

Max-Linear Regression by Convex Programming

Seonho Kim, *Student Member, IEEE*, Sohail Bahmani, and Kiryung Lee, *Senior Member, IEEE*

Abstract—We consider the multivariate max-linear regression problem where the model parameters $\beta_1, \dots, \beta_k \in \mathbb{R}^p$ need to be estimated from n independent samples of the (noisy) observations $y = \max_{1 \leq j \leq k} \beta_j^\top x + \text{noise}$. The max-linear model vastly generalizes the conventional linear model, and it can approximate any convex function to an arbitrary accuracy when the number of linear models k is large enough. However, the inherent nonlinearity of the max-linear model renders the estimation of the regression parameters computationally challenging. Particularly, no estimator based on convex programming is known in the literature. We formulate and analyze a scalable convex program given by anchored regression (AR) as the estimator for the max-linear regression problem. Under the standard Gaussian observation setting, we present a non-asymptotic performance guarantee showing that the convex program recovers the parameters with high probability. When the k linear components are equally likely to achieve the maximum, our result shows a sufficient number of noise-free observations for exact recovery scales as $k^4 p$ up to a logarithmic factor. This sample complexity coincides with that by alternating minimization (Ghosh et al., 2021). Moreover, the same sample complexity applies when the observations are corrupted with arbitrary deterministic noise. We provide empirical results that show that our method performs as our theoretical result predicts, and is competitive with the alternating minimization algorithm particularly in presence of multiplicative Bernoulli noise. Furthermore, we also show empirically that a recursive application of AR can significantly improve the estimation accuracy.

Index Terms—nonlinear regression, convex programming, max-linear model, empirical processes, and sample complexity

I. INTRODUCTION

WE consider the problem of estimating the parameters $\beta_{*,1}, \dots, \beta_{*,k} \in \mathbb{R}^p$ that determine the *max-linear* function

$$x \in \mathbb{R}^p \mapsto \max_{j \in [k]} \langle \beta_{*,j}, x \rangle, \quad (\text{I.1})$$

from independent and identically distributed (i.i.d.) observations, where $[k]$ denotes the set $\{1, \dots, k\}$. Specifically, given the data points $x_1, \dots, x_n \in \mathbb{R}^p$, and denoting the value of a max-linear function, with parameter β , at these points by

$$f_i(\beta) := \max_{j \in [k]} \langle x_i, \beta_j \rangle, \quad (\text{I.2})$$

we observe the nonlinear measurement

$$y_i = f_i(\beta_*) + w_i,$$

of the parameter vector $\beta_* = [\beta_{*,1}; \dots; \beta_{*,k}] \in \mathbb{R}^{kp}$ where w_i denotes noise for $i \in [n]$.

The most relevant prior work studied an *alternating minimization* (AM) algorithm to solve a slightly more general problem

of max-affine regression [1]. Each iteration consists of a step to identify the maximizing linear models followed by least-squares update of model parameters. However, we observed that their empirical performance significantly degrades with outliers, mainly due to the sensitivity of the “maximizer identification” step. Leveraging recent theory for convexifying nonlinear inverse problems in the original domain [2], [3], [4], we propose an alternative approach by convex programming. Due to the inherent geometry of the formulation, the convex estimator provides stable performance in the presence of adversarial noise. It is worth mentioning that Ghosh et al. [1] considered a random noise model, whereas we consider a deterministic “gross error” model. Nevertheless, in the noiseless case, both results achieve exact parameter recovery at comparable sample complexities.

A. Convex estimator

The common estimators for β_* such as the *least absolute deviation* (LAD), i.e.,

$$\underset{\beta}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n |f_i(\beta) - y_i|, \quad (\text{I.3})$$

are generally hard to compute as they involve nonconvex optimization. Given an “anchor vector” θ , we study the estimation of β_* through *anchored regression* (AR) that formulates the estimation by the convex program

$$\begin{aligned} &\underset{\beta}{\text{maximize}} \quad \langle \theta, \beta \rangle \\ &\text{subject to} \quad \frac{1}{n} \sum_{i=1}^n (f_i(\beta) - y_i)_+ \leq \eta, \end{aligned} \quad (\text{I.4})$$

where $(\cdot)_+$ denotes the positive-part function. The parameter η should be chosen so that the feasible set of (I.4) is not empty. The anchored regression can be interpreted as a convexification of the LAD estimator. Since the observation functions (I.2) are convex, the LAD is nonconvex mainly due to the effect of the absolute value operator in (I.3). This source of nonconvexity is removed in anchored regression by relaxing the absolute deviation to the positive part of the error. The linear objective that is determined by the anchor vector θ acts as a “regularizer” to prevent degenerate solutions and guarantees exact recovery of the true parameter β_* under certain conditions on the measurement model in the noiseless scenario.

Anchored regression has been originally developed as a scalable convex program to solve the phase retrieval problem [2], [5] with provable guarantees. Anchored regression is highly scalable compared to other convex relaxations in this context [6], [7] that rely on semidefinite programming. The idea of anchored regression is further studied in a broader class of nonlinear parametric regression problems with convex observations [4] and *difference of convex* functions [3].

S.K. and K.L. are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43220 USA (e-mail: kim.7604@osu.edu; kiryung@ece.osu.edu).

S.B. is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sohail.bahmani@ece.gatech.edu).

B. Background and motivation

A closely related problem is the max-affine regression problem. A max-affine model generalizes the max-linear model in (I.1) by introducing an extra offset parameter to each component. Alternatively, by fixing any regressor to constant 1, each linear component in (I.1) has its range away from the origin, which turns the model into a max-affine model. Thus methods developed for the two models are compatible. For the sake of simplicity, we use the description in (I.1). If necessary, a coordinate of \mathbf{x}_i can be fixed to 1 for all $i \in [n]$.

Since the max-affine model can approximate a convex function to an arbitrary accuracy, it has been utilized in numerous applications, particularly in machine learning and optimization. Recently it has been shown that an extension called the *max-affine spline operators* (MASOs) can represent a large class of *deep neural networks* (DNNs) [8], [9], [10]. They leveraged the model to analyze the expressive power of various DNNs. Max-affine model has also been leveraged to approximate Bregman divergences in metric learning [11] and utility functions in energy storage and beer brewery optimization problems [12].

As mentioned above, the max-affine and max-linear regression problems are challenging due to the inherent nonlinearity in the model. In the literature, the max-affine regression problem has been studied mostly as a nonlinear least squares [13], [14], [15], [12], [1], [16]:

$$\hat{\beta} = \operatorname{argmin}_{[\beta_1, \dots, \beta_k]} \sum_{i=1}^n \left(\max_{1 \leq j \leq k} \langle \mathbf{x}_i, \beta_j \rangle - \mathbf{y}_i \right)^2. \quad (\text{I.5})$$

By utilizing a special structure in (I.5), a suite of iterative optimization algorithms have been developed [13], [14], [15]. The fact that (I.1) is a special case of piecewise linear function allows us to divide $\mathbf{x}_1, \dots, \mathbf{x}_n$ into k partitions based on their membership in the polyhedral cones

$$\mathcal{C}_j := \{\mathbf{w} \in \mathbb{R}^p : \langle \mathbf{w}, \beta_{*,j} - \beta_{*,l} \rangle \geq 0, \forall l \neq j\}, \quad j \in [k], \quad (\text{I.6})$$

which are pairwise *almost disjoint*¹ and cover the entire space \mathbb{R}^p . In other words, \mathcal{C}_j is determined according to which component achieves the maximum in the max-linear model in (I.1). If this oracle information is known a priori, then the estimation is divided into k decoupled linear least squares given by

$$\hat{\beta}_j = \operatorname{argmin}_{\beta_j} \sum_{i \in \mathcal{C}_j} (\langle \mathbf{x}_i, \beta_j \rangle - \mathbf{y}_i)^2, \quad j \in [k]. \quad (\text{I.7})$$

However, since the oracle partition information is not available in practice, various adaptive partitioning methods have been studied. Magnani and Boyd [13] proposed the *least-squares partition algorithm*, which is an alternating minimization algorithm that progressively refines the estimates for both model parameters and partitions similar to the k -means clustering algorithm. Hannah and Dunson [15] proposed the *convex adaptive partitioning* (CAP) method, which is a greedy algorithm that builds a partitioning of covariates through

dyadic splitting and refit. They have shown that the CAP method is asymptotically consistent. Balász [12] proposed the *adaptive max-affine partitioning algorithm*, which combines AM and CAP using a cross-validation scheme and significantly reduces computation time by partitioning at the median. The performance of all these algorithms critically depends on the initialization. Moreover, they proposed an initialization scheme based on a random search, but its search space grows exponentially in p . In a later work, Ghosh et al. [1] further improved the random search method by using a spectral method so that the size of the search space does not depend on p , even though it grows exponentially in k . Their initialization scheme remains a practical method when $k = O(1)$. Toriello and Vielma [14] formulated (I.5) as a mixed integer program based on the “big-M” method. Similar to other methods based on mixed integer programming, their method also suffers from a high computational cost and does not scale well to large instances. Ho et al. [16] applied the *DC algorithm* [17] to a reformulation of (I.5) as a *difference-of-convex* program. They showed, empirically, fast convergence of their algorithm to a local minimum.

The aforementioned methods demonstrated satisfactory empirical performance on selected benchmark sets at a tractable computational cost. However, except for the method of Ghosh et al. [1], these methods lack non-asymptotic statistical guarantees even under reasonable simplifying assumptions. A non-asymptotic analysis for the alternating minimization method is first established in [1] which also provides a provably accurate initialization scheme.

C. Contributions

We provide a scalable convex estimator for the max-linear regression problem that is formulated as a linear program and is backed by statistical guarantees. Under the standard Gaussian covariate model, the convex estimator (I.4) is guaranteed to recover the regression parameters exactly with high probability if the number of observations n scales as $\pi_{\min}^{-4} p$ up to some logarithmic factors where π_{\min} is defined as $\min_{j \in [k]} \mathbb{P}(\mathbf{g} \in \mathcal{C}_j)$ for $\mathbf{g} \in \text{Normal}(\mathbf{0}, \mathbf{I}_p)$. This sample complexity implicitly depends on k (i.e., the number of components) through π_{\min} . Particularly, when the k linear components form a “well-balanced partition” in the sense that they are equally likely to achieve the maximum, the smallest probability π_{\min} is close to $1/k$ and the derived sample complexity reduces to $k^4 p$ up to the logarithmic factors. This is comparable to the sufficient condition for exact recovery $n = \mathcal{O}(kp\pi_{\min}^{-3})$ of alternating minimization algorithm [1] in the noise-free scenario. Monte Carlo simulations show that our proposed convex estimator, as a convexification of the LAD estimator, exhibits robustness against outliers, whereas AM appears to be fragile in the presence of impulsive noise. Furthermore, the repetition of AR significantly improves the accuracy of the estimation.

II. ACCURACY OF THE CONVEX ESTIMATOR

In this section, we provide our main results on the estimation error of the convex program in (I.4). We consider the

¹We say two sets are almost disjoint whenever their intersection has measure zero with respect to an underlying measure.

anchor vector θ constructed from a given initial estimate $\tilde{\beta} = [\tilde{\beta}_1; \dots; \tilde{\beta}_k] \in \mathbb{R}^{kp}$ as

$$\theta = \frac{1}{2n} \sum_{i=1}^n \nabla f_i(\tilde{\beta}) = \frac{1}{2n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\{x_i \in \tilde{\mathcal{C}}_j\}} e_j \otimes x_i, \quad (\text{II.1})$$

where

$$\tilde{\mathcal{C}}_j := \left\{ w \in \mathbb{R}^p : \langle w, \tilde{\beta}_j - \tilde{\beta}_l \rangle \geq 0, \forall l \neq j \right\}, \quad j \in [k] \quad (\text{II.2})$$

and $e_j \in \mathbb{R}^k$ denotes the j th column of the k -by- k identity matrix I_k for $j \in [k]$. Since f_i is differentiable except on a set of measure zero, with a slight abuse of terminology, ∇f_i in (II.1) is referred to as the “gradient”. In (II.1), the choice of anchor vector follows from the geometry of convex equations [4, Section 1.4]. In particular, in the noiseless case, β_* would be a solution to

$$\begin{aligned} & \underset{\beta}{\text{maximize}} && \langle \theta, \beta \rangle \\ & \text{subject to} && f_i(\beta) \leq y_i, \quad \forall i \in [n]. \end{aligned}$$

if it satisfies the Karush–Kuhn–Tucker condition

$$-\theta + \sum_{i=1}^n \lambda_i \nabla f_i(\beta_*) = 0$$

for some $\lambda_1, \dots, \lambda_n \geq 0$. In other words, the anchor vector θ needs to be in the cone $(\{\nabla f_i(\beta)\}_{i=1}^n)$. The choice of θ in (II.1) is inspired by this condition.

The following theorem illustrates the sample complexity and the corresponding estimation error achieved by the estimator in (I.4). The estimation error is measured as the sum of the ℓ_2 norms of the difference between the corresponding components of the ground truth β_* and the estimate $\tilde{\beta}$.

