Projected Gradient Descent Algorithms for Solving Nonlinear Inverse Problems with Generative Priors

Zhaoqiang Liu\(^1\) and Jun Han\(^2\)

\(^1\)National University of Singapore
\(^2\)PCG, Tencent
dcsлизha@nus.edu.sg, junhanjh@tencent.com

Abstract

In this paper, we propose projected gradient descent (PGD) algorithms for signal estimation from noisy nonlinear measurements. We assume that the unknown signal lies near the range of a Lipschitz continuous generative model with bounded inputs. In particular, we consider two cases when the nonlinear link function is either unknown or known. For unknown nonlinearity, we make the assumption of sub-Gaussian observations and propose a linear least-squares estimator. We show that when there is no representation error, the sensing vectors are Gaussian, and the number of samples is sufficiently large, with high probability, a PGD algorithm converges linearly to a point achieving the optimal statistical rate using arbitrary initialization. For known nonlinearity, we assume monotonicity, and make much weaker assumptions on the sensing vectors and allow for representation error. We propose a nonlinear least-squares estimator that is guaranteed to enjoy an optimal statistical rate. A corresponding PGD algorithm is provided and is shown to also converge linearly to the estimator using arbitrary initialization. In addition, we present experimental results on image datasets to demonstrate the performance of our PGD algorithms.

1 Introduction

Over the past two decades, the theoretical and algorithmic aspects of high-dimensional linear inverse problems have been studied extensively. The standard compressed sensing (CS) problem, which models low-dimensional structure via the sparsity assumption, is particularly well-understood [Foucart and Rauhut, 2013].

Despite the popularity of linear CS, in many real-world applications, nonlinearities may arise naturally, and it is more desirable to adopt nonlinear measurement models. For example, the semi-parametric single index model (SIM), which is formulated below, is a popular nonlinear measurement model that has long been studied [Han, 1987]:

\[
y_i = f_i(a_i^T x^*), \quad i = 1, 2, \ldots, n,
\]

where \(x^* \in \mathbb{R}^p\) is an unknown signal that is close to some structured set \(K\), \(a_i \in \mathbb{R}^p\) are the sensing vectors, and \(f_i : \mathbb{R} \to \mathbb{R}\) are i.i.d. realizations of an unknown (possibly random) function \(f_i\). In general, \(f_i\) plays the role of a nonlinearity, and it is called a link function. The goal is to estimate \(x^*\) despite this unknown link function. Note that since the norm of \(x^*\) can be absorbed into the unknown \(f_i\), the signal \(x^*\) is typically assumed to have unit \(\ell_2\)-norm.

In addition, inspired by the tremendous success of deep generative models in numerous real-world applications, recently, for the CS problem, it has been of interest to replace the sparsity assumption with the generative model assumption. More specifically, instead of being assumed to be sparse, the signal is assumed to lie near the range of a generative model, typically corresponding to a deep neural network [Bora et al., 2017]. Along with several theoretical developments, the authors of [Bora et al., 2017] perform extensive numerical experiments on image datasets to demonstrate that for a given accuracy, generative priors can reduce the required number of measurements by a factor of 5 to 10. There are a variety of follow-up works of [Bora et al., 2017], including [Heckel and Hand, 2019; Ongie et al., 2020; Jalal et al., 2021], among others.

In this paper, following the developments in both sparsity-based nonlinear inverse problems and inverse problems with generative priors, we provide theoretical guarantees for projected gradient descent (PGD) algorithms devised for nonlinear inverse problems using generative models.

1.1 Related Work

The most relevant existing works can roughly be divided into (i) nonlinear inverse problems without generative priors, and (ii) inverse problems with generative priors.

**Nonlinear inverse problems without generative priors:**

The SIM has long been studied in the low-dimensional setting where \(p \ll n\), based on various assumptions on the sensing vector or link function. For example, the maximum rank correlation estimator has been proposed in [Han, 1987] under the assumption of a monotonic link function. In recent years, the SIM has also been studied in [Plan and Vershynin, 2016; Genzel, 2016; Plan et al., 2017; Oymak and Soltanolkotabi, 2017] in the high-dimensional setting where an accurate estimate can be obtained when \(n \ll p\), with the sensing vectors being assumed to be Gaussian. In particular, the au-
thors of [Plan and Vershynin, 2016] show that the general-
ized Lasso approach works for high-dimensional SIM under
the assumption that the set of structured signal \( \mathcal{K} \) is convex,
which is in general not satisfied for the range of a generative
model with the Lipschitz continuity.

Nonetheless, the generality of the unknown link function
in SIM comes at a price. Specifically, as mentioned above, it
is necessary for the works studying SIM to assume the dis-
tribution of the sensing vector to be Gaussian or symmetric
elliptical, and for nonlinear signal estimation problems with
general sensing vectors, in order to achieve consistent estima-
tion, knowledge of the link function is required [Zhang et al.,
2018]. In addition, when the link function is unknown, since
the norm of the signal may be absorbed into this link function,
there is an identifiability issue and we are only able to esti-
mate the direction of the signal. In practice, this can be unsat-
sfactory and may lead to large estimation errors. Moreover,
for an unknown nonlinearity, it remains an open problem to
handle signals with representation error [Plan and Vershynin,
2016], i.e., the signal (up to a fixed scale factor) is not exactly
contained in \( \mathcal{K} \).

