On the statistical rate of nonlinear recovery in generative models with heavy-tailed data

Xiaohan Wei [1] **Zhuoran Yang** [2] **Zhaoran Wang** [3]

Abstract

We consider estimating a high-dimensional vector from non-linear measurements where the unknown vector is represented by a generative model $G : \mathbb{R}^k \to \mathbb{R}^d$ with $k \ll d$. Such a model poses structural priors on the unknown vector without having a dedicated basis, and in particular allows new and efficient approaches solving recovery problems with number of measurements far less than the ambient dimension of the vector. While progresses have been made recently regarding theoretical understandings on the linear Gaussian measurements, much less is known when the model is possibly misspecified and the measurements are non-Gaussian. In this paper, we make a step towards such a direction by considering the scenario where the measurements are non-Gaussian, subject to possibly unknown nonlinear transformations and the responses are heavy-tailed. We then propose new estimators via score functions based on the first and second order Stein's identity, and prove the sample size bound of $m = \mathcal{O}(k\varepsilon^{-2}\log(L/\varepsilon))$ achieving an ε error in the form of exponential concentration inequalities. Furthermore, for the special case of multi-layer ReLU generative model, we improve the sample bound by a logarithm factor to $m = \mathcal{O}(k\varepsilon^{-2}\log(d))$, matching the state-of-art statistical rate in compressed sensing for estimating k-sparse vectors. On the technical side, we develop new chaining methods bounding heavy-tailed processes, which could be of independent interest.

[1]Department of Electrical Engineering, University of Southern California, Los Angeles, CA, USA. [2]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA. [3]Department of Industrial Engineering and Management Sciences, Northwestern University, Chicago, IL, USA. Correspondence to: Xiaohan Wei <xiaohanw@usc.edu>, Zhuoran Yang <zy6@princeton.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

1. Introduction

Consider a random pair $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ with the distribution satisfying the following semi-parametrized single index model:

$$y = f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big), \tag{1}$$

where $\theta^* \in \mathbb{R}^k$, the vector $\mathbf{x} \in \mathbb{R}^d$ is a random vector following a probability density $p(\mathbf{x})$, ξ is a random noise variable assumed to be independent of $\mathbf{x}$, and y is the response variable. The function $G : \mathbb{R}^k \to \mathbb{R}^d$ is an L-Lipschitz mapping generating a high-dimensional vector $G(\theta^*) \in \mathbb{R}^d$ from a low-dimensional signal $\theta^* \in \mathbb{R}^k$. Such a mapping can be arbitrary but otherwise fixed generator, which implicitly captures the structural property of the high-dimensional vector. For example, in neural nets based generative models such as generative adversarial nets (GANs) (Goodfellow et al., 2014), the generator is responsible for generating high dimensional vectors in $\mathbb{R}^d$ that resemble the samples in the training image dataset from a low dimensional representation space via a deep neural network. Since the image samples are often distributed near a low-dimensional manifold of $\mathbb{R}^d$, a well-trained generative model is able to capture such a manifold structure, which can be difficult to represent in canonical basis otherwise.

The non-linear link function $f : \mathbb{R}^2 \to \mathbb{R}$ is *unknown*. Throughout the paper, we make no specific assumption on the form of $f(\cdot)$ apart from differentiability. Our goal is to recover the high dimensional vector $G(\theta^*)$ from a sequence of i.i.d. copies $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ of $(\mathbf{x}, y)$. Since $f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big) = f\big(a^{-1}\langle \mathbf{x}, a \cdot G(\theta^*)\rangle,\ \xi\big)$ for any $a > 0$, one can only hope to recover $G(\theta^*)$ up to constant scaling. For simplicity of presentation, we assume that $\|G(\theta^*)\|_2 = 1$ throughout the paper.

Despite sharing some similarities with compressed sensing, computing a consistent estimator of $G(\theta^*)$ in such a scenario is challenging mainly because of the coupling of two aspects: First, in contrast to the classical structured signal recovery (Hastie et al., 2015; Lecué & Mendelson, 2014; Candes et al., 2006) which rely on the notion of sparsity on some chosen basis, the structure of the unknown here is inherited in the generative model $G(\cdot)$. The proposed method is expected to capture a complexity measure of $G(\cdot)$

similar to sparsity level in the classical setting. Second, even for the classical compressed sensing scenario, the unknown distortion $f(\cdot)$ creates an additional source of ambiguity, resulting in a non-diminishing bias for least square estimators (Ai et al., 2014; Goldstein & Wei, 2016).

1.1. Related works

Deep generative models have been applied to a variety of modern machine learning areas. Some notable applications include synthesizing images that resemble the realistic ones via generative adversarial nets (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017) and sampling from high dimensional posterior distributions using variational auto-encoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014). In both scenarios, a deep neural network takes Gaussian random vectors as inputs and outputs generative samples whose distribution is close to the target image/signal distribution, when trained using a sufficiently large number of target samples.

Another line of works, which is more related to this paper, focuses on using deep generative models to solve inverse problems, and has find extensive empirical successes in image reconstructions such as super-resolution (Sønderby et al., 2016; Ledig et al., 2017), image impainting (Yeh et al., 2017) and medical imaging (Hammernik et al., 2018; Yang et al., 2018). In particular, these generative model based methods have been shown to produce comparable results to the classical sparsity based methods with much fewer (sometimes 5-10x fewer) measurements, which will greatly benefit application areas such as magnetic resonance imaging (MRI) and computed tomography (CT), where the measurements are usually quite expensive to obtain. These recent, yet impressive empirical results drive researchers to look for theoretical justifications from various perspectives.

In a recent work (Bora et al., 2017), the authors consider a linear model

$$\mathbf{y} = \mathbf{A}G(\theta^*) + \eta,$$

where $\mathbf{A}$ is a Gaussian measurement matrix and η is a bounded noise term. By showing that the Gaussian measurement matrix satisfies a restricted eigenvalue condition (REC) over the range of the generative model $G(\cdot)$, the authors prove the L_2 empirical risk minimizer

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \|\mathbf{A}G(\theta) - \mathbf{y}\|_2^2 \tag{2}$$

satisfies an estimation error bound $\|\eta\|_2 + \varepsilon$ when the number of samples is of the order $\mathcal{O}(k \log \frac{L}{\varepsilon})$. They further show that the ε term in the error bound can be removed when $G(\cdot)$ is a multilayer ReLU network. The works (Hand & Voroninski, 2017; Huang et al., 2018) consider the same linear model with the aforementioned L_2 empirical risk minimizer and a multilayer ReLU network $G(\cdot)$. They show under the REC condition on the measurement matrix and other suitable conditions on the weights of the ReLU function, the gradient descent algorithm essentially converges to the global minimum of the the empirical risk objective. Furthermore, the work (Hand et al., 2018) considers the phase retrieval problem under an n-layer ReLU generator $G(\cdot)$ and show that when the measurement matrix satisfies a range restricted concentration, which is stronger than REC, and the number of measurements is greater than $\mathcal{O}(kn \log d)$, the gradient descent algorithm on the empirical amplitude flow risk converges to neighborhoods of the global minimum.

