Max-Affine Regression: Parameter Estimation for Gaussian Designs

Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran, *Fellow, IEEE*

***Abstract*— Max-affine regression refers to a model where the unknown regression function is modeled as a maximum of k unknown affine functions for a fixed $k \geq 1$. This generalizes linear regression and (real) phase retrieval, and is closely related to convex regression. We study this problem in the high-dimensional setting assuming that k is a fixed constant, and focus on the estimation of the unknown coefficients of the affine functions underlying the model. We analyze a natural alternating minimization (AM) algorithm for the non-convex least squares objective when the design is Gaussian. We show that the AM algorithm, when initialized suitably, converges with high probability and at a geometric rate to a small ball around the optimal coefficients. In order to initialize the algorithm, we propose and analyze a combination of a spectral method and a search algorithm in a low-dimensional space, which may be of independent interest. The final rate that we obtain is near-parametric and minimax optimal (up to a polylogarithmic factor) as a function of the dimension, sample size, and noise variance. In that sense, our approach should be viewed as a *direct* and implementable method of enforcing regularization to alleviate the curse of dimensionality in problems of the convex regression type. Numerical experiments illustrate the sharpness of our bounds in the various problem parameters.**

***Index Terms*— Max-affine regression, alternating minimization, dimension reduction, iterative optimization.**

I. Introduction

MAX-AFFINE regression refers to the regression model

$$Y = \max_{1 \leq j \leq k} \left(\langle X, \theta_j^* \rangle + b_j^* \right) + \epsilon \tag{1}$$

where Y is a univariate response, X is a d-dimensional vector of covariates and ϵ models zero-mean noise that is independent of X. We assume that $k \geq 1$ is a known integer and study the problem of estimating the unknown parameters $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$ and $b_1^*, \ldots, b_k^* \in \mathbb{R}$ from independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn according to the model (1). Furthermore, we assume for concreteness[1] in this paper that the covariate distribution is standard Gaussian, with $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$.

Let us provide some motivation for studying the model (1). When $k = 1$, equation (1) corresponds to the classical linear regression model. When $k = 2$, the intercepts $b_2^* = b_1^* = 0$, and $\theta_2^* = -\theta_1^* = \theta^*$, the model (1) reduces to

$$Y = |\langle X, \theta^* \rangle| + \epsilon. \tag{2}$$

The problem of recovering θ^* from observations drawn according to the above model is known as (real) phase retrieval—variants of which arise in a diverse array of science and engineering applications [2]–[5]—and has associated with it an extensive statistical and algorithmic literature.

To motivate the model (1) for general k, note that the function $x \mapsto \max_{1 \leq j \leq k} (\langle x, \theta_j^* \rangle + b_j^*)$ is always a convex function and, thus, estimation under the model (1) can be used to fit convex functions to the observed data. Indeed, the model (1) serves as a parametric approximation to the non-parametric convex regression model

$$Y = \phi^*(X) + \epsilon, \tag{3}$$

where $\phi^* : \mathbb{R}^d \to \mathbb{R}$ is an unknown convex function. It is well-known that convex regression suffers from the curse of dimensionality unless d is small, which is basically a consequence of the fact that the metric entropy of natural totally bounded sub-classes of convex functions grows exponentially in d (see, e.g., [6]–[8]). To overcome this curse of dimensionality, one would need to work with more structured sub-classes of convex functions. Since convex functions can be approximated to arbitrary accuracy by maxima of affine functions, it is reasonable to regularize the problem by considering only those convex functions that can be written as a maximum of a fixed number of affine functions. Constraining the number of affine pieces in the function therefore presents a simple method to enforce structure, and such function classes have been introduced and studied in the convex regression literature (see e.g., [9]). This assumption directly leads to our model (1), and it has been argued by [10]–[12] that the parametric model (1) is a tractable alternative to the full non-parametric

Manuscript received February 23, 2020; revised September 22, 2021; accepted November 8, 2021. Date of publication November 25, 2021; date of current version February 17, 2022. The work of Avishek Ghosh and Kannan Ramchandran was supported in part by the National Science Foundation (NSF) under Grant CCF-1527767. The work of Ashwin Pananjady was supported in part by Grants ONR-N00014-18-1-2640, NSF DMS-1612948, NSF CCF-1704967, and NSF CCF-2107455. The work of Adityanand Guntuboyina was supported in part by the NSF CAREER Grant DMS-16-54589. *(Avishek Ghosh and Ashwin Pananjady contributed equally to this work.) (Corresponding author: Ashwin Pananjady.)*

Avishek Ghosh was with the Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA. He is now with the Halicioglu Data Science Institute, University of California at San Diego, San Diego, CA 92093 USA.

Ashwin Pananjady was with the Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA 94720 USA. He is now with the School of Industrial and Systems Engineering and the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: ashwinpm@gatech.edu).

Adityanand Guntuboyina and Kannan Ramchandran are with the Department of Statistics and the Department of Electrical Engineering and Computer Sciences (EECS), University of California at Berkeley, Berkeley, CA 94720 USA.

Communicated by E. Gassiat, Associate Editor for Probability and Statistics.

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TIT.2021.3130717.

Digital Object Identifier 10.1109/TIT.2021.3130717

[1]Our companion paper [1] weakens distributional assumptions on the covariates, but this requires significantly more technical effort.

convex regression model (3) in common applications of convex regression to data arising in economics, finance and operations research where d is often moderate to large.

Another motivation for the model (1) comes from the problem of estimating convex sets from support function measurements. The support function of a compact convex set $K \subseteq \mathbb{R}^d$ is defined by $h_K(x) := \sup_{u \in K} \langle x, u \rangle$ for d-dimensional unit vectors x. The problem of estimating an unknown compact, convex set K^* from noisy measurements of $h_{K^*}(\cdot)$ arises in certain engineering applications such as robotic tactile sensing and projection magnetic resonance imaging (see, e.g., [13]–[15]). Specifically, the model considered here is

$$Y = h_{K^*}(X) + \epsilon,$$

and the goal is to estimate the set $K^* \subseteq \mathbb{R}^d$. As in convex regression, this problem suffers from a curse of dimensionality unless d is small, as is evident from known minimax lower bounds [16]. To alleviate this curse, it is natural to restrict K^* to the class of all polytopes with at most k extreme points for a fixed k; such a restriction has been studied as a special case of enforcing structure in these problems by Soh and Chandrasekharan [17]. Under this restriction, one is led to the model (1) with $b_1^* = \cdots = b_k^* = 0$, since if K^* is the polytope given by the convex hull of $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$, then its support function is equal to $x \mapsto \max_{1 \le j \le k} \langle x, \theta_j^* \rangle$.

Equipped with these motivating examples, our goal is to study a computationally efficient estimation methodology for the unknown parameters of the model (1) from i.i.d observations $(x_i, y_i)_{i=1}^n$. Before presenting our contributions, let us first rewrite the observation model (1) by using more convenient notation, and use it to describe existing estimation procedures for this model. Denote the unknown parameters by $\beta_j^* := (\theta_j^*, b_j^*) \in \mathbb{R}^{d+1}$ for $j = 1, \ldots, k$ and the observations by (ξ_i, y_i) for $i = 1, \ldots, n$, where $\xi_i := (x_i, 1) \in \mathbb{R}^{d+1}$. In this notation, the observation model takes the form

$$y_i = \max_{1 \le j \le k} \langle \xi_i, \beta_j^* \rangle + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{4}$$

Throughout the paper, we assume that in addition to the covariates being i.i.d. standard Gaussian, the noise variables $\epsilon_1, \ldots, \epsilon_n$ are independent random variables drawn from a (univariate) distribution that is zero-mean and sub-Gaussian, with unknown sub-Gaussian parameter σ.

Let us now describe existing estimation procedures for max-affine regression. The most obvious approach is the global least squares estimator, defined as any minimizer of the least squares criterion

$$(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) \in \operatorname*{argmin}_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \sum_{i=1}^n \left(y_i - \max_{1 \le j \le k} \langle \xi_i, \beta_j \rangle \right)^2. \tag{5}$$

It is easy to see (see Lemma 1 to follow) that a global minimizer of the least squares criterion above always exists but it will not—at least in general—be unique, since any relabeling of the indices of a minimizer will also be a minimizer. While the least squares estimator has appealing statistical properties (see, e.g. [16]–[18]), the optimization problem (5) is non-convex. Furthermore, for a *worst-case* choice of covariates, the problem can be shown to be NP-hard[2] via a reduction from the subset-sum problem. Consequently, we focus on settings where the covariates are drawn i.i.d. (in which this hardness no longer applies), and in particular, we assume that the covariate distribution is Gaussian.

It is interesting to compare (5) to the optimization problem used to compute the least squares estimator in the more general convex regression model (3), given by

$$\widehat{\phi}^{(\mathsf{ls})} \in \operatorname*{argmin}_{\phi} \sum_{i=1}^n \left(y_i - \phi(x_i) \right)^2, \tag{6}$$

where the minimization is over all convex functions ϕ. In sharp contrast to the problem (5), the optimization problem (6) is convex [19], [20] and can be solved efficiently for fairly large values of the pair (d, n) [21]. Unfortunately however, the utility of $\widehat{\phi}^{(\mathsf{ls})}$ in estimating the parameters of the max-affine model is debatable, as it is unclear how one may obtain estimates of the true parameters $\beta_1^*, \ldots, \beta_k^*$ from $\widehat{\phi}^{(\mathsf{ls})}$, which typically will *not* be a maximum of only k affine functions.

Three heuristic techniques for solving the non-convex optimization problem (5) were empirically evaluated by Balázs [12, Chapters 6 and 7], who compared running times and performance of these techniques on a wide variety of real and synthetic datasets for convex regression. The first technique is the alternating minimization algorithm of Magnani and Boyd [10], the second technique is the convex adaptive partitioning (or CAP) algorithm of Hannah and Dunson [11], and the third is the adaptive max-affine partitioning algorithm proposed by Balázs himself [12]. The simplest and most intuitive of these three methods is the first alternating minimization (AM) algorithm, which is an iterative algorithm for estimating the parameters $\beta_1^*, \ldots, \beta_k^*$ and forms the focus of our study. In the t-th iteration of the algorithm, the current estimates $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ are used to partition the observation indices $1, \ldots, n$ into k sets $S_1^{(t)}, \ldots, S_k^{(t)}$ such that $j \in \operatorname{argmax}_{u \in [k]} \langle \xi_i, \beta_u^{(t)} \rangle$ for every $i \in S_j^{(t)}$. For each $1 \le j \le k$, the next estimate $\beta_j^{(t+1)}$ is then obtained by performing a least squares fit (or equivalently, linear regression) to the data $(\xi_i, y_i), i \in S_j^{(t)}$. More intuition and a formal description of the algorithm are provided in Section II. Balázs found that when this algorithm was run on a variety of datasets with multiple random initializations, it compared favorably with the state of the art in terms of its final predictive performance—see, for example, Figures 7.4 and 7.5 in the thesis [12], which show encouraging results when the algorithm is used to fit convex functions to datasets of average wages and aircraft profile drag data, respectively. In the context of fitting convex sets to support function measurements, Soh and Chandrasekaran [17] recently proposed and empirically evaluated a similar algorithm in the case of isotropic covariates. However, to the best of our knowledge, no theoretical results exist to support the performance of such a technique.

In this paper, we present a theoretical analysis of the AM algorithm for recovering the parameters of the max-affine

[2]We provide a proof of this in Appendix I for completeness.

regression model when the covariate distribution is Gaussian.[3] This assumption forms a natural starting point for the study of many iterative algorithms in related problems [22]–[25], and is also quite standard in theoretical investigations of multidimensional regression problems. Note that the AM algorithm described above can be seen as a generalization of classical AM algorithms for (real) phase retrieval [26], [27], which have recently been theoretically analyzed in a series of papers [22]–[24] for Gaussian designs. The AM—and the closely related expectation maximization,[4] or EM—methodology is widely used for parameter estimation in missing data problems [28], [29] and mixture models [30], including those with covariates such as mixtures-of-experts [31] and mixtures-of-regressions [32] models. Theoretical guarantees for such algorithms have been established in multiple statistical contexts [25], [33]–[35]; in the case when the likelihood is not unimodal, these are typically of the local convergence type. In particular, algorithms of the EM type return, for many such latent variable models, minimax-optimal parameter estimates when initialized in a neighborhood of the optimal solution (e.g., [32], [36], [37]); conversely, these algorithms can get stuck at spurious fixed points when initialized at random [38]. In some specific applications of EM to mixtures of two Gaussians [39], [40] and mixtures of two regressions [41], however, it has been shown that randomly initializing the EM algorithm suffices in order to obtain consistent parameter estimates. Here, we establish guarantees on the AM algorithm for max-affine regression that are of the former type: we prove local geometric convergence of the AM iterates when initialized in a neighborhood of the optimal solution. We analyze the practical variant of the algorithm in which the steps are performed without sample-splitting. As in the case of mixture models [32], [42], we use spectral methods to obtain such an initialization.

A. Contributions

Let us now describe our results in more detail. To simplify the exposition, we state simplified corollaries of our theorems; for precise statements, see Section III. We prove in Theorem 1 that for each $\epsilon > 0$, the parameter estimates $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ returned by the AM algorithm at iteration t satisfy, with high probability, the inequality

$$\sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^*\|^2 \leq \epsilon + C(\beta_1^*, \ldots, \beta_k^*) \frac{\sigma^2 kd}{n} \log(kd) \log\Big(\frac{n}{kd}\Big) \tag{7}$$

for every $t \geq \log_{4/3}\left(\frac{\sum_{j=1}^{k} \|\beta_j^{(0)} - \beta_j^*\|^2}{\epsilon}\right)$, provided that the sample size n is sufficiently large and that the initial estimates satisfy the condition

$$\min_{c>0} \max_{1\leq j\leq k} \|c\beta_j^{(0)} - \beta_j^*\|^2 \leq \frac{1}{k} c(\beta_1^*, \ldots, \beta_k^*). \tag{8}$$

Here $C(\beta_1^*, \ldots, \beta_k^*)$ and $c(\beta_1^*, \ldots, \beta_k^*)$ are constants depending only on the true parameters $\beta_1^*, \ldots, \beta_k^*$, and their explicit values are given in Theorem 1. The constant c in equation (8) endows the initialization with a scale-invariance property: indeed, scaling all parameters $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ by the same positive constant c produces the same initial partition of subsets $S_1^{(0)}, \ldots, S_k^{(0)}$, from which the algorithm proceeds identically.

Treating k as a fixed constant, inequality (7) implies, under the initialization condition (8), that the parameter estimates returned by AM converge geometrically to within a small ball of the true parameters, and that this error term is nearly the parametric risk $\frac{\sigma^2 d}{n}$ up to a logarithmic factor. The initialization condition (8) requires the distance between the initial estimates and the true parameters to be at most a specific (k-dependent) constant. It has been empirically observed that there exist bad initializations under which the AM algorithm behaves poorly (see, e.g., [10], [12]) and assumption (8) is one way to rule these out.

A natural question based on our Theorem 1 is whether it is possible to produce preliminary estimates $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ satisfying the initialization condition (8). Indeed, one such method is to repeatedly initialize parameters (uniformly) at random within the unit ball $\mathbb{B}^{d+1}$; Balázs empirically observed in a close relative of such a scheme (see Figure 6.6 in his thesis [12]) that increasing the number of random initializations is often sufficient to get the AM algorithm to succeed. However, reasoning heuristically, the number of repetitions required to ensure that one such random initialization generates parameters that satisfy condition (8) increases exponentially in the ambient dimension d, and so it is reasonable to ask if, in large dimensions, there is some natural form of dimensionality reduction that allows us to perform this step in a lower-dimensional space.

When[5] $k < d$, we show that a natural spectral method (described formally in Algorithm 2) is able to reduce the dimensionality of our problem from d to k. In particular, this method returns an orthonormal basis of vectors $\widehat{U}_1, \ldots, \widehat{U}_k$ such that the k-dimensional linear subspace spanned by these vectors accurately estimates the subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$. We form the matrix $\widehat{U} := [\widehat{U}_1 : \cdots : \widehat{U}_k]$ by collecting these vectors as its columns, and in order to account for the intercepts, further append such a matrix to form the matrix $\widehat{V} := \begin{bmatrix} \widehat{U} & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(k+1)}$. Finally, we construct a covering of the $(k+1)$-dimensional unit ball $\mathbb{M} = \{\nu^\ell, \ell = 1, \ldots, M\}$ and search it for a "good" set of initial parameters. To that end, we evaluate (on an independent set of samples) the goodness-of-fit statistic $\min_{c\geq 0} \sum_i (y_i - c \max_{1\leq j\leq k} \langle \xi_i, \widehat{V}\nu_j \rangle)^2$ for each $\nu_1, \ldots, \nu_k \in \mathbb{M}$, where the minimization over the constant c

[3] In our companion paper [1], we weaken this assumption on the covariate distribution.

[4] Indeed, for many problems, the EM algorithm reduces to AM in the noiseless limit, and AM should thus be viewed as a variant of EM that uses hard-thresholding to determine values of the latent variables.

[5] If $k \geq d$, then this dimensionality reduction step can be done away with and one can implement the search routine directly.

accounts for the scale-invariance property alluded to above. Letting $\nu_1^\sharp, \ldots, \nu_k^\sharp$ denote the minimizers, we then return the initializer $\beta_j^{(0)} = \widehat{V}\nu_j^\sharp$ for $j = 1, \ldots, k$.

Our algorithm can thus be viewed as a variant of the repeated random initialization evaluated by Balázs [12], but incurs significantly smaller computational cost, since we only run the full-blown iterative AM algorithm once. Note that our algorithm treats the radius of the covering (and subsequently its size M) as a tuning parameter to be chosen by the statistician, similar to Balázs [12], but we show a concrete upper bound on M that is sufficient to guarantee convergence. In particular, we show that in order to produce an initialization satisfying condition (8) with high probability, it suffices to choose M as a function *only* of the number of affine pieces k and other geometric parameters of the problem (and independently of the sample size n and ambient dimension d when $k \leq d$).

To produce our overall guarantee, we combine the initialization with the AM algorithm in Corollary 1, showing that provided the sample size scales linearly in the dimension (with a multiplicative pre-factor that depends polynomially on k and other problem-dependent parameters), we obtain estimates that are accurate up to the parametric risk. Our algorithm is also computationally efficient when k is treated as a fixed constant.

From a technical standpoint, our results for the AM algorithm are significantly more challenging to establish than related results in the literature [23], [25], [43], [44]. First, it is technically very challenging to compute the *population operator* [25]—corresponding to running the AM update in the infinite sample limit—in this setting, since the $\max$ function introduces intricate geometry in the problem that is difficult to reason about in closed form. Second, we are interested in analyzing the AM update without sample-splitting, and so cannot assume that the iterates are independent of the covariates; the latter assumption has been used fruitfully in the literature to simplify analyses of such algorithms [22], [24], [43]. Third, and unlike algorithms for phase retrieval [23], [44], our algorithm performs least squares using sub-matrices of the covariate matrix that are chosen *depending* on our random iterates. Accordingly, a key technical difficulty of the proof, which may be of independent interest, is to control the spectrum of these random matrices, rows of which are drawn from (randomly) truncated variants of the Gaussian distribution.

Our spectral initialization algorithm is also a natural estimator based on the method-of-moments, and has been used in a variety of non-convex problems [32], [36], [37]. However, our guarantees for this step are once again non-trivial to establish. In particular, the eigengap of the population moment (on which the rates of the estimator depend) is difficult to compute in our case since the $\max$ function is not differentiable, and so it is not clear that higher order moments return reasonable estimates even in the infinite sample limit (see Section II). However, since we operate exclusively with Gaussian covariates, we are able to use some classical moment calculations for truncated Gaussian distributions [45] in order to bound the eigengap. Translating these calculations into an eigengap is quite technical, and involves the isolation of many properties of the population moments that may be of independent interest.

Finally, it is important to note that owing to the scale invariance of our initialization condition (8) and goodness-of-fit statistic, our search scheme does not require a bound on the size of the parameters; it suffices to initialize parameters uniformly within the *unit* ball. This is in contrast to other search procedures employed for similar problems [46], [47], which are based on covering arguments and require a bound on the maximum norm of the unknown parameters.

B. Organization

The rest of the paper is organized as follows. Section II describes the problem setup and our methodology (including the AM algorithm and initialization methods) in more detail. In Section III, we present our main theoretical results and their consequences, complementing our discussion with figures that verify that our results are borne out in simulation. An overview of the main ideas behind our proofs is given in Section III-D. We conclude the main paper with a discussion in Section IV of some related models and future directions. Full proofs of our results are presented in the supplementary material in Sections B-D, with further technical details relegated to the later sections of the appendix.

C. Notation

For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. All logarithms are to the natural base unless otherwise mentioned. For two sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there is a universal constant C such that $a_n \leq Cb_n$ for all $n \geq 1$. The relation $a_n \gtrsim b_n$ is defined analogously, and we use $a_n \sim b_n$ to indicate that both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold simultaneously. We use $c, C, c_1, c_2, \ldots$ to denote universal constants that may change from line to line. For a pair of vectors (u, v), we let $u \otimes v := uv^\top$ denote their outer product. We use $\|\cdot\|$ to denote the ℓ_2 norm unless otherwise stated. Denote by I_d the $d \times d$ identity matrix. We let $\mathbf{1}\{\mathcal{E}\}$ denote the indicator of an event $\mathcal{E}$. Let $\mathsf{sgn}(t)$ denote the sign of a scalar t, with the convention that $\mathsf{sgn}(0) = 1$. Let $\lambda_i(\Gamma)$ denote the i-th largest eigenvalue of a symmetric matrix Γ. Let $\mathbb{S}^{d-1} := \{v \in \mathbb{R}^d : \|v\| = 1\}$ denote the unit sphere in d-dimensions, and use $\mathbb{B}^d := \{v \in \mathbb{R}^d : \|v\| \leq 1\}$ to denote the d-dimensional unit ball. Finally, we use the shorthand $a \wedge b := \min(a, b)$ and $a \vee b := \max(a, b)$ for two scalars a and b.

II. Background and Problem Formulation

In this section, we formally introduce the geometric parameters underlying the max-affine regression model, as well as the methodology we use to perform parameter estimation.

A. Model and Geometric Parameters

We work throughout with the observation model defined in equation (4); recall that our covariates are drawn i.i.d. from

a standard Gaussian distribution, and that our noise is σ-sub-Gaussian. We let $X \in \mathbb{R}^{n\times d}$ denote the covariate matrix with row i given by the vector x_i, and collect the responses in a vector $y \in \mathbb{R}^n$.

Recall that $\xi_i = (x_i,\ 1) \in \mathbb{R}^{d+1}$ for each $i \in [n]$; the matrix of appended covariates $\Xi \in \mathbb{R}^{n\times(d+1)}$ is defined by appending a vector of ones to the right of the matrix X. Our primary goal is to use the data (X, y)—or equivalently, the pair (Ξ, y)—to estimate the underlying parameters $\{\beta_j^*\}_{j=1}^k$.

An important consideration in achieving such a goal is the "effective" sample size with which we observe the parameter β_j^*. Toward that end, for $X \sim \mathcal{N}(0, I_d)$, let

$$\pi_j(\beta_1^*, ., \beta_k^*) := \Pr\{\langle X,\, \theta_j^*\rangle + b_j^* = \max_{j'\in[k]} \left(\langle X,\, \theta_{j'}^*\rangle + b_{j'}^*\right)\} \tag{9}$$

denote the probability with which the j-th parameter $\beta_j^* = (\theta_j^*\ ,\ b_j^*)$ attains the maximum. Note that the event on which more than one of the parameters attains the maximum has measure zero, except in the case where $\beta_i^* = \beta_j^*$ for some $i \neq j$. We explicitly disallow this case and assume that the parameters $\beta_1^*, \ldots, \beta_k^*$ are distinct. Let

$$\pi_{\min}(\beta_1^*, \ldots, \beta_k^*) := \min_{j\in[k]} \pi_j(\beta_1^*, \ldots, \beta_k^*), \tag{10}$$

and assume that we have $\pi_{\min}(\beta_1^*, \ldots, \beta_k^*) > 0$; in other words, we ignore vacuous cases in which some parameter is never observed. Roughly speaking, the sample size of the parameter that is observed most rarely is given by $\min_{j\in[k]} \pi_j n \sim n \cdot \pi_{\min}(\beta_1^*, \ldots, \beta_k^*)$, and so the error in estimating this parameter should naturally depend on $\pi_{\min}(\beta_1^*, \ldots, \beta_k^*)$. By definition, we always have $\pi_{\min}(\beta_1^*, \ldots, \beta_k^*) \leq 1/k$.