Theorem 1: Let $\{\mathcal{C}_j\}_{j=1}^k$ and $\{\tilde{\mathcal{C}}_j\}_{j=1}^k$ be respectively defined as in (I.6) and (II.2). Let θ be as in (II.1) and $\{x_i\}_{i=1}^n$ be independent copies of $g \sim \text{Normal}(\mathbf{0}, I_p)$. Then there exist absolute constants $c, C > 0$, for which the following statement holds for all $w \in \mathbb{R}^n$ with probability at least $1 - \delta$: Suppose that $\tilde{\beta}$ is independent of $\{x_i\}_{i=1}^n$ satisfies

$$\begin{aligned} & \frac{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{*,j} - \beta_{*,j'})\|_2}{\|\beta_{*,j} - \beta_{*,j'}\|_2} \leq \\ & \min \left(0.1, \frac{c\pi_{\min}^4}{2k} \log^{-1/2} \left(\frac{k}{c\pi_{\min}^4} \right) \right), \quad \forall j, j' \in [k] : j \neq j'. \end{aligned} \quad (\text{II.3})$$

If the feasible set of the optimization problem in (I.4) is not empty and the number of observations satisfies

$$n \geq C \zeta^{-2} (4p \log^3 p \log^5 k + 4 \log(1/\delta) \log k), \quad (\text{II.4})$$

where

$$\zeta := \min_{j \in [k]} \sqrt{\frac{\pi}{32}} \mathbb{P}^2\{g \in \mathcal{C}_j\} - 2 \max_{j \in [k]} \sqrt{\mathbb{P}\{g \in \tilde{\mathcal{C}}_j \triangle \mathcal{C}_j\}},$$

then the solution $\hat{\beta}$ to (I.4) obeys

$$\sum_{j=1}^k \|\beta_{*,j} - \hat{\beta}_j\|_2 \leq \frac{2}{\zeta} \left(\eta + \frac{1}{n} \sum_{i=1}^n (w_i)_+ \right). \quad (\text{II.5})$$

To make the optimization problem in (I.4) feasible, it suffices to include the ground-truth β_* in the feasible set, i.e.

$$\eta \geq \frac{1}{n} \sum_{i=1}^n (-w_i)_+, \quad (\text{II.6})$$

The error bound in (II.5) reduces to $\frac{2}{\zeta n} \sum_{i=1}^n |w_i|$ when the parameter η is chosen so that the equality in (II.6) is achieved. In practice, the noise entries are unknown and this error cannot be achieved. If η , as a parameter that determines the power of the adversary, is chosen so that $\eta \geq \|w\|_1/n$, then the resulting error bound becomes $\frac{4\|w\|_1}{n\zeta}$. In particular, if η satisfies $\eta \geq \|w\|_\infty$, then the resulting error bound will be $\frac{4\|w\|_\infty}{\zeta}$. The latter condition will be readily satisfied in practical applications. Furthermore, as shown in the empirical sensitivity analysis in Section III, the estimation error does not crucially depend on the choice of η .

A. Comparison with an oracle estimator

Assuming that the additive noise is i.i.d. sub-Gaussian with zero mean and variance σ^2 , the error bound in (II.5) becomes $\tilde{O}(\sigma/\zeta)$, which implies that our estimator is not consistent. However, in the adversarial noise setting which is our focus, we can compare the performance case of our estimator with an oracle-assisted estimator, similar to the analysis carried out in [18] for the matrix completion problem. In this scenario, the error bound by the convex estimator nearly matches the performance of an oracle-assisted estimator (up to a factor determined by β_*).

Lemma 1: Consider the same regression problem as in Theorem 1 with $\{x_i\}_{i=1}^n$ being independent copies of $g \sim \text{Normal}(\mathbf{0}, I_p)$. Suppose that $\{\mathcal{C}_j\}_{j=1}^k$ in (I.6) is given as the oracle information. Then there exists an absolute constant $C > 0$ such that if

$$n \geq C \pi_{\min}^{-2} \max(kp \log(n/p), \log(1/\delta)), \quad (\text{II.7})$$

then the estimates $\{\hat{\beta}_j\}_{j=1}^k$ obtained through the decoupled least-squares (I.7) satisfy

$$\sup_{\|w\|_\infty \leq \eta'} \sum_{j=1}^k \|\beta_{*,j} - \hat{\beta}_j\|_2 \gtrsim \frac{\pi_{\min}^{3/2} \eta'}{\pi_{\max}} \quad (\text{II.8})$$

with probability at least $1 - \delta$, where $\pi_{\max} := \max_{j \in [k]} \mathbb{P}(g \in \mathcal{C}_j)$.

Proof: See Appendix A. ■

One expects that the oracle estimator nearly achieves the optimal performance. However, since the lower bound by Lemma 1 does not vanish as n increases to infinity, the oracle estimator is also biased in the presence of adversarial noise. Note that the lower bound in (II.8) remains the same with the feasible set substituted by $\|w\|_1 \leq n\eta'$. Furthermore, if η achieves the equality in (II.6), then the error bound in (II.5) implies

$$\sup_{\|w\|_1 \leq n\eta'} \sum_{j=1}^k \|\beta_{*,j} - \hat{\beta}_j\|_2 \leq \frac{2\eta'}{\zeta}. \quad (\text{II.9})$$

Therefore, in this scenario, the error bound in (II.9) matches that by the oracle estimator up to an extra factor $O(\pi_{\max}/\zeta \pi_{\min}^{3/2})$.

In particular, if $\pi_{\max} \approx \pi_{\min} \approx 1/k$, then the error by the convex estimator is sub-optimal up to a factor $k^{5/2}$ relative to the oracle estimator.

B. Initialization

Theorem 1 provides an error bound by the convex estimator given an initial estimate satisfying (II.3). Finding such an initial estimate is not a trivial task. Ghosh et al. [1] proposed an initialization scheme that consists of dimensionality reduction by a spectral method [1, Algorithm 2], followed by a low-dimensional random search [1, Algorithm 3]. It has been shown that if the observations are corrupted with independent sub-Gaussian noise, then the initialization scheme provides an estimate within a certain neighborhood of the ground-truth in a polynomial time when $k = O(1)$. Their proof only uses the fact that the maximum magnitude of sub-Gaussian noise entries is bounded with high probability. Below, we extend the analysis of their initialization scheme to the scenario where the noise vector \mathbf{w} is a fixed *deterministic vector* under the only condition that $\|\mathbf{w}\|_\infty \leq \eta'$.

To this end, we first recall the first stage in their initialization scheme that extracts the eigenvectors corresponding to the k dominant eigenvalues of the following matrix:

$$\widehat{\mathbf{M}} = \frac{2}{n} \left(\sum_{i=1}^{n/2} y_i \mathbf{x}_i \right) \left(\sum_{i=1}^{n/2} y_i \mathbf{x}_i \right)^\top + \frac{2}{n} \sum_{i=1}^{n/2} y_i (\mathbf{x}_i \mathbf{x}_i^\top - \mathbf{I}_p). \quad (\text{II.10})$$

Let $\widetilde{\mathbf{M}}$ denote the noise-free version of $\widehat{\mathbf{M}}$, i.e.,

$$\begin{aligned} \widetilde{\mathbf{M}} &= \frac{2}{n} \sum_{i=1}^{n/2} \left(\max_{j \in [k]} \langle \beta_{*,j}, \mathbf{x}_i \rangle \right) (\mathbf{x}_i \mathbf{x}_i^\top - \mathbf{I}_p) + \\ &\frac{2}{n} \left(\sum_{i=1}^{n/2} \left(\max_{j \in [k]} \langle \beta_{*,j}, \mathbf{x}_i \rangle \right) \mathbf{x}_i \right) \left(\sum_{i=1}^{n/2} \left(\max_{j \in [k]} \langle \beta_{*,j}, \mathbf{x}_i \rangle \right) \mathbf{x}_i \right)^\top. \end{aligned}$$

Then the ground-truth parameter vectors $\beta_1^*, \dots, \beta_k^*$ are in the columns space of $\mathbb{E}\widetilde{\mathbf{M}}$. Ghosh et al. [1] derived a tail bound on the perturbation of those eigenvectors due to sub-Gaussian noise. We provide an analogous perturbation analysis in the deterministic noise setting. The following lemma provides upper bounds on the contributions of the noise to the two summands in the right-hand side of (II.10).

Lemma 2: Suppose that $\mathbf{x}_1, \dots, \mathbf{x}_n \stackrel{\text{i.i.d.}}{\sim} \text{Normal}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{w} := (w_1, \dots, w_n) \in \mathbb{R}^n$ are arbitrary fixed. Then the following inequalities hold with probability at least $1 - \delta$:

$$\begin{aligned} \left\| \frac{1}{n} \sum_{i=1}^n w_i \mathbf{x}_i \right\|_2 &\lesssim \|\mathbf{w}\|_\infty \cdot \sqrt{\frac{p + \log(1/\delta)}{n}}, \\ \left\| \frac{1}{n} \sum_{i=1}^n w_i (\mathbf{x}_i \mathbf{x}_i^\top - \mathbf{I}_p) \right\| &\lesssim \\ \|\mathbf{w}\|_\infty \cdot \max \left(\sqrt{\frac{p + \log(1/\delta)}{n}}, \frac{p + \log(1/\delta)}{n} \right). \end{aligned} \quad (\text{II.11})$$

Proof: See Appendix C. ■

Let $\widehat{\mathbf{U}}$ be a matrix whose columns are the k dominant eigenvectors of $\widehat{\mathbf{M}}$. Furthermore, let the columns of \mathbf{U}^* be

the eigenvectors of the noise-free component of $\mathbb{E}\widetilde{\mathbf{M}}$. Then, plugging the results in Lemma 2 into the proof of [1, Lemma 8] yields that

$$\left\| \widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top - \mathbf{U}^* (\mathbf{U}^*)^\top \right\|_F^2 \lesssim \left(\frac{\|\mathbf{w}\|_\infty^2 + \max_{j \in [k]} \|\beta_{*,j}\|_1^2}{\lambda_k^2(\mathbb{E}\widetilde{\mathbf{M}})} \right) \frac{kp \log^3(pk/\delta)}{n} \quad (\text{II.12})$$

holds with probability at least $1 - \delta$. This is analogous to [1, Theorem 2] which addresses the case of the sub-Gaussian noise. The remainder of their initialization scheme does not depend on any assumption on the noise model. Therefore, the resulting initial estimate satisfies (II.3) if

$$\begin{aligned} n &\gtrsim \frac{k^6 \log(k/\pi_{\min})}{\pi_{\min}^{13}}, \\ \max \left\{ \|\mathbf{w}\|_\infty^2 \log \left(1 + \frac{\max_{j \in [k]} \|\beta_{*,j}\|_2 k^4 \log^{1/2}(k/\pi_{\min})}{\pi_{\min}^{5.5}} \right), \right. \\ &\left. \left(\|\mathbf{w}\|_\infty^2 + \max_{j \in [k]} \|\beta_{*,j}\|_1^2 \right) \frac{k^3 p \log^3(n/k) \cdot \max_{j \in [k]} \|\beta_{*,j}\|_2}{\lambda_k^2(\mathbb{E}\widetilde{\mathbf{M}})} \right\}. \end{aligned} \quad (\text{II.13})$$

The condition in (II.13) is obtained by applying the initialization condition (II.3) and substituting σ by $\|\mathbf{w}\|_\infty$ in [1, Equation 20]. The requirement for the initial point of anchored regression (II.3) is more relaxed in terms of the dependence on π_{\min} , compared to the similar requirement for the alternating minimization method [1, Theorem 1]. Furthermore, for both anchored regression and alternating minimization, the sample complexity of the initialization dominates that of the subsequent stages of the algorithms.

In the above paragraphs, we have shown that the anchored regression combined with the spectral initialization provides a stable estimate in the presence of an arbitrarily fixed deterministic noise of bounded magnitudes. However, this result does not extend to the adversarial noise setting in Theorem 1 and Lemma 1. Maximization over \mathbf{w} s that obey $\|\mathbf{w}\|_\infty \leq \eta'$ in (II.11), can be addressed effectively by taking the union bound over extreme points of ℓ_∞^n ball with the radius η' and choosing $\delta = 2^{-n\bar{\delta}}$ with $\bar{\delta} \in [0, 1]$ denoting overall error probability. Therefore, the terms $\frac{p + \log(1/\delta)}{n}$ in (II.11) are equal to $\frac{p + n \log(2) + \log(1/\bar{\delta})}{n}$, which are clearly bounded from below by $\log 2$. Consequently, in the adversarial setting, the error in the spectral method does not vanish as n grows, and the desired accuracy for the initialization scheme cannot be established. Considering a relaxed condition $\|\mathbf{w}\|_1 \leq n\eta'$ exacerbates the situation and the error bound in the spectral method becomes even larger.

C. Comparison with alternating minimization in computational cost

This section compares AR and AM in their computational costs. First, AR is implemented via an equivalent formulation with auxiliary variables $\mathbf{t} := [t_1; \dots; t_n] \in \mathbb{R}^n$ as

$$\begin{aligned}
& \underset{(\beta_j)_{j=1}^k, (t_i)_{i=1}^n}{\text{maximize}} \quad \langle \theta, [\beta_1; \dots; \beta_k] \rangle \\
& \text{subject to} \quad t_i \geq 0, \quad \langle x_i, \beta_j \rangle - y_i \leq t_i, \quad \frac{1}{n} \sum_{i=1}^n t_i \leq \eta, \\
& \quad \quad \quad \forall i \in [n], \forall j \in [k].
\end{aligned} \tag{II.14}$$

To compute the computational costs for (II.14), we further reformulate it into the form of a linear program $\min_{\mathbf{A}\mathbf{s}=\mathbf{b}, \mathbf{s} \geq \mathbf{0}} \langle \mathbf{c}, \mathbf{s} \rangle$ by introducing an additional $nk + 1$ auxiliary variables to convert the second and third inequality constraints into equality constraints. Then, we have $nk + 1$ equality constraints and $2pk + nk + n + 1$ variables. By [19], finding its exact solution costs $\tilde{O}(((n+p)k)^c)$ with $c \approx 2.38$. In contrast, with finitely many operations, AM can find only an approximate solution. The per-iteration cost of AM is $O(nkp^2)$. In the noiseless case, due to the linear convergence of AM, the total cost to obtain an ϵ -accurate solution is $O(nkp^2 \log(1/\epsilon))$.

In a special case where the observations are almost equally distributed over the linear components of the max-linear model, we have $\pi_{\min} \approx \pi_{\max} \approx 1/k$. Consequently, the sample complexity for both estimators is $\tilde{O}(pk^4)$. Thus, the computational costs for AR and AM become $\tilde{O}(p^{2.38}k^{12})$ and $\tilde{O}(p^3k^5)$, respectively. When p is much larger than k (specifically, $p > k^{14}$, the computational cost of AR is significantly lower than that of AM. However, in the opposite scenario, AM is more cost-effective. We summarize the comparison with respect to the computational cost, sample complexity and model assumption in Section II.

