{x}\|_2\big\}_{t\in[0,T_0]}$ is monotonically decreasing, with the following inequality holds for all $t \le T_0$:

$$\left\|\mathbf{x}^{(t)} - \mathbf{x}\right\|_2 < (1-\beta_2)^t \cdot \left\|\mathbf{x}^{(0)} - \mathbf{x}\right\|_2 + \frac{CK_y}{\beta_2}\sqrt{\frac{(k\log(nLr)) \cdot (\log m)}{m}}. \tag{19}$$

In addition, we have for all $t \ge T_0$ that

$$\left\|\mathbf{x}^{(t)} - \mathbf{x}\right\|_2 \le \frac{CK_y}{\beta_2}\sqrt{\frac{(k\log(nLr)) \cdot (\log m)}{m}} = O\left(\sqrt{\frac{(k\log(nLr)) \cdot (\log m)}{m}}\right). \tag{20}$$

Remark 3. *In our analysis, we need to impose the assumption* (18) *on the step size ζ and initial vector $\mathbf{x}^{(0)}$. This makes the first step of Algorithm 1 necessary since when $\|\mathbf{x}^{(0)} - \mathbf{x}\|_2$ is not small, say $\|\mathbf{x}^{(0)} - \mathbf{x}\|_2 = 1$, the condition* (18) *cannot be satisfied. In comparison, to attain the near-optimal statistical rate $O(\sqrt{(k\log L) \cdot (\log m)/m})$, the initialization condition of the first step of Algorithm 1 is milder, and $\mathbf{x}^T\mathbf{w}^{(t_0)}$ (see Theorem 1) only needs to be lower bounded by a sufficiently small positive constant (thus $\|\mathbf{w}^{(t_0)} - \mathbf{x}\|_2$ can be close to $\sqrt{2}$). However, although the second step of Algorithm 1 requires a more restrictive initialization condition, we observe from our experimental results that it clearly refines the estimate of the first step. Such a phenomenon is also observed in various works related to PR, including [54, 12, 9, 90, 89, 56].*

Remark 4. *The condition in* (18) *requires $|1 - \zeta\nu| < \frac{1}{2}$. This reveals that we should choose ζ such that $\zeta \in \big(\frac{1}{2\nu}, \frac{3}{2\nu}\big)$, and a good choice of ζ is $\zeta = \frac{1}{\nu}$ (for this case, the condition* (18) *reduces to $\|\mathbf{x}^{(0)} - \mathbf{x}\|_2 < \frac{1}{5}$). Since the knowledge of the link function f (and thus the knowledge of ν, which is dependent on f. See* (8)*) cannot be assumed, in our experiments, we use $\hat{\nu}^{(t)}$ to approximate ν. That is, ζ is set to be $\frac{1}{\hat{\nu}^{(t)}}$ in the t-th iteration of the second step of Algorithm 1 (though it is slightly varying instead of being fixed).*

5 Numerical Results

In this section, we demonstrate the empirical performance of our Algorithm 1 (denoted by `MPRG`). We remark that these numerical experiments are proof-of-concept rather than seeking to be comprehensive since our contributions are primarily theoretical. We perform the experiments on the MNIST [43] and CelebA [50] datasets, with the experimental results for the CelebA dataset being presented in the supplementary material.

The MNIST dataset contains $60,000$ images of handwritten digits. The size of each image is 28×28, and thus the dimension of the image vector is $n = 784$. For the MNIST dataset, the generative model G is set to be (the normalized version of) a pre-trained variational autoencoder (VAE) model with the latent dimension being $k = 20$. We make use of the VAE model pre-trained by the authors of [5] directly. For this VAE model, both the encoder and decoder are set to be fully connected neural networks with two hidden layers, and the architecture is $20 - 500 - 500 - 784$. The VAE model is trained by the Adam optimizer with a mini-batch size 100 and a learning rate of 0.001, and is trained from the images in the training set. The projection step $\mathcal{P}_G(\cdot)$ (*cf.* (11)) is approximated by the Adam optimizer with a learning rate of 0.1 and 120 steps.

We report the results on 10 testing images that are selected from the test set. Note that these images are unseen by the pre-trained generative model. We perform 10 random restarts. For reconstructed images, we choose the best among these 10 random restarts to reduce the impact of local minima. We also provide quantitative comparisons with respect to the reconstruction error $\|\hat{\mathbf{x}} - \mathbf{x}\|_2$, where $\mathbf{x}$ is the underlying signal and $\hat{\mathbf{x}}$ refers to the estimated (normalized) vector produced by each of the methods described below. The reconstruction error is averaged over both the 10 testing images and

10 random restarts. All experiments are run using Python 3.6 and TensorFlow 1.5.0, with a NVIDIA GeForce GTX 1080 Ti 11GB GPU.

