Quickly Finding the Best Linear Model in High Dimensions

Yahya Sattar Samet Oymak

***Abstract*—We study the problem of finding the best linear model that can minimize least-squares loss given a dataset. While this problem is trivial in the low-dimensional regime, it becomes more interesting in high-dimensions where the population minimizer is assumed to lie on a manifold such as sparse vectors. We propose projected gradient descent (PGD) algorithm to estimate the population minimizer in the finite sample regime. We establish linear convergence rate and data-dependent estimation error bounds for PGD. Our contributions include: 1) The results are established for heavier tailed sub-exponential distributions besides sub-gaussian. 2) We directly analyze the empirical risk minimization and do not require a realizable model that connects input data and labels. 3) Our PGD algorithm is augmented to learn the bias terms which boosts the performance. The numerical experiments validate our theoretical results.**

***Index Terms*—high-dimensional estimation, projected gradient descent, one-bit compressed sensing, gaussian width.**

I. Introduction

Supervised learning is concerned with finding a relation between the input-output pairs $(\boldsymbol{x}_i, y_i)_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$. The simplest relations are linear functions where the output y_i is estimated by a linear function of the input, that is, $\hat{y}_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle$. Using quadratic loss, we can find the optimal $\boldsymbol{\theta}$ with a simple linear regression which minimizes.

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle)^2.$$

If the samples are i.i.d. and input has identity covariance, the population minimizer ($n \to \infty$) is simply given by

$$\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta})] = \mathbb{E}[y\boldsymbol{x}].$$

where $(\boldsymbol{x}, y)$ is drawn from same distribution as data. In many applications, we operate in the high-dimensional regime where we have fewer samples than the parameter dimension i.e. $n \ll p$. In this case, the problem is ill-posed; however, if $\boldsymbol{\theta}^\star$ lies on a low-dimensional manifold, we can take advantage of this information to solve the problem. We assume $\boldsymbol{\theta}^\star$ is structured-sparse, for instance, it can be a signal that is sparse in a dictionary or it can be a low-rank matrix. If $\mathcal{R}$ is a regularization function that promotes this structure, we can solve the regularized empirical risk minimization (ERM)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{\theta}) \le R. \quad (1)$$

where $\boldsymbol{y} = [y_1 \ \dots \ y_n]^T \in \mathbb{R}^n$ and $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times p}$ are the output labels and data matrix respectively. This problem is well-studied in the statistics and compressed sensing (CS) literature. However, much of the theory literature is concerned with the scenario where the problem is realizable i.e. the outputs are explicitly generated with respect to some ground truth vector $\boldsymbol{a}$. In the simplest scenario, input/output relation can be $y = \langle \boldsymbol{x}, \boldsymbol{a} \rangle + \boldsymbol{z}$ where $\boldsymbol{z}$ is independent zero-mean noise vector. In this case, one simply has $\boldsymbol{\theta}^\star = \boldsymbol{a}$. Such realizability assumption is also common in the single-index models [1], [2]. One contribution of this paper will be analyzing regularized ERM without the realizability assumption.

Bias in the data can negatively affect the estimation quality. Assuming input is zero-mean, instead of solving (1) we can solve a modified problem which accounts for the mean of the output as well. Again, denoting the regularization function by $\mathcal{R}$, we will solve the modified problem

$$\hat{\boldsymbol{\theta}}, \hat{\mu} = \arg\min_{\boldsymbol{\theta},\mu} \mathcal{L}(\boldsymbol{\theta}, \mu) \text{ subject to } \mathcal{R}(\boldsymbol{\theta}) \le R. \quad (2)$$

where the loss is given by $\mathcal{L}(\boldsymbol{\theta}, \mu) = \frac{1}{2}\left\|\boldsymbol{y} - [\boldsymbol{X}\ \boldsymbol{1}]\begin{bmatrix}\boldsymbol{\theta}\\ \mu\end{bmatrix}\right\|_{\ell_2}^2$. We will show that solving problem (2) is essentially equivalent to solving (1) with debiased output hence it will result in more accurate estimation. The goal of this paper is studying problem (2) under a general algorithmic framework, establishing finite-sample statistical and algorithmic convergence, and addressing practical considerations on the data distribution. In particular, we are interested in how well one can estimate the best linear model (BLM) given by the pair $(\boldsymbol{\theta}^\star = \mathbb{E}[y\boldsymbol{x}], \mu^\star = \mathbb{E}[y])$.

Department of Electrical and Computer Engineering, University of California, Riverside, CA 92521, USA. Email: ysatt001@ucr.edu, oymak@ece.ucr.edu.

For estimation, we will utilize the projected gradient descent algorithm given by the iterates

$$\begin{aligned}\boldsymbol{\theta}_{\tau+1} &= \mathcal{P}_{\mathcal{K}}(\boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_\tau,\mu_\tau)),\\ \mu_{\tau+1} &= \mu_\tau - \eta\nabla\mathcal{L}_{\mu}(\boldsymbol{\theta}_\tau,\mu_\tau),\end{aligned} \tag{3}$$

where $\mathcal{P}_{\mathcal{K}}$ projects onto the constraint set $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p \;\big|\; \mathcal{R}(\boldsymbol{\theta}) \le R\}$ and η is the step size.

A. Relation to Prior Work

There is a significant amount of literature on nonlinear (or one-bit) CS [2]–[12]. [4], [13]–[16] study algorithmic and statistical convergence rates for first order methods such as projected/proximal gradient descent. For nonlinear CS, [4], [5], [7], [17] provide statistical analysis of single index estimation with a focus on Gaussian data. Recently, one-bit CS techniques have been extended to sub-gaussian distributions using dithering trick which adds noise before quantization [18]–[21]. Dithering is introduced to guarantee consistent estimation of the ground-truth parameter. The papers [22]–[26] address non-gaussianity by utilizing Stein identity which requires access to the distribution of the input samples. Closer to us [27] studies the constrained empirical risk minimization with linear functions and squared loss with a focus on convex problems. In comparison our analysis applies to a broader class of distributions and focus on first order algorithms. Much of our analysis focuses on addressing subexponential samples, which requires tools from high-dimensional probability [28], [29].

Our results apply to general regularizers and borrow ideas from [4]–[7]. Similar to these, we view the nonlinearity between input and output as an additive noise. The convergence analysis of projected gradient descent is a rather well-understood topic and we utilize insights from [13]–[16] for our analysis.

B. Contributions

At a high-level our work has three distinguishing features compared to the prior literature.

- Subexponential samples: Most nonlinear CS results apply to Gaussian or subgaussian data when dithering trick is utilized [18]–[21]. We take advantage of the recent techniques for subexponential distributions to provide statistical/computational guarantees for heavier-tailed distributions.
- No realizability assumption: Nonlinear CS literature is typically concerned with a ground-truth vector to be recovered. For instance, one-bit CS aims to learn $\boldsymbol{\theta}$ from samples of type $y = \mathrm{sgn}(\boldsymbol{\theta}^T\boldsymbol{x})$. Unlike these, we do not enforce such relationship to exist between input and output, hence the results apply under much weaker assumptions. Instead of a ground-truth $\boldsymbol{\theta}$, we work with the population BLM $\boldsymbol{\theta}^\star$. However, $\boldsymbol{\theta}^\star$ can be shown to coincide with ground truth when it exists, if the input distribution is *nice* (e.g. Gaussian) [4]–[7].
- Bias estimation: Our analysis addresses the bias in the output by solving the modified problem (2). We show that (2) can be studied in a similar fashion to (1) by studying the statistical properties of the concatenated data matrix. However, empirically this modification results in a substantial improvement in estimation.

C. Paper Organization

We review mathematical background and formulate the problem in Section II. We introduce our main results on statistical and computational convergence guarantees in Section III. Section IV provides numerical experiments to corroborate our theoretical results. Proofs of the main results are provided in Section V and finally the concluding remarks are made in Section VI.

II. Preliminaries and Problem Formulation

In this section we introduce statistical quantities which are utilized to characterize the benefits of the regularization $\mathcal{R}$.

We first set the notation. $c, c_0, \ldots, C$ denote positive absolute constants. For a vector $\boldsymbol{v}$, we denote its Euclidean norm by $\|\boldsymbol{v}\|_{\ell_2}$. Similarly for a matrix $\boldsymbol{X}$, we denote its spectral norm by $\|\boldsymbol{X}\|$. Given a set S, let $\mathrm{cl}(S)$ and $\mathrm{clconv}(S)$ be the minimal closed set and minimal closed-convex set containing S respectively. Let $\mathrm{rad}(S)$ denote the set radius $\sup_{\boldsymbol{v}\in S}\|\boldsymbol{v}\|_{\ell_2}$. For closed sets, let $\mathcal{P}_S(\cdot)$ be the projection operator defined as $\mathcal{P}_S(\boldsymbol{a}) = \arg\min_{\boldsymbol{v}\in S}\|\boldsymbol{a}-\boldsymbol{v}\|_{\ell_2}$. $\mathcal{N}(\mu,\sigma^2)$ denotes the normal distribution and $\mathcal{B}^p$ denote the unit ball in $\mathbb{R}^p$. $\mathbf{1}$ is the all ones vector of proper dimension. We will use $\gtrsim$ and $\lesssim$ for inequalities that hold up to a constant factor.