III. NUMERICAL RESULTS

We present a set of Monte Carlo simulations to evaluate the performance of the estimator by anchored regression numerically. The experiments were designed to illustrate the following perspectives on the estimation performance: i) The empirical phase transition on exact recovery without noise corroborates Theorem 1; ii) Further iterations of AR with updated anchor vectors significantly reduce the estimation error; iii) AR provides a competitive empirical performance with additive Gaussian noise to AM; iv) AR provides a stable estimation in the presence of sparse noise, where the performance of AM significantly deteriorates. We implement AR by the linear program given in (II.14). Since (II.14) is in the standard form of a linear program, it can be solved efficiently by readily available software such as CPLEX and Gurobi [20]. AR is compared to the version of AM by Ghosh et al. [1]. For a fair comparison, we let both methods start from the same initial estimate, which will be specified later.

In the Monte Carlo simulations, the regressors x_1, \dots, x_n are generated as independent copies of a random vector following $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$, as assumed in Theorem 1. For each run, the estimation error is measured up to permutation

ambiguity, that is, the error is calculated as the minimum of $\sum_{j=1}^k \|\hat{\beta}_{\pi(j)} - \beta_{\star,j}\|_2 / \sum_{j=1}^k \|\beta_{\star,j}\|_2$ over all possible permutation π over segment indices, where $(\beta_{\star,j})_{j=1}^k$ and $(\hat{\beta}_j)_{j=1}^k$ denote the ground-truth parameters and their estimates, respectively.

Since both AR and AM algorithms operate provided suitably initialized parameter, it is crucial to obtain an initial estimate, which lands near the ground-truth parameter. To this end, throughout the simulations, we apply the heuristic known as the *AM with repeated random initialization* in [12], summarized as follows: One repeats the following procedure for $q \in [m]$: i) Randomly generate parameters $\beta_{q,1}^r, \dots, \beta_{q,k}^r \in \mathbb{R}^p$. ii) Run the AM algorithm from given initial estimates for I_{init} iterations and obtain estimates $\beta_{q,1}^o, \dots, \beta_{q,k}^o$. Then choose the set of parameters $\beta_{q',1}^o, \dots, \beta_{q',k}^o$, which achieves the least empirical loss in (I.5), i.e.

$$q' = \underset{q \in [m]}{\text{argmin}} \sum_{i=1}^n \left(\max_{1 \leq j \leq k} \langle x_i, \beta_{q,j}^o \rangle - y_i \right)^2.$$

Throughout all simulations, the initialization parameters are set to $m = 200$ and $I_{\text{init}} = 10$. Moreover, the maximum iteration number for the AM algorithm, denoted by I_{AM} , is set to $I_{\text{AM}} = 120$.

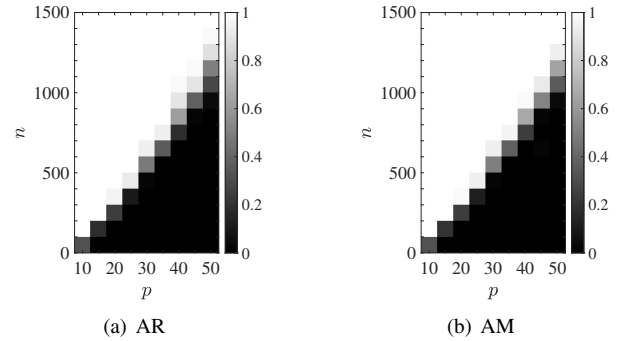


Fig. 1: Phase transition of recovery rate for varying n and p in the noiseless case ($k = 5$).

Figures 1 and 2 illustrate the empirical phase transition of exact recovery in the noise-free scenario as a function of the sample size n per varying dimension parameters, which

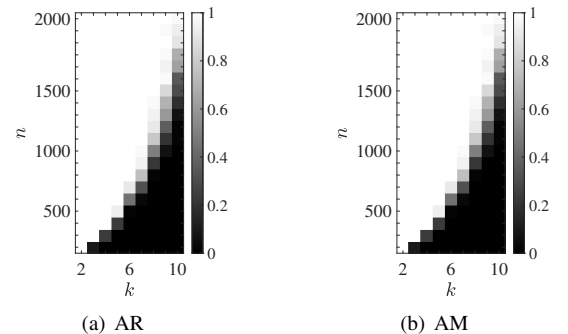


Fig. 2: Phase transition of recovery rate for varying n and k in the noiseless case ($p = 20$).

²The spectral initialization is not included in this comparison. To incorporate the initialization into the analysis, it is necessary to modify the noise model from an adversarial noise model to a gross error model as discussed in Section II-B.

TABLE I: Comparison of local convergence of AR and AM.²

Method	Cost for ϵ -accuracy	Cost for an ideal instance	Sample complexity	Covariate model	Noise model
AR	$\tilde{O}\left(((n+p)k)^{2.38}\right)$	$\tilde{O}(p^{2.38}k^{12})$	$\tilde{O}\left(\pi_{\min}^{-4}p\right)$	Gaussian	Adversarial
AM [1]	$O(nkp^2 \log(1/\epsilon))$	$\tilde{O}(p^3k^5)$	$O(\pi_{\min}^{-3}kp)$	Gaussian	Sub-Gaussian

Algorithm 1: Iterative Anchored Regression (IAR)

-
- 1: **Input:** data $\{\mathbf{x}_i, y_i\}_{i=1}^n$; initialized parameter $\tilde{\beta} \in \mathbb{R}^{kp}$; fidelity upper bound η ; max. number of iterations I_{IAR}
 - 2: **Output:** estimated parameter $\hat{\beta} \in \mathbb{R}^{kp}$
 - 3: **for** $i = 1$ **to** I_{IAR} **do**
 - 4: Compute anchor vector θ from $\tilde{\beta}$ by (II.1)
 - 5: Estimate $\hat{\beta}$ by anchored regression in (II.14)
 - 6: $\tilde{\beta} \leftarrow \hat{\beta}$
 - 7: **end for**
-

are the ambient dimension p and the number of segments k . The reconstruction is determined as success if the normalized estimation error is below 10^{-5} . The recovery rate is calculated as the ratio of success out of 50 trials. In this simulation, we assume that $k \leq p$. To satisfy the “well-balance partition” condition, we generate the ground-truth parameter vectors so that they are mutually orthogonal one another.

Figure 1 shows that for both AR and AM, the phase transition occurs when n grows linearly with p while k is fixed to 5. This observation qualitatively coincides with the sample complexity by Theorem 1. A complementary view is provided by Figure 2 for varying k while p is fixed to 20. Here, the phase transition occurs when n is proportional to k^t for some constant $t \in (1, 2)$. The order of this polynomial is smaller than the corresponding result by Theorem 1, where n is proportional to k^4 . A similar gap between theoretical sufficient condition and empirical phase transition was observed for AM in the noise-free setting [21, Appendix L]. Overall, as shown in these figures, AR and AM provide similar empirical performance in the noiseless scenario.

In practice, observations are often corrupted with noise. Next, we study the estimation under two noise models. In these experiments, the ground-truth parameter vectors are i.i.d Normal($\mathbf{0}, \mathbf{I}_{kp}$). Furthermore, to deduce statistical performance, the median of the estimation error in 50 trials is observed.

First, we consider the i.i.d. Gaussian noise model, i.e. $y_i = f_i(\beta_*) + \epsilon_i$, where $\{\epsilon_i\}_{i=1}^n$ are i.i.d following Normal($0, \sigma^2$). To track the change of the estimation performance as a function of the noise strength, the dimension parameters are fixed as $p = 30$ and $k = 6$. AM has shown to be consistent, with an error rate that vanishes as n grows[1]. Its empirical estimation error decays similarly in the experiment. However, we observe that AR has a larger estimation error compared to AM, which remains nontrivial even for large n . We conjecture that this bias term is due to the regularizer with an imperfect anchor vector. In fact, as the anchor vector is obtained from a more accurate initial estimate, the result estimation error decays accordingly. Motivated by this observation, we consider a modification of AR with further iterative refinements, which we

call the *iterative anchored regression* (IAR). The first iteration of IAR is equivalent to AR, but in the subsequent iterations, the anchor vector is refined by using the estimate from the previous iteration. The entire IAR algorithm is summarized in Algorithm 1. The number of iterations in IAR is set to $I_{\text{IAR}} = 40$. Figure 3 shows that with more iterations the performance of iterative AR becomes as good as that of AM. Moreover, for small n (e.g. $n \leq 1,000$), IAR provides a smaller estimation error than AM. Moreover, we also study the sensitivity to the choice of the parameter η in (I.4). The need to tune this parameter can be a weakness of AR since AM does not involve any such parameter. As shown in Figure 4, the estimation error by AR does not critically depend on η . In this experiment, we vary η around η_* that achieves the equality in (II.6) with $\pm 50\%$ margin. Within this range, the estimation error remains small. Also, note that the minimum estimation error is achieved when η is slightly smaller than η_* . It still remains to set the value of η within this range. Since the observations are corrupted with i.i.d. noise in this experiment, we applied a 5-fold cross-validation to estimate the validation error. Figure 4 suggests that choosing an η value that yields the smallest prediction error will likely result in the smallest estimation error.

Next, we study the empirical performance of the estimators under a gross error model. In Section II, we have shown that the theoretical analysis of AR combined with the initialization by Ghosh et al. [1] applies to this model. Specifically, each observation is corrupted by a sparse noise according to the multiplicative Bernoulli model with probability φ , that is, $\mathbb{P}\{y_i = -f_i(\beta_*)\} = \varphi$ and $\mathbb{P}\{y_i = f_i(\beta_*)\} = 1 - \varphi$ for $i \in [n]$. The multiplicative Bernoulli noise model has a similarity with the Massart noise [22, Definition 1.1]. Similar to the previous experiment, we compare AR to IAR and AM. Furthermore, we also study the performance of a variation of AM in which the least squares update is substituted by LAD. It will be denoted by AM-LAD. Figure 5 illustrates the estimation error in this setting where $p = 30$ and $k = 6$. Unlike the case of Gaussian noise, AR outperforms AM in the presence of multiplicative Bernoulli noise. Furthermore, IAR and AM-LAD achieve exact recovery over the range of φ in this experiment.

IV. PROOF OF THEOREM 1

We prove Theorem 1 in two steps. First, in the following proposition, we present a sufficient condition for stable estimation by convex program in (I.4). Then we derive an upper bound on ϱ in the proposition, which provides the sample complexity condition along with the corresponding error bound in Theorem 1.

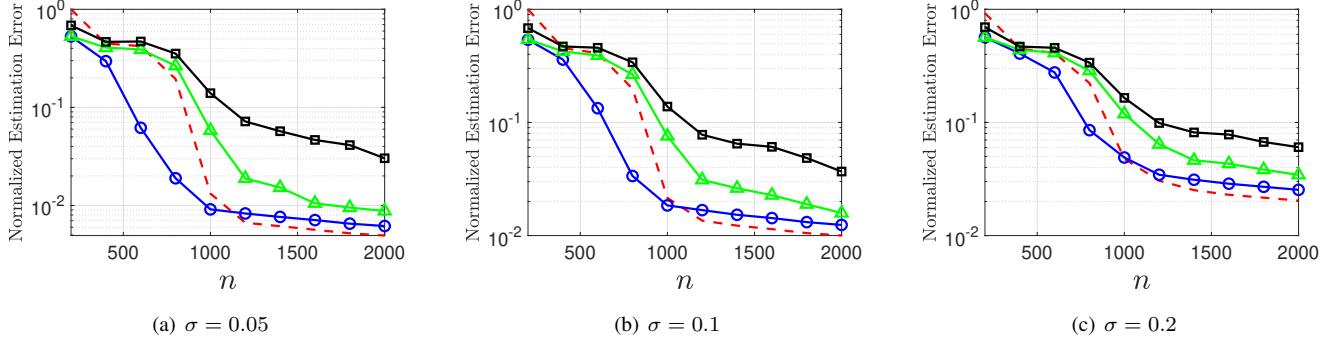


Fig. 3: Estimation error versus the number of observations n under Gaussian noise of variance σ^2 ($k = 6$ and $p = 30$): repeated random initialization (black line with square markers), AR (green line with triangle markers), iterative AR (blue line and circle markers), and AM (red dashed line). All methods start from the repeated random initialization.

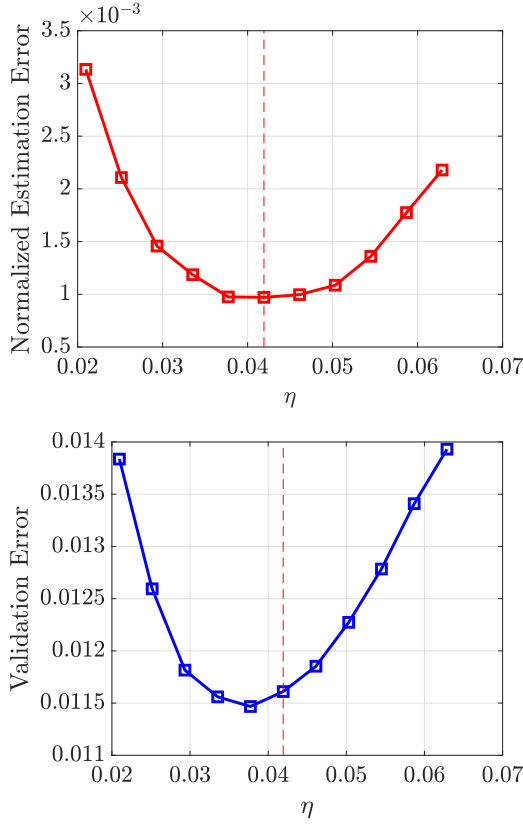


Fig. 4: Estimation error and validation error via cross-validation by AR for varying η ($k = 3, p = 30$, and $n = 1, 500$): The dotted vertical line indicates the location of η_* that achieves the equality in (II.6).