For Algorithm 1, we set $T_1 = 20$ and $T_2 = 30$. As mentioned in Section 3, the starting point $\mathbf{w}^{(0)}$ is set to be the column of $\frac{1}{m}\sum_{i=1}^{m} y_i \mathbf{a}_i \mathbf{a}_i^T$ (i.e., a shifted version of $\mathbf{V}$ defined in (10)) that corresponds to the largest diagonal entry. In addition, as mentioned in Remark 4, we set the step size ζ as $\zeta = \frac{1}{\hat{\nu}^{(t)}}$ (*cf.* (13)) in the t-th iteration of the second step of Algorithm 1. We compare our Algorithm 1 (denoted by `MPRG`) with the following methods: 1) The method proposed in [89], which is for misspecified phase retrieval with sparse priors and is denoted by `MPRS`. All the involved parameters are set to be the same as those in [89]. 2) Simply performing the first step of Algorithm 1 for $T_1 + T_2 = 50$ iterations. The corresponding method is denoted by `PPower`. The purpose of comparing to `PPower` is to verify the benefit of the second step of Algorithm 1. 3) The Alternating Phase Projected Gradient Descent (denoted by `APPGD`) algorithm proposed in [36]. This algorithm is specifically designed for phase retrieval with magnitude-only measurements (*cf.* (2)) and generative priors, and the corresponding iterative procedure is

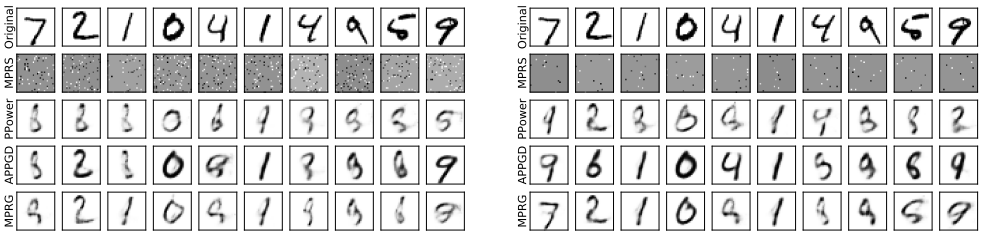
$$\mathbf{x}^{(t+1)} = \mathcal{P}_G\left(\mathbf{x}^{(t)} - \frac{\tau}{m}\sum_{i=1}^{m}\left(\left(\mathbf{a}_i^T\mathbf{x}^{(t)}\right) - y_i \cdot \operatorname{sign}\left(\mathbf{a}_i^T\mathbf{x}^{(t)}\right)\right)\mathbf{a}_i\right), \tag{21}$$

where $\tau > 0$ is the step size. We follow [36] to set $\tau = 0.9$. For a fair comparison, we use the vector produced after $T_1 = 20$ iterations of the first step of Algorithm 1 as the initialization vector of `APPGD`, and then we run `APPGD` for $T_2 = 30$ iterations.

We first consider the noisy magnitude-only measurement model

$$y_i = |\mathbf{a}_i^T\mathbf{x}| + \eta_i, \quad i = 1, 2, \ldots, m, \tag{22}$$

where η_i are i.i.d. realizations of an $\mathcal{N}(0, \sigma^2)$ random variable. For such a measurement model, the corresponding random nonlinear link function f is $f(x) = |x| + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$. The numerical results for (22) are demonstrated in Figures 1 and 2. From these figures, we observe that the sparsity based method `MPRS` attains poor reconstructions, and the results of `PPower` are not desirable. In contrast, our method `MPRG` leads to mostly accurate reconstructions that are competitive compared to those of `APPGD`, even if we do not make use of the knowledge of the link function f and `MPRG` is not specifically designed for the magnitude-only measurement model (22).



(a) $m = 200$ and $\sigma = 0$ (b) $m = 400$ and $\sigma = 0.1$

Figure 1: Examples of reconstructed images of the MNIST dataset for the measurement model (22).