Suppose we are given n i.i.d. samples $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$. To keep the exposition clean, we assume that $\boldsymbol{x}$ is whitened, that is, it has zero-mean and identity covariance. We will aim to find a linear relation between the modified input-output pairs $([\boldsymbol{x}_i^T \;\; 1]^T, y_i)_{i=1}^n$. Let us consider the statistical properties of our modified estimate in the population limit which is given by

$$\begin{aligned}\boldsymbol{\theta}^\star, \mu^\star &= \arg\min_{\boldsymbol{\theta},\mu}\mathbb{E}[\mathcal{L}(\boldsymbol{\theta},\mu)]\\ &= \mathbb{E}[y\boldsymbol{x}], \mathbb{E}[y].\end{aligned}$$

Thus, in the limiting case, $\mu^\star$ captures the mean of the output and $\boldsymbol{\theta}^\star$ is the ideal solution of the problem with debiased output. Our goal is estimating the population minimizer $\boldsymbol{\theta}^\star, \mu^\star$; which minimizes the expected quadratic loss $\mathbb{E}[(y - \boldsymbol{\theta}^T\boldsymbol{x} - \mu)^2]$. As discussed in Section I, assuming $\boldsymbol{\theta}^\star$ is structured sparse, we consider

a non-asymptotic estimation of $\boldsymbol{\theta}^\star, \mu^\star$ via problem (2). To proceed with analysis, set

$$\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p \;\big|\; \mathcal{R}(\boldsymbol{\theta}) \le R\}, \tag{4}$$

$$\mathcal{K}_{\text{ext}} = \{[\boldsymbol{\theta}^T \;\; \mu]^T \in \mathbb{R}^{p+1} \;\big|\; \mathcal{R}(\boldsymbol{\theta}) \le R\}. \tag{5}$$

We investigate the PGD algorithm (3) which can be written as

$$\begin{bmatrix}\boldsymbol{\theta}_{\tau+1}\\ \mu_{\tau+1}\end{bmatrix} = \mathcal{P}_{\mathcal{K}_{\text{ext}}}\left(\begin{bmatrix}\boldsymbol{\theta}_{\tau}\\ \mu_{\tau}\end{bmatrix} + \eta[\boldsymbol{X}\;\mathbf{1}]^T\left(\boldsymbol{y} - [\boldsymbol{X}\;\mathbf{1}]\begin{bmatrix}\boldsymbol{\theta}_{\tau}\\ \mu_{\tau}\end{bmatrix}\right)\right), \tag{6}$$

where η is a fixed learning rate and $[\boldsymbol{X}\;\mathbf{1}] \in \mathbb{R}^{n\times(p+1)}$ is the modified data matrix constructed as follows

$$[\boldsymbol{X}\;\mathbf{1}] = \begin{bmatrix}\boldsymbol{x}_1^T & 1\\ \vdots & \\ \boldsymbol{x}_n^T & 1\end{bmatrix}.$$

Following [4], [30] PGD analysis can be related to the *tangent ball* around the population parameter $\boldsymbol{\theta}^\star$ which is given by

$$\mathcal{C} = \text{cl}(\{\alpha\boldsymbol{v} \;\big|\; \boldsymbol{v} + \boldsymbol{\theta}^\star \in \mathcal{K},\; \alpha \ge 0\}) \bigcap \mathcal{B}^p. \tag{7}$$

Similarly, we define the extended tangent ball as follows

$$\mathcal{C}_{\text{ext}} = \left\{\begin{bmatrix}\alpha\boldsymbol{v}\\ \gamma\end{bmatrix} \;\big|\; \alpha \ge 0,\; \boldsymbol{v} \in \mathcal{C},\; \gamma \in \mathbb{R}\right\} \bigcap \mathcal{B}^{p+1}. \tag{8}$$

The two definitions above are closely related. For any vector $\boldsymbol{v} \in \mathcal{C}$, we have that $[\sqrt{1-\gamma^2}\boldsymbol{v}^T \;\gamma]^T \in \mathcal{C}_{\text{ext}}$ for $0 \le \gamma \le 1$. In the following we will express the convergence rates and residual errors of the PGD algorithm (3) in terms of the statistical properties of the tangent balls .

Technical approach: Denoting the parameter estimation error in (6) by $\boldsymbol{h}_\tau = [\boldsymbol{\theta}_\tau{}^T \;\; \mu_\tau]^T - [\boldsymbol{\theta}^{\star T} \;\; \mu^\star]^T$ and the effective noise by $\boldsymbol{w} = \boldsymbol{y} - [\boldsymbol{X}\;\mathbf{1}][\boldsymbol{\theta}^{\star T} \;\; \mu^\star]^T$, the PGD update can be shown to obey [14] (see Eq. (VI.10))

$$\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} \le \kappa\left(\|\boldsymbol{h}_\tau\|_{\ell_2}\rho(\mathcal{C}) + \eta\nu(\mathcal{C})\right) \tag{9}$$

where κ is a numerical constant which is equal to 1 for convex regularizer $\mathcal{R}$ and 2 for arbitrary $\mathcal{R}$ and

$$\rho(\mathcal{C}) = \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}_{\text{ext}}} |\boldsymbol{u}^T(\boldsymbol{I} - \eta[\boldsymbol{X}\;\mathbf{1}]^T[\boldsymbol{X}\;\mathbf{1}])\boldsymbol{v}|, \tag{10}$$

$$\nu(\mathcal{C}) = \sup_{\boldsymbol{v}\in\mathcal{C}_{\text{ext}}} |\boldsymbol{v}^T[\boldsymbol{X}\;\mathbf{1}]^T\boldsymbol{w}|. \tag{11}$$

Here ρ captures the algorithmic convergence and ν captures the statistical accuracy in terms of regularization. To achieve statistical learning bounds, we need to characterize the quantities above in finite sample. Existing literature provides a fairly good understanding of the related terms when $\boldsymbol{X}$ has subgaussian rows or $\boldsymbol{w}$ is independent of $\boldsymbol{X}$. The technical contributions of this work are i) extending these results to subexponential samples, ii) allowing for nonlinear dependencies between the noise and data, and iii) addressing the bias term by studying the concatenated matrix $[\boldsymbol{X}\;\mathbf{1}]$. To proceed with statistical analysis, we introduce Gaussian width.

Definition 2.1 ((Perturbed) Gaussian width [29]): The Gaussian width of a set $S \subset \mathcal{B}^p$ is defined as

$$\omega(S) = \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,\boldsymbol{I}_p)}[\sup_{\boldsymbol{v}\in S} \boldsymbol{v}^T\boldsymbol{g}].$$

Let $C > 0$ be an absolute constant. Given an integer $n \ge 1$, the perturbed Gaussian width $\omega_n(T)$ of $T \subset \mathcal{B}^d$ is defined as

$$\omega_n(T) = \min_{\substack{\text{clconv}(S)\supseteq T\\ \text{rad}(S)\le C}} \omega(S) + \frac{\gamma_1(S)}{\sqrt{n}}$$

where $\gamma_1(S)$ is Talagrand's γ_1-functional (see [28]) with ℓ_2-metric.

Gaussian width helps to quantify the complexity of the regularized problem and determines the sample complexity of the linear inverse problems i.e. high-dimensional problems become manageable in the regime $n \gtrsim \omega^2(\mathcal{C})$ [30], [31]. Perturbed width is introduced more recently in [29] to address subexponential samples. [29] shows that, for standard regularizers such as ℓ_0, ℓ_1, subspace, and rank constraints, we have that

$$\omega^2(\mathcal{C}) \sim \omega_n^2(\mathcal{C}) \tag{12}$$

in the interesting regime $n \ge \omega^2(\mathcal{C})$. Hence, perturbed width has the same statistical accuracy of Gaussian width but applies to subexponential samples.

As illustrated in Table I, square of the Gaussian width captures the degrees of freedom for practical regularizers. Table I is obtained by setting $R = \mathcal{R}(\boldsymbol{\theta}^\star)$ in (4). In practice, a good choice for R can be found by using cross validation. It is also known that the performance of PGD is robust to choice of R (see Thm 2.6 of [14]).

Constraint	**Parameter vector model**	$\boldsymbol{\omega^2(\mathcal{C})}$
None	$\boldsymbol{\theta}^\star \in \mathbb{R}^p$	p
Sparsity $\|\cdot\|_{\ell_0}$	s non-zero entries	$s\log(6p/s)$
ℓ_1 norm $\|\cdot\|_{\ell_1}$	s non-zero entries	$s\log(6p/s)$
Subspace	$\boldsymbol{\theta}^\star \in \mathcal{S}$, $\dim(\mathcal{S}) = k$	k
Matrix rank	$\text{rank}(\text{mat}(\boldsymbol{\theta}^\star)) \le r$	$rp^{1/2}$

TABLE I: List of low-dimensional models and corresponding Gaussian widths (up to a constant factor) for the constraint sets $\mathcal{K} = \{\boldsymbol{\theta} \;\big|\; \mathcal{R}(\boldsymbol{\theta}) \le \mathcal{R}(\boldsymbol{\theta}^\star)\}$. If constraint is set membership such as subspace, $\mathcal{R}(\boldsymbol{\theta}) = 0$ inside the set and ∞ outside. Furthermore, we represent the vector $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ in matrix form as $\text{mat}(\boldsymbol{\theta}^\star) \in \mathbb{R}^{\sqrt{p}\times\sqrt{p}}$.

The next statistical quantity required in our analysis is the Orlicz norm defined as.

Definition 2.2 (Orlicz norms): For a scalar random variable Orlicz-a norm is defined as

$$\|X\|_{\psi_a} = \sup_{p\geq 1} p^{-1/a}(\mathbb{E}[|X|^p])^{1/p}$$

Orlicz-a norm of a vector $\boldsymbol{x} \in \mathbb{R}^d$ is defined as $\|\boldsymbol{x}\|_{\psi_a} = \sup_{\boldsymbol{v}\in\mathcal{B}^d} \|\boldsymbol{v}^T\boldsymbol{x}\|_{\psi_a}$. Subexponential and subgaussian norms are special cases of Orlicz-a norm given by $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ respectively.

Based on perturbed Gaussian width definition, we will show that one can upper bound the critical quantities (10) and (11). In return, this will reveal the statistical and computational performance of the PGD algorithm. This is the topic of the next section which states our main results.

III. Main Results

In this section we estimate the convergence rate and the statistical accuracy of the PGD algorithm as a function of sample size, complexity of the parameter (e.g. sparsity level), and the distribution of the data (whether subgaussian or subexponential). Our main theorem establishes a linear convergence rate of PGD and shows that PGD achieves statistically efficient error rates. We first describe the data model.