Based on these issues of SIM and some applications in
machine learning such as the activation functions of deep
neural networks [Yang et al., 2016], nonlinear measure-
ment models with known and monotonic link functions\(^1\) have
been studied in [Yang et al., 2016; Zhang et al., 2018;
Soltani and Hegde, 2017]. In [Yang et al., 2016], an \( \ell_1 \)-
regularized nonlinear least-squares estimator is proposed, and
an iterative soft thresholding algorithm is provided to effi-
ciently approximate this estimator. The authors of [Zhang
et al., 2018] propose an iterative hard thresholding (IHT)
algorithm to minimize a nonlinear least-squares loss function
subject to a combinatorial constraint. In the work [Soltani
and Hegde, 2017], the demixing problem is formulated as
minimizing a special (not the typical least-squares) loss func-
tion under a combinatorial constraint, and a corresponding
IHT algorithm is designed to approximately find a mini-
mizer. All the algorithms proposed in [Yang et al., 2016;
Zhang et al., 2018; Soltani and Hegde, 2017] can be regarded
as special cases of the PGD algorithm.

Inverse problems with generative priors: Bora et al.
show that when the generative model is \( L \)-Lipschitz contin-
uous with bounded \( k \)-dimensional inputs, roughly
\( O(k \log L) \) random Gaussian linear measurements are sufficient to attain
accurate estimates [Bora et al., 2017]. Their analysis is based
on minimizing a linear least-squares loss function, and the
objective function is minimized directly over the latent variable
in \( \mathbb{R}^k \) using gradient descent. A PGD algorithm in the ambi-
ent space in \( \mathbb{R}^k \) has been proposed in [Shah and Hegde, 2018;
Peng et al., 2020] for noiseless and noisy Gaussian linear
measurements respectively. It has been empirically demonstr-
ated that this PGD algorithm leads to superior reconstruc-
tion performance over the algorithm used in [Bora et al.,
2017]. Various nonlinear measurement models with known
nonlinearity have also been studied for generative priors.
Specifically, near-optimal sample complexity bounds for 1-
bit measurement models have been presented in [Qiu et al.,
2020; Liu et al., 2020]. Furthermore, the works [Wei et al.,
2019; Liu and Scarlett, 2020a] have provided near-optimal
non-uniform recovery guarantees for nonlinear compressed
sensing with an unknown nonlinearity. More specifically,
the authors of [Wei et al., 2019] assume that the link func-
tion is differentiable and propose estimators via score func-
tions based on the first and second order Steins identity. The
differentiability assumption fails to hold for 1-bit and other
quantized measurement models. To take such measurement
models into consideration, the work [Liu and Scarlett, 2020a]
instead makes the assumption that the (uncorrupted) obser-
vations are sub-Gaussian, and proposes to use a simple lin-
ear least-squares estimator despite the unknown nonlinearity.
While obtaining these estimators is practically hard due to
the typical non-convexity of the range of a generative model, both
works are primarily theoretical, and no practical algorithm is
provided to approximately find the estimators.

1.2 Contributions
Throughout this paper, we make the assumption that the gen-
erative model is \( L \)-Lipschitz continuous with bounded \( k \)-
dimensional inputs (see, e.g., [Bora et al., 2017]). The main
contributions of this paper are as follows:

- For the scenario where \( \mathcal{K} \) is convex, we
  assume that the sensing vector is Gaussian and the signal
  is exactly contained in the range of the generative model,
  and propose a PGD algorithm for a linear least-squares
  estimator. We show that roughly \( O(k \log L) \) samples
  suffice to ensure that this PGD algorithm converges lin-
  early and yields an estimator with optimal statistical
  rate, which is roughly of order \( \sqrt{k \log L/n} \). While this
  PGD algorithm is identical to the PGD algorithm for
  solving linear inverse problems using generative mod-
  els as proposed in [Shah and Hegde, 2018; Peng et al.,
  2020], the corresponding analysis is significantly differ-
  ent since we consider the SIM with an unknown non-
  linear function \( f \), instead of the simple linear measure-
  ment model. Moreover, unlike [Shah and Hegde, 2018;
Peng et al., 2020], we have provided a neat theoretical
guarantee for choosing the step size.

- For the scenario where there is no prior knowledge,
  we make much weaker assumptions for sensing vectors
  and allow for representation error, i.e., the signal do not
  reside in the range of the generative model, and proposes
  a nonlinear least-squares estimator. We prove that the
  estimator enjoys optimal statistical rate, and show that a
  corresponding PGD algorithm converges linearly to this
  estimator. To the best of our knowledge, the correspond-
  ing PGD algorithm (cf. (18)) is novel.

- We perform various numerical experiments on image
  and audio datasets to back up our theoretical results.