Proposition 1: Under the hypothesis of Theorem 1, suppose that $\hat{\beta}$ satisfies

$$\varrho := \inf_{\substack{j \in [k] \\ \mathbf{w} \in \mathbb{S}^{p-1}}} \mathbb{E} \mathbb{1}_{\mathcal{C}_j}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| - \sup_{\substack{j \in [k] \\ \mathbf{w} \in \mathbb{S}^{p-1}}} \mathbb{E} \mathbb{1}_{\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ - \sup_{\substack{j \in [k] \\ \mathbf{w} \in \mathbb{S}^{p-1}}} \mathbb{E} \mathbb{1}_{\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ > 0. \quad (\text{IV.1})$$

Then there exists an absolute constant $c > 0$ such that if

$$n \geq c \varrho^{-2} (4p \log^3 p \log^5 k + 4 \log(\delta^{-1}) \log k), \quad (\text{IV.2})$$

then the solution $\hat{\beta}$ to the optimization problem in (I.4) obeys

$$\sum_{j=1}^k \|\beta_{\star, j} - \hat{\beta}_j\|_2 \leq \frac{2}{\varrho n} \sum_{i=1}^n |w_i| \quad (\text{IV.3})$$

with probability $1 - \delta$.

Proof: We first show that there exists a constant $c > 0$ such that the condition in (II.3) implies $\zeta > 0$. Hence, we consider

$$\underbrace{\min_{j \in [k]} \sqrt{\frac{\pi}{32}} \mathbb{P}^2\{\mathbf{g} \in \mathcal{C}_j\}}_{(i)} - 2 \underbrace{\max_{j \in [k]} \sqrt{\mathbb{P}\{\mathbf{g} \in \tilde{\mathcal{C}}_j \Delta \mathcal{C}_j\}}}_{(ii)} > 0. \quad (\text{IV.4})$$

It follows from the definition of π_{\min} that (i) in (IV.4) is bounded from below as

$$(i) \geq \sqrt{\frac{\pi}{32}} \pi_{\min}^2. \quad (\text{IV.5})$$

It only remains to find an appropriate upper bound on (ii). Since $\{\mathcal{C}_j\}_{j=1}^k$ consists of disjoint sets (except their boundaries corresponding to sets of measure zero), for a fixed $j \in [k]$, the symmetric difference between $\tilde{\mathcal{C}}_j$ and \mathcal{C}_j is written as

$$\tilde{\mathcal{C}}_j \Delta \mathcal{C}_j = \left(\cup_{j' \neq j} \tilde{\mathcal{C}}_j \cap \mathcal{C}_{j'} \right) \cup \left(\cup_{j' \neq j} \mathcal{C}_j \cap \tilde{\mathcal{C}}_{j'} \right).$$

Therefore, we obtain

$$(ii) \leq 2\sqrt{2k} \max_{j \in [k]} \max_{j' \in [k] \setminus \{j\}} \sqrt{\mathbb{P}(\mathbf{g} \in \tilde{\mathcal{C}}_j \cap \mathcal{C}_{j'})}. \quad (\text{IV.6})$$

Moreover, since

$$\begin{aligned} \mathbf{g} \in \tilde{\mathcal{C}}_j \cap \mathcal{C}_{j'} &\implies \mathbf{g}^\top \tilde{\beta}_j \geq \mathbf{x}_i^\top \tilde{\beta}_{j'}, \quad \mathbf{x}_i^\top \beta_{\star, j'} \geq \mathbf{x}_i^\top \beta_{\star, j} \\ &\implies \mathbf{g}^\top (\tilde{\beta}_j - \tilde{\beta}_{j'}) \geq 0, \quad \mathbf{g}^\top (\beta_{\star, j} - \beta_{\star, j'}) \leq 0 \\ &\implies \mathbf{g}^\top (\tilde{\beta}_j - \tilde{\beta}_{j'}) \cdot \mathbf{g}^\top (\beta_{\star, j} - \beta_{\star, j'}) \leq 0, \end{aligned} \quad (\text{IV.7})$$

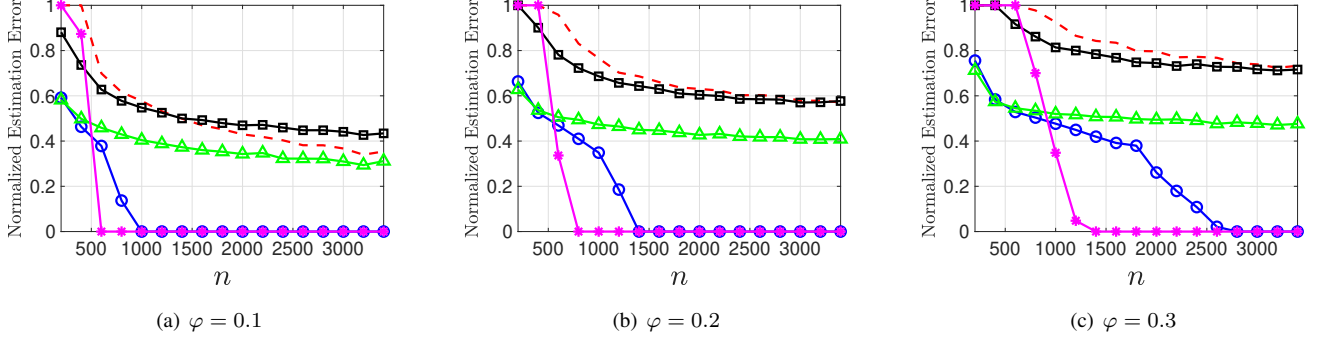


Fig. 5: Estimation error versus the number of observations n under multiplicative Bernoulli noise model with probability φ ($k = 6$ and $p = 30$): repeated random initialization (black line with square markers), AR (green line with triangle markers), IAR (blue line with circle markers), AM (red dashed line), and AM-LAD (magenta line with asterisk markers). All methods start from repeated random initialization.

with [1, Lemma 9], (ii) in (IV.6) is further upper-bounded by

$$\begin{aligned}
 & \text{(ii)} \\
 & \leq 2\sqrt{2k} \cdot \sqrt{\mathbb{P}\left(\mathbf{g}^\top(\tilde{\beta}_j - \tilde{\beta}_{j'}) \cdot \mathbf{g}^\top(\beta_{\star,j} - \beta_{\star,j'}) \leq 0\right)} \\
 & \leq C\sqrt{k} \cdot \max_{j \in [k]} \max_{j' \in [k] \setminus \{j\}} \left(\sqrt{\frac{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{\star,j} - \beta_{\star,j'})\|_2}{\|\beta_{\star,j} - \beta_{\star,j'}\|_2}} \right. \\
 & \quad \left. \cdot \log^{1/4} \left(\frac{2\|\beta_{\star,j} - \beta_{\star,j'}\|_2}{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{\star,j} - \beta_{\star,j'})\|_2} \right) \right), \tag{IV.8}
 \end{aligned}$$

for an absolute constant $C > 0$. Then, by plugging in (IV.5) and (IV.8) to (IV.4), we obtain a sufficient condition for (IV.4) as

$$\begin{aligned}
 & C\sqrt{k} \max_{j \in [k]} \max_{j' \in [k] \setminus \{j\}} \sqrt{\frac{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{\star,j} - \beta_{\star,j'})\|_2}{\|\beta_{\star,j} - \beta_{\star,j'}\|_2}} \cdot \\
 & \log^{1/4} \left(\frac{2\|\beta_{\star,j} - \beta_{\star,j'}\|_2}{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{\star,j} - \beta_{\star,j'})\|_2} \right) < \sqrt{\frac{\pi}{32}} \pi_{\min}^2. \tag{IV.9}
 \end{aligned}$$

For a fixed $j' \in [k] \setminus \{j\}$, let

$$a = \frac{\|(\tilde{\beta}_j - \tilde{\beta}_{j'}) - (\beta_{\star,j} - \beta_{\star,j'})\|_2}{\|\beta_{\star,j} - \beta_{\star,j'}\|_2} \quad \text{and} \quad b = \frac{\pi_{\min}^4}{k}.$$

Since $a, b \in (0, 0.1]$ and $a \leq \frac{b}{2} \log^{-1/2}(1/b)$ imply $a \log^{1/2}(2/a) \leq b$, if one chooses c in (II.3) so that $c < \frac{\pi}{32C^2}$, then (II.3) implies (IV.9) for all distinct $j, j' \in [k]$. In the remainder of the proof, we will assume that (IV.4) holds.

We show that, for a sufficiently large $\rho > 0$, the following three conditions cannot hold simultaneously:

$$\frac{1}{n} \sum_{i=1}^n (f_i(\beta_{\star} + \mathbf{z}) - y_i)_+ \leq \eta, \tag{IV.10}$$

$$\|\mathbf{z}\|_{1,2} > \rho, \tag{IV.11}$$

$$\langle \boldsymbol{\theta}, \mathbf{z} \rangle \geq 0. \tag{IV.12}$$

Therefore, assuming (IV.11) and (IV.12) hold, it suffices to show

$$\mathcal{L}(\mathbf{z}) := \frac{1}{n} \sum_{i=1}^n (f_i(\beta_{\star} + \mathbf{z}) - y_i)_+ > \eta. \tag{IV.13}$$

To this end, we derive a lower bound on $\mathcal{L}(\mathbf{z})$ as follows:

$$\begin{aligned}
 \mathcal{L}(\mathbf{z}) & \geq \frac{1}{n} \sum_{i=1}^n (f_i(\beta_{\star} + \mathbf{z}) - f_i(\beta_{\star}))_+ - \frac{1}{n} \sum_{i=1}^n (w_i)_+ \\
 & \stackrel{(a)}{\geq} \frac{1}{n} \sum_{i=1}^n (\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle)_+ - \frac{1}{n} \sum_{i=1}^n (w_i)_+ \\
 & = \frac{1}{n} \sum_{i=1}^n \frac{|\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle|}{2} + \frac{1}{n} \sum_{i=1}^n \frac{\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle}{2} \\
 & \quad - \frac{1}{n} \sum_{i=1}^n (w_i)_+ \\
 & = \frac{1}{n} \sum_{i=1}^n \frac{|\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle|}{2} + \frac{1}{n} \sum_{i=1}^n \frac{\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle}{2} \\
 & \quad - \langle \boldsymbol{\theta}, \mathbf{z} \rangle + \langle \boldsymbol{\theta}, \mathbf{z} \rangle - \frac{1}{n} \sum_{i=1}^n (w_i)_+ \\
 & \stackrel{(b)}{=} \langle \boldsymbol{\theta}, \mathbf{z} \rangle - \frac{1}{n} \sum_{i=1}^n (w_i)_+ + \frac{1}{n} \sum_{i=1}^n \frac{|\langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle|}{2} \\
 & \quad + \frac{1}{n} \sum_{i=1}^n \frac{\langle \nabla f_i(\beta_{\star}) - \nabla f_i(\tilde{\beta}), \mathbf{z} \rangle}{2} \\
 & \stackrel{(c)}{=} \langle \boldsymbol{\theta}, \mathbf{z} \rangle - \frac{1}{n} \sum_{i=1}^n (w_i)_+ + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \frac{\mathbb{1}_{C_j}(\mathbf{x}_i) |\langle \mathbf{x}_i, \mathbf{z}_j \rangle|}{2} \\
 & \quad + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \frac{\{\mathbb{1}_{C_j}(\mathbf{x}_i) - \mathbb{1}_{\tilde{C}_j}(\mathbf{x}_i)\} \langle \mathbf{x}_i, \mathbf{z}_j \rangle}{2}, \tag{IV.14}
 \end{aligned}$$

where (a) holds by the convexity of f_i , which implies

$$f_i(\beta_{\star} + \mathbf{z}) \geq f_i(\beta_{\star}) + \langle \nabla f_i(\beta_{\star}), \mathbf{z} \rangle,$$

(b) follows from (II.1), and (c) is obtained by calculating $\nabla f_i(\beta)$ at $\beta = \beta_{\star}$ and $\beta = \tilde{\beta}$. We further proceed by obtaining lower bounds on the last two terms in (IV.14) by the following lemmas, which are proved in Appendices D and E.

Lemma 3: Let $(V_{\mathbf{z}})_{\mathbf{z} \in \mathbb{R}^{kp}}$ be a random process defined by

$$V_{\mathbf{z}} := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{C_j}(\mathbf{x}_i) |\langle \mathbf{x}_i, \mathbf{z}_j \rangle|,$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$. Then, for $\mathbf{g} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_p)$ and any $\delta \in (0, 1)$, there exists an absolute constant $c_1 > 0$ such that

$$\underline{V} := \inf_{\|\mathbf{z}\|_{1,2}=1} V_{\mathbf{z}} \geq \min_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{C_j}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| - c_1 \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2}$$

holds with probability at least $1 - \delta/2$.

Proof: See Appendix D. ■

Lemma 4: Let $(Q_{\mathbf{z}})_{\mathbf{z} \in B_{1,2}}$ be a random process defined by

$$Q_{\mathbf{z}} := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \left\{ \mathbb{1}_{\tilde{C}_j}(\mathbf{x}_i) - \mathbb{1}_{C_j}(\mathbf{x}_i) \right\} \langle \mathbf{x}_i, \mathbf{z}_j \rangle,$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$. Then, for $\mathbf{g} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_p)$ and any $\delta \in (0, 1)$, there exists an absolute constant $c_2 > 0$ such that

$$\begin{aligned} \bar{Q} := \sup_{\|\mathbf{z}\|_{1,2}=1} Q_{\mathbf{z}} &\leq \max_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \\ &\quad + \max_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{C_j \setminus \tilde{C}_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \\ &\quad + c_2 \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \end{aligned}$$

holds with probability at least $1 - \delta/2$.