Definition 3.1 (Isotropic vector): $\boldsymbol{x} \in \mathbb{R}^p$ is called an isotropic Orlicz-a vector if it is zero-mean with identity covariance and if its Orlicz-a norm $\|\boldsymbol{x}\|_{\psi_a}$ is bounded by an absolute constant.

Definition 3.2 (σ-noisy datasets): We assume the samples $(y_i, \boldsymbol{x}_i)_{i=1}^n$. We call a dataset σ-Orlicz-a if the input samples are isotropic Orlicz-a vectors and the residual at the ground truth obeys

$$\|y - \boldsymbol{x}^T\boldsymbol{\theta}^\star - \mu^\star\|_{\psi_a} \leq \sigma.$$

We call σ-Orlicz-1 dataset σ-subexponential and σ-Orlicz-2 dataset σ-subgaussian.

Note that residual at the ground truth is the noise in our problem which may be function of the nonlinearity. Our main results capture the PGD performance for different dataset models.

Theorem 3.3 (Subgaussian): Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subgaussian dataset. Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$ and set learning rate $\eta = 1/n$. Let $\mathcal{R}$ be an arbitrary regularizer. Starting form any initial estimate $[\boldsymbol{\theta}_0^T \ \mu_0]^T$, with probability at least $1 - 6\exp(-C_0t^2/2) - 4n^{-100}$, all PGD iterates (6) obeys

$$\left\|\begin{bmatrix}\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^\star \\ \mu_\tau - \mu^\star\end{bmatrix}\right\|_{\ell_2} \leq (c\frac{\omega(\mathcal{C}) + t}{\sqrt{n}})^\tau \left\|\begin{bmatrix}\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star \\ \mu_0 - \mu^\star\end{bmatrix}\right\|_{\ell_2} + C\sigma\frac{(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}}.$$

Similarly, for subexponential samples, we have the following theorem which applies to convex regularizers.

Theorem 3.4 (Subexponential): Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subexponential dataset. Set $q = (n+p)\log^3(n+p)$. Set learning rate $\eta = c_0/q$, suppose $\mathcal{R}$ is convex and $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$. Starting from initialization $[\boldsymbol{\theta}_0^T \ \mu_0]^T$, with probability at least $1-7\exp(-C_0\min\{n, t\sqrt{n}, t^2\})-6p(n+p)^{-100}$, all PGD iterates (6) obey

$$\left\|\begin{bmatrix}\boldsymbol{\theta}_\tau - \boldsymbol{\theta}^\star \\ \mu_\tau - \mu^\star\end{bmatrix}\right\|_{\ell_2} \leq \left(1 - \frac{cn}{q}\right)^\tau \left\|\begin{bmatrix}\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star \\ \mu_0 - \mu^\star\end{bmatrix}\right\|_{\ell_2} + C\sigma\frac{(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}}.$$

Both of these results show that PGD iterates converge to population parameters $\boldsymbol{\theta}^\star$, $\mu^\star$ at a linear rate. Subexponential theorem requires a more conservative choice of learning rate. The statistical estimation error grows as $\omega(\mathcal{C})/\sqrt{n}$ for subgaussian and $\omega_n(\mathcal{C})/\sqrt{n}$ for subexponential. Since our results apply in the regime $n \gtrsim \omega^2(\mathcal{C})$, following (12), statistical errors associated with subgaussian and subexponential are same up to a constant for typical regularizers.

Our main results follow from Theorems 3.5 and 3.6 which are the topics of the following sections.

A. Bounding the Error due to Nonlinearity

In this section we provide a bound on the effective noise level $\nu(\mathcal{C})$; which is crucial for assessing statistical accuracy. This term arises from the nonlinearity and noise associated with the relation between input and output. For example, for single-index models, we have $\mathbb{E}[y \mid \boldsymbol{x}] = \phi(\boldsymbol{x}^T\boldsymbol{\theta}_{\text{GT}})$ for some link function ϕ and ground truth $\boldsymbol{\theta}_{\text{GT}}$, and ϕ becomes the source of the nonlinearity. Our approach is similar to [4]–[7], [27] and treats the nonlinearity as a noise. The finite sample noise is captured by the residual vector

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}^\star - \mathbf{1}\mu^\star. \tag{13}$$

Following $\nu(\mathcal{C})$ term in (11), the contribution of the residual $\boldsymbol{w}$ to the estimated parameter is captured by the vector

$$\boldsymbol{e} = [\boldsymbol{X} \ \mathbf{1}]^T\boldsymbol{w} = \sum_{i=1}^n (y_i - \mu^\star - \boldsymbol{x}_i^T\boldsymbol{\theta}^\star)\begin{bmatrix}\boldsymbol{x}_i \\ 1\end{bmatrix}. \tag{14}$$

Our key observation is that the properties of $\boldsymbol{e}$ can be characterized under fairly general assumptions compared to the existing literature; which is mostly restricted to zero-mean subgaussian samples.

Theorem 3.5 (Statistical error): Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$ is a σ-subgaussian dataset. Let the tangent balls $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in (7) and (8) respectively.

Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$. Then, with probability at least $1 - 2\exp(-t^2/2) - 4n^{-100}$, we have

$$\frac{\nu(\mathcal{C})}{n} \lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}}.$$

where $\nu(\mathcal{C})$ is the effective noise given by (11). If $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subexponential dataset and $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$, with probability at least $1 - 2\exp(-c\ \min\{t\sqrt{n}, t^2\}) - 4n^{-100}$, we have

$$\frac{\nu(\mathcal{C})}{n} \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}}.$$

This theorem establishes the crucial finite sample upper bounds on $\nu(\mathcal{C})$ for both subgaussian and subexponential data as a function of Gaussian width of the tangent ball.

B. Controlling the Convergence Rate of PGD

Next, we study the convergence rate characterized by the $\rho(\mathcal{C})$ term. The challenges we address are (i) characterizing the *restricted singular values* of the subexponential data matrices and (ii) addressing the concatenated all ones vector.

Theorem 3.6 (Convergence rate): Suppose $(\boldsymbol{x}_i, y_i)_{i=1}^n$ is a σ-subgaussian dataset and $[\boldsymbol{X}\ \mathbf{1}]$ is the modified-data matrix, where $\mathbf{1}$ is a vector of all ones. Let $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be the tangent balls as defined in (7) and (8) respectively. Assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$. Setting $\eta = 1/n$, with probability at least $1 - 4e^{-ct^2}$ we have

$$\rho(\mathcal{C}) \lesssim \frac{\omega(\mathcal{C}) + t}{\sqrt{n}}.$$

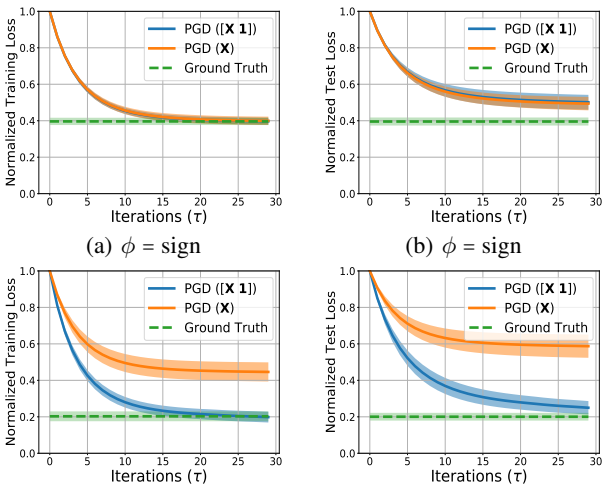
If the dataset is σ-subexponential, then setting $\eta = c_0/(n+p)\log^3(n+p)$ and assuming $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$, with probability $1 - 5\exp(-c\min(n, t\sqrt{n}, t^2)) - 3(n+p)^{-100}$, we have

$$\rho(\mathcal{C}) \leq 1 - C_0\eta n.$$

Note that, subexponential requires a smaller choice of learning rate which results in slower convergence. Combining our bounds on $\rho(\mathcal{C})$ and $\nu(\mathcal{C})$ and utilizing the recursion (9), we can obtain the PGD convergence characteristics and prove the main theorems.

IV. Numerical Experiments

In this section, we discuss experiments that corroborate our theoretical results. We consider a standard single-index model where for some ground truth vector $\boldsymbol{\beta}$ and link function ϕ, the input/output relation is given by $y_i = \phi(\boldsymbol{\beta}^T\boldsymbol{x}_i)$. We pick $\boldsymbol{\beta}$ to be a sparse vector with $s = 20$ nonzeros and $p = 800$ and set sample size to be $n = 500$. Because of sparsity prior, we run PGD as iterative hard thresholding where $\boldsymbol{\theta}_\tau$ is projected to be s-sparse after every iteration. As link functions, we considered ReLU (i.e. $\max(x, 0)$) and sign functions (maps to ± 1); which are of interest for deep learning and quantization respectively. We generate $\boldsymbol{x}_i$'s with i.i.d. exponentially distributed entries (with parameter $\lambda = 1$) and then remove the mean and normalize the covariance to identity. We pick a learning rate of $\eta = 1/5n$ in all experiments. The shaded areas in the plots correspond to one standard deviation.



(a) ϕ = sign (b) ϕ = sign

(c) ϕ = ReLU (d) ϕ = ReLU

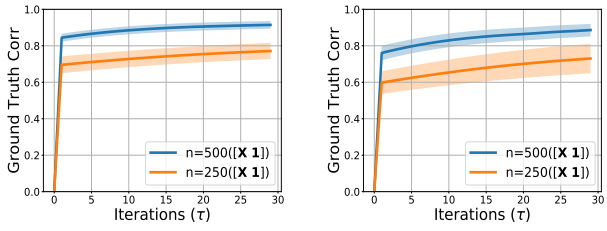
Fig. 1: We run PGD with different activations (ReLU and sign) using $\boldsymbol{X}$ and $[\boldsymbol{X}\ \mathbf{1}]$ as data matrices. In both cases, train and test errors decay gracefully to the ground truth baseline (with debiased output). However, PGD using $[\boldsymbol{X}\ \mathbf{1}]$ outperforms PGD using $\boldsymbol{X}$ alone for ReLU.

