Structured Signal Recovery from Non-linear and Heavy-tailed Measurements

Larry Goldstein, Stanislav Minsker and Xiaohan Wei

***Abstract*—We study high-dimensional signal recovery from non-linear measurements with design vectors having elliptically symmetric distribution. Special attention is devoted to the situation when the unknown signal belongs to a set of low statistical complexity, while both the measurements and the design vectors are heavy-tailed. We propose and analyze a new estimator that adapts to the structure of the problem, while being robust both to the possible model misspecification characterized by arbitrary non-linearity of the measurements as well as to data corruption modeled by the heavy-tailed distributions. Moreover, this estimator has low computational complexity. Our results are expressed in the form of exponential concentration inequalities for the error of the proposed estimator. On the technical side, our proofs rely on the generic chaining methods, and illustrate the power of this approach for statistical applications. Theory is supported by numerical experiments demonstrating that our estimator outperforms existing alternatives when data is heavy-tailed.**

***Index Terms*—Signal reconstruction, Nonlinear measurements, Heavy-tailed noise, Elliptically symmetric distribution.**

I. INTRODUCTION.

Let $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ be a random couple with distribution P governed by the *semi-parametric single index model*

$$y = f(\langle \mathbf{x}, \theta_* \rangle, \delta), \tag{1}$$

where $\mathbf{x}$ is a measurement vector with marginal distribution Π, δ is a noise variable that is assumed to be independent of $\mathbf{x}$, $\theta_* \in \mathbb{R}^d$ is a fixed but otherwise unknown signal ("index vector"), and $f : \mathbb{R}^2 \mapsto \mathbb{R}$ is an unknown link function; here and in what follows, $\langle \cdot, \cdot \rangle$ denotes the Euclidean dot product. We impose mild conditions on f, and in particular it is not assumed that f is convex, or even continuous.[1] Our goal is to estimate the signal θ_* from the training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ - a sequence of i.i.d. copies of $(\mathbf{x}, y)$ defined on a probability space $(\Omega, \mathcal{B}, \mathbb{P})$. As $f(a^{-1}\langle \mathbf{x}, a\theta_* \rangle, \delta) = f(\langle \mathbf{x}, \theta_* \rangle, \delta)$ for any $a > 0$, the best one can hope for is to recover θ_* up to a scaling factor. Hence, without loss of generality, we will assume that θ_* satisfies $\|\mathbf{\Sigma}^{1/2}\theta_*\|_2^2 := \left\langle \mathbf{\Sigma}^{1/2}\theta_*, \mathbf{\Sigma}^{1/2}\theta_* \right\rangle = 1$, where $\mathbf{\Sigma} = \mathbb{E}(\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{x} - \mathbb{E}\mathbf{x})^T$ is the covariance matrix of $\mathbf{x}$.

In many applications, θ_* possesses special structure, such as sparsity or low rank (when $\theta_* \in \mathbb{R}^{d_1 \times d_2}$, $d_1 d_2 = d$, is a matrix). To incorporate such structural assumptions into the problem, we will assume that θ_* is an element of a closed set $\Theta \subseteq \mathbb{R}^d$ of small "statistical complexity" that is characterized by its Gaussian mean width [1]. The past decade has witnessed significant progress related to estimation in high-dimensional spaces, both in theory and applications. Notable examples include sparse linear regression ([2], [3], [4]), low-rank matrix recovery ([5], [6], [7]), and mixed structure recovery [8]. However, the majority of the aforementioned works assume that the link function f is linear, and their results apply only to this particular case.

Generally, the task of estimating the index vector requires approximating the link function f or its derivative, assuming that it exists (the so-called Average Derivative Method), see [9], [10], [11]. However, when the measurement vector $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$, a somewhat surprising result states that one can estimate θ_* directly, avoiding preliminary link function estimation step completely. More specifically, it has been proven in [12] that $\eta\theta_* = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \mathbb{E}\left(y - \langle \theta, \mathbf{x} \rangle\right)^2$, where $\eta = \mathbb{E}\, y \langle \mathbf{x}, \theta_* \rangle$. Here is the short proof of this fact:

$$\begin{aligned}
&\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \mathbb{E}\left(y - \langle \theta, \mathbf{x} \rangle\right)^2 \\
=&\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left[\|\theta\|_2^2 - 2\mathbb{E}y \langle \mathbf{x}, \theta \rangle\right] \\
=&\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left[\|\theta\|_2^2 - 2\mathbb{E}y \langle \mathbf{x}, \theta_* \rangle \langle \theta, \theta_* \rangle - 2\mathbb{E}y \left\langle \mathbf{x}, \theta_*^\perp \right\rangle \left\langle \theta, \theta_*^\perp \right\rangle\right] \\
=&\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \left[\|\theta\|_2^2 - 2\mathbb{E}y \langle \mathbf{x}, \theta_* \rangle \langle \theta, \theta_* \rangle\right] \\
=&\operatorname*{argmin}_{\theta\in\mathbb{R}^d} \|\theta - \eta\theta_*\|_2^2,
\end{aligned}$$

where $\theta_*^\perp$ denotes the unit vector in the (θ_*, θ) plane orthogonal to θ_*, and the third equality follows from

L. Goldstein and S. Minsker are with the Department of Mathematics, University of Southern California, Los Angeles, CA (email: larry@usc.edu; minsker@usc.edu).

X. Wei is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA (email: xiaohanw@usc.edu).

S. Minsker and X. Wei were supported in part by the National Science Foundation grant DMS-1712956.

[1]The precise assumption on f is specified after (16).

the fact that $\langle \mathbf{x}, \theta_*^\perp \rangle$ is a centered Gaussian random variable independent of $\langle \mathbf{x}, \theta_* \rangle$, hence also independent of y.

Later, in [13], this result was extended to the more general class of elliptically symmetric distributions, which includes Gaussian distributions as a special case; see Lemma 5. In general, it is not always possible to recover θ_*: see [14] for an example in the case when $f(x,\delta) = sign(x) + \delta$ (so-called "1-bit compressed sensing" [15]).

Y. Plan, R. Vershynin and E. Yudovina recently presented the non-asymptotic study for the case of Gaussian measurements in the context of high-dimensional structured estimation, see [16], [17], [18]; we refer the reader to [19], [14], [20], [21] for further details. On a high level, these works show that when $\mathbf{x_j}$'s are Gaussian, nonlinearity can be treated as an additional noise term. To give an example, works [17] and [16] demonstrate that under the same model as (1), when $\mathbf{x}_j \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ and $\theta_* \in \Theta$, solving the constrained problem

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \ \|\mathbf{y} - \mathbf{X}\theta\|_2^2, \tag{2}$$

with $\mathbf{y} = [y_1 \ \cdots \ y_m]^T$ and $\mathbf{X} = \frac{1}{\sqrt{m}}[\mathbf{x}_1 \ \cdots \ \mathbf{x}_m]^T$, results in the following bound: with probability ≥ 0.99,

$$\left\| \widehat{\theta} - \eta\theta_* \right\|_2 \leq C \frac{\sigma_1 \omega(D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}) + \sigma_2}{\sqrt{m}}, \tag{3}$$

where (with formal definitions to follow in Section II) C is an absolute constant,

$$\begin{aligned} &g \sim \mathcal{N}(0,1), \quad \eta = \mathbb{E}(f(g,\delta)g), \\ &\sigma_1^2 = \mathbb{E}\big((f(g,\delta) - \eta g)^2\big), \\ &\sigma_2^2 = \mathbb{E}\big((f(g,\delta) - \eta g)^2 g^2\big), \end{aligned}$$

$\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$, $D(\Theta, \theta)$ is the descent cone of Θ at point θ, and $\omega(T)$ is the Gaussian mean width of a subset $T \subset \mathbb{R}^d$. A different approach to estimation of the index vector in model (1) with similar recovery guarantees has been developed in [21]. However, the key assumption adopted in all these works that the vectors $\{\mathbf{x}_j\}_{j=1}^m$ follow Gaussian distributions preclude situations where the measurements are heavy tailed, and hence might be overly restrictive for some practical applications; for example, noise and outliers observed in high-dimensional image recovery often exhibit heavy-tailed behavior, see [22].

As we mentioned above, in [13] authors have shown that direct consistent estimation of θ_* is possible when Π belongs to a family of elliptically symmetric distributions. Our main contribution is the non-asymptotic analysis for this scenario, with a particular focus on the case when $d > n$ and θ_* possesses special structure, such as sparsity. Moreover, we make very mild assumptions on the tails of the response variable y: for example, when the link function satisfies $f(\langle \mathbf{x}, \theta_* \rangle, \delta) = \tilde{f}(\langle \mathbf{x}, \theta_* \rangle) + \delta$, it is only assumed that δ possesses $2+\varepsilon$ moments, for some $\varepsilon > 0$. In [17], Y. Plan and R. Vershynin present analysis for the Gaussian case and ask "Can the same kind of accuracy be expected for random non-Gaussian matrices?" In this paper, we give a positive answer to their question. To achieve our goal, we propose a Lasso-type estimator that admits tight probabilistic guarantees in spirit of (3) despite weak tail assumptions (see Theorem 1 below for details).

Proofs of related non-asymptotic results in the literature rely on special properties of Gaussian measures (see, for example, [23]). To handle a wider class of elliptically symmetric distributions, we rely on recent developments in generic chaining methods, see [24], [25]. These general tools could prove useful in developing further extensions to a wider class of design distributions.

II. Definitions and background material.

This section introduces main notation and the key facts related to elliptically symmetric distributions, convex geometry and empirical processes. The results of this section will be used repeatedly throughout the paper. For the unified treatment of vectors and matrices, it will be convenient to treat a vector $v \in \mathbb{R}^{d\times 1}$ as a $d \times 1$ matrix. Let $d_1, d_2 \in \mathbb{N}$ be such that $d_1 d_2 = d$. Given $v_1, v_2 \in \mathbb{R}^{d_1 \times d_2}$, the Euclidean dot product is then defined as $\langle v_1, v_2 \rangle = \operatorname{tr}(v_1^T v_2)$, where $\operatorname{tr}(\cdot)$ stands for the trace of a matrix and v^T denotes the transpose of v.
The ℓ_1-norm of $v \in \mathbb{R}^d$ is defined as $\|v\|_1 = \sum_{j=1}^d |v_j|$. The nuclear norm of a matrix $v \in \mathbb{R}^{d_1 \times d_2}$ is $\|v\|_* = \sum_{j=1}^{\min(d_1,d_2)} \sigma_j(v)$, where $\sigma_j(v)$, $j = 1, \ldots, \min(d_1, d_2)$ stand for the singular values of v, and the operator norm is defined as $\|v\| = \max_{j=1,\ldots,\min(d_1,d_2)} \sigma_j(v)$.

A. Elliptically symmetric distributions.

A centered random vector $\mathbf{x} \in \mathbb{R}^d$ has elliptically symmetric (alternatively, elliptically contoured or just elliptical) distribution with parameters $\mathbf{\Sigma}$ and F_μ, denoted $\mathbf{x} \sim \mathcal{E}(0, \ \mathbf{\Sigma}, \ F_\mu)$, if

$$\mathbf{x} \stackrel{d}{=} \mu \mathbf{B} U, \tag{4}$$

where $\stackrel{d}{=}$ denotes equality in distribution, μ is a scalar random variable with cumulative distribution function F_μ, $\mathbf{B}$ is a fixed $d \times d$ matrix such that $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T$, and U is uniformly distributed over the unit sphere $\mathbb{S}^{d-1}$ and independent of μ. Note that distribution $\mathcal{E}(0, \ \mathbf{\Sigma}, \ F_\mu)$ is well defined, as if $\mathbf{B}_1\mathbf{B}_1^T = \mathbf{B}_2\mathbf{B}_2^T$, then there exists a unitary matrix $\mathbf{Q}$ such that $\mathbf{B}_1 =$

$\mathbf{B}_2\mathbf{Q}$, and $\mathbf{Q}U \stackrel{d}{=} U$. Along these same lines, we note that representation (4) is not unique, as one may replace the pair $(\mu,\ \mathbf{B})$ with $\left(c\mu,\ \frac{1}{c}\mathbf{B}\mathbf{Q}\right)$ for any constant $c > 0$ and any orthogonal matrix $\mathbf{Q}$. To avoid such ambiguity, in the following we allow $\mathbf{B}$ to be any matrix satisfying $\mathbf{B}\mathbf{B}^T = \mathbf{\Sigma}$, and noting that the covariance matrix of U is a multiple of the identity, we further impose the condition that the covariance matrix of $\mathbf{x}$ is equal to $\mathbf{\Sigma}$, i.e. $\mathbb{E}\left(\mathbf{x}\mathbf{x}^T\right) = \mathbf{\Sigma}$.

Alternatively, the mean-zero elliptically symmetric distribution can be defined uniquely via its characteristic function

$$\mathbf{s} \to \psi\left(\mathbf{s}^T\mathbf{\Sigma}\mathbf{s}\right),\ \mathbf{s} \in \mathbb{R}^d,$$

where $\psi : \mathbb{R}^+ \to \mathbb{R}$ is called the characteristic generator of $\mathbf{x}$. See [26] for further details about elliptical distribution.

An important special case of the family $\mathcal{E}(0,\ \mathbf{\Sigma},\ F_\mu)$ of elliptical distributions is the Gaussian distribution $\mathcal{N}(0,\mathbf{\Sigma})$, where $\mu = \sqrt{z}$ with $z \stackrel{d}{=} \chi^2_d$, and the characteristic generator is $\psi(x) = e^{-x/2}$.

The following elliptical symmetry property, generalizing the well known fact for the conditional distribution of the multivariate Gaussian, plays an important role in our subsequent analysis, see [27]:

Proposition 1. *Let* $\mathbf{x} = [\mathbf{x}_1,\ \mathbf{x}_2] \sim \mathcal{E}_d(0,\mathbf{\Sigma},F_\mu)$, *where are of dimension* d_1 *and* d_2 *respectively, with* $d_1 + d_2 = d$. *Let* $\mathbf{\Sigma}$ *be partitioning accordingly as*

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

Then, whenever $\mathbf{\Sigma}_{22}$ *has full rank, the conditional distribution of* $\mathbf{x}_1$ *given* $\mathbf{x}_2$ *is elliptical* $\mathcal{E}_{d_1}(0,\mathbf{\Sigma}_{1|2},F_{\mu_{1|2}})$, *where*

$$\mathbf{\Sigma}_{1|2} = \mathbf{\Sigma}_{11} - \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21},$$

and $F_{\mu_{1|2}}$ *is the cumulative distribution function of* $(\mu^2 - \mathbf{x}_2^T\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_2)^{1/2}$ *given* $\mathbf{x}_2$.

Note that $\mu^2 - \mathbf{x}_2^T\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_2$ is always nonnegative, hence $F_{\mu_{1|2}}$ is well defined, since by (4) we have

$$\begin{aligned}\mathbf{x}_2^T\mathbf{\Sigma}_{22}^{-1}\mathbf{x}_2 &= \mu^2(\mathbf{B}_2U)^T(\mathbf{B}_2\mathbf{B}_2^T)^{-1}(\mathbf{B}_2U) \\ &= \mu^2U^T\mathbf{B}_2^T(\mathbf{B}_2\mathbf{B}_2^T)^{-1}\mathbf{B}_2U \le \mu^2U^TU = \mu^2,\end{aligned}$$

where $\mathbf{B}_2$ is the matrix consisting of the last d_2 rows of $\mathbf{B}$ in (4), and where the inequality holds due to the fact that $\mathbf{B}_2^T(\mathbf{B}_2\mathbf{B}_2^T)^{-1}\mathbf{B}_2$ is a projection matrix. The following corollary is easily deduced from the theorem above:

Corollary 1. *If* $\mathbf{x} \sim \mathcal{E}_d(0,\mathbf{\Sigma},F_\mu)$ *with* $\mathbf{\Sigma}$ *of full rank, then for any two fixed vectors* $\mathbf{y}_1,\mathbf{y}_2 \in \mathbb{R}^d$ *with* $\|\mathbf{y}_2\|_2 = 1$,

$$\mathbb{E}(\langle\mathbf{x},\mathbf{y}_1\rangle \mid \langle\mathbf{x},\mathbf{y}_2\rangle) = \langle\mathbf{y}_1,\mathbf{y}_2\rangle\langle\mathbf{x},\mathbf{y}_2\rangle.$$

Proof. Let $\{\mathbf{v}_1,\cdots,\mathbf{v}_d\}$ be an orthonormal basis in $\mathbb{R}^d$ such that $\mathbf{v}_d = \mathbf{y}_2$. Let $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_d]$ and consider the linear transformation

$$\widetilde{\mathbf{x}} = \mathbf{V}^T\mathbf{x}.$$

Then, by (4), $\widetilde{\mathbf{x}} = \mu\mathbf{V}^T\mathbf{B}U$, which is centered elliptical with full rank covariance matrix $\mathbf{V}^T\mathbf{\Sigma}\mathbf{V}$. Applications of Theorem 1 with $\mathbf{x}_1 = [\langle\mathbf{x},\mathbf{v}_1\rangle,\ \cdots,\ \langle\mathbf{x},\mathbf{v}_{d-1}\rangle]$ and $\mathbf{x}_2 = \langle\mathbf{x},\mathbf{v}_d\rangle = \langle\mathbf{x},\mathbf{y}_2\rangle$ yields

$$\begin{aligned}&\mathbb{E}(\langle\mathbf{x},\mathbf{y}_1\rangle \mid \langle\mathbf{x},\mathbf{y}_2\rangle) \\ =&\mathbb{E}\left(\sum_{i=1}^{d}\langle\mathbf{x},\mathbf{v}_i\rangle\langle\mathbf{y}_1,\mathbf{v}_i\rangle \;\middle|\; \langle\mathbf{x},\mathbf{v}_d\rangle\right) \\ =&\mathbb{E}\left(\sum_{i=1}^{d-1}\langle\mathbf{x},\mathbf{v}_i\rangle\langle\mathbf{y}_1,\mathbf{v}_i\rangle \;\middle|\; \langle\mathbf{x},\mathbf{v}_d\rangle\right) + \langle\mathbf{x},\mathbf{v}_d\rangle\langle\mathbf{y}_1,\mathbf{v}_d\rangle \\ =&\langle\mathbf{x},\mathbf{v}_d\rangle\langle\mathbf{y}_1,\mathbf{v}_d\rangle = \langle\mathbf{y}_1,\mathbf{y}_2\rangle\langle\mathbf{x},\mathbf{y}_2\rangle,\end{aligned}$$

where in the second to last equality we have used the fact that the conditional distribution of $[\langle\mathbf{v}_1,\mathbf{x}\rangle,\ \cdots,\ \langle\mathbf{v}_{d-1},\mathbf{x}\rangle]$ given $\langle\mathbf{x},\mathbf{v}_d\rangle$ is elliptical with mean zero. □

B. Geometry.

In this section, we recall the definitions of several quantities that control the "complexity" of the estimation problem in model (1).