Proof: See Appendix E. ■

Since $V_{\mathbf{z}}$ and $Q_{\mathbf{z}}$ are homogeneous in \mathbf{z} , we obtain that the third term in the right-hand side of (IV.14) is written as $V_{\mathbf{z}}$ and lower-bounded by

$$\frac{V_{\mathbf{z}}}{2} \geq \frac{\underline{V} \|\mathbf{z}\|_{1,2}}{2}. \quad (\text{IV.15})$$

Similarly, the last term in the right-hand side of (IV.14) is written as $-Q_{\mathbf{z}}$ and lower-bounded by

$$-\frac{Q_{\mathbf{z}}}{2} \geq -\frac{\bar{Q} \|\mathbf{z}\|_{1,2}}{2}. \quad (\text{IV.16})$$

Furthermore, by Lemmas 3 and 4, the condition in (IV.2) implies that

$$\underline{V} - \bar{Q} \geq c_3 \varrho > 0 \quad (\text{IV.17})$$

holds with probability $1 - \delta$ for an absolute constant $c_3 > 0$. Then we choose ρ so that it satisfies

$$\rho = \frac{2}{\underline{V} - \bar{Q}} \cdot \left(\eta + \frac{1}{n} \sum_{i=1}^n (w_i)_+ \right).$$

Next, by plugging in the above estimates to (IV.14), we obtain that, under the event in (IV.17), the conditions in (IV.11) and (IV.12) imply

$$\begin{aligned} \mathcal{L}(\mathbf{z}) &\geq \langle \boldsymbol{\theta}, \mathbf{z} \rangle - \frac{1}{n} \sum_{i=1}^n (w_i)_+ + \frac{(V - \bar{Q}) \|\mathbf{z}\|_{1,2}}{2} \\ &> -\frac{1}{n} \sum_{i=1}^n (w_i)_+ + \frac{(V - \bar{Q}) \rho}{2} \\ &= -\frac{1}{n} \sum_{i=1}^n (w_i)_+ + \frac{1}{n} \sum_{i=1}^n (w_i)_+ + \eta \\ &= \eta. \end{aligned}$$

This lower bound implies (IV.13). Therefore we have shown that the three conditions in (IV.10), (IV.11), and (IV.12) cannot hold simultaneously. It remains to apply the claim to a special case.

Let $\hat{\mathbf{z}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*$. Recall that both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_*$ are feasible for the optimization problem in (I.4). Moreover, since $\hat{\boldsymbol{\beta}}$ is the maximizer, it follows that $\langle \boldsymbol{\theta}, \hat{\boldsymbol{\beta}} \rangle \geq \langle \boldsymbol{\theta}, \boldsymbol{\beta}_* \rangle$, which implies $\langle \boldsymbol{\theta}, \hat{\mathbf{z}} \rangle \geq 0$. Therefore the conditions in (IV.10) and (IV.12) are satisfied with \mathbf{z} substituted by $\hat{\mathbf{z}}$. Since the three conditions cannot be satisfied simultaneously, the condition in (IV.11) cannot hold, i.e. $\hat{\mathbf{z}}$ satisfies

$$\|\hat{\mathbf{z}}\|_{1,2} \leq \rho \leq \frac{2}{\varrho} \left(\eta + \frac{1}{n} \sum_{i=1}^n (w_i)_+ \right). \quad (\text{IV.18})$$

Since the noise vector \mathbf{w} was arbitrary, (IV.18) holds for any \mathbf{w} . Furthermore, since the random processes in Lemma 3 and Lemma 4 do not depend on the noise \mathbf{w} , the conclusion of the theorem applies to an adversarial noise without amplifying the error probability. ■

Next, we use the following lemma to obtain a lower bound on ϱ in (IV.1).

Lemma 5: Let $\mathcal{A} \subset \mathbb{R}^p$ be of finite Gaussian measure and $\mathbf{g} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_p)$. Then we have

$$\inf_{\mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\mathcal{A}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| \geq \sqrt{\frac{\pi}{32}} \mathbb{P}^2\{\mathbf{g} \in \mathcal{A}\}$$

and

$$\sup_{\mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\mathcal{A}}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \leq \sqrt{\mathbb{P}\{\mathbf{g} \in \mathcal{A}\}}.$$

Proof: For an arbitrarily fixed $\epsilon > 0$, let $S_{\epsilon} \subset \mathbb{R}^p$ denote the set defined by

$$S_{\epsilon} := \{\mathbf{x} \in \mathbb{R}^p : |\langle \mathbf{x}, \mathbf{w} \rangle| < \epsilon\}.$$

Then we have

$$\begin{aligned} \mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| &\geq \epsilon \mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) \mathbb{1}_{S_{\epsilon}^c}(\mathbf{g}) \\ &= \epsilon \mathbb{E} (\mathbb{1}_{\mathcal{C}}(\mathbf{g}) - \mathbb{1}_{\mathcal{C}}(\mathbf{g}) \mathbb{1}_{S_{\epsilon}}(\mathbf{g})) \\ &\geq \epsilon \mathbb{E} (\mathbb{1}_{\mathcal{C}}(\mathbf{g}) - \mathbb{1}_{S_{\epsilon}}(\mathbf{g})) \\ &= \epsilon (\mathbb{P}\{\mathbf{g} \in \mathcal{C}\} - \mathbb{P}\{\mathbf{g} \in S_{\epsilon}\}). \end{aligned} \quad (\text{IV.19})$$

Moreover, since $\langle \mathbf{g}, \mathbf{w} \rangle \sim \text{Normal}(0, 1)$, $\mathbb{P}\{\mathbf{g} \in S_{\epsilon}\}$ is upper-bounded by

$$\mathbb{P}\{\mathbf{g} \in S_{\epsilon}\} = \mathbb{P}\{|\langle \mathbf{g}, \mathbf{w} \rangle| < \epsilon\} = \int_{-\epsilon}^{\epsilon} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \leq \epsilon \sqrt{\frac{2}{\pi}}. \quad (\text{IV.20})$$

By plugging in (IV.20) to (IV.19), we obtain

$$\mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| \geq \epsilon \left(\mathbb{P}\{\mathbf{g} \in \mathcal{C}\} - \epsilon \sqrt{\frac{2}{\pi}} \right). \quad (\text{IV.21})$$

Since the parameter $\epsilon > 0$ was arbitrary, one can we maximize the right-hand side of (IV.21) with respect to ϵ to obtain the tightest lower bound. Note that the objective is a concave quadratic function and the maximum is attained at $\epsilon = \sqrt{\pi/8} \mathbb{P}\{\mathbf{g} \in \mathcal{C}\}$. This provides the lower bound in the first assertion. Next, by the Cauchy-Schwarz inequality, we obtain the upper bound in the second assertion as follows:

$$\begin{aligned} \mathbb{E} \mathbb{1}_{\mathcal{A}}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ &\leq \sqrt{\mathbb{E} (\mathbb{1}_{\mathcal{A}}(\mathbf{g}))^2} \sqrt{\mathbb{E} \langle \mathbf{g}, \mathbf{w} \rangle_+^2} \\ &= \sqrt{\mathbb{E} \mathbb{1}_{\mathcal{A}}(\mathbf{g})} \sqrt{\frac{\mathbb{E} \langle \mathbf{g}, \mathbf{w} \rangle^2}{2}} \\ &= \sqrt{\frac{\mathbb{P}\{\mathbf{g} \in \mathcal{A}\}}{2}}. \end{aligned}$$

Finally, by applying Lemma 5 to each of the expectation terms in ϱ , we obtain a lower bound on ϱ given by

$$\begin{aligned} \varrho &\geq \min_{j \in [k]} \sqrt{\frac{\pi}{32}} \mathbb{P}^2 \{\mathbf{g} \in \mathcal{C}_j\} - \max_{j \in [k]} \sqrt{\mathbb{P}\{\mathbf{g} \in \mathcal{C}_j \setminus \tilde{\mathcal{C}}_j\}} \\ &\quad - \max_{j \in [k]} \sqrt{\mathbb{P}\{\mathbf{g} \in \tilde{\mathcal{C}}_j \setminus \mathcal{C}_j\}} \\ &\geq \min_{j \in [k]} \sqrt{\frac{\pi}{32}} \mathbb{P}^2 \{\mathbf{g} \in \mathcal{C}_j\} - 2 \max_{j \in [k]} \sqrt{\mathbb{P}\{\mathbf{g} \in \mathcal{C}_j \triangle \tilde{\mathcal{C}}_j\}}, \end{aligned} \quad (\text{IV.22})$$

where the second inequality holds since $\tilde{\mathcal{C}}_j \triangle \mathcal{C}_j = (\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j) \cup (\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j)$ for all $j \in [k]$. This implies that (II.4) is a sufficient condition for (IV.2). Moreover, substituting ϱ in (IV.3) by the lower bound in (IV.22) provides (II.5). This completes the proof of Theorem 1.

A. Tightness of the lower bound on ϱ

In (IV.22), we obtain a lower bound on ϱ by Lemma 5. We show through the following example that the lower bound is tight in terms of its dependence on $\mathbb{P}\{\mathbf{g} \in \mathcal{C}_j\}$ for $j \in [k]$.

Example 1: Let $p = 2$. Then $\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j$ and $\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j$ are Lorentz cones. Let $\theta_{\mathcal{C}_j}$, $\theta_{\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j}$ and $\theta_{\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j}$ denote the angular width of \mathcal{C}_j , $\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j$, and $\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j$ respectively. Furthermore, we assume that

$$\min_{j \in [k]} \mathbb{P}\{\mathbf{g} \in \mathcal{C}_j\} \geq \max_{j \in [k]} \mathbb{P}\{\mathbf{g} \in \mathcal{C}_j \triangle \tilde{\mathcal{C}}_j\}. \quad (\text{IV.23})$$

In this case, the parameter ϱ in Proposition 1 is expressed as

$$\begin{aligned} \varrho &= \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(1)} \left[\min_{j \in [k]} \frac{2}{\pi} \sin^2 \left(\frac{\theta_{\mathcal{C}_j}}{4} \right) - \max_{j \in [k]} \frac{1}{\pi} \sin \left(\frac{\theta_{\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j}}{2} \right) \right. \\ &\quad \left. - \max_{j \in [k]} \frac{1}{\pi} \sin \left(\frac{\theta_{\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j}}{2} \right) \right]. \end{aligned} \quad (\text{IV.24})$$

When $\theta_{\mathcal{C}}$ is small enough, $\sin(\theta_{\mathcal{C}}) \approx \theta_{\mathcal{C}}$ holds by the Taylor series approximation. Hence, there exists absolute constants $c_1 > 0$ and $c_2 > 0$ such that

$$\varrho \geq c_1 \min_{j \in [k]} \mathbb{P}^2 \{\mathbf{g} \in \mathcal{C}_j\} - c_2 \max_{j \in [k]} \mathbb{P}\{\mathbf{g} \in \tilde{\mathcal{C}}_j \triangle \mathcal{C}_j\}.$$

This example shows that ζ in Theorem 1 is tight in the sense that the dominating term in both ϱ and ζ is proportional to the squared probability measure of the smallest \mathcal{C}_j .

Let $\theta_{\mathcal{C}_j}$ denote the angular width of \mathcal{C}_j . Without loss of generality, we may assume that $\min_{j \in [k]} \theta_{\mathcal{C}_j} \leq \pi$. Furthermore, the assumption in (IV.23) implies that the angular width of $\mathcal{C}_j \triangle \tilde{\mathcal{C}}_j$ is at most π for all $j \in [k]$. Therefore, the identity in (IV.24) is obtained by applying the following lemma, proved in Appendix B, to the infimum/supremum of expectation terms in (IV.1).

Lemma 6: Let \mathcal{C} be a polyhedral cone in \mathbb{R}^2 and $\mathbf{g} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_2)$. Suppose that the angular width of \mathcal{C} , denoted by $\theta_{\mathcal{C}}$ satisfies $0 \leq \theta_{\mathcal{C}} \leq \pi$. Then we have

$$\inf_{\mathbf{w} \in \mathbb{S}^1} \mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| = \frac{2\sqrt{2}\Gamma(3/2)}{\pi\Gamma(2)} \sin^2 \left(\frac{\theta_{\mathcal{C}}}{4} \right)$$

and

$$\sup_{\mathbf{w} \in \mathbb{S}^1} \mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ = \frac{\sqrt{2}\Gamma(3/2)}{\pi\Gamma(2)} \sin \left(\frac{\theta_{\mathcal{C}}}{2} \right).$$

Proof: See Appendix B. ■

V. DISCUSSION

As discussed in Section III, the proposed convex estimator provides a comparable error bound relative to an oracle estimator in the adversarial noise case. However, it does not provide a consistent estimator with random noise. This inconsistency arises due to the maximization of the correlation with the anchor vector $\boldsymbol{\theta}$. Since the direction of the anchor vector does not coincide with the ground truth, the convex estimator introduces a bias. As a way to mitigate the bias in the convex estimator, we propose the iterative anchored regression that recursively refines the anchor vector to better align its direction with that of the ground truth. We have demonstrated that the iterative anchored regression empirically provides an exact recovery of the ground-truth parameters in the presence of outliers. Hence, it would be fruitful to pursue the theoretical analysis of the iterative anchored regression, particularly in terms of its behavior in the presence of outliers and random noise. Each iteration solves a linear program, which costs $\tilde{O}((n+p)k)^c$ with $c \approx 2.38$ as discussed in Section II-C. Therefore, the per-iteration cost of the iterative anchored regression might be higher than that of the alternating minimization, which is $O(nkp^2)$. To further alleviate the computational cost of the iterative version, one might consider warm-start strategies in interior-point methods for linear programming (e.g. [23]).