Next, we consider the following two measurement models:

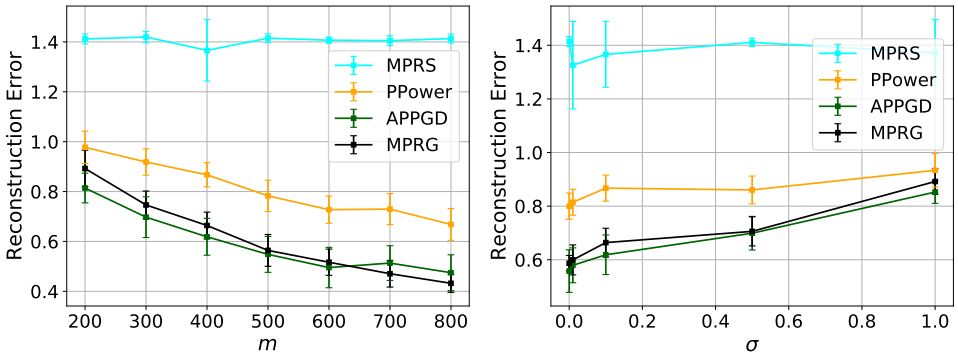
$$y_i = |\mathbf{a}_i^T\mathbf{x}| + 2\tanh(|\mathbf{a}_i^T\mathbf{x}|) + \eta_i, \tag{23}$$

$$y_i = 2(\mathbf{a}_i^T\mathbf{x})^2 + 3\sin(|\mathbf{a}_i^T\mathbf{x}|) + \eta_i, \tag{24}$$

where again η_i are i.i.d. realizations of an $\mathcal{N}(0, \sigma^2)$ random variable. For both models in (23) and (24), the corresponding link functions satisfy the condition (8) for MPR [89]. The numerical results are presented in Figures 3 and 4. We observe from these figures that for the measurement models (23) and (24), our method `MPRG` achieves the best reconstructions. In particular, it significantly outperforms `APPGD` in terms of recovery quality and/or reconstruction error.

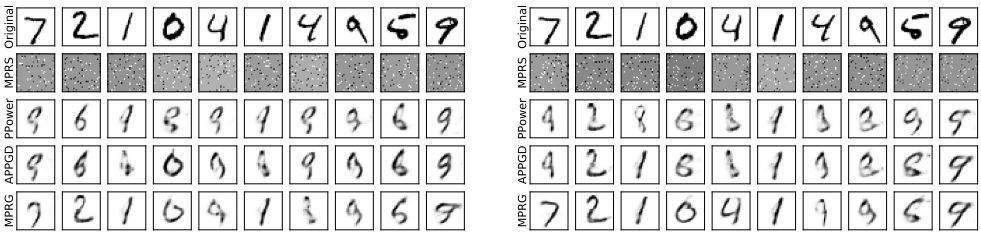
6 Conclusion and Future Work

We have proposed a two-step approach for phase retrieval under model misspecification and generative priors. We show that under suitable conditions, both steps of our approach obtain estimated vectors



(a) Fixing $\sigma = 0.1$ and varying m (b) Fixing $m = 400$ and varying σ

Figure 2: Quantitative comparisons of the performance for the measurement model (22).

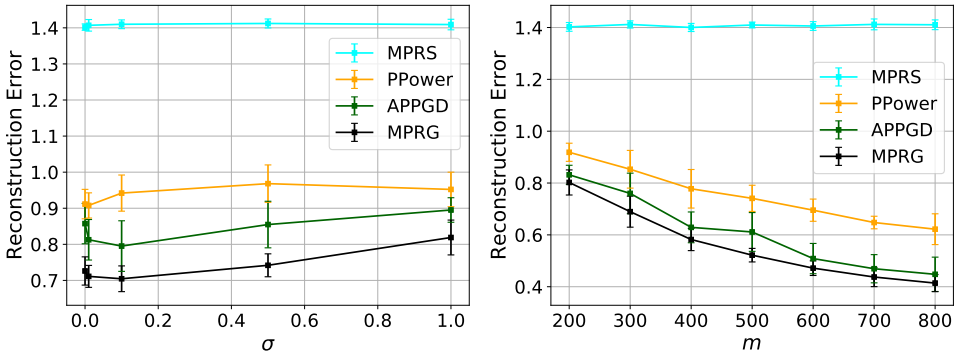


(a) (23) with $m = 200$ and $\sigma = 0.01$ (b) (24) with $m = 300$ and $\sigma = 0.5$

Figure 3: Examples of reconstructed MNIST images for the models in (23) and (24).

that achieve the near-optimal statistical rate of order $\sqrt{(k \log L) \cdot (\log m)/m}$, where k is the latent dimension and L is the Lipschitz constant of the generative model respectively, and m refers to the number of samples.