Since we are interested in performing parameter estimation under the max-affine regression model, a few geometric quantities also appear in our bounds, and serve as natural notions of "signal strength" and "condition number" of the estimation problem. The signal strength is given by the minimum separation

$$\Delta(\beta_1^*, \ldots, \beta_k^*) = \min_{j,j':j\neq j'} \left\|\theta_j^* - \theta_{j'}^*\right\|^2;$$

we also assume that Δ is strictly positive, since otherwise, a particular parameter is never observed. To denote a natural form of conditioning, define the quantities

$$\kappa_j(\beta_1^*, \ldots, \beta_k^*) = \frac{\max_{j'\neq j} \left\|\theta_j^* - \theta_{j'}^*\right\|^2}{\min_{j'\neq j} \left\|\theta_j^* - \theta_{j'}^*\right\|^2},$$

with $\kappa(\beta_1^*, \ldots, \beta_k^*) = \max_{j\in[k]} \kappa_j(\beta_1^*, \ldots, \beta_k^*)$. Finally, let $\mathsf{B}_{\mathsf{max}}(\beta_1^*, \ldots, \beta_k^*) := \max_{j\in[k]} \|\beta_j^*\|$ denote the maximum norm of any unknown parameter. We often use the shorthand

$$\pi_{\min} = \pi_{\min}(\beta_1^*, \ldots, \beta_k^*),\ \ \Delta = \Delta(\beta_1^*, \ldots, \beta_k^*),$$
$$\kappa = \kappa(\beta_1^*, \ldots, \beta_k^*), \quad \text{and } \mathsf{B}_{\mathsf{max}} = \mathsf{B}_{\mathsf{max}}(\beta_1^*, \ldots, \beta_k^*)$$

when the true parameters $\beta_1^*, \ldots, \beta_k^*$ are clear from context.

B. Methodology

As discussed in the introduction, the most natural estimation procedure from i.i.d. samples $(\xi_i, y_i)_{i=1}^n$ of the model (4) is the least squares estimator (5). The following lemma (which does not seem to have been explicitly stated previously in the literature, except in the case $k = 2$ [18], [48]) proves that the least squares estimator $(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})})$ always exists. Note, however, that it will not be unique in general since any relabeling of a minimizer is also a minimizer.

Lemma 1: The least squares estimator $\left(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}\right)$ exists for every dataset (Ξ, y).

We postpone the proof of Lemma 1 to Appendix A. In spite of the fact that the least squares estimator always exists, the problem (5) is non-convex and NP-hard in general. The AM algorithm presents a tractable approach towards solving it in the statistical setting that we consider.

1) Alternating Minimization: We now formally describe the AM algorithm proposed by Magnani and Boyd [10]. For each $\beta_1, \ldots, \beta_k$, define the sets

$$S_j(\beta_1, \ldots, \beta_k) := \left\{ i \in [n] : j = \min \operatorname*{argmax}_{1\leq u\leq k} \left(\langle \xi_i,\, \beta_u\rangle\right) \right\} \tag{11}$$

for $j = 1, \ldots, k$. In words, the set $S_j(\beta_1, \ldots, \beta_k)$ contains the indices of samples on which parameter β_j attains the maximum; in the case of a tie, samples having multiple parameters attaining the maximum are assigned to the set with the smallest corresponding index (i.e., ties are broken in the lexicographic order[6]). Thus, the sets $\{S_j(\beta_1, \ldots, \beta_k)\}_{j=1}^k$ define a partition of $[n]$. The AM algorithm employs an iterative scheme where one first constructs the partition $S_j\left(\beta_1^{(t)}, \ldots, \beta_k^{(t)}\right)$ based on the current iterates $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ and then calculates the next parameter estimate $\beta_j^{(t+1)}$ by a least squares fit to the dataset $\{(\xi_i, y_i), i \in S_j(\beta_1^{(t)}, \ldots, \beta_k^{(t)})\}$. The algorithm (also described below as Algorithm 1) is, clearly, quite intuitive and presents a natural approach to solving (5).

As a sanity check, Lemma 2 (stated and proved in Appendix A) shows that the global least squares estimator (5) is a fixed-point of this iterative scheme under a mild technical assumption.

We also note that the AM algorithm was proposed by Soh [49] in the context of estimating structured convex sets from support function measurements. It should be viewed as a generalization of a classical algorithm for (real) phase retrieval due to Fienup [27], which has been more recently analyzed in a series of papers [22], [23] for Gaussian designs. While some analyses of AM algorithms assume sample-splitting across iterations (e.g. [22], [24], [43]), we consider the more practical variant of AM run without sample-splitting, since the update (12a)-(12b) is run on the full data (Ξ, y) in every iteration.

[6]In principle, it is sufficient to define the sets $S_j(\beta_1, \ldots, \beta_k), j \in [k]$ as any partition of $[n]$ having the property that $\langle \xi_i,\, \beta_j\rangle = \max_{u\in[k]} \langle \xi_i,\, \beta_u\rangle$ for every $j \in [k]$ and $i \in S_j(\beta_1, \ldots, \beta_k)$; here "any" means that ties can be broken according to an arbitrary rule, and we have chosen this rule to be the lexicographic order in equation (11).

Algorithm 1 Alternating Minimization for Estimating Maximum of k Affine Functions

Input: Data $\{\xi_i, y_i\}_{i=1}^n$; initial parameter estimates $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$; number of iterations T.

Output: Final estimator of parameters $\widehat{\beta}_1, \ldots, \widehat{\beta}_k$.

Initialize $t \leftarrow 0$.

repeat

Compute maximizing index sets

$$\mathsf{S}_j^{(t)} = S_j(\beta_1^{(t)}, \ldots, \beta_k^{(t)}), \tag{12a}$$

for each $j \in [k]$, according to equation (11).

Update

$$\beta_j^{(t+1)} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \sum_{i \in \mathsf{S}_j^{(t)}} \left(y_i - \langle \xi_i, \beta \rangle\right)^2, \tag{12b}$$

for each $j \in [k]$.

until $t = T$;

Return $\widehat{\beta}_j = \beta_j^{(T)}$ for each $j \in [k]$.

2) Initialization: The alternating minimization algorithm described above requires an initialization. While the algorithm was proposed to be run from a random initialization with restarts [10], [17], we propose to initialize the algorithm from parameter estimates that are sufficiently close to the optimal parameters. This is similar to multiple procedures to solve non-convex optimization problems in statistical settings (e.g., [25], [50]), that are based on iterative algorithms that exhibit *local* convergence to the unknown parameters. Such algorithms are typically initialized by using a moment method, which (under various covariate assumptions) returns useful parameter estimates.

Algorithm 2 PCA for k-Dimensional Subspace Initialization When $k < d$

Input: Data $\{\xi_i, y_i\}_{i=1}^n$.

Output: Matrix $\widehat{U} \in \mathbb{R}^{d \times k}$ having orthonormal columns that (approximately) span the k dimensional subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$.

Compute the quantities

$$\widehat{M}_1 = \frac{2}{n} \sum_{i=1}^{n/2} y_i x_i \text{ and } \widehat{M}_2 = \frac{2}{n} \sum_{i=1}^{n/2} y_i \left(x_i x_i^\top - I_d\right), \tag{13}$$

and let $\widehat{M} = \widehat{M}_1 \otimes \widehat{M}_1 + \widehat{M}_2$; here, I_d denotes the $d \times d$ identity matrix and $\otimes$ denotes the outer product.

Perform the eigendecomposition $\widehat{M} = \widehat{P}\widehat{\Lambda}\widehat{P}^\top$, and use the first k columns of $\widehat{P}$ (corresponding to the k largest eigenvalues) to form the matrix $\widehat{U} \in \mathbb{R}^{d \times k}$. Return $\widehat{U}$.

Our approach to the initialization problem is similar, in that we combine a moment method with search in a $\min\{k, d\}$ space. For convenience of analysis, we split the n samples into two equal parts—assume that n is even without loss of generality—and perform each of the above steps on different samples so as to maintain independence between the two steps. The formal algorithm is presented in two parts as Algorithms 2 and 3. In Algorithm 2, we address the case $k < d$; if $k \geq d$, then it suffices to return $\widehat{U} = I_d$ and proceed directly to Algorithm 3.

In related problems [32], [36], [43], [51], a combination of a second order and third order method (involving tensor decomposition) is employed to obtain parameter estimates in one shot. Take the problem of learning generalized linear models [51] as an example; here, the analysis of the moment method relies on the link function being (at least) three times differentiable so that the population moment quantities can be explicitly computed. After showing that these expectations are closed form functions of the unknown parameters, matrix/tensor perturbation tools are then applied to show that the empirical moments concentrate about their population counterparts. However, in our setting, the $\max$ function is not differentiable, and so it is not clear that higher order moments return reasonable estimates even in expectation since Stein's lemma (on which many of these results rely) is not applicable[7] in this setting. Nevertheless, we show that the second order moment returns a k-dimensional subspace that is close to the true span of the parameters $\{\theta_j^*\}_{j=1}^k$; the degree of closeness depends only on the geometric properties of these parameters.

Algorithm 3 Low-Dimensional Search

Input: Data $\{\xi_i, y_i\}_{i=1}^n$, subspace estimate $\widehat{U} \in \mathbb{R}^{d \times k \wedge d}$ having orthonormal columns that (approximately) span the $k \wedge d$ dimensional subspace $\mathsf{span}(\theta_1^*, \ldots, \theta_k^*)$, and radius of covering r.

Output: Initial estimator of parameters $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$.

Choose M points $\mathbb{M} = \{\nu^\ell, \ell = 1, \ldots, M\}$ such that they form an r-covering of the $(k \wedge d + 1)$-dimensional unit ball $\mathbb{B}^{k \wedge d + 1}$, i.e., with $\min_{\ell \in [M]} \|v - \nu^\ell\| \leq r$ for all $v \in \mathbb{B}^{k+1}$. Let

$$\widehat{V} = \begin{bmatrix} \widehat{U} & 0 \\ 0 & 1 \end{bmatrix}$$

be a matrix in $\mathbb{R}^{(d+1) \times (k \wedge d + 1)}$ having orthonormal columns.

Compute the k parameters

$$\nu_1^\sharp, ., \nu_k^\sharp \in \operatorname*{argmin}_{\nu_1, ., \nu_k \in \mathbb{M}} \frac{2}{n} \left\{ \min_{c \geq 0} \sum_{i=n/2+1}^{n} (y_i - c \max_{j \in [k]} \langle \xi_i, \widehat{V}\nu_j \rangle)^2 \right\}.$$

Return the $(d+1)$-dimensional parameters

$$\beta_j^{(0)} = \widehat{V}\nu_j^\sharp \qquad \text{for each } j \in [k].$$

Let us also briefly discuss Algorithm 3, which corresponds to performing a brute force search in $(k \wedge d + 1)$-dimensional

[7] A natural workaround is to use Stein's lemma on the infinitely differentiable "softmax" surrogate function, but our approach to this involved balancing the estimation error (which, in turn, involves derivatives of the softmax function) and approximation error terms, and led to suboptimal dependence on the dimension.

space to obtain the final initialization. First, note that a covering of the set $\mathbb{B}^{k\wedge d+1}$ can be constructed in time that is exponential in $k\wedge d$ in a variety of ways including repeated random trials. Second, note that we use the mean squared error on a holdout set (corresponding to samples $n/2+1$ through n) to select the final parameter estimates. In particular, we evaluate the error in a scale-invariant fashion; the computation of the optimal constant c in step 2 of the algorithm can be performed in closed form for each fixed choice of the tuple $(\nu_1,\ldots,\nu_k)$, since for a pair of vectors (u,v) having equal dimension, we have

$$\operatorname*{argmin}_{c\geq 0}\|u-cv\|^2=\max\left\{\frac{\langle u,\,v\rangle}{\|v\|^2},0\right\}.$$

A key parameter that governs the performance of our search procedure is the radius of the covering r, and the resulting cardinality of the covering set M. We show in the sequel that it suffices to take r depending only on $k\wedge d$ and other geometric parameters in the problem, which also bounds M independently of the ambient dimension for problems in which $k\ll d$.

Our overall algorithm should be viewed as a variant of the AM algorithm with random restarts. When the covering set $\mathbb{M}$ is generated by random sampling and k is small relative to the dimension, the algorithm inherits similar empirical performance (see panel (b) of Figure 2 to follow), while significantly reducing the computational cost, since operations are now performed in ambient dimension $k+1$, and the iterative AM algorithm is run only once overall. It also produces parameter estimates with theoretical error guarantees. Having stated the necessary background and described our methodology, we now proceed to statements and discussions of our main results.

III. Main Results

In this section, we present our main theoretical results for the methodology introduced in Section II.

A. Local Geometric Convergence of Alternating Minimization

We now establish local convergence results for the AM algorithm. Recall the definition of the parameters $(\pi_{\min},\Delta,\kappa)$ introduced in Section II, and the assumption that the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. from the standard Gaussian distribution $\mathcal{N}(0,I_d)$. Throughout the paper, we assume that the true parameters $\beta_1^*,\ldots,\beta_k^*$ are fixed.

Theorem 1: There exists a tuple of universal constants (c_1,c_2) such that if the sample size satisfies the bound

$$n\geq c_1\max\{d,10\log n\}\max\left\{\frac{k\kappa}{\pi_{\min}^3},\ \sigma^2\frac{k^5\kappa^2}{\Delta\pi_{\min}^{15}}\right\},$$

then for all initializations $\beta_1^{(0)},\ldots,\beta_k^{(0)}$ satisfying the bound

$$\min_{c>0}\max_{1\leq j\neq j'\leq k}\frac{\left\|c\left(\beta_j^{(0)}-\beta_{j'}^{(0)}\right)-\left(\beta_j^*-\beta_{j'}^*\right)\right\|}{\|\theta_j^*-\theta_{j'}^*\|}\leq c_2\frac{\pi_{\min}^6}{k^2\kappa}\log^{-3/2}\left(\frac{k^2\kappa}{\pi_{\min}^6}\right),\tag{14a}$$

the estimation error at all iterations $t\geq 1$ is simultaneously bounded as

$$\sum_{j=1}^k\|\beta_j^{(t)}-\beta_j^*\|^2\leq\left(\frac{3}{4}\right)^t\left(\sum_{j=1}^k\|c^*\beta_j^{(0)}-\beta_j^*\|^2\right)+c_1\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(kd)\log(n/kd)\tag{14b}$$

with probability exceeding $1-c_2\left(k\exp\left(-c_1 n\frac{\pi_{\min}^6}{k^2}\right)+\frac{k^2}{n^7}\right)$. Here, the positive scalar c^* minimizes the LHS of inequality (14a).

See Appendix B for a concise mathematical statement of the probability bound.

Let us interpret the various facets of Theorem 1. As mentioned before, it is a local convergence result, which requires the initialization $\beta_1^{(0)},\ldots,\beta_k^{(0)}$ to satisfy condition (14a). In the well-balanced case (with $\pi_{\min}\sim 1/k$) and treating k as a fixed constant, the initialization condition (14a) posits that the parameters are a constant "distance" from the true parameters. Notably, closeness is measured in a relative sense, and between pairwise differences of the parameter estimates as opposed to the parameters themselves; the intuition for this is that the initialization $\beta_1^{(0)},\ldots,\beta_k^{(0)}$ induces the initial partition of samples $S_1(\beta_1^{(0)},\ldots,\beta_k^{(0)}),\ldots,S_k(\beta_1^{(0)},\ldots,\beta_k^{(0)})$, whose closeness to the true partition depends only on the relative pairwise differences between parameters, and is also invariant to a global scaling of the parameters. It is also worth noting that local geometric convergence of the AM algorithm is guaranteed uniformly from *all* initializations satisfying condition (14a). In particular, the initialization parameters are not additionally required to be independent of the covariates or noise, and this allows us to use the same n samples for initialization of the parameters.

Let us now turn our attention to the bound (14b), which consists of two terms. In the limit $t\to\infty$, the final parameters provide an estimate of the true parameters that is accurate to within the second term of the bound (14b). Up to a constant, this is the statistical error term

$$\delta_{n,\sigma}(d,k,\pi_{\min})=\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(kd)\log(n/kd)\tag{15}$$

that converges to 0 as $n\to\infty$, thereby providing a consistent estimate in the large sample limit. Notice that the dependence of $\delta_{n,\sigma}(d,k,\pi_{\min})$ on the tuple (σ,d,n) is minimax-optimal up to the logarithmic factor $\log(n/d)$, since a matching lower bound can be proved for the linear regression problem when $k=1$. In Proposition 2, (see Appendix E) we also show a parametric lower bound on the minimax estimation error for general k, of the order $\sigma^2\ kd/n$. Panel (c) of Figure 1 verifies in a simulation that the statistical error depends linearly on d/n. The dependence of the statistical error on the pair $(k,\pi_{\min})$ is more involved, and we do not yet know if these are optimal. As discussed before, a linear dependence of $\pi_{\min}$ is immediate from a sample-size argument; the cubic dependence arises because the sub-matrices of Ξ chosen over the course of the algorithm are not always well-conditioned, and their condition number scales (at most) as $\pi_{\min}^2$. In Appendix E-B, we show a low-dimensional example (with $d=2$ and $k=3$)

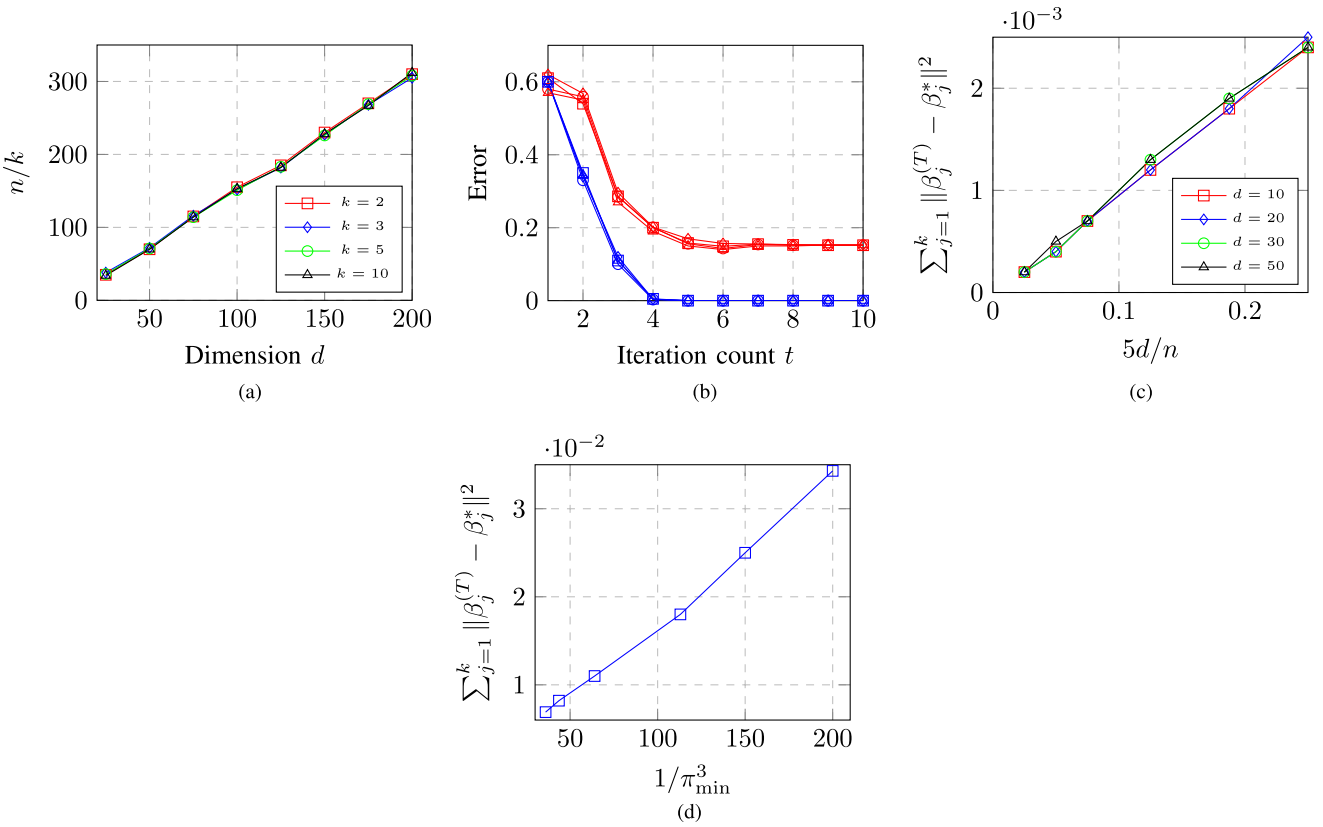

Fig. 1. Convergence of the AM with Gaussian covariates— in panel (a), we plot the noiseless sample complexity of AM; we fix $\left\|\beta_i^*\right\| = 1$ for all $i \in [k]$, $\sigma = 0$ and $\pi_{\min} = 1/k$. We say β_i^* is recovered if $\left\|\beta_i^{(t)} - \beta_i^*\right\| \leq 0.01$. For a fixed dimension d, we run a linear search on the number of samples n, such that the empirical probability of success over 100 trials is more than 0.95, and output the least such n. In panel (b), we plot the optimization error (in blue) $\sum_{j=1}^k \|\beta_j^{(t)} - \beta_j^{(T)}\|^2$ and the deviation from the true parameters (in red) $\sum_{j=1}^k \left\|\beta_j^{(t)} - \beta_j^*\right\|^2 / \sigma^2$ over iterations t for different σ $(0.15, 0.25, 0.4, 0.5)$, with $k = 5$, $d = 100$, $T = 50$ and $n = 5d$, and averaged over 50 trials. Panel (c) shows that the estimation error at $T = 50$ scales at the parametric rate d/n, where we have chosen a fixed $k = 5$ and $\sigma = 0.25$. Panel (d) shows the variation of this error as a function of $\pi_{\min}$ where we fix $k = 3, d = 2, n = 10^3, \sigma = 0.4$.

in which the least squares estimator incurs a parameter estimation error of the order $\frac{1}{\pi_{\min}^3 n}$ even when provided with the *true* partition of covariates $\{S_j(\beta_1^*, \ldots, \beta_k^*)\}_{j=1}^3$. While this does not constitute an information theoretic lower bound, it provides strong evidence to suggest that our dependence on $\pi_{\min}$ is optimal at least when viewed in isolation. We verify this intuition via simulation: in panel (d) of Figure 1, we observe that on this example, the error of the final AM iterate varies linearly with the quantity $1/\pi_{\min}^3$.

The first term of the bound (14b) is an optimization error that is best interpreted in the noiseless case $\sigma = 0$, wherein the parameters $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ converge at a geometric rate to the true parameters $\beta_1^*, \ldots, \beta_k^*$, as verified in panel (a) of Figure 1. In particular, in the noiseless case, we obtain exact recovery of the parameters provided $n \geq C \frac{kd}{\pi_{\min}^3} \log(n/d)$. Thus, the "sample complexity" of parameter recovery is linear in the dimension d, which is optimal (panel (a) verifies this fact). In the well-balanced case, the dependence on k is quartic, but lower bounds based on parameter counting suggest that the true dependence ought to be linear. Again, we are not aware of whether the dependence on $\pi_{\min}$ in the noiseless case is optimal; our simulations shown in panel (a) suggests that the sample complexity depends inversely on $\pi_{\min}$, and so closing this gap is an interesting open problem. When $\sigma > 0$, we have an overall sample size requirement

$$n \geq c \max\{d, 10 \log n\} \cdot \max\left\{\frac{k\kappa}{\pi_{\min}^3}, \sigma^2 \frac{k^5 \kappa^2}{\Delta \pi_{\min}^{15}}\right\} := n_{\mathsf{AM}}(c). \tag{16}$$

As a final remark, note that Theorem 1 holds under Gaussian covariates and when the true parameters $\beta_1^*, \ldots, \beta_k^*$ are fixed independently of the covariates. In our companion paper [1], it is shown that both of these features of the result can be relaxed, i.e., AM converges geometrically even under a milder covariate assumption, and this convergence occurs *for all* true parameters that are geometrically similar.

B. Initialization

In this section, we provide guarantees on the initialization method described in Algorithms 2 and 3 in Theorems 2 and 3, respectively.

Consider the matrices $\widehat{U}$ and $\widehat{M}$ defined in Algorithm 2. Algorithm 2 is a moment method: we extract the top k principal components of a carefully chosen moment statistic of the data to obtain a subspace estimate $\widehat{U}$. Spectral algorithms such as these have been used to obtain initializations in a wide variety of non-convex problems [43], [52], [53] to obtain an

accurate estimate of the subspace spanned by the unknown parameters. It is well-known that the performance of the algorithm in recovering a k-dimensional subspace depends on $\lambda_k(\mathbb{E}[\widehat{M}])$, which is the k-th largest eigenvalue of the population moment $\mathbb{E}[\widehat{M}]$. We show in the proof (see the discussion following Lemma 7) that there is a strictly positive scalar γ such that

$$\lambda_k(\mathbb{E}[\widehat{M}]) \geq \gamma. \tag{17}$$

It should be stressed that we obtain an explicit expression for γ as a function of the various problem parameters (in equation (49) of the proof) that is, a priori,[8] independent of the ambient dimension d.