APPENDIX

A. Proof of Lemma 1

For brevity, we introduce the shorthand notations

$$\mathbf{A}_j = \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^T, \quad \text{and} \quad \mathbf{b}_j = \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) y_i \mathbf{x}_i.$$

Then, since each \mathcal{C}_j is given by the intersection of $(k-1)$ half-planes in \mathbb{R}^p , by [24, Theorem 2], it holds with probability at least $1 - \delta/3$ that

$$\sup_{j \in [k]} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) - \mathbb{P}(\mathbf{g} \in \mathcal{C}_j) \right| \leq \frac{C_1 \sqrt{\log(3/\delta) + kp \log(n/p)}}{n}, \quad (\text{A.1})$$

which implies

$$c_2 n \pi_{\min} \leq \sum_{i=1}^n \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) \leq C_3 n \pi_{\max}, \quad \forall j \in [k]. \quad (\text{A.2})$$

Moreover, by [25, Theorem 5.7], with probability at least $1 - \delta/3$, we have

$$\sup_{\mathcal{I}: |\mathcal{I}| \leq \alpha n} \lambda_{\max} \left(\sum_{i \in \mathcal{I}} \mathbf{x}_i \mathbf{x}_i^\top \right) \leq C_4 \sqrt{\alpha n}$$

provided

$$n \geq \max \left(p, \frac{\log(3/\delta)}{\alpha} \right). \quad (\text{A.3})$$

We also use the following claim: If

$$n \geq C_5 \beta^{-2} \max(p \log(n/p), \log(3/\delta)), \quad (\text{A.4})$$

then it holds with probability $1 - \delta/3$ that

$$\inf_{\mathcal{I} \subset [n]: |\mathcal{I}| \geq \beta n} \lambda_{\min} \left(\sum_{i \in \mathcal{I}} \mathbf{x}_i \mathbf{x}_i^\top \right) \geq c_6 n \beta^3. \quad (\text{A.5})$$

Proof of Claim: For an arbitrarily fixed $T > 0$, we have

$$\frac{1}{n} \sum_{i \in \mathcal{I}} \langle \mathbf{x}_i, \mathbf{v} \rangle^2 \geq \frac{\beta T}{2}, \quad \forall \mathcal{I} \subset [n]: |\mathcal{I}| \geq \beta n \quad (\text{A.6})$$

provided

$$N(\mathbf{v}) := \sum_{i=1}^n \mathbb{1}_{\{\langle \mathbf{x}_i, \mathbf{v} \rangle^2 > T\}}(\mathbf{x}_i) > n - \frac{\beta n}{2}. \quad (\text{A.7})$$

Since $\{\mathbf{x} : \langle \mathbf{x}, \mathbf{v} \rangle^2 > T\}$ consists of two half-spaces in \mathbb{R}^p , by [24, Theorem 2], there exists an absolute constant $C_7 > 0$, for which it holds with probability at least $1 - \delta/3$ that

$$\frac{1}{n} N(\mathbf{v}) \geq \frac{1}{n} \mathbb{E} N(\mathbf{v}) - C_7 \sqrt{\frac{p \log(n/p) + \log(3/\delta)}{n}}, \quad \forall \mathbf{v} \in \mathbb{S}^{p-1}. \quad (\text{A.8})$$

Moreover, due to [21, Lemma 15], we have

$$\frac{1}{n} \mathbb{E} N(\mathbf{v}) = \mathbb{P}(|\langle \mathbf{x}, \mathbf{v} \rangle|^2 > T) \geq 1 - \sqrt{eT}. \quad (\text{A.9})$$

Plugging in (A.9) into (A.8) yields

$$\frac{1}{n} N(\mathbf{v}) \geq 1 - \sqrt{eT} - C_7 \sqrt{\frac{p \log(n/p) + \log(3/\delta)}{n}}, \quad \forall \mathbf{v} \in \mathbb{S}^{p-1}.$$

Then (A.7) is satisfied for all $\mathbf{v} \in \mathbb{S}^p$ by $T = \frac{\beta^2}{16e}$ and $C_5 = (4C_7)^2$. ■

Since (II.7) implies (A.3) and (A.4), combining the above results provides that

$$c_8 n \pi_{\min}^3 \leq \lambda_{\min}(\mathbf{A}_j) \leq \lambda_{\max}(\mathbf{A}_j) \leq C_9 n \sqrt{\pi_{\max}}, \quad \forall j \in [k],$$

holds with probability $1 - \delta$. Then the least squares solution in (I.7) is written as $\hat{\beta}_j = \mathbf{A}_j^{-1} \mathbf{b}_j$ and satisfies

$$\begin{aligned} \|\beta_{\star, j} - \hat{\beta}_j\|_2 &\geq \frac{\lambda_{\min}^{1/2}(\mathbf{A}_j)}{\lambda_{\max}(\mathbf{A}_j)} \left\| (w_i)_{i: \mathbf{x}_i \in \mathcal{C}_j} \right\|_2 \\ &\geq \frac{c_{10} \pi_{\min}^{3/2}}{\sqrt{n \pi_{\max}}} \left\| (w_i)_{i: \mathbf{x}_i \in \mathcal{C}_j} \right\|_2 \geq \frac{c_{11} \pi_{\min}^{3/2}}{\pi_{\max}} \left\| (w_i)_{i: \mathbf{x}_i \in \mathcal{C}_j} \right\|_{\infty}. \end{aligned} \quad (\text{A.10})$$

Then taking a sum over $j \in [k]$ and maximizing over \mathbf{w} satisfying $\|\mathbf{w}\|_{\infty} \leq \eta'$, we obtain

$$\sup_{\|\mathbf{w}\|_{\infty} \leq \eta'} \sum_{j=1}^k \|\beta_{\star, j} - \hat{\beta}_j\|_2 \geq \frac{c_{12} \pi_{\min}^{3/2} \eta'}{\pi_{\max}}.$$

This completes the proof.

B. Proof of Lemma 6

We first prove the first assertion. Since \mathcal{C} is a cone, it follows that $\mathbf{g} \in \mathcal{C}$ if and only $\mathbf{g}/\|\mathbf{g}\|_2 \in \mathcal{C}$. Moreover, Bayes' rule implies

$$\mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| = \mathbb{P}\{\mathbf{g} \in \mathcal{C}\} \mathbb{E}[|\langle \mathbf{g}, \mathbf{w} \rangle| | \mathbf{g} \in \mathcal{C}].$$

Therefore we have

$$\begin{aligned} &\inf_{\mathbf{w} \in \mathbb{S}^1} \mathbb{E} \mathbb{1}_{\mathcal{C}}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle| \\ &= \inf_{\mathbf{w} \in \mathbb{S}^1} \mathbb{P}\{\mathbf{g} \in \mathcal{C}\} \mathbb{E} \left[\|\mathbf{g}\|_2 \left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \mid \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \\ &\stackrel{(a)}{=} \inf_{\mathbf{w} \in \mathbb{S}^1} \mathbb{P}\{\mathbf{g} \in \mathcal{C}\} \mathbb{E} [\|\mathbf{g}\|_2] \mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \mid \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \\ &\stackrel{(b)}{=} \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(2)} \inf_{\mathbf{w} \in \mathbb{S}^1} \frac{\theta_{\mathcal{C}}}{2\pi} \mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \mid \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right], \end{aligned} \quad (\text{A.11})$$

where (a) holds since $\|\mathbf{g}\|_2$ and $\mathbf{g}/\|\mathbf{g}\|_2$ are independent and (b) follows from $\mathbb{E}\|\mathbf{g}\|_2 = \sqrt{2}\Gamma(3/2)/\Gamma(2)$ and

$$\mathbb{P}\{\mathbf{g} \in \mathcal{C}\} = \mathbb{P}\left\{ \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right\} = \frac{\theta_{\mathcal{C}}}{2\pi}.$$

Then it remains to compute the expectation in (A.11). Below we show that

$$\inf_{\mathbf{w} \in \mathbb{S}^2} \mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \mid \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] = \frac{4}{\theta_{\mathcal{C}}} \sin^2 \left(\frac{\theta_{\mathcal{C}}}{2} \right) \quad (\text{A.12})$$

and

$$\sup_{\mathbf{w} \in \mathbb{S}^2} \mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \mid \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] = \frac{2}{\theta_{\mathcal{C}}} \sin \left(\frac{\theta_{\mathcal{C}}}{2} \right). \quad (\text{A.13})$$

Let $\mathcal{T} = \{\mathbf{a}, \mathbf{b}\} \subset \mathbb{S}^1$ satisfy that \mathcal{C} is the conic hull of \mathcal{T} . Then let \mathbf{z} be the unit vector obtained by normalizing $(\mathbf{a} + \mathbf{b})/2$. Then we have $\angle(\mathbf{a}, \mathbf{z}) = \theta_{\mathcal{C}}/2$ and $\angle(\mathbf{b}, \mathbf{z}) = \theta_{\mathcal{C}}/2$. Let $\phi: \mathbb{S}^1 \rightarrow \mathbb{R}$ be defined by $\phi(\mathbf{w}) := \angle(\mathbf{z}, \mathbf{w})$. Since the conditional expectation applies to $|\langle \mathbf{g}/\|\mathbf{g}\|_2, \mathbf{w} \rangle|$, which is invariant under the global sign change in \mathbf{w} , it suffices to consider \mathbf{w} that satisfies $0 \leq \phi(\mathbf{w}) \leq \pi$. Since $\mathbf{g}/\|\mathbf{g}\|_2$ is

uniformly distributed on the unit sphere, the expectation term in (A.12) is written as

$$\mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] = \frac{1}{\theta_C} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} |\cos \theta| d\theta. \quad (\text{A.14})$$

It follows from the assumption on the range of θ_C and $\phi(\mathbf{w})$ that $-\pi/2 \leq \phi(\mathbf{w}) - \theta_C/2 \leq \pi$ and $0 \leq \phi(\mathbf{w}) + \theta_C/2 \leq 3\pi/2$. We proceed by separately considering the complementary cases for $(\theta_C, \phi(\mathbf{w}))$ given below.

Case 1: Suppose that

$$-\frac{\pi}{2} \leq \phi(\mathbf{w}) - \frac{\theta_C}{2} < \phi(\mathbf{w}) + \frac{\theta_C}{2} \leq \frac{\pi}{2}. \quad (\text{A.15})$$

Then $\phi(\mathbf{w})$ is constrained by

$$0 \leq \phi(\mathbf{w}) \leq \pi/2 - \theta_C/2. \quad (\text{A.16})$$

Furthermore, the integral in (A.14) is rewritten as

$$\begin{aligned} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} |\cos \theta| d\theta &= \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} \cos \theta d\theta \\ &= \sin \left(\phi(\mathbf{w}) + \frac{\theta_C}{2} \right) - \sin \left(\phi(\mathbf{w}) - \frac{\theta_C}{2} \right) \sup_{\mathbf{w} \in \mathbb{S}^1} \mathbb{E} 1_{\mathcal{C}_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \\ &= 2 \cos(\phi(\mathbf{w})) \sin \left(\frac{\theta_C}{2} \right). \end{aligned} \quad (\text{A.17})$$

Since $\sin(\theta_C/2) \geq 0$, the expression in (A.17) monotonically decreases in $\phi(\mathbf{w})$ for the interval given in (A.16). Thus the maximum (resp. minimum) is attained as $2 \sin(\theta_C/2)$ at $\phi(\mathbf{w}) = 0$ (resp. $2 \sin^2(\theta_C/2)$ at $\phi(\mathbf{w}) = \pi/2 - \theta_C/2$).

Case 2: Suppose that

$$-\frac{\pi}{2} \leq \phi(\mathbf{w}) - \frac{\theta_C}{2} < \frac{\pi}{2} < \phi(\mathbf{w}) + \frac{\theta_C}{2} \leq \frac{3\pi}{2}. \quad (\text{A.18})$$

Then $\phi(\mathbf{w})$ satisfies

$$\frac{\pi}{2} - \frac{\theta_C}{2} \leq \phi(\mathbf{w}) \leq \frac{\pi}{2} + \frac{\theta_C}{2} \quad (\text{A.19})$$

and the integral in (A.14) reduces to

$$\begin{aligned} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} |\cos \theta| d\theta &= \int_{\phi(\mathbf{w})-\frac{\theta_C}{2}}^{\frac{\pi}{2}} \cos \theta d\theta - \int_{\frac{\pi}{2}}^{\phi(\mathbf{w})+\frac{\theta_C}{2}} \cos \theta d\theta \\ &= 2 - \sin \left(\phi(\mathbf{w}) - \frac{\theta_C}{2} \right) - \sin \left(\phi(\mathbf{w}) + \frac{\theta_C}{2} \right) \\ &= 2 - 2 \sin(\phi(\mathbf{w})) \cos \left(\frac{\theta_C}{2} \right). \end{aligned} \quad (\text{A.20})$$

Since $\cos(\theta_C/2) \geq 0$ for all $\theta_C \in [0, \pi]$, the maximum (resp. minimum) is attained as $2 \sin^2(\theta_C/2)$ at $\phi(\mathbf{w}) = \pi/2 - \theta_C/2$ (resp. $4 \sin^2(\theta_C/4)$ at $\phi(\mathbf{w}) = \pi/2$).

Case 3: Suppose that

$$\frac{\pi}{2} \leq \phi(\mathbf{w}) - \frac{\theta_C}{2} < \phi(\mathbf{w}) + \frac{\theta_C}{2} \leq \frac{3\pi}{2}. \quad (\text{A.21})$$

Then we have

$$\frac{\pi}{2} + \frac{\theta_C}{2} \leq \phi(\mathbf{w}) \leq \pi \quad (\text{A.22})$$

and

$$\begin{aligned} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} |\cos \theta| d\theta &= \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} (-\cos \theta) d\theta \\ &= \sin \left(\phi(\mathbf{w}) - \frac{\theta_C}{2} \right) - \sin \left(\phi(\mathbf{w}) + \frac{\theta_C}{2} \right) \\ &= -2 \cos \phi(\mathbf{w}) \sin \frac{\theta_C}{2}. \end{aligned} \quad (\text{A.23})$$

The maximum (resp. minimum) of (A.23) is attained as $2 \sin(\theta_C/2)$ at $\phi(\mathbf{w}) = \pi$ (resp. $2 \sin^2(\theta_C/2)$ at $\phi(\mathbf{w}) = \pi/2 + \theta_C/2$).

By combining the results in the above three cases, we obtain (A.12) and (A.13). Then substituting the expectation term in (A.11) by (A.12) provides the first assertion.