Remark 1. Generative model based phase retrieval has been
studied in [Hand et al., 2018; Jagatap and Hegde, 2019;
Hyder et al., 2019; Liu et al., 2021b]. However, for the sce-
nario of SIM with unknown nonlinearity, we follow the settings in [Liu and Scarlett, 2020a], and as mentioned therein, phase retrieval is beyond the scope of this setup. For the case of a known link function, phase retrieval is also not applicable since its corresponding nonlinear functions are not monotonic. Moreover, it is typically unavoidable for phase retrieval with generative priors to require the strong assumption about the existence of a good initial vector, whereas for the nonlinear function (whether it is unknown or known) and the corresponding PGD algorithm considered in our work, the initial vector can be arbitrary.

1.3 Notation

We use upper and lower case boldface letters to denote matrices and vectors respectively. For any positive integer $N$, we write $[N] = \{1, 2, \ldots, N\}$ and we use $I_N$ to represent an identity matrix in $\mathbb{R}^{N \times N}$. A generative model is a function $G : \mathcal{D} \to \mathbb{R}^p$, with latent dimension $k$, ambient dimension $p$, and input domain $\mathcal{D} \subseteq \mathbb{R}^k$. We focus on the setting where $k \ll p$. For a set $S \subseteq \mathbb{R}^k$ and a generative model $G : \mathbb{R}^k \to \mathbb{R}^p$, we write $G(S) = \{G(z) : z \in S\}$. We use $\|X\|_{\mathcal{D} \to \mathcal{X}}$ to denote the spectral norm of a matrix $X$. We define the $\ell_q$-ball $B^{(q)}_\mathcal{K}(r) := \{z \in \mathbb{R}^k : \|z\|_q \leq r\}$ for $q \in [0, +\infty]$. The symbols $c, C$ are absolute constants whose values may be different per appearance.

2 Preliminaries

We present the definition for a sub-Gaussian random variable.

**Definition 1.** A random variable $X$ is said to be sub-Gaussian if there exists a positive constant $C$ such that $(\mathbb{E} \|X|^{q}\|)^{1/q} \leq C \sqrt{q}$ for all $q \geq 1$. The sub-Gaussian norm of a sub-Gaussian random variable $X$ is defined as $\|X\|_{\mathcal{D} \to \mathcal{X}} := \sup_{q \geq 1} q^{-1/2}(\mathbb{E} \|X|^{q}\|)^{1/q}$.

Throughout this paper, we make the assumption that the generative model $G : B^{(2)}_\mathcal{K}(r) \to \mathbb{R}^p$ is $L$-Lipschitz continuous, and we fix the structured set $\mathcal{K}$ to be the range of $G$, i.e., $\mathcal{K} := G(B^{(2)}_\mathcal{K}(r))$.

In the following, we state the definition of the Two-sided Set-Restricted Eigenvalue Condition (TS-REC), which is adapted from the S-REC proposed in [Bora et al., 2017].

**Definition 2.** Let $S \subseteq \mathbb{R}^p$. For parameters $\epsilon \in (0, 1)$, $\delta \geq 0$, a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ is said to satisfy the TS-REC($\mathcal{S}, \epsilon, \delta$) if, for every $x_1, x_2 \in S$, it holds that

$$(1 - \epsilon)\|x_1 - x_2\|_2 - \delta \leq \left\|\mathbf{A}(x_1 - x_2)\right\|_2 \leq (1 + \epsilon)\|x_1 - x_2\|_2 + \delta.$$  

(2)

Suppose that $\mathbf{B} \in \mathbb{R}^{n \times p}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. We have the following lemma, which says that $\frac{1}{\sqrt{n}} \mathbf{B}$ satisfies TS-REC for the set $\mathcal{K} = G(B^{(2)}_\mathcal{K}(r))$ with high probability.

**Lemma 1.** (Adapted from [Bora et al., 2017, Lemma 4.1]) For $\epsilon \in (0, 1)$ and $\delta > 0$, if $n = \Theta \left(\frac{k}{\epsilon^2} \log \frac{L}{\delta}\right)^2$, then a random matrix $\frac{1}{\sqrt{n}} \mathbf{B} \in \mathbb{R}^{n \times p}$ with $\mathcal{B}_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ satisfies the TS-REC($\mathcal{K}, \epsilon, \delta$) with probability $1 - e^{-\Omega(n^2)}$.

3 Here and in subsequent statements of lemmas and theorems, the implied constant is assumed to be sufficiently large.

3 $\mu$ is important for analyzing the recovery performance, but note that the knowledge of $\mu$ cannot be assumed since $f$ is unknown.

3 PGD for Unknown Nonlinearity

In this section, we provide theoretical guarantees for a PGD algorithm in the case that the nonlinear link function $f$ is unknown. For this case, we follow [Liu and Scarlett, 2020a] and make the assumptions:

- Let $x^* \in \mathbb{R}^p$ be the signal to estimate. We assume that $\mu x^*$ is contained in the set $\mathcal{K} = G(B^{(2)}_\mathcal{K}(r))$, where

$$\mu := \mathbb{E}_{g \sim \mathcal{N}(0, 1)}[f(g)g]$$

(3)

is a fixed parameter depending solely on $f$. Since the norm of $x^*$ may be absorbed into the unknown $f$, for simplicity of presentation, we assume that $\|x^*\|_2 = 1$.