To assess test and training performance of PGD, we use the following three metrics

- **the normalized training error** defined as $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau - \mu_\tau\mathbf{1}\|_{\ell_2}^2/\|\boldsymbol{y}\|_{\ell_2}^2$ and
- **the normalized test error** that is similarly defined but evaluated on a fresh dataset of size n using the training model $\boldsymbol{\theta}_\tau$
- **correlation to ground truth** vector $\boldsymbol{\beta}$ defined as $\frac{\boldsymbol{\theta}_\tau^T\boldsymbol{\beta}}{\|\boldsymbol{\theta}_\tau\|_{\ell_2}\|\boldsymbol{\beta}\|_{\ell_2}}$.

We compare two baselines. First one is running PGD with only $\boldsymbol{X}$ i.e. without mean estimation. Second one assumes knowledge of ground truth $\boldsymbol{\beta}$ and fits a model $\gamma\boldsymbol{\beta}$ by finding γ to minimize the training loss. Numerically, we minimize $\|\overline{\boldsymbol{y}} - \gamma\boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2$ over γ where $\overline{y}_i = y_i - (1/n)\sum_{i=1}^n y_i$. This sets $\gamma = \overline{\boldsymbol{y}}^T\boldsymbol{X}\boldsymbol{\beta}/\|\boldsymbol{X}\boldsymbol{\beta}\|_{\ell_2}^2$.

Figure 1 plots the loss as a function of the PGD iterations τ. Both training and test errors gracefully decays with more iterations for both choices of link functions. The dashed values corresponds to $\gamma\boldsymbol{\beta}$'s performance.



(a) ϕ = sign (b) ϕ = ReLU

(c) ϕ = sign (d) ϕ = ReLU

Fig. 2: We estimate correlation of PGD estimate with the ground truth vector $\boldsymbol{\beta}$. Correlation increases with more samples. Correlation also improves when we use mean estimation $[\boldsymbol{X}\ \mathbf{1}]$ instead of $\boldsymbol{X}$.

While there is a slight mismatch between train/test performances (due to finite samples), high-dimensional estimation via PGD works well and performs on par with ground truth. Observe that for ReLU, $\mathbb{E}[y]$ is nonzero and estimating mean should be beneficial. Indeed, Figures 1c and 1d demonstrates that $[\boldsymbol{X}\ \mathbf{1}]$ substantially outperforms using $\boldsymbol{X}$ alone. There is no improvement for sign function since $\mathbb{E}[y] \approx 0$ (as the sign is symmetric).

In Figure 2 we focus on the parameter estimation question by plotting the correlation between $\boldsymbol{\theta}_\tau$ and $\boldsymbol{\beta}$. Correlation is always between $-1, 1$ and quantifies how well we can estimate direction of the ground truth vector via PGD. This experiment is conducted with two values of n namely 250 and 500 while $p = 800$ in both cases. Observe that, a larger sample size results in more stable estimation (smaller standard deviations) and higher correlation with output. Additionally Figure 2b shows that ReLU problem achieves better correlation once we account for the bias term. Hence, mean estimation is not only beneficial for test performance but also for parameter estimation.

V. Proofs of Main Theorems

This section proves our main results and outlines the proof of Theorems 3.3, 3.4 3.5 and 3.6. Throughout we use the same notation as described in II.

A. Proof of Theorem 3.4

We provide our analysis for subexponential samples. The extension to subgaussian samples is accomplished in an identical fashion. Set the estimation error at iteration τ to be $\boldsymbol{h}_\tau = \boldsymbol{\theta}_\tau - \boldsymbol{\theta}^\star$. Note that, when $\rho(\mathcal{C}) < 1$ and $\mathcal{R}$ is a convex regularizer, then the recursion (9) can be iteratively expanded as

$$\begin{aligned}\|\boldsymbol{h}_\tau\|_{\ell_2} &\le \|\boldsymbol{h}_0\|_{\ell_2}\rho(\mathcal{C})^\tau + \eta\nu(\mathcal{C})\sum_{k=0}^{\tau-1}\rho(\mathcal{C})^k \\ &\le \|\boldsymbol{h}_0\|_{\ell_2}\rho(\mathcal{C})^\tau + \eta\nu(\mathcal{C})\sum_{k=0}^{\infty}\rho(\mathcal{C})^k \\ &= \|\boldsymbol{h}_0\|_{\ell_2}\rho(\mathcal{C})^\tau + \frac{\eta\nu(\mathcal{C})}{1-\rho(\mathcal{C})} \end{aligned} \tag{15}$$

With the advertised probability, subexponential statements of Theorems 3.5 and 3.6 hold. Hence, for some constants, we have that $\rho(\mathcal{C}) \le 1 - c_0\eta n$, $\nu(\mathcal{C}) \le C\sqrt{n}\sigma(\omega_n(\mathcal{C}) + t)\log(n)$ and $\eta = c/q$ with $q = (n+p)\log^3(n+p)$. Plugging these in (15), we find the following upper bound on the right hand side,

$$\begin{aligned}\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} &\le (1 - c_0\eta n)^\tau\|\boldsymbol{h}_0\|_{\ell_2} \\ &\quad + \frac{\eta}{c_0\eta n}C\sqrt{n}\sigma(\omega_n(\mathcal{C}) + t)\log(n) \\ &= (1 - \frac{c_0 cn}{q})^\tau\|\boldsymbol{h}_0\|_{\ell_2} \\ &\quad + \sigma\eta\frac{C}{c_0}\frac{(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}},\end{aligned}$$

which is the desired bound. The case of subgaussian samples is again a corollary of Theorems 3.5 and 3.6. This concludes the proof of our main result.

B. Proof of Theorem 3.6 for subgaussian samples

We start our proof with the following lemma.

Lemma 5.1: Let $(\boldsymbol{x}_i)_{i=1}^n \mathbb{R}^p$ be i.i.d. isotropic subgaussian samples. Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be concatenated data and $[\boldsymbol{X}\ \mathbf{1}]$ is modified-data matrix, where $\mathbf{1}$ is a vector of all ones. Let $\mathcal{T}$ be a closed set with Euclidian radius bounded by a constant and

$$\mathcal{T}_{\text{ext}} = \{\widetilde{\boldsymbol{v}} \mid \widetilde{\boldsymbol{v}} = [\beta\boldsymbol{v}^T\ \gamma]^T\}.$$

where $\beta \le C_1$, $\gamma \le C_2$ for some positive constants C_1, C_2. Assume $n \gtrsim (\omega(\mathcal{T}) + t)^2$. Then, with probability at least $1 - 2e^{-t^2}$ we have

$$\sup_{\widetilde{\boldsymbol{v}}\in\mathcal{T}_{\text{ext}}} |\widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}])\widetilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}}.$$

The proof of Lemma 5.1 is deferred to Section VII-A. Next using the result of Lemma 5.1, we obtain the following lemma which bounds the convergence rate for subgaussian samples.

Lemma 5.2: Consider the setup of Lemma 5.1. Furthermore, let the tangent balls $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in (7)

and (8) respectively. With probability at least $1-4e^{-t^2}$, the following holds

$$\sup_{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} |\widetilde{\boldsymbol{u}}^T(\boldsymbol{I}-\frac{1}{n}[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}])\widetilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{C})+t}{\sqrt{n}}.$$

The proof of Lemma 5.2 is deferred to Section VII-B. This completes the proof for subgaussian samples.

C. Proof of Theorem 3.6 for subexponential samples

Let $(\boldsymbol{x}_i)_{i=1}^n \sim \boldsymbol{x}$ be i.i.d. isotropic subexponential vectors and $\boldsymbol{X}$ be the associated design matrix as previously. Let $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in 7 and 8 respectively. Assume $n \gtrsim \omega_n^2(\mathcal{C})$. Our proof strategy is based on the observation that, we can bound the (restricted) singular values of $[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]$ with high probability for subexponential data as follows.

1) Upper bounding the singular values: In this section we will bound the largest eigenvalue of the matrix $[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]$ with high probability. Towards this goal, we utilize Matrix Chernoff bound from [33].

Theorem 5.3 (Matrix Chernoff [33]): Consider a finite sequence $\{\boldsymbol{X}_i\}_{i=1}^n$ of independent, random, Hermitian matrices with common dimension d. Assume that

$$0 \le \sigma_{\min}(\boldsymbol{X}_i) \text{ and } \|\boldsymbol{X}_i\| \le L \quad \text{for } i=1,\dots,n$$

Define the sum $\boldsymbol{M} = \sum_{i=1}^n \boldsymbol{X}_i$ and $\zeta_{\max}$ be an upper bound on the spectral norm of the expectation $\mathbb{E}[\boldsymbol{M}]$ i.e. $\zeta_{\max} \ge \|\mathbb{E}[\boldsymbol{M}]\| = \|\sum_{i=1}^n \mathbb{E}[\boldsymbol{X}_i]\|$. We have that

$$\mathbb{P}\{\|\boldsymbol{M}\| \ge (1+\epsilon)\zeta_{\max}\} \le d\left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\frac{\zeta_{\max}}{L}}, \epsilon \ge 0$$

We will use Theorem 5.3 to bound the largest eigenvalue of $[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]$. Observe that

$$[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}] = \sum_{i=1}^n \begin{bmatrix}\boldsymbol{x}_i \\ 1\end{bmatrix}[\boldsymbol{x}_i^T\ 1].$$

Clearly this matrix is positive semidefinite. To bound $\|[\boldsymbol{x}_i^T\ 1]^T[\boldsymbol{x}_i^T\ 1]\|$ we use the following lemma.