This characterization is the main novelty of our contribution, and allows us to establish the following guarantee on the PCA algorithm. We let $U^* \in \mathbb{R}^{d\times k}$ denote a matrix whose orthonormal columns span the linear subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$, and define the quantity

$$\varsigma := \max_{j\in[k]} \left\{ \|\theta_j^*\|_1 + |b_j^*| \right\}. \tag{18}$$

Theorem 2: There is a universal constant C such that $\widehat{U}$ satisfies the bound

$$\|\widehat{U}\widehat{U}^\top - U^*(U^*)^\top\|_{\mathrm{F}}^2 \leq C\left(\frac{\sigma^2+\varsigma^2}{\gamma^2}\right)\frac{kd\log^3(nk)}{n}$$

with probability greater than $1-Cn^{-10}$.

The proof of Theorem 2 is provided in Appendix C. We have thus shown that the projection matrix $U^*(U^*)^\top$ onto the true subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$ can be estimated at the parametric rate via our PCA procedure. Note that this is useful when $k \leq d$, since otherwise we have $\widehat{U} = U^* = I_d$. The guarantee of this theorem is illustrated via simulation in panel (a) of Figure 2.

Let us now turn to establishing a guarantee on Algorithm 3 when it is given a (generic) subspace estimate $\widehat{U}$ as input. Since the model (1) is only identifiable up to a relabeling of the individual parameters, we can only hope to show that a suitably permuted set of the initial parameters is close to the true parameters. Toward that end, let $\mathcal{P}_k$ denote the set of all permutations from $[k] \to [k]$, and let

$$\mathsf{dist}\left(\left\{\beta_j^{(0)}\right\}_{j=1}^k, \left\{\beta_j^*\right\}_{j=1}^k\right) := \min_{P\in\mathcal{P}_k}\sum_{j=1}^k \|\beta_{P(j)}^{(0)} - \beta_j^*\|^2 \tag{19}$$

denote the minimum distance attainable via a relabeling of the parameters. With this notation in place, we are now ready to state our result for parameter initialization. In it, we assume that the input matrix $\widehat{U}$ is fixed independently of the samples used to carry out the search procedure; again, recall that $\widehat{U} = U^* = I_d$ if $k > d$.

[8]While this may seem surprising—after all, the unknown parameters $\theta_1^*, \ldots, \theta_k^*$ live in dimension d—all the interesting action is confined to the k dimensional subspace spanned by these parameters and γ is a function of the geometry induced by the parameters on this subspace.

Theorem 3: Let $\bar{k} = k \wedge d$. Suppose we set $0 \leq r \leq \frac{\Delta\pi_{\min}^{5/2}\log^{-1/2}(k/\pi_{min})}{8\mathsf{B}_{\max}k^3}$, that

$$\|\widehat{U}\widehat{U}^\top - U^*(U^*)^\top\|_{\mathrm{op}} \leq \frac{\Delta\pi_{\min}^{3/2}}{8\mathsf{B}_{\max}k^2}$$

and note that it suffices to set $M = \left(1+\frac{1}{r}\right)^{\bar{k}}$. Then there is a tuple of universal constants (c_1, c_2) such that if

$$n \geq c_1 \max\left\{ d\frac{k^3}{\pi_{\min}^3}\log^2(\pi_{\min}/k), \sigma^2\frac{k^3}{\pi_{\min}^3\Delta^2}\log M\right\},$$

then

$$\begin{aligned}&\min_{c>0}\ \mathsf{dist}\left(\left\{c\beta_j^{(0)}\right\}_{j=1}^k, \left\{\beta_j^*\right\}_{j=1}^k\right)\\ &\leq c_1\left(\frac{k}{\pi_{\min}}\right)^3\left\{4k\mathsf{B}_{\max}^2\left(r^2 + \|\widehat{U}\widehat{U}^\top - U^*(U^*)^\top\|_{\mathrm{op}}^2\right) + \frac{\sigma^2\log M}{n}\right\}\end{aligned}$$

with probability exceeding $1-c_1k\exp\left(-c_2n\frac{\pi_{\min}^4}{k^4\log^2(k/\pi_{\min})}\right)$.

We prove Theorem 3 in Appendix D. Combining Theorems 2 and 3 with some algebra then allows us to prove a guarantee for the initialization procedure that combines Algorithms 2 and 3 in sequence. In particular, fix a positive scalar $\epsilon \leq \Delta$. Then combining the theorems shows that if (for an appropriately large universal constant c), we have $M = \left(1 + c\frac{\mathsf{B}_{\max}k^3\log^{1/2}(k/\pi_{\min})}{\epsilon\pi_{\min}^{5/2}}\right)^k$, and the sample size n is greater than

$$\begin{aligned}n_{\mathsf{init}}(\epsilon, M, c) := c\max\Bigg\{&\frac{dk}{\pi_{\min}}, \frac{\sigma^2k^5}{\pi_{\min}^5\epsilon^2}\log(\frac{k}{\pi_{\min}})\log(\frac{M}{\delta}),\\ &d\log^3(nk)\log(\frac{k}{\pi_{\min}})\frac{k^7\mathsf{B}_{\max}^2}{\gamma^2\pi_{\min}^5\epsilon^2}(\sigma^2+\varsigma^2)\Bigg\},\end{aligned} \tag{20}$$

then $\min_{c>0}\ \mathsf{dist}\left(\left\{c\beta_j^{(0)}\right\}_{j=1}^k, \left\{\beta_j^*\right\}_{j=1}^k\right) \leq \epsilon^2$ with probability greater than $1-cn^{-10}$. Equipped with this guarantee on our initialization step, we are now in a position to state an end-to-end guarantee on our overall methodology in the next section.

C. Overall Algorithmic Guarantee

Assume without loss of generality that the identity permutation minimizes the distance measure dist, so that $\beta_j^{(0)}$ is the estimate of the parameter β_j^* for each $j \in [k]$. Recall the statistical error $\delta_{n,\sigma}(d,k,\pi_{\min})$ defined in equation (15), which is, up to a constant factor, the final (squared) radius of the ball to which the AM update converges when initialized suitably, and the notation $n_{\mathsf{AM}}(c)$ and $n_{\mathsf{init}}(\epsilon, M, c)$ from equations (16) and (20), respectively. We now state a guarantee for our overall procedure that runs Algorithms 2, 3, and 1 in that sequence; we omit the proof since it follows by simply putting together the pieces from Theorem 1 and the discussion above.

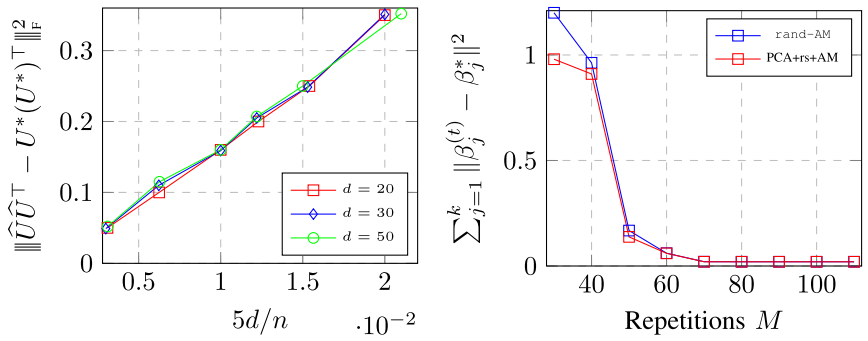

Fig. 2. Simulation of the PCA and overall guarantees. We assume that the true parameter matrix $\Theta^* = A^*(U^*)^\top$ for a $\mathbb{R}^{d\times k}$ matrix U^* and an invertible $A^* \in \mathbb{R}^{k\times k}$, and that Algorithm 2 returns a subspace estimate $\widehat{U}$. Panel (a) reveals the subspace estimation error as a function of d/n, which is corroborated by Theorem 2. In panel (b), we compare the performance of our overall algorithm (in red) with that of AM with repeated random initialization [12] (in blue) averaged over 50 trials. We fix $k = 3$, $d = 50$, $n = 35kd$ and $\sigma = 0.1$. For a sufficiently large M, both schemes perform in a similar fashion.

Corollary 1: Let $\bar{k} = k \wedge d$. There exist universal constants c_1 and c_2 such that with

$$M = \left(1 + c_1 \frac{\mathsf{B}_{\max}\bar{k}^4 \log^{1/2}(k/\pi_{\min})}{\pi_{\min}^{11/2}}\right)^{\bar{k}},$$
$$n \geq \max\left\{n_{\mathsf{init}}\left(c_2\frac{\pi_{\min}^3}{k}, c_1, M\right), n_{\mathsf{AM}}(c_1)\right\}$$
$$\text{and} \quad T_0 = c_1 \log\left(\frac{1}{\delta_{n,\sigma}(d,k,\pi_{\min})}\right),$$

then the combined algorithm satisfies, simultaneously for all $T \geq T_0$, the bound

$$\Pr\left\{\sum_{j=1}^k \|\beta_j^{(T)} - \beta_j^*\|^2 \geq c_1\delta_{n,\sigma}(d,k,\pi_{\min})\right\}$$
$$\leq c_1\left(n^{-10} + k\exp\left(-c_2 n\frac{\pi_{\min}^6}{k^2}\right) + \frac{k^2}{n^7}\right).$$

We thus obtain, an algorithm that when given a number of samples that is near-linear in the ambient dimension, achieves the rate $\delta_{n,\sigma}(d,k,\pi_{\min}) = \frac{\sigma^2 kd}{\pi_{\min}^3 n}\log(kd)\log(n/kd)$ of estimation of all kd parameters in squared ℓ_2 norm. This convergence is illustrated in simulation in Figure 2, in which we choose $k = 3$, $d = 50$ and $n = 35kd$. Interestingly, panel (b) of this figure shows that our provable multi-step algorithm has performance similar to the algorithm that runs AM with repeated random initializations.

The computational complexity of our overall algorithm (with exact matrix inversions) is given by $\mathcal{O}\left(knd^2\log\left(\frac{1}{\delta_{n,\sigma}(d,k,\pi_{\min})}\right) + \binom{M}{k}\cdot nd\right)$, where we also assume that the k top eigenvectors of the matrix $\widehat{M}$ are computed exactly in Algorithm 2. This guarantee can also be extended to the case where the linear system is solved up to some numerical precision by (say) a conjugate gradient method and the eigenvectors of $\widehat{M}$ are computed using the power method, thereby reducing the computational complexity. Such an extension is standard and we do not detail it here.

D. Proof Sketch and Technical Challenges

Let us first sketch, at a high level, the ideas required to establish guarantees on the AM algorithm. We need to control the iterates of the AM algorithm without sample-splitting across iterations, and so the iterates themselves are random and depend on the sequence of random variables $(\xi_i, \epsilon_i)_{i=1}^n$. A popular and recent approach to handling this issue in related iterative algorithms (e.g., [25]) goes through two steps: first, the population update, corresponding to running (12a)-(12b) in the case $n \to \infty$, is analyzed, after which the random iterates in the finite-sample case are shown to be close to their (non-random) population counterparts by using concentration bounds for the associated empirical process. The main challenge in our setting is that the population update is quite non-trivial to write down since it involves a delicate understanding of the geometry of the covariate distribution induced by the maxima of affine functions. We thus resort to handling the random iterates directly, thereby sidestepping the calculation of the population operator entirely.

In order to convey the principal difficulties associated with our approach, let us present a bound on the error obtained after running a single step of the algorithm, starting at the parameters $\beta_1, \ldots, \beta_k$ and obtaining, as a result of one step of the algorithm, the parameters $\beta_1^+, \ldots, \beta_k^+$. We use the shorthand notation $S_j := S_j(\beta_1, \ldots, \beta_k)$, and let $P_{\Xi^j(\beta_1,\ldots,\beta_k)}$ denote the projection matrix onto the range of the matrix Ξ_{S_j}.

Let y^* denote the vector with entry i given by $\max_{\ell\in[k]}\langle\xi_i, \beta_\ell^*\rangle$. We have

$$\begin{aligned}
&\|\Xi_{S_j}(\beta_j^+ - \beta_j^*)\|^2 = \|P_{\Xi^j(\beta_1,\ldots,\beta_k)}y_{S_j} - \Xi_{S_j}\beta_j^*\|^2 \\
&= \|P_{\Xi^j(\beta_1,\ldots,\beta_k)}y_{S_j}^* + P_{\Xi^j(\beta_1,\ldots,\beta_k)}\epsilon_{S_j} - \Xi_{S_j}\beta_j^*\|^2 \\
&\leq 2\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}(y_{S_j}^* - \Xi_{S_j}\beta_j^*)\|^2 + 2\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}\epsilon_{S_j}\|^2 \\
&\leq 2\|y_{S_j}^* - \Xi_{S_j}\beta_j^*\|^2 + 2\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}\epsilon_{S_j}\|^2, \qquad (21)
\end{aligned}$$

where we have used the fact that the projection operator is non-expansive on a convex set.

Let

$$\{\langle\xi_i, \beta_\ell\rangle = \max\} := \left\{\langle\xi_i, \beta_\ell\rangle = \max_{u\in[k]}\langle\xi_i, \beta_u\rangle\right\}$$

for each $i \in [n]$, $\ell \in [k]$ denote a convenient shorthand for these events. The first term on the RHS of inequality (21) can be written as

$$\sum_{i \in S_j} (y_i^* - \langle \xi_i, \beta_j^* \rangle)^2$$
$$\leq \sum_{i=1}^{n} \sum_{j' : j' \neq j} \mathbf{1}\left\{ \langle \xi_i, \beta_j \rangle = \max \text{ and } \langle \xi_i, \beta_{j'}^* \rangle = \max \right\} \times \langle \xi_i, \beta_{j'}^* - \beta_j^* \rangle^2,$$

where the inequality accounts for ties. Each indicator random variable is bounded, in turn, as

$$\mathbf{1}\left\{ \langle \xi_i, \beta_j \rangle = \max \text{ and } \langle \xi_i, \beta_{j'}^* \rangle = \max \right\}$$
$$\leq \mathbf{1}\left\{ \langle \xi_i, \beta_j \rangle \geq \langle \xi_i, \beta_{j'} \rangle \text{ and } \langle \xi_i, \beta_{j'}^* \rangle \geq \langle \xi_i, \beta_j^* \rangle \right\}$$
$$\leq \mathbf{1}\left\{ \langle \xi_i, \beta_j - \beta_{j'} \rangle \cdot \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle \leq 0 \right\}.$$

Switching the order of summation yields the bound

$$\sum_{i \in S_j} (y_i^* - \langle \xi_i, \beta_j^* \rangle)^2$$
$$\leq \sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i, \beta_j - \beta_{j'} \rangle \cdot \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle \leq 0 \right\} \times \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle^2.$$

Recalling our notation for the minimum eigenvalue of a symmetric matrix, the LHS of inequality (21) can be bounded as

$$\|\Xi_{S_j}(\beta_j^+ - \beta_j^*)\|^2 \geq \lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j} \right) \cdot \|\beta_j^+ - \beta_j^*\|^2.$$

Putting together the pieces yields, for each $j \in [k]$, the pointwise bound

$$\frac{1}{2}\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j} \right) \cdot \|\beta_j^+ - \beta_j^*\|^2$$
$$\leq \sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i, \beta_j - \beta_{j'} \rangle \cdot \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle \leq 0 \right\} \times \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle^2 + \|P_{\Xi^j(\beta_1, \ldots, \beta_k)} \epsilon_{S_j}\|^2. \quad (22)$$

Up to this point, note that all steps of the proof were deterministic. Observe from equation (22) that in order to obtain an error bound on the next parameter, we need to control three distinct quantities: (a) the noise term $\|P_{\Xi^j(\beta_1, \ldots, \beta_k)} \epsilon_{S_j}\|^2$, (b) the prediction error of the noiseless problem, given by a pairwise sum of terms of the form $\mathbf{1}\left\{ \langle \xi_i, \beta_j - \beta_{j'} \rangle \cdot \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle \leq 0 \right\} \langle \xi_i, \beta_j^* - \beta_{j'}^* \rangle^2$, and (c) the minimum eigenvalue of the covariate matrix restricted to the set S_j, denoted by $\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j} \right)$. Since the set S_j is in itself random and depends on the pair (Ξ, ϵ) (since the current parameters were obtained over the course of running the algorithm), obtaining such a bound is especially challenging.

For step (a)—handled by Lemma 3—we apply standard concentration bounds for quadratic forms of sub-Gaussian random variables in conjunction with bounds on the *growth functions* of multi-class classifiers [54]. Crucially, this affords a uniform bound on the noise irrespective of which iterate the alternating minimization update is run from. To show step (b)—in Lemma 4—we generalize a result of Waldspurger [23]. Finally, the key difficulty in step (c) is to control the spectrum of random matrices, rows of which are drawn from (randomly) truncated variants of the Gaussian distribution. The expectation of such a random matrix can be characterized by appealing to tail bounds on the non-central χ^2 distribution, and the Gaussian covariate assumption additionally allows us to show that an analogous result holds for the random matrix with high probability (see Lemma 5). Here, our initialization condition is crucial: the aforementioned singular value control suffices for the sub-matrices formed by the *true* parameters, and we translate these bounds to the sub-matrices generated by random parameters by appealing to the fact that the initialization is sufficiently close to the truth.

Having discussed our proof of the AM update in some detail, let us now turn to a brief discussion of the techniques used to prove Theorems 2 and 3. As mentioned before, our proof of Theorem 2 relies on a lower bound on the eigengap of the population moment. We obtain such a lower bound by appealing to classical moment calculations for suitably truncated Gaussian distributions [45]. Translating these calculations into an eigengap is quite technical, and involves the isolation of many properties of the population moments that may be of independent interest. As briefly alluded to in Section II, the heart of the technical difficulty is due to the fact that $\max$ function is not differentiable, and so moments cannot be calculated by repeated applications of Stein's lemma like in related problems [43], [52], [55].

In order to establish Theorem 3, we crucially use the scale-invariance property of the initialization along with some arguments involving empirical process theory to show that the goodness-of-fit statistic employed in the algorithm is able to isolate a good initialization. Establishing these bounds requires us to relate the prediction and estimation errors in the problem (in Lemma 16), which may be of independent interest.

IV. Discussion

We conclude this portion of the paper with short discussions of prediction error guarantees, a comparison with adaptivity in convex regression, related models, and future directions.

A. Guarantees on Prediction Error

While our principal focus in this paper was on estimation of the unknown parameters $\{\theta_j^*, b_j^*\}_{j=1}^k$, the complementary question of prediction error is also interesting and important. In particular, suppose that we produce the max-affine function estimate $\widehat{\phi}^{(\mathsf{MA})}$ given by $\widehat{\phi}^{(\mathsf{MA})}(x) := \max_{j \in [k]} (\langle x, \widehat{\theta}_j \rangle + \widehat{b}_j)$ for each $x \in \mathbb{R}^d$, and measure its performance via the prediction error

$$\frac{1}{n} \sum_{i=1}^{n} (\widehat{\phi}^{(\mathsf{ls})}(x_i) - \phi^*(x_i))^2, \quad (23)$$

where $\phi^*(x) = \max_{j \in [k]} (\langle x, \theta_j^* \rangle + b_j^*)$ denotes the "true" function. When ϕ^* belongs to the sub-class of k-piece affine functions induced by parameters in the set $\mathsf{B}_{\mathsf{vol}}(\pi_{\min}, \Delta, \kappa)$

and the covariates are drawn from a Gaussian distribution, our results imply (via Theorem 1, and by using Lemma 16 to translate our estimation error guarantee into a prediction error guarantee) the rate

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\phi}^{(\mathsf{MA})}(x_i)-\phi^*(x_i))^2 \le C(\pi_{\min},\Delta,\kappa)\frac{kd}{n} \times \log(kd)\log(n/kd). \quad (24)$$

At least in principle, an explicit dependence on $\pi_{\min}$ should not be expected in the prediction error, since if a particular pair of parameters (θ, b) attains the maximum extremely rarely (resulting in a small value of $\pi_{\min}$), then we may simply drop these parameters from the estimate (and estimate the function with the maximum of the remaining $k-1$ pieces) without affecting the prediction error significantly. Indeed the minimax risk of prediction (without any requirements of computational efficiency) is known to be independent of the geometry of the problem instance (see, e.g., [16]).

We also note that polynomial-time algorithms with small prediction error are known, without any dependence on $\pi_{\min}$. In particular, [56, Theorem 1.8] shows that the sample complexity for obtaining ϵ-accurate estimates in prediction error is bounded by $n \le \exp\left\{c_1(k/\epsilon)^{\log k}\right\} d^{c_2}$ for absolute constants c_1 and c_2. While the dependence on both ϵ and d can likely be improved,[9] these results provide additional evidence that the prediction error is much less sensitive to the geometry of the instance than the estimation error considered in this paper.

B. Comparison With Algorithms for Convex Regression

As mentioned earlier, the most standard estimator in convex regression is the convex least squares estimator $\widehat{\phi}^{(\mathsf{ls})}$ defined as in (6) which can be computed efficiently as shown in [19], [21]. The performance of $\widehat{\phi}^{(\mathsf{ls})}$ in the max-affine regression model (1) has been the subject of some interest in the literature on adaptivity of shape-constrained estimators (see [57] for an overview of results of this type). These results mainly focus on the prediction error:

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\phi}^{(\mathsf{ls})}(x_i)-\phi^*(x_i))^2$$

as opposed to estimation of the parameters $\theta_j^*, b_j^*, j=1,\ldots,k$ which is our main focus. There is actually no natural way of obtaining parameter estimates from $\widehat{\phi}^{(\mathsf{ls})}$ as $\widehat{\phi}^{(\mathsf{ls})}$ will typically be a maximum of a strictly larger than k number of affine functions. Let us now compare our results with the existing results on the prediction error of the convex least squares estimator. When $d=1$, it has been showed by [58] and [59] that

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\phi}^{(\mathsf{ls})}(x_i)-\phi^*(x_i))^2 \le \frac{k\log n}{n}$$

with high probability assuming that $\{x_i\}_{i=1}^n$ are uniformly spaced on the interval $[0,1]$. For $d \ge 2$, Han and Wellner [9] studied the adaptivity properties of $\widehat{\phi}^{(\mathsf{bls})}$ which is the least squares estimator over the class of *bounded* convex functions which is different from $\widehat{\phi}^{(\mathsf{ls})}$ and computationally tricky to compute. However, [9] showed that unless $d \le 4$,

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\phi}^{(\mathsf{bls})}(x_i)-\phi^*(x_i))^2 \le C_m n^{-4/d}(\log n)^{d+4}$$

with high probability assuming that the covariates are drawn from a distribution supported on a convex polytope with m simplices (the constant pre-factor C_m depends on m). Comparing these two results with our result on the prediction error (24), we see that when $d=1$, our results in the prediction error are strictly weaker than those of prior work [58], [59], but as soon as $d \ge 2$, they are significantly stronger than existing adaptivity results [9], at least for a sub-class of k-affine functions. We emphasize once again that the focus of the body of work differs from ours, and so the comparison presented above is necessarily incomplete.

A parallel line of work (including our own) eschews the c-LSE (and its variants) entirely and pursues a different avenue to alleviate the curse of dimensionality,[10] by directly fitting convex functions consisting of a certain number of affine pieces [10], or more broadly, by treating the number of affine functions as a tuning parameter to be chosen in a data-dependent fashion via cross-validation [11]. Hannah and Dunson [11] showed that performing estimation under a carefully chosen sequence of models of the form (1) via their "convex adaptive partitioning", or CAP estimator is able to obtain consistent prediction rates for general convex regression problems. However, it is unclear if the CAP estimator is able to avoid the curse of dimensionality in the special case when the true function is k-piece affine.

C. Related Models

Models closely related to (1) also appear in second price auctions, where an item having d features is bid on and sold to the highest bidder at the second highest bid [60], [61]. Assuming that each of k user groups bids on an item and that each bid is a linear function of the features, one can use a variant of the model (1) with the $\max$ function replaced by the second order statistic to estimate the individual bids of the user groups based on historical data. Another related problem is that of multi-class classification [54], in which one of k labels is assigned to each sample based on the argmax function, i.e., for a class of functions $\mathcal{F}$, we have the model $Y = \operatorname{argmax}_{1\le j\le k} f_j(X)$ for j distinct functions $f_1,\ldots,f_k \in \mathcal{F}$. When $\mathcal{F}$ is the class of linear functions based on d features, this can be viewed as the "classification" variant of our regression problem.