In the absence of generative models, the problem of recovering a vector θ^* from the measurements of the form $y = f\big(\langle \mathbf{x}, \theta^* \rangle\big) + \xi$, which is usually referred to as the single index model (SIM), has been studied in several previous works. One of the most classical, yet surprising results dates back to (Brillinger, 1983), which states that if the measurement vector $\mathbf{x}$ is Gaussian, then, one can simply "ignore" the nonlinearity $f(\cdot)$ and estimate θ^* up to scalar scaling by solving the ordinary least square: $\operatorname{argmin}_{\theta \in \mathbb{R}^d} \mathbb{E}\big[(y - \langle \mathbf{x}, \theta \rangle)^2\big]$. Later, (Li & Duan, 1989) shows the same result holds for the class of elliptical symmetric measurements. More recently, the works (Plan et al., 2017; Plan & Vershynin, 2016) study the SIM with Gaussian measurements when the true vector θ^* is high dimensional but lies in a conic set, whose recovery error bound can then be represented in terms of the Gaussian complexity measure of the set. The works (Goldstein et al., 2018; Wei, 2018) consider high-dimensional SIM with heavy-tailed elliptical symmetric measurements and propose thresholded least square estimators with a similar performance guarantee as the Gaussian case. As is discussed in (Ai et al., 2014; Goldstein & Wei, 2016), for SIM with general non-Gaussian measurements, least square estimators can incur a fixed bias term regardless of how many measurements being taken. To obtain a consistent estimator for the general non-Gaussian measurements, the works (Yang et al., 2017a;b) propose thresholded score function estimators via Stein's identity. While treating heavy-tailed measurements, their methods and analysis depend heavily on the chosen basis and thus apply to estimation of sparse and low-rank signals only.

1.2. Our contributions

In this paper, we make a step towards understanding the high-dimensional generative model $G(\cdot)$ by proposing new estimation methods and proving recovery guarantees without involving any REC type conditions, which allows the measurements to be non-Gaussian and the responses to be heavy-tailed. More specifically, under the assumption that the distribution of the measurement $\mathbf{x}$ is known a priori and the link function $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable on the first argument,

- we propose a LASSO type estimator based on the first

order Stein's identity and show a sample size bound of $m = \mathcal{O}\left(k\varepsilon^{-2}\log(L/\varepsilon)\right)$ achieving an ε estimation error, which applies to the scenario when the link function satisfies $\mathbb{E}[f'(\langle \mathbf{x}, G(\theta^*)\rangle, \xi)] \neq 0$.[1] In particular, this error bound depends logarithmically on the Lipschitz constant L, matching the sample complexity of the linear Gaussian scenario.

- we propose a PCA type estimator based on the second order Stein's identity and show the same $m = \mathcal{O}\left(k\varepsilon^{-2}\log(L/\varepsilon)\right)$ bound achieving ε estimation error, allowing $\mathbb{E}[f'(\langle \mathbf{x}, G(\theta^*)\rangle, \xi)] = 0$ (and in particular the phase retrieval scenario) but at the cost of solving a potentially more challenging optimization problem.
- In the special case where the generative model $G(\cdot)$ is an n-layer ReLU network with each layer having at most d nodes, the Lipschitz constant can be bounded by $\mathcal{O}(d^n)$. We show that for each of the two methods above, the sample bound can be improved by a logarithm factor to $m = \mathcal{O}\left(kn\varepsilon^{-2}\log(d)\right)$. In particular, this error bound depends logarithmically on the output dimension d, matching state-of-art statistical rate in compressed sensing for estimating k-sparse vectors when the number of layers n is small.

On the technical side, our results are built upon new chaining arguments bounding heavy-tailed processes over the range of $G(\cdot)$ as well as piecewise linearity analysis of the ReLU network, which could be of independent interests.

1.3. Notations

Throughout the paper, for any vector $\mathbf{v} \in \mathbb{R}^d$, $\|\mathbf{v}\|_p$, $p \geq 1$ denotes the Euclidean vector p-norm. For any real random variable X, the L_p norm $(\mathbb{E}[|X|^p]^{1/p}$, $p \geq 1)$ is denoted as $\|X\|_{L_p}$. The Orlicz ψ_2-norm is denoted as $\|X\|_{\psi_2} := \sup_{p\geq 1} p^{-1/2}\|X\|_{L_p}$. We use $\mathcal{S}^{d-1}$ to denote the unit sphere in $\mathbb{R}^d$ and use $\mathcal{B}^k(r)$ to denote the ball of radius r centered at the origin in $\mathbb{R}^k$. For any matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, $\|\mathbf{A}\|_F := (\sum_{i=1}^d \sum_{j=1}^d |A_{ij}|^2)^{1/2}$. For any two matrices $\mathbf{A}$, $\mathbf{B} \in \mathbb{R}^{d\times d}$, the matrix inner product $\langle \mathbf{A}, \mathbf{B}\rangle := \text{Trace}(\mathbf{A}^T\mathbf{B})$. For any (twice) differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, ∇g and $\nabla^2 g$ denote the gradient and Hessian of g, respectively. For a random variable $X \in \mathbb{R}$ with density p, we use $p^{\otimes d} : \mathbb{R}^d \to \mathbb{R}$ to denote the joint density of $X_1, \cdots, X_d$, which are d identical copies of X. Finally, the notations C, C_1, C_2, C', C'' are all absolute constants, and their values can be different per appearance.

2. Model definitions

We start by defining the score function of a distribution. Let $p : \mathbb{R}^d \to \mathbb{R}$ be a differentiable probability density function on $\mathbb{R}^d$. Define the score function $S_p(\mathbf{x})$ as

$$S_p(\mathbf{x}) = -\nabla \log p(\mathbf{x}) = -\nabla p(\mathbf{x})/p(\mathbf{x}), \;\; \mathbf{x} \in \mathbb{R}^d.$$

Moreover, a mapping $G : \mathbb{R}^k \to \mathbb{R}^d$ is Lipschitz with a constant L if it satisfies

$$\|G(\mathbf{a}) - G(\mathbf{b})\|_2 \leq L\|\mathbf{a} - \mathbf{b}\|_2, \;\; \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^k.$$

2.1. First order link function and Stein's identity

We first discuss the statistical model based on the first order link function, which results from the first order Stein's identity. Recall the following Stein's identity:

Proposition 2.1 ((Stein et al., 2004)). *Let $X \in \mathbb{R}^d$ be a real valued random vector with density p. Assume that $p : \mathbb{R}^d \to \mathbb{R}$ is differentiable. In addition, let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function such that $\mathbb{E}[\nabla g(X)]$ exists. Then it holds that $\mathbb{E}[g(X)S_p(X)] = \mathbb{E}[\nabla g(X)]$, where $S_p(X) = -\nabla p(X)/p(X)$.*

To see how Stein's identity can help us estimate $G(\theta^*)$ from $f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)$, we consider the following correlation:

$$\begin{aligned}
\mathbb{E}[y \cdot S_p(\mathbf{x})] =& \mathbb{E}\big[f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)S_p(\mathbf{x})\big] \\
=& \mathbb{E}\big[\mathbb{E}\big[\, f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big) \cdot S_p(\mathbf{x}) \,\big|\, \xi\big]\big] \\
=& \mathbb{E}\big[\mathbb{E}\big[\, \nabla_{\mathbf{x}} f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big) \,\big|\, \xi\big]\big] \\
=& \mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \cdot G(\theta^*), \qquad (3)
\end{aligned}$$

where for $f : \mathbb{R}^2 \to \mathbb{R}$, and we use $f'(x, y)$ to denote $\partial f(x, y)/\partial x$ use $\nabla_{\mathbf{x}}$ to denote taking a gradient on $\mathbf{x}$ argument only, and the third equality follows from the Stein's identity. Thus, when $\mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \neq 0$, one can estimate $G(\theta^*)$ from the expectation $\mathbb{E}[y \cdot S_p(\mathbf{x})]$, which serves as our motivation for the following first order link definition:

Assumption 2.1 (First order link). *Consider the model (1). The function f is differentiable on the first argument. The entries of $\mathbf{x}$ are i.i.d. with differentiable density p_0 and $\lambda := \mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \neq 0$, where $f'(x, y) = \partial f(x, y)/\partial x$. Furthermore, there exists $r > 0$, such that $\theta^* \in \mathcal{B}^k(r)$ and $\exists \overline{\theta} \in \mathcal{B}^k(r)$ satisfying $\lambda \cdot G(\theta^*) = G(\overline{\theta})$.*