In the statistical rate $\sqrt{(k \log L) \cdot (\log m)/m}$ and sample complexity $O((k \log L) \cdot (\log m))$, there is a $\log m$ term. This logarithmic term plays a minor role, but it is better to get rid of it. In addition, we assume accurate projections in our analysis and use a gradient-based method to approximate the projection step $\mathcal{P}_G(\cdot)$ in our experiments. Although the exact projection assumption is commonly made in relevant works [70, 36, 37, 63, 47], it is a very interesting future research direction to design provably-guaranteed efficient methods for the projection step.



(a) (23) with fixed $m = 200$ (b) (24) with fixed $\sigma = 0.5$

Figure 4: Quantitative comparisons of the performance for the measurement models in (23) and (24).

References

[1] E. Arias-Castro, E. J. Candes, and M. A. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 472–481, 2012.

[2] M. Asim, M. Daniels, O. Leong, A. Ahmed, and P. Hand, "Invertible generative models for inverse problems: Mitigating representation error and dataset bias," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2020, pp. 399–409.

[3] B. Aubin, B. Loureiro, A. Baker, F. Krzakala, and L. Zdeborová, "Exact asymptotics for phase retrieval and compressed sensing with random generative priors," in *Math. Sci. Mach. Learn. (MSML)*. PMLR, 2020, pp. 55–73.

[4] 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.

[5] 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.

[6] J.-F. Cai, Y. Jiao, X. Lu, and J. You, "Sample-efficient sparse phase retrieval via stochastic alternating minimization," *https://arxiv.org/2112.07919*, 2021.

[7] 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.

[8] J. Cai, M. Huang, D. Li, and Y. Wang, "Solving phase retrieval with random initial guess is nearly as good as by spectral initialization," *https://arxiv.org/abs/2101.03540*, 2021.

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

[10] E. J. Candes and M. A. Davenport, "How well can we estimate a sparse vector?" *Appl. Comp. Harm. Analysis*, vol. 34, no. 2, pp. 317–323, 2013.

[11] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Rev.*, vol. 57, no. 2, pp. 225–251, 2015.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] L. Cheng, P. Zeng, and Y. Zhu, "BS-SIM: An effective variable selection method for high-dimensional single index model," *Electron. J. Stat.*, vol. 11, no. 2, pp. 3522–3548, 2017.

[17] Y. Deshpande, A. Montanari, and E. Richard, "Cone-constrained principal component analysis," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2014, pp. 2717–2725.

[18] M. Dhar, A. Grover, and S. Ermon, "Modeling sparse deviations for compressed sensing using generative models," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2018, pp. 1214–1223.

[19] H. Eftekhari, M. Banerjee, and Y. Ritov, "Inference in high-dimensional single-index models under symmetric designs," *J. Mach. Learn. Res.*, vol. 22, pp. 27–1, 2021.

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

[21] J. C. Foster, J. M. Taylor, and B. Nan, "Variable selection in monotone single-index models via the adaptive Lasso," *Stat. Med.*, vol. 32, no. 22, pp. 3944–3954, 2013.

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

[23] R. Ganti, N. Rao, R. M. Willett, and R. Nowak, "Learning single index models in high dimensions," *https://arxiv.org/1506.08910*, 2015.

[24] B. Gao, X. Sun, Y. Wang, and Z. Xu, "Perturbed amplitude flow for phase retrieval," *IEEE Trans. Sig. Proc.*, vol. 68, pp. 5427–5440, 2020.

[25] M. Genzel, "High-dimensional estimation of structured signals from non-linear observations with general convex loss functions," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1601–1619, 2016.

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

[27] L. Goldstein, S. Minsker, and X. Wei, "Structured signal recovery from non-linear and heavy-tailed measurements," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5513–5530, 2018.

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

[29] A. K. Han, "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *J. Econom.*, vol. 35, no. 2-3, pp. 303–316, 1987.

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

[31] P. Hand and V. Voroninski, "Global guarantees for enforcing deep generative priors by empirical risk," in *Conf. Learn. Theory (COLT)*. PMLR, 2018, pp. 970–978.

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

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

[34] J. L. Horowitz, *Semiparametric and nonparametric methods in econometrics*. Springer, 2009, vol. 12.

[35] 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.

[36] 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.

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

[38] 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.

[39] A. Jalal, S. Karmalkar, A. Dimakis, and E. Price, "Instance-optimal compressed sensing via posterior sampling," in *Int. Conf. Mach. Learn. (ICML)*, 2021.

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

[41] A. Kamath, S. Karmalkar, and E. Price, "On the power of compressed sensing with generative models," in *Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 5101–5109.