The model (1) can also be seen as a special case of multi-index models [62], [63] as well as mixture-of-experts models [64], [65]. Multi-index models are of the form $Y = g(\langle\theta_1^*, X\rangle,\ldots,\langle\theta_k^*, X\rangle) + \epsilon$ for an *unknown* function g and

[9]Note that unlike our paper, this work makes only boundedness assumptions on the covariates, and their focus is not on achieving the optimal dimension/sample size dependence.

[10]Note that setting $k \sim n^{d/(4+d)}$, we can (essentially) recover the entire class of convex functions from the maxima of k affine functions (see, e.g., Balázs [12]), so interesting parametric structure is only expected to emerge when k is essentially constant, or grows very slowly with n.

this function g is taken to be the $\max(\cdot)$ function in the model (1). In the mixture-of-experts model, the covariate space is partitioned into k regions via certain *gating functions*, and the observation model is given by k distinct regression functions: one on each region. The model (1) is clearly a member of this class, since the $\max(\cdot)$ function implicitly defines a partition of $\mathbb{R}^d$ depending on which of the k linear functions of X attains the maximum, and on each of these partitions, the regression function is linear in X.

D. Future Directions

In this paper, we analyzed a natural alternating minimization algorithm for estimating the maximum of unknown affine functions, and established that it enjoys local linear convergence to a ball around the optimal parameters. We also proposed an initialization based on PCA followed by random search in a lower-dimensional space. An interesting open question is if there are other efficient methods besides random search that work just as well post dimensionality reduction. Another interesting question has to do with the necessity of dimensionality reduction: in simulations (see, e.g., Figure 2), we have observed that if the AM algorithm is repeatedly initialized in $(d+1)$-dimensional space without dimensionality reduction, then the number of repetitions required to obtain an initialization from which it succeeds (with high probability) is similar to the number of repetitions required after dimensionality reduction. This suggests that our (sufficient) initialization condition (14a) may be too stringent, and that the necessary conditions on the initialization to ensure convergence of the AM algorithm are actually much weaker. We leave such a characterization for future work, but note that some such conditions must exist: the AM algorithm when run from a single random initialization, for instance, fails with constant probability when $k \geq 3$. Understanding the behavior of the randomly initialized AM algorithm is also an open problem in the context of phase retrieval [23], [66].

We note that once again that the Gaussian assumption made in this paper for convenience of analysis can be relaxed to allow (for instance) log-concave covariate distributions, which includes the uniform distribution on $[-1, 1]^d$ common in nonparametric statistics. Such an extension requires significant technical effort and the structure of the proof also changes slightly; simultaneously, the dependence of the eventual error bounds on the parameter $\pi_{\min}$ is also different in the more general setting. In particular, Lemmas 4-6 in the current paper must be extended, and this requires, among other things, an analysis of random matrices whose rows are drawn from a (truncated) small-ball distribution. Our companion paper [1] is also concerned with the *universal* setting in which guarantees are proved uniformly over all choices of parameters once the covariates have been drawn, in contrast to the setting of the current paper in which parameters are fixed in advance. Universal guarantees are commonly sought out in statistical signal processing applications, including phase retrieval [53].

In the broader context of max-affine estimation, it is also interesting to analyze other non-convex procedures (e.g. gradient descent) to obtain conditions under which they obtain accurate parameter estimates. The CAP estimator of Hannah and Dunson [11] and the adaptive max-affine partitioning algorithm of Balázs [12] are also interesting procedures for estimation under these models, and it would be interesting to analyze their performance when the number of affine pieces k is fixed and known. For applications in which the dimension d is very large, it is also interesting to study the model with additional restrictions of sparsity on the unknown parameters—such problems are known to exhibit interesting statistical-computational gaps even in the special case of sparse phase retrieval (see, e.g., Cai *et al.* [67]). We also note that in practice, and especially for convex regression, the parameter k would be unknown and must be estimated (say) via cross-validation, and understanding such a data-driven estimator is an important direction of future work.

We now present proofs of our main results. We assume throughout that the sample size n is larger than some universal constant. Values of constants $c, c_1, c', \ldots$ may change from line to line. Statements of our theorems, for instance, minimize the number of constants by typically using one of these to denote a large enough constant, and another to denote a small enough constant.

APPENDIX A
TECHNICAL RESULTS CONCERNING THE GLOBAL LSE

In this section, we provide a proof of the existence of the global least squares estimator that was stated in the main text. We also state and prove a lemma that shows that the global LSE is a fixed point of the AM update under a mild technical condition.

A. Proof of Lemma 1

Fix data $(x_1, y_1), \ldots, (x_n, y_n)$ and let

$$L(\gamma_1, \ldots, \gamma_k) := \sum_{i=1}^{n} \left(y_i - \max_{j \in [k]} \langle \xi_i, \gamma_j \rangle \right)^2$$

denote the objective function in (5) with $\xi_i := (x_i, 1)$. The goal is to show that a global minimizer of $L(\gamma_1, \ldots, \gamma_k)$ over $\gamma_1, \ldots, \gamma_k \in \mathbb{R}^{d+1}$ exists. For $\gamma_1, \ldots, \gamma_k \in \mathbb{R}^{d+1}$, let $S_1^\gamma, \ldots, S_k^\gamma$ denote a fixed partition of $[n]$ having the property that

$$\langle \xi_i, \gamma_j \rangle = \max_{u \in [k]} \langle \xi_i, \gamma_u \rangle \qquad \text{for every } j \in [k] \text{ and } i \in S_j^\gamma.$$

Also, let $\widehat{\beta}_1^\gamma, \ldots, \widehat{\beta}_k^\gamma$ denote the solution to the following constrained least squares problem:

$$\begin{aligned} \underset{\beta_1, \ldots, \beta_k}{\text{minimize}} \quad & \sum_{j=1}^{k} \sum_{i \in S_j^\gamma} (y_i - \langle \xi_i, \beta_j \rangle)^2 \\ \text{subject to} \quad & \langle \xi_i, \beta_j \rangle \geq \langle \xi_i, \beta_u \rangle, u, j \in [k], i \in S_j^\gamma. \end{aligned}$$

Note that the above quadratic problem is feasible as $\gamma_1, \ldots, \gamma_k$ satisfies the constraint and, consequently, $\widehat{\beta}_1^\gamma, \ldots, \widehat{\beta}_k^\gamma$ exists uniquely for every $\gamma_1, \ldots, \gamma_k \in \mathbb{R}^{d+1}$. Note further that, by construction,

$$L\left(\widehat{\beta}_1^\gamma, \ldots, \widehat{\beta}_k^\gamma\right) \leq L(\gamma_1, \ldots, \gamma_k).$$

and that the set

$$\Delta := \left\{ (\widehat{\beta}_1^{\gamma}, \ldots, \widehat{\beta}_k^{\gamma}) : \gamma_1, \ldots, \gamma_k \in \mathbb{R}^{d+1} \right\}$$

is finite because $\widehat{\beta}_1^{\gamma}, \ldots, \widehat{\beta}_k^{\gamma}$ depends on $\gamma_1, \ldots, \gamma_k$ only through the partition $S_1^{\gamma}, \ldots, S_k^{\gamma}$ and the number of possible such partitions of $[n]$ is obviously finite. Finally, it is evident that

$$(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) = \operatorname*{argmin}_{(\beta_1, \ldots, \beta_k) \in \Delta} L(\beta_1, \ldots, \beta_k)$$

is a global minimizer of $L(\gamma_1, \ldots, \gamma_k)$ as

$$L\left(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}\right) \leq L\left(\widehat{\beta}_1^{\gamma}, \ldots, \widehat{\beta}_k^{\gamma}\right) \leq L(\gamma_1, \ldots, \gamma_k)$$

for every $\gamma_1, \ldots, \gamma_k$. This concludes the proof of Lemma 1.

B. Fixed Point of AM Update

The following lemma establishes that the global LSE is a fixed point of the AM update under a mild technical condition.

Lemma 2: Consider the global least squares estimator (5). Suppose that the k values $\langle \xi_i, \widehat{\beta}_j^{\mathsf{ls}} \rangle$ for $j = 1, \ldots, k$ are distinct for each $i \in [n]$. Then

$$\widehat{\beta}_j^{(\mathsf{ls})} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \sum_{i \in S_j(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} (y_i - \langle \xi_i, \beta \rangle)^2 \text{ for every} j \in [k]. \tag{25}$$

Proof: It is clearly enough to prove (25) for $j = 1$. Suppose that $\widehat{\beta}_1^{(\mathsf{ls})}$ does not minimize the least squares criterion over $S_1(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})})$. Let

$$\widehat{\gamma}_1^{(\mathsf{ls})} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \sum_{i \in S_1(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \beta \rangle\right)^2$$

be any other least squares minimizer over $S_1(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})})$ and let, for $\epsilon > 0$,

$$\widetilde{\beta}_1 := \widehat{\beta}_1^{(\mathsf{ls})} + \epsilon \left(\widehat{\gamma}_1^{(\mathsf{ls})} - \widehat{\beta}_1^{(\mathsf{ls})} \right).$$

When $\epsilon > 0$ is sufficiently small, we have

$$S_j(\widetilde{\beta}_1, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) = S_j(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) \text{ for every } j \in [k]$$

due to the no ties assumption and the fact that $\widetilde{\beta}_1$ and $\widehat{\beta}_1^{(\mathsf{ls})}$ can be made arbitrarily close as ϵ becomes small. Thus, if

$$\begin{aligned} U(\beta_1, \ldots, \beta_k) &:= \sum_{i=1}^{n} \left(y_i - \max_{j \in [k]} \langle \xi_i, \beta_j \rangle \right)^2 \\ &= \sum_{j \in [k]} \sum_{i \in S_j(\beta_1, \ldots, \beta_k)} \left(y_i - \langle \xi_i, \beta_j \rangle\right)^2, \end{aligned}$$

then

$$\begin{aligned} U(\widetilde{\beta}_1, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) &= \sum_{i \in S_1(\widetilde{\beta}_1, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widetilde{\beta}_1 \rangle \right)^2 \\ &+ \sum_{j \geq 2} \sum_{i \in S_j(\widetilde{\beta}_1, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widehat{\beta}_j^{(\mathsf{ls})} \rangle \right)^2 \\ &= \sum_{i \in S_1(\widehat{\beta}_1^{(\mathsf{ls})}, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widetilde{\beta}_1 \rangle \right)^2 \\ &+ \sum_{j \geq 2} \sum_{i \in S_j(\widehat{\beta}_1^{(\mathsf{ls})}, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widehat{\beta}_j^{(\mathsf{ls})} \rangle \right)^2 \\ &< \sum_{i \in S_1(\widehat{\beta}_1^{(\mathsf{ls})}, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widehat{\beta}_1^{(\mathsf{ls})} \rangle \right)^2 \\ &+ \sum_{j \geq 2} \sum_{i \in S_j(\widehat{\beta}_1^{(\mathsf{ls})}, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})})} \left(y_i - \langle \xi_i, \widehat{\beta}_j^{(\mathsf{ls})} \rangle \right)^2 \\ &= U(\widehat{\beta}_1^{(\mathsf{ls})}, \widehat{\beta}_2^{(\mathsf{ls})} \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) \end{aligned}$$

where the strict inequality above comes from the fact that $\widetilde{\beta}_1$ is closer to the least squares solution $\widehat{\gamma}_1^{(\mathsf{ls})}$ compared to $\widehat{\beta}_1^{(\mathsf{ls})}$. This leads to a contradiction as the criterion function is smaller than its value at a global minimizer, thereby concluding the proof. □

APPENDIX B
PROOF OF THEOREM 1

Let us begin by introducing some shorthand notation, and providing a formal statement of the probability bound guaranteed by the theorem. For a scalar w^*, vectors $u^* \in \mathbb{R}^d$ and $v^* = (u^*, w^*) \in \mathbb{R}^{d+1}$, and a positive scalar r, let $\mathcal{B}_{v^*}(r) = \left\{ v \in \mathbb{R}^{d+1} : \frac{\|v - v^*\|}{\|u^*\|} \leq r \right\}$, and let

$$\begin{aligned} &\mathcal{I}\left(r; \{\beta_j^*\}_{j=1}^k\right) \\ &= \left\{ \beta_1, \ldots \beta_k \in \mathbb{R}^{d+1} : \exists c > 0 : c(\beta_i - \beta_j) \in \mathcal{B}_{\beta_i^* - \beta_j^*}(r) \right\} \end{aligned}$$

for all $1 \leq i \neq j \leq k$. Also, use the shorthand

$$\begin{aligned} \vartheta_t\left(r; \{\beta_j^*\}_{j=1}^k\right) &:= \sup_{\beta_1^{(0)}, \ldots, \beta_k^{(0)} \in \mathcal{I}(r)} \sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^*\|^2 \\ &\quad - \left(\frac{3}{4}\right)^t \left(\sum_{j=1}^{k} \|c^* \beta_j^{(0)} - \beta_j^*\|^2 \right), \text{ and} \\ \delta_{n,\sigma}(d, k, \pi_{\min}) &:= \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(kd) \log(n/kd) \end{aligned}$$

to denote the error tracked over iterations (with c^* denoting the smallest $c > 0$ such that $c(\beta_i - \beta_j) \in \mathcal{B}_{\beta_i^* - \beta_j^*}(r)$ for all $1 \leq i \neq j \leq k$), and a proxy for the final statistical rate, respectively.

Theorem 1 states that there are universal constants c_1 and c_2 such that if the sample size obeys the condition

$n \geq n_{\mathsf{AM}}(c_1)$, then we have

$$\Pr\left\{\max_{t\geq 1}\ \vartheta_t\left(c_2\frac{\pi_{\min}^6}{k^2\kappa};\left\{\beta_j^*\right\}_{j=1}^k\right)\geq c_1\delta_{n,\sigma}(d,k,\pi_{\min})\right\}$$
$$\leq c_2\left(k\exp\left(-c_1 n\frac{\pi_{\min}^6}{k^2}\right)+\frac{k^2}{n^7}\right). \tag{26}$$

Let us assume, without loss of generality, that the scalar c^* above is equal to 1. It is convenient to state and prove another result that guarantees a one-step contraction, from which Theorem 1 follows as a corollary. In order to state this result, we assume that one step of the alternating minimization update (12a)-(12b) is run starting from the parameters $\{\beta_j\}_{j=1}^k$ to produce the next iterate $\left\{\beta_j^+\right\}_{j=1}^k$. We use the shorthand

$$v_{i,j}^* = \beta_i^* - \beta_j^*,$$
$$v_{i,j} = \beta_i - \beta_j, \text{ and}$$
$$v_{i,j}^+ = \beta_i^+ - \beta_j^+.$$

Also recall the definitions of the geometric quantities (Δ,κ). The following proposition guarantees the one step contraction bound.

Proposition 1: There exist universal constants c_1 and c_2 such that

(a) If the sample size satisfies the bound $n \geq c_1\max\left\{\frac{dk}{\pi_{\min}^3}, \log n\cdot\frac{k^2}{\pi_{\min}^6}\right\}$, then for all parameters $\{\beta_j\}_{j=1}^k$ satisfying

$$\max_{1\leq j\neq j'\leq k}\ \frac{\left\|v_{j,j'}-v_{j,j'}^*\right\|}{\|\theta_j^*-\theta_{j'}^*\|}\log^{3/2}\left(\frac{\|\theta_j^*-\theta_{j'}^*\|}{\left\|v_{j,j'}-v_{j,j'}^*\right\|}\right)\leq c_2\frac{\pi_{\min}^6}{k^2\kappa}, \tag{27a}$$

we have, simultaneously for all pairs $1\leq j\neq\ell\leq k$, the bound

$$\frac{\left\|v_{j,\ell}^+-v_{j,\ell}^*\right\|^2}{\|\theta_j^*-\theta_\ell^*\|^2}\leq\max\left\{\frac{d\kappa}{\pi_{\min}^3 n},\frac{1}{4k}\right\}\times\left(\sum_{j'=1}^k\frac{\left\|v_{j,j'}-v_{j,j'}^*\right\|^2}{\|\theta_j^*-\theta_{j'}^*\|^2}+\frac{\left\|v_{\ell,j'}-v_{\ell,j'}^*\right\|^2}{\|\theta_\ell^*-\theta_{j'}^*\|^2}\right)+c_1\frac{\sigma^2}{\Delta}\frac{kd}{\pi_{\min}^3 n}\log(n/d) \tag{27b}$$

with probability exceeding $1-c_1\left(k\exp\left(-c_2 n\frac{k^2}{\pi_{\min}^6}\right)+\frac{k^2}{n^7}\right)$.

(b) If the sample size satisfies the bound $n\geq c_1\max\left\{\frac{dk}{\pi_{\min}^3}, \log n\cdot\frac{k^2}{\pi_{\min}^6}\right\}$, then for all parameters $\{\beta_j\}_{j=1}^k$ satisfying

$$\max_{1\leq j\neq j'\leq k}\ \frac{\left\|v_{j,j'}-v_{j,j'}^*\right\|}{\|\theta_j^*-\theta_{j'}^*\|}\log^{3/2}\left(\frac{\|\theta_j^*-\theta_{j'}^*\|}{\left\|v_{j,j'}-v_{j,j'}^*\right\|}\right)\leq c_2\frac{\pi_{\min}^6}{k^2}, \tag{28a}$$

we have the overall estimation error bound

$$\sum_{i=1}^k\|\beta_j^+-\beta_j^*\|^2\leq\frac{3}{4}\cdot\left(\sum_{i=1}^k\|\beta_j-\beta_j^*\|^2\right)+c_1\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(k)\log(n/dk) \tag{28b}$$

with probability exceeding $1-c_1\left(k\exp\left(-c_2 n\frac{k^2}{\pi_{\min}^6}\right)+\frac{k^2}{n^7}\right)$.

Let us briefly comment on why Proposition 1 implies Theorem 1 as a corollary. Clearly, equations (28a) and (28b) in conjunction show that the estimation error decays geometrically after running one step of the algorithm. The only remaining detail to be verified is that the next iterates $\left\{\beta_j^+\right\}_{j=1}^k$ also satisfy condition (27a) provided the sample size is large enough; in that case, the one step estimation bound (28b) can be applied recursively to obtain the final bound (14b).

With the constant c_2 from the proposition, let r_b be the largest value in the interval $[0,e^{-3/2}]$ such that $r_b\log^{3/2}(1/r_b)\leq c_2\frac{\pi_{\min}^6}{k^2}$. Similarly, let r_a be the largest value in the interval $[0,e^{-3/2}]$ such that $r_a\log^{3/2}(1/r_a)\leq c_2\frac{\pi_{\min}^6}{k^2\kappa}$. Bounds on both of these values will be used repeatedly later on.

Assume that the current parameters satisfy the bound (27a). Choosing $n\geq 4\kappa d/\pi_{\min}^3$ and applying inequality (27b), we have, for each pair $1\leq j\neq\ell\leq k$, the bound

$$\begin{aligned}\frac{\left\|v_{j,\ell}^+-v_{j,\ell}^*\right\|^2}{\|\theta_j^*-\theta_\ell^*\|^2}&\leq\frac{1}{4k}\left(\sum_{j'=1}^k\frac{\left\|v_{j,j'}-v_{j,j'}^*\right\|^2}{\|\theta_j^*-\theta_{j'}^*\|^2}+\frac{\left\|v_{\ell,j'}-v_{\ell,j'}^*\right\|^2}{\|\theta_\ell^*-\theta_{j'}^*\|^2}\right)+c_1\frac{\sigma^2}{\Delta}\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(n/d)\\&\leq\frac{1}{2}r_a^2+c_1\frac{\sigma^2}{\Delta}\frac{kd}{\pi_{\min}^3 n}\log(n/d).\end{aligned}$$

Further, if $n\geq C\sigma^2\frac{kd}{\pi_{\min}^3\Delta r_a^2}$ for a sufficiently large constant C, we have

$$\frac{\left\|v_{j,\ell}^+-v_{j,\ell}^*\right\|^2}{\|\theta_j^*-\theta_\ell^*\|^2}\leq r_a^2.$$

Thus, the parameters $\left\{\beta_j^+\right\}_{j=1}^k$ satisfy inequality (27a) for the sample size choice required by Theorem 1. Finally, noting, for a pair of small enough scalars (a,b), the implication

$$a\leq\frac{b}{2}\log^{-3/2}(1/b)\implies a\log^{3/2}(1/a)\leq b,$$

and adjusting the constants appropriately to simplify the probability statement completes the proof of the theorem. It now remains to establish Proposition 1.

A. Proof of Proposition 1

Recall that we denote by

$$S_j(\beta_1,\ldots,\beta_k):=\left\{1\leq i\leq n:\langle\xi_i,\,\beta_j\rangle=\max_{1\leq u\leq k}\left(\langle\xi_i,\,\beta_u\rangle\right)\right\},$$

the indices of the rows for which β_j attains the maximum, and we additionally keep this sets disjoint by breaking ties lexicographically. To lighten notation, we use the shorthand

$$\Xi^j(\beta_1, \ldots, \beta_k) := \Xi_{S_j(\beta_1,\ldots,\beta_k)}.$$

Recall the notation

$$\mathcal{B}_{v^*}(r) \; = \left\{ v \in \mathbb{R}^{d+1} : \frac{\|v - v^*\|}{\|u^*\|} \leq r \right\}$$

introduced before, and the definitions of the pair of scalars (r_a, r_b). To be agnostic to the scale invariance of the problem, we set $c^* = 1$ and define the set of parameters

$$\mathcal{I}(r) = \left\{ \beta_1, \ldots, \beta_k : v_{i,j} \in \mathcal{B}_{v^*_{i,j}}(r) \;\; \text{for all } 1 \leq i \neq j \leq k \right\},$$

and use the shorthand $\mathcal{I}_a := \mathcal{I}(r_a)$ and $\mathcal{I}_b := \mathcal{I}(r_b)$, to denote the set of parameters satisfying conditions (27a) and (28a), respectively,

Finally, recall the deterministic bound (22) established in Section III-D, restated below for convenience.

$$\begin{aligned}
&\frac{1}{2}\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) \cdot \|\beta_j^+ - \beta_j^*\|^2 \\
&\leq \sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\} \langle \xi_i,\, v^*_{j,j'} \rangle^2 \\
&\quad + \|P_{\Xi^j(\beta_1,\ldots,\beta_k)} \epsilon_{S_j}\|^2.
\end{aligned}$$

It suffices to show high probability bounds on the various quantities appearing in this bound. First, we claim that the noise terms are uniformly bounded as

$$\begin{aligned}
&\Pr\left\{ \sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \sum_{j=1}^{k} \|P_{\Xi^j(\beta_1,\ldots,\beta_k)} \epsilon_{S_j(\beta_1,\ldots,\beta_k)}\|^2 \right. \\
&\left. \geq 2\sigma^2 k(d+1)\log(kd)\log(n/kd) \right\} \leq \binom{n}{kd}^{-1}, \;\; \text{and}
\end{aligned} \tag{29a.I}$$

$$\begin{aligned}
&\Pr\left\{ \sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \|P_{\Xi^j(\beta_1,\ldots,\beta_k)} \epsilon_{S_j(\beta_1,\ldots,\beta_k)}\|^2 \right. \\
&\left. \geq 2\sigma^2 k(d+1)\log(n/d) \right\} \leq \binom{n}{d}^{-1} \;\; \text{for each } j \in [k].
\end{aligned} \tag{29a.II}$$

Second, we show that the indicator quantities are simultaneously bounded for all j, j' pairs. In particular, we claim that there exists a tuple of universal constants (C, c_1, c_2, c') such that for each positive scalar $r \leq 1/24$, we have

$$\begin{aligned}
&\Pr\left\{ \exists 1 \leq j \neq j' \leq k,\; v_{j,j'} \in \mathcal{B}_{v^*_{j,j'}}(r) : \right. \\
&\sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\} \langle \xi_i,\, v^*_{j,j'} \rangle^2 \\
&\quad \left. \geq C \max\{d, nr \;\; \log^{3/2}(1/r)\} \sum_{j' : j' \neq j} \|v_{j,j'} - v^*_{j,j'}\|^2 \right\} \\
&\leq c_1 \binom{k}{2} \left\{ n e^{-c_2 n} + e^{-c' \max\{d, 10 \log n\}} \right\}.
\end{aligned} \tag{29b}$$