Remark 2.1. *Note that under the first order link model, one has $S_p(\mathbf{x}) = -\nabla p(\mathbf{x})/p(\mathbf{x}) = s_0 \circ (\mathbf{x})$, where $s_0 \circ (\mathbf{x})$ denotes the entry-wise application of the 1-d score function $s_0 = -p_0'/p_0$ to $\mathbf{x}$. Furthermore, the condition $\lambda \cdot G(\theta^*) = G(\overline{\theta})$ is mild. For example, consider the case G is a multi-layer ReLU network $G(\theta) = \sigma \circ (\mathbf{W}_n\sigma \circ (\mathbf{W}_{n-1} \cdots \sigma \circ (\mathbf{W}_1\theta)))$, where $\sigma(x) = \max(x, 0)$. Suppose $\lambda \geq 0$, then we have $\lambda \cdot G(\theta^*) = G(\lambda \cdot \theta^*)$. More generally, the condition $\lambda \cdot G(\theta^*) = G(\overline{\theta})$ holds when $G(\mathbb{R}^k)$ coincides with the cone of $G(\mathbb{R}^k)$ in $\mathbb{R}^d$.*[2]

[1] We use $f'(x, y)$ to denote $\frac{\partial f(x,y)}{\partial x}$.

[2] For any set $\Theta \in \mathbb{R}^d$, the cone of Θ is the set $\{\tau\mathbf{h} : \tau \in \mathbb{R}, \mathbf{h} \in \Theta\}$.

2.2. Second order link function and Stein's identity

The previous model requires $\mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \neq 0$, which is sometimes restrictive. For example, it excludes the phase retrieval link where $f(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi) = |\langle \mathbf{x}, G(\theta^*)\rangle|^2 + \xi$. In view of this limitation, we consider the following second order Stein's identity.

Proposition 2.2 ((Janzamin et al., 2014)). *Let $X \in \mathbb{R}^d$ be a real valued random vector with density p. Assume that $p : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable. Define the second order score function $T(X) := \nabla^2 p(X)/p(X)$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous function such that $\mathbb{E}[\nabla^2 g(X)]$ exists. Then, $\mathbb{E}[g(X)T(X)] = \mathbb{E}[\nabla^2 g(X)]$.*

To see how one can extract $G(\theta^*)$ from $f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)$ using this Stein's identity, we consider

$$\begin{aligned}\mathbb{E}[yT(\mathbf{x})] =& \mathbb{E}\big[f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)T(\mathbf{x})\big]\\ =& \mathbb{E}\big[\mathbb{E}\big[\,f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)T(\mathbf{x}) \mid \xi\big]\big]\\ =& \mathbb{E}\big[\mathbb{E}\big[\,\nabla^2_{\mathbf{x}} f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big) \mid \xi\big]\big]\\ =& \mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big]G(\theta^*)G(\theta^*)^T, \quad (4)\end{aligned}$$

where for $f : \mathbb{R}^2 \to \mathbb{R}$, $f''(x, y) = \partial^2 f(x, y)/\partial x^2$, $\nabla^2_{\mathbf{x}}$ denotes taking Hessian on $\mathbf{x}$ argument only, and the third equality follows from the Stein's identity. Thus, when $\mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] > 0$, one can estimate $G(\theta^*)$ by estimating the principle component of the matrix $\mathbb{E}[yT(\mathbf{x})]$, which motivates the following definition.

Assumption 2.2 (Second order link). *Consider the model $y = f\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)$. The function f is twice differentiable on the first argument. The entries of $\mathbf{x}$ are i.i.d. with twice differentiable density p_0 and the expectation $\mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] > 0$.*

Remark 2.2. *Note that when $\mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] < 0$, we can replace y by $-y$ and the above condition is satisfied. Thus, the condition we require is essentially $\mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \neq 0$. Furthermore, under the second order link model, the joint density is $p(\mathbf{x}) := p_0^{\otimes d}(\mathbf{x}) = \prod_{i=1}^d p_0(x_i)$. Recall the 1-d score function $s_0(x) = p_0'(x)/p(x)$, and the second order score function $T(\mathbf{x})$ can be expressed by $S_p(\mathbf{x})$ as*

$$T(\mathbf{x}) = S_p(\mathbf{x})S_p(\mathbf{x})^T - \mathit{diag}(s_0' \circ (\mathbf{x})), \quad (5)$$

where $s_0' \circ (\mathbf{x})$ denotes the entry-wise application of $s_0' = \big((p_0')^2 - p_0'' p_0\big)\big/\, p_0^2$ to $\mathbf{x}$.

3. Main results on first order links

3.1. Theoretical results for Lipschitz generative models

In this section, we present theoretical results which hold for any arbitrary but fixed L-Lipschitz function $G(\cdot)$. According to the derivation (3), one could estimate $G(\theta^*)$ by solving the following optimization problem:

$$\theta_0 \in \operatorname*{argmin}_{\theta \in \mathcal{B}^k(r)} \|G(\theta)\|_2^2 - 2\mathbb{E}[y\langle S_p(\mathbf{x}), G(\theta)\rangle], \quad (6)$$

for some radius $r > 0$ large enough. Indeed, by (3),

$$\begin{aligned}\mathbb{E}[yS_p(\mathbf{x})] &= \mathbb{E}[f(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi)S_p(\mathbf{x})]\\ &= \mathbb{E}[f'(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi)] \cdot G(\theta^*) = \lambda G(\theta^*),\end{aligned}$$

thus, $\theta_0 \in \operatorname{argmin}_{\theta \in \mathcal{B}^k(r)} \|G(\theta) - \lambda G(\theta^*)\|_2^2$, which implies $G(\theta_0) = \lambda G(\theta^*)$. In the current scenario, we propose to solve an empirical version of (6), i.e.

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathcal{B}^k(r)} \|G(\theta)\|_2^2 - \frac{2}{m}\sum_{i=1}^m y_i \langle S_p(\mathbf{x}_i), G(\theta)\rangle. \quad (7)$$

Remark 3.1. *There is no guarantee that the solution to the minimization problems (6) and (7) are unique. However, since the objectives in these problems are quadratic on $G(\theta)$, the solution is unique on the range of $G(\cdot)$. Thus our estimator $G(\widehat{\theta})$ of $G(\theta^*)$ is uniquely defined.*

Since the function $f(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi)$ is arbitrary and unknown, it is unrealistic to assume that the responses $\{y_i\}_{i=1}^m$ have nice statistical properties (i.e. higher order moments exist). Throughout the paper, we make the following moment assumptions on the measurements:

Assumption 3.1. *Each entry of the random vector $S_p(\mathbf{x})$ is subgaussian, i.e. $\|[S_p(\mathbf{x})]_i\|_{\psi_2} < \infty,\ i = 1, 2, \cdots, d$. Furthermore, we assume $y \in L_q$, i.e. $\mathbb{E}[|y|^q]^{1/q} < \infty$, for some $q > 4$.*

Remark 3.2. *This assumption implies that the random vector $S_p(\mathbf{x})$ is subgaussian, i.e. $\forall t \in \mathcal{S}^{d-1}$, $\|\langle S_p(\mathbf{x}), t\rangle\|_{\psi_2} \leq C_0$ for some absolution constant C_0. [3] For the rest of the paper, let $\|S_p(\mathbf{x})\|_{\psi_2} := \sup_{t \in \mathcal{B}^d(1)} \|\langle S_p(\mathbf{x}), t\rangle\|_{\psi_2}$.*

Since the response y is heavy-tailed, it is not guaranteed that the resulting estimator $\widehat{\theta}$ is well concentrated around the ground truth. To this point, we propose a thresholded version of the above empirical estimator. Let $\tilde{y}_i := \operatorname{sign}(y_i) \cdot |y_i| \wedge \tau$ be the truncated version of y_i, with the truncation level τ to be determined below. We then consider learning $\lambda G(\theta^*)$ via the following optimization problem:

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathcal{B}^k(r)} \|G(\theta)\|_2^2 - \frac{2}{m}\sum_{i=1}^m \widetilde{y}_i \langle S_p(\mathbf{x}_i), G(\theta)\rangle. \quad (8)$$

The following theorem characterizes the statistical rate of the proposed estimator.