Lemma 5.4 (Spectral norm bound): Let $(\boldsymbol{x}_i)_{i=1}^n$ be i.i.d. isotropic subexponential samples in $\mathbb{R}^p$. Then, with probability at least $1-2(n+p)^{-100}$ the spectral norm of all $\boldsymbol{x}_i\boldsymbol{x}_i^T$ matrices can be bounded as

$$\|\boldsymbol{x}_i\boldsymbol{x}_i^T\| \le \|\boldsymbol{x}_i\|_{\ell_2}^2 \le cp\log^2(n+p)$$

The proof of lemma 5.4 is deferred to Section VII-C. Lemma 5.4 guarantees that $\|[\boldsymbol{x}_i^T\ 1]^T[\boldsymbol{x}_i^T\ 1]\| \le \|[\boldsymbol{x}_i^T\ 1]^T\|_{\ell_2}^2 = \|\boldsymbol{x}_i\|_{\ell_2}^2+1 \le Cp\log^2(n+p)$. Hence, we do satisfy the conditions required by Theorem 5.3. Before using Theorem 5.3 we will upper bound the spectral norm of the expectation $\mathbb{E}[[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]]$ as follows.

Lemma 5.5 (Spectral norm bound of expectation): Let $\boldsymbol{x}$ be an isotropic subexponential vector, $\tilde{\boldsymbol{x}} = [\boldsymbol{x}^T\ 1]^T$ and let $B = Cp\log^2(n+p)$ for sufficiently large constant $C > 0$. Then we have

$$\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \mid \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le \sqrt{B}\right] \le 2\boldsymbol{I}_p.$$

The proof of Lemma 5.5 is deferred to Section VII-D. Thus, applying Lemma 5.5 on the set of all $[\boldsymbol{x}_i^T\ 1]^T$ satisfying $\|[\boldsymbol{x}_i^T\ 1]^T[\boldsymbol{x}_i^T\ 1]\| \le Cp\log^2(n+p)$, we find that with $1-2(n+p)^{-100}$ probability the following holds

$$\begin{aligned}\|\mathbb{E}[[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]]\| &= \|\mathbb{E}[\sum_{i=1}^n \begin{bmatrix}\boldsymbol{x}_i\\1\end{bmatrix}[\boldsymbol{x}_i^T\ 1]]\| \\ &\le \|\sum_{i=1}^n 2\boldsymbol{I}_p\| \\ &= 2n.\end{aligned}$$

Hence we can pick $\zeta_{\max} \ge 2n$ to upper bound the largest eigenvalue of $\mathbb{E}[[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]]$. Now, using Theorem 5.3 with $\zeta_{\max} = C_0C(n+p)\log^3(n+p), L = Cp\log^2(n+p)$ and $\epsilon = e-1$ we get

$$\begin{aligned}&\mathbb{P}\left\{\|[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]\| \ge eC_0C(n+p)\log^3(n+p)\right\} \\ &\le p\left[\frac{e^{e-1}}{e^e}\right]^{C_0\frac{n+p}{p}\log(n+p)} \\ &= pe^{-C_0\frac{n+p}{p}\log(n+p)} \le (n+p)^{-100}. \end{aligned} \tag{16}$$

Union bounding, with probability at least $1-3(n+p)^{-100}$,

$$\|[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}]\| \lesssim (n+p)\log^3(n+p). \tag{17}$$

2) Lower bounding the singular values: In this section we will lower bound the gain of $[\boldsymbol{X}\ \boldsymbol{1}]$ restricted to the tangent ball $\mathcal{C}_{\text{ext}}$. We will utilize the notion of restricted singular value (RSV) to proceed.

Definition 5.6 (Restricted singular value): Given a matrix $\boldsymbol{M}$ and a closed set $\mathcal{C}$, the RSV of $\boldsymbol{M}$ at $\mathcal{C}$ is defined as

$$\sigma(\boldsymbol{M},\mathcal{C}) = \min_{\boldsymbol{v}\in\mathcal{C}} \frac{\|\boldsymbol{M}\boldsymbol{v}\|_{\ell_2}}{\|\boldsymbol{v}\|_{\ell_2}}$$

In the following, we will lower bound $\min_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}}\|[\boldsymbol{X}\ \boldsymbol{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}$ which is the RSV of $[\boldsymbol{X}\ \boldsymbol{1}]$ at $\mathcal{C}_{\text{ext}}$. Recall that any $\widetilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}$ with unit Euclidian norm obeys $\widetilde{\boldsymbol{v}} = [\sqrt{1-\gamma^2}\boldsymbol{v}^T\ \gamma]^T \in \mathcal{C}_{\text{ext}}$ for $0 \le \gamma \le 1$ and $\|\boldsymbol{v}\|_{\ell_2} = 1$. Consequently

$$\begin{aligned}\|[\boldsymbol{X}\ \boldsymbol{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}^2 &= \|\sqrt{1-\gamma^2}\boldsymbol{X}\boldsymbol{v}+\gamma\boldsymbol{1}\|_{\ell_2}^2 \\ &= (1-\gamma^2)\ \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2\boldsymbol{1}^T\boldsymbol{1} + 2\gamma\sqrt{1-\gamma^2}\boldsymbol{1}^T\boldsymbol{X}\boldsymbol{v} \\ &\ge (1-\gamma^2)\ \|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2 n + 2\gamma\sqrt{1-\gamma^2}\boldsymbol{v}^T\sum_{i=1}^n \boldsymbol{x}_i.\end{aligned}$$

Setting $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$ and taking infimum of both sides, we get

$$\min_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} \|[\boldsymbol{X}\ \mathbf{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}^2 \geq \tag{18}$$

$$\geq \min_{|\gamma|\leq 1}\left((1-\gamma^2)\min_{\boldsymbol{v}\in\mathcal{C}}\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2 n\right) - 2n\sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^T\overline{\boldsymbol{x}}|$$

$$\geq \min(\min_{\boldsymbol{v}\in\mathcal{C}}\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2, n) - 2n\sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^T\overline{\boldsymbol{x}}|. \tag{19}$$

In essence, (19) bounds RSV of $[\boldsymbol{X}\ \mathbf{1}]$ in terms of the RSV of $\boldsymbol{X}$ and some simpler terms. The following theorem from [29] (Theorem D.11) gives a lower lower bound on the RSV of a matrix $\boldsymbol{X}$ with i.i.d. subexponential rows.

Theorem 5.7 (Bounding RSV [29]): Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ be a random matrix with i.i.d. isotropic subexponential rows. Let $\mathcal{C}$ be a tangent ball as in (7) and suppose the sample size obeys $n \gtrsim (\omega_n(\mathcal{C}) + t)$. Then with probability at least $1 - 3\exp(-c\min(n, t^2, t\sqrt{n}))$, we have that

$$\min_{\boldsymbol{u}\in\mathcal{C}}\|\boldsymbol{X}\boldsymbol{u}\|_{\ell_2}^2 \geq c_0 n$$

Next, we shall state a lemma from [29] (Lemma D.7) to upper bound the term involving the sample average $\overline{\boldsymbol{x}}$.

Lemma 5.8 (Bounding empirical width [29]): Suppose $\mathcal{C}$ is a subset of the unit Euclidian ball and $(\boldsymbol{x}_i)_{i=1}^n$ are i.i.d. zero-mean vectors with bounded subexponential norm. Define the empirical average vector $\overline{\boldsymbol{x}} = n^{-1}\sum_i \boldsymbol{x}_i$. We have that

$$\mathbb{P}\left(\sup_{\boldsymbol{u}\in\mathcal{C}}|\boldsymbol{u}^T\overline{\boldsymbol{x}}| \leq C\frac{(\omega_n(\mathcal{C})+t)}{\sqrt{n}}\right) \geq 1 - 2\exp(-c\ \min(t\sqrt{n}, t^2))$$

Combining Theorem 5.7 and Lemma 5.8 into (19) we find that, there exist constants $c, c_0, C_0 > 0$ such that with probability at least $1 - 5\exp(-c\min(n, t\sqrt{n}, t^2))$, we can lower bound the RSV of $[\boldsymbol{X}\ \mathbf{1}]$ following (19)

$$\min_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}}\|[\boldsymbol{X}\ \mathbf{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}^2 \geq c_0 n - C_0 n\frac{\omega_n(\mathcal{C})+t}{\sqrt{n}} \geq c_0 n/2. \tag{20}$$

where last line follows from the assumption that $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$.

3) Upper bounding the convergence rate: Union bounding the events (17) and (20), we obtain upper and lower bounds on the singular values of $[\boldsymbol{X}\ \mathbf{1}]$ with the desired probability. Hence, we can bound the convergence rate of PGD as follows. Setting $q = (n+p)\log^3(n+p)$, we have (17) $\|[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}]\| \leq Cq$. Therefore, choosing learning rate $\eta = 1/Cq$, the matrix $\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}]$ is positive semidefinite (PSD). Hence, applying the generalized Cauchy-Schwarz inequality for PSD matrix, we find

$$\begin{aligned}
\rho(\mathcal{C}) &= \sup_{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} \widetilde{\boldsymbol{u}}^T(\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}])\widetilde{\boldsymbol{v}} \\
&\leq \sup_{\widetilde{\boldsymbol{u}},\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} [(\widetilde{\boldsymbol{u}}^T(\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}])\widetilde{\boldsymbol{u}})^{1/2} \\
&\qquad (\widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}])\widetilde{\boldsymbol{v}})^{1/2}] \\
&= \sup_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} \widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \eta[\boldsymbol{X}\ \mathbf{1}]^T[\boldsymbol{X}\ \mathbf{1}])\widetilde{\boldsymbol{v}} \\
&= 1 - \eta\min_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}}\|[\boldsymbol{X}\ \mathbf{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}^2 \\
&\leq 1 - c_0\eta n/2.
\end{aligned}$$

Here the last inequality follows from (20). This completes the proof for subexponential samples.