Finally, we show a bound on the LHS of the bound (22) by handling the singular values of (random) sub-matrices of Ξ with a uniform bound. In particular, we claim that there are universal constants (C, c, c') such that if $n \geq C \max\left\{ \frac{dk}{\pi^3_{\min}}, \log n \cdot \frac{k^2}{\pi^6_{\min}} \right\}$, then for each $j \in [k]$, we have

$$\begin{aligned}
&\Pr\left\{ \inf_{\beta_1,\ldots,\beta_k \in \mathcal{I}_b} \lambda_{\min}\left(\Xi^j(\beta_1,\ldots,\beta_k)^\top \cdot \Xi^j(\beta_1,\ldots,\beta_k) \right) \right. \\
&\quad \left. \leq C\pi^3_{\min} n \right\} \leq c \exp\left(-cn \frac{\pi^6_{\min}}{k^2} \right) + c' n^{-10}.
\end{aligned} \tag{29c}$$

Notice that claim (29a.I) implicitly defines a high probability event $\mathcal{E}^{(a.I)}$, claim (29a.II) defines high probability events $\mathcal{E}_j^{(a.II)}$, claim (29b) defines a high probability event $\mathcal{E}^{(b)}(r)$, and claim (29c) defines high probability events $\mathcal{E}_j^{(c)}$. Define the intersection of these events as

$$\mathcal{E}(r) := \mathcal{E}^{(a.I)} \bigcap \left(\bigcap_{j \in [k]} \mathcal{E}_j^{(a.II)} \right) \bigcap \mathcal{E}^{(b)}(r) \bigcap \left(\bigcap_{j \in [k]} \mathcal{E}_j^{(c)} \right),$$

and note that the claims in conjunction with the union bound guarantee that if the condition on the sample size $n \geq c_1 \max\left\{ \frac{dk}{\pi^3_{\min}}, \log n \cdot \frac{k^2}{\pi^6_{\min}} \right\}$ holds, then for all $r \leq r_b$, we have

$$\Pr\{\mathcal{E}(r)\} \geq 1 - c_1 \left(k \exp\left(-c_2 n \frac{\pi^6_{\min}}{k^2} \right) + \frac{k^2}{n^7} \right),$$

where we have adjusted constants appropriately in stating the bound. We are now ready to prove the two parts of the proposition.

a) Proof of part (a): Work on the event $\mathcal{E}(r_a)$. Normalizing inequality (22) by n and using claims (29a.II). (29b), and (29c) with $r = r_a$ then yields, simultaneously for all $j \in [k]$, the bound

$$\begin{aligned}
&\|\beta_j^+ - \beta_j^*\|^2 \\
&\leq C \max\left\{ \frac{d}{\pi^3_{\min} n}, \frac{r_a}{\pi^3_{\min}} \log^{3/2}(1/r_a) \right\} \sum_{j' : j' \neq j} \|v_{j,j'} - v^*_{j,j'}\|^2 \\
&\qquad + C'\sigma^2 \frac{kd}{\pi^3_{\min} n} \log(n/d) \\
&\overset{\text{(i)}}{\leq} \max\left\{ \frac{Cd}{\pi^3_{\min} n}, \frac{1}{4k\kappa} \right\} \sum_{j' : j' \neq j} \|v_{j,j'} - v^*_{j,j'}\|^2 \\
&\qquad + C'\sigma^2 \frac{kd}{\pi^3_{\min} n} \log(n/d),
\end{aligned}$$

where in step (i), we have used the definition of the quantity r_a. Using this bound for the indices j, ℓ in conjunction with the definition of the quantity κ proves inequality (27b). □

b) Proof of part (b): We now work on the event $\mathcal{E}(r_b)$ and proceed again (see equation (22)) from the bound

$$\begin{aligned}
\|\beta_j^+ - \beta_j^*\|^2 &\leq C \max\left\{ \frac{d}{\pi^3_{\min} n}, \frac{r_b}{\pi^3_{\min}} \log^{3/2}(1/r_b) \right\} \\
&\quad \sum_{j' : j' \neq j} \|v_{j,j'} - v^*_{j,j'}\|^2 + \frac{C}{\pi^3_{\min} n} \|P_{\Xi^j(\beta_1,\ldots,\beta_k)} \epsilon_{S_j}\|^2.
\end{aligned}$$

Summing over $j \in [k]$ and using the fact that $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\sum_{j=1}^{k} \|\beta_j^+ - \beta_j^*\|^2 \leq C \max\left\{\frac{kd}{\pi_{\min}^3 n}, \frac{kr_b}{\pi_{\min}^3}\log^{3/2}(1/r_b)\right\} \times \left(\sum_{j=1}^{k}\|\beta_j - \beta_j^*\|^2\right) + \frac{C}{\pi_{\min}^3 n}\sum_{j\in[k]}\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}\epsilon_{S_j}\|^2 \overset{\text{(ii)}}{\leq} \frac{3}{4}\left(\sum_{j=1}^{k}\|\beta_j - \beta_j^*\|^2\right) + C'\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(k)\log(n/kd),$$

where in step (ii), we have used the definition of the quantity r_b, the bound $n \geq Ckd/\pi_{\min}^3$, and claim (29a.I). This completes the proof. □

We now prove each of the claims in turn. This constitutes the technical meat of our proof, and involves multiple technical lemmas whose proofs are postponed to the end of the section.

c) Proof of claims (29a.I) *and* (29a.II)*:* We begin by stating a general lemma about concentration properties of the noise.

Lemma 3: Consider a random variable $z \in \mathbb{R}^n$ with i.i.d. σ-sub-Gaussian entries, and a fixed matrix $\Xi \in \mathbb{R}^{n\times(d+1)}$. Then, we have

$$\sup_{\beta_1,\ldots,\beta_k\in\mathbb{R}^{d+1}}\sum_{j=1}^{k}\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}z\|^2 \leq 2\sigma^2 k(d+1)\log(kd)\log(n/kd) \tag{30a}$$

with probability greater than $1 - \binom{n}{kd}^{-1}$ and

$$\sup_{\beta_1,\ldots,\beta_k\in\mathbb{R}^{d+1}}\max_{j\in[k]}\|P_{\Xi^j(\beta_1,\ldots,\beta_k)}z_{S_j(\beta_1,\ldots,\beta_k)}\|^2 \leq 2\sigma^2 k(d+1)\log(n/d) \tag{30b}$$

with probability greater than $1 - \binom{n}{d}^{-1}$.

Here $z_{S_j(\beta_1,\ldots,\beta_k)}$ denotes the restriction of z onto the coordinates denoted by the set $S_j(\beta_1,\ldots,\beta_k)$. The proof of the claims follows directly from Lemma 3, since the noise vector (here we instantiate z by ϵ, which is sub-Gaussian) ϵ is independent of the matrix Ξ, and $\mathcal{I}_b \subseteq \left(\mathbb{R}^{d+1}\right)^{\otimes k}$. □

d) Proof of claim (29b)*:* We now state a lemma that directly handles indicator functions as they appear in the claim.

Lemma 4: Let $u^* \in \mathbb{R}^d$ and $w^* \in \mathbb{R}$, and consider a fixed parameter $v^* = (u^*,\ w^*) \in \mathbb{R}^{d+1}$. Then there are universal constants (c_1, c_2, c_3, c_4) such that for all positive scalars $r \leq 1/24$, we have

$$\sup_{v\in\mathcal{B}_{v^*}(r)}\frac{\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left\{\langle\xi_i,\, v\rangle\cdot\langle\xi_i,\, v^*\rangle \leq 0\right\}\langle\xi_i,\, v^*\rangle^2\right)}{\|v-v^*\|^2} \leq c_1\cdot\max\left\{\frac{d}{n}, r\log^{3/2}\left(\frac{1}{r}\right)\right\}$$

with probability exceeding $1 - c_1 e^{-c_2\max\{d, 10\log n\}} - c_3\, ne^{-c_4 n}$. Here, we adopt the convention that $0/0 = 0$.

Applying Lemma 4 with $v = v_{j,j'}$ and $v^* = v^*_{j,j'}$ for all pairs (j, j') and using a union bound directly yields the claim. □

e) Proof of claim (29c)*:* For this claim, we state three technical lemmas pertaining to the singular values of random matrices whose rows are formed by truncated Gaussian random vectors. We let $\mathfrak{vol}(K)$ denote the volume of a set $K \subseteq \mathbb{R}^d$ with respect to d-dimensional standard Gaussian measure, i.e., with $\mathfrak{vol}(K) = \Pr\{Z \in K\}$ for $Z \sim \mathcal{N}(0, I_d)$.

Lemma 5: Suppose n vectors $\{x_i\}_{i=1}^n$ are drawn i.i.d. from $\mathcal{N}(0, I_d)$, and $K \subseteq \mathbb{R}^d$ is a fixed convex set. Then there exists a tuple of universal constants (c_1, c_2) such that if $\mathfrak{vol}^3(K)n \geq c_1 d\log^2(1/\mathfrak{vol}(K))$, then

$$\lambda_{\min}\left(\sum_{i:x_i\in K}\xi_i\xi_i^\top\right) \geq c_2\,\mathfrak{vol}^3(K)\cdot n$$

with probability greater than $1 - c_1\exp\left(-c_2 n\frac{\mathfrak{vol}^4(K)}{\log^2(1/\mathfrak{vol}(K))}\right) - c_1\exp(-c_2 n\cdot\mathfrak{vol}(K))$.

For a pair of scalars (w, w') and d-dimensional vectors (u, u'), define the *wedge* formed by the $d+1$-dimensional vectors $v = (u,\ w)$ and $v' = (u',\ w')$ as the region

$$W(v, v') = \{x \in \mathbb{R}^d : (\langle x,\, u\rangle + w)\cdot(\langle x,\, u'\rangle + w') \leq 0\},$$

and let $\mathcal{W}_\delta = \{W = W(v, v') : \mathfrak{vol}(W) \leq \delta\}$ denote the set of all wedges with Gaussian volume less than δ. The next lemma bounds the maximum singular value of a sub-matrix formed by any such wedge.

Lemma 6: There is a tuple of universal constants (c_1, c_2) such that if $n \geq c_1\max\left\{d, \frac{\log n}{\delta^2}\right\}$, then

$$\sup_{W\in\mathcal{W}_\delta}\lambda_{\max}\left(\sum_{i:x_i\in W}\xi_i\xi_i^\top\right) \leq c_1 n\sqrt{\delta}$$

with probability greater than $1 - \exp(-c_2 n\delta^2) - n^{-10}$.

We are now ready to proceed to a proof of claim (29c). For convenience, introduce the shorthand notation

$$S_j^* := S_j\left(\beta_1^*, \ldots, \beta_k^*\right)$$

to denote the set of indices corresponding to observations generated by the true parameter β_j^*. Letting $A\Delta B := (A \setminus B)\bigcup(B \setminus A)$ denote the symmetric difference between two sets A and B, we have

$$\lambda_{\min}\left(\Xi_{S_j}^\top\Xi_{S_j}\right) \geq \lambda_{\min}\left(\Xi_{S_j^*}^\top\Xi_{S_j^*}\right) - \lambda_{\max}\left(\Xi_{S_j^*\Delta S_j}^\top\Xi_{S_j^*\Delta S_j}\right).$$

Recall that by definition, we have

$$\begin{aligned} S_j^*\Delta S_j &= \Big\{i : \langle\xi_i,\, \beta_j^*\rangle\} = \max \text{ and } \langle\xi_i,\, \beta_j\rangle \neq \max\} \\ &\qquad \bigcup\{i : \langle\xi_i,\, \beta_j^*\rangle \neq \max \text{ and } \langle\xi_i,\, \beta_j\rangle = \max\Big\} \\ &\subseteq \bigcup_{j'\in[k]\setminus j}\{i : \langle\xi_i,\, v_{j,j'}^*\rangle\cdot\langle\xi_i,\, v_{j,j'}\rangle < 0\} \\ &\subseteq \bigcup_{j'\in[k]\setminus j}\{i : x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)\}. \end{aligned} \tag{31}$$

Putting together the pieces, we have

$$\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) \geq \lambda_{\min}\left(\Xi_{S_j^*}^\top \Xi_{S_j^*}\right) - \sum_{j' \neq j} \lambda_{\max}\left(\sum_{i: x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)} \xi_i \xi_i^\top\right). \tag{32}$$

Now by Lemma 9, the definition of the set $\mathcal{I}_b$, and the definition of r_b, we have

$$\mathfrak{vol}\left(W\left(v_{j,j'}^*, v_{j,j'}\right)\right) \leq n r_b \log^{1/2}(1/r_b) \leq C\frac{\pi_{\min}^6}{k^2}.$$

Owing to the sample size assumption $n \geq C\max\left\{d, k^2\frac{\log n}{\pi_{\min}^6}\right\}$, the conditions of Lemma 6 are satisfied, and applying it yields

$$\sup_{v_{j,j'} \in \mathcal{B}_{v_{j,j'}^*}(r_a)} \lambda_{\max}\left(\sum_{i: x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)} \xi_i \xi_i^\top\right) \leq nC\frac{\pi_{\min}^3}{k}$$

with probability exceeding $1 - n^{-10} - \exp\left(-cn\frac{\pi_{\min}^6}{k^2}\right)$. Moreover, Lemma 5 guarantees the bound $\lambda_{\min}\left(\Xi_{S_j^*}^\top \Xi_{S_j^*}\right) \geq c_2 n \cdot \pi_{\min}^3$, so that putting together the pieces, we have

$$\inf_{\beta_1,\ldots,\beta_k \in \mathcal{I}_b} \lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) \geq c_2 n\pi_{\min}^3 - Cnk\frac{\pi_{\min}^3}{k} \geq C\pi_{\min}^3\, n, \tag{33}$$

with probability greater than $1 - c\exp\left(-cn\frac{\pi_{\min}^6}{k^2}\right) - n^{-10}$. These assertions hold provided

$$n \geq C\max\left\{d \cdot \frac{k}{\pi_{\min}^3}, \frac{k^2 \log n}{\pi_{\min}^6}\right\},$$

and this completes the proof. □

Having proved the claims, we turn to proofs of our technical lemmas.

1) Proof of Lemma 3: In this proof, we assume that $\sigma = 1$; our bounds can finally be scaled by σ^2.

It is natural to prove the bound (30b) first followed by bound (30a). First, consider a fixed set of parameters $\{\beta_1, \ldots, \beta_k\}$. Then, we have

$$\left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2 = \left\|UU^\top z_{S^j}\right\|^2,$$

where $U \in \mathbb{R}^{|\Xi^j| \times (d+1)}$ denotes a matrix with orthonormal columns that span the range of $\Xi^j(\beta_1, \ldots, \beta_k)$.

Applying the Hanson-Wright inequality for independent sub-Gaussians (see [68, Theorem 2.1]) and noting that $\left\|UU^\top\right\|_{\mathrm{F}} \leq \sqrt{d+1}$ we obtain

$$\Pr\left\{\left\|UU^\top z_{S^j}\right\|^2 \geq (d+1) + t\right\} \leq e^{-ct},$$

for each $t \geq 0$. In particular, this implies that the random variable $\left\|UU^\top z_{S^j}\right\|^2$ is sub-exponential.

This tail bound holds for a fixed partition of the rows of Ξ; we now take a union bound over all possible partitions. Toward that end, define the sets

$$\mathcal{S}^j = \left\{S_j(\beta_1, \ldots, \beta_k) : \beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}\right\}, \text{ for each } j \in [k].$$

From Lemma 11, we have the bound $|\mathcal{S}^j| \leq 2^{ckd\log(en/d)}$. Thus, applying the union bound, we obtain

$$\Pr\left\{\sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2 \geq (d+1) + t\right\} \leq |\mathcal{S}^j| e^{-ct},$$

and substituting $t = ck(d+1)\log(n/d)$ and performing some algebra establishes bound (30b).

In order to establish bound (30a), we once again consider the random variable $\sum_{j=1}^k \left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2$ for a fixed set of parameters $\{\beta_1, \ldots, \beta_k\}$. Note that this is the sum of k independent sub-exponential random variables and can be thought of as a quadratic form of the entire vector z. So once again from the Hanson-Wright inequality, we have

$$\Pr\left\{\sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \sum_{j=1}^k \left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2 \geq k(d+1) + t\right\} \leq e^{-ct/k}$$

for all $t \geq 0$.

Also define the set of all possible partitions of the n points via the max-affine function; we have the set

$$\mathcal{S} = \left\{S_1(\beta_1, \ldots, \beta_k), ., S_k(\beta_1, \ldots, \beta_k) : \beta_1, ., \beta_k \in \mathbb{R}^{d+1}\right\}.$$

Lemma 12 yields the bound $|\mathcal{S}| \leq 2^{ckd\log(kd)\log(n/kd)}$, and combining a union bound with the high probability bound above establishes bound (30a) after some algebraic manipulation. □

2) Proof of Lemma 4: Let $\gamma_v = v - v^*$; we have

$$\begin{aligned}\mathbf{1}\left\{\langle \xi_i, v\rangle \cdot \langle \xi_i, v^*\rangle \leq 0\right\}\langle \xi_i, v^*\rangle^2 &\leq \mathbf{1}\left\{\langle \xi_i, v\rangle \cdot \langle \xi_i, v^*\rangle \leq 0\right\}\langle \xi_i, \gamma_v\rangle^2 \\ &\leq \mathbf{1}\left\{\langle \xi_i, \gamma_v\rangle^2 \geq \langle \xi_i, v^*\rangle^2\right\}\langle \xi_i, \gamma_v\rangle^2.\end{aligned}$$

Define the (random) set $K_v = \{i : \langle \xi_i, \gamma_v\rangle^2 > \langle \xi_i, v^*\rangle^2\}$; we have the bound

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{\langle \xi_i, v\rangle \cdot \langle \xi_i, v^*\rangle \leq 0\right\}\langle \xi_i, v^*\rangle^2 \leq \frac{1}{n}\|\Xi_{K_v}\gamma_v\|^2.$$

We now show that the quantity $\|\Xi_{K_v}\gamma_v\|^2$ is bounded uniformly for all $v \in \mathcal{B}_{v^*}(r)$ for small enough r. Recall that u^* is the "linear" portion of v^*, and let $m = \max\{d, 10\log n, n \cdot (16r \cdot \sqrt{\log(1/r)}\}$ (note that m depends implicitly on r). We claim that for all $r \in (0, 1/24]$, we have

$$\Pr\left\{\sup_{v \in \mathcal{B}_{v^*}(r)} |K_v| > m\right\} \leq 4e^{-c\max\{d, 10\log n\}} + cne^{-c'n}, \tag{34a}$$

$$\Pr\left\{\bigcup_{\substack{T \subseteq [n]: \\ |T| \leq m}} \sup_{\substack{\omega \in \mathbb{R}^{d+1} \\ \omega \neq 0}} \frac{\|\Xi_T \omega\|^2}{\|\omega\|^2} \geq (2d + 20m\log(n/m))\right\} \leq e^{-c\max\{d, 10\log n\}}. \tag{34b}$$

Taking these claims as given, the proof of the lemma is immediate, since $\frac{n}{m} \leq \frac{1}{16r\log(1/r)}$, so that $\log(n/m) \leq C\log(1/r)$.

a) Proof of claim (34a)*:* By definition of the set K_v, we have

$$\begin{aligned}
&\Pr\{\sup_{v\in\mathcal{B}_{v^*}(r)}|K_v|>m\}\\
&\le \sum_{\substack{T\subseteq[n]:\\|T|>m}}\Pr\left\{\exists v\in\mathcal{B}_{v^*}(r):\|\Xi_T\gamma_v\|^2\ge\|\Xi_T v^*\|^2\right\}\\
&= \sum_{\substack{T\subseteq[n]:\\|T|>m}}\Pr\left\{\exists v\in\mathcal{B}_{v^*}(r):\frac{\|\gamma_v\|^2}{\|u^*\|^2}\frac{\|\Xi_T\gamma_v\|^2}{\|\gamma_v\|^2}\ge\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\right\}\\
&\le \sum_{\substack{T\subseteq[n]:\\|T|>m}}\Pr\left\{\exists v\in\mathcal{B}_{v^*}(r):r^2\frac{\|\Xi_T\gamma_v\|^2}{\|\gamma_v\|^2}\ge\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\right\}\\
&\le \sum_{\substack{T\subseteq[n]:\\|T|>m}}\Bigg(\Pr\left\{\exists v\in\mathcal{B}_{v^*}(r):\frac{\|\Xi_T\gamma_v\|^2}{\|\gamma_v\|^2}\right.\\
&\qquad\left.\ge(\sqrt{d}+\sqrt{|T|}+t_T)^2\right\}\\
&\qquad+\Pr\left\{\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\le r^2(\sqrt{d}+\sqrt{|T|}+t_T)^2\right\}\Bigg),
\end{aligned}$$

where the final step follows by the union bound and holds for all positive scalars $\{t_T\}_{T\subseteq[n]}$. For some *fixed* subset T of size ℓ, we have the tail bounds,

$$\Pr\left\{\sup_{\substack{\omega\in\mathbb{R}^{d+1}\\\omega\neq0}}\frac{\|\Xi_T\omega\|^2}{\|\omega\|^2}\ge(\sqrt{d}+\sqrt{\ell}+t)^2\right\}\overset{\text{(i)}}{\le}2e^{-t^2/2},\tag{35a}$$

$$\Pr\left\{\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\le\delta\ell\right\}\overset{\text{(ii)}}{\le}(e\delta)^{\ell/2}\text{ for all }\delta\ge0,\tag{35b}$$

for all $t\ge0$, where step (i) follows from the sub-Gaussianity of the covariate matrix (see Lemma 13), and step (ii) from a tail bound for the non-central χ^2 distribution (see Lemma 14).

Substituting these bounds yields

$$\begin{aligned}
&\Pr\{\sup_{v\in\mathcal{B}_{v^*}(r)}|K_v|>m\}\\
&\le\sum_{\ell=m+1}^{n}\binom{n}{\ell}\left[2e^{-t_\ell^2/2}+\left(er^2\cdot\frac{(\sqrt{d}+\sqrt{\ell}+t_\ell)^2}{\ell}\right)^{\ell/2}\right]\\
&\le\sum_{\ell=m+1}^{n}\binom{n}{\ell}\left[2e^{-t_\ell^2/2}+\left(2r\cdot\frac{\sqrt{d}+\sqrt{\ell}+t_\ell}{\sqrt{\ell}}\right)^{\ell}\right].
\end{aligned}$$

Recall that t_ℓ was a free (non-negative) variable to be chosen. We now split the proof into two cases and choose this parameter differently for the two cases.

b) Case 1, $m\le\ell<n/e$: Substituting the choice $t_\ell=4\sqrt{\ell\log(n/\ell)}$, we obtain

$$\begin{aligned}
&\binom{n}{\ell}\left[2e^{-t_\ell^2/2}+\left(2r\cdot\frac{\sqrt{d}+\sqrt{\ell}+t_\ell}{\sqrt{\ell}}\right)^{\ell}\right]\\
&\le\left(\frac{n}{\ell}\right)^{-c\ell}+\binom{n}{\ell}\cdot\left(2r\cdot\frac{\sqrt{d}+5\sqrt{\ell\log(n/\ell)}}{\sqrt{\ell}}\right)^{\ell}\\
&\overset{\text{(i)}}{\le}\left(\frac{n}{\ell}\right)^{-c\ell}+\binom{n}{\ell}\cdot\left(2r\cdot(1+5\sqrt{\log(n/\ell)})\right)^{\ell}\\
&\overset{\text{(ii)}}{\le}\left(\frac{n}{\ell}\right)^{-c\ell}+\binom{n}{\ell}\cdot\left(12r\cdot\sqrt{\log(n/\ell)}\right)^{\ell}\\
&\le\left(\frac{n}{\ell}\right)^{-c\ell}+\left(12\left(\frac{en}{\ell}\right)r\cdot\sqrt{\log(n/\ell)}\right)^{\ell},
\end{aligned}$$

where step (i) follows from the bound $m\ge d$, and step (ii) from the bound $\ell\le n/e$.

Now note that the second term is only problematic for small ℓ. For all $\ell\ge m=n\cdot(16r\cdot\sqrt{\log(1/r)})$, we have

$$\left(12\left(\frac{en}{\ell}\right)r\cdot\sqrt{\log(n/\ell)}\right)^{\ell}\le(3/4)^{\ell}.$$

The first term, on the other hand, satisfies the bound $\left(\frac{n}{\ell}\right)^{-c\ell}\le(3/4)^{\ell}$ for sufficiently large n.

c) Case 2, $\ell\ge n/e$: In this case, setting $t_\ell=2\sqrt{n}$ for each ℓ yields the bound

$$\begin{aligned}
&\binom{n}{\ell}\left[2e^{-t_\ell^2/2}+\left(2r\cdot\frac{\sqrt{d}+\sqrt{\ell}+t_\ell}{\sqrt{\ell}}\right)^{\ell}\right]\\
&\le2\binom{n}{n/2}e^{-2n}+(12r)^{\ell}\\
&\le ce^{-c'n},
\end{aligned}$$

where we have used the fact that $d\le n/2$ and $r\le1/24$.