[3] For any $u > 0$, $\mathbb{E}[\exp(u\langle S_p(\mathbf{x}), t\rangle)] = \prod_{i=1}^d \mathbb{E}[\exp(u[S_p(\mathbf{x})]_i t_i)] \leq \prod_{i=1}^d \exp(Cu^2 \|[S_p(\mathbf{x})]_i\|_{\psi_2} t_i^2) \leq \prod_{i=1}^d \exp(Cu^2 \cdot \max_i \|[S_p(\mathbf{x})]_i\|_{\psi_2})$, for some absolute constant $C > 0$, , which implies the subgaussianity of the vector.

Theorem 3.1. *Let $\delta \in (0, r)$, $\eta, \beta \geq 2$ be any constants. Suppose $\tau = m^{1/2(1+\kappa)}\sigma_y$, where $\sigma_y \geq \|y\|_{L_q}$, $\kappa \in (0, q/4 - 1)$ are any chosen constants, then, with probability at least $1 - e^{-\eta} - e^{-\beta}$, the solution to (8) satisfies*

$$\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 \leq \overline{C}\sqrt{\frac{\eta + k\log(4Lr/\delta)}{m}} + \sqrt{\overline{C}\delta}\Big(\frac{\eta + k\log(4Lr/\delta)}{m}\Big)^{1/4},$$

where $\overline{C} = C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta(1+\kappa)/\kappa}$, $C > 0$ is an absolute constant and $\lambda = \mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big]$. Furthermore, for any fixed accuracy level $\varepsilon \in (0, 1]$, take $\delta = \varepsilon$ and we have $\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 \leq \varepsilon$ with a sample size $m = \mathcal{O}\left(k\varepsilon^{-2}\log(4Lr/\varepsilon)\right)$.

Remark 3.3. *Note that by Assumption (2.1), $\lambda \neq 0$. The previous theorem shows that one can estimate $G(\theta^*)$ up to constant scaling. Using the assumption that $\|G(\theta^*)\|_2 = 1$, one can further estimate $G(\theta^*)$ itself by a simple normalization: Define the normalization estimator*

$$\overline{G(\theta)} = \begin{cases} G(\widehat{\theta})/\big\|G(\widehat{\theta})\big\|_2, & \big\|G(\widehat{\theta})\big\|_2 > 0, \\ 0, & \big\|G(\widehat{\theta})\big\|_2 = 0. \end{cases}$$

Then, by a similar argument as Corollary 1.1 of (Goldstein & Wei, 2016), we can show that with probability at least $1 - e^{-\eta} - e^{-\beta}$, $\|\overline{G(\theta)} - G(\theta^*)\|_2 \leq \varepsilon/|\lambda|$ with the same sample size $m = \mathcal{O}(k\varepsilon^{-2}\log(4Lr/\varepsilon))$.

3.2. Theoretical results for ReLU generative model

In this section, we show that if we make more assumptions on the generative model, then, a slightly sharper result can be obtained. Specifically, we will consider one of the most widely used generative model, namely, the multilayer ReLU network with n-layers and at most d nodes at each layer,

$$G(\theta) = \sigma \circ (\mathbf{W}_n\sigma \circ (\mathbf{W}_{n-1} \cdots \sigma \circ (\mathbf{W}_1\theta))), \quad (9)$$

where $\sigma(x) = \max(x, 0)$ and $\sigma \circ (\mathbf{x})$ denotes the entry-wise application of $\sigma(\cdot)$. One can easily show that in this scenario, the Lipschitz constant of $G(\cdot)$ is bounded by $\mathcal{O}(d^n)$. We consider the following optimization:

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \|G(\theta)\|_2^2 - \frac{2}{m}\sum_{i=1}^m \widetilde{y}_i\langle S_p(\mathbf{x}_i), G(\theta)\rangle. \quad (10)$$

The following theorem provides the statistical rate:

Theorem 3.2. *Let $\eta, \beta \geq 2$ be any constants. Suppose $G(\cdot)$ is a ReLU generative model with n-layers and at most d nodes at each layer, $\tau = m^{1/2(1+\kappa)}\sigma_y$, where $\sigma_y \geq \|y\|_{L_q}$, $\kappa \in (0, q/4 - 1)$ are any chosen constants, then, with probability at least $1 - e^{-\eta} - e^{-\beta}$, the solution to (10) satisfies*

$$\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 \leq \overline{C}\sqrt{\frac{\eta + kn\log(2d)}{m}},$$

where $\overline{C} = C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta}\sqrt{\frac{1+\kappa}{\kappa}}$, $C > 0$ is an absolute constant and $\lambda = \mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] > 0$. Furthermore, for any fixed accuracy level $\varepsilon \in (0, 1]$, take $\delta = \varepsilon$ and we have $\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 \leq \varepsilon$ with a sample bound $m = \mathcal{O}(kn\varepsilon^{-2}\log d)$.

Similar to Theorem 3.1, one can further normalize $G(\widehat{\theta})$ and estimate $G(\theta^*)$ itself. Finally, we note that (10) is slightly simpler than (8) in that we do not impose the constraint $\theta \in \mathcal{B}^k(r)$ in the optimization (i.e. no prior knowledge on the radius r is required), and the result in Theorem 3.2 is tightened by a logarithm factor compared to that of Theorem 3.1. The intuition is that ReLU generative model is piecewise linear and within each single piece it is a linear map. Thus, within each piece, we are essentially doing optimization over a k-dimensional subspace of $\mathbb{R}^d$, which is relatively easy to understand. Hence, our result in this section can actually be generalized to any piecewise linear function without too much overhead. Furthermore, our bound matches the optimal statistical rate $\mathcal{O}(k\varepsilon^{-2}\log d)$ for sparse recovery when the number of layers n is small compared to the input dimension k.

Remark 3.4. *From a computation point of view, both (8) and (10) can be approximately solved via gradient descent. In particular, the works (Huang et al., 2018; Hand & Voroninski, 2017) show that under suitable conditions on the weights of the ReLU function, gradient descent on the L_2 empirical risk objective (2) converges to the global minimum of that objective. Our objective (10) is arguably simpler than (2) in the sense that the quadratic term does not involve the measurement matrix. Thus, the same argument in (Huang et al., 2018) carries through with little modification, showing that gradient descent also converges to the global minimum of (10). We omitted the details for brevity.*

4. Main results on second order links

As is mentioned previously, the first order link assumes $\mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] \neq 0$, which precludes the phase retrieval link function, and more generally misspecified phase retrieval problems considered in a recent work (Yang et al., 2017c). We now introduce new estimators which work for these cases and establish their statistical rates of convergence. Our method is of PCA type and works under the second order link conditions.