Definition 1 (Gaussian mean width)**.** *The Gaussian mean width of a set* $T \subseteq \mathbb{R}^d$ *is defined as*

$$\omega(T) := \mathbb{E}\left(\sup_{t\in T}\ \langle\mathbf{g},t\rangle\right),$$

where $\mathbf{g} \sim \mathcal{N}(0,\mathbf{I}_{d\times d})$.

Definition 2 (Descent cone)**.** *The descent cone of a set* $\Theta \subseteq \mathbb{R}^d$ *at a point* $\theta \in \mathbb{R}^d$ *is defined as*

$$D(\Theta,\theta) = \{\tau\mathbf{h} :\ \tau \ge 0, \mathbf{h} \in \Theta - \theta\}.$$

For example, when $T = \mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$, it is easy to see that $\omega(\mathbb{S}^{d-1}) = \mathbb{E}(\|\mathbf{g}\|_2) \sim \sqrt{d}$. We will be interested in the Gaussian mean widths of subsets of the unit sphere of the form $T = \mathbb{S}^{d-1} \cap D(\Theta,\theta)$, where θ lies on the boundary of Θ; the importance of such subsets in structured recovery is explained in [1].

Definition 3 (Restricted set)**.** *Given* $c_0 > 1$, *the* c_0*-restricted set of the norm* $\|\cdot\|_{\mathcal{K}}$ *at* $\theta \in \mathbb{R}^d$ *is defined as*

$$\begin{aligned}&\mathbb{S}_{c_0}(\theta) := \mathbb{S}_{c_0}(\theta;\mathcal{K}) \\ &= \left\{\mathbf{v} \in \mathbb{R}^d :\ \|\theta+\mathbf{v}\|_{\mathcal{K}} \le \|\theta\|_{\mathcal{K}} + \frac{1}{c_0}\|\mathbf{v}\|_{\mathcal{K}}\right\}. \quad (5)\end{aligned}$$

Restricted set is similar to the "cone of dominant coordinates" that appears in the analysis of sparse recovery problems; we provide more details and examples in the Appendix.

Definition 4 (Restricted compatibility)**.** *The restricted compatibility constant of a set $A \subseteq \mathbb{R}^d$ with respect to the norm $\|\cdot\|_{\mathcal{K}}$ is given by*

$$\Psi(A) := \Psi(A; \mathcal{K}) = \sup_{\mathbf{v} \in A \setminus \{0\}} \frac{\|\mathbf{v}\|_{\mathcal{K}}}{\|\mathbf{v}\|_2}.$$

The restricted compatibility concept is introduced to capture the dependence of the equivalence constants between two norms on the geometry of the set under consideration.

Remark 1. *The restricted set from the Definition 3 is not necessarily convex. However, if the norm $\|\cdot\|_{\mathcal{K}}$ is decomposable (see Definition 8), then the restricted set is contained in a convex cone, and the corresponding restricted compatibility constant is easier to estimate. Decomposable norms have been introduced by [28] and later appeared in a number of works, e.g. see [29] and references therein. For the reader's convenience, we provide a self-contained discussion in the Appendix.*

III. Main results.

In this section, we define a version of Lasso estimator that is well-suited for heavy-tailed measurements, and state its performance guarantees.

We will assume that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^d$ are i.i.d. copies of an **isotropic** vector $\mathbf{x}$ with spherically symmetric distribution $\mathcal{E}_d(0, \mathbf{I_{d\times d}}, F_\mu)$. If $\mathbf{x} \sim \mathcal{E}_d(0, \mathbf{\Sigma}, F_\mu)$ for some positive definite matrix $\mathbf{\Sigma}$, then by definition $\mathbf{x} \overset{d}{=} \mu \mathbf{\Sigma}^{1/2} U$, and $\langle \mathbf{x}, \theta_* \rangle = \left\langle \mathbf{\Sigma}^{-1/2}\mathbf{x}, \mathbf{\Sigma}^{1/2}\theta_* \right\rangle$, where $\mathbf{\Sigma^{-1/2}x} = \mu U \sim \mathcal{E}_d(0, \mathbf{I_{d\times d}}, F_\mu)$. Hence, if we set $\tilde{\theta}_* := \mathbf{\Sigma}^{1/2}\theta_*$, then all results that we establish for isotropic measurements hold with θ_* replaced by $\tilde{\theta}_*$; remark after Theorem 1 includes more details.

A. Description of the proposed estimator.

We first introduce an estimator under the scenario that $\theta_* \in \Theta$, for some known closed set $\Theta \subseteq \mathbb{R}^d$. Define the loss function $L_m^0(\cdot)$ as

$$L_m^0(\theta) := \|\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \langle y_i \mathbf{x}_i, \theta \rangle, \tag{6}$$

which is the unbiased estimator of

$$L^0(\theta) := \|\theta\|_2^2 - 2\mathbb{E}\langle y\mathbf{x}, \theta \rangle = \mathbb{E}\left(y - \langle x, \theta \rangle\right)^2 - \mathbb{E}y^2,$$

where the last equality follows since x is isotropic. Clearly, minimizing $L^0(\theta)$ over any set $\Theta \subseteq \mathbb{R}^d$ is equivalent to minimizing the quadratic loss $\mathbb{E}\left(y - \langle \mathbf{x}, \theta \rangle\right)^2$. If distribution F_μ has heavy tails, the sample average $\frac{1}{m}\sum_{i=1}^m y_i \mathbf{x}_i$ might not concentrate sufficiently well around its mean, hence we replace it by a more "robust" version obtained via truncation. Let $\mu \in \mathbb{R}$, $U \in \mathbb{S}^{d-1}$ be such that $\mathbf{x} = \mu U$ (so that $\mu = \|\mathbf{x}\|_2$), and set

$$\begin{aligned} \widetilde{U} &= \sqrt{d}U, \\ q &= \mu y / \sqrt{d}, \end{aligned} \tag{7}$$

so that $q\widetilde{U} = y\mathbf{x}$ and $\widetilde{U}$ is uniformly distributed on the sphere of radius $\sqrt{d}$, implying that its covariance matrix is I_d, the identity matrix. Next, define the truncated random variables

$$\widetilde{q}_i = \operatorname{sign}(q_i)(|q_i| \wedge \tau), \ i = 1, \ldots, m, \tag{8}$$

where $\tau = m^{\frac{1}{2(1+\kappa)}}$ for some $\kappa \in (0,1)$ that is chosen based on the integrability properties of q, see (17). Finally, set

$$L_m^\tau(\theta) = \|\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i \widetilde{U}_i, \theta \right\rangle, \tag{9}$$

and define the estimator $\widehat{\theta}_m$ as the solution to the constrained optimization problem:

$$\widehat{\theta}_m := \operatorname*{argmin}_{\theta \in \Theta} L_m^\tau(\theta). \tag{10}$$

We will also denote

$$L^\tau(\theta) := \mathbb{E}L_m^\tau(\theta) = \|\theta\|_2^2 - 2\mathbb{E}\left\langle \widetilde{q}\widetilde{U}, \theta \right\rangle. \tag{11}$$

For the scenarios where structure on the unknown θ_* is induced by a norm $\|\cdot\|_{\mathcal{K}}$ (e.g., if θ_* is sparse, then $\|\cdot\|_{\mathcal{K}}$ could be the $\|\cdot\|_1$ norm), we will also consider the estimator $\widehat{\theta}_m^\lambda$ defined via

$$\widehat{\theta}_m^\lambda := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left[L_m^\tau(\theta) + \lambda \|\theta\|_{\mathcal{K}} \right], \tag{12}$$

where $\lambda > 0$ is a regularization parameter to be specified, and $L_m^\tau(\theta)$ is defined in (9).

Let us note that truncation approach has previously been successfully implemented in [30] to handle heavy-tailed noise in the context of matrix recovery with sub-Gaussian design. In the present paper, we show that truncation-based approach is also useful in the situations where the measurements are heavy-tailed.

Remark 2.
1) *In the special case when the measurement vector $\mathbf{x}$ is Gaussian, Y. Plan and R. Vershynin [18] proposed and analyzed an estimator similar to (6) in the framework of 1-bit compressed sensing.*
2) *Note that our estimator (12) is in general much easier to implement than some other popular alternatives, such as the usual Lasso estimator*

[2]. For example, when the signal θ is sparse, our estimator takes the form

$$\widehat{\theta}_m^\lambda := \underset{\theta\in\mathbb{R}^d}{\mathit{argmin}} \left[\|\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i\widetilde{U}_i, \theta\right\rangle + \lambda\|\theta\|_1\right],$$

which yields a closed form solution in the form of "soft-thresholding". Specifically, let $\mathbf{b} = \frac{1}{m}\sum_{i=1}^m \widetilde{q}_i\widetilde{U}_i$*, then, the* k*-th entry of* $\widehat{\theta}_m^\lambda$ *takes the form:*

$$\left(\widehat{\theta}_m^\lambda\right)_k = \begin{cases} b_k - \lambda/2, & \text{if } b_k \geq \lambda/2, \\ 0, & \text{if } -\lambda/2 \leq b_k \leq \lambda/2, \\ b_k + \lambda/2, & \text{if } b_k \leq -\lambda/2. \end{cases} \quad (13)$$

We should note however that such simplification comes at the cost of prior knowledge that the measurement vector $\mathbf{x}$ *is isotropic. Despite having low computational complexity, our estimator can still exploit the structure of the problem, while being robust both to the possible model misspecification as well as to data corruption modeled by the heavy-tailed distributions. We demonstrate this in the following sections.*

Remark 3 (Non-isotropic measurements)**.** *When* $\mathbf{x} \sim \mathcal{E}_d(0, \mathbf{\Sigma}, F_\mu)$ *for some* $\Sigma \succ 0$*, then estimator (10) has to be replaced by*

$$\widehat{\theta}_m := \underset{\theta\in\Theta}{\mathit{argmin}} \left[\|\mathbf{\Sigma}^{1/2}\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i\widetilde{U}_i, \mathbf{\Sigma}^{1/2}\theta\right\rangle\right], \quad (14)$$

which is equivalent to

$$\tilde{\theta}_m := \underset{\theta\in\mathbf{\Sigma}^{1/2}\Theta}{\mathit{argmin}} \left[\|\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i\widetilde{U}_i, \theta\right\rangle\right],$$

is a sense that $\tilde{\theta}_m = \mathbf{\Sigma}^{1/2}\hat{\theta}_m$*. Hence, results obtained for isotropic measurements easily extend to the more general case. Similarly, estimator (12) should be replaced by*

$$\hat{\theta}_m^\lambda := \underset{\theta\in\mathbb{R}^d}{\mathit{argmin}} \left[\|\mathbf{\Sigma}^{1/2}\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i\widetilde{U}_i, \Sigma^{1/2}\theta\right\rangle + \lambda\|\mathbf{\Sigma}^{1/2}\theta\|_{\mathcal{K}}\right], \quad (15)$$

which is equivalent to

$$\tilde{\theta}_m^\lambda := \underset{\theta\in\mathbb{R}^d}{\mathit{argmin}} \left[\|\theta\|_2^2 - \frac{2}{m}\sum_{i=1}^m \left\langle \widetilde{q}_i\widetilde{U}_i, \theta\right\rangle + \lambda\|\theta\|_{\mathbf{\Sigma}^{1/2}\mathcal{K}}\right],$$

meaning that $\tilde{\theta}_m^\lambda = \mathbf{\Sigma}^{1/2}\hat{\theta}_m^\lambda$*.*

B. Estimator performance guarantees.

In this section, we present the probabilistic guarantees for the performance of the estimators $\widehat{\theta}_m$ and $\widehat{\theta}_m^\lambda$ defined by (10) and (12) respectively.
Everywhere below, C, c, C_j denote numerical constants; when these constants depend on parameters of the problem, we specify this dependency by writing $C_j = C_j(\text{parameters})$. Let

$$\eta = \mathbb{E}\left\langle y\mathbf{x}, \theta_*\right\rangle, \quad (16)$$

and assume that $\eta \neq 0$ and $\eta\theta_* \in \Theta$.

Theorem 1. *Suppose that* $\mathbf{x} \sim \mathcal{E}(0, \mathbf{I}_{d\times d}, F_\mu)$*. Moreover, suppose that for some* $\kappa > 0$

$$\phi := \mathbb{E}|q|^{2(1+\kappa)} < \infty. \quad (17)$$

Then there exist constants $C_1 = C_1(\kappa, \phi), C_2 = C_2(\kappa, \phi) > 0$ *such that* $\widehat{\theta}_m$ *satisfies*

$$\mathbb{P}\left(\left\|\widehat{\theta}_m - \eta\theta_*\right\|_2 \geq C_1\frac{(\omega(D(\Theta, \eta\theta_*)\cap\mathbb{S}^{d-1}) + 1)\beta}{\sqrt{m}}\right) \leq C_2 e^{-\beta/2},$$

for any $m \geq \beta^2\left(\omega(D(\Theta, \eta\theta_*)\cap\mathbb{S}^{d-1}) + 1\right)^2$*,* $\beta \geq 8$*.*

Remark 4.
1) *Unknown link function* f *enters the bound only through the constant* η *defined in (16).*
2) *Aside from independence, conditions on the noise* δ *are implicit and follow from assumptions on* y*. In the special case when the error is additive, that is, when* $y = f(\langle\mathbf{x}, \theta_*\rangle) + \delta$*, the moment condition (17) becomes* $\mathbb{E}\Big|\|\mathbf{x}\|_2 f(\langle\mathbf{x}, \theta_*\rangle) + \|\mathbf{x}\|_2\delta\Big|^{2(1+\kappa)} < \infty$*, for which it is sufficient to assume that* $\mathbb{E}\Big|\|\mathbf{x}\|_2 f(\langle\mathbf{x}, \theta_*\rangle)\Big|^{2(1+\kappa)} < \infty$ *and* $\mathbb{E}\left|\|\mathbf{x}\|_2\delta\right|^{2(1+\kappa)} < \infty$*.*
3) *Theorem 1 is mainly useful when* $\eta\theta_*$ *lies on the boundary of the set* Θ*. Otherwise, if* $\eta\theta_*$ *belongs to the relative interior of* Θ*, the descent cone* $D(\Theta, \eta\theta_*)$ *is the affine hull of* Θ *(which will often be the whole space* $\mathbb{R}^d$*). Thus, in such cases the Gaussian mean width* $\omega(D(\Theta, \eta\theta_*)\cap\mathbb{S}^{d-1})$ *can be on the order of* $\sqrt{d}$*, which is prohibitively large when* $d \gg m$*. We refer the reader to [17], [16] for a discussion of related result and possible ways to tighten them.*
4) *As we mentioned in the introduction, for the special case* $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$*, the work [17] shows that solution of the problem (2) satisfies bound (3) with high probability. It is worth comparing the aforementioned bound with our results. Note that (3) explicitly captures the effect of nonlinearity on the error bound via two parameters* σ_1 *and* σ_2*, whereas in our bound is the dependence on these parameters is merged into an (unspecified)*

constant $C_1(\kappa, \phi)$, mainly due to the limitations imposed by the more involved proof techniques required to handle the heavy-tailed distributions.