D. Proof of Theorem 3.5 for subgaussian samples

Suppose the dataset $(\boldsymbol{x}_i, y_i)_{i=1}^n \sim (\boldsymbol{x}, y)$ is σ-subgaussian. Let $\boldsymbol{X}, [\boldsymbol{X}\ \mathbf{1}], \mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in Section II, recall $\boldsymbol{w}$ from (13) and assume $n \gtrsim (\omega(\mathcal{C}) + t)^2$. Representing $\widetilde{\boldsymbol{v}} \in \mathcal{C}_{\text{ext}}$ as $\widetilde{\boldsymbol{v}} = [\sqrt{1-\gamma^2}\boldsymbol{v}^T\ \gamma]^T$ for $\boldsymbol{v} \in \mathcal{C}$ and $|\gamma| \leq 1$, we have

$$\begin{aligned}
\nu(\mathcal{C}) &= \sup_{\widetilde{\boldsymbol{v}}\in\mathcal{C}_{\text{ext}}} |\widetilde{\boldsymbol{v}}^T[\boldsymbol{X}\ \mathbf{1}]^T\boldsymbol{w}| \\
&= \sup_{\boldsymbol{v}\in\mathcal{C},|\gamma|\leq 1} |\sqrt{1-\gamma^2}\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w} + \gamma\mathbf{1}^T\boldsymbol{w}| \\
&\leq \sup_{\boldsymbol{v}\in\mathcal{C},|\gamma|\leq 1} |\sqrt{1-\gamma^2}\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w}| + \sup_{|\gamma|\leq 1}|\gamma\mathbf{1}^T\boldsymbol{w}| \\
&\leq \sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w}| + |\mathbf{1}^T\boldsymbol{w}|
\end{aligned} \tag{21}$$

In the following we will upper bound the terms $\sup_{\boldsymbol{v}\in\mathcal{C}}|\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w}|$ and $|\mathbf{1}^T\boldsymbol{w}|$ separately and will combine them to get an upper bound on the residual error.

1) Upper bounding the first term in (21): In order to upper bound the first term in (21), define the clipping function

$$\text{clip}(a, B) = \begin{cases} a \text{ if } |a| \leq B \\ \text{sign}(a)B \text{ else} \end{cases}.$$

Following lemma immediately follows from union bounding the large deviations of subgaussian and subexponential variables and shows that $X = \text{clip}(X, B)$ with high probability.

Lemma 5.9: Let $(w_i)_{i=1}^n$ be i.i.d. subgaussian random variables with $\|w_i\|_{\psi_2} \leq \sigma$. There exists a constant $C > 0$ such that picking $B = C\sqrt{\log(n)}$, with probability $1 - 2n^{-100}$ for all i, we have

$$w_i = \text{clip}(w_i, \sigma B).$$

If instead $(w_i)_{i=1}^n$ are i.i.d. subexponential with $\|w_i\|_{\psi_1} \leq \sigma$, then picking $B = C\log(n)$ leads to the same result.

Using Lemma 5.9, $\|\boldsymbol{w}\|_\infty \le \sigma B$ with probability $1 - 2n^{-100}$. Conditioned on this event, we have

$$\sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{w}| = \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n \text{clip}(w_i, \sigma B)\boldsymbol{x_i}|. \quad (22)$$

Setting $\boldsymbol{z}_i = \text{clip}(w_i, \sigma B)\boldsymbol{x_i} = w_i \boldsymbol{x}_i$, (22) can be rewritten as

$$\begin{aligned}\sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{w}| &= \frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n \boldsymbol{z}_i| \\ &\le \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| + \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n \mathbb{E}[\boldsymbol{z}_i]| \\ &\le \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| + n\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}. \quad (23)\end{aligned}$$

Note that $\boldsymbol{z}_i = w_i \boldsymbol{x}_i$ is subgaussian since w_i is bounded. The subgaussian norm obeys

$$\|\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]\|_{\psi_2} \lesssim \|\boldsymbol{z}_i\|_{\psi_2} \lesssim \sigma\sqrt{\log(n)}.$$

Define the average vector $\overline{\boldsymbol{z}} = n^{-1/2} \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])$ which is still subgaussian with same norm (up to a constant). Standard results from functional analysis [28] guarantee

$$\begin{aligned}\frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| &= n^{-1/2} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \overline{\boldsymbol{z}}| \\ &\lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}} \quad (24)\end{aligned}$$

with probability at least $1 - 2e^{-t^2/2}$. This bounds the first term of (23). Next, we address the expectation term $\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}$ via following lemma.

Lemma 5.10: Suppose $\boldsymbol{x}$ is an isotropic Orlicz-a vector and $\|w\|_{\psi_a} \le \sigma$. Let $B = C \log^{1/a}(n)$ for sufficiently large constant $C > 0$. For $a = 1, 2$, we have that

$$\|\mathbb{E}[w\boldsymbol{x} \mid |w| \le \sigma B]\|_{\ell_2} \lesssim \sigma p^2 n^{-201}.$$

The proof of Lemma 5.10 is deferred to Section VII-F. Combining (24) and Lemma 5.10 into (23), with probability at least $1 - 2e^{-t^2/2} - 2n^{-100}$, we find that,

$$\begin{aligned}\frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{w}| &\lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}} \\ &\quad + \sigma p^2 n^{-201} \\ &\lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}} \quad (25)\end{aligned}$$

which is the desired bound for the first term in (21).

2) Upper bounding the second term in (21): The vector $\boldsymbol{w}$ is zero-mean with $\|\boldsymbol{w}\|_{\psi_2} \le \sigma$. Hence, $\|\boldsymbol{1}^T \boldsymbol{w}\|_{\psi_2} \le \sigma\sqrt{n}$ which implies that with probability $1 - 2n^{-100}$,

$$|\boldsymbol{1}^T \boldsymbol{w}| \lesssim \sigma\sqrt{n \log n}.$$

Combining the bound above with (25), we get the advertised bound on the residual, namely

$$\frac{1}{n}\nu(\mathcal{C}) \lesssim \frac{\sigma(\omega(\mathcal{C}) + t)\sqrt{\log(n)}}{\sqrt{n}}, \quad (26)$$

with probability at least $1 - 2\exp(-t^2/2) - 4n^{-100}$.

E. Proof of Theorem 3.5 for subexponential samples

Suppose the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is σ-subexponential. Let $\boldsymbol{X}, [\boldsymbol{X}\ \boldsymbol{1}], \boldsymbol{w}, \mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in Section II. Assume $n \gtrsim (\omega_n(\mathcal{C}) + t)^2$. Similar to the subgaussian case, we split the residual into two terms via (21) and bound each term separately to get a final bound.

1) Upper bounding the first term in (21): Let $\boldsymbol{z}_i = w_i \boldsymbol{x}_i$. With probability $1 - 2n^{-100}$, we have that $\|\boldsymbol{w}\|_\infty \lesssim \sigma \log n$. We continue the analysis conditioned on this event. With bounded w_i, $\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]$ is subexponential via

$$\|\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i]\|_{\psi_1} \lesssim \|\boldsymbol{z}_i\|_{\psi_1} \lesssim \sigma \log n \|\boldsymbol{x}_i\|_{\psi_1} \lesssim \sigma \log n.$$

Lemma 5.8 guarantees that

$$\frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \sum_{i=1}^n (\boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i])| \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}} \quad (27)$$

with probability at least $1 - 2\exp(-\mathcal{O}(\min\{t\sqrt{n}, t^2\}))$. Using Theorem 5.10, we also upper bound $\|\mathbb{E}[\boldsymbol{z}_1]\|_{\ell_2}$ by $C\sigma p^2 n^{-201}$. Combining this with (27) and substituting into the (deterministic inequality) (23), with probability at least $1 - 2\exp(-\mathcal{O}(\min\{t\sqrt{n}, t^2\})) - 2n^{-100}$ we have,

$$\frac{1}{n} \sup_{\boldsymbol{v}\in\mathcal{C}} |\boldsymbol{v}^T \boldsymbol{X}^T \boldsymbol{w}| \lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}}. \quad (28)$$

2) Upper bounding overall error of (21): Using $\|\boldsymbol{w}\|_{\psi_1} \lesssim \sigma$ and applying Lemma 5.8 (over one-dimensional $\mathbb{R}$), we find that $|\boldsymbol{1}^T \boldsymbol{w}| \lesssim \sigma(1 + t)\sqrt{n}$ with probability $1 - 2\exp(-c\ \min(t\sqrt{n}, t^2))$. Combining this with (28) and plugging into (21), we get the advertised upper bound

$$\begin{aligned}\frac{1}{n}\nu(\mathcal{C}) &\lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}} + \frac{(1 + t)\sigma}{\sqrt{n}} \\ &\lesssim \frac{\sigma(\omega_n(\mathcal{C}) + t)\log(n)}{\sqrt{n}} \quad (29)\end{aligned}$$

which holds with probability at least $1 - 4\exp(-c\min\{t\sqrt{n}, t^2\}) - 2n^{-100}$. This completes the proof for σ-subexponential data.

VI. Conclusion

We studied the problem of finding the best linear model from n input-output samples under quadratic loss in the high-dimensional regime $n \ll p$. For estimation, we utilized the projected gradient descent algorithm and showed its fast convergence as well as statistical accuracy in a data-dependent fashion. Our results are established for subexponential design which is heavier tailed compared to well-studied subgaussian. In both cases, we prove that *nonlinearity of the problem behaves like independent noise* and we establish favorable statistical guarantees as if the problem is linear. We also modified the original regression problem to allow for mean estimation and demonstrated its practical benefit when output labels have nonzero mean via simulations.

It would be desirable to extend our results to general loss function. If a loss function ℓ has the potential to better capture input/output relation, we can solve for

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(y_i, \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle).$$

Specifically this function can still be quadratic but characterized by a nonlinear link function ϕ i.e. $\ell(y_i, \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle) = (y_i - \phi(\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle))^2$. We believe that much of the results presented here extends to strongly-increasing ϕ where the derivative is lower bounded by a constant i.e. $\phi' \geq \alpha$ for some $\alpha > 0$. These functions are shown to behave like linear regression [32]. However, it is not immediately clear if strong statistical and computational guarantees established in this paper (as well as related literature) can be established.