4.1. Theoretical results for Lipschitz generative models

According to the derivation (4), our goal is to estimate the principle eigenvector of $\mathbb{E}[yT(\mathbf{x})]$. Let $\lambda_1 := \mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] > 0$, then, using the fact that $\|G(\theta^*)\|_2 = 1$, one can estimate $G(\theta^*)$ by solving the

following optimization:

$$\theta_0 \in \underset{\theta\in\mathcal{B}^k(r):\ \|G(\theta)\|_2=1}{\operatorname{argmax}} G(\theta)^T\mathbb{E}[yT(\mathbf{x})]G(\theta)$$

Substituting the form of $T(\mathbf{x})$ in (5), the above optimization is equivalent to the following:

$$\theta_0 \in \underset{\theta\in\mathcal{B}^k(r):\ \|G(\theta)\|_2=1}{\operatorname{argmax}} G(\theta)^T\mathbb{E}\big[yS_p(\mathbf{x})S_p(\mathbf{x})^T\big]G(\theta).$$

Under Assumption 3.1, similar to the previous first order links, we replace $\mathbb{E}[yS_p(\mathbf{x})S_p(\mathbf{x})]$ by an empirical version with y being thresholded, i.e.

$$\hat{\theta} \in \underset{\theta\in\mathcal{B}^k(r):\ \|G(\theta)\|_2=1}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^m \widetilde{y}_i|\langle G(\theta), S_p(\mathbf{x}_i)\rangle|^2, \quad (11)$$

where $\widetilde{y}_i := \operatorname{sign}(y_i)\cdot|y_i|\wedge\tau$. Note that by assumption $\theta^*\in\mathcal{B}^k(r)$ and $\|G(\theta^*)\|_2=1$, the above feasible set is not empty. The following theorem characterizes the statistical rate of the proposed estimator.

Theorem 4.1. *Let $\delta\in(0,1)$, $\eta,\beta\geq 2$ be any constants. Suppose $m\geq k\log(4Lr)$, the truncation level $\tau=(k\log(4Lr)/m)^{-1/2(1+\kappa)}\sigma_y$, where $\sigma_y\geq\|y\|_{L_q}$, $\kappa\in(0,q/4-1)$ are any chosen constants, and $\lambda_1=\mathbb{E}\big[f''\big(\langle\mathbf{x},G(\theta^*)\rangle,\ \xi\big)\big]>0$, then, with probability at least $1-e^{-\eta}-e^{-\beta}$, the solution to (11) satisfies*

$$\|G(\widehat{\theta})G(\widehat{\theta})^T-G(\theta^*)G(\theta^*)^T\|_F \leq \overline{C}\sqrt{\frac{\eta+k\log(4Lr/\delta)}{m}}+\sqrt{\overline{C}\delta}\Big(\frac{\eta+k\log(4Lr/\delta)}{m}\Big)^{1/4},$$

where $\overline{C}=C\frac{1+\kappa}{\kappa}\frac{(\|S_p(\mathbf{x})\|_{\psi_2}+\|S_p(\mathbf{x})\|_{\psi_2}^2)\sigma_y\sqrt{\beta}}{\lambda_1}$ and $C>0$ is an absolute constant. Moreover, taking $\delta=\varepsilon$ for some accuracy level $\varepsilon\in(0,1)$ and the number of samples $m=\mathcal{O}\left(k\varepsilon^{-2}\log(4Lr/\varepsilon)\right)$ gives $\|G(\widehat{\theta})-G(\theta^*)\|_2\leq\varepsilon$.

4.2. Theoretical results for ReLU generative model

In this section, we consider the special case when $G(\cdot)$ is a multilayer ReLU function (9). We propose to estimate $G(\theta^*)$ from the following optimization problem:

$$\hat{\theta} \in \underset{\|G(\theta)\|_2=1}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^m \widetilde{y}_i|\langle G(\theta), S_p(\mathbf{x}_i)\rangle|^2, \quad (12)$$

which is the same as (11) except that we drop the constraint $\theta\in\mathcal{B}^k(r)$. The following theorem provides the statistical convergence rate.

Theorem 4.2. *Let $\eta,\beta\geq 2$ be any constants. Suppose $m\geq kn\log(2d)$, the truncation level $\tau=(kn\log(2d)/m)^{-1/2(1+\kappa)}\sigma_y$, where $\sigma_y\geq\|y\|_{L_q}$, $\kappa\in(0,q/4-1)$ are chosen constants, and $\lambda_1=\mathbb{E}\big[f''\big(\langle\mathbf{x},G(\theta^*)\rangle,\ \xi\big)\big]>0$, then, with probability at least $1-e^{-\eta}-e^{-\beta}$, the solution to (12) satisfies*

$$\|G(\widehat{\theta})G(\widehat{\theta})^T-G(\theta^*)G(\theta^*)^T\|_F\leq\overline{C}\sqrt{\frac{\eta+kn\log(2d)}{m}}$$

where $\overline{C}=C\frac{1+\kappa}{\kappa}\frac{(\|S_p(\mathbf{x})\|_{\psi_2}+\|S_p(\mathbf{x})\|_{\psi_2}^2)\sigma_y\sqrt{\beta}}{\lambda_1}$ and $C>0$ is absolute constant. In particular, taking the number of samples $m=\mathcal{O}(kn\varepsilon^{-2}\log d)$ for some accuracy level $\varepsilon\in(0,1)$ gives $\|G(\widehat{\theta})-G(\theta^*)\|_2\leq\varepsilon$.

Remark 4.1. *Note that by a change of variable $\mathbf{b}=G(\theta)$, the method (11) can be rewritten as $\max_{\mathbf{b}\in G(\mathcal{B}^k(r))\cap\mathcal{S}^{d-1}} m^{-1}\sum_{i=1}^m\widetilde{y}_i|\langle\mathbf{b},S_p(\mathbf{x}_i)\rangle|^2$. This is related to the sparse PCA over ℓ_q-norm ball ($q\leq 1$) considered in earlier works, e.g. (Jolliffe et al., 2003; Witten et al., 2009; Vu & Lei, 2012). More specifically, given a random vector $X\in\mathbb{R}^d$ with zero mean and covariance matrix $\mathbf{\Sigma}:=\mathbb{E}\big[XX^T\big]$, the sparse PCA problem focuses on estimating the principle eigenvector of $\mathbf{\Sigma}$, under the assumption that it is approximately sparse and stays in an ℓ_q ball ($q\leq 1$): $\mathcal{B}_q^d(r):=\{\mathbf{b}\in\mathbb{R}^d:\ \sum_{i=1}^d|b_i|^q\leq r\}$. Then, the aforementioned works propose to estimate the principle eigenvector from m samples $X_1,\ \cdots,\ X_m$ by solving the following problem: $\max_{\mathbf{b}\in\mathcal{B}_q^d(r)\cap\mathcal{S}^{d-1}} m^{-1}\sum_{i=1}^m|\langle\mathbf{b},X_i\rangle|^2$. In particular, it is known that the solution to the non-convex $q=0$ case achieves the min-max optimal statistical rate $m=\mathcal{O}(k\varepsilon^{-2}\log d)$ for sparse PCA. Our method replaces the sparse inducing ℓ_q-ball with a generative model $G(\mathcal{B}^k(r))$, matching the aforementioned statistical rate when $G(\cdot)$ is a ReLU generative function.*

Remark 4.2. *Note that computing (11) or (12) is challenging mainly because of the constraint $\|G(\theta)\|_2=1$. A somewhat promising way of solving (12) for $G(\theta^*)$ directly is the Rayleigh flow method (Tan et al., 2018), which avoids explicitly treating the constraint and has been shown to converge for sparse PCA under a proper initialization. Adapting it to our scenario, one could iteratively take the gradient ascent step on the "Rayleigh quotient" $m^{-1}\sum_{i=1}^m\widetilde{y}_i|\langle G(\theta),S_p(\mathbf{x}_i)\rangle|^2/\|G(\theta)\|_2^2$, push forward the update through G and then normalize it. It is interesting to see if such a method can still converge in the presence of generative models. We leave it for future works.*

5. Proofs of main results

In this section, we provide details on the proofs of Theorem 3.1 and 4.1. The proofs of Theorem 3.2 and 4.2 build upon the previous two results by further taking into account the piecewise linearity of the ReLU function. For simplicity of presentation, we delay the proofs of them as well as other concentration results to the supplementary.