5) *In the special case where only the noise variable has heavy-tailed distribution while the measurement vector $\mathbf{x}$ is sub-Gaussian, an alternative approach based on LAD regression or Huber's regression is possible; non-asymptotic analysis has been performed in several recent works, for example see [31], [32]. However, construction of the resulting estimators typically requires either the known upper bound on the variance of the noise, or the symmetry of the distribution of the noise variable.*

Next, we present performance guarantees for the unconstrained estimator (12).

Theorem 2. *Assume that the norm $\|\cdot\|_\mathcal{K}$ dominates the 2-norm, i.e. $\|\mathbf{v}\|_\mathcal{K} \geq \|\mathbf{v}\|_2,\ \forall \mathbf{v} \in \mathbb{R}^d$. Let $\mathbf{x} \sim \mathcal{E}(0,\ \mathbf{I}_{d\times d},\ F_\mu)$, and suppose that for some $\kappa > 0$*

$$\phi := \mathbb{E}|q|^{2(1+\kappa)} < \infty.$$

Then there exist constants $C_3 = C_3(\kappa, \phi), C_4 = C_4(\kappa, \phi) > 0$ such that for all $\lambda \geq \frac{C_3\beta}{\sqrt{m}}(1 + \omega(\mathcal{G}))$

$$\mathbb{P}\left(\left\|\widehat{\theta}_m^\lambda - \eta\theta_*\right\|_2 \geq \frac{3}{2}\lambda \cdot \Psi\left(\mathbb{S}_2\left(\eta\theta_*\right)\right)\right) \leq C_4 e^{-\beta/2},$$

for any $\beta \geq 8$ and $m \geq (\omega(\mathcal{G}) + 1)^2\beta^2$, where $\mathcal{G} := \{\mathbf{x} \in \mathbb{R}^d :\ \|\mathbf{x}\|_\mathcal{K} \leq 1\}$ is the unit ball of $\|\cdot\|_\mathcal{K}$ norm, and $\mathbb{S}_2(\cdot)$ and $\Psi(\cdot)$ are given in Definitions 3 and 4 respectively.

Remark 5 (Non-isotropic measurements)**.** *It follows from remark 3 and (14) that, whenever $\mathbf{x} \sim \mathcal{E}_d(0, \mathbf{\Sigma}, F_\mu)$, inequality of Theorem 1 has the form*

$$\mathbb{P}\left(\left\|\mathbf{\Sigma}^{1/2}\left(\widehat{\theta}_m - \eta\theta_*\right)\right\|_2 \geq C_1\frac{\left(\omega\left(\mathbf{\Sigma}^{1/2}D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}\right) + 1\right)\beta}{\sqrt{m}}\right) \leq C_2 e^{-\beta/2},$$

which can be further combined with the bound

$$\omega\left(\mathbf{\Sigma}^{1/2}D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}\right) \leq \|\mathbf{\Sigma}^{1/2}\| \cdot \|\mathbf{\Sigma}^{-1/2}\|\, \omega\left(D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}\right),$$

that follows from remark 1.7 in [17]. Similarly, the inequality of Theorem 2 holds with

$$\mathcal{G}_{\mathbf{\Sigma}^{1/2}} := \{\mathbf{x} \in \mathbb{R}^d :\ \|\mathbf{x}\|_{\mathbf{\Sigma}^{1/2}\mathcal{K}} \leq 1\},$$

the unit ball of $\|\cdot\|_{\mathbf{\Sigma}^{1/2}\mathcal{K}}$ norm, in place of $\mathcal{G}$. Namely, for all $\lambda \geq \frac{C_3\beta}{\sqrt{m}}(1 + \omega(\mathcal{G}_{\mathbf{\Sigma}^{1/2}}))$,

$$\mathbb{P}\left(\left\|\mathbf{\Sigma}^{1/2}\left(\widehat{\theta}_m^\lambda - \eta\theta_*\right)\right\|_2 \geq \frac{3}{2}\lambda \cdot \Psi\left(\mathbb{S}_2\left(\eta\mathbf{\Sigma}^{1/2}\theta_*\right); \mathbf{\Sigma}^{1/2}\mathcal{K}\right)\right) \leq C_4 e^{-\beta/2}$$

Note that $\omega\left(\mathcal{G}_{\mathbf{\Sigma}^{1/2}}\right) \leq \|\mathbf{\Sigma}^{1/2}\|\,\omega(\mathcal{G})$. Moreover, we show in the Appendix that for a class of decomposable norms (which includes $\|\cdot\|_1$ and nuclear norm), the upper bounds for $\Psi\left(\mathbb{S}_2\left(\eta\mathbf{\Sigma}^{1/2}\theta_*\right); \mathbf{\Sigma}^{1/2}\mathcal{K}\right)$ and $\Psi\left(\mathbb{S}_2(\eta\theta_*)\right)$ differ by the factor of $\left\|\mathbf{\Sigma}^{-1/2}\right\|$.

C. Examples.

We discuss two popular scenarios: estimation of the sparse vector and estimation of the low-rank matrix.

Estimation of the sparse signal. Assume that there exists $J \subseteq \{1, \ldots, d\}$ of cardinality $s \leq d$ such that $\theta_{*,j} = 0$ for $j \notin J$. Let $\Theta = \left\{\theta \in \mathbb{R}^d :\ \|\theta\|_1 \leq \|\eta\theta_*\|_1\right\}$, with η defined in (16). In this case, it is well-known that $\omega^2\left(D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}\right) \leq 2s\log(d/s) + \frac{3}{2}s$, see equation (8) in [33], hence Theorem 1 implies that, with high probability,

$$\left\|\widehat{\theta}_m - \eta\theta_*\right\|_2 \lesssim \sqrt{\frac{s\log(d/s)}{m}} \tag{18}$$

as long as $m \gtrsim s\log(d/s)$.

We compare this bound to result of Theorem 2 for constrained estimator. Let $\|\cdot\|_\mathcal{K}$ be the ℓ_1 norm. It is well-know that $\omega(\mathcal{G}) = \mathbb{E}\max_{j=1,\ldots,d}|g_j| \leq \sqrt{2\log(2d)}$, where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$. Moreover, we show in the Appendix that $\Psi\left(\mathbb{S}_2\left(\eta\theta_*\right)\right) \leq 4\sqrt{s}$. Hence, for $\lambda \simeq \sqrt{\frac{\log(2d)}{m}}$, Theorem 2 implies that

$$\left\|\widehat{\theta}_m^\lambda - \eta\theta_*\right\|_2 \lesssim \sqrt{\frac{s\log(d)}{m}}$$

with high probability whenever $m \gtrsim \log(2d)$. This bound is only marginally weaker than (18) due to the logarithmic factor, however, definition of $\widehat{\theta}_m^\lambda$ does not require the knowledge of $\|\eta\theta_*\|_1$, as we have already mentioned before.

Estimation of low-rank matrices. Assume that $d = d_1 d_2$ with $d_1 \leq d_2$, and $\theta_* \in \mathbb{R}^{d_1\times d_2}$ has rank $r \leq \min(d_1, d_2)$. Let $\Theta = \left\{\theta \in \mathbb{R}^{d_1\times d_2} :\ \|\theta\|_* \leq \|\eta\theta_*\|_*\right\}$. Then the Gaussian mean width of the intersection of a descent cone with a unit ball is bounded as $\omega^2\left(D(\Theta, \eta\theta_*) \cap \mathbb{S}^{d-1}\right) \leq 3r(d_1 + d_2 - r)$ (see proposition 3.11 in [7]). Hence, Theorem 1 yields that with high probability,

$$\left\|\widehat{\theta}_m - \eta\theta_*\right\|_2 \lesssim \sqrt{\frac{r(d_1 + d_2)}{m}}$$

as long as the number of observations satisfies $m \gtrsim r(d_1 + d_2)$.
Finally, we derive the corresponding bound from Theorem 2. The Gaussian mean width of the unit ball in the nuclear norm is bounded by $2(\sqrt{d_1} + \sqrt{d_2})$, see proposition 10.3 in [1]. It follows from results in the Appendix that $\Psi(\mathbb{S}_2(\eta\theta_*)) \le 4\sqrt{2r}$. Theorem 2 now implies that with high probability

$$\left\| \widehat{\theta}_m - \eta\theta_* \right\|_2 \lesssim \sqrt{\frac{r(d_1 + d_2)}{m}},$$

which matches the bound of Theorem 1.

IV. Numerical experiments

In this section, we demonstrate the performance of proposed robust estimator (12) for one-bit compressed sensing model. The model takes the following form:

$$y = sign(\langle \mathbf{x}, \theta_* \rangle) + \delta, \tag{19}$$

where δ is the additive noise and the parameter θ^* is assumed to be s-sparse. This model is highly non-linear because one can only observe the sign of each measurement.

The 1-bit compressed sensing model was previously discussed extensively in a number of works, e.g. [16], [14], [17]. It was shown that when the measurement vectors are either Gaussian or sub-Gaussian, the Lasso estimator recovers the support of θ^* with high probability. Here, we show that under the heavy-tailed elliptically distributed measurements, our estimator numerically outperforms the standard Lasso estimator

$$\theta_{\text{Lasso}} = \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \quad \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \lambda\|\theta\|_1,$$

while taking the form of a simple soft-thresholding as explained in (13).

In the first numerical experiment, data are simulated in the following way: $\mathbf{x}_1, \ \mathbf{x}_2, \ \cdots, \ \mathbf{x}_{128} \in \mathbb{R}^{512}$ are i.i.d. with spherically symmetric distribution $\mathbf{x}_i \overset{d}{=} \mu_i U_i, \ i = 1, \ldots, n$. The random vectors $U_i \in \mathbb{R}^{512}$ are i.i.d. with uniform distribution over the sphere of radius $\sqrt{512}$, and the random variables $\mu_i \in \mathbb{R}$ are also i.i.d., independent of U_i and such that

$$\mu_i \overset{d}{=} \frac{1}{\sqrt{2c(q)}}(\xi_{i,1} - \xi_{i,2}), \tag{20}$$

where $\xi_{i,1}$ and $\xi_{i,2}, \ i = 1, 2, \cdots, 128$ are i.i.d. with Pareto distribution, meaning that their probability density function is given by

$$p(t; q) = \frac{q}{(1+t)^{1+q}} I_{\{t>0\}},$$

$c(q) := \text{Var}(\xi) = \frac{q}{(q-1)^2(q-2)}$, and $q = 2.1$. The true signal θ^* has sparsity level $s = 5$, with index of each non-zero coordinate chosen uniformly at random, and the magnitude having uniform distribution on $[0, 1]$.

Since we can only recover the original signal θ^* up to scaling, define the relative error for any estimator $\hat{\theta}$ with respect to θ^* as follows:

$$\text{Relative error} = \left| \frac{\hat{\theta}}{\|\hat{\theta}\|_2} - \frac{\theta^*}{\|\theta^*\|_2} \right|. \tag{21}$$

In each of the following two scenarios, we run the experiment 200 times for both the Lasso estimator and the estimator defined in (12) with $\|\cdot\|_{\mathcal{K}}$ being the $\|\cdot\|_1$ norm. We set the truncation level as $\tau = cm^{\frac{1}{2(1+\kappa)}}$, and the values of c and regularization parameter λ are obtained via the standard 2-fold cross validation for the relative error (21). We then plot the histogram of obtained results over 200 runs of the experiment.

In the first scenario, we set the additive error $\delta_i = 0, \ i = 1, 2, \cdots, 128$ in the 1-bit model (19) and plot the histogram in Fig. 1. We can see from the plot that the robust estimator (12) noticeably outperforms the Lasso estimator.

In the second scenario, we set the additive error $\delta_i, \ i = 1, 2, \cdots, 128$ to be i.i.d. heavy tailed noise with signal-to-noise ratio (SNR)[2] equal to 10dB, so that the noise has the distribution

$$\delta_i \overset{d}{=} h_i/\sqrt{10},$$

and $h_i, \ i = 1, 2, \cdots, 128$ are i.i.d. random variables with Pareto-type distribution (20). The results are plotted in Fig. 2. The histogram shows that, while performance of the Lasso estimator becomes worse, results of robust estimator (12) are relatively stable.

In the second simulation study, the simulation framework similar to the second scenario above, the only difference being the increased sample size m. The results are plotted in Fig. 3, 4 with sample sizes $m = 256$ and $m = 512$ respectively.

V. Final remarks

In this paper, we investigated the problem of structured signal recovery from nonlinear and heavy-tailed measurements. In particular, we focus on the scenario where the measurement vectors have an elliptical symmetric distribution, and propose an estimator that is robust both to non-linearity and heavy-tailed nature of the measurements.

Several questions remain open: first, the proposed estimator relies heavily on the prior knowledge of the true covariance matrix of the measurements, whereas the usual LASSO estimator is does not require such

[2]The signal-to-noise ratio (dB) is defined as SNR $:= 10\log_{10}(\sigma^2_{\text{signal}}/\sigma^2_{\text{noise}})$. In our case, since $\langle \mathbf{x}_i, \theta^* \rangle$ can be positive or negative with equal probability, $\sigma^2_{\text{signal}} = 1$, and thus, $\sigma^2_{\text{noise}} = 1/10$.

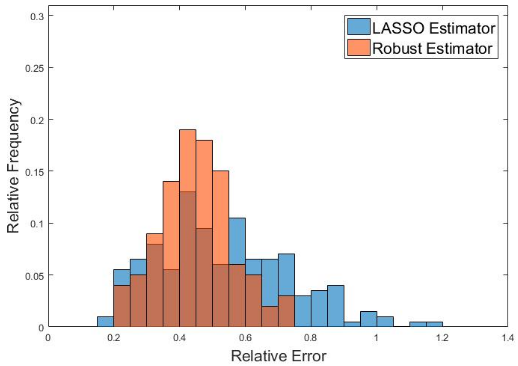

Fig. 1. Lasso vs robust estimator in the noiseless case.

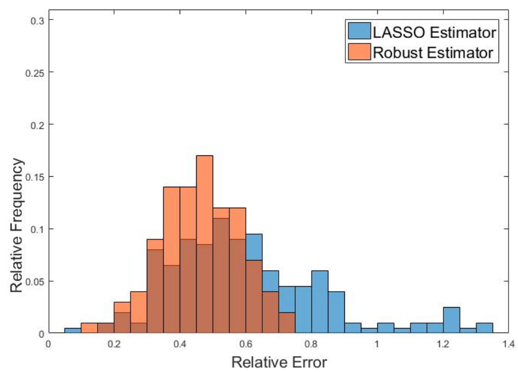

Fig. 2. Lasso vs robust estimator under heavy-tailed noise with signal-to-noise ratio(SNR) equal to $10dB$.

prior information. Is it possible to obtain strong theoretical guarantees in the case of heavy-tailed measurements when the covariance matrix is unknown?

Second (and perhaps more important) question asks whether one can extend results of the present paper beyond the class of elliptically symmetric distributions. We note that the case of non-Gaussian measurement vectors with i.i.d. entries has been investigated in several works including [14], [34] which showed that the least squares-type estimator is often biased, and the bias can be controlled by a certain distance (for instance, the total variation distance) between the distribution of the entries and the Gaussian law.

VI. Proofs.

This section is devoted to the proofs of Theorems 1 and 2.

A. Preliminaries.

We recall several useful facts from probability theory that we rely on in the subsequent analysis.
The following well-known bound shows that the uniform distribution on a high-dimensional sphere enjoys strong concentration properties.

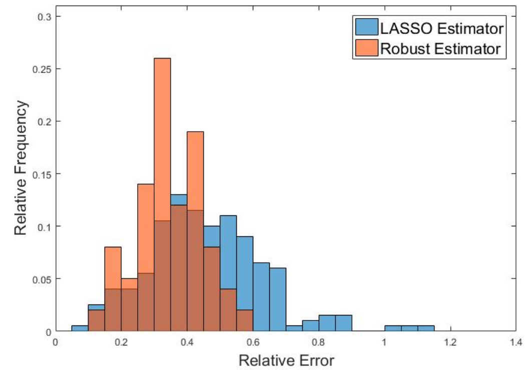

Fig. 3. Lasso vs robust estimator under heavy-tailed noise with signal-to-noise ratio(SNR) equal to $10dB$ and sample size $m = 256$.

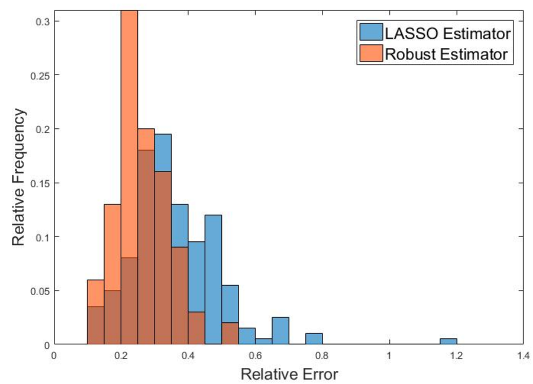

Fig. 4. Lasso vs robust estimator under heavy-tailed noise with signal-to-noise ratio(SNR) equal to $10dB$ and sample size $m = 512$.