Putting together the pieces from both cases, we have shown that for all $r\in(0,1/24]$, we have

$$\begin{aligned}
\Pr\{\sup_{v\in\mathcal{B}_{v^*}(r)}|K_v|>m\}&\le cne^{-c'n}+\sum_{\ell=m+1}^{n/e}(3/4)^{\ell}\\
&\le cne^{-c'n}+4(3/4)^{\max\{d,10\log n\}},
\end{aligned}$$

thus completing the proof of the claim.

d) Proof of claim (34b)*:* The proof of this claim follows immediately from the steps used to establish the previous claim. In particular, writing

$$\begin{aligned}
&\Pr\left\{\bigcup_{\substack{T\subseteq[n]:\\|T|\le m}}\bigcup_{\omega:\|\omega\|=1}\|\Xi_T\omega\|^2\ge2d+20m\log(n/m)\right\}\\
&\le\Pr\left\{\bigcup_{\substack{T\subseteq[n]:\\|T|\le m}}\bigcup_{\omega:\|\omega\|=1}\|\Xi_T\omega\|^2\right.\\
&\qquad\left.\ge\left(\sqrt{d}+\sqrt{m}+\sqrt{4m\log(n/m)}\right)^2\right\}\\
&\le\sum_{\ell=1}^{m}\Pr\left\{\bigcup_{\substack{T\subseteq[n]:\\|T|=\ell}}\bigcup_{\omega:\|\omega\|=1}\|\Xi_T\omega\|^2\right.\\
&\qquad\left.\ge\left(\sqrt{d}+\sqrt{m}+\sqrt{4m\log(n/m)}\right)^2\right\}\\
&\overset{\text{(iv)}}{\le}2\sum_{\ell=1}^{m}\binom{n}{\ell}\exp\{-2m\log(n/m)\}\\
&\le2\left(\frac{n}{m}\right)^{-cm}\le2e^{-c\max\{d,10\log n\}},
\end{aligned}$$

where step (iv) follows from the tail bound (35a). □

3) Proof of Lemma 5: The lemma follows from some structural results on the truncated Gaussian distribution. Using the shorthand $\mathfrak{vol} := \mathfrak{vol}(K)$ and letting ψ denote the d-dimensional Gaussian density, consider a random vector τ drawn from the distribution having density $h(y) = \frac{1}{\mathfrak{vol}}\psi(y)\mathbf{1}\{y \in K\}$, and denote its mean and second moment matrix by μ_τ and Σ_τ, respectively. Also denote the recentered random variable by $\widetilde{\tau} = \tau - \mu_\tau$. We claim that

$$\|\mu_\tau\|^2 \le C \log(1/\mathfrak{vol}), \tag{36a}$$

$$C\,\mathfrak{vol}^2 \cdot I \preceq \Sigma_\tau \preceq (1 + C\log(1/\mathfrak{vol}))\, I, \text{ and} \tag{36b}$$

$$\widetilde{\tau} \text{ is } c\text{-sub-Gaussian for a universal constant } c. \tag{36c}$$

Taking these claims as given for the moment, let us prove the lemma.

The claims (36a) and (36c) taken together imply that the random variable τ is sub-Gaussian with parameter $\zeta^2 \le 2c^2 + 2C\log(1/\mathfrak{vol})$. Now consider m i.i.d. draws of τ given by $\{\tau_i\}_{i=1}^m$; standard results (see, e.g., Vershynin [69, Remark 5.40], or Wainwright [70, Theorem 6.2]) yield the bound

$$\Pr\left\{\Big\|\frac{1}{m}\sum_{i=1}^m \tau_i\tau_i^\top - \Sigma_\tau\Big\|_{\text{op}} \ge \zeta^2\left(\frac{d}{m} + \sqrt{\frac{d}{m}} + \delta\right)\right\} \le 2\exp\left(-cn\min\{\delta, \delta^2\}\right).$$

Using this bound along with claim (36b) and Weyl's inequality yields

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \tau_i\tau_i^\top\right) \ge C\,\mathfrak{vol}^2 - \zeta^2\left(\frac{d}{m} + \sqrt{\frac{d}{m}} + \delta\right) \tag{37}$$

with probability greater than $1 - 2\exp\left(-cn\min\{\delta, \delta^2\}\right)$.

Furthermore, when n samples are drawn from a standard Gaussian distribution, the number m of them that fall in the set K satisfies $m \ge \frac{1}{2}n\cdot\mathfrak{vol}$ with high probability. In particular, this follows from a straightforward binomial tail bound, which yields

$$\Pr\left\{m \le \frac{n\cdot\mathfrak{vol}}{2}\right\} \le \exp(-cn\cdot\mathfrak{vol}). \tag{38}$$

Recall our choice $n \ge Cd\frac{\log^2(1/\mathfrak{vol})}{\mathfrak{vol}^3}$, which in conjunction with the bound (38) ensures that $C\,\mathfrak{vol}^2 \ge \frac{1}{8}\sigma^2\sqrt{\frac{d}{m}}$ with high probability. Setting $\delta = C\,\mathfrak{vol}^2/\sigma^2$ in inequality (37), we have

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^m \tau_i\tau_i^\top\right) \ge \frac{C}{2}\,\mathfrak{vol}^2$$

with probability greater than $1 - 2\exp\left(-cn\,\mathfrak{vol}^4/\sigma^4\right)$. Putting together the pieces thus proves the lemma. It remains to show the various claims. □

a) Proof of claim (36a)*:* Let $\tau_{\mathcal{A}}$ denote a random variable formed as a result of truncating the Gaussian distribution to a (general) set $\mathcal{A}$ with volume $\mathfrak{vol}$. Letting $\mu_{\mathcal{A}}$ denote its mean, the dual norm definition of the ℓ_2 norm yields

$$\|\mu_{\mathcal{A}}\| = \sup_{v\in\mathbb{S}^{d-1}} \langle v, \mu_{\mathcal{A}}\rangle \le \sup_{v\in\mathbb{S}^{d-1}} \mathbb{E}|\langle v, \tau_{\mathcal{A}}\rangle|.$$

Let us now evaluate an upper bound on the quantity $\mathbb{E}|\langle v, \tau_{\mathcal{A}}\rangle|$. In the calculation, for any d-dimensional vector y, we use the shorthand $y_v := v^\top y$ and $y_{\setminus v} := U_{\setminus v}^\top y$ for a matrix $U_{\setminus v} \in \mathbb{R}^{d\times(d-1)}$ having orthonormal columns that span the subspace orthogonal to v. Letting $\mathcal{A}_v \subseteq \mathbb{R}$ denote the projection of $\mathcal{A}$ onto the direction v, define the set $\mathcal{A}_{\setminus v}(w) \subseteq \mathbb{R}^{d-1}$ via

$$\mathcal{A}_{\setminus v}(w) = \{y_{\setminus v} \in \mathbb{R}^{d-1} : y \in \mathcal{A} \text{ and } y_v = w\}.$$

Letting ψ_d denote the d-dimensional standard Gaussian pdf, we have

$$\begin{aligned}
\mathbb{E}|\langle v, \tau_{\mathcal{A}}\rangle| &= \frac{1}{\mathfrak{vol}}\int_{y\in\mathcal{A}} |y^\top v|\psi_d(y)dy \\
&= \frac{1}{\mathfrak{vol}}\int_{y\in\mathcal{A}} |y_v|\psi(y_v)\psi_{d-1}(y_{\setminus v})dy \\
&= \frac{1}{\mathfrak{vol}}\int_{y_v\in\mathcal{A}_v} |y_v|\psi(y_v) \\
&\quad\times \underbrace{\left(\int_{y_{\setminus v}\in\mathcal{A}_{\setminus v}(y_v)} \psi_{d-1}(y_{\setminus v} \in \mathcal{A}_{\setminus v}(y_v))dy_{\setminus v}\right)}_{f(y_v)} dy_v \\
&\overset{\text{(i)}}{\le} \frac{1}{\mathfrak{vol}}\int_{y_v\in\mathcal{A}_v} |y_v|\psi(y_v)dy_v,
\end{aligned} \tag{39}$$

where step (i) follows since $f(y_v) \le 1$ point-wise. On the other hand, we have

$$\begin{aligned}
\mathfrak{vol} &= \int_{y_v\in\mathcal{A}_v} \psi(y_v)\left(\int_{y_{\setminus v}\in\mathcal{A}_{\setminus v}(y_v)} \psi_{d-1}dy_{\setminus v}\right)dy_v \\
&\le \int_{y_v\in\mathcal{A}_v} \psi(y_v)dy_v.
\end{aligned} \tag{40}$$

Combining inequalities (39) and (40) and letting $w = y_v$, an upper bound on $\|\mu_\tau\|$ can be obtained by solving the one-dimensional problem given by

$$\|\mu_\tau\| \le \sup_{\mathcal{S}\subseteq\mathbb{R}} \frac{1}{\mathfrak{vol}}\int_{w\in\mathcal{S}} |w|\psi(w)dw \quad \text{s.t.} \int_{w\in\mathcal{S}} \psi(w)dw \ge \mathfrak{vol}.$$

It can be verified that the optimal solution to the problem above is given by choosing the truncation set $\mathcal{S} = (\infty, -\beta)\cup[\beta, \infty)$ for some threshold $\beta > 0$. With this choice, the constraint can be written as

$$\mathfrak{vol} \le \int_{|w|\ge\beta} \psi(w)dw \le 2\sqrt{\frac{2}{\pi}}\frac{1}{\beta}e^{-\beta^2/2},$$

where we have used a standard Gaussian tail bound. Simplifying yields the bound

$$\beta \le 2\sqrt{\log(C/\mathfrak{vol})}.$$

Furthermore, we have

$$\begin{aligned}
\frac{1}{\mathfrak{vol}}\int_{|w|\ge\beta} |w|\psi(w)dw &= \frac{C}{\mathfrak{vol}}e^{-\beta^2/2} \\
&\overset{\text{(ii)}}{\lesssim} \frac{\beta^3}{\beta^2 - 1} \\
&\le c\sqrt{\log(1/\mathfrak{vol})},
\end{aligned}$$

where step (ii) follows from the bound $\Pr\{Z \geq z\} \geq \psi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right)$ valid for a standard Gaussian variate Z. Putting together the pieces, we have

$$\|\mu_\tau\|^2 \leq c \log(1/\,\mathfrak{vol}).$$

□

b) Proof of claim (36b)*:* Let us first show the upper bound. Writing $\mathrm{cov}(\tau)$ for the covariance matrix, we have

$$\begin{aligned}\|\Sigma_\tau\|_{\mathrm{op}} &\leq \|\mathrm{cov}(\tau)\|_{\mathrm{op}} + \|\mu_\tau\|^2 \\ &\overset{\text{(iii)}}{\leq} \|I\|_{\mathrm{op}} + C\log(1/\,\mathfrak{vol}),\end{aligned}$$

where step (iii) follows from the fact that $\mathrm{cov}(\tau) \preceq \mathrm{cov}(Z)$, since truncating a Gaussian to a convex set reduces its variance along all directions [71], [72].

We now proceed to the lower bound. Let $\mathbb{P}_K$ denote the Gaussian distribution truncated to the set K. Recall that we denoted the probability that a Gaussian random variable falls in the set K by $\mathfrak{vol}(K)$; use the shorthand $\mathfrak{vol} = \mathfrak{vol}(K)$. Define the polynomial

$$p_u(x) \ = \langle x - \mathbb{E}_{X\sim\mathbb{P}_K}[X],\, u\rangle^2;$$

note that we are interested in a lower bound on $\inf_{u\in\mathbb{S}^{d-1}} \mathbb{E}_{X\sim\mathbb{P}_K}[p_u(X)]$.

For $\delta > 0$, define the set

$$S_\delta := \{x \in \mathbb{R}^d : p_u(x) \ \leq \delta\} \subseteq \mathbb{R}^d.$$

Letting Z denote a d-dimensional standard Gaussian random vector and using the shorthand $\alpha := \mathbb{E}_{X\sim\mathbb{P}_K}[X]$, we have

$$\begin{aligned}\Pr\{Z \in S_\delta\} &= \Pr\left\{\langle Z-\alpha,\, u\rangle^2 \leq \delta\right\} &(41)\\ &= \Pr\left\{\langle \alpha,\, u\rangle - \sqrt{\delta} \leq \langle Z,\, u\rangle \leq \langle \alpha,\, u\rangle + \sqrt{\delta}\right\} &(42)\\ &= \int_{\langle \alpha,\, u\rangle - \sqrt{\delta}}^{\langle \alpha,\, u\rangle + \sqrt{\delta}} \psi(x)dx \leq \sqrt{\frac{2}{\pi}}\delta, &(43)\end{aligned}$$

where in the final step, we have used the fact that $\psi(x) \leq 1/\sqrt{2\pi}$ for all $x \in \mathbb{R}$.

Consequently, we have

$$\begin{aligned}\mathbb{E}_{X\sim\mathbb{P}_K}[p_u(X)] &= \frac{1}{\mathfrak{vol}}\mathbb{E}_Z\left[p_u(Z)\ \mathbf{1}\left\{Z \in K\right\}\right] \\ &\geq \frac{1}{\mathfrak{vol}}\mathbb{E}_Z\left[p_u(Z)\ \mathbf{1}\left\{Z \in K \cap S_\delta^c\right\}\right] \\ &\overset{\text{(iv)}}{\geq} \frac{1}{\mathfrak{vol}}\mathbb{E}_Z\left[\delta \mathbf{1}\left\{Z \in K \cap S_\delta^c\right\}\right] \\ &= \frac{\delta}{\mathfrak{vol}}\Pr\{Z \in K \cap S_\delta^c\} \\ &\overset{\text{(v)}}{\geq} \delta\frac{\mathfrak{vol} - \sqrt{\frac{2}{\pi}\delta}}{\mathfrak{vol}}.\end{aligned}$$

Here, step (iv) follows from the definition of the set S_δ, which ensures that $p_u(x) \ \geq \delta$ for all $x \in S_\delta^c$. Step (v) follows as a consequence of equation (43), since

$$\Pr\{Z \in K \cap S_\delta^c\} = \Pr\{Z \in K\} - \Pr\{Z \in S_\delta\} \geq \mathfrak{vol} - \sqrt{\frac{2}{\pi}\delta}.$$

Finally, choosing $\delta = c\,\mathfrak{vol}^2$ for a suitably small constant c, we have $\mathbb{E}_{X\sim\mathbb{P}_K}[p_u(X)] \geq C\,\mathfrak{vol}^2$ for a fixed $u \in \mathbb{S}^{d-1}$. Since u was chosen arbitrarily, this proves the claim. □

c) Proof of claim (36c)*:* Since the random variable ξ is obtained by truncating a Gaussian random variable to a convex set, it is 1-strongly log-concave. Thus, standard results [73, Theorem 2.15] show that the random variable $\widetilde{\xi}$ is c-sub-Gaussian. □

4) Proof of Lemma 6: For a pair of $d+1$-dimensional vectors (v, v'), denote by

$$n_{W(v,v')} = \#\{i : x_i \in W(v, v')\} \tag{44}$$

the random variable that counts the number of points that fall within the wedge $W(v, v')$; recall our notation W_δ for the set of all wedges with Gaussian volume less than δ. Since each wedge is formed by the intersection of two hyperplanes, applying Lemmas 10 and 11 in conjunction yields that there are universal constants (c, c', C) such that

$$\sup_{W\in\mathcal{W}_\delta} n_W \leq c\delta n \tag{45}$$

with probability exceeding $1 - \exp(-c'n\delta^2)$, provided $n \geq \frac{C}{\delta^2}$. In words, the maximum number of points that fall in *any* wedge of volume δ is linear in δn with high probability.

It thus suffices to bound, simultaneously, the maximum singular value of every sub-matrix of Ξ having (at most) $c\delta n$ rows. Applying [44, Theorem 5.7] yields the bound[11]

$$\Pr\left\{\max_{S:|S|\leq c\delta n} \lambda_{\max}\left(\sum_{i\in S}\xi_i\xi_i^\top\right) \geq c_1 n\sqrt{\delta}\right\} \leq n^{-10},$$

where we have used the lower bound $n \geq c\max\{d, \log n/\delta\}$ on the sample size.

Putting together the pieces, we have that if $n \geq c\max\left\{d, \frac{\log n}{\delta^2}\right\}$, then

$$\sup_{W\in\mathcal{W}_\delta} \lambda_{\max}\left(\sum_{i:x_i\in W}\xi_i\xi_i^\top\right) \leq c_1 n\sqrt{\delta}$$

with probability exceeding $1 - n^{-10} - \exp(-c'n\delta^2)$. □

Appendix C
Proof of Theorem 2

We dedicate the first portion of the proof to a precise definition of the quantity γ.

Let $\Theta^* \in \mathbb{R}^{k\times d}$ denote a matrix with rows $(\theta_j^*)^T, j = 1, \ldots, k$ and let $\Sigma = \Theta^*(\Theta^*)^\top \in \mathbb{R}^{k\times k}$. We employ the decomposition $\Theta^* = A^*(U^*)^\top$, where $A^* \in \mathbb{R}^{k\times k}$ is the invertible matrix of coefficients and $U^* \in \mathbb{R}^{d\times k}$ is a matrix of orthonormal columns. Note that for $X \sim N(0, I_d)$, the vector in $\mathbb{R}^k$ with j-th component $\langle X,\, \theta_j^*\rangle + b_j^*$ is distributed

[11] Strictly speaking, [44, Theorem 5.7] applies to Gaussian random matrices, i.e., without the appended column of ones. By multiplying each row of Ξ with an independent Rademacher RV (see the proof of Lemma 13) to obtain a sub-Gaussian random matrix with the same singular values, and noting also that the proof technique of [44, Theorem 5.7] relies on chaining and holds for a sub-Gaussian random matrix, one can show that the same result also holds for the matrix Ξ.

as $Z+b^*$ where $Z \sim N(0,\Sigma)$ and the vector $b^* \in \mathbb{R}^k$ collects the scalars $\{b_j^*\}_{j=1}^k$ in its entries. For $Z \sim \mathcal{N}(0,\Sigma)$, let

$$\rho = \frac{\mathbb{E}\left[\max(Z+b^*)Z^\top \Sigma^{-1}\mathbf{1}\right]}{\sqrt{\mathbb{E}\left[(\max(Z+b^*))^2\right]\cdot \mathbb{E}\left[(Z^\top\Sigma^{-1}\mathbf{1})^2\right]}} \tag{46}$$

denote the correlation coefficient between the maximum and a particular linear combination of a multivariate Gaussian distribution. Variants of such quantities have been studied extensively in the statistical literature (see, e.g., James [74]). For our purposes, the fact that $\max(Z+b^*)Z \neq 0$ for any finite b^*, coupled with a full-rank Σ, ensure that $\rho \neq 0$ for any fixed k. Also define the positive scalar $\varrho := \sqrt{\mathbb{E}[(\max(Z+b^*))^2]}$, which tracks the average size of our observations. Also recall the quantity ς defined in the main section.

For each $j \in [k]$ consider the zero-mean Gaussian random vector with covariance $(\mathbf{1}\cdot e_j^\top - I)A^*(A^*)^\top(\mathbf{1}\cdot e_j^\top - I)^\top$. This is effectively a Gaussian that lives in $k-1$ dimensions, with density that we denote by $\widetilde{\psi}_j(x_1,x_2,\ldots,x_{j-1},0,x_{j+1},\ldots x_k)$ at point $(x_1,x_2,\ldots,x_{j-1},0,x_{j+1},\ldots x_k)$ (the density is not defined elsewhere). Truncate this random vector to the region $\{x_i \geq b_i^* - b_j^* : i \in [k]\}$; this results in the *truncated* Gaussian density $\psi_j(x_1,x_2,\ldots,x_{j-1},0,x_{j+1},\ldots x_k)$ for each $j \in [k]$.

For any $x \in \mathbb{R}^k$ such that $x_j = 0$, define

$$\begin{aligned} F_i^j(x) \;=& \int_{b_1^*-b_j^*}^{\infty} \cdots \int_{b_{i-1}^*-b_j^*}^{\infty} \int_{b_{i+1}^*-b_j^*}^{\infty} \cdots \\ & .. \int_{b_k^*-b_j^*}^{\infty} \psi_j(x_1,\ldots,x_{i-1},x,x_{i+1},\ldots,x_k)dx_k \\ & \ldots dx_{i+1}dx_{i-1}\ldots dx_1 \end{aligned} \tag{47}$$

to be the i-th marginal density of this truncated Gaussian evaluated at the point x, with the convention that $F_j^j(\cdot) = 0$ everywhere. Also define the vector F^j by setting its i-th entry to $(F^j)_i = F_i^j(b_i^* - b_j^*)$.

Now let P denote the matrix with entries

$$P_{i,j} = \begin{cases} (F^j)_i / \sum_{k\neq j}(F^j)_k & \text{if } i \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{48}$$

Note that the matrix P is the transition matrix of an irreducible, aperiodic Markov chain, with one eigenvalue equal to 1. Consequently, the matrix $I-P$ is rank $k-1$. With this setup in place, let

$$\begin{aligned} \gamma := \min\Bigg\{ & \rho^2\varrho^2, \min_{j\in k}\left(\sum_{k\neq j}(F^j)_k\right)\lambda_k(\Sigma) \\ & \times \sqrt{\lambda_{k-1}\left((I-P^\top)(I-P)\right)}\Bigg\} \end{aligned} \tag{49}$$

denote a positive scalar that will serve as a bound on our eigengap.

Let $M_1 = \mathbb{E}[\max(\Theta^* X + b^*)X]$ and $M_2 = \mathbb{E}\left[\max(\Theta^* X + b^*)(XX^\top - I_d)\right]$ denote the expectations of the first and second moment estimators, respectively.

For a random variable $W \sim \mathcal{N}(b^*,\Sigma)$, we often use the shorthand

$$\{W_j = \max\} := \{W_j \geq W_i \text{ for all } 1 \leq i \leq k\}.$$

Finally, collect the probabilities $\{\pi_j\}_{j=1}^k$ defined in equation (9) in a vector $\pi \in \mathbb{R}^k$. We use $\mathbf{1}$ to denote the all-ones vector in k dimensions.

We are ready to state our two main lemmas.

Lemma 7: (a) The first moment satisfies

$$M_1 = (\Theta^*)^\top \pi \quad \text{and} \quad \langle M_1,\, (\Theta^*)^\top\Sigma^{-1}\mathbf{1}\rangle = \rho\varrho\left\|(\Theta^*)^\top\Sigma^{-1}\mathbf{1}\right\|.$$

(b) The second moment satisfies

$$\begin{aligned} M_2 \succeq 0, \quad M_2(\Theta^*)^\top\Sigma^{-1}\mathbf{1} = 0, \quad \mathsf{rank}(M_2) = k-1 \quad \text{and} \\ \lambda_{k-1}(M_2) \geq \min_{j\in k}\left(\sum_{k\neq j}(F^j)_k\right)\lambda_k(\Sigma) \\ \times \sqrt{\lambda_{k-1}\left((I-P^\top)(I-P)\right)}. \end{aligned}$$

We combine this lemma with a result that shows that the empirical moments concentrate about their expectations.

Lemma 8: For an absolute constant C, we have

$$\Pr\left\{\left\|\widehat{M}_1 - M_1\right\|^2 \geq C\left(\sigma^2+\varsigma^2\right)\frac{d\log^2(nk)}{n}\right\} \leq 5dn^{-12}, \tag{50a}$$

$$\Pr\left\{\|\!|\widehat{M}_2 - M_2|\!\|_{\text{op}}^2 \geq C\left(\sigma^2+\varsigma^2\right)\frac{d\log^3(nk)}{n}\right\} \leq 5dn^{-12}. \tag{50b}$$

Lemma 7 is proved at the end of this section, and Lemma 8 is proved in Appendix G-A. For now, we take both lemmas as given and proceed to a proof of Theorem 2.

Recall the matrix $\widehat{M} = \widehat{M}_1 \otimes \widehat{M}_1 + \widehat{M}_2$ and let $M = M_1 \otimes M_1 + M_2$. By Lemma 7, the matrix M is positive semidefinite with k non-zero eigenvalues. In particular, using the shorthand $\bar{\theta} := (\Theta^*)^\top\Sigma^{-1}\mathbf{1}$, we have $\bar{\theta} \in \mathsf{nullspace}(M_2)$, and so

$$\bar{\theta}^\top M\bar{\theta} = \langle\bar{\theta},\, M_1\rangle^2 = \rho^2\varrho^2\|\bar{\theta}\|^2,$$

where the final inequality follows by part (a) of Lemma 7.