5.1. Proof of Theorem 3.1

We start by defining the following empirical average $\mathbb{E}_m[\widetilde{y}\langle S_p(\mathbf{x}),\theta\rangle] := m^{-1}\sum_{i=1}^m \widetilde{y}_i\langle S_p(\mathbf{x}_i),\theta\rangle$, and letting

$$\begin{aligned}\widetilde{L}(G(\theta)) &:= \|G(\theta)\|_2^2 - 2\mathbb{E}_m[\widetilde{y}\langle S_p(\mathbf{x}), G(\theta)\rangle],\\ \widetilde{L}^0(G(\theta)) &:= \|G(\theta)\|_2^2 - 2\mathbb{E}[\widetilde{y}\langle S_p(\mathbf{x}), G(\theta)\rangle].\\ L^0(G(\theta)) &:= \|G(\theta)\|_2^2 - 2\mathbb{E}[y\langle S_p(\mathbf{x}), G(\theta)\rangle].\end{aligned}$$

By the first order Stein's identity (3), we have $\mathbb{E}[y\langle S_p(\mathbf{x}), G(\theta)\rangle] = \lambda\langle G(\theta^*), G(\theta)\rangle$, where $\lambda = \mathbb{E}\big[f'\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big]$. Thus, we have for any $\theta \in \mathcal{B}^k(r)$,

$$\begin{aligned}&L^0(G(\theta)) - L^0(\lambda G(\theta^*))\\ =&\|G(\theta)\|_2^2 - 2\mathbb{E}[y\langle S_p(\mathbf{x}), G(\theta) - \lambda G(\theta^*)\rangle] - \lambda^2\|G(\theta^*)\|_2^2\\ =&\|G(\theta)\|_2^2 - \lambda^2\|G(\theta^*)\|_2^2 - 2\lambda\langle G(\theta^*), G(\theta) - \lambda G(\theta^*)\rangle\\ =&\|G(\theta) - \lambda G(\theta^*)\|_2^2.\end{aligned}$$

In particular, for an empirical minimizer $\widehat{\theta} \in \mathcal{B}^k(r)$,

$$\begin{aligned}&\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2^2\\ =&L^0(G(\widehat{\theta})) - L^0(\lambda G(\theta^*))\\ =&L^0(G(\widehat{\theta})) - \widetilde{L}^0(G(\widehat{\theta})) + \widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*))\\ &+ \widetilde{L}^0(\lambda G(\theta^*)) - L^0(\lambda G(\theta^*))\\ \le&|L^0\big(G(\widehat{\theta}) - \lambda G(\theta^*)\big) - \widetilde{L}^0\big(G(\widehat{\theta}) - \lambda G(\theta^*)\big)|\\ &+ |\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*))|.\end{aligned}$$

Note that the bias term

$$\begin{aligned}&|L^0\big(G(\widehat{\theta}) - \lambda G(\theta^*)\big) - \widetilde{L}^0\big(G(\widehat{\theta}) - \lambda G(\theta^*)\big)|\\ =&\left|\mathbb{E}\Big[(y-\widetilde{y})\cdot\langle S_p(\mathbf{x}), G(\widehat{\theta}) - \lambda G(\theta^*)\rangle\Big]\right|.\end{aligned}$$

Thus, we have the following bias-variance decomposition:

$$\begin{aligned}&\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2^2\\ &\quad\le \left|\mathbb{E}\Big[(y-\widetilde{y})\cdot\langle S_p(\mathbf{x}), G(\widehat{\theta}) - \lambda G(\theta^*)\rangle\Big]\right|\\ &\qquad + |\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*))|. \quad (13)\end{aligned}$$

The following lemma, whose proof can be found in Section 9 of the supplementary, gives a bound for the first term in (13) under a certain truncation level τ:

Lemma 5.1. *Suppose $\tau = m^{1/2(1+\kappa)}\sigma_y$, where $\sigma_y \ge \|y\|_{L_q}$, $\kappa \in (0, q/4-1)$ are chosen constants. Let $\overline{C}_{S,\sigma,\kappa} = \sigma_y\|S_p(\mathbf{x})\|_{\psi_2}\sqrt{(1+\kappa)/\kappa}$. Then, for any $t \in \mathbb{R}^d$,*

$$|\mathbb{E}[y\langle S_p(\mathbf{x}), t\rangle - \widetilde{y}\langle S_p(\mathbf{x}), t\rangle]| \le \overline{C}_{S,\sigma,\kappa}m^{-1/2}\|t\|_2.$$

In view of (13), it remains to bound the term $|\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*))|$. First of all, we can further decompose the term into differences of empirical averages:

$$\begin{aligned}&\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*)) = \underbrace{\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}(G(\widehat{\theta}))}_{(I)}\\ &+ \underbrace{\widetilde{L}(G(\widehat{\theta})) - \widetilde{L}(\lambda G(\theta^*))}_{(II)} + \underbrace{\widetilde{L}(\lambda G(\theta^*)) - \widetilde{L}^0(\lambda G(\theta^*))}_{(III)}\end{aligned}$$

Note that by the first order link condition, $\exists\overline{\theta} \in \mathcal{B}^k(r)$ such that $\lambda G(\theta^*) = G(\overline{\theta})$. Since $\widehat{\theta}$ minimizes the empirical risk $\widetilde{L}(G(\theta))$ over all $\theta \in \mathcal{B}^k(r)$, we have $(II) \le 0$. Substituting the definitions of $\widetilde{L}(\cdot)$ and $\widetilde{L}^0(\cdot)$ gives

$$\begin{aligned}&|\widetilde{L}^0(G(\widehat{\theta})) - \widetilde{L}^0(\lambda G(\theta^*))|\\ \le&\left|\langle\mathbb{E}[\widetilde{y}S_p(\mathbf{x})] - \mathbb{E}_m[\widetilde{y}S_p(\mathbf{x})], G(\widehat{\theta}) - \lambda G(\theta^*)\rangle\right|\\ =&\left|\langle\mathbb{E}[\widetilde{y}S_p(\mathbf{x})] - \mathbb{E}_m[\widetilde{y}S_p(\mathbf{x})], G(\widehat{\theta}) - G(\overline{\theta})\rangle\right|, \quad (14)\end{aligned}$$

where the last equality follows again from the first order link condition $\exists\overline{\theta} \in \mathcal{B}^k(r)$ such that $\lambda G(\theta^*) = G(\overline{\theta})$.

Using the Lipschitz structure of $G(\mathcal{B}^k(r))$, we establish the following key uniform concentration lemma in Section 9 of the supplementary regarding the above term via a new chaining argument on a *heavy-tailed multiplier process*.

Lemma 5.2. *Let $\delta \in (0, r)$, $\eta, \beta \ge 2$ be constants. Suppose $\tau = m^{1/2(1+\kappa)}\sigma_y$, where $\sigma_y \ge \|y\|_{L_q}$, $\kappa \in (0, q/4-1)$ are any chosen constants. Then, with probability at least $1 - e^{-\eta} - e^{-\beta}$, for any $t, t' \in G(\mathcal{B}^k(r))$, we have*

$$\begin{aligned}&\left|\frac{1}{m}\sum_{i=1}^m \widetilde{y}_i\langle S_p(\mathbf{x}_i), t - t'\rangle - \mathbb{E}[\widetilde{y}\langle S_p(\mathbf{x}), t - t'\rangle]\right|\\ &\qquad\le \overline{C}'_{S,\sigma,\beta,\kappa}\sqrt{\frac{\eta + k\log(4Lr/\delta)}{m}}\left(\|t - t'\|_2 + \delta\right),\end{aligned}$$

where $\overline{C}'_{S,\sigma,\beta,\kappa} = C\|S_p(\mathbf{x})\|_{\psi_2}\sigma_y\sqrt{\beta(1+\kappa)/\kappa}$ and $C > 0$ is an absolute constant.