Next we prove the second assertion. Similarly to (A.11), we have

$$\begin{aligned} &= \sup_{\mathbf{w} \in \mathbb{S}^1} P \{ \mathbf{g} \in \mathcal{C} \} \mathbb{E} \left[\|\mathbf{g}\|_2 \cdot \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle_+ \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \\ &\stackrel{(a)}{=} \sup_{\mathbf{w} \in \mathbb{S}^1} P \{ \mathbf{g} \in \mathcal{C} \} \mathbb{E} [\|\mathbf{g}\|_2] \mathbb{E} \left[\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle_+ \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \\ &\stackrel{(b)}{=} \frac{\sqrt{2}\Gamma(3/2)}{\Gamma(2)} \sup_{\mathbf{w} \in \mathbb{S}^1} \frac{\theta_C}{2\pi} \mathbb{E} \left[\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle_+ \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right], \end{aligned}$$

where (a) holds since $\|\mathbf{g}\|_2$ and $\mathbf{g}/\|\mathbf{g}\|_2$ are independent, (b) follows from $\mathbb{E} \|\mathbf{g}\|_2 = \sqrt{2}\Gamma(3/2)/\Gamma(2)$, and

$$\mathbb{P} \{ \mathbf{g} \in \mathcal{C} \} = \mathbb{P} \left\{ \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right\} = \frac{\theta_C}{2\pi}.$$

If suffices to show that

$$\max_{\mathbf{w} \in \mathbb{S}^1} \mathbb{E} \left[\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle_+ \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] = \frac{2}{\theta_C} \sin \left(\frac{\theta_C}{2} \right). \quad (\text{A.24})$$

Since $\mathbf{g}/\|\mathbf{g}\|_2$ is uniformly distributed on the unit sphere \mathbb{S}^1 and $u_+ = (u + |u|)/2$ for all $u \in \mathbb{R}$, we have

$$\begin{aligned} &\mathbb{E} \left[\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle_+ \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \\ &= \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} \frac{\cos \theta + |\cos \theta|}{2\theta_C} d\theta \\ &= \frac{1}{2} \left(\frac{1}{\theta_C} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} |\cos \theta| d\theta + \frac{1}{\theta_C} \int_{\phi(\mathbf{w})-\theta_C/2}^{\phi(\mathbf{w})+\theta_C/2} \cos \theta d\theta \right) \\ &= \frac{1}{2} \left(\mathbb{E} \left[\left| \left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \right| \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \right. \\ &\quad \left. + \mathbb{E} \left[\left\langle \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \mathbf{w} \right\rangle \middle| \frac{\mathbf{g}}{\|\mathbf{g}\|_2} \in \mathcal{C} \right] \right). \end{aligned} \quad (\text{A.25})$$

As shown above, the first term in (A.25) is maximized at $\phi(\mathbf{w}) = 0$ and the maximum is given in (A.13). Furthermore, the second term in (A.25) is rewritten as

$$\begin{aligned} \int_{\phi(\mathbf{w})-\theta_c/2}^{\phi(\mathbf{w})+\theta_c/2} \cos \theta d\theta &= \sin\left(\phi(\mathbf{w}) + \frac{\theta_c}{2}\right) - \sin\left(\phi(\mathbf{w}) - \frac{\theta_c}{2}\right) \\ &= 2 \cos \phi(\mathbf{w}) \sin\left(\frac{\theta_c}{2}\right). \end{aligned} \quad (\text{A.26})$$

Since $\sin(\theta_c/2) \geq 0$, the expression in (A.26) is a decreasing function of $\phi(\mathbf{w}) \in [0, \pi]$. Hence, the maximum is attained at $\phi(\mathbf{w}) = 0$ as

$$\max_{\mathbf{w} \in \mathbb{S}^1} 2 \cos \phi(\mathbf{w}) \sin\left(\frac{\theta_c}{2}\right) = 2 \sin\left(\frac{\theta_c}{2}\right). \quad (\text{A.27})$$

Since the two terms in (A.25) are maximized simultaneously, by plugging in the above results to (A.24), the second assertion is obtained.

C. Proof of Lemma 2

By construction, we have

$$\frac{1}{n} \sum_{i=1}^N w_i \mathbf{x}_i \sim \text{Normal}\left(\mathbf{0}, \frac{\|\mathbf{w}\|_2^2}{n^2} \mathbf{I}_p\right).$$

Then, the concentration of the Euclidean norm of a standard Gaussian vector guarantees, with probability at least $1 - \delta/2$, that

$$\left\| \frac{1}{n} \sum_{i=1}^n w_i \mathbf{x}_i \right\|_2 \lesssim \frac{\|\mathbf{w}\|_2}{n} (\sqrt{p} + \sqrt{\log(1/\delta)}) \quad (\text{A.28})$$

for some absolute constant C . This implies the first bound in (II.11).

Next, we want to obtain an upper bound on the second term in (II.11). By the variational characterization of the spectral norm, we have

$$\left\| \frac{1}{n} \sum_{i=1}^n w_i (\mathbf{x}_i \mathbf{x}_i^\top - \mathbf{I}_p) \right\| \leq \sup_{\mathbf{u} \in \mathbb{B}_2^p} \left| \frac{1}{n} \sum_{i=1}^n w_i ((\mathbf{x}_i^\top \mathbf{u})^2 - 1) \right|. \quad (\text{A.29})$$

For brevity, we introduce a shorthand notation to denote the following random process

$$Y_{\mathbf{u}} := \sum_{i=1}^n w_i ((\mathbf{x}_i^\top \mathbf{u})^2 - 1),$$

indexed by $\mathbf{u} \in \mathbb{B}_2^p$. Then, for $\mathbf{u}, \mathbf{u}' \in \mathbb{B}_2^p$, we have

$$Y_{\mathbf{u}} - Y_{\mathbf{u}'} = \sum_{i=1}^n w_i \langle \mathbf{x}_i, \mathbf{u} - \mathbf{u}' \rangle \langle \mathbf{x}_i, \mathbf{u} + \mathbf{u}' \rangle.$$

Therefore, we bound the subexponential norm of each summand as

$$\begin{aligned} &\|w_i \langle \mathbf{x}_i, \mathbf{u} - \mathbf{u}' \rangle \langle \mathbf{x}_i, \mathbf{u} + \mathbf{u}' \rangle\|_{\psi_1} \\ &\leq w_i \|\langle \mathbf{x}_i, \mathbf{u} - \mathbf{u}' \rangle\|_{\psi_2} \cdot \|\langle \mathbf{x}_i, \mathbf{u} + \mathbf{u}' \rangle\|_{\psi_2} \lesssim w_i \|\mathbf{u} - \mathbf{u}'\|_2. \end{aligned}$$

Applying the Bernstein inequality (e.g. see [26, Theorem 2.8.1]) then yields

$$\begin{aligned} \mathbb{P}\left(|Y_{\mathbf{u}} - Y_{\mathbf{u}'}| \geq c \left(\sqrt{t} \|\mathbf{w}\|_2 \|\mathbf{u} - \mathbf{u}'\|_2 + t \|\mathbf{w}\|_\infty \|\mathbf{u} - \mathbf{u}'\|_2\right)\right) \\ \leq 2 \exp(-t), \end{aligned} \quad (\text{A.30})$$

for any $t \geq 0$ and an absolute constant c . Then, the process $Y_{\mathbf{u}}$ has mixed tail increments (i.e. see [27, Equation 12]) with respect to the metrics (d_1, d_2) where $d_1(\mathbf{a}, \mathbf{b}) = \|\mathbf{w}\|_\infty \|\mathbf{a} - \mathbf{b}\|_2$ and $d_2(\mathbf{a}, \mathbf{b}) = \|\mathbf{w}\|_2 \|\mathbf{a} - \mathbf{b}\|_2$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{B}_2^p$. Hence, applying [27, Corollary 5.2] with the bound on γ -functional (i.e. see [27, Equation 4]) provides

$$\begin{aligned} &\sup_{\mathbf{u} \in \mathbb{B}_2^p} |Y_{\mathbf{u}}| \\ &\lesssim \|\mathbf{w}\|_2 \left(\int_0^\infty \sqrt{\log N(\mathbb{B}_2^p, \|\cdot\|_2, \eta)} d\eta + \sqrt{\log(1/\delta)} \right) \\ &\quad + \|\mathbf{w}\|_\infty \left(\int_0^\infty \log N(\mathbb{B}_2^p, \|\cdot\|_2, \eta) d\eta + \log(1/\delta) \right) \\ &\stackrel{(b)}{\leq} \|\mathbf{w}\|_2 (\sqrt{p} + \sqrt{\log(1/\delta)}) + \|\mathbf{w}\|_\infty (p + \log(1/\delta)), \end{aligned}$$

holds with probability at least $1 - \delta/2$ where (b) holds due to an upper bound on the covering number $N(\mathbb{B}_2^p, \|\cdot\|_2, \eta) \leq (3/\eta)^p$ (e.g. see [26, Example 8.1.11]). This implies the second bound in (II.11).

D. Proof of Lemma 3

For any \mathbf{z} satisfying $\|\mathbf{z}\|_{1,2} = 1$, we have

$$V_{\mathbf{z}} \geq \min_{\|\mathbf{z}\|_{1,2}=1} \mathbb{E} V_{\mathbf{z}} - \sup_{\mathbf{z} \in B_{1,2}} |V_{\mathbf{z}} - \mathbb{E} V_{\mathbf{z}}|. \quad (\text{A.31})$$

In what follows, we derive lower estimates of the summands in the right-hand side of (A.31).

First, we derive a lower bound on $\min_{\|\mathbf{z}\|_{1,2}=1} \mathbb{E} V_{\mathbf{z}}$. Since $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$, we have

$$\begin{aligned} \mathbb{E} V_{\mathbf{z}} &= \mathbb{E} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{C_j}(\mathbf{x}_i) |\langle \mathbf{x}_i, \mathbf{z}_j \rangle| = \mathbb{E} \sum_{j=1}^k \mathbb{1}_{C_j}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{z}_j \rangle| \\ &= \sum_{j=1}^k \|\mathbf{z}_j\|_2 \mathbb{E} \mathbb{1}_{C_j}(\mathbf{g}) \left| \left\langle \mathbf{g}, \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\rangle \right|, \end{aligned}$$

where $\mathbf{z} = [\mathbf{z}_1; \dots; \mathbf{z}_k]$. Then $\mathbb{E} V_{\mathbf{z}}$ is lower-bounded by

$$\mathbb{E} V_{\mathbf{z}} \geq \|\mathbf{z}\|_{1,2} \inf_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{C_j}(\mathbf{g}) |\langle \mathbf{g}, \mathbf{w} \rangle|.$$

Next, we show that $(V_{\mathbf{z}} - \mathbb{E} V_{\mathbf{z}})_{\mathbf{z} \in B_{1,2}}$ is concentrated around 0 with high probability by using the following lemma.

Lemma 7: Suppose that $\mathcal{A}_1, \dots, \mathcal{A}_k$ be disjoint subsets in \mathbb{R}^p . Let $(U_{\mathbf{z}})_{\mathbf{z} \in B_{1,2}}$ be a random process defined by

$$U_{\mathbf{z}} := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{A}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, \mathbf{z}_j \rangle, \quad (\text{A.32})$$

where $\mathbf{x}_1, \dots, \mathbf{x}_n$ are i.i.d. $\text{Normal}(\mathbf{0}, \mathbf{I}_p)$. Then, for any $\delta \in (0, 1)$, there exists an absolute constant $c > 0$ such that

$$\sup_{\mathbf{z} \in B_{1,2}} |U_{\mathbf{z}} - \mathbb{E} U_{\mathbf{z}}| \leq c \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \quad (\text{A.33})$$

holds with probability at least $1 - \delta$.

Proof: We first show that $U_{\mathbf{z}}$ has sub-Gaussian increments with respect to the $\ell_{\infty}^k(\ell_2^p)$ -norm, i.e.

$$\|U_{\mathbf{z}} - U_{\mathbf{z}'}\|_{\psi_2} \lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \left\| (\mathbf{z}_j)_{j=1}^k - (\mathbf{z}'_j)_{j=1}^k \right\|_{\ell_{\infty}^k(\ell_2^p)}. \quad (\text{A.34})$$

Since $\mathcal{A}_1, \dots, \mathcal{A}_k$ are disjoint, it follows that

$$\begin{aligned} |U_{\mathbf{z}} - U_{\mathbf{z}'}| &\leq \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{A}_j}(\mathbf{x}_i) |\langle \mathbf{x}_i, \mathbf{z}_j - \mathbf{z}'_j \rangle| \\ &\leq \frac{1}{n} \sum_{i=1}^n \max_{1 \leq j \leq k} |\langle \mathbf{x}_i, \mathbf{z}_j - \mathbf{z}'_j \rangle| \end{aligned} \quad (\text{A.35})$$

holds almost surely, where the last step follows from Hölder's inequality. We proceed with the following lemma.

Lemma 8 ([28, Lemma 2.2.2]): Let $\mathbf{g} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_p)$ and $\mathbf{a}_1, \dots, \mathbf{a}_k \in \mathbb{R}^p$. Then

$$\left\| \max_{j \in [k]} |\langle \mathbf{g}, \mathbf{a}_j \rangle| \right\|_{\psi_2} \lesssim \sqrt{\log k} \max_{j \in [k]} \|\mathbf{a}_j\|_2.$$

It follows from (A.35) and Lemma 8 that

$$\begin{aligned} \|U_{\mathbf{z}} - U_{\mathbf{z}'}\|_{\psi_2} &\leq \left\| \frac{1}{n} \sum_{i=1}^n \max_{j \in [k]} |\langle \mathbf{x}_i, \mathbf{z}_j - \mathbf{z}'_j \rangle| \right\|_{\psi_2} \\ &\lesssim \frac{1}{n} \sqrt{\sum_{i=1}^n \left\| \max_{j \in [k]} |\langle \mathbf{x}_i, \mathbf{z}_j - \mathbf{z}'_j \rangle| \right\|_{\psi_2}^2} \\ &\lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \max_{j \in [k]} \|\mathbf{z}_j - \mathbf{z}'_j\|_2 \\ &= \frac{\sqrt{\log k}}{\sqrt{n}} \left\| (\mathbf{z}_j)_{j=1}^k - (\mathbf{z}'_j)_{j=1}^k \right\|_{\ell_{\infty}^k(\ell_2^p)}, \end{aligned}$$

where the second inequality follows from [26, Proposition 2.6.1].