[42] 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.

[43] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[44] K.-C. Li and N. Duan, "Regression analysis under link violation," *Ann. Stat.*, pp. 1009–1052, 1989.

[45] Z. Liu, S. Ghosh, and J. Scarlett, "Towards sample-optimal compressive phase retrieval with sparse and generative priors," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2021.

[46] 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.

[47] Z. Liu, J. Liu, S. Ghosh, J. Han, and J. Scarlett, "Generative principal component analysis," in *Int. Conf. Learn. Repr. (ICLR)*, 2022.

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

[49] 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.

[50] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3730–3738.

[51] S. Luo and S. Ghosal, "Forward selection and estimation in high dimensional single index models," *Stat. Methodol.*, vol. 33, pp. 172–179, 2016.

[52] P. McCullagh and J. A. Nelder, *Generalized linear models*. Routledge, 2019.

[53] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recognit. (CVPR)*, 2020, pp. 2437–2445.

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

[55] M. Neykov, J. S. Liu, and T. Cai, "L_1-regularized least squares for support recovery of high dimensional single index models with Gaussian designs," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2976–3012, 2016.

[56] M. Neykov, Z. Wang, and H. Liu, "Agnostic estimation for misspecified phase retrieval models," *J. Mach. Learn. Res.*, vol. 21, pp. 1–39, 2020.

[57] T. V. Nguyen, G. Jagatap, and C. Hegde, "Provable compressed sensing with generative priors via Langevin dynamics," *https://arxiv.org/2102.12643*, 2021.

[58] 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.

[59] 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.

[60] S. Oymak and M. Soltanolkotabi, "Fast and reliable parameter estimation from nonlinear observations," *SIAM J. Optim.*, vol. 27, no. 4, pp. 2276–2300, 2017.

[61] A. Pananjady and D. P. Foster, "Single-index models in the high signal regime," *IEEE Trans. Inf. Theory*, vol. 67, no. 6, pp. 4092–4124, 2021.

[62] E. J. R. Pauwels, A. Beck, Y. C. Eldar, and S. Sabach, "On fienup methods for sparse phase retrieval," *IEEE Trans. Sig. Proc.*, vol. 66, no. 4, pp. 982–991, 2017.

[63] P. Peng, S. Jalali, and X. Yuan, "Solving inverse problems via auto-encoders," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 312–323, 2020.

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

[65] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *Inf. Inference*, vol. 6, no. 1, pp. 1–40, 2017.

[66] P. Radchenko, "High dimensional single index models," *J. Multivar. Anal.*, vol. 139, pp. 266–282, 2015.

[67] A. Raj, Y. Li, and Y. Bresler, "GAN-based projector for faster recovery with convergence guarantees in linear inverse problems," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019.

[68] J. Rick Chang, C.-L. Li, B. Poczos, B. Vijaya Kumar, and A. C. Sankaranarayanan, "One network to solve them all–solving linear inverse problems using deep projection models," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 5888–5897.

[69] J. Scarlett and V. Cevher, "Limits on support recovery with probabilistic models: An information-theoretic framework," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 593–620, 2016.

[70] V. Shah and C. Hegde, "Solving linear inverse problems using GAN priors: An algorithm with provable guarantees," in *Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*. IEEE, 2018, pp. 4609–4613.

[71] 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.

[72] R. P. Sherman, "The limiting distribution of the maximum rank correlation estimator," *Econometrica*, pp. 123–137, 1993.

[73] 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.

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

[75] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Lasso with non-linear measurements is equivalent to one with linear measurements," in *Conf. Neur. Inf. Proc. Sys. (NeurIPS)*, 2015, pp. 3420–3428.

[76] 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.

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

[78] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[79] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.

[80] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[81] 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.

[82] 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.

[83] 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.

[84] X. Wei, "Structured recovery with heavy-tailed measurements: A thresholding procedure and optimal rates," *https://arxiv.org/abs/1804.05959*, 2018.

[85] 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.

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

[87] Y. Wu, M. Rosca, and T. Lillicrap, "Deep compressed sensing," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2019, pp. 6850–6860.

[88] Z. Yang, K. Balasubramanian, and H. Liu, "High-dimensional non-Gaussian single index models via thresholded score function estimation," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2017, pp. 3851–3860.

[89] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov, "Misspecified nonconvex statistical optimization for sparse phase retrieval," *Math. Program.*, vol. 176, no. 1, pp. 545–571, 2019.

[90] 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.

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 are relevant to our future work, 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? [Yes]

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 code is 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]