- $\mathbf{a}_i$ are i.i.d. realizations of a random vector $\mathbf{a} \sim \mathcal{N}(0, \mathbf{I}_p)$, with $\mathbf{a}$ being independent of $f$. We write the sensing matrix as $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^T \in \mathbb{R}^{n \times p}$.

- We assume the SIM for the (unknown) uncorrupted measurements $y_1, y_2, \ldots, y_n$ as in (1).

- Similarly to [Plan et al., 2017; Liu and Scarlett, 2020a], the random variable $y := f(\mathbf{a}^T x^*)$ is assumed to be sub-Gaussian with sub-Gaussian norm $\psi$, i.e.,

$$\psi := \|f(\mathbf{a}^T x^*)\|_{\ell_2} = \|f(g)\|_{\ell_2},$$

(4)

where $g \sim \mathcal{N}(0, 1)$. Such an assumption will be satisfied, e.g., when $f$ does not grow faster than linearly, i.e., for any $x \in \mathbb{R}$, $|f(x)| \leq a + b|x|$ for some $a$ and $b$. Hence, various noisy 1-bit measurement models and non-binary quantization schemes satisfy this assumption [Liu and Scarlett, 2020a].

- In addition to possible random noise in $f$, we allow for adversarial noise that may depend on $\mathbf{a}$. In particular, instead of observing $y$ directly, we only assume access to (corrupted) measurements $\hat{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$ satisfying

$$\frac{1}{\sqrt{n}} \|\hat{y} - y\|_2 \leq \tau$$

(5)

for some $\tau \geq 0$, where $y = [y_1, y_2, \ldots, y_n]^T \in \mathbb{R}^n$.

- To derive an estimate of the signal $x^*$ (up to constant scaling), we minimize the linear $\ell_2$ loss over $\mathcal{K}$:

$$\minimize \mathcal{L}_1(\mathbf{x}) := \frac{1}{2n}\|\hat{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{K}.$$  

(6)

The above optimization problem is referred to as the generalized Lasso or $\mathcal{K}$-Lasso. The idea behind using the $\mathcal{K}$-Lasso to derive an accurate estimate even for nonlinear observations is that the nonlinearity is regarded as noise and the nonlinear observation model can be converted into a scaled linear model with unconventional noise [Plan and Vershynin, 2016].

The authors of [Liu and Scarlett, 2020a] provide recovery guarantees with respect to globally optimal solutions of (6), but they have not designed practical algorithms to find an optimal solution. Solving (6) may be practically difficult since
Algorithm 1 A PGD algorithm for approximately solving (6) (PGD–Glasso)

**Input:** \( A, \tilde{y}, \nu > 0, \) number of iterations \( T, \) generative model \( G, \) arbitrary initial vector \( \mathbf{x}^{(0)} \)

**Procedure:** Iterate as in (8) for \( t = 0, \ldots, T-1; \) return \( \mathbf{x}^{(T)} \)

in general, \( K = G(B^R_2(r)) \) is not a convex set. In this section, we consider using the following iterative procedure to approximately solve (6):

\[
\mathbf{x}^{(t+1)} = \mathcal{P}_K \left( \mathbf{x}^{(t)} - \nu \cdot \nabla L_1 \left( \mathbf{x}^{(t)} \right) \right)
\]

(7)

\[
= \mathcal{P}_K \left( \mathbf{x}^{(t)} - \frac{L}{n} \cdot A^T (Ax^{(t)} - \tilde{y}) \right),
\]

(8)

where \( \mathcal{P}_K (\cdot) \) is the projection function onto \( K \) and \( \nu > 0 \) is a tuning parameter. For convenience, the corresponding algorithm is described in Algorithm 1.

**Remark 2.** We will implicitly assume the exact projection in analysis. Our proof technique does not require \( \mathcal{P}_K \) to be unique, but only requires it to be a retraction onto the manifold of the generative prior. The exact projection assumption is also made in relevant works including [Hyder et al., 2019; Liu et al., 2022; Peng et al., 2020; Shah and Hegde, 2018].

In practice approximate methods might be needed, and both gradient-based projection [Shah and Hegde, 2018] and GAN-based projection [Raj et al., 2019] have been shown to be highly effective. Compared to exact projection, in practice global optima of optimization problems like (6) are typically much more difficult to approximate, and projection-based methods may serve as powerful tools for approximating the global optima. For example, for the simple linear measurement model, it has been numerically verified that performing gradient descent over the latent variable (without projection) leads to inferior performance and cannot approximate the global optima of (6) well, whereas a projection-based gradient descent method gives better reconstruction [Shah and Hegde, 2018].

Algorithm 1 is identical to the PGD algorithm for solving linear inverse problems using generative models as proposed in [Shah and Hegde, 2018; Peng et al., 2020]. However, the corresponding analysis is significantly different, since we consider the SIM with an unknown nonlinear function \( f \) instead of the simple linear measurement model. In particular, we have the following theorem showing that if \( 2\mu_1 < 1, \) Algorithm 1 converges linearly and achieves optimal statistical rate, which is roughly of order \( \sqrt{k \log L/n}. \) The proof of Theorem 1 is provided in the supplementary material.