Lemma 1 (Lemma 2.2 of [35]). *Let U have the uniform distribution on $\mathbb{S}^{d-1}$. Then for any $\Delta \in (0,1)$ and any fixed $\mathbf{v} \in \mathbb{S}^{d-1}$,*

$$\mathbb{P}\left(\langle U, \mathbf{v}\rangle \geq \Delta\right) \leq e^{-d\Delta^2/2}.$$

Next, we state several useful results from the theory of empirical processes.

Definition 5 (ψ_q-norm). *For $q \geq 1$, the ψ_q-norm of a random variable $\xi \in \mathbb{R}$ is given by*

$$\|\xi\|_{\psi_q} = \sup_{p\geq 1} p^{-\frac{1}{q}} \left(\mathbb{E}(|X|^p)\right)^{\frac{1}{p}}.$$

Specifically, the cases $q = 1$ and $q = 2$ are known as the sub-exponential and sub-Gaussian norms respectively. We will say that ξ is sub-exponential if $\|\xi\|_{\psi_1} < \infty$, and X is sub-Gaussian if $\|\xi\|_{\psi_2} < \infty$.

Remark 6. *It is easy to check that ψ_q-norm is indeed a norm.*

Remark 7. *A useful property, equivalent to the previous definition of a sub-Gaussian random variable ξ, is* *that there exists a positive constant C such that*

$$\mathbb{P}(|\xi| \geq u) \leq \exp(1 - Cu^2).$$

For the proof, see Lemma 5.5 in [36].

Definition 6 (sub-Gaussian random vector)**.** *A random vector $\mathbf{x} \in \mathbb{R}^d$ is called sub-Gaussian if there exists $C > 0$ such that $\|\langle \mathbf{x}, \mathbf{v}\rangle\|_{\psi_2} \leq C$ for any $\mathbf{v} \in \mathbb{S}^{d-1}$. The corresponding sub-Gaussian norm is then*

$$\|\mathbf{x}\|_{\psi_2} := \sup_{\mathbf{v} \in \mathbb{S}^{d-1}} \|\langle \mathbf{x}, \mathbf{v}\rangle\|_{\psi_2}.$$

Next, we recall the notion of the generic chaining complexity. Let (T, d) be a metric space. We say a collection $\{\mathcal{A}_l\}_{l=0}^{\infty}$ of subsets of T is increasing when $\mathcal{A}_l \subseteq \mathcal{A}_{l+1}$ for all $l \geq 0$.

Definition 7 (Admissible sequence)**.** *An increasing sequence of subsets $\{\mathcal{A}_l\}_{l=0}^{\infty}$ of T is admissible if $|\mathcal{A}_l| \leq N_l,\ \forall l$, where $N_0 = 1$ and $N_l = 2^{2^l},\ \forall l \geq 1$.*

For each $\mathcal{A}_l$, define the map $\pi_l : T \to \mathcal{A}_l$ as $\pi_l(t) = \arg\min_{s \in \mathcal{A}_l} d(s, t),\ \ \forall t \in T$. Note that, since each $\mathcal{A}_l$ is a finite set, the minimum is always achieved. When the minimum is achieved for multiple elements in $\mathcal{A}_l$, we break the ties arbitrarily. The generic chaining complexity γ_2 is defined as

$$\gamma_2(T, d) := \inf \sup_{t \in T} \sum_{l=0}^{\infty} 2^{l/2} d(t, \pi_l(t)), \tag{22}$$

where the infimum is over all admissible sequences. The following theorem tells us that γ_2-functional controls the "size" of a Gaussian process.

Lemma 2 (Theorem 2.4.1 of [24])**.** *Let $\{G(t),\ t \in T\}$ be a centered Gaussian process indexed by the set T, and let*

$$d(s, t) = \mathbb{E}\big((G(s) - G(t))^2\big)^{1/2},\ \forall s, t \in T.$$

Then, there exists a universal constant L such that

$$\frac{1}{L}\gamma_2(T, d) \leq \mathbb{E}\left(\sup_{t \in T} G(t)\right) \leq L\gamma_2(T, d).$$

Let (T, d) be a semi-metric space, and let $X_1(t), \cdots, X_m(t)$ be independent stochastic processes indexed by T such that $\mathbb{E}|X_j(t)| < \infty$ for all $t \in \mathbb{T}$ and $1 \leq j \leq m$. We are interested in bounding the supremum of the empirical process

$$Z_m(t) = \frac{1}{m}\sum_{i=1}^{m}\left[X_i(t) - \mathbb{E}(X_i(t))\right]. \tag{23}$$

The following well-known symmetrization inequality reduces the problem to bounds on a (conditionally) Rademacher process $R_m(t) = \frac{1}{m}\sum_{i=1}^m \varepsilon_i X_i(t),\ t \in T$, where $\varepsilon_1, \ldots, \varepsilon_m$ are i.i.d. Rademacher random variables (meaning that they take values $\{-1, +1\}$ with probability $1/2$ each), independent of X_i's.

Lemma 3 (Symmetrization inequalities)**.**

$$\mathbb{E}\sup_{t \in T}|Z_m(t)| \leq 2\mathbb{E}\sup_{t \in T}|R_m(t)|,$$

and for any $u > 0$, we have

$$\mathbb{P}\left(\sup_{t \in T}|Z_m(t)| \geq 2\mathbb{E}\sup_{t \in T}|Z_m(t)| + u\right) \leq 4\mathbb{P}\left(\sup_{t \in T}|R_m(t)| \geq u/2\right).$$

Proof. See Lemmas 6.3 and 6.5 in [37] □

Finally, we recall Bernstein's concentration inequality.

Lemma 4 (Bernstein's inequality)**.** *Let $X_1, \cdots, X_m$ be a sequence of independent centered random variables. Assume that there exist positive constants σ and D such that for all integers $p \geq 2$*

$$\frac{1}{m}\sum_{i=1}^m \mathbb{E}(|X_i|^p) \leq \frac{p!}{2}\sigma^2 D^{p-2},$$

then

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m X_i\right| \geq \frac{\sigma}{\sqrt{m}}\sqrt{2u} + \frac{D}{m}u\right) \leq 2\exp(-u).$$

In particular, if $X_1, \cdots, X_m$ are all sub-exponential random variables, then σ and D can be chosen as $\sigma = \frac{1}{m}\sum_{i=1}^m \|X_i\|_{\psi_1}$ and $D = \max\limits_{i=1\ldots m}\|X_i\|_{\psi_1}$.

B. Roadmap of the proof of Theorem 1.

We outline the main steps in the proof of Theorem 1, and postpone some technical details to sections VI-D and VI-E.
As it will be shown below in Lemma 5, $\operatorname{argmin}_{\theta \in \Theta} L^0(\theta) = \eta\theta_*$ for $\eta = \mathbb{E}(\langle y\mathbf{x}, \theta_*\rangle)$ and $L^0(\widehat{\theta}_m) - L^0(\eta\theta_*) = \|\widehat{\theta}_m - \eta\theta_*\|_2^2$, hence

$$\begin{aligned}
&\|\widehat{\theta}_m - \eta\theta_*\|_2^2 \\
=& L^\tau(\widehat{\theta}_m) - L^\tau(\eta\theta_*) \\
&+ \left(L^0(\widehat{\theta}_m) - L^\tau(\widehat{\theta}_m) - L^0(\eta\theta_*) + L^\tau(\eta\theta_*)\right) \\
=& L^\tau(\widehat{\theta}_m) - L^\tau(\eta\theta_*) + (L_m^\tau(\widehat{\theta}_m) - L_m^\tau(\eta\theta_*)) \\
&- (L_m^\tau(\widehat{\theta}_m) - L_m^\tau(\eta\theta_*)) - 2\mathbb{E}_m\left\langle y\mathbf{x} - \widetilde{q}\widetilde{U}, \widehat{\theta}_m - \eta\theta_*\right\rangle,
\end{aligned} \tag{24}$$

where $\mathbb{E}_m(\cdot)$ stands for the conditional expectation given $(\mathbf{x}_i, y_i)_{i=1}^m$, and where we used the equality $L^0(\widehat{\theta}_m) - L^\tau(\widehat{\theta}_m) - L^0(\eta\theta_*) + L^\tau(\eta\theta_*) =$

$-2\mathbb{E}_m\left(\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}_m-\eta\theta_*\right\rangle\right)$ in the last step. Since $\widehat{\theta}_m$ minimizes L^τ_m, $L^\tau_m(\widehat{\theta}_m)-L^\tau_m(\eta\theta_*)\le 0$, and

$$\|\widehat{\theta}_m-\eta\theta_*\|_2^2\le \frac{2}{m}\sum_{i=1}^m\left(\left\langle \widetilde{q}_i\widetilde{U}_i,\widehat{\theta}_m-\eta\theta_*\right\rangle-\mathbb{E}_m\left(\left\langle \widetilde{q}\widetilde{U},\widehat{\theta}_m-\eta\theta_*\right\rangle\right)\right) - 2\mathbb{E}_m\left(\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}_m-\eta\theta_*\right\rangle\right).$$

Note that $\widehat{\theta}_m-\eta\theta_*\in D(\Theta,\eta\theta_*)$; dividing both sides of the inequality by $\|\widehat{\theta}_m-\eta\theta_*\|_2$, we obtain

$$\|\widehat{\theta}_m-\eta\theta_*\|_2\le \sup_{\mathbf{v}\in D(\Theta,\eta\theta_*)\cap\mathbb{S}^{d-1}}\left|\frac{2}{m}\sum_{i=1}^m\left\langle \widetilde{q}_i\widetilde{U}_i,\mathbf{v}\right\rangle-\mathbb{E}\left\langle \widetilde{q}\widetilde{U},\mathbf{v}\right\rangle\right| + 2\sup_{\mathbf{v}\in\mathbb{S}^{d-1}}\mathbb{E}\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\mathbf{v}\right\rangle. \quad (25)$$

To get the desired bound, it remains to estimate two terms above. The bound for the first term is implied by Lemma 8: setting $T=D(\Theta,\eta\theta_*)\cap\mathbb{S}^{d-1}$, and observing that the diameter $\Delta_d(T):=\sup_{t\in T}\|t\|_2=1$, we get that with probability $\ge 1-ce^{-\beta/2}$,

$$\sup_{\mathbf{v}\in D(\Theta,\eta\theta_*)\cap\mathbb{S}^{d-1}}\left|\frac{2}{m}\sum_{i=1}^m\left\langle \widetilde{q}_i\widetilde{U}_i,\mathbf{v}\right\rangle-\mathbb{E}\left\langle \widetilde{q}\widetilde{U},\mathbf{v}\right\rangle\right| \le C\frac{(\omega(T)+1)\beta}{\sqrt{m}}.$$

To estimate the second term, we apply Lemma 7:

$$2\sup_{\mathbf{v}\in\mathbb{S}^{d-1}}\mathbb{E}\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\mathbf{v}\right\rangle \le \frac{\tilde{C}}{\sqrt{m}}.$$

Result of Theorem 1 now follows from the combination of these bounds. □

C. Roadmap of the proof of Theorem 2.

Once again, we will present the main steps while skipping the technical parts. Lemma 5 implies that $\operatorname{argmin}_{\theta\in\Theta} L^0(\theta)=\eta\theta_*$ for $\eta=\mathbb{E}\langle y\mathbf{x},\theta_*\rangle$ and

$$L^0(\widehat{\theta}^\lambda_m)-L^0(\eta\theta_*)=\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_2^2.$$

Thus, arguing as in (24),

$$\begin{aligned}\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_2^2 &= L^\tau(\widehat{\theta}^\lambda_m)-L^\tau(\eta\theta_*)+(L^\tau_m(\widehat{\theta}^\lambda_m)-L^\tau_m(\eta\theta_*)) \\ &\quad -(L^\tau_m(\widehat{\theta}^\lambda_m)-L^\tau_m(\eta\theta_*)) \\ &\quad - 2\mathbb{E}_m\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle.\end{aligned}$$

Since $\widehat{\theta}^\lambda_m$ is a solution of problem (12), it follows that

$$L^\tau_m(\theta^\lambda_m)+\lambda\left\|\theta^\lambda_m\right\|_{\mathcal{K}}\le L^\tau_m(\eta\theta_*)+\lambda\|\eta\theta_*\|_{\mathcal{K}},$$

which further implies that

$$\begin{aligned}&\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_2^2 \\ &\le\frac{2}{m}\sum_{i=1}^m\left(\left\langle \widetilde{q}_i\widetilde{U}_i,\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle-\mathbb{E}_m\left\langle \widetilde{q}\widetilde{U},\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle\right) \\ &\quad -2\mathbb{E}_m\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle+\lambda\left(\|\eta\theta_*\|_{\mathcal{K}}-\|\widehat{\theta}^\lambda_m\|_{\mathcal{K}}\right) \\ &=\left\langle\frac{2}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right),\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle \\ &\quad -2\mathbb{E}_m\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle+\lambda\left(\|\eta\theta_*\|_{\mathcal{K}}-\|\widehat{\theta}^\lambda_m\|_{\mathcal{K}}\right).\end{aligned} \quad (26)$$

Letting $\|\cdot\|^*_{\mathcal{K}}$ be the dual norm of $\|\cdot\|_{\mathcal{K}}$ (meaning that $\|\mathbf{x}\|^*_{\mathcal{K}}=\sup\{\langle\mathbf{x},\mathbf{z}\rangle,\ \|\mathbf{z}\|_{\mathcal{K}}\le 1\}$), the first term in (26) can be estimated as

$$\left\langle\frac{1}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right),\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle \le\left\|\frac{1}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right)\right\|^*_{\mathcal{K}}\cdot\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_{\mathcal{K}}. \quad (27)$$

Since

$$\left\|\frac{1}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right)\right\|^*_{\mathcal{K}} =\sup_{\|t\|_{\mathcal{K}}\le 1}\left\langle\frac{1}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right),t\right\rangle,$$

Lemma 8 applies with $T=\mathcal{G}:=\{\mathbf{x}\in\mathbb{R}^d:\ \|\mathbf{x}\|_{\mathcal{K}}\le 1\}$. Together with an observation that $\Delta_d(T)\le\sup_{t\in T}\|t\|_{\mathcal{K}}=1$ (due to the assumption $\|\mathbf{v}\|_2\le\|\mathbf{v}\|_{\mathcal{K}},\ \forall\mathbf{v}\in\mathbb{R}^d$), this yiels

$$\mathbb{P}\left(\sup_{\|t\|_{\mathcal{K}}\le 1}\left|\left\langle\frac{1}{m}\sum_{i=1}^m\widetilde{q}_i\widetilde{U}_i-\mathbb{E}\left(\widetilde{q}\widetilde{U}\right),t\right\rangle\right| \ge C'\frac{(\omega(\mathcal{G})+1)\,\beta}{\sqrt{m}}\right)\le c'e^{-\beta/2},$$

for any $\beta\ge 8$ and some constants $C',c>0$. For the second term in (26), we use Lemma 7 to obtain

$$2\mathbb{E}_m\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\widehat{\theta}^\lambda_m-\eta\theta_*\right\rangle\le\frac{C''}{\sqrt{m}}\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_2 \le\frac{C''}{\sqrt{m}}\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_{\mathcal{K}},$$

for some constant $C''>0$, where we have again applied the inequality $\|\mathbf{v}\|_2\le\|\mathbf{v}\|_{\mathcal{K}}$. Combining the above two estimates gives that with probability at least $1-ce^{-\beta/2}$,

$$\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_2^2\le C\frac{(\omega(\mathcal{G})+1)\,\beta}{\sqrt{m}}\|\widehat{\theta}^\lambda_m-\eta\theta_*\|_{\mathcal{K}} +\lambda\left(\|\eta\theta_*\|_{\mathcal{K}}-\|\widehat{\theta}^\lambda_m\|_{\mathcal{K}}\right), \quad (28)$$

for some constant $C > 0$ and any $\beta \geq 8$. Since $\lambda \geq 2C(\omega(\mathcal{G})+1)\beta/\sqrt{m}$ by assumption, and the right hand side of (28) is nonnegative, it follows that

$$\frac{1}{2}\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_{\mathcal{K}} + \|\eta\theta_*\|_{\mathcal{K}} - \|\widehat{\theta}_m^\lambda\|_{\mathcal{K}} \geq 0.$$

This inequality implies that $\widehat{\theta}_m^\lambda - \eta\theta_* \in \mathbb{S}_2(\eta\theta_*)$. Finally, from (28) and the triangle inequality,

$$\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_2^2 \leq \frac{3}{2}\lambda\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_{\mathcal{K}}.$$

Dividing both sides by $\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_2$ gives

$$\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_2 \leq \frac{3}{2}\lambda\frac{\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_{\mathcal{K}}}{\|\widehat{\theta}_m^\lambda - \eta\theta_*\|_2} \leq \frac{3}{2}\lambda\cdot\Psi\left(\mathbb{S}_2(\eta\theta_*)\right).$$

This finishes the proof of Theorem 2.