References

[1] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *Information and Inference: A Journal of the IMA*, vol. 6, no. 1, pp. 1–40, 2016.

[2] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*. IEEE, 2008, pp. 16–21.

[3] R. Ganti, N. Rao, R. M. Willett, and R. Nowak, "Learning single index models in high dimensions," *arXiv preprint arXiv:1506.08910*, 2015.

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

[5] Y. Plan, R. Vershynin, and E. Yudovina, "High-dimensional estimation with geometric constraints," *Information and Inference: A Journal of the IMA*, vol. 6, no. 1, pp. 1–40, 2017.

[6] Y. Plan and R. Vershynin, "The generalized lasso with non-linear observations," *IEEE Transactions on information theory*, vol. 62, no. 3, pp. 1528–1537, 2016.

[7] C. Thrampoulidis, E. Abbasi, and B. Hassibi, "Lasso with non-linear measurements is equivalent to one with linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 3420–3428.

[8] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[9] R. Vershynin, "Estimation in high dimensions: a geometric perspective," in *Sampling theory, a renaissance*. Springer, 2015, pp. 3–66.

[10] S. Dirksen, H. C. Jung, and H. Rauhut, "One-bit compressed sensing with partial gaussian circulant matrices," *arXiv preprint arXiv:1710.03287*, 2017.

[11] S. Dirksen and S. Mendelson, "Robust one-bit compressed sensing with partial circulant matrices," *arXiv preprint arXiv:1812.06719*, 2018.

[12] Y. Plan and R. Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *Information Theory, IEEE Transactions on*, vol. 59, no. 1, pp. 482–494, 2013.

[13] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence rates of gradient methods for high-dimensional statistical recovery," in *Advances in Neural Information Processing Systems*, 2010, pp. 37–45.

[14] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4129–4158, 2018.

[15] R. Giryes, Y. C. Eldar, A. M. Bronstein, and G. Sapiro, "Tradeoffs between convergence speed and reconstruction accuracy in inverse problems," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1676–1690, 2018.

[16] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[17] M. Genzel, "High-dimensional estimation of structured signals from non-linear observations with general convex loss functions," *IEEE Transactions on Information Theory*, vol. 63, no. 3, pp. 1601–1619, 2017.

[18] S. Dirksen and S. Mendelson, "Robust one-bit compressed sensing with non-gaussian measurements," *arXiv preprint arXiv:1805.09409*, 2018.

[19] C. Thrampoulidis and A. S. Rawat, "The generalized lasso for sub-gaussian measurements with dithered quantization," *arXiv preprint arXiv:1807.06976*, 2018.

[20] L. Jacques and V. Cambareri, "Time for dithering: fast and quantized random embeddings via the restricted isometry property," *Information and Inference: A Journal of the IMA*, vol. 6, no. 4, pp. 441–476, 2017.

[21] C. Xu and L. Jacques, "Quantized compressive sensing with rip matrices: The benefit of dithering," *arXiv preprint arXiv:1801.05870*, 2018.

[22] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation under nonconvexity," in *International Conference on Machine Learning*, 2016, pp. 2472–2481.

[23] Z. Yang, K. Balasubramanian, and H. Liu, "High-dimensional non-gaussian single index models via thresholded score function estimation," in *International Conference on Machine Learning*, 2017, pp. 3851–3860.

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

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

[26] H. L. Yap, M. B. Wakin, and C. J. Rozell, "Stable manifold embeddings with structured random matrices," *IEEE Journal on Selected Topics in Signal Processing,*, vol. 7, no. 4, pp. 720–730, 2013.

[27] M. Genzel and G. Kutyniok, "The mismatch principle: Statistical learning under large model uncertainties," *arXiv preprint arXiv:1808.06329*, 2018.

[28] M. Talagrand, "Gaussian processes and the generic chaining," in *Upper and Lower Bounds for Stochastic Processes*. Springer, 2014, pp. 13–73.

[29] S. Oymak, "Learning compact neural networks with regularization," *International Conference on Machine Learning*, 2018.
[30] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.
[31] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Inform. Inference*, 2014.
[32] S. Oymak, "Stochastic gradient descent learns state equations with nonlinear activations," *arXiv preprint arXiv:1809.03019*, 2018.
[33] J. A. Tropp *et al.*, "An introduction to matrix concentration inequalities," *Foundations and Trends® in Machine Learning*, vol. 8, no. 1-2, pp. 1–230, 2015.

VII. Appendix

This section provides the proofs of supporting results.

A. Proof of Lemma 5.1

We start by expanding the convergence term as follows by substituting $\widetilde{\boldsymbol{v}} = [\beta \boldsymbol{v} \ \gamma]^T$

$$
\begin{aligned}
&|\widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}])\widetilde{\boldsymbol{v}}| \\
&= |\frac{1}{n}\|[\boldsymbol{X}\ \boldsymbol{1}]\widetilde{\boldsymbol{v}}\|_{\ell_2}^2 - \|\widetilde{\boldsymbol{v}}\|_{\ell_2}^2| \\
&= |\frac{1}{n}\|\beta\boldsymbol{X}\boldsymbol{v} + \gamma\boldsymbol{1}\|_{\ell_2}^2 - \|[\beta\boldsymbol{v}^T\ \gamma]^T\|_{\ell_2}^2| \\
&= |\frac{1}{n}(\beta^2\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 + \gamma^2\boldsymbol{1}^T\boldsymbol{1} + 2\beta\gamma\boldsymbol{1}^T\boldsymbol{X}\boldsymbol{v}) - \beta^2\|\boldsymbol{v}\|_{\ell_2}^2 - \gamma^2| \\
&= |\frac{1}{n}\beta^2\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 - \beta^2\|\boldsymbol{v}\|_{\ell_2}^2 + \frac{1}{n}\gamma^2 n - \gamma^2 + 2\beta\gamma\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^T\boldsymbol{v} \\
&\le \beta^2|\frac{1}{n}\|\boldsymbol{X}\boldsymbol{v}\|_{\ell_2}^2 - \|\boldsymbol{v}\|_{\ell_2}^2| + |2\beta\gamma|\boldsymbol{v}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i}{n}| \\
&\lesssim |\boldsymbol{v}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{v}| + |\boldsymbol{v}^T\overline{\boldsymbol{x}}|, \qquad (30)
\end{aligned}
$$

where, $\overline{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i$ is the empirical average vector of i.i.d. subgaussian rows $(\boldsymbol{x}_i)_{i=1}^n$. Thus, using (30), we can write

$$
\begin{aligned}
&\sup_{\widetilde{\boldsymbol{v}}\in\mathcal{T}_{\text{ext}}} |\widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}])\widetilde{\boldsymbol{v}}| \\
&\qquad \lesssim \sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{v}| + \sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^T\overline{\boldsymbol{x}}|. \qquad (31)
\end{aligned}
$$

Given $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is isotropic subgaussian, Lemma 6.14 in [14] guarantees

$$
\sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{v}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}} \qquad (32)
$$

with probability at least $1 - e^{-t^2}$. Furthermore, since $(\boldsymbol{x}_i)_{i=1}^n$'s have bounded subgaussian norm, $\overline{\boldsymbol{x}}$ is also bounded and standard results from functional analysis guarantee [28]

$$
\sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^T\frac{\sum_{i=1}^n \boldsymbol{x}_i}{n}| = \sup_{\boldsymbol{v}\in\mathcal{T}} |\boldsymbol{v}^T\overline{\boldsymbol{x}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}} \qquad (33)
$$

with probability at least $1 - e^{-t^2}$. Combining the results (32) and (33) into (31), we find that

$$
\sup_{\widetilde{\boldsymbol{v}}\in\mathcal{T}_{\text{ext}}} |\widetilde{\boldsymbol{v}}^T(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}])\widetilde{\boldsymbol{v}}| \lesssim \frac{\omega(\mathcal{T}) + t}{\sqrt{n}}. \qquad (34)
$$

holds with probability at least $1 - 2e^{-ct^2}$, where $c > 0$ is a fixed constant.

B. Proof of Lemma 5.2

Let the tangent cones $\mathcal{C}$ and $\mathcal{C}_{\text{ext}}$ be as defined in (7) and (8) respectively. Define the sets

$$
\mathcal{T}_- = \mathcal{C}_{\text{ext}} - \mathcal{C}_{\text{ext}} \quad \text{and} \quad \mathcal{T}_+ = \mathcal{C}_{\text{ext}} + \mathcal{C}_{\text{ext}}
$$

and note that,

$$
\begin{aligned}
\omega(\mathcal{C} - \mathcal{C}) &= \mathbb{E}[\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}} \boldsymbol{g}^T(\boldsymbol{u} - \boldsymbol{v})] \\
&\le \mathbb{E}[\sup_{\boldsymbol{u}\in\mathcal{C}} \boldsymbol{g}^T\boldsymbol{u} + \sup_{\boldsymbol{v}\in-\mathcal{C}} \boldsymbol{g}^T\boldsymbol{v}] = 2\omega(\mathcal{C}).
\end{aligned}
$$

Similarly, $\omega(\mathcal{C} + \mathcal{C}) \le 2\omega(\mathcal{C})$. Applying Lemma 5.1 on $\mathcal{T}_+$ and $\mathcal{T}_-$, with advertised probability, we have

$$
\sup_{\boldsymbol{a}\in\mathcal{T}_+\cup\mathcal{T}_-} |\Lambda(\boldsymbol{a},\boldsymbol{a})| \lesssim \frac{\omega(\mathcal{C}) + t}{\sqrt{n}}.
$$

where $\Lambda(\boldsymbol{a},\boldsymbol{b}) = \boldsymbol{a}^T(\boldsymbol{I} - \frac{1}{n}[\boldsymbol{X}\ \boldsymbol{1}]^T[\boldsymbol{X}\ \boldsymbol{1}])\boldsymbol{b}$. Now, for any $\boldsymbol{u},\boldsymbol{v} \in \mathcal{C}_{\text{ext}}$, picking $\boldsymbol{u} + \boldsymbol{v} \in \mathcal{T}_+$, $\boldsymbol{u} - \boldsymbol{v} \in \mathcal{T}_-$, we have

$$
|\boldsymbol{\Lambda}(\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{u} + \boldsymbol{v})|, |\boldsymbol{\Lambda}(\boldsymbol{u} - \boldsymbol{v}.\boldsymbol{u} - \boldsymbol{v})| \le \frac{\omega(\mathcal{C}) + t}{\sqrt{n}}.
$$

To proceed, note that

$$
\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{v}) = \frac{\boldsymbol{\Lambda}(\boldsymbol{u} + \boldsymbol{v}, \boldsymbol{u} + \boldsymbol{v}) - \boldsymbol{\Lambda}(\boldsymbol{u} - \boldsymbol{v}, \boldsymbol{u} - \boldsymbol{v})}{4}.
$$

Hence $|\boldsymbol{\Lambda}(\boldsymbol{u},\boldsymbol{v})| \lesssim (\omega(\mathcal{C}) + t)/\sqrt{n}$.