**Theorem 1.** Recall that \( \mu \) and \( \psi \) are defined in (3) and (4) respectively. For any \( \epsilon \in (0, 1), \) letting

\[
\mu_1 := \max \{1 - \nu (1 - \epsilon), \nu (1 + \epsilon) - 1\}.
\]

(9)

For any \( \delta > 0 \) satisfying \( Lr = \Omega(\delta p), \) if \( n = \Omega \left( \frac{k \log L}{\delta^2} \right) \) and \( 2\mu_1 < 1 \) with \( 1 - 2\mu_1 = \Theta(1), \) then for any \( t \in \mathbb{N}, \) with probability \( 1 - e^{\Omega(\epsilon^2 n)} \), we have

\[
\|\mathbf{x}^{(t)} - \mu \mathbf{x}^\star\|_2 \leq \left( 2\mu_1 \right)^t \cdot \|\mathbf{x}^{(0)} - \mu \mathbf{x}^\star\|_2 \\
+ C \left( \psi \sqrt{\frac{k \log L}{n}} + \delta + \tau \right). \]  

(10)

To ensure that \( 2\mu_1 < 1, \) if \( \epsilon \) is chosen to be a sufficiently small positive constant, we should select the parameter \( \nu \) from the interval \( (0.5, 1.5), \) and a good choice of \( \nu \) is \( \nu = 1. \) In addition, a \( d \)-layer neural network generative model typically has Lipschitz constant \( L = p^{\Theta(d)} [Bora et al., 2017], \) and thus we may set \( r = p^{\Theta(d)} \) and \( \delta = \frac{1}{p^{\Theta(d)}} \) without affecting the scaling of the term \( \log \frac{L}{r} \) (and the assumption \( Lr = \Omega(\delta p) \) is certainly satisfied for fixed \( (r, \delta) \)).

Hence, if there is no adversarial noise, i.e., \( \tau = 0, \) and \( \psi \) is a fixed constant, we see that after a sufficient number of iterations, Algorithm 1 will return a point \( \mathbf{x}^{(T)} \) satisfying

\[
\|\mathbf{x}^{(T)} - \mu \mathbf{x}^\star\|_2 = O \left( \sqrt{\frac{k \log L}{n}} \right). \]

By the analysis of sample complexity lower bounds for noisy linear CS using generative models [Liu and Scarlett, 2020b; Kamath et al., 2020], this statistical rate is optimal and cannot be improved without extra assumptions.

4 PGD for Known Nonlinearity

In this section, we provide theoretical guarantees for the case when the nonlinear link function \( f \) is known. Throughout this section, we make the following assumptions:

- Unlike in the case of unknown nonlinearity, we now allow for representation error and assume that the signal \( \mathbf{x}^\star \) lies near (but does not need to be exactly contained in) \( K = G(B^R_2(\nu)) \). Note that since we have precise knowledge of \( f, \) we do not need to make any assumption on the norm of \( \mathbf{x}^\star \).

- The sensing matrix \( A \in \mathbb{R}^{n \times p} \) satisfies the following two assumptions with high probability:
  1. Johnson-Lindenstrauss embeddings (JLE): For any \( \epsilon \in (0, 1) \) and any finite set \( E \subseteq \mathbb{R}^p \) satisfying
     \[
     n = \Omega \left( \frac{1}{\epsilon^2 \cdot \log E} \right)
     \]
     for some absolute constants \( c_1, c_2, \) we have for all \( x \in E \) that
     \[
     (1 - \epsilon)\|x\|_2^2 \leq \left\| \frac{1}{\sqrt{n}} A x \right\|_2^2 \leq (1 + \epsilon)\|x\|_2^2.
     \]
   2. Bounded spectral norm: For some absolute constant \( a, \) it holds that
      \[
      \|A\|_{2 \rightarrow 2} = O(p^a).
      \]

When \( A \) has independent isotropic sub-Gaussian rows,\(^4\) from Lemma [Vershynin, 2010, Proposition 5.16], we have that when \( n = \Omega \left( \frac{1}{\epsilon^2 \cdot \log E} \right) \) (thus \( c_1 = 2 \) and \( c_2 = 1 \)), the event corresponding to (11) occurs with probability \( 1 - e^{-\Omega(n^2)} \). In addition, similarly to [Vershynin, 2010, Corollary 5.35], we have that with probability \( 1 - e^{-\Omega(n)} \), the assumption about bounded spectral norm is satisfied with \( a = 0.5 \). Moreover, from

\(^4\)A random vector \( v \) is said to be isotropic if \( E[vv^T] = I_p. \)
the theoretical results concerning JLE in [Krahmer and Ward, 2011] and the standard inequality \( \|A\|_{2 \to 2} \leq \max \{\|A\|_{1 \to 1}, \|A\|_{\infty \to \infty}\} \), we know that when \( A \) is a subsampled Fourier matrix or a partial Gaussian circulant matrix with random column sign flips and isotropic rows, these two assumptions are also satisfied with high probability for appropriate absolute constants \( c_1, c_2 \) and \( a \). Hence, these assumptions on \( A \) are significantly more generalized than the i.i.d. Gaussian assumption made for the case of unknown nonlinearity. Notably, when the two assumptions are satisfied, by a chaining argument [Bora et al., 2017; Liu et al., 2021a], the random matrix \( \frac{1}{\sqrt{n}} A \) satisfies the TS-REC for \( K \).