Thus, there is a k-dimensional subspace orthogonal to the nullspace of M (and so the range of M is k dimensional). For any unit vector v in this subspace, we have

$$v^\top M v \geq \min\{\rho^2\varrho^2, \lambda_{k-1}(M_2)\}.$$

Thus, the kth eigenvalue of M satisfies

$$\begin{aligned} \lambda_k(M) \geq \min\Bigg\{ & \rho^2\varrho^2, \min_{j\in k}\left(\sum_{k\neq j}(F^j)_k\right)\lambda_k(\Sigma) \\ & \sqrt{\lambda_{k-1}\left((I-P^\top)(I-P)\right)}\Bigg\} = \gamma, \end{aligned}$$

where the equality follows by definition (49). By Lemma 8, we have

$$\begin{aligned} & \|\!|\widehat{M} - M|\!\|_{\text{op}}^2 \leq 2\|\!|\widehat{M}_2 - M_2|\!\|_{\text{op}}^2 + 2\|\!|\widehat{M}_1\otimes\widehat{M}_1 - M_1\otimes M_1|\!\|_{\text{op}}^2 \\ & \leq 2C\left(\sigma^2+\varsigma^2\log^2(nk)\right)\frac{d\log(nk)}{n} \\ & \quad + 16\left\|\widehat{M}_1 - M_1\right\|^2\|M_1\|^2 + 4\left\|\widehat{M}_1 - M_1\right\|^4 \\ & \leq C'\left(\sigma^2+\varsigma^2\log^2(nk)\right)\frac{d\log(nk)}{n}, \end{aligned}$$

where the last two inequalities each hold with probability greater than $1 - 2n^{-10}$.

We denote the estimated and true eigenspaces by $\widehat{U}$ and U^*, respectively. Applying [75, Theorem 2] yields the bound

$$\|\|U^*(U^*)^\top - \widehat{U}\widehat{U}^\top\|\|_{\mathrm{F}}^2 \leq C\left(\frac{\sigma^2+\varsigma^2}{\gamma^2}\right)\frac{kd\log^3(nk)}{n},$$

thereby proving the required result. □

We now proceed to a proof of Lemma 7.

A. Proof of Lemma 7

Recall our decomposition $\Theta^* = A^*(U^*)^\top$, where $U^* \in \mathbb{R}^{d\times k}$ is a matrix of orthonormal columns, and $A^* \in \mathbb{R}^{k\times k}$ is an invertible matrix of coefficients. Since we are always concerned with random variables of the form $\Theta^* X$ with X Gaussian, we may assume without loss of generality by the rotation invariance of the Gaussian distribution that $U^* = [e_1^d\ e_2^d \dots\ e_k^d]$, where e_i^d denotes the ith standard basis vector in $\mathbb{R}^d$.

We let $X_i^j = (X_i, X_{i+1}, \dots, X_j)$ denote a sub-vector of the random vector X, so that by the above argument, we have $\Theta^* X \stackrel{d}{=} A^* X_1^k$.

a) Calculating M_1: Using the shorthand $Z = A^* X_1^k$, we have

$$\begin{aligned} M_1 &= \mathbb{E}[\max(\Theta^* X + b^*)X] \\ &= U^*\mathbb{E}[\max(A^* X_1^k + b^*)X] \\ &= U^*(A^*)^{-1}\mathbb{E}[\max(Z + b^*)Z]. \end{aligned}$$

Now using Stein's lemma,[12] by a calculation similar to the one performed also in Seigel [76] and Liu [77], we have

$$\mathbb{E}[\max(Z + b^*)Z] = \Sigma\pi,$$

where $\pi \in \mathbb{R}^k$ is the vector of probabilities, the j-th of which is given by equation (9), and we have used $\Sigma = A^*(A^*)^\top = (\Theta^*)(\Theta^*)^\top$ to denote the covariance matrix of Z.

Therefore, we have the first moment

$$M_1 = U^*(A^*)^{-1}A^*(A^*)^\top\pi = (\Theta^*)^\top\pi.$$

b) Correlation bound: By computation, we have

$$\begin{aligned} &\langle M_1,\, (\Theta^*)^\top\Sigma^{-1}\mathbf{1}\rangle = \mathbb{E}\left[\max(Z + b^*)\langle Z,\, \Sigma^{-1}\mathbf{1}\rangle\right] \\ &\stackrel{\text{(i)}}{=} \rho\cdot\sqrt{\mathbb{E}\left[(\max(Z + b^*))^2\right]\cdot\mathbb{E}\left[\langle Z,\, \Sigma^{-1}\mathbf{1}\rangle^2\right]} \\ &\stackrel{\text{(ii)}}{=} \rho\varrho\cdot\left\|(\Theta^*)^\top\Sigma^{-1}\mathbf{1}\right\|, \end{aligned}$$

where step (i) follows from the definition (46) of the quantity ρ, and step (ii) from explicitly calculating the expectation and recalling the definition of ϱ.

[12]One can also derive $M_1 = (\Theta^*)^\top\pi$ directly applying Stein's lemma $\mathbb{E}Xf(X) = \mathbb{E}\nabla f(X)$ to $f(x) := \max(\Theta^* X + b^*)$ so that $\nabla f(x)$ equals θ_j^* whenever x belongs to the region when j is maximized.

c) Positive semidefiniteness of M_2: For some $u \in \mathbb{R}^d$, let $f(X) = \max(\Theta^* X + b)$ and $g_u(X) = \langle u,\, X\rangle^2$. Since g_u is an even function, we have $\mathbb{E}[g_u(X)X] = 0$. Furthermore, since both f and g_u are convex, applying Lemma 15 (see Appendix G) yields the bound

$$\mathbb{E}[f(X)g_u(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[g_u(X)],$$

so that substituting yields the bound

$$u^\top\mathbb{E}[\max(\Theta^* X + b)XX^\top]u \geq u^\top\mathbb{E}[\max(\Theta^* X + b)I]u.$$

Since this holds for all $u \in \mathbb{R}^d$, we have shown that the matrix $\mathbb{E}[\max(\Theta^* X + b)(XX^\top - I)]$ is positive semidefinite.

d) Calculating M_2: We now use Stein's lemma to compute an explicit expression for the moment M_2. By the preceding substitution, we have

$$\begin{aligned} M_2 &= \mathbb{E}\Bigg[\max(A^* X_1^k + b^*) \\ &\quad\times\begin{bmatrix} X_1^k(X_1^k)^\top - I_k & X_1^k(X_{k+1}^d)^\top \\ X_{k+1}^d(X_1^k)^\top & X_{k+1}^d(X_{k+1}^d)^\top - I_{d-k}\end{bmatrix}\Bigg] \\ &= \begin{bmatrix}\mathbb{E}\left[\max(A^* X_1^k + b^*)(X_1^k(X_1^k)^\top - I_k)\right] & 0 \\ 0 & 0\end{bmatrix} \end{aligned}$$

Once again using the substitution $Z = A^* X_1^k$ and $\Sigma = A^*(A^*)^\top$, we have

$$M_2 = U^*(A^*)^{-1}\mathbb{E}\left[\max(Z + b^*)(ZZ^\top - \Sigma)\right](A^*)^{-\top}(U^*)^\top,$$

and applying Stein's lemma yields

$$\mathbb{E}\left[\max(Z + b^*)(ZZ^\top - \Sigma)\right] = \Sigma\Pi^\top = \Pi\Sigma,$$

where $\Pi \in \mathbb{R}^{k\times k}$ denotes a matrix with entry i, j given by $\Pi_{i,j} = \mathbb{E}[Z_i\mathbf{1}\left\{Z_j + b_j^* = \max\right\}]$, and the final equality follows by symmetry of the matrix.

Simplifying further, we have

$$M_2 = U^*(A^*)^{-1}\Pi A^*(U^*)^\top.$$

e) Nullspace of M_2: Notice that $\Pi\mathbf{1} = \mathbb{E}[Z] = 0$, so that

$$M_2(\Theta^*)^\top\Sigma^{-1}\mathbf{1} = U^*(A^*)^{-1}\Pi A^*(U^*)^\top U^*(A^*)^\top\Sigma^{-1}\mathbf{1} = 0.$$

f) Rank of M_2 and bound on $\lambda_{k-1}(M_2)$: By the previous claim, we have $\mathsf{rank}(M_2) \leq k - 1$. Furthermore, the matrix M_2 has $d - k$ eigenvalues equal to zero, and the other k of its eigenvalues equal to those of Π, all of which are positive (by the PSD property of M_2), and at least one of which is zero. Thus, it suffices to work with the eigenvalues of Π; in particular, a lower bound on $\lambda_{k-1}(\Pi)$ directly implies a lower bound on $\lambda_{k-1}(M_2)$.

Let us first show that $\lambda_{k-1}(\Pi) > 0$. Since we know that a zero-eigenvector of Π is the all-ones vector $\mathbf{1}$, it suffices to show that $x^\top\Pi x \neq 0$ when $\langle x,\, \mathbf{1}\rangle = 0$. We use the shorthand $x \perp \mathbf{1}$ to denote any such vector.

We now explicitly evaluate the entries of the matrix Π. We denote the jth column of this matrix by Π_j. We have

$$\begin{aligned}\Pi_j &= \mathbb{E}[Z\mathbf{1}\left\{Z_j + b_j^* = \max\right\}] \\ &= \mathbb{E}[\mathbf{1}\cdot Z_j\mathbf{1}\left\{Z_j + b_j^* = \max\right\}] \\ &\quad - \mathbb{E}[(\mathbf{1}\cdot Z_j - Z)\mathbf{1}\left\{Z_j + b_j^* = \max\right\}] \\ &= \mathbf{1}\cdot\mathbb{E}[Z_j\mathbf{1}\left\{Z_j + b_j^* = \max\right\}] \\ &\quad - \mathbb{E}[(\mathbf{1}\cdot Z_j - Z)\mathbf{1}\left\{Z_j + b_j^* = \max\right\}]. \end{aligned} \tag{51}$$

For any $x \perp \mathbf{1}$, we have $x^\top\mathbf{1}\mathbb{E}[Z\mathbf{1}\left\{Z + b^* = \max\right\}]^\top\mathbf{1} = 0$, so that in order to show that $x^\top\Pi x \neq 0$, it suffices to consider just the second term in the expression (51).

In order to focus on this term, consider the matrix Φ with column j given by

$$\Phi_j = \mathbb{E}[(\mathbf{1}\cdot Z_j - Z)\mathbf{1}\left\{Z_j - Z \geq b^* - b_j^*\right\}].$$

where the indicator random variable above is computed element-wise. We are interested in evaluating the eigenvalues of the matrix $-\Phi$.

The quantity Φ_j can be viewed as the first moment of a (lower) truncated, multivariate Gaussian with (original) covariance matrix

$$\kappa_j = (\mathbf{1}\cdot e_j^\top - I)A^*(A^*)^\top(\mathbf{1}\cdot e_j^\top - I)^\top.$$

Recalling the column vectors F^j defined (in equation (47)) for each $j \in [k]$ and applying [78, (11)] (see also Tallis [45] for a similar classical result), we may explicitly evaluate the vector Φ_j, as

$$\begin{aligned}\Phi_j &= \kappa_j F^j \\ &\overset{\text{(iii)}}{=} (\mathbf{1}\cdot e_j^\top - I)A^*(A^*)^\top G_j\end{aligned}$$

where in step (iii), we have let G_j denote a vector in $\mathbb{R}^k$ with entry i given by

$$(G_j)_i = \begin{cases} -(F^j)_i \text{ if } j \neq i \\ \sum_{k\neq j}(F^j)_k \text{ otherwise.}\end{cases}$$

Letting $G \in \mathbb{R}^{k\times k}$ denote the matrix with G_j as its jth column, and for $x \perp \mathbf{1}$, we have

$$x^\top(-\Phi)x = x^\top\Sigma G x,$$

since once again, for each $x \perp \mathbf{1}$, we have $x^\top\mathbf{1}\cdot e_j^\top A^*(A^*)^\top(\mathbf{1}\cdot e_j^\top - I)^\top x = 0$.

Now consider the matrix ΣG. In order to show the claimed bound, it suffices to show that $x^\top\Sigma G x \neq 0$ if $x \perp \mathbf{1}$. We show this by combining two claims:

Claim 1: The nullspace of G is one-dimensional.

Claim 2: Both the left and right eigenvectors of ΣG that correspond to this nullspace are not orthogonal to the $\mathbf{1}$ vector.

We show both claims concurrently. The nullspace of G is clearly non-trivial, since $\mathbf{1}^\top G = 0$. Let us first show, by contradiction, that the left eigenvector corresponding to this nullspace dimension is not orthogonal to the all-ones vector. Toward that x_ℓ denote the aforementioned left eigenvector which also satisfies $\langle x_\ell, \mathbf{1}\rangle = 0$. By virtue of being a left eigenvector, x_ℓ satisfies $\Sigma x_\ell = \mathbf{1}$, or in other words, $x_\ell = \Sigma^{-1}\mathbf{1}$. Since $x_\ell \perp \mathbf{1}$, we have $\mathbf{1}^\top\Sigma\mathbf{1} = 0$, but this contradicts the positive definiteness of Σ.

It remains to establish that the null-space of G is in fact only one-dimensional, and that its right eigenvector is not orthogonal to the all-ones vector. Notice that we may write the matrix as

$$G = (I - P^\top)\operatorname{diag}(G),$$

where we recall the matrix P from equation (48). Since all of the entries of P are positive and sum to 1 along the rows, the matrix P can be viewed as the transition matrix of a Markov chain. Furthermore, since this Markov chain communicates, it is irreducible and aperiodic, with only one eigenvalue equal to 1. Thus, the matrix $I - P^\top$ is rank $k-1$, thereby establishing that the nullspace of G is one-dimensional. Furthermore, the right eigenvector x_r of G is a non-negative vector by the Perron-Frobenius theorem, so that it cannot satisfy $\langle x_r, \mathbf{1}\rangle = 0$.

We have thus established both claims, which together show that $\lambda_{k-1}(M_2) \neq 0$. Further noting that the matrix M_2 is positive semi-definite, we have

$$\lambda_{k-1}(M_2) \geq \min_{j\in[k]} G_{j,j}\cdot\lambda_{\min}(\Sigma)\sqrt{\lambda_{k-1}[(I-P^\top)(I-P)]},$$

and this completes the proof of the claim, and consequently, the lemma. □

Appendix D
Proof of Theorem 3

Recall the matrix $\widehat{V}$ formed by appending a standard basis vector to $\widehat{U}$. First, we show that there is a point among the randomly chosen initializations that is sufficiently close to the true parameters. Toward that end, let $c_0 := \mathsf{B}_{\mathsf{max}}$ (for reasons that will be apparent shortly) and define $\beta^\ell = \widehat{V}\nu^\ell$ for each $\ell \in [M]$ with $\mathbb{M}' = \{\beta^\ell\}_{\ell\in[M]}$. Let

$$\bar{\beta}_j := \operatorname*{argmin}_{\beta\in\mathbb{M}'}\|c_0\beta - \beta_j^*\|.$$

We claim that

$$\max_{j\in[k]}\|c_0\bar{\beta}_j - \beta_j^*\| \leq c_0 r + \mathsf{B}_{\mathsf{max}}\|\widehat{U}\widehat{U}^\top - U^*(U^*)^\top\|_{\text{op}}. \tag{52}$$

Taking this claim as given for the moment, let us proceed with the rest of the proof. Define the shorthand

$$\mathcal{P}(\beta_1,\ldots,\beta_k) := \frac{2}{n}\sum_{i=n/2+1}^{n}\left(\max_{j\in[k]}\langle\xi_i,\,\beta_j\rangle - \max_{j\in[k]}\langle\xi_i,\,\beta_j^*\rangle\right)^2$$

for each set of parameters $\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}$. Let

$$c(\nu_1,\ldots,\nu_k) := \operatorname*{argmin}_{c\geq 0}\frac{2}{n}\sum_{i=n/2+1}^{n}\left(y_i - c\max_{j\in[k]}\langle\xi_i,\,\widehat{V}\nu_j\rangle\right)^2,$$

and recall that $(\nu_1^\sharp,\ldots,\nu_k^\sharp)$ are the minimizers returned by the algorithm; use the shorthand $c^\sharp := c(\nu_1^\sharp,\ldots,\nu_k^\sharp)$. Note that trivially, we have $c^\sharp > 0$ with probability tending to 1 exponentially in n, so that this pathological case in which the initial partition is random can be ignored.

Applying Lemma 17 from Appendix H-B yields the bound

$$\Pr\left\{\mathcal{P}(c^\sharp\beta_1^{(0)},.,c^\sharp\beta_k^{(0)})\geq c_1\left\{\min_{\substack{c\geq 0\\ \nu_1,.,\nu_k\in[M]}}\mathcal{P}(c\widehat{V}\nu_1,\ldots,c\widehat{V}\nu_k)+\frac{\sigma^2 t(\sqrt{\log M}+c_1)}{n}\right\}\right\}\leq e^{-c_2 nt(\sqrt{\log M}+c_1)},$$

valid for all $t\geq\sqrt{\log M}+c_1$ and suitable universal constants c_1 and c_2. Setting $t=\sqrt{\log M}+c_1$, we have on the complementary event that

$$\mathcal{P}(c^\sharp\beta_1^{(0)},\ldots,c^\sharp\beta_k^{(0)})\leq c_1\mathcal{P}(c_0\bar{\beta}_1,\ldots,c_0\bar{\beta}_k)+c_1\frac{\sigma^2\log M}{n}$$

with probability greater than $1-e^{-c_2 n}$.

To complete the proof, let $C(\pi_{\min},k):=c_2\left(\frac{k}{\pi_{\min}}\right)^3$ for a suitable constant c_2 and apply Lemma 16 twice (note that here we use the assumption $n\geq Cd\frac{k^3}{\pi_{\min}^3}\log^2(k/\pi_{\min})$) in order to obtain

$$\begin{aligned}
&\sum_{j\in[k]}\min_{j'\in[k]}\|\beta_j^*-c^\sharp\beta_{j'}^{(0)}\|^2\leq C(\pi_{\min},k)\cdot\mathcal{P}(c^\sharp\beta_1^{(0)},\ldots,\beta_k^{(0)})\\
&\leq c_1\cdot C(\pi_{\min},k)\cdot\left\{\mathcal{P}(c_0\bar{\beta}_1,\ldots,c_0\bar{\beta}_k)+\frac{\sigma^2\log M}{n}\right\}\\
&\leq c_1\cdot C(\pi_{\min},k)\cdot\left\{2\sum_{j=1}^k\|c_0\bar{\beta}_j-\beta_j^*\|^2+\frac{\sigma^2\log M}{n}\right\}\\
&\leq c_1\cdot C(\pi_{\min},k)\cdot\left\{2k\max_{j\in[k]}\|c_0\bar{\beta}_j-\beta_j^*\|^2+\frac{\sigma^2\log M}{n}\right\}\\
&\overset{\text{(ii)}}{\leq} c_1\cdot C(\pi_{\min},k)\left\{4k\left(c_0^2r^2+\mathsf{B}_{\mathsf{max}}^2\|\widehat{U}\widehat{U}^\top-U^*(U^*)^\top\|_{\text{op}}^2\right)+\frac{\sigma^2\log M}{n}\right\}
\end{aligned}$$

on an event of suitably high probability, where step (ii) follows from claim (52).

Finally, note that provided the RHS above is less than $\Delta^2/4$, each minimum on the LHS is attained for a unique index j'. This condition is ensured by the sample size assumption of the theorem; thus, we have

$$\begin{aligned}
&\min_{c>0}\ \mathsf{dist}\left(\left\{c\beta_j^{(0)}\right\}_{j=1}^k,\left\{\beta_j^*\right\}_{j=1}^k\right)\leq c_1\cdot C(\pi_{\min},k)\\
&\times\left\{4k\mathsf{B}_{\mathsf{max}}^2\left(r^2+\|\widehat{U}\widehat{U}^\top-U^*(U^*)^\top\|_{\text{op}}^2\right)+\frac{\sigma^2\log M}{n}\right\}.
\end{aligned}$$

Combining the various probability bounds then completes the proof. □

g) Proof of claim (52)*:* Recall that U^* is a matrix of orthonormal columns spanning the k-dimensional subspace spanned by the vectors $\{\theta_1^*,\ldots,\theta_k^*\}$. Define the matrix

$$V^*=\begin{bmatrix}U^* & 0\\ 0 & 1\end{bmatrix};$$

for each $j\in[k]$, we have $\beta_j^*=V^*\nu_j^*$ for some vector $\nu_j^*\in\mathbb{R}^{k+1}$. Also define the rotation matrix

$$O=\begin{bmatrix}\widehat{U}^\top U^* & 0\\ 0 & 1\end{bmatrix},$$

so that $\widehat{V}O-V^*=\begin{bmatrix}\widehat{U}\widehat{U}^\top U^*-U^* & 0\\ 0 & 0\end{bmatrix}$ and we have $\|\widehat{V}O-V^*\|=\|\widehat{U}\widehat{U}^\top-U^*(U^*)^\top\|$ for any unitarily invariant norm $\|\cdot\|$.

Now for each $j\in[k]$ and $\ell\in[M]$, applying the triangle inequality yields

$$\begin{aligned}
\|c_0\beta^\ell-\beta_j^*\|&\leq\|c_0\widehat{V}\nu^\ell-\widehat{V}O\nu_j^*\|+\|\widehat{V}O\nu_j^*-V^*\nu_j^*\|\\
&\leq\|c_0\nu^\ell-O\nu_j^*\|+\|\nu_j^*\|\|\widehat{V}O-V^*\|_{\text{op}}\\
&\leq c_0r+\mathsf{B}_{\mathsf{max}}\|\widehat{U}\widehat{U}^\top-U^*(U^*)^\top\|_{\text{op}},
\end{aligned}$$

where the last line follows by definition of the r-covering of the set $\mathbb{B}^{k+1}$, which ensures the existence of some index ℓ such that $\|c_0\nu^\ell-O\nu_j^*\|\leq c_0r$. □

APPENDIX E
FUNDAMENTAL LIMITS

In this section, we present two lower bounds: one on the minimax risk of parameter estimation, and another on the risk of the least squares estimator with side-information.

A. Minimax Lower Bounds

Recall our notation Θ^* for the matrix whose columns consist of the parameters $\theta_1^*,\ldots,\theta_k^*$. Assume that the intercepts $b_1^*,\ldots,b_k^*$ are identically zero, so that $\xi_i=x_i$ and $\Xi=X$. For a fixed matrix X, consider the observation model

$$y=\max(X\Theta^*)+\epsilon,\tag{53}$$

where $y\in\mathbb{R}^n$, the noise $\epsilon\sim\mathcal{N}(0,\sigma^2I_n)$ is chosen independently of X, and the max function is computed row-wise.

Proposition 2: There is an absolute constant C such that the minimax risk of estimation satisfies

$$\inf_{\widehat{\Theta}}\sup_{\Theta^*\in\mathbb{R}^{k\times d}}\mathbb{E}\left[\frac{1}{n}\|X(\widehat{\Theta}-\Theta^*)\|_{\text{F}}^2\right]\geq C\frac{\sigma^2kd}{n}.$$

Here, the expectation is taken over the noise ϵ, and infimum is over all measurable functions of the observations (X,y). Indeed, when X is a random Gaussian matrix, it is well conditioned and has singular values of the order $\sqrt{n}$, so that this bound immediately yields

$$\inf_{\widehat{\Theta}}\sup_{\Theta^*\in\mathbb{R}^{k\times d}}\mathbb{E}\left[\frac{1}{n}\|\widehat{\Theta}-\Theta^*\|_{\text{F}}^2\right]\geq C\frac{\sigma^2kd}{n}.$$

Let us now provide a proof of the proposition.

Proof: The proof is based on a standard application of Fano's inequality (see, e.g., Wainwright [70, Chapter 15] and Tsybakov [79, Chapter 2]). For a tolerance level $\delta>0$ to be chosen, we choose the local set

$$F=\left\{X\Theta\in\mathbb{R}^{n\times k}\ \middle|\ \|X\Theta^\top\|_{\text{F}}\leq 4\delta\sqrt{kn}\right\}$$

and let $\left\{X\Theta^1,\ldots,X\Theta^M\right\}$ be a $2\delta\sqrt{kn}$-packing of the set in the Frobenius norm. This can be achieved by packing the j-th column $Q_j:=\left\{X\theta_j\mid\|X\theta_j\|_2\leq 4\delta\sqrt{n}\right\}$ at level $2\delta\sqrt{n}$ in ℓ_2 norm for all $j\in[k]$. Standard results yield the bound $\log M\geq C_1\cdot kd\log 2$.