D. Bias of the truncated mean.

The following lemma is motivated by and is similar to Theorem 2.1 in [13].

Lemma 5. *Let $\eta = \mathbb{E}\langle y\mathbf{x}, \theta_*\rangle$. Then*

$$\eta\theta_* = \underset{\theta\in\Theta}{\textit{argmin}}\, L^0(\theta),$$

and for any $\theta \in \Theta$,

$$L^0(\theta) - L^0(\eta\theta_*) = \|\theta - \eta\theta_*\|_2^2.$$

Proof. Since $y = f(\langle \mathbf{x}, \theta_*\rangle, \delta)$, we have that for any $\theta \in \mathbb{R}^d$

$$\begin{aligned}\mathbb{E}\langle y\mathbf{x}, \theta\rangle =& \mathbb{E}\langle \mathbf{x}, \theta\rangle f(\langle \mathbf{x}, \theta_*\rangle, \delta)\\ =& \mathbb{E}\mathbb{E}(\langle \mathbf{x}, \theta\rangle f(\langle \mathbf{x}, \theta_*\rangle, \delta) \mid \langle \mathbf{x}, \theta_*\rangle, \delta)\\ =& \mathbb{E}\mathbb{E}\left(\langle \mathbf{x}, \theta\rangle \mid \langle \mathbf{x}, \theta_*\rangle\right)\cdot f(\langle \mathbf{x}, \theta_*\rangle, \delta)\\ =& \mathbb{E}\Big(\langle \theta_*, \theta\rangle\langle \mathbf{x}, \theta_*\rangle f(\langle \mathbf{x}, \theta_*\rangle, \delta)\Big)\\ =& \eta\langle\theta_*, \theta\rangle,\end{aligned}$$

where the third equality follows from the fact that the noise δ is independent of the measurement vector $\mathbf{x}$, the second to last equality from the properties of elliptically symmetric distributions (Corollary 1), and the last equality from the definition of η. Thus,

$$\begin{aligned}L^0(\theta) = \|\theta\|_2^2 - 2\mathbb{E}(\langle y\mathbf{x}, \theta\rangle) &= \|\theta\|_2^2 - 2\eta\langle\theta_*, \theta\rangle\\ &= \|\theta - \eta\theta_*\|_2^2 - \|\eta\theta_*\|_2^2,\end{aligned}$$

which is minimized at $\theta = \eta\theta^*$. Furthermore, $L^0(\eta\theta^*) = -\|\eta\theta_*\|_2^2$, hence

$$L^0(\theta) - L^0(\eta\theta_*) = \|\theta - \eta\theta_*\|_2^2,$$

finishing the proof. □

Next, we estimate the "bias term" $\sup_{\mathbf{v}\in\mathbb{S}^{d-1}}\mathbb{E}\left\langle y\mathbf{x} - \widetilde{q}\widetilde{U}, \mathbf{v}\right\rangle$ in inequality (25). In order to do so, we need the following preliminary result.

Lemma 6. *If $\mathbf{x} \sim \mathcal{E}(0, \mathbf{I}_{d\times d}, F_\mu)$, then the unit random vector $\mathbf{x}/\|\mathbf{x}\|_2$ is uniformly distributed over the unit sphere $\mathbb{S}^{d-1}$. Furthermore, $\widetilde{U} = \sqrt{d}\mathbf{x}/\|\mathbf{x}\|_2$ is a sub-Gaussian random vector with sub-Gaussian norm $\|\widetilde{U}\|_{\psi_2}$ independent of the dimension d.*

Proof. First, we use decomposition (4) for elliptical distribution together with our assumption that $\mathbf{\Sigma}$ is the identity matrix, to write $\mathbf{x} \overset{d}{=} \mu U$, which implies that

$$\mathbf{x}/\|\mathbf{x}\|_2 \overset{d}{=} \text{sign}(\mu)U/\|U\|_2 = \text{sign}(\mu)U \overset{d}{=} U,$$

with the final distributional equality holding as $\mathbb{S}^{d-1}$, and hence its uniform distribution, is invariant with respect to reflections across any hyperplane through the origin.

To prove the second claim, it is enough to show that $\left\|\left\langle \widetilde{U}, \mathbf{v}\right\rangle\right\|_{\psi_2} \leq C,\ \forall \mathbf{v} \in \mathbb{S}^{d-1}$ with constant C independent of d. By the first claim and Lemma 1, we have

$$\mathbb{P}\left(\langle \mathbf{x}, \mathbf{v}\rangle/\|\mathbf{x}\|_2 \geq \Delta\right) \leq e^{-d\Delta^2/2},\ \forall \mathbf{v} \in \mathbb{S}^{d-1}.$$

Choosing $\Delta = u/\sqrt{d}$ gives

$$\mathbb{P}\left(\left\langle \widetilde{U}, \mathbf{v}\right\rangle \geq u\right) \leq e^{-u^2/2},\ \forall \mathbf{v} \in \mathbb{S}^{d-1},\ \forall u > 0.$$

By an equivalent definition of sub-Gaussian random variables (Lemma 5.5 in [36]), this inequality implies that $\left\|\left\langle \widetilde{U}, \mathbf{v}\right\rangle\right\|_{\psi_2} \leq C$, hence finishing the proof. □

With the previous lemma in hand, we now establish the following result.

Lemma 7. *Under the assumptions of Theorem 1, there exists a constant $C = C(\kappa, \phi) > 0$ such that*

$$\left|\mathbb{E}\left\langle y\mathbf{x} - \widetilde{q}\widetilde{U}, \mathbf{v}\right\rangle\right| \leq C/\sqrt{m},$$

for all $\mathbf{v} \in \mathbb{S}^{d-1}$.

Proof. By (7), we have that $y\mathbf{x} = q\widetilde{U}$, thus the claim is equivalent to

$$\left|\mathbb{E}\left(\left\langle \widetilde{U}, \mathbf{v}\right\rangle(\widetilde{q} - q)\right)\right| \leq C/\sqrt{m}.$$

Since $\widetilde{q} = \text{sign}(q)(|q| \wedge \tau)$, we have $|\widetilde{q} - q| = (|q| - \tau)\mathbf{1}(|q| \geq \tau) \leq |q|\mathbf{1}(|q| \geq \tau)$, and it follows that

$$\begin{aligned}&\left|\mathbb{E}\left\langle \widetilde{U}, \mathbf{v}\right\rangle(\widetilde{q} - q)\right|\\ \leq&\mathbb{E}\left|\left\langle \widetilde{U}, \mathbf{v}\right\rangle(\widetilde{q} - q)\right|\\ \leq&\mathbb{E}\left(\left|\left\langle \widetilde{U}, \mathbf{v}\right\rangle q\right|\cdot\mathbf{1}_{\{|q|\geq\tau\}}\right)\\ \leq&\mathbb{E}\left(\left|\left\langle \widetilde{U}, \mathbf{v}\right\rangle q\right|^2\right)^{1/2}\mathbb{P}\left(|q| \geq \tau\right)^{1/2}\\ \leq&\mathbb{E}\left(\left|\left\langle \widetilde{U}, \mathbf{v}\right\rangle\right|^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}}\mathbb{E}\left(|q|^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\mathbb{P}\left(|q| \geq \tau\right)^{1/2},\end{aligned}$$

where the second to last inequality uses Cauchy-Schwarz, and the last inequality follows from Hölder's inequality.

For the first term, by Lemma 6, $\widetilde{U}$ is sub-Gaussian with $\|\widetilde{U}\|_{\psi_2}$ independent of d. Thus, by the definition of the $\|\cdot\|_{\psi_2}$ norm and the fact that $\mathbf{v}\in\mathbb{S}^{d-1}$,

$$\mathbb{E}\left(\left|\left\langle\widetilde{U},\mathbf{v}\right\rangle\right|^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}} \leq \sqrt{\frac{2(1+\kappa)}{\kappa}}\|\widetilde{U}\|_{\psi_2}.$$

Recall that $\phi=\mathbb{E}|q|^{2(1+\kappa)}$. Then, the second term is bounded by $\phi^{\frac{1}{2(1+\kappa)}}$. For the final term, since $\tau=m^{\frac{1}{2(1+\kappa)}}$, Markov's inequality implies that

$$\left(\mathbb{P}\left(|q|>\tau\right)\right)^{1/2} \leq \left(\frac{\mathbb{E}|q|^{2(1+\kappa)}}{\tau^{2(1+\kappa)}}\right)^{1/2} \leq \frac{\phi^{1/2}}{\sqrt{m}}.$$

Combining these inequalities yields

$$\begin{aligned}&\left|\mathbb{E}\left\langle y\mathbf{x}-\widetilde{q}\widetilde{U},\mathbf{v}\right\rangle\right| \\ &\qquad\leq \frac{\sqrt{\frac{2(1+\kappa)}{\kappa}}\|\widetilde{U}\|_{\psi_2}\phi^{\frac{2+\kappa}{2(1+\kappa)}}}{\sqrt{m}} := C(\kappa,\phi)/\sqrt{m},\end{aligned}$$

completing the proof. □

E. Concentration via generic chaining.

In the following sections, we will use c, C, C', C'' to denote constants that are either absolute, or depend on underlying parameters κ and ϕ (in the latter case, we specify such dependence). To make notation less cumbersome, constants denoted by the same letter $(c, C, C',$ etc.) might be different in various parts of the proof.

The goal of this subsection is to prove the following inequality:

Lemma 8. *Suppose $\widetilde{U}_i$ and $\widetilde{q}_i$ are as defined according to (7) and (8) respectively. Then, for any bounded subset $T\subset\mathbb{R}^d$,*

$$\begin{aligned}\mathbb{P}\Bigg(\sup_{t\in T}\left|\frac{1}{m}\sum_{i=1}^m\left\langle\widetilde{U}_i,t\right\rangle\widetilde{q}_i-\mathbb{E}\left(\left\langle\widetilde{U},t\right\rangle\widetilde{q}\right)\right| \\ \geq C\frac{(\omega(T)+\Delta_d(T))\beta}{\sqrt{m}}\Bigg) \leq ce^{-\beta/2},\end{aligned}$$

for any $\beta\geq 8$, a positive constant $C=C(\kappa,\phi)$ and an absolute constant $c>0$. Here

$$\Delta_d(T) := \sup_{t\in T}\|t\|_2. \tag{29}$$

The main technique we apply is the generic chaining method developed by M. Talagrand [24] for bounding the supremum of stochastic processes. Later, works [38] and [39] advanced the technique to obtain a sharp bound for supremum of processes index by squares of functions. More recently, S. Mendelson [25] proved a concentration result for the supremum of multiplier processes under weak moment assumptions. In the current work, we show that exponential-type concentration inequalities for multiplier processes, such as the one in Lemma 8, are achievable by applying truncation under a bounded $2(1+\kappa)$-moment assumption.

Define

$$\begin{aligned}\overline{Z}(t) &= \frac{1}{m}\sum_{i=1}^m\left\langle\widetilde{U}_i,t\right\rangle\widetilde{q}_i-\mathbb{E}\left(\left\langle\widetilde{U},t\right\rangle\widetilde{q}\right), \\ Z(t) &= \frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{q}_i\left\langle\widetilde{U}_i,t\right\rangle,\ \forall t\in T,\end{aligned}$$

where T is a bounded set in $\mathbb{R}^d$ and $\{\varepsilon_i\}_{i=1}^m$ is a sequence i.i.d. Rademacher random variables taking values ±1 with probability $1/2$ each, and independent of $\{\widetilde{U}_i,\widetilde{q}_i,\ i=1,\dots,m\}$. Result of Lemma 8 easily follows from the following concentration inequality:

Lemma 9. *For any $\beta\geq 8$,*

$$\mathbb{P}\left[\sup_{t\in T}|Z(t)|\geq C\frac{(\omega(T)+\Delta_d(T))\beta}{\sqrt{m}}\right]\leq ce^{-\beta/2}, \tag{30}$$

where $C=C(\kappa,\phi)$ is another constant possibly different from that of Lemma 8, and $c>0$ is an absolute constant.

To deduce the inequality of Lemma 8, we first apply the symmetrization inequality (Lemma 3), followed by Lemma 13 with $\beta_0=8$. It implies that

$$\begin{aligned}\mathbb{E}\left(\sup_{t\in T}\left|\overline{Z}(t)\right|\right) &\leq 2\mathbb{E}\left(\sup_{t\in T}|Z(t)|\right) \\ &\leq 2C\left(8+2ce^{-4}\right)\frac{\omega(T)+\Delta_d(T)}{\sqrt{m}}.\end{aligned}$$

Application of the second bound of the symmetrization lemma with $u=2C(\omega(T)+\Delta_d(T))\beta/\sqrt{m}$ and (30) completes the proof of Lemma 8.

It remains to justify (30). We start by picking an arbitrary point $t_0\in T$ such that there exists an admissible sequence $\{t_0\}=\mathcal{A}_0\subseteq\mathcal{A}_1\subseteq\mathcal{A}_2\subseteq\cdots$ satisfying

$$\sup_{t\in T}\sum_{l=0}^{\infty}2^{l/2}\|\pi_l(t)-t\|_2\leq 2\gamma_2(T), \tag{31}$$

where we recall that π_l is the closest point map from T to $\mathcal{A}_l$ and the factor 2 is introduced so as to deal with the case where the infimum in the definition (22) of $\gamma_2(T)$ is not achieved. Then, write $Z(t)-Z(t_0)$ as the telescoping sum:

$$\begin{aligned}Z(t)-Z(t_0) &= \sum_{l=1}^{\infty}Z(\pi_l(t))-Z(\pi_{l-1}(t)) \\ &= \sum_{l=1}^{\infty}\frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{q}_i\left\langle\widetilde{U}_i,\pi_l(t)-\pi_{l-1}(t)\right\rangle.\end{aligned}$$

We claim that the telescoping sum converges with probability 1 for any $t \in T$. Indeed, note that for each fixed set of realizations of $\{\mathbf{x}_i\}_{i=1}^m$ and $\{\varepsilon_i\}_{i=1}^m$, each summand is bounded as

$$\begin{aligned}&|\varepsilon_i\widetilde{q}_i\langle\widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\rangle|\\ \le&|\widetilde{q}_i|\|\widetilde{U}_i\|_2\|\pi_l(t) - \pi_{l-1}(t)\|_2\\ \le&|\widetilde{q}_i|\|\widetilde{U}_i\|_2(\|\pi_l(t) - t\|_2 + \|\pi_{l-1}(t) - t\|_2).\end{aligned}$$

Furthermore, since T is a compact subset of $\mathbb{R}^d$, its Gaussian mean width is finite. Thus, by lemma 2, $\gamma_2(T) \le L\omega(T) < \infty$. This inequality further implies that the sum on the left hand side of (31) converges with probability 1.

Next, with $\beta \ge 8$ being fixed, we split the index set $\{l \ge 1\}$ into the following three subsets:

$$\begin{aligned}I_1 &= \{l \ge 1 : 2^l\beta < \log em\};\\ I_2 &= \{l \ge 1 : \log em \le 2^l\beta < m\};\\ I_3 &= \{l \ge 1 : 2^l\beta \ge m\}.\end{aligned}$$

By the assumptions in Theorem 1 and the bound $\beta \ge 8$, we have that $m \ge (\omega(T) + 1)^2\beta^2 \ge 64$, implying that $\log em = 1 + \log m < m$, and hence these three index sets are well defined. Depending on β, some of them might be empty, but this only simplifies our argument by making the partial sum over such an index set equal 0.

The following argument yields a bound for $Z(\pi_l(t)) - Z(\pi_{l-1}(t))$, assuming all three index sets are nonempty. Specifically, we show that

$$\mathbb{P}\left(\sup_{t\in T}\left|\sum_{l\in I_j}(Z(\pi_l(t)) - Z(\pi_{l-1}(t)))\right| \ge C\frac{\gamma_2(T)\beta}{\sqrt{m}}\right) \le ce^{-\beta/2}, \quad (32)$$

for $C = C(\kappa, \phi)$ and $j = 1, 2, 3$, respectively.

1) The case $l \in I_1$:

Proof of inequality (32) *for the index set I_1.* Recall that $\tau = m^{\frac{1}{2(1+\kappa)}}$.