Substituting Lemma 5.1 with $t = G(\hat{\theta}) - \lambda G(\theta^*)$ and Lemma 5.2 with $t = G(\hat{\theta})$, $t' = \lambda G(\theta^*) = G(\overline{\theta})$ into (13) gives with probability at least $1 - e^{-\beta} - e^{-\eta}$,

$$\begin{aligned}&\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2^2 \le (\overline{C}'_{S,\sigma,\beta,\kappa} + \overline{C}_{S,\sigma,\kappa})\\ &\quad\cdot\sqrt{\frac{\eta + k\log(4Lr/\delta)}{m}}\left(\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 + \delta\right).\end{aligned}$$

Define $a_m = \overline{C}\sqrt{(\eta + k\log(4Lr/\delta))/m}$, and $b_m = \overline{C}\delta\sqrt{(\eta + k\log(4Lr/\delta))/m}$, where $\overline{C} = (\overline{C}'_{S,\sigma,\beta,\kappa} + \overline{C}_{S,\delta,\kappa})$. Then, we have

$$\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2^2 \le a_m\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 + b_m.$$

Solving this quadratic inequality gives

$$\|G(\widehat{\theta}) - \lambda G(\theta^*)\|_2 \leq \frac{a_m + \sqrt{a_m^2 + 4b_m}}{2} \leq a_m + \sqrt{b_m},$$

which implies the claim.

5.2. Proof of Theorem 4.1

By Stein's identity (4), we have

$$\mathbf{\Sigma} := \mathbb{E}[yT(\mathbf{x})] = \mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] G(\theta^*)G(\theta^*)^T,$$

where $\lambda_1 := \mathbb{E}\big[f''\big(\langle \mathbf{x}, G(\theta^*)\rangle,\ \xi\big)\big] > 0$ by the second order link condition. Next, recall that $T(\mathbf{x})$ can be written as (5). Thus, we define

$$\mathbf{S} = \frac{1}{m}\sum_{i=1}^m \tilde{y}_i S_p(\mathbf{x}_i) S_p(\mathbf{x}_i)^T - \mathbb{E}[y \cdot \mathrm{diag}(s_0' \circ (\mathbf{x}))], \tag{15}$$

$$\mathbf{S}_0 = \mathbb{E}\big[\tilde{y} S_p(\mathbf{x}) S_p(\mathbf{x})^T\big] - \mathbb{E}[y \cdot \mathrm{diag}(s_0' \circ (\mathbf{x}))]. \tag{16}$$

To proceed, we need the following bound:

Lemma 5.3 (Lemma 3.2.1 of (Vu & Lei, 2012))**.** *Let $\mathbf{v} \in \mathcal{S}^{d-1}$. If the matrix $\mathbf{\Sigma}$ is positive semi-definite and has a unique largest eigenvalue λ_1 with the corresponding eigenvector $\mathbf{v}_1$, then,*

$$\frac{1}{2}(\lambda_1 - \lambda_2)\|\mathbf{v}\mathbf{v}^T - \mathbf{v}_1\mathbf{v}_1^T\|_F^2 \leq \langle \mathbf{\Sigma}, \mathbf{v}_1\mathbf{v}_1^T - \mathbf{v}\mathbf{v}^T\rangle.$$

Then, we have the following chain of inequalities:

$$\begin{aligned}
&\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F^2 \\
\leq&\langle \mathbf{\Sigma}, G(\theta^*)G(\theta^*)^T - G(\hat{\theta})G(\hat{\theta})^T\rangle \\
=&\langle \mathbf{S}, G(\theta^*)G(\theta^*)^T\rangle - \langle \mathbf{\Sigma}, G(\hat{\theta})G(\hat{\theta})^T\rangle \\
&- \langle \mathbf{S} - \mathbf{\Sigma}, G(\theta^*)G(\theta^*)^T\rangle \\
\leq&\langle \mathbf{S} - \mathbf{\Sigma}, G(\hat{\theta})G(\hat{\theta})^T\rangle - \langle \mathbf{S} - \mathbf{\Sigma}, G(\theta^*)G(\theta^*)^T\rangle \\
=&\langle \mathbf{S} - \mathbf{\Sigma}, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle,
\end{aligned}$$

where the first inequality follows from Lemma 5.3 and the second inequality follows from the fact that $G(\hat{\theta})$ is the solution to (11) and thus, it is also the eigenvector corresponding to the largest eigenvalue of $\mathbf{S}$. Now, we decompose the above error into bias and variance:

$$\begin{aligned}
&\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F^2 \\
\leq&|\langle \mathbf{S} - \mathbf{\Sigma}, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle| \\
\leq&|\langle \mathbf{S} - \mathbf{S}_0, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle| \\
&+ |\langle \mathbf{S}_0 - \mathbf{\Sigma}, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle|
\end{aligned} \tag{17}$$

Note that the second term, which is the bias, is equal to

$$\Big|\mathbb{E}\Big[(\tilde{y} - y)\left(|\langle G(\hat{\theta}), S_p(\mathbf{x})\rangle|^2 - |\langle G(\theta^*), S_p(\mathbf{x})\rangle|^2\right)\Big]\Big|.$$

The following lemma bounds this term, whose proof can be found in Section 10 of the supplementary.

Lemma 5.4. *Suppose $\tau = \big(k\log(Lr)/m\big)^{-1/2(1+\kappa)}\sigma_y$, where $\sigma_y \geq \|y\|_{L_q}$, $\kappa \in (0, q/4-1)$ are any chosen constants. Then, for any $t, t' \in \mathcal{S}^{d-1}$,*

$$\begin{aligned}
&\big|\mathbb{E}\big[(\tilde{y} - y)\left(|\langle t, S_p(\mathbf{x})\rangle|^2 - |\langle t', S_p(\mathbf{x})\rangle|^2\right)\big]\big| \leq \\
&\qquad \overline{C}_{S,\sigma,\kappa}\sqrt{\frac{k\log(Lr)}{m}}\|t - t'\|_2.
\end{aligned}$$

where $\overline{C}_{S,\sigma,\kappa} = \sigma_y\|S_p(\mathbf{x})\|_{\psi_2}^2\sqrt{4(1+\kappa)/\kappa}$ and C is an absolute constant.

Now, for any $t, t' \in \mathbb{R}^d$, we define

$$\begin{aligned}
D(t,t') = \Big|&\frac{1}{m}\sum_{i=1}^m \tilde{y}_i\left(|\langle t, S_p(\mathbf{x}_i)\rangle|^2 - |\langle t', S_p(\mathbf{x}_i)\rangle|^2\right) \\
&- \mathbb{E}\big[\tilde{y}\left(|\langle t, S_p(\mathbf{x})\rangle|^2 - |\langle t', S_p(\mathbf{x})\rangle|^2\right)\big]\Big|.
\end{aligned}$$

Then, for the first term in (17), we have

$$\begin{aligned}
&|\langle \mathbf{S} - \mathbf{S}_0, G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\rangle| \\
&\qquad = D(G(\hat{\theta}), G(\theta^*)),
\end{aligned}$$

We establish the following key uniform concentration lemma in Section 10 of the supplementary regarding the above term via a new chaining argument on a *heavy-tailed quadratic process*:

Lemma 5.5. *Let $\delta \in (0,1)$, $\eta, \beta \geq 2$ be constants. Suppose $\tau = \big(k\log(Lr)/m\big)^{-1/2(1+\kappa)}\sigma_y$, where $\sigma_y \geq \|y\|_{L_q}$, $\kappa \in (0, q/4-1)$ are any chosen constants. Then, with probability at least $1 - e^{-\eta} - e^{-\beta}$, for any $t, t' \in G(\mathcal{B}^k(r)) \cap \mathcal{S}^{d-1}$,*

$$D(t,t') \leq \overline{C}'_{S,\sigma,\beta,\kappa}\sqrt{\frac{\eta + k\log(4Lr/\delta)}{m}}\left(\|t - t'\|_2 + \delta\right),$$

where $\overline{C}'_{S,\sigma,\beta,\kappa} = C(\|S_p(\mathbf{x})\|_{\psi_2} + \|S_p(\mathbf{x})\|_{\psi_2}^2)\sigma_y\sqrt{\beta}(1+\kappa)/\kappa$ and C is an absolute constant.