Since $U_{\mathbf{z}}$ has a sub-Gaussian increment as in (A.34), by [26, Lemma 2.6.8], which says that centering does not harm the sub-gaussianity, we also have

$$\begin{aligned} \|(U_{\mathbf{z}} - \mathbb{E}U_{\mathbf{z}}) - (U_{\mathbf{z}'} - \mathbb{E}U_{\mathbf{z}'})\|_{\psi_2} \\ \lesssim \frac{\sqrt{\log k}}{\sqrt{n}} \left\| (\mathbf{z}_j)_{j=1}^k - (\mathbf{z}'_j)_{j=1}^k \right\|_{\ell_{\infty}^k(\ell_2^p)}. \end{aligned} \quad (\text{A.36})$$

Therefore Dudley's inequality [29] applies to provide a tail bound on the left-hand side of (A.33). Specifically it follows from a version of Dudley's inequality [26, Theorem 8.1.6] that

$$\begin{aligned} \sup_{\mathbf{z} \in B_{1,2}} |U_{\mathbf{z}} - \mathbb{E}U_{\mathbf{z}}| &\lesssim \\ &\frac{\sqrt{\log k}}{\sqrt{n}} \left(\int_0^\infty \sqrt{\log N(B_{1,2}, \|\cdot\|_{\ell_{\infty}^k(\ell_2^p)}, \eta)} d\eta + u \text{diam}(B_{1,2}) \right) \end{aligned} \quad (\text{A.37})$$

holds with probability at least $1 - 2\exp(-u^2)$. Note that the diameter term in (A.37) is trivially upper-bounded by

$$\text{diam}(B_{1,2}) = \sup_{\mathbf{z}, \mathbf{z}' \in B_{1,2}} \|\mathbf{z} - \mathbf{z}'\|_{\ell_{\infty}^k(\ell_2^p)} \leq 2.$$

Moreover, since $B_{1,2} \subseteq \sqrt{p}B_1$, where B_1 denotes the unit ball in ℓ_1 , we have

$$\begin{aligned} &\int_0^\infty \sqrt{\log N(B_{1,2}, \|\cdot\|_{\ell_{\infty}^k(\ell_2^p)}, \eta)} d\eta \\ &\leq \int_0^\infty \sqrt{\log N(\sqrt{p}B_1, \|\cdot\|_{\ell_{\infty}^k(\ell_2^p)}, \eta)} d\eta \\ &\lesssim \sqrt{p} \log^{3/2} p \log^2 k, \end{aligned}$$

where the second inequality follows from Maurey's empirical method [30] (also see [31, Lemma 3.4]). By plugging in these estimates to (A.37), we obtain that

$$\sup_{\mathbf{z} \in B_{1,2}} |U_{\mathbf{z}} - \mathbb{E}U_{\mathbf{z}}| \lesssim \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2}$$

holds with probability at least $1 - \delta$. \blacksquare

Note that $\mathcal{C}_1, \dots, \mathcal{C}_k$ are disjoint except on a boundary, which corresponds to a set of measure zero. Since the standard multivariate normal distribution is absolutely continuous relative to the Lebesgue measure, these null sets can be ignored in getting a tail bound on the infimum of the random process $(V_{\mathbf{z}})_{\mathbf{z} \in B_{1,2}}$. Moreover, $V_{\mathbf{z}}$ is written as $V_{\mathbf{z}} = V_{\mathbf{z}}^+ + V_{\mathbf{z}}^-$, where

$$V_{\mathbf{z}}^+ := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, \mathbf{z}_j \rangle_+$$

and

$$V_{\mathbf{z}}^- := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, -\mathbf{z}_j \rangle_+.$$

Since $(V_{\mathbf{z}}^+)_{\mathbf{z} \in B_{1,2}}$ and $(V_{\mathbf{z}}^-)_{\mathbf{z} \in B_{1,2}}$ are in the form of (A.32), by Lemma 7, we obtain that

$$\begin{aligned} \sup_{\mathbf{z} \in B_{1,2}} |V_{\mathbf{z}} - \mathbb{E}V_{\mathbf{z}}| &\leq \sup_{\mathbf{z} \in B_{1,2}} |V_{\mathbf{z}}^+ - \mathbb{E}V_{\mathbf{z}}^+| + \sup_{\mathbf{z} \in B_{1,2}} |V_{\mathbf{z}}^- - \mathbb{E}V_{\mathbf{z}}^-| \\ &\lesssim \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \end{aligned} \quad (\text{A.38})$$

holds with probability at least $1 - \delta/2$.

Finally, the assertion is obtained by plugging in the above estimates to (A.31).

E. Proof of Lemma 4

Note that $Q_{\mathbf{z}}$ is decomposed into

$$\begin{aligned} Q_{\mathbf{z}} &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, \mathbf{z}_j \rangle \\ &\quad + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\mathcal{C}_j \setminus \tilde{\mathcal{C}}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, -\mathbf{z}_j \rangle. \end{aligned} \quad (\text{A.39})$$

Then the summands in the right-hand side of (A.39) are respectively upper-bounded by

$$Q'_{\mathbf{z}} := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\tilde{\mathcal{C}}_j \setminus \mathcal{C}_j}(\mathbf{x}_i) \langle \mathbf{x}_i, \mathbf{z}_j \rangle_+$$

and

$$Q''_z := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{x}_i) \langle \mathbf{x}_i, -\mathbf{z}_j \rangle_+.$$

We upper-bound $\sup_{\mathbf{z} \in B_{1,2}} Q'_z$ and $\sup_{\mathbf{z} \in B_{1,2}} Q''_z$ to get an upper bound on $\sup_{\mathbf{z} \in B_{1,2}} Q_z$ through (A.39) by the triangle inequality. Specifically, we show that there exists an absolute constant $c > 0$ such that

$$\begin{aligned} \sup_{\|\mathbf{z}\|_{1,2}=1} Q'_z &\leq \sup_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \\ &\quad + c \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \end{aligned} \quad (\text{A.40})$$

and

$$\begin{aligned} \sup_{\|\mathbf{z}\|_{1,2}=1} Q''_z &\leq \sup_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ \\ &\quad + c \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \end{aligned}$$

hold simultaneously with probability at least $1 - \delta/2$.

Due to the symmetry, it suffices to show that (A.40) holds with probability $1 - \delta/4$. By the triangle inequality, it follows that

$$\sup_{\|\mathbf{z}\|_{1,2}=1} Q'_z \leq \sup_{\|\mathbf{z}\|_{1,2}=1} \mathbb{E} Q'_z + \sup_{\mathbf{z} \in B_{1,2}} |Q'_z - \mathbb{E} Q'_z|.$$

Then, similar to Lemma 3, we derive (A.40) through the concentration of the maximum deviation, that is, $\sup_{\mathbf{z} \in B_{1,2}} |Q'_z - \mathbb{E} Q'_z|$, and an upper bound on $\sup_{\mathbf{z} \in B_{1,2}} \mathbb{E} Q'_z$. The supremum of the expectation is upper-bounded as

$$\begin{aligned} \mathbb{E} Q'_z &= \mathbb{E} \sum_{j=1}^k \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{z}_j \rangle_+ \\ &\leq \max_{j \in [k], \mathbf{w} \in \mathbb{S}^{p-1}} \mathbb{E} \mathbb{1}_{\tilde{C}_j \setminus C_j}(\mathbf{g}) \langle \mathbf{g}, \mathbf{w} \rangle_+ + \sum_{j=1}^k \|\mathbf{z}_j\|_2. \end{aligned}$$

Moreover, since $\tilde{C}_1, \dots, \tilde{C}_k$ are disjoint (except on a set of measure zero), by Lemma 7, we obtain that

$$\sup_{\mathbf{z} \in B_{1,2}} |Q'_z - \mathbb{E} Q'_z| \lesssim \left(\frac{p \log^3 p \log^5 k + \log(\delta^{-1}) \log k}{n} \right)^{1/2} \quad (\text{A.41})$$

holds with probability at least $1 - \delta/4$. This provides the assertion in (A.40).

ACKNOWLEDGEMENTS

The authors appreciate anonymous reviewers for their insightful feedback and constructive suggestions that helped significantly improve the technical contributions and presentation of the manuscript. S.K. and K.L. were supported in part by NSF CAREER award CCF-1943201. S.B. was supported in part by Semiconductor Research Corporation (SRC) and Defense Advanced Research Projects Agency (DARPA).

REFERENCES

- [1] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression: Parameter estimation for gaussian designs," *IEEE Transactions on Information Theory*, vol. 68, no. 3, pp. 1851–1885, 2021.
- [2] S. Bahmani and J. Romberg, "Phase retrieval meets statistical learning theory: A flexible convex relaxation," in *Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 252–260.
- [3] S. Bahmani, "Estimation from nonlinear observations via convex programming with application to bilinear regression," *Electronic Journal of Statistics*, vol. 13, no. 1, pp. 1978–2011, 2019.
- [4] S. Bahmani and J. Romberg, "Solving equations of random convex functions via anchored regression," *Foundations of Computational Mathematics*, vol. 19, no. 4, pp. 813–841, 2019.
- [5] T. Goldstein and C. Studer, "Phasemax: Convex phase retrieval via basis pursuit," *IEEE Transactions on Information Theory*, vol. 64, no. 4, pp. 2675–2689, 2018.
- [6] E. J. Candes, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1241–1274, 2013.
- [7] I. Waldspurger, A. d'Aspremont, and S. Mallat, "Phase recovery, maxcut and complex semidefinite programming," *Mathematical Programming*, vol. 149, no. 1-2, pp. 47–81, 2015.
- [8] R. Balestrierio and R. Baraniuk, "Mad max: Affine spline insights into deep learning," *arXiv preprint arXiv:1805.06576*, 2018.
- [9] R. Balestrierio, R. Cosentino, B. Aazhang, and R. Baraniuk, "The geometry of deep networks: Power diagram subdivision," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 806–15 815.
- [10] R. Balestrierio, S. Paris, and R. Baraniuk, "Max-affine spline insights into deep generative networks," *arXiv preprint arXiv:2002.11912*, 2020.
- [11] A. Siahkamari, V. Saligrama, D. Castanon, and B. Kulis, "Learning Bregman divergences," *arXiv preprint arXiv:1905.11545*, 2019.
- [12] G. Balázs, "Convex regression: Theory, practice, and applications," Ph.D. dissertation, Dept. Comput. Sci., University of Alberta, Edmonton, AB, Canada, 2016.
- [13] A. Magnani and S. P. Boyd, "Convex piecewise-linear fitting," *Optimization and Engineering*, vol. 10, no. 1, pp. 1–17, 2009.
- [14] A. Toriello and J. P. Vielma, "Fitting piecewise linear continuous functions," *European Journal of Operational Research*, vol. 219, no. 1, pp. 86–95, 2012.
- [15] L. A. Hannah and D. B. Dunson, "Multivariate convex regression with adaptive partitioning," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 3261–3294, 2013.
- [16] V. T. Ho, H. A. Le Thi, and T. P. Dinh, "DCA with successive DC decomposition for convex piecewise-linear fitting," in *International Conference on Computer Science, Applied Mathematics and Applications*. Springer, 2019, pp. 39–51.
- [17] P. D. Tao and L. T. H. An, "A DC optimization algorithm for solving the trust-region subproblem," *SIAM Journal on Optimization*, vol. 8, no. 2, pp. 476–505, 1998.
- [18] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925–936, 2010.
- [19] J. van den Brand, "A deterministic linear program solver in current matrix multiplication time," in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2020, pp. 259–278.
- [20] L. Gurobi Optimization, "Gurobi optimizer reference manual," 2021. [Online]. Available: <http://www.gurobi.com>
- [21] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression: Provable, tractable, and near-optimal statistical estimation," *arXiv preprint arXiv:1906.09255*, 2019.
- [22] I. Diakonikolas, J. H. Park, and C. Tzamos, "Relu regression with massart noise," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [23] E. John and E. A. Yildirim, "Implementation of warm-start strategies in interior-point methods for linear programming in fixed dimension," *Computational Optimization and Applications*, vol. 41, no. 2, pp. 151–183, 2008.
- [24] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.
- [25] Y. S. Tan and R. Vershynin, "Phase retrieval via randomized kaczmarz: theoretical guarantees," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 97–123, 2019.
- [26] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

- [27] S. Dirksen, “Tail bounds via generic chaining,” 2015.
- [28] A. W. van der Vaart and J. A. Wellner, *Weak convergence and empirical processes*, ser. Springer Series in Statistics. Springer, 1996.
- [29] R. M. Dudley, “The sizes of compact subsets of Hilbert space and continuity of Gaussian processes,” *Journal of Functional Analysis*, vol. 1, no. 3, pp. 290–330, 1967.
- [30] B. Carl, “Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces,” in *Annales de l’institut Fourier*, vol. 35, no. 3, 1985, pp. 79–118.
- [31] M. Junge and K. Lee, “Generalized notions of sparsity and restricted isometry property. Part I: A unified framework,” *Information and Inference: A Journal of the IMA*, vol. 9, no. 1, pp. 157–193, 2020.

Seonho Kim received the B.S. and M.S. degrees in electrical and computer engineering from Ajou University, Suwon, South Korea, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with The Ohio State University, Columbus, OH, USA. His main research interests include the areas of optimization, machine learning, and data science.

Sohail Bahmani is a researcher broadly interested in the areas of statistics and optimization, and their interplay. He has made contributions in areas such as nonlinear regression, robust estimation, high-dimensional statistics, and inverse problems. He has received a best paper award from the AISTats conference in 2017 for his work on the phase retrieval problem. He was a postdoctoral researcher in the School of Electrical and Computer Engineering at Georgia Tech from 2013 to 2022, and received his PhD in Electrical and Computer Engineering from Carnegie Mellon University in 2013.

Kiryung Lee (Senior Member 2019, IEEE) received the B.S. and the M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 2000 and 2002, and the Ph.D. in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA. Since 2018, he has been an Assistant Professor with the Electrical and Computer Engineering Department, Ohio State University, Columbus. His research focuses on developing mathematical theory and optimization algorithms for inverse problems in signal processing, imaging, machine learning, statistics, and data science. He is a recipient of the NSF CAREER Award.