For each $t \in T$ we apply Bernstein's inequality (Lemma 4) to estimate each summand

$$Z(\pi_l(t)) - Z(\pi_{l-1}(t)) = \frac{1}{m}\sum_{i=1}^m \varepsilon_i\widetilde{q}_i\left\langle\widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle.$$

For any integer $p \ge 2$, we have the following chains of inequalities:

$$\begin{aligned}&\mathbb{E}\left(\left|\varepsilon\widetilde{q}\left\langle\widetilde{U}, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right|^p\right)\\ \le&\mathbb{E}\left(\left|\varepsilon\left\langle\widetilde{U}, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right|^p q^2 \cdot |\widetilde{q}|^{p-2}\right)\\ \le&\mathbb{E}\left(\left|\left\langle\widetilde{U}, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right|^p q^2\right) \cdot \tau^{p-2}\\ \le&\tau^{p-2}\mathbb{E}\left(\left|\left\langle\widetilde{U}, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right|^{\frac{1+\kappa}{\kappa}p}\right)^{\frac{\kappa}{1+\kappa}}\mathbb{E}\left(q^{2(1+\kappa)}\right)^{\frac{1}{1+\kappa}}\\ \le&\tau^{p-2}\|\widetilde{U}\|_{\psi_2}^p\left(\frac{(1+\kappa)p}{\kappa}\right)^{p/2}\phi^{\frac{1}{1+\kappa}}\|\pi_l(t) - \pi_{l-1}(t)\|_2^p,\end{aligned}$$

where the second inequality follows from the truncation bound, the third from Hölder's inequality, and the last from the assumption that $\mathbb{E}\left(q^{2(1+\kappa)}\right) \le \phi$ and the following bound: by Lemma 6, $\widetilde{U}_i$ is sub-Gaussian, hence for any $p \ge 2$

$$\left(\mathbb{E}\left\langle\widetilde{U}_i, \mathbf{v}\right\rangle^{\frac{1+\kappa}{\kappa}p}\right)^{\frac{\kappa}{(1+\kappa)p}} \le \left(\frac{(1+\kappa)p}{\kappa}\right)^{1/2}\|\widetilde{U}_i\|_{\psi_2}\|\mathbf{v}\|_2, \forall\mathbf{v} \in \mathbb{R}^d.$$

We also note that $\|\widetilde{U}_i\|_{\psi_2}$ does not depend on d by Lemma 6. Next, by Stirling's approximation, $p! \ge \sqrt{2\pi}\sqrt{p}(p/e)^p$, thus there exist constants $C' = C'(\kappa, \phi)$ and $C'' = C''(\kappa)$ such that

$$\begin{aligned}&\mathbb{E}\left|\varepsilon\widetilde{q}\left\langle\widetilde{U}, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right|^p\\ \le& \frac{p!}{2}C'\|\pi_l(t) - \pi_{l-1}(t)\|_2^2(C''\tau\|\pi_l(t) - \pi_{l-1}(t)\|_2)^{p-2}.\end{aligned}$$

Bernstein's inequality (Lemma 4), with $\sigma = C'\|\pi_l(t) - \pi_{l-1}(t)\|_2$, $D = C''\tau\|\pi_l(t) - \pi_{l-1}(t)\|_2$ and $\tau = m^{1/2(1+\kappa)}$ now implies

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m\varepsilon_i\widetilde{q}_i\left\langle\widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right| \ge \left(\frac{C'\sqrt{2u}}{\sqrt{m}} + \frac{C''u}{m^{1-\frac{1}{2(1+\kappa)}}}\right)\|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \le 2e^{-u},$$

for any $u > 0$. Taking $u = 2^l\beta$, noting that as $\beta \ge 8$ by assumption, we have $m \ge (\omega(T) + 1)^2\beta^2 \ge 64$, and since $l \in I_1$, $2^l \le 2^l\beta < \log em$. In turn, this implies

$$\begin{aligned}\frac{2^l}{m^{1-\frac{1}{2(1+\kappa)}}} &= \frac{2^{l/2}}{m^{1/2}} \cdot \frac{2^{l/2}}{m^{\kappa/2(1+\kappa)}}\\ &\le \frac{2^{l/2}}{m^{1/2}} \cdot \sqrt{\frac{\log em}{m^{\kappa/(1+\kappa)}}} \le \sqrt{\frac{1+\kappa}{\kappa}}\frac{2^{l/2}}{m^{1/2}},\end{aligned}$$

where the last inequality follows from the fact that $\log em$ is dominated by $\frac{1+\kappa}{\kappa}m^{\kappa/(1+\kappa)}$ for all $m \ge$

1. This inequality implies that there exists a positive constant $C = C(\kappa, \phi)$ such that for any $\beta \geq 8$

$$\mathbb{P}\left(\Omega_{l,t}\right) \leq 2\exp(-2^l \beta), \tag{33}$$

where for all $l \geq 1$ and $t \in T$ we let

$$\Omega_{l,t} = \left\{\omega : \left|\frac{1}{m}\sum_{i=1}^m \varepsilon_i \widetilde{q}_i \left\langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle\right| \geq C\frac{2^{l/2}\beta}{\sqrt{m}}\|\pi_l(t) - \pi_{l-1}(t)\|_2\right\}.$$

Notice that for each $l \geq 1$ the number of pairs $(\pi_l(t), \pi_{l-1}(t))$ appearing in the sum in (32) can be bounded by $|\mathcal{A}_l| \cdot |\mathcal{A}_{l-1}| \leq 2^{2^{l+1}}$. Thus, by a union bound and (33),

$$\mathbb{P}\left(\bigcup_{t\in T}\Omega_{l,t}\right) \leq \ 2\cdot 2^{2^{l+1}}\exp(-2^l\beta),$$

and hence,

$$\begin{aligned}\mathbb{P}\left(\bigcup_{l\in I_1, t\in T}\Omega_{l,t}\right) &\leq \sum_{l\in I_1} 2\cdot 2^{2^{l+1}}\exp(-2^l\beta)\\ &\leq \sum_{l\in I_1} 2\cdot 2^{2^{l+1}}\exp\left(-2^{l-1}\beta - \beta/2\right)\\ &\leq c e^{-\beta/2},\end{aligned}$$

for some absolute constant $c > 0$, where in the last inequality we use the fact $\beta \geq 8$ to get a geometrically decreasing sequence. Thus, on the complement of the event $\cup_{l\in I_1, t\in T}\Omega_{l,t}$, we have that with probability at least $1 - ce^{-\beta/2}$,

$$\begin{aligned}&\sup_{t\in T}\left|\sum_{l\in I_1}\left(Z(\pi_l(t)) - Z(\pi_{l-1}(t))\right)\right|\\ \leq&\sup_{t\in T}\sum_{l\in I_1}\left|Z(\pi_l(t)) - Z(\pi_{l-1}(t))\right|\\ \leq&\sup_{t\in T} C\sum_{l\in I_1}\frac{2^{l/2}\beta}{\sqrt{m}}\|\pi_l(t) - \pi_{l-1}(t)\|_2\\ \leq&\sup_{t\in T} C\sum_{l=1}^{\infty}\frac{2^{l/2}\beta}{\sqrt{m}}\|\pi_l(t) - \pi_{l-1}(t)\|_2\\ \leq&4C\frac{\gamma_2(T)\beta}{\sqrt{m}},\end{aligned}$$

for $C = C(\kappa, \phi)$, where the last inequality follows from triangle inequality $\|\pi_l(t) - \pi_{l-1}(t)\|_2 \leq \|\pi_{l-1}(t) - t\|_2 + \|\pi_l(t) - t\|_2$ and (31). This proves the inequality (32) for $l \in I_1$. □

2) The case $l \in I_2$: This is the most technically involved case of the three. For any fixed $t \in T$ and $l \in I_2$, we let $X_i = \widetilde{q}_i\left\langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle$ and $w_i = \langle\widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\rangle$. Then $X_i = \widetilde{q}_i w_i$ and

$$Z(\pi_l(t)) - Z(\pi_{l-1}(t)) = \frac{1}{m}\sum_{i=1}^m \varepsilon_i X_i = \frac{1}{m}\sum_{i=1}^m \varepsilon_i w_i \widetilde{q}_i. \tag{34}$$

For every fixed $k \in \{1, 2, \cdots, m-1\}$ and fixed $u > 0$, we bound the summation using the following inequality

$$\mathbb{P}\left(\left|\sum_{i=1}^m \varepsilon_i X_i\right| \geq \sum_{i=1}^k X_i^* + u\left(\sum_{i=k+1}^m (X_i^*)^2\right)^{1/2}\right) \leq 2\exp(-u^2/2),$$

where $\{X_i^*\}_{i=1}^m$ is the *non-increasing* rearrangement of $\{|X_i|\}_{i=1}^m$ and $\{\varepsilon_i\}_{i=1}^m$ is a sequence of i.i.d. Rademancher random variables independent of $\{X_i\}_{i=1}^m$.

Remark 8. *This bound was first stated and proved in [40] with a sequence of fixed constants $\{X_i\}_{i=1}^m$. The current form can be obtained using independence property and conditioning on $\{X_i\}_{i=1}^m$. Furthermore, paper [40] tells us that the optimal choice of k is at $\mathcal{O}(u^2)$ Applications of this inequality to generic chaining-type arguments were previously introduced in [25].*

Letting J be the set of indices of the variables corresponding to the k largest coordinates of $\{|w_i|\}_{i=1}^m$ and of $\{|\widetilde{q}_i|\}_{i=1}^m$, we have $|J| \leq 2k$ and with probability at least $1 - 2\exp(-u^2/2)$

$$\begin{aligned}&\left|\sum_{i=1}^m \varepsilon_i X_i\right|\\ \leq&\sum_{i\in J} X_i^* + u\left(\sum_{i\in J^c}(X_i^*)^2\right)^{1/2}\\ \leq&2\sum_{i=1}^k w_i^*\widetilde{q}_i^* + u\left(\sum_{i\in J^c}(w_i^*\widetilde{q}_i^*)^2\right)^{1/2}\\ \leq&2\left(\sum_{i=1}^k (w_i^*)^2\right)^{1/2}\left(\sum_{i=1}^k(\widetilde{q}_i^*)^2\right)^{1/2}\\ &+ u\left(\sum_{i=k+1}^m (w_i^*)^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=k+1}^m(\widetilde{q}_i^*)^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\\ \leq&2\left(\sum_{i=1}^k (w_i^*)^2\right)^{1/2}\left(\sum_{i=1}^m\widetilde{q}_i^2\right)^{1/2}\\ &+ u\left(\sum_{i=k+1}^m (w_i^*)^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^m\widetilde{q}_i^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\end{aligned} \tag{35}$$

where the second to last inequality is a consequence of Hölder's inequality. We take $u = 2^{(l+1)/2}\sqrt{\beta}$. The key is to pick an appropriate cut point k for each $l \in I_2$. Here, we choose $k = \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor$, which makes $k = \mathcal{O}(2^l\beta)$ and also guarantees that $k \in \{1, 2, \cdots, m-1\}$; see Lemma 16. Under this choice, we have the following lemma:

Lemma 10. *Let $k = \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor$, $w_i = \left\langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle$ and $\{w_i^*\}_{i=1}^m$ be the nonincreasing rearrangement of $\{|w_i|\}_{i=1}^m$. Then there exists an absolute constant $C > 1$ such that for all $\beta \geq 8$,*

$$\mathbb{P}\left(\left(\sum_{i=1}^{k}(w_i^*)^2\right)^{1/2} \geq C2^{l/2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\sqrt{\beta}\right) \leq 2\exp(-2^l\beta).$$

Proof. By Lemma 6, we know that $\{w_i\}_{i=1}^m$ are i.i.d. sub-Gaussian random variables. Thus, by Lemma 14, w_i^2 is sub-exponential with norm

$$\|w_i^2\|_{\psi_1} = 2\|w_i\|_{\psi_2}^2 \leq 2\|\widetilde{U}_i\|_{\psi_2}^2\|\pi_l(t) - \pi_{l-1}(t)\|_2^2. \tag{36}$$

It then follows from Bernstein's inequality (Lemma 4) that for any fixed set $J \subseteq \{1, 2, \cdots, m\}$ with $|J| = k$,

$$\mathbb{P}\left(\left|\frac{1}{k}\sum_{i\in J}\left(w_i^2 - \mathbb{E}(w_i^2)\right)\right| \geq 2\|\widetilde{U}_i\|_{\psi_2}^2\|\pi_l(t) - \pi_{l-1}(t)\|_2^2\left(\sqrt{\frac{2u}{k}} + \frac{u}{k}\right)\right) \leq 2\exp(-u).$$

We choose $u = 4 \cdot 2^l\beta = 2^{l+2}\beta$. Since $2^l\beta \geq \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor = k \geq 1$, the factor u/k dominates the right hand side. Noting that $\mathbb{E}(w_i^2) = \|\pi_l(t) - \pi_{l-1}(t)\|_2^2$, we obtain

$$\mathbb{P}\left(\left(\sum_{i\in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\sqrt{\beta}\right) \leq 2\exp(-4\cdot 2^l\beta),$$

where $C \leq 4\|\widetilde{U}_i\|_{\psi_2}$; note that the upper bound for C is independent of d by Lemma 1. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\left(\sum_{i=1}^{k}(w_i^*)^2\right)^{1/2} \geq C2^{l/2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\sqrt{\beta}\right)\\
=&\mathbb{P}\Bigg(\exists J \subseteq \{1, \cdots, m\},\ |J| = k :\\
&\left(\sum_{i\in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\sqrt{\beta}\Bigg)\\
\leq&\binom{m}{k}\mathbb{P}\left(\left(\sum_{i\in J} w_i^2\right)^{1/2} \geq C2^{l/2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\sqrt{\beta}\right)\\
\leq&2\binom{m}{k}\exp(-4\cdot 2^l\beta)\\
\leq&2\left(\frac{em}{k}\right)^k\exp(-4\cdot 2^l\beta) \leq 2\exp(-2^l\beta),
\end{aligned}$$

where the last step follows from $\left(\frac{em}{k}\right)^k \leq \exp(3\cdot 2^l\beta)$, an inequality proved in Lemma 15 in the Appendix. ☐

Lemma 11. *Let $k = \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor$, $w_i = \left\langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle$ and $\{w_i^*\}_{i=1}^m$ be the nonincreasing rearrangement of $\{|w_i|\}_{i=1}^m$. Then*

$$\mathbb{P}\left(\left(\sum_{i=k+1}^{m}(w_i^*)^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}} \geq C(\kappa)m^{\frac{\kappa}{2(1+\kappa)}}\|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \leq \exp(-2^l\beta),$$

for any $\beta \geq 8$ and some constant $C(\kappa) > 0$.

Proof. To avoid possible confusion, we use i to index the nonincreasing rearrangement and j for the original sequence. We start by noting that $\{w_j\}_{j=1}^m$ are i.i.d. sub-Gaussian random variables with $\|w_j\|_{\psi_2} \leq \|\widetilde{U}_j\|_{\psi_2}\|\pi_l(t) - \pi_{l-1}(t)\|_2$. By an equivalent definition of sub-Gaussian random variables (Lemma 5.5. in [36]), we have for any fixed $j \in \{1, 2, \ldots, m\}$,

$$\mathbb{P}\left(|w_j| - \mathbb{E}(|w_j|) \geq Cu\|\widetilde{U}_j\|_{\psi_2}\|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \leq e^{-u^2}, \tag{37}$$

for any $u > 0$ and an absolute constant $C > 0$.

To establish the claim of the lemma, we bound each w_i^* separately for $i = 1, 2\ldots, m$ and then combine individual bounds. Instead of using a fixed value of u in (37), our choice of u will depend on the index i. Specifically, for each w_i^*, we choose $u = c_\kappa(m/i)^{\kappa/4(1+\kappa)}$ with

$$c_\kappa := \max\left\{\frac{\sqrt{5}\left(2 + \frac{4}{\kappa}\right)^{\frac{2+\kappa}{4(1+\kappa)}}}{e^{1/2(1+\kappa)}}, \sqrt{\frac{4(1+\kappa)}{\kappa}}\right\}. \tag{38}$$

The reason for this choice will be clear as we proceed.