• The (unknown) uncorrupted measurements are generated from the following measurement model:

\[
y_i = f(a_i^T x^*) + \eta_i, \quad i = 1, 2, \ldots, n, \tag{13}
\]

where \( f : \mathbb{R} \to \mathbb{R} \) is a known (deterministic) nonlinear function, and \( \eta_i \) are additive noise terms. Similarly to [Yang et al., 2016; Soltani and Hegde, 2017; Zhang et al., 2018], we assume that \( f \) is monotonic, differentiable, and for all \( x \in \mathbb{R} \), \( f'(x) \in [l, u] \) with \( u \geq l > 0 \) being fixed constants.\(^{5}\) In addition, we assume that \( \eta_i \) are independent realizations of zero-mean sub-Gaussian random variables with maximum sub-Gaussian norm \( \sigma \).

• We also allow for adversarial noise and assume that for some \( \tau \geq 0 \), the observed (corrupted) vector \( \tilde{y} \) satisfies

\[
\frac{1}{\sqrt{n}} \|\tilde{y} - y\|_2 \leq \tau. \tag{14}
\]

• To estimate the signal \( x^* \), we utilize the knowledge of \( f \), and consider minimizing the nonlinear \( \ell_2 \) loss over \( \mathcal{K} \):

\[
\min \mathcal{L}_2(x) := \frac{1}{2n} \|y - f(Ax)\|_2^2 \quad \text{s.t. } x \in \mathcal{K}. \tag{15}
\]

Under the preceding assumptions, we have the following theorem that gives a recovery guarantee for optimal solutions to (15). The proof is placed in the supplementary material.

**Theorem 2.** Let \( \hat{x} = \arg \min_{x \in \mathcal{K}} \|x - x^*\|_2 \). For any \( \delta > 0 \), we have that any solution \( \hat{x} \) to (15) satisfies

\[
\|\hat{x} - x^*\|_2 \leq O \left( \|\tilde{x} - x^*\|_2 + \sigma \sqrt{\frac{k \log \frac{Lr}{\delta}}{n}} + \tau + \delta \right). \tag{16}
\]

In (16), the term \( \|\tilde{x} - x^*\|_2 \) corresponds to the representation error. Similarly to the discussion after Theorem 1, we see that when there is no representation error or adversarial noise, and considering \( \delta \) being sufficiently small and \( \sigma \) being a fixed constant, we obtain the optimal statistical rate, i.e.,

\[
\|\hat{x} - x^*\|_2 = O \left( \sqrt{\frac{k \log \frac{Lr}{\delta}}{n}} \right).
\]

Theorem 2 is concerned with globally optimal solutions of (15), which are intractable to obtain due to the non-convexity of the corresponding objective function. In the following, we use a PGD algorithm to approximately minimize (15). In particular, we select an initial vector \( x^{(0)} \) arbitrarily, and for any non-negative integer \( t \), letting

\[
x^{(t+1)} = \mathcal{P}_\mathcal{K} \left( x^{(t)} - \frac{\zeta}{n} \cdot \nabla \mathcal{L}_2 \left( f(Ax^{(t)}) - y \right) \right), \tag{17}
\]

where \( \zeta > 0 \) is the step size and \( \mathcal{P}_\mathcal{K} \) represents elementwise projection. The corresponding algorithm is described in Algorithm 2 for convenience.

Next, we present the following theorem, which establishes a theoretical guarantee similar to that of Theorem 1, except that there is an extra \( \|\hat{x} - x^*\|_2 \) term corresponding to representation error in the upper bound. The proof of Theorem 3 can be found in the supplementary material.

**Theorem 3.** Let \( \hat{x} = \arg \min_{x \in \mathcal{K}} \|x - x^*\|_2 \). For any \( \delta > 0 \) and \( \varepsilon \in (0, 1) \) that is sufficiently small, letting

\[
\mu_2 := \max \left\{ 1 - \zeta \varepsilon^2 (1 - \epsilon), \zeta \epsilon^2 (1 + \epsilon) - 1 \right\}. \tag{19}
\]

If \( 2 \mu_2 < 1 \) with \( 1 - 2 \mu_2 = \Theta(1) \), we have for all \( t \in \mathbb{N} \) that

\[
\|x^{(t)} - x^*\|_2 \leq (2 \mu_2)^t \|x^{(0)} - x^*\|_2 + \mu_2 \left( \|\hat{x} - x^*\|_2 + \sigma \sqrt{\frac{k \log \frac{Lr}{\delta}}{n}} + \tau + \delta \right). \tag{20}
\]

5 Experiments

In this section, we empirically evaluate the performance of Algorithm 1 (PGD-GLasso; abbreviated to PGD-G) for the linear least-squares estimator (6) and Algorithm 2

![Figure 1: Examples of reconstructed images of the MNIST dataset with \( n = 100 \) measurements and \( p = 784 \) dimensional vectors.](image-url)
Figure 2: Reconstructed images of the CelebA dataset with \( n = 500 \) measurements and \( p = 12288 \) dimensional vectors.