C. Proof of Lemma 5.4

To start our proof let $\{\boldsymbol{x}_i\}_{i=1}^n \sim \boldsymbol{x}$ be i.i.d. isotropic subgaussian samples in $\mathbb{R}^p$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the concatenated design matrix. Let x_{ij} denotes the ij^{th} element of the matrix $\boldsymbol{X}$. Since each x_{ij} has subexponential norm bounded by a constant, there exists a constant $C > 0$ such that $|x_{ij}| \le C\log(n + p)$ holds with probability at least $1 - 2(n+p)^{-102}$ using subexponential tail bound. Union bounding over all entries of $\boldsymbol{X}$ yields that $|x_{ij}| \le C\log(n + p)$ holds for all i, j with probability at least $1 - 2(n+p)^{-100}$. Hence, we can bound each row $\boldsymbol{x}_i$ of $\boldsymbol{X}$ with probability at least $1 - 2(n + p)^{-100}$ via

$$
\|\boldsymbol{x}_i\|_{\ell_2} \le C\sqrt{p}\log(n + p), \qquad (35)
$$

or equivalently, we have

$$
\|\boldsymbol{x}_i\boldsymbol{x}_i^T\| \le \|\boldsymbol{x}_i\|_{\ell_2}^2 \le Cp\log^2(n + p)
$$

This completes the proof of Lemma 5.4.

D. Proof of Lemma 5.5

Recall that $\boldsymbol{x}_i$'s are i.i.d. isotropic subexponential. We can estimate the covariance matrix of $\tilde{\boldsymbol{x}}$ given $\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B$ using law of total probability as follows

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right] &= \mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \mid \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right]\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right) \\
&\quad + \mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \mid \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right]\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right) \qquad (36)
\end{aligned}
$$

Since a covariance matrix is positive-semidefinite, each term in (36) is individually positive semidefinite. Hence, we will drop

the second term in (36) to get the following lower bound on the covariance matrix

$$\mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] \succeq \mathbb{E}[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \mid \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B]\,\mathbb{P}(\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B) \tag{37}$$

Using Lemma 5.4, it follows that $\|\tilde{\boldsymbol{x}}\|_{\ell_2}^2 = \|[\boldsymbol{x}^T\ 1]\|_{\ell_2}^2 \le Cp\log^2(n+p) = B^2$ holds with probability at least $1-2(n+p)^{-100}$. Hence, following (37), we get

$$\begin{aligned}\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T \mid \|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right] &\preceq \frac{\mathbb{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right]}{\mathbb{P}\left(\|\tilde{\boldsymbol{x}}\|_{\ell_2} \le B\right)} \\ &\preceq \frac{1}{1-2(n+p)^{-100}}\boldsymbol{I}_p \preceq 2\boldsymbol{I}_p.\end{aligned}$$

This completes the proof of Lemma 5.5.

E. Proof of Lemma 5.9

Subgaussian case: Using subgaussian tail, for large enough constant $C > 0$, for each i, we have $|w_i| \le C\sigma\sqrt{\log(n)} = \sigma B$ with probability at least $1 - 2n^{-101}$. This implies $\text{clip}(w_i, \sigma B) = w_i$. Union bounding over all entries of $\boldsymbol{w}$, we find the result which holds with probability at least $1-2n^{-100}$. **Subexponential case** follows similarly with $B = C\log(n)$.

F. Proof of Lemma 5.10

We prove the result for subexponential samples. Subgaussian case follows similarly. Without losing generality let $\sigma = 1$ as everything can be scaled accordingly. Defining clip function as previously, set $\boldsymbol{z} = \text{clip}(w, B)\boldsymbol{x}$. Furthermore, let w_{tail} denotes the tail of $|w|$, such that,

$$w_{\text{tail}} = \begin{cases} |w| & \text{if } |w| > B \\ 0 & \text{otherwise} \end{cases} \tag{38}$$

w_{tail} is an upper bound on the error due to clipping i.e.

$$|w - \text{clip}(w, B)| \le w_{\text{tail}} \tag{39}$$

We proceed by upper bounding $\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2}$ in terms of w_{tail}, using subadditive property of ℓ_2-norm and the orthogonality of w and $\boldsymbol{x}$ (i.e., $\mathbb{E}[w\boldsymbol{x}] = 0$) as follows

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} &= \|\mathbb{E}[\text{clip}(w, B)\boldsymbol{x}]\|_{\ell_2} \\ &= \|\mathbb{E}[(w - \text{clip}(w, B))\boldsymbol{x}]\|_{\ell_2} \\ &\le \mathbb{E}[|w - \text{clip}(w, B)|\|\boldsymbol{x}\|_{\ell_2}] \\ &\le \mathbb{E}[w_{\text{tail}} \max(\|\boldsymbol{x}\|_{\ell_2}, \sqrt{p}B)]\end{aligned} \tag{40}$$

Using subexponentiality, for some constant $c > 0$, we have that $\mathbb{P}(w_{\text{tail}} > \sqrt{ct}) \le 2e^{-t}$ and $\mathbb{P}\{\|\boldsymbol{x}\|_{\ell_2} > \sqrt{cpt}\} \le 2pe^{-t}$ where latter follows from union bounding entries of $\boldsymbol{x}$. Union bounding, we get the following tail bound on the product,

$$\mathbb{P}\{w_{\text{tail}}\|\boldsymbol{x}\|_{\ell_2} > c\sqrt{p}t^2\} \le 4pe^{-t}. \tag{41}$$

For notational convenience, set

$$g = w_{\text{tail}} \max(\|\boldsymbol{x}\|_{\ell_2}, \sqrt{p}B). \tag{42}$$

and note that g satisfies the following property due to (38)

$$\begin{cases}\text{either} & g > \sqrt{p}B^2 \\ \text{or} & g = 0\end{cases}. \tag{43}$$

Furthermore, from (41) we get the following tail distribution

$$Q_g(t) = \mathbb{P}(g > t) \le 4pe^{-[\frac{t}{c\sqrt{p}}]^{1/2}}. \tag{44}$$

for $t \ge \alpha := \sqrt{p}B^2$. Combining (42), (43) and (44) into (40) and denoting probability density function of g by f_g, we get

$$\begin{aligned}\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \le \mathbb{E}[g] &= \int_\alpha^\infty t f_g(t)dt = -\int_\alpha^\infty t dQ_g(t) \\ &= -tQ_g(t)\big|_\alpha^\infty + \int_\alpha^\infty Q_g(t)dt \\ &= \sqrt{p}B^2 Q_g(\sqrt{p}B^2) + \int_\alpha^\infty Q_g(t)dt \\ &\overset{(a)}{\le} e^{-B/\sqrt{c}} + 4p\int_{\sqrt{p}B^2}^\infty e^{-[\frac{t}{c\sqrt{p}}]^{1/2}}dt.\end{aligned} \tag{45}$$

where (a) follows from (44). To bound the term on the right hand side, we do a change of variable in (45) by setting $\tau = [t/(c\sqrt{p})]^{1/2}$ to get,

$$\begin{aligned}4p\int_{\sqrt{p}B^2}^\infty e^{-[\frac{t}{c\sqrt{p}}]^{1/2}}dt &\le 8cp^2\int_{B/\sqrt{c}}^\infty \tau e^{-\tau}d\tau \\ &\le 8cp^2\big[-\tau e^{-\tau}\big|_{B/\sqrt{c}}^\infty + \int_{B/\sqrt{c}}^\infty e^{-\tau}d\tau\big] \\ &= 8cp^2\big[\frac{B}{\sqrt{c}}e^{-B/\sqrt{c}} + e^{-B/\sqrt{c}}\big] \\ &\le 8cp^2(B/\sqrt{c}+1)e^{-B/\sqrt{c}}.\end{aligned} \tag{46}$$

Overall, we found

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \le 4p^2(B^2 + 2c(B/\sqrt{c}+1))e^{-B/\sqrt{c}}$$

which is upper bounded by $\mathcal{O}(n^{-100})$ by picking $B = C\log n$. Finally, note that conditioned on $|w| \le B$, $\boldsymbol{z} = w\boldsymbol{x}$ and

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \ge \|\mathbb{E}[w\boldsymbol{x} \mid |w| \le B]\|_{\ell_2}\mathbb{P}(|w| \le B).$$

Since $\mathbb{P}(|w| \le B) > 1/2$, this yields $\|\mathbb{E}[w\boldsymbol{x} \mid |w| \le B]\|_{\ell_2} \lesssim p^2 n^{-201}$ which is the advertised result.

Similarly for subgaussian samples, one can show that

$$\|\mathbb{E}[\boldsymbol{z}]\|_{\ell_2} \lesssim p^2 B^2 e^{-B^2/c}. \tag{47}$$

Picking $B = C\sqrt{\log n}$, we conclude with the same result concluding the proof of Lemma 5.10.