First, for a fixed nonincreasing rearrangement index $i > k$, by (37) and the fact that

$$\mathbb{E}(|w_j|) \le \mathbb{E}\left(w_j^2\right)^{1/2} = \|\pi_l(t) - \pi_{l-1}(t)\|_2, \quad \forall j \in \{1, 2, \cdots, m\},$$

we have $\forall j \in \{1, 2, \cdots, m\}$,

$$\mathbb{P}\left(|w_j| \ge \left(1 + Cc_\kappa \|\widetilde{U}_j\|_{\psi_2}\right) \cdot \left(\frac{m}{j}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \le \exp\left(-c_\kappa^2 \left(\frac{m}{j}\right)^{\frac{\kappa}{2(1+\kappa)}}\right).$$

To simplify notation, let $C' = 1 + Cc_\kappa \|\widetilde{U}_j\|_{\psi_2}$ (note that it depends only on κ). It then follows that

$$\begin{aligned}
&\mathbb{P}\left(w_i^* \ge C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \\
=&\mathbb{P}\Big(\exists J \subseteq \{1, \cdots, m\},\ |J| = i : \\
&\quad w_j \ge C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2,\ \forall j \in J\Big) \\
\le&\binom{m}{i} \cdot \mathbb{P}\left(|w_j| \ge C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2\right)^i \\
\le&\binom{m}{i} \exp\left(-c^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right) \\
\le&\left(\frac{em}{i}\right)^i \exp\left(-c^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right).
\end{aligned}$$

Union bound gives

$$\begin{aligned}
&\mathbb{P}\left(\exists i > k : w_i^* \ge C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \\
\le&\sum_{i=k+1}^{m} \left(\frac{em}{i}\right)^i \exp\left(-c^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right) \\
=&\sum_{i=k+1}^{m} \exp\left(i \log\left(\frac{em}{i}\right) - c^2 m^{\frac{\kappa}{2(1+\kappa)}} i^{\frac{2+\kappa}{2(1+\kappa)}}\right) \\
\le&m \cdot \exp\left(k \log\left(\frac{em}{k}\right) - c^2 m^{\frac{\kappa}{2(1+\kappa)}} k^{\frac{2+\kappa}{2(1+\kappa)}}\right) \\
\le&\exp\left(4 \cdot 2^l \beta - c^2 m^{\frac{\kappa}{2(1+\kappa)}} k^{\frac{2+\kappa}{2(1+\kappa)}}\right),
\end{aligned}$$

where the second to last inequality follows since by the definition (38) of c_κ, $c_\kappa \ge \sqrt{4(1+\kappa)/\kappa}$, the function $v(i) = i \log\left(\frac{em}{i}\right) - c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} \cdot i^{\frac{2+\kappa}{2(1+\kappa)}}$ is monotonically decreasing with respect to i (recall that $i \le m$), and thus is dominated by $v(k)$. The final inequality follows from Lemma 15 as well as the fact that $\log m \le \log(em) \le 2^l \beta$. Furthermore, by Lemma 16 in the Appendix and (38) implying $c_\kappa \ge \sqrt{5}\left(2 + \frac{4}{\kappa}\right)^{\frac{2+\kappa}{4(1+\kappa)}} / e^{1/2(1+\kappa)}$, we have

$$c_\kappa^2 m^{\frac{\kappa}{2(1+\kappa)}} k^{\frac{2+\kappa}{2(1+\kappa)}} \ge 5 \cdot 2^l \beta.$$

Overall, we have the following bound:

$$\begin{aligned}
&\mathbb{P}\left(\exists i > k : w_i^* \ge C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2\right) \\
&\quad \le \exp\left(4 \cdot 2^l \beta - 5 \cdot 2^l \beta\right) \le \exp(-2^l \beta).
\end{aligned}$$

Thus, with probability at least $1 - \exp(-2^l \beta)$,

$$w_i^* \le C' \left(\frac{m}{i}\right)^{\frac{\kappa}{4(1+\kappa)}} \|\pi_l(t) - \pi_{l-1}(t)\|_2,\ \forall i > k,$$

hence with the same probability

$$\begin{aligned}
&\left(\sum_{i=k+1}^{m} (w_i^*)^{\frac{2(1+\kappa)}{\kappa}}\right)^{\frac{\kappa}{2(1+\kappa)}} \\
\le&C' \|\pi_l(t) - \pi_{l-1}(t)\|_2 \left(\sum_{i=k+1} \left(\frac{m}{i}\right)^{1/2}\right)^{\frac{\kappa}{2(1+\kappa)}} \\
\le&C' \|\pi_l(t) - \pi_{l-1}(t)\|_2 m^{\frac{\kappa}{4(1+\kappa)}} \left(\int_1^m \frac{dx}{x^{1/2}}\right)^{\frac{\kappa}{2(1+\kappa)}} \\
\le&2^{\frac{\kappa}{2(1+\kappa)}} C' \|\pi_l(t) - \pi_{l-1}(t)\|_2 m^{\frac{\kappa}{2(1+\kappa)}},
\end{aligned}$$

and the desired result follows. □

Lemma 12. *The following inequalities hold for any $\beta \ge 8$:*

$$\mathbb{P}\left(\left(\sum_{i=1}^{m} \widetilde{q}_i^2\right)^{1/2} \ge C' \sqrt{\beta m}\right) \le 2e^{-\beta},$$

$$\mathbb{P}\left(\left(\sum_{i=1}^{m} \widetilde{q}_i^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}} \ge C''(\beta m)^{\frac{1}{2(1+\kappa)}}\right) \le 2e^{-\beta},$$

for some positive constants $C' = C'(\phi, \kappa)$, $C'' = C''(\phi, \kappa)$.

Proof. Recall that $\widetilde{q}_i = \text{sign}(q_i)(|q_i| \wedge \tau)$, $\tau = m^{1/2(1+\kappa)}$, and $\phi = \mathbb{E}\left(q_i^{2(1+\kappa)}\right)$. Thus, $\mathbb{E}(\widetilde{q}_i^2) \le \mathbb{E}(q_i^2) \le \phi^{1/1+\kappa}$, and for any integer $p \ge 2$, we have

$$\begin{aligned}
\mathbb{E}\left(\widetilde{q}_i^{2p}\right) &= \mathbb{E}\left(\widetilde{q}_i^{2p-2(1+\kappa)} \widetilde{q}_i^{2(1+\kappa)}\right) \\
&\le m^{\frac{p-1-\kappa}{1+\kappa}} \mathbb{E}\left(q_i^{2(1+\kappa)}\right) \le m^{\frac{p-1-\kappa}{1+\kappa}} \phi.
\end{aligned}$$

Thus, for any $p \ge 2$,

$$\begin{aligned}
&\mathbb{E}\left(|\widetilde{q}_i^2 - \mathbb{E}(\widetilde{q}_i^2)|^p\right) \le \mathbb{E}\left(\widetilde{q}_i^{2p}\right) + \left(\mathbb{E}(q_i^2)\right)^p \\
&\le m^{\frac{p-1-\kappa}{1+\kappa}} \phi + \phi^{\frac{p}{1+\kappa}} \le (m+\phi)^{\frac{1-\kappa}{1+\kappa}} \phi (m+\phi)^{\frac{p-2}{1+\kappa}}.
\end{aligned}$$

By Bernstein's inequality (Lemma 4), with probability at least $1-2e^{-\beta}$,

$$\left|\frac{1}{m}\sum_{i=1}^m \widetilde{q}_i^2 - \mathbb{E}(\widetilde{q}_i^2)\right| \leq \left(\frac{\sqrt{2\beta}(m+\phi)^{\frac{1-\kappa}{2(1+\kappa)}}\phi^{1/2}}{m^{1/2}} + \frac{\beta(m+\phi)^{\frac{1}{1+\kappa}}}{m}\right) \leq \frac{\sqrt{2\beta}(1+\phi)^{\frac{1-\kappa}{2(1+\kappa)}}\phi^{1/2} + \beta(1+\phi)^{\frac{1}{1+\kappa}}}{m^{\frac{\kappa}{1+\kappa}}},$$

which implies the first claim. To establish the second claim, note that for any $p \geq 2$,

$$\begin{aligned}
&\mathbb{E}\left|\widetilde{q}_i^{2(1+\kappa)} - \mathbb{E}\left(\widetilde{q}_i^{2(1+\kappa)}\right)\right|^p \\
\leq& C(p)\left(\mathbb{E}\left|\widetilde{q}_i^{2(1+\kappa)p}\right| + \left(\mathbb{E}\left|q_i^{2(1+\kappa)}\right|\right)^p\right) \\
\leq& C(p)\left(\mathbb{E}\left|\widetilde{q}_i^{2(1+\kappa)(p-1)}q_i^{2(1+\kappa)}\right| + \phi^p\right) \\
\leq& C(p)(m^{p-1}\phi + \phi^p) \leq C(p)(m+\phi)^{p-2}(m+\phi)\phi,
\end{aligned}$$

where we used the fact that $|\widetilde{q}_i| \leq m^{1/2(1+\kappa)}$ to obtain the third inequality. Bernstein's inequality implies that with probability at least $1-2e^{-\beta}$,

$$\left|\frac{1}{m}\sum_{i=1}^m \widetilde{q}_i^{2(1+\kappa)} - \mathbb{E}\left(\widetilde{q}_i^{2(1+\kappa)}\right)\right| \leq \sqrt{2\beta}(1+\phi)\phi^{1/2} + \beta(1+\phi),$$

which yields the second part of the claim. □

Proof of inequality (32) for the index set I_2. Combining Lemmas 10 and 11 with the inequality (35), and setting $u = 2^{l/2}\sqrt{\beta}$, we get that with probability at least $1-4\exp(-2^l\beta)$, for all $l \in I_2$,

$$|Z(\pi_l(t)) - Z(\pi_{l-1}(t))| \leq C\|\pi_l(t)-\pi_{l-1}(t)\|_2\frac{2^{l/2}\sqrt{\beta}}{m}\left(\left(\sum_{i=1}^m \widetilde{q}_i^2\right)^{1/2} + m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^m \widetilde{q}_i^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right),$$

for some constant $C = C(\kappa,\phi) > 0$; note that the factor $1/m$ appears due to equality (34). Next, we apply a chaining argument similar to the one used in Section VI-E1, we obtain that with probability at least $1-ce^{-\beta/2}$,

$$\sup_{t\in T}\left|\sum_{l\in I_2}(Z(\pi_l(t)) - Z(\pi_{l-1}(t)))\right| \leq C\frac{\gamma_2(T)\sqrt{\beta}}{m}\cdot \left(\left(\sum_{i=1}^m \widetilde{q}_i^2\right)^{1/2} + m^{\frac{\kappa}{2(1+\kappa)}}\left(\sum_{i=1}^m \widetilde{q}_i^{2(1+\kappa)}\right)^{\frac{1}{2(1+\kappa)}}\right), \tag{39}$$

for a positive constant $C = C(\kappa,\phi)$ and an absolute constant $c > 0$. In order to handle the remaining terms involving $\widetilde{q}_i$ in (39), we apply Lemma 12, which gives

$$\sup_{t\in T}\left|\sum_{l\in I_2}(Z(\pi_l(t)) - Z(\pi_{l-1}(t)))\right| \leq C\frac{\gamma_2(T)\beta}{\sqrt{m}},$$

with probability at least $1-ce^{-\beta/2}$, where $C = C(\kappa,\phi)$ and $c > 0$ are positive constants and $\beta \geq 8$. This completes the second part of the chaining argument. □

3) The case $l \in I_3$:

Proof of inequality (32) for the index set I_3. Direct application of Cauchy-Schwartz on (34) yields, for all $t \in T$,

$$|Z(\pi_l(t)) - Z(\pi_{l-1}(t))| \leq \left(\frac{1}{m}\sum_{i=1}^m w_i^2\right)^{1/2}\left(\frac{1}{m}\sum_{i=1}^m \widetilde{q}_i^2\right)^{1/2},$$

where $w_i = \left\langle \widetilde{U}_i, \pi_l(t) - \pi_{l-1}(t)\right\rangle$ are sub-Gaussian random variables. Thus, by Lemma 14, ω_i^2 are sub-exponential with norm bounded as in (36). Using Bernstein's inequality again, we deduce that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m (w_i^2 - \mathbb{E}(w_i^2))\right| \geq 2\|\widetilde{U}_i\|_{\psi_2}^2\|\pi_l(t)-\pi_{l-1}(t)\|_2^2\left(\sqrt{\frac{2u}{m}} + \frac{u}{m}\right)\right) \leq 2\exp(-u).$$

Let $u = 2^l\beta$. Using the fact that $2^l\beta/m \geq 1$ as well as $\mathbb{E}(w_i^2) = \|\pi_l(t)-\pi_{l-1}(t)\|_2^2$, we see that the term u/m dominates the right hand side and

$$\mathbb{P}\left(\left(\frac{1}{m}\sum_{i=1}^m w_i^2\right)^{1/2} \geq C\|\pi_l(t)-\pi_{l-1}(t)\|_2\frac{2^{l/2}\sqrt{\beta}}{\sqrt{m}}\right) \leq 2\exp(-2^l\beta),$$

for some absolute constant $C > 0$. Thus, repeating a chaining argument of section VI-E1 (namely, the argument following (33)), we obtain

$$\sup_{t\in T}\left|\sum_{l\in I_3}(Z(\pi_l(t)) - Z(\pi_{l-1}(t)))\right| \leq C\frac{\gamma_2(T)\sqrt{\beta}}{\sqrt{m}}\left(\frac{1}{m}\sum_{i=1}^m \widetilde{q}_i^2\right)^{1/2}$$

with probability at least $1-ce^{-\beta/2}$ for some absolute constants $C, c > 0$. Combining this inequality with the first claim of Lemma 12 gives

$$\sup_{t\in T}\left|\sum_{l\in I_3}(Z(\pi_l(t)) - Z(\pi_{l-1}(t)))\right| \leq C\frac{\gamma_2(T)\beta}{\sqrt{m}},$$

with probability at least $1 - ce^{-\beta/2}$ for absolute constants $C, c > 0$ and any $\beta \geq 8$. This finishes the bound for the third (and final) segment of the "chain". □

4) Finishing the proof of Lemma 8:

Proof. So far, we have shown that

$$\begin{aligned}
&\sup_{t\in T} |Z(t) - Z(t_0)| \\
&= \sup_{t\in T} \left| \sum_{l\geq 1} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \\
&\leq \sum_{j\in\{1,2,3\}} \sup_{t\in T} \left| \sum_{l\in I_j} (Z(\pi_l(t)) - Z(\pi_{l-1}(t))) \right| \\
&\leq C \frac{\gamma_2(T)\beta}{\sqrt{m}}, \qquad (40)
\end{aligned}$$

with probability at least $1 - ce^{-\beta/2}$ for some positive constants $C = C(\kappa, \phi)$ and c, and any $\beta \geq 8$. To finish the proof, it remains to bound $|Z(t_0)| = \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \widetilde{q}_i \left\langle \widetilde{U}_i, t_0 \right\rangle \right|$. With $\Delta_d(T)$ defined in (29), and since t_0 is an arbitrary point in T, we trivially have $\|t_0\|_2 \leq \Delta_d(T)$. Applying Bernstein's inequality in a way similar to Section VI-E1 yields

$$\begin{aligned}
&\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \widetilde{q}_i \left\langle \widetilde{U}_i, t_0 \right\rangle \right| \right. \\
&\left. \geq \left(\frac{C'\sqrt{2u}}{\sqrt{m}} + \frac{C''u}{m^{1-\frac{1}{2(1+\kappa)}}} \right) \Delta_d(T) \right) \leq 2e^{-u},
\end{aligned}$$

for some constants $C' = C'(\kappa, \phi)$, $C'' = C''(\kappa, \phi) > 0$ and any $u > 0$. Choosing $u = \beta$ gives

$$\mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \widetilde{q}_i \left\langle \widetilde{U}_i, t_0 \right\rangle \right| \geq \frac{C\Delta_d(T)\beta}{\sqrt{m}} \right) \leq 2e^{-\beta},$$

for a constant $C = C(\kappa, \phi) > 0$ and any $\beta \geq 0$. Combining this bound with (40) shows that with probability at least $1 - ce^{-\beta/2}$,

$$\begin{aligned}
\sup_{t\in T} \left| \frac{1}{m} \sum_{i=1}^m \varepsilon_i \langle \widetilde{U}_i, t \rangle \widetilde{q}_i \right| &\leq C \frac{(\gamma_2(T) + \Delta_d(T))\beta}{\sqrt{m}} \\
&\leq C \frac{(L\omega(T) + \Delta_d(T))\beta}{\sqrt{m}},
\end{aligned}$$

for $C = C(\kappa, \phi)$, an absolute constant $L > 0$ and all $\beta \geq 8$; note that the last inequality follows from Lemma 2. We have established (30), thus completing the proof. □

References

[1] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*, pages 3–66. Springer, 2015.

[2] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[3] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,. *IEEE Transaction on Information Theory*, 52(2):5406–5425, 2006.

[4] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58:3, 2011.

[6] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[7] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[8] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory,*, 61(5):2886–2908, 2015.

[9] W. Hardle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178, 1993.

[10] T. M. Stoker. Consistent estimation of scaled coefficients. *Econometrica: Journal of the Econometric Society*, pages 1461–1481, 1986.

[11] M. Hristache, A. Juditsky, and V. Spokoiny. Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623, 2001.

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

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

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

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

[16] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

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

[18] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2013.

[19] M. Genzel. High-dimensional estimation of structured signals from non-linear observations with general convex loss functions. *arXiv preprint arXiv:1602.03436*, 2016.

[20] 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*, pages 3420–3428, 2015.

[21] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. In *Advances in Neural Information Processing Systems*, pages 1549–1557, 2015.

[22] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. PAMI*, 31(2):210–227, 2009.