(PGD–NLasso; abbreviated to PGD–N) for the nonlinear least-squares estimator (15) on the MNIST [LeCun et al., 1998] and CelebA [Liu et al., 2015] datasets. For both datasets, we use the generative models pre-trained by the authors of [Bora et al., 2017]. For PGD–G, the step size \( \nu \) is set to be 1. For PGD–N, we set the step size \( \zeta \) to be 0.2. For both algorithms, the total number of iterations \( T \) is set to be 30. On the MNIST dataset, we do 5 random restarts, and pick the best estimate among these random restarts. The generative model \( G \) is set to be a variational autoencoder (VAE) model with a latent dimension of \( k = 20 \). The projection step \( \mathcal{P}_K(\cdot) \) with \( K \) being the range of \( G \) is approximated using gradient descent, performed using the Adam optimizer with a learning rate of 0.03 and 200 steps. On the CelebA dataset, we use a Deep Convolutional Generative Adversarial Networks (DCGAN) generative model with a latent dimension of \( k = 100 \). We select the best estimate among 2 random restarts. An Adam optimizer with 100 steps and a learning rate of 0.1 is used for the projection operator \( \mathcal{P}_K(\cdot) \). Throughout this section, for simplicity, we consider the case that there is no adversarial noise, i.e., \( \tau = 0 \). Since for unknown nonlinearity, the signal is recovered up to a scalar ambiguity, to compare performance across algorithms, we use a scale-invariant metric named Cosine Similarity defined as

\[
\text{Cos}(\mathbf{x}^*, \mathbf{x}^{(T)}) := \frac{\langle \mathbf{x}^*, \mathbf{x}^{(T)} \rangle}{\|\mathbf{x}^*\|_2 \|\mathbf{x}^{(T)}\|_2},
\]

where \( \mathbf{x}^* \) is the signal vector to estimate, and \( \mathbf{x}^{(T)} \) denotes the output vector of an algorithm. We use Python 2.7 and TensorFlow 1.0.1, with a NVIDIA Tesla K80 24GB GPU.

We follow the measurement model (13) to generate the observations, with \( f(x) = 2x + 0.5 \cos(x) \) and the noises being independent zero-mean Gaussian with standard deviation \( \sigma \). We observe that \( f \) is smooth and monotonically increasing with \( t = 1.5 \) and \( u = 2.5 \). Thus, we perform PGD–N on the generated data. In addition, since the assumption made for unknown nonlinearity that \( f(g) \) for \( g \sim \mathcal{N}(0, 1) \) is sub-Gaussian is satisfied, we also compare with PGD–G. The standard deviation \( \sigma \) is set to be 0.1 for the MNIST dataset, and 0.01 for the CelebA dataset. The baseline is the Lasso using 2D Discrete Cosine Transform (2D-DCT) basis [Tibshirani, 1996] and the method for linear inverse problem with generative models proposed in [Bora et al., 2017] (denoted by CSGM). For Lasso, CSGM and PGD–G, the sensing matrix \( A \) is assumed to contain i.i.d. standard Gaussian entries. Moreover, since for a known and monotonic nonlinear link function, we allow for a wide variety of distributions on the sensing vectors, for PGD–N, we consider both cases where \( A \) is a standard Gaussian matrix or a partial Gaussian circulant matrix similar to that in [Liu et al., 2021b]. The corresponding Algorithm 2 is denoted by PGD–N (G) or PGD–N (C) respectively, where “G” refers to the standard Gaussian matrix and “C” refers to the partial Gaussian circulant matrix.

We perform experiments to compare the performance of these algorithms, and the reconstructed results are reported in Figures 1, 2, and 3. We observe that for our settings, the sparsity-based method Lasso always attains poor reconstructions, while all three generative model based PGD methods attain accurate reconstructions even when the number of measurements \( n \) is small compared to the ambient dimension \( p \). In addition, from Figure 3(a), we observe that on the MINST dataset, PGD–N (G) and PGD–N (C) lead to similar Cosine Similarity, and they are clearly better than CSGM and PGD–G when \( n < 300 \). This is not surprising since both CSGM and PGD–G does not make use of the knowledge of the nonlinear link function. From Figures 2 and 3(b), we see that for the CelebA dataset, three generative prior based PGD methods give similar reconstructed images, with the Cosine Similarity corresponding to PGD–N (G) and PGD–N (C) being slightly higher than that of PGD–G, and all of them lead to clearly higher Cosine Similarity compared to that of CSGM.

Numerical results for noisy 1-bit measurements are presented in the supplementary material.

6 Conclusion

We have proposed PGD algorithms to solve generative model based nonlinear inverse problems, and we have provided theoretical guarantees for these algorithms for both unknown and known link functions. In particular, these algorithms are guaranteed to converge linearly to points achieving optimal statistical rate in spite of the model nonlinearity.

Acknowledgments

We sincerely thank the three anonymous reviewers for their careful reading and insightful comments. We are extremely grateful to Dr. Jonathan Scarlett for proofreading the manuscript and giving valuable suggestions.
References

[Bora et al., 2017] A. Bora, A. Jalal, E. Price, and A. Dimakis. Compressed sensing using generative models. In ICML, 2017.