Substituting the previous two lemmas into (17), we have

$$\begin{aligned}
&\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F^2 \\
&\quad \leq (\overline{C}'_{S,\sigma,\beta,\kappa} + \overline{C}_{S,\sigma,\kappa}) \cdot \sqrt{\frac{\eta + k\log(4Lr/\delta)}{m}} \\
&\qquad \cdot \Big(\|G(\hat{\theta}) - \lambda G(\theta^*)\|_2 + \delta\Big).
\end{aligned} \tag{18}$$

Finally, note that since $G(\hat{\theta}),\ G(\theta^*) \in \mathcal{S}^{d-1}$, it follows from a linear algebraic fact (Lemma 11.7 in the Supplementary) that

$$\|G(\hat{\theta}) - G(\theta^*)\|_2 \leq \|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F. \tag{19}$$

Define $a_m = \overline{C}\sqrt{(\eta + k\log(4Lr/\delta))/m}$, and $b_m = \overline{C}\delta\sqrt{(\eta + k\log(4Lr/\delta))/m}$, where $\overline{C} = (\overline{C}'_{S,\sigma,\beta,\kappa} + \overline{C}_{S,\delta,\kappa})$. Then, (18), together with (19), implies

$$\frac{1}{2}\lambda_1\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F^2 \leq a_m\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F + b_m.$$

Solving this quadratic inequality gives

$$\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F \leq 2\big(a_m + \sqrt{a_m^2 + 2\lambda_1 b_m}\big)/\lambda_1 \leq 4a_m/\lambda_1 + 2\sqrt{2b_m/\lambda_1},$$

which proves the first part of the Theorem. Furthermore, when m is large enough such that $m = \mathcal{O}(\frac{k\log(Lr/\varepsilon)}{\varepsilon^2})$ for some $\varepsilon < 1$, we have $\|G(\hat{\theta})G(\hat{\theta})^T - G(\theta^*)G(\theta^*)^T\|_F \leq \varepsilon$, and by Lemma 11.7 again, $\|G(\hat{\theta}) - G(\theta^*)\|_2 \leq \varepsilon$.

6. Conclusions

In this paper, we propose and analyze two types of new estimators for nonlinear recovery with a generative model when the measurements are non-Gaussian and the responses are heavy-tailed. The estimators are constructed via first and second order Stein's identity, respectively, and shown theoretically to achieve an ε estimation error with a sample size bound $m = \mathcal{O}\left(k\varepsilon^{-2}\log(L/\varepsilon)\right)$. Furthermore, when the generative model is a multilayer ReLU function, the sample bound of the aforementioned two estimators can be improved to $m = \mathcal{O}\left(k\varepsilon^{-2}\log(d)\right)$, matching the state-of-art statistical rates of sparse recovery and sparse PCA.

References

Ai, A., Lapanowski, A., Plan, Y., and Vershynin, R. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Babichev, D., Bach, F., et al. Slice inverse regression with score functions. *Electronic Journal of Statistics*, 12(1): 1507–1543, 2018.

Bora, A., Jalal, A., Price, E., and Dimakis, A. G. Compressed sensing using generative models. *arXiv preprint arXiv:1703.03208*, 2017.

Brillinger, D. R. A generalized linear model with "Gaussian" regressor variables. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pp. 97–114. Wadsworth, Belmont, CA, 1983.

Candes, E. J., Romberg, J. K., and Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59 (8):1207–1223, 2006.

Dirksen, S. et al. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015.

Goldstein, L. and Wei, X. Non-gaussian observations in nonlinear compressed sensing via stein discrepancies. *Information and Inference: A Journal of the IMA*, 2016.

Goldstein, L., Minsker, S., and Wei, X. Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Transactions on Information Theory*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., and Knoll, F. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Han, Y., Özgür, A., and Weissman, T. Geometric lower bounds for distributed parameter estimation under communication constraints. *arXiv preprint arXiv:1802.08417*, 2018.

Hand, P. and Voroninski, V. Global guarantees for enforcing deep generative priors by empirical risk. *arXiv preprint arXiv:1705.07576*, 2017.

Hand, P., Leong, O., and Voroninski, V. Phase retrieval under a generative prior. In *Advances in Neural Information Processing Systems*, pp. 9154–9164, 2018.

Hastie, T., Tibshirani, R., and Wainwright, M. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Huang, W., Hand, P., Heckel, R., and Voroninski, V. A provably convergent scheme for compressive sensing under random generative priors. *arXiv preprint arXiv:1812.04176*, 2018.

Janzamin, M., Sedghi, H., and Anandkumar, A. Score function features for discriminative learning: Matrix and tensor framework. *arXiv preprint arXiv:1412.2863*, 2014.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3): 531–547, 2003.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lecué, G. and Mendelson, S. Sparse recovery under weak moment assumptions. *arXiv preprint arXiv:1401.2188*, 2014.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114. IEEE, 2017.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Li, K.-C. and Duan, N. Regression analysis under link violation. *The Annals of Statistics*, pp. 1009–1052, 1989.

Montgomery-Smith, S. J. The distribution of rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

Plan, Y. and Vershynin, R. The generalized lasso with nonlinear observations. *IEEE Transactions on information theory*, 62(3):1528–1537, 2016.

Plan, Y., Vershynin, R., and Yudovina, E. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Stein, C., Diaconis, P., Holmes, S., Reinert, G., et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pp. 1–25. Institute of Mathematical Statistics, 2004.

Tan, K. M., Wang, Z., Liu, H., and Zhang, T. Sparse generalized eigenvalue problem: Optimal statistical rates via truncated rayleigh flow. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5): 1057–1086, 2018.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Vu, V. and Lei, J. Minimax rates of estimation for sparse pca in high dimensions. In *Artificial Intelligence and Statistics*, pp. 1278–1286, 2012.

Wei, X. Structured recovery with heavy-tailed measurements: A thresholding procedure and optimal rates. *arXiv preprint arXiv:1804.05959*, 2018.

Winder, R. Partitions of n-space by hyperplanes. *SIAM Journal on Applied Mathematics*, 14(4):811–818, 1966.

Witten, D. M., Tibshirani, R., and Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., Liu, F., Arridge, S., Keegan, J., Guo, Y., et al. Dagan: Deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2018.

Yang, Z., Balasubramanian, K., and Liu, H. On stein's identity and near-optimal estimation in high-dimensional index models. *arXiv preprint arXiv:1709.08795*, 2017a.

Yang, Z., Balasubramanian, K., Wang, Z., and Liu, H. Learning non-gaussian multi-index model via second-order stein's method. *Advances in Neural Information Processing Systems*, 2017b.

Yang, Z., Yang, L. F., Fang, E. X., Zhao, T., Wang, Z., and Neykov, M. Misspecified nonconvex statistical optimization for phase retrieval. *arXiv preprint arXiv:1712.06245*, 2017c.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5485–5493, 2017.