[23] Roman Vershynin. High-dimensional probability. *Cambridge Univ Press*, 2017.
[24] M. Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer, 2014.
[25] S. Mendelson. Upper bounds on product and multiplier empirical processes. *arXiv preprint arXiv:1410.8003*, 2014.
[26] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Etats-Unis Mathématicien. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.
[27] S. Cambanis, S. Huang, and G. Simons. On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11:368–385, 1981.
[28] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
[29] A. Banerjee, S. Chen, F. Fazayeli, and V. Sivakumar. Estimation with norm regularization. *Advances Neural Information Processing Systems (NIPS) 27*, 2014.
[30] J. Fan, W. Wang, and Z. Zhu. Robust low-rank matrix recovery. *arXiv:1603.08315*, 2016.
[31] L. Wang. The l_1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.
[32] Q. Sun, W. Zhou, and J. Fan. Adaptive huber regression: Optimality and phase transition. *arXiv preprint arXiv:1706.06991*, 2017.
[33] R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.
[34] L. Goldstein and X. Wei. Non-gaussian observations in nonlinear compressed sensing via Stein discrepancies. *arXiv preprint arXiv:1609.08512*, 2016.
[35] K. Ball. *An elementary introduction to modern convex geometry*. Cambridge University Press, New York,, 1997.
[36] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2010.
[37] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer-Verlag, Berlin, 1991.
[38] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.
[39] S. Dirksen. Tail bounds via generic chaining. *arXiv preprint arXiv:1309.3522*, 2013.
[40] S. J. Montgomery-Smith. The distribution of rademacher sums. In *Proceedings of the AMS*, pages 517–522, 109(2), 1990.
[41] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

APPENDIX

Lemma 13. *For any nonnegative random variable X, if $\mathbb{P}(X > K\beta) \le ce^{-\beta/2}$ for some constants $K, c > 0$ and all $\beta \ge \beta_0 \ge 0$, then,*

$$\mathbb{E}(X) \le K\left(\beta_0 + 2ce^{-\beta_0/2}\right).$$

Proof. Using a well known identity for the expectation of non-negative random variables,

$$\begin{aligned}\mathbb{E}(X) = &\int_0^\infty \mathbb{P}(X > u)\,du = K\int_0^\infty \mathbb{P}(X > K\beta)\,d\beta \\ \le& K\left(\beta_0 + \int_{\beta_0}^\infty \mathbb{P}(X > K\beta)\,d\beta\right) \\ \le& K\left(\beta_0 + \int_{\beta_0}^\infty ce^{-\beta/2}d\beta\right) \\ =& K\left(\beta_0 + 2ce^{-\beta_0/2}\right).\end{aligned}$$

□

Lemma 14. *If X and Y are sub-Gaussian random variables, then the product XY is a subexponential random variable, and*

$$\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

Proof. See [41].

□

Lemma 15. *Let $k = \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor$ and $l \in I_2$, then, $\left(\frac{em}{k}\right)^k \le \exp(3\cdot 2^l\beta)$.*

Proof. If $k \ge 2$, then, $2^l\beta/\log(em/2^l\beta) \ge 2$, which implies $2^l\beta \ge 2\log(em/2^l\beta)$. Thus,

$$\begin{aligned}\left(\frac{em}{k}\right)^k \le& 2\exp\left(\frac{2^l\beta}{\log\frac{em}{2^l\beta}}\log\left(\frac{em}{\frac{2^l\beta}{\log\frac{em}{2^l\beta}}-1}\right)\right) \\ \le& 2\exp\left(\frac{2^l\beta}{\log\frac{em}{2^l\beta}}\log\left(\frac{em}{2^l\beta-\log\frac{em}{2^l\beta}}\log\frac{em}{2^l\beta}\right)\right) \\ \le& 2\exp\left(\frac{2^l\beta}{\log\frac{em}{2^l\beta}}\log\left(\frac{2em}{2^l\beta}\log\frac{em}{2^l\beta}\right)\right) \\ \le& \exp(3\cdot 2^l\beta),\end{aligned}$$

where the second from last inequality follows from $\left(\frac{em}{k}\right)^k \le \exp(3\cdot 2^l\beta)$, and the last inequality follows from $m \ge 2^l\beta$, thus, $\log(2em/2^l\beta)/\log(em/2^l\beta) \le 2$.

On the other hand, if $k = 1$, then, since $\log em \le 2^l\beta$, $\left(\frac{em}{k}\right)^k = em = \exp(\log em) \le \exp(2^l\beta)$, finishing the proof. □

Lemma 16. *With $m \ge 1, \beta \ge 1, \kappa \in (1, 0)$ and $l \in I_2 = \{l \ge 1 : \log em \le 2^l\beta < m\}$, the integer $k = \lfloor 2^l\beta/\log(em/2^l\beta)\rfloor$ satisfies $k \ge 1$, and*

$$\frac{\left(2+\frac{4}{\kappa}\right)^{\frac{2+\kappa}{2(1+\kappa)}}}{e^{1/(1+\kappa)}} m^{\frac{\kappa}{2(1+\kappa)}} k^{\frac{2+\kappa}{2(1+\kappa)}} \ge 2^l\beta.$$

Proof. Since $2^l\beta \ge \log(em) \ge 1$, it follows that $k \ge 1$, and thus $k \ge 2^l\beta/2\log(em/2^l\beta)$. It is then enough to show that

$$\frac{\left(1+\frac{2}{\kappa}\right)^{\frac{2+\kappa}{2(1+\kappa)}}}{e^{1/(1+\kappa)}}\left(\frac{m}{2^l\beta}\right)^{\frac{\kappa}{2(1+\kappa)}} \ge \left(\log\frac{em}{2^l\beta}\right)^{\frac{2+\kappa}{2(1+\kappa)}}.$$

Raising both sides to the power of $2(1+\kappa)/\kappa$, equivalently

$$\left(1+\frac{2}{\kappa}\right)^{\frac{2+\kappa}{\kappa}} \Big/ e^{\frac{2}{\kappa}} \geq \left(\log\frac{em}{2^l\beta}\right)^{\frac{2+\kappa}{\kappa}} \Big/ \frac{m}{2^l\beta}.$$

Consider the function $g(x) = (\log ex)^{\frac{2+\kappa}{\kappa}}/x$. Note that as $m > 2^l\beta$, to prove the inequality above it suffices to show that the $\sup_{x\geq 1} g(x)$ is upper bounded by the left hand side. Taking the derivative of $g(x)$ yields

$$g'(x) = \frac{\frac{2+\kappa}{\kappa}(1+\log x)^{2/\kappa} - (1+\log x)^{(2+\kappa)/\kappa}}{x^2}.$$

Since $x \geq 1$, the only critical point at which the global maximum occurs is given by $x = e^{2/\kappa}$. As $g\left(e^{2/\kappa}\right)$ is exactly equal to the left hand side the proof is complete. □

Finally, we discuss some facts about decomposable norms that have been introduced in [28].

Definition 8. *Suppose that $\mathcal{L} \subseteq \mathcal{L}_1$ are two subspace of $\mathbb{R}^d$, and let $\mathcal{L}_1^\perp$ be the orthogonal complement of $\mathcal{L}_1$. Norm $\|\cdot\|_\mathcal{K}$ is said to be decomposable with respect to $(\mathcal{L},\ \mathcal{L}_1^\perp)$ if for any $\theta \in \mathbb{R}^d$,*

$$\|\theta_1+\theta_2\|_\mathcal{K} = \|\Pi_\mathcal{L}\theta_1\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\theta\|_\mathcal{K},$$

where $\Pi_\mathcal{L}$ and $\Pi_{\mathcal{L}_1^\perp}$ stand for the orthogonal projectors onto $\mathcal{L}$ and $\mathcal{L}_1^\perp$ respectively.

It is well known that many frequently used norms, including the ℓ_1 norm of a vector and the nuclear norm of a matrix, are decomposable with respect to the appropriately chosen pair of subspaces. For instance, the ℓ_1 norm is decomposable with respect to the pair of subspaces $(\mathcal{L}(J), \mathcal{L}(J)^\perp)$, where

$$\mathcal{L}(J) := \left\{v \in \mathbb{R}^d :\ v_j = 0 \text{ for all } j \notin J\right\} \quad (41)$$

consists of sparse vectors with non-zero coordinates indexed by a set $J \subseteq \{1,\dots,d\}$.

Let $W_1 \subseteq \mathbb{R}^{d_1},\ W_2 \subseteq \mathbb{R}^{d_2}$ be two linear subspaces. Then we define the subspace $\mathcal{L}(W_1, W_2) \subseteq \mathbb{R}^{d_1\times d_2}$ via

$$\mathcal{L}(W_1,W_2) := \left\{M \in \mathbb{R}^{d_1\times d_2} :\ \text{row}(M) \subseteq W_1,\ \text{col}(M) \subseteq W_2\right\},$$

where $\text{row}(M)$ and $\text{col}(M)$ are the linear subspaces spanned by the rows and columns of M respectively, and

$$\mathcal{L}_1^\perp(W_1,W_2) := \left\{M \in \mathbb{R}^{d_1\times d_2} :\ \text{row}(M) \subseteq W_1^\perp,\ \text{col}(M) \subseteq W_2^\perp\right\}. \quad (42)$$

Then the nuclear norm $\|\cdot\|_*$ is decomposable with respect to $\left(\mathcal{L}(W_1,W_2), \mathcal{L}_1^\perp(W_1,W_2)\right)$ (see [28] for details).

Assume that the norm $\|\cdot\|_\mathcal{K}$ is decomposable with respect to $(\mathcal{L}, \mathcal{L}_1^\perp)$, and let $\theta \in \mathcal{L}$. It is clear that for any $\mathbf{v} \in \mathbb{S}_{c_0}(\theta)$

$$\begin{aligned}\|\theta+\mathbf{v}\|_\mathcal{K} &= \|\Pi_\mathcal{L}\theta + \Pi_{\mathcal{L}_1}\mathbf{v} + \Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K} \\ &\leq \|\Pi_\mathcal{L}\theta\|_\mathcal{K} + \frac{1}{c_0}\|\Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K}. \quad (43)\end{aligned}$$

Since $\theta \in \mathcal{L}$, decomposability and the triangle inequality imply that

$$\begin{aligned}&\|\Pi_\mathcal{L}\theta + \Pi_{\mathcal{L}_1}\mathbf{v} + \Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K} \\ &= \|\Pi_\mathcal{L}\theta + \Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K} \\ &\geq \|\Pi_\mathcal{L}\theta\|_\mathcal{K} - \|\Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K}.\end{aligned}$$

Substituting this bound into (43) gives

$$\begin{aligned}&-\|\Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K} + \|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K} \\ &\qquad\leq \frac{1}{c_0}\|\Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K} + \frac{1}{c_0}\|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K},\end{aligned}$$

which implies that for any $\mathbf{v} \in \mathbb{S}_{c_0}(\theta)$

$$\|\Pi_{\mathcal{L}_1^\perp}\mathbf{v}\|_\mathcal{K} \leq \frac{c_0+1}{c_0-1}\|\Pi_{\mathcal{L}_1}\mathbf{v}\|_\mathcal{K}.$$

It is easy to see that the set of all $\mathbf{v}$ satisfying the inequality above is a convex cone, which we will denote by $C_{c_0} = C_{c_0}(\mathcal{K})$. Since $\mathbb{S}_{c_0}(\theta) \subseteq C_{c_0}$,

$$\Psi\left(\mathbb{S}_{c_0}(\theta)\right) \leq \Psi\left(C_{c_0}\right)$$

by definition of the restricted compatibility constant. This inequality is useful due to the fact that it is often easier to estimate $\Psi\left(C_{c_0}\right)$.

Finally, we make a remark that is useful when dealing with non-isotropic measurements. Let $\mathbf{\Sigma} \succ 0$ be a $d\times d$ matrix, and consider the norm corresponding to the convex set $\mathbf{\Sigma}^{1/2}\mathcal{K}$, so that $\|\mathbf{v}\|_{\mathbf{\Sigma}^{1/2}\mathcal{K}} = \|\mathbf{\Sigma}^{-1/2}\mathbf{v}\|_\mathcal{K}$. It is easy to see that $C_{c_0}(\mathbf{\Sigma}^{1/2}\mathcal{K}) = \mathbf{\Sigma}^{1/2}C_{c_0}(\mathcal{K})$, hence

$$\begin{aligned}&\Psi\left(C_{c_0}(\mathbf{\Sigma}^{1/2}\mathcal{K}); \mathbf{\Sigma}^{1/2}\mathcal{K}\right) \\ &= \sup_{\mathbf{v}\in\mathbf{\Sigma}^{1/2}\mathcal{K}\setminus\{0\}} \frac{\|\mathbf{v}\|_{\mathbf{\Sigma}^{1/2}\mathcal{K}}}{\|\mathbf{v}\|_2} \\ &= \sup_{\mathbf{u}\in\mathcal{K}\setminus\{0\}} \frac{\|\mathbf{u}\|_\mathcal{K}}{\|\mathbf{\Sigma}^{1/2}\mathbf{u}\|_2} \\ &\leq \|\mathbf{\Sigma}^{-1/2}\|\,\Psi\left(C_{c_0}(\mathcal{K}); \mathcal{K}\right).\end{aligned}$$

Example 1: ℓ_1 norm. Let $\mathcal{L}(J)$ be as in (41) with $|J| = s \leq d$. If $v \in \mathbb{R}^d$ belongs to the corresponding cone $C(c_0)$, then clearly $\|v\|_1 \leq \frac{2c_0}{c_0-1}\|v_J\|_1$, where $v_J := \Pi_{\mathcal{L}(J)}v$. Hence

$$\|v\|_1 \leq \frac{2c_0}{c_0-1}\|v_J\|_1 \leq \frac{2c_0}{c_0-1}\sqrt{|J|}\|v\|_2,$$

and $\Psi(C_{c_0}) \leq \frac{2c_0}{c_0-1}\sqrt{s}$.

Example 2: nuclear norm. Let $\mathcal{L}_1^\perp(W_1, W_2)$ be as in

(42). Note that for any $v \in \mathbb{R}^{d_1\times d_2}$, $\Pi_{\mathcal{L}_1^{\perp}(W_1,W_2)}v = \Pi_{W_2^{\perp}}v\Pi_{W_1^{\perp}}$, where $\Pi_{W_1^{\perp}}$ and $\Pi_{W_2^{\perp}}$ are the orthogonal projectors onto subspaces $W_1 \subseteq \mathbb{R}^{d_1}$ and $W_2 \subseteq \mathbb{R}^{d_2}$ respectively. Then for any $v \in C_{c_0}$, we have that

$$\begin{aligned}\|v\|_* &\leq \|\Pi_{\mathcal{L}_1^{\perp}(W_1,W_2)}v\|_* + \|\Pi_{\mathcal{L}_1(W_1,W_2)}v\|_* \\ &\leq \frac{2c_0}{c_0-1}\|\Pi_{\mathcal{L}_1(W_1,W_2)}v\|_*. \qquad (44)\end{aligned}$$

Note that

$$\begin{aligned}\Pi_{\mathcal{L}_1(W_1,W_2)}v &= v - \Pi_{W_2^{\perp}}v\Pi_{W_1^{\perp}} \\ &= \Pi_{W_2^{\perp}}v\Pi_{W_1} + \Pi_{W_2}v,\end{aligned}$$

hence

$$\text{rank}\left(\Pi_{\mathcal{L}_1(W_1,W_2)}v\right) \leq 2\max\left(\dim(W_1),\dim(W_2)\right),$$

which yields together with (44) that

$$\begin{aligned}\|v\|_* &\leq \frac{2c_0}{c_0-1}\|\Pi_{\mathcal{L}_1(W_1,W_2)}v\|_* \\ &\leq \frac{2c_0}{c_0-1}\sqrt{2\max\left(\dim(W_1),\dim(W_2)\right)}\|v\|_2,\end{aligned}$$

and $\Psi(C_{c_0}) \leq \frac{2\sqrt{2}c_0}{c_0-1}\sqrt{\max\left(\dim(W_1),\dim(W_2)\right)}$.

Larry Goldstein is Professor in the Department of Mathematics at the University of Southern California. Goldstein completed his Ph.D at UCSD in 1984, where he also earned a Masters degree in Electrical Enginering. His current research mainly focuses on the applications of Stein's method to modern day statistics.

Stanislav Minsker is an assistant professor in the Department of Mathematics at the University of Southern California. Before joining USC, he held the Visiting Assistant Professor position at Duke University and was the Quantitative Associate at Wells Fargo Securities. Stanislav completed his Ph.D at Georgia Institute of Technology in 2012. His current research focuses on high dimensional statistics and statistical learning theory, specifically on the analysis of robust algorithms and distributed statistical estimation.

Xiaohan Wei received B.S. degree in electrical engineering and information science from University of Science and Technology of China, Hefei, Anhui, China in 2012, and M.S. degree with honor in electrical engineering from University of Southern California, Los Angeles, in 2014. He is now working towards Ph.D. degree in electrical engineering. His research is in the area of stochastic optimization and statistical learning theory.