[Foucart and Rauhut, 2013] S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Springer New York, 2013.

[Genzel, 2016] M. Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. IEEE Trans. Inf. Theory, 63(3):1601–1619, 2016.

[Han, 1987] A. Han. Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. J. Econom., 35(2-3):303–316, 1987.

[Hand et al., 2018] P. Hand, O. Leong, and V. Voroninski. Phase retrieval under a generative prior. In NeurIPS, 2018.

[Heckel and Hand, 2019] R. Heckel and P. Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In ICLR, 2019.

[Hyder et al., 2019] R. Hyder, V. Shah, C. Hegde, and S. Asif. Alternating phase projected gradient descent with generative priors for solving compressive phase retrieval. In ICASSP, 2019.

[Jagatap and Hegde, 2019] G. Jagatap and C. Hegde. Algorithmic guarantees for inverse imaging with untrained network priors. In NeurIPS, 2019.

[Jalal et al., 2021] A. Jalal, S. Karmalkar, A. Dimakis, and E. Price. Instance-optimal compressed sensing via posterior sampling. In ICML, 2021.

[Kamath et al., 2020] A. Kamath, S. Karmalkar, and E. Price. On the power of compressed sensing with generative priors. In ICML, 2020.

[Krahmer and Ward, 2011] F. Krahmer and R. Ward. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. SIAM J. Math. Anal., 43(3):1269–1281, 2011.

[LeCun et al., 1998] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proc. IEEE, 86(11), 1998.

[Liu and Scarlett, 2020a] Z. Liu and J. Scarlett. The generalized Lasso with nonlinear observations and generative priors. In NeurIPS, 2020.

[Liu and Scarlett, 2020b] Z. Liu and J. Scarlett. Information-theoretic lower bounds for compressive sensing with generative models. IEEE J. Sel. Areas Inf. Theory, 1(1):292–303, 2020.

[Liu et al., 2015] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In ICCV, 2015.

[Liu et al., 2020] 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 ICML, 2020.

[Liu et al., 2021a] Z. Liu, S. Ghosh, J. Han, and J. Scarlett. Robust 1-bit compressive sensing with partial Gaussian circulant matrices and generative priors. https://arxiv.org/2108.03570, 2021.

[Liu et al., 2021b] Z. Liu, S. Ghosh, and J. Scarlett. Towards sample-optimal compressive phase retrieval with sparse and generative priors. In NeurIPS, 2021.

[Liu et al., 2022] Z. Liu, J. Liu, S. Ghosh, J. Han, and J. Scarlett. Generative principal component analysis. In ICLR, 2022.

[Ongie et al., 2020] G. Ongie, A. Jalal, C. Metzler, R. Baraniuk, A. Dimakis, and R. Willett. Deep learning techniques for inverse problems in imaging. IEEE J. Sel. Areas Inf. Theory, 1(1):39–56, 2020.

[Oymak and Soltanolkotabi, 2017] S. Oymak and M. Soltanolkotabi. Fast and reliable parameter estimation from nonlinear observations. SIAM J. Optim., 27(4):2276–2300, 2017.

[Peng et al., 2020] P. Peng, S. Jalal, and X. Yuan. Solving inverse problems via auto-encoders. IEEE J. Sel. Areas Inf. Theory, 1(1):312–323, 2020.

[Plan and Vershynin, 2016] Y. Plan and R. Vershynin. The generalized Lasso with non-linear observations. IEEE Trans. Inf. Theory, 62(3):1528–1537, 2016.

[Plan et al., 2017] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. Inf. Inference, 6(1):1–40, 2017.

[Qiu et al., 2020] S. Qiu, X. Wei, and Z. Yang. Robust one-bit recovery via ReLU generative networks: Improved statistical rates and global landscape analysis. In IJCAI, 2020.

[Raj et al., 2019] A. Raj, Y. Li, and Y. Bresler. GAN-based projector for faster recovery with convergence guarantees in linear inverse problems. In ICCV, 2019.

[Shah and Hegde, 2018] V. Shah and C. Hegde. Solving linear inverse problems using GAN priors: An algorithm with provable guarantees. In ICASSP, 2018.

[Soltani and Hegde, 2017] M. Soltani and C. Hegde. Fast algorithms for demixing sparse signals from nonlinear observations. IEEE Trans. Sig. Proc., 65(16):4209–4222, 2017.

[Shah and Hegde, 2018] V. Shah and C. Hegde. Solving linear inverse problems using GAN priors: An algorithm with provable guarantees. In ICASSP, 2018.

[Tibshirani, 1996] R Tibshirani. Regression shrinkage and selection via the Lasso. JRSSB, pages 267–288, 1996.

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

[Wei et al., 2019] X. Wei, Z. Yang, and Z. Wang. On the statistical rate of nonlinear recovery in generative models with heavy-tailed data. In ICML, 2019.

[Yang et al., 2016] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. Sparse nonlinear regression: Parameter estimation under nonconvexity. In ICML, 2016.

[Zhang et al., 2018] K. Zhang, Z. Yang, and Z. Wang. Nonlinear structured signal estimation in high dimensions via iterative hard thresholding. In AISTATS, 2018.