For each $i \neq j$, we have

$$2\delta\sqrt{k} \leq \frac{\|X(\Theta^i - \Theta^j)\|_{\mathrm{F}}}{\sqrt{n}} \leq 8\delta\sqrt{k}. \tag{54}$$

Let $\mathbb{P}_j = \mathcal{N}\left(\max(X(\Theta^j)), \sigma^2 I_n\right)$ denote the distribution of the observation vector y when the true parameter is Θ^j. We thus obtain

$$\begin{aligned} D_{\mathsf{KL}}(\mathbb{P}_j \parallel \mathbb{P}_i) &= \frac{1}{2\sigma^2}\left\|\max(X(\Theta^j)) - \max(X(\Theta^i))\right\|_2^2 \\ &\leq \frac{1}{2\sigma^2}\|X(\Theta^j - \Theta^i)\|_{\mathrm{F}}^2, \end{aligned}$$

where the inequality follows since the $\max$ function is 1-Lipschitz in ℓ_2 norm. Putting together the pieces yields

$$D_{\mathsf{KL}}(\mathbb{P}_j \parallel \mathbb{P}_i) \leq \frac{32k\delta^2 n}{\sigma^2},$$

so that the condition

$$\frac{\frac{1}{M^2}\sum_{i,j} D_{\mathsf{KL}}(\mathbb{P}_{\Theta^j} \parallel \mathbb{P}_{\Theta^k}) + \log 2}{\log M} \leq \frac{1}{2}$$

is satisfied with the choice $\delta^2 = C\frac{\sigma^2 d}{n}$. Finally, applying Fano's inequality (see, e.g., [70, Proposition 15.2]) yields the minimax lower bound

$$\inf_{\widehat{\Theta}} \sup_{\Theta^*} \mathbb{E}\left[\frac{1}{n}\|X(\widehat{\Theta} - \Theta^*)\|_{\mathrm{F}}^2\right] \geq C\frac{\sigma^2 kd}{n}. \tag{55}$$

□

B. Performance of Unconstrained Least Squares With Side-Information

In this section, we perform an explicit computation when $k = 3$ and $d = 2$ to illustrate the cubic $\pi_{\min}$ dependence of the error incurred by the unconstrained least squares estimator, even when provided access to the true partition $\{S_j(\beta_1^*, \ldots, \beta_3^*)\}_{j=1}^3$.

We begin by defining our unknown parameters. For a scalar $\alpha \in (0, \pi/4)$, let

$$\theta_1^* = \sin(\alpha) \cdot e_1, \theta_2^* = \cos(\alpha) \cdot e_2, \text{ and } \theta_3^* = -\cos(\alpha) \cdot e_2,$$

and set $b_j^* = 0$ for $j = 1, 2, 3$.

Now an explicit computation yields that the cone on which θ_1^* attains the maximum is given by

$$\begin{aligned} \mathcal{C}_1 &:= \left\{x \in \mathbb{R}^2 : \langle x, \theta_1^*\rangle \geq \max_{j\in[k]}\langle x, \theta_j^*\rangle\right\} \\ &= \left\{x \in \mathbb{R}^2 : x_1 \geq 0,\ |x_2| \leq x_1 \tan(\alpha)\right\}. \end{aligned}$$

Now consider a Gaussian random vector in $\mathbb{R}^2$ truncated to that cone. In particular, consider a two-dimensional random variable W with density $\psi(x)\mathbf{1}\{x \in \mathcal{C}_1\}/\operatorname{vol}(\mathcal{C}_1)$, where ψ is the two-dimensional standard Gaussian density and $\operatorname{vol}(S)$ denotes the Gaussian volume of a set S. Note that we have $\operatorname{vol}(\mathcal{C}_1) = \alpha/\pi$ by construction.

Let us now compute the second order statistics of W, using polar coordinates with R^2 denoting a χ_2^2 random variable. The individual second moments take the form

$$\mathbb{E}[W_1^2] = \frac{\pi}{\alpha}\mathbb{E}[R^2]\left(\frac{1}{2\pi}\int_{-\alpha}^{\alpha}\cos^2\phi d\phi\right) = 1,$$

and

$$\begin{aligned} \mathbb{E}[W_2^2] &= \frac{\pi}{\alpha}\mathbb{E}[R^2]\left(\frac{1}{2\pi}\int_{-\alpha}^{\alpha}\sin^2\phi d\phi\right) \\ &= \frac{1}{\alpha}\left(\alpha - \sin(2\alpha)/2\right) \sim \alpha^2. \end{aligned}$$

On the other hand, the cross terms are given by

$$\mathbb{E}[W_1 W_2] = \frac{\pi}{\alpha}\mathbb{E}[R^2]\left(\frac{1}{2\pi}\int_{-\alpha}^{\alpha}\sin(\phi)\cos(\phi)d\phi\right) = 0.$$

Thus, it can be verified that for all $\alpha \in [0, \pi/4]$, the second moment matrix of W has a tuple of singular values $(1, c\alpha^2)$ for an absolute constant c.

Let us now use this calculation to reason about the least squares estimator. Drawing n samples from the Gaussian distribution on $\mathbb{R}^2$, we expect $n_1 \sim \frac{\alpha}{\pi}n$ of them to fall in the set $\mathcal{C}_1$ with high probability. Collect these samples as rows of a matrix X_1. When n is large enough, i.e., on the order of α^{-3}, standard bounds (as in Section B-A.3) can be applied to explicitly evaluate the singular values of the matrix $\frac{1}{n_1}X_1^\top X_1$. In particular, we have

$$\lambda_1\left(\frac{1}{n_1}X_1^\top X_1\right) = c' \qquad \text{and} \qquad \lambda_2\left(\frac{1}{n_1}X_1^\top X_1\right) = c\alpha^2.$$

We now provide the $n_1 \times 2$ matrix X_1 as side information to a procedure whose goal is to estimate the unknown parameters. Clearly, given this matrix, a natural procedure to run in order to estimate θ_1^* is the (unconstrained) least squares estimator on these samples, which we denote by $\widehat{\theta}_1$. As is well known, the rate obtained (in the fixed design setting) by this estimator with σ-sub-Gaussian noise is given by

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\theta}_1 - \theta_1^*\|^2\right] &= \sigma^2\,\mathsf{tr}(X_1^\top X_1)^{-1} \\ &= \sigma^2\frac{1}{n_1}\left(c\alpha^{-2} + c'\right) \\ &\sim \sigma^2\frac{1}{\alpha^3 n}, \end{aligned}$$

where the last two relations hold with exponentially high probability in n. We have thus shown that the unconstrained least squares estimator (even when provided with additional side information) attains an error having cubic dependence on $\alpha \sim \pi_{\min}$. While this does not constitute an information theoretic lower bound, our calculation provides some evidence for the fact that, at least when viewed in isolation, the dependence of our statistical error bound (15) on $\pi_{\min}$ is optimal for Gaussian covariates.

APPENDIX F
BACKGROUND AND TECHNICAL LEMMAS USED IN THE PROOF OF THEOREM 1

In this section, we collect statements and proofs of some technical lemmas used in the proofs of our results concerning the AM algorithm.

A. Bounds on the "Volumes" of Wedges in $\mathbb{R}^d$

For a pair of scalars (w, w') and d-dimensional vectors (u, u'), recall that we define the *wedge* formed by the $d+1$-dimensional vectors $v = (u,\ w)$ and $v' = (u',\ w')$ as the region

$$W(v, v') = \{x \in \mathbb{R}^d : (\langle x,\ u\rangle + w) \cdot (\langle x,\ u'\rangle + w') \leq 0\}.$$

Note that the wedge is a purely geometric object.

For any set $\mathcal{C} \subseteq \mathbb{R}^d$, let

$$\mathfrak{vol}(\mathcal{C}) = \Pr_{X \sim \mathcal{N}(0, I_d)} \{X \in \mathcal{C}\}$$

denote the volume of the set under the measure corresponding to the covariate distribution.

We now bound the volume of a wedge for the Gaussian distribution.

Lemma 9: Suppose that for a pair of scalars (w, w'), d-dimensional vectors (u, u'), and $v = (u,\ w)$ and $v' = (u',\ w')$, we have $\frac{\|v - v'\|}{\|u\|} < 1/2$. Then, there is a positive constant C such that

$$\mathfrak{vol}(W(v, v')) \leq C \frac{\|v - v'\|}{\|u\|} \log^{1/2}\left(\frac{2\|u\|}{\|v - v'\|}\right).$$

1) Proof of Lemma 9: Using the notation $\xi = (x,\ 1) \in \mathbb{R}^{d+1}$ to denote the appended covariate, we have

$$\mathfrak{vol}(W(v, v')) = \Pr\left\{\langle \xi,\ v\rangle \cdot \langle \xi,\ v'\rangle \leq 0\right\},$$

where the probability is computed with respect to Gaussian measure.

In order to prove a bound on this probability, we begin by bounding the associated indicator random variable as

$$\begin{aligned}\mathbf{1}\left\{\langle \xi,\ v\rangle \cdot \langle \xi,\ v'\rangle \leq 0\right\} &\leq \mathbf{1}\left\{\langle \xi,\ v' - v\rangle^2 \geq \langle \xi,\ v\rangle^2\right\} \\ &\leq \mathbf{1}\left\{\langle \xi,\ v' - v\rangle^2 \geq t\right\} + \mathbf{1}\left\{\langle \xi,\ v\rangle^2 \leq t\right\}, \qquad (56)\end{aligned}$$

where inequality (56) holds for all $t \geq 0$. In order to bound the expectation of the second term, we write

$$\begin{aligned}\Pr\left\{\langle \xi,\ v\rangle^2 \leq t\right\} &= \Pr\left\{\|u\|^2 \chi^2_{nc} \leq t\right\} \\ &\overset{\text{(i)}}{\leq} \left(\frac{et}{\|u\|^2}\right)^{1/2}\end{aligned}$$

where χ^2_{nc} is a non-central chi-square random variable centered at $\frac{w}{\|u\|}$, and step (i) follows from standard χ^2 tail bounds (see Lemma 14).

It remains to control the expectation of the first term on the RHS of inequality (56). We have

$$\begin{aligned}&\Pr\left\{\langle \xi,\ v' - v\rangle^2 \geq t\right\} \\ &\quad\leq \Pr\left\{2\langle x,\ u' - u\rangle^2 + 2(w' - w)^2 \geq t\right\} \\ &\quad\leq \Pr\left\{\|u - u'\|^2 \chi^2 \geq \frac{t}{2} - \|v - v'\|^2\right\}.\end{aligned}$$

Now, invoking a standard sub-exponential tail bound on the upper tail of a χ^2 random variable yields

$$\begin{aligned}&\Pr\left\{\langle \xi,\ v' - v\rangle^2 \geq t\right\} \\ &\leq c_1 \exp\left(-\frac{c_2}{\|u - u'\|^2}\left\{\frac{t}{2} - \|v - v'\|^2\right\}\right) \\ &\leq c_1 \exp\left(-\frac{c_2}{\|v - v'\|^2}\left\{\frac{t}{2} - \|v - v'\|^2\right\}\right).\end{aligned}$$

Putting all the pieces together, we obtain

$$\begin{aligned}\mathfrak{vol}(W(v, v')) \leq c_1 \exp&\left(-\frac{c_2}{\|v - v'\|^2}\left\{\frac{t}{2} - \|v - v'\|^2\right\}\right) \\ &+ \left(\frac{et}{\|u\|^2}\right)^{1/2}.\end{aligned}$$

Substituting $t = 2c\|v - v'\|^2 \log(2\|u\|/\|v - v'\|)$, which is a valid choice provided $\frac{\|v - v'\|}{\|u\|} < 1/2$, yields the desired result. □

B. Growth Functions and Uniform Empirical Concentration

We now briefly introduce growth functions and uniform laws derived from them, and refer the interested reader to Mohri *et al.* [80] for a more in-depth exposition on these topics.

We define growth functions in the general multi-class setting [54]. Let $\mathcal{X}$ denote a set, and let $\mathcal{F}$ denote a family of functions mapping $\mathcal{X} \mapsto \{0, 1, \ldots, k-1\}$. The *growth function* $\Pi_{\mathcal{F}} : \mathbb{N} \to \mathbb{R}$ of $\mathcal{F}$ is defined via

$$\Pi_{\mathcal{F}}(n) := \max_{x_1, \ldots, x_n \in \mathcal{X}} \left|\{\{f(x_1), f(x_2), \ldots, f(x_n)\} : f \in \mathcal{F}\}\right|.$$

In words, it is the cardinality of all possible labelings of n points in the set $\mathcal{X}$ by functions in the family $\mathcal{F}$.

A widely studied special case arises in the case $k = 2$, with the class of binary functions. In this case, a natural function class $\mathcal{F}$ is formed by defining $\mathcal{C}$ to be a family of subsets of $\mathcal{X}$, and identifying each set $C \in \mathcal{C}$ with its indicator function $f_C := 1_C : \mathcal{X} \to \{0, 1\}$. In this case, define $\mathcal{F}_{\mathcal{C}} = \{f_C : C \in \mathcal{C}\}$. A bound on the growth function for such binary function provides following guarantee for the uniform convergence for the empirical measures of sets belonging to $\mathcal{C}$.

Lemma 10 (Theorem 2 in [81]): Let $\mathcal{C}$ be a family of subsets of a set $\mathcal{X}$. Let μ be a probability measure on $\mathcal{X}$, and let $\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ be the empirical measure obtained from m independent copies of a random variable X with distribution μ. For every u such that $m \geq 2/u^2$, we have

$$\Pr\left\{\sup_{C \in \mathcal{C}} |\hat{\mu}_m(C) - \mu(C)| \geq u\right\} \leq 4\Pi_{\mathcal{F}_{\mathcal{C}}}(2m)\exp(-mu^2/16). \qquad (57)$$

We conclude this section by collecting some results on the growth functions of various function classes. For our development, it will be specialized to the case $\mathcal{X} = \mathbb{R}^d$.

Define the class of binary functions $\mathcal{F}_{\mathcal{H}}$ as the set of all functions of the form

$$f_{\theta,b}(x) \ := \frac{\mathsf{sgn}(\langle x,\ \theta\rangle + b) + 1}{2};$$

specifically, let $\mathcal{F}_{\mathcal{H}} := \left\{f_{\theta,b} : \theta \in \mathbb{R}^d, b \in \mathbb{R}\right\}$. In particular, these are all functions that can be formed by a d-dimensional hyperplane.

Using the shorthand $B_1^k = \{B_1, \ldots, B_k\}$, define the binary function

$$g_{\theta_1^k, b_1^k}(x) := \prod_{i=1}^{k} f_{\theta_i, b_i}(x),$$

and the binary function class corresponding to the intersection of k hyperplanes

$$\mathcal{G}_{\mathcal{H}^k} := \left\{g_{\theta_1^k, b_1^k} : \theta_1, \ldots, \theta_k \in \mathbb{R}^d,\ b_1, \ldots, b_k \in \mathbb{R}\right\}.$$

Finally, we are interested in the argmax function over hyperplanes. Here, define the function

$$m_{\theta_1^k, b_1^k}(x) := \operatorname*{argmax}_{j \in [k]} \left(\langle \theta_j,\, x\rangle + b_j\right) - 1,$$

mapping $\mathbb{R}^d \mapsto \{0, \ldots, k-1\}$. The function class that collects all such functions is given by

$$\mathcal{M}_k := \left\{m_{\theta_1^k, b_1^k} : \theta_1, \ldots, \theta_k \in \mathbb{R}^d,\ b_1, \ldots, b_k \in \mathbb{R}\right\}.$$

The following results bound the growth functions of each of these function classes. We first consider the function classes $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{G}_{\mathcal{H}^k}$, for which bounds on the VC dimension directly yield bounds on the growth function.

Lemma 11 (Sauer-Shelah (e.g. Section 3 of Mohri et al. [80])): We have

$$\Pi_{\mathcal{F}_{\mathcal{H}}}(n) \leq \left(\frac{en}{d+1}\right)^{d+1}, \text{ and} \tag{58}$$

$$\Pi_{\mathcal{G}_{\mathcal{H}^k}}(n) \leq \left(\frac{en}{d+1}\right)^{k(d+1)}. \tag{59}$$

The second bound can be improved (see, e.g. [82]), but we state the version obtained by a trivial composition of individual halfspaces.

The following bound on the growth function of the class $\mathcal{M}_k$ is also known.

Lemma 12 (Theorem 3.1 of Daniely et al. [54]): For an absolute constant C, we have

$$\Pi_{\mathcal{M}_k}(n) \ \leq \left(\frac{en}{Ck(d+1)\log(kd)}\right)^{Ck(d+1)\log(kd)}.$$

C. Singular Value Bound

We now state and prove a technical lemma that bound the maximum singular value of a matrix whose rows are drawn from a sub-Gaussian distribution.

Lemma 13: Suppose that the covariates are drawn i.i.d. from a η-sub-Gaussian distribution. Then for a fixed subset $S \in [n]$ of size ℓ and each $t \geq 0$, we have

$$\Pr\left\{\lambda_{\max}\left(\Xi_S^\top \Xi_S\right) \geq \ell + \widetilde{\eta}^2(\sqrt{\ell d} + d + \ell t)\right\} \leq 2e^{-\ell \min\{t, t^2\}},$$

where $\widetilde{\eta} = \max\{\eta, 1\}$.

1) Proof of Lemma 13: Let $\{z_i\}_{i=1}^{\ell}$ denote i.i.d. Rademacher variables, and collect these in an ℓ-dimensional vector z. Let $D = \operatorname{diag}(z)$ denote a diagonal matrix, and note that by unitary invariance of the singular values, the singular values of the matrix $\widetilde{\Xi}_S = D\Xi_S$ are the same as those of Ξ_S.

By construction, the matrix $\widetilde{\Xi}_S$ has i.i.d. rows, and the i-th row is given by $z_i(x_i,\ 1)$. For a $d+1$ dimensional vector $\widetilde{\lambda} = (\lambda,\ w)$ with $\lambda \in \mathbb{R}^d$ and $w \in \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\exp(\langle \widetilde{\lambda},\, z_i(x_i,\ 1)\rangle)\right] \\
&= \frac{e^w}{2} \cdot \mathbb{E}\left[\exp(\langle \lambda,\, x_i\rangle)\right] + \frac{e^{-w}}{2} \cdot \mathbb{E}\left[\exp(-\langle \lambda,\, x_i\rangle)\right] \\
&= \exp(\|\lambda\|^2 \eta^2/2) \cdot \frac{1}{2}\left(e^w + e^{-w}\right) \\
&\leq \exp(\|\lambda\|^2 \eta^2/2) \cdot \exp(w^2/2) \leq \exp(\|\widetilde{\lambda}\|^2 \widetilde{\eta}^2/2).
\end{aligned}$$

where we have used the fact that x_i is zero-mean and η sub-Gaussian.

Since the rows of $\widetilde{\Xi}_S$ are i.i.d., zero-mean, and $\widetilde{\eta}$-sub-Gaussian, applying [70, Theorem 6.15] immediately yields the lemma. □

D. Anti-Concentration of χ^2 Random Variable

The following lemma shows the anti-concentration of the central and non-central χ^2 random variable.

Lemma 14: Let Z_ℓ and Z'_ℓ denote central and non-central χ^2 random variables with ℓ degrees of freedom, respectively. Then for all $p \in [0, \ell]$, we have

$$\begin{aligned}
\Pr\{Z'_\ell \leq p\} \leq \Pr\{Z_\ell \leq p\} &\leq \left(\frac{p}{\ell}\exp\left(1 - \frac{p}{\ell}\right)\right)^{\ell/2} \\
&= \exp\left(-\frac{\ell}{2}\left[\log\frac{\ell}{p} + \frac{p}{\ell} - 1\right]\right)
\end{aligned} \tag{60}$$

1) Proof of Lemma 14: The fact that $Z'_\ell \overset{st.}{\leq} Z_\ell$ follows from standard results that guarantee that central χ^2 random variables stochastically dominate their non-central counterparts.

The tail bound is a simple consequence of the Chernoff bound. In particular, we have for all $\lambda > 0$ that

$$\begin{aligned}
\Pr\{Z_\ell \leq p\} &= \Pr\{\exp(-\lambda Z_\ell) \geq \exp(-\lambda p)\} \\
&\leq \exp(\lambda p)\mathbb{E}\left[\exp(-\lambda Z_\ell)\right] \\
&= \exp(\lambda p)(1 + 2\lambda)^{-\frac{\ell}{2}}.
\end{aligned} \tag{61}$$

where in the last step, we have used $\mathbb{E}\left[\exp(-\lambda Z_\ell)\right] = (1 + 2\lambda)^{-\frac{\ell}{2}}$, which is valid for all $\lambda > -1/2$. Minimizing the last expression over $\lambda > 0$ then yields the choice $\lambda^* = \frac{1}{2}\left(\frac{\ell}{p} - 1\right)$, which is greater than 0 for all $0 \leq p \leq \ell$. Substituting this choice back into equation (61) proves the lemma. □

APPENDIX G
BACKGROUND AND TECHNICAL LEMMAS USED IN THE PROOF OF THEOREM 2

We begin by stating a result of Harge [83, Theorem 1.2] (see also Hu [84]) that guarantees that convex functions of a Gaussian random vector are positively correlated. We state it below in the notation of the current paper.

Lemma 15 ([83]): Let f and g be two convex functions on $\mathbb{R}^d$, and let X be a standard d-dimensional Gaussian vector. Then

$$\mathbb{E}[f(X)g(X)] \geq (1 + \langle m(g),\, m(f)\rangle)\mathbb{E}[f(X)]\mathbb{E}[g(X)], \tag{62}$$

where for any d-variate function h, we have $m(h) = \frac{\mathbb{E}[Xh(X)]}{\mathbb{E}[h(X)]}$.

We also prove Lemma 8, which was used in the proof of Theorem 2.

A. Proof of Lemma 8

We prove each bound separately. First, by the rotation invariance of the Gaussian distribution, we may assume that $U^* = [e_1^d \ \dots \ e_k^d]$, so that the max is computed as a function of the k coordinates $X_1, \dots X_k$.

We also define some events that we make use of repeatedly in the proofs. For each $i \in [n]$, define the events

$$\mathcal{E}_i = \{|x_{i,j}| \leq 5\sqrt{\log(2nk)} \text{ for all } 1 \leq j \leq k\}, \text{ and}$$
$$\mathcal{F}_i = \{|\epsilon_i| \leq 5\sigma\sqrt{\log(2n)}\}.$$

Note that by standard sub-Gaussian tail bounds, we have $\Pr\{\mathcal{E}_i^c\} \leq 2n^{-12}$ and $\Pr\{\mathcal{F}_i^c\} \leq 2n^{-12}$ for each $i \in [n]$. For notational convenience, define for each i the modified covariate $z_i = x_i \cdot \mathbf{1}\{\mathcal{E}_i\}$.

We have

$$\begin{aligned}|\max(\Theta^* z_i + b^*)| &\leq C \max_{j\in[k]} \|\theta_j^*\|_1 \sqrt{\log(nk)} + |b_j^*| \\ &\leq \left(C\sqrt{\log(nk)}\right)\varsigma\end{aligned}$$

almost surely, where in the second bound, we have used the shorthand $\varsigma = \max_j \left(\|\theta_j^*\|_1 + \|b_j^*\|_1\right)$ as defined in equation (18). With this setup in place, we are now ready to prove both deviation bounds.

1) Proof of Bound (50a)*:* Let us first bound the deviation of the first moment. We work with the decomposition

$$\widehat{M}_1 - M_1 = \frac{2}{n}\sum_{i=1}^{n/2} \underbrace{\max(\Theta^* x_i + b^*)x_i - \mathbb{E}[\max(\Theta^* X + b^*)X]}_{T_i^1} + \frac{2}{n}\sum_{i=1}^{n/2} \underbrace{\epsilon_i x_i}_{T_i^2}.$$

By triangle inequality, it suffices to bound the norms of each of the two sums separately. We now use the further decomposition

$$\begin{aligned}T_i^1 = &\underbrace{\max(\Theta^* x_i + b^*)x_i - \max(\Theta^* z_i + b^*)z_i}_{P_i} \\ &+ \underbrace{\max(\Theta^* z_i + b^*)z_i - \mathbb{E}[\max(\Theta^* z_i + b^*)z_i]}_{Q_i} \\ &+ \underbrace{\mathbb{E}[\max(\Theta^* z_i + b^*)z_i] - \mathbb{E}[\max(\Theta^* x_i + b^*)x_i]}_{R_i}.\end{aligned}$$

Since $z_i = x_i$ with probability greater than $1 - 2n^{-12}$, the term $P_i = 0$ on this event.

Also, for each fixed $j \in [k]$, applying the Hoeffding inequality yields the bound

$$\Pr\left\{\left|\frac{2}{n}\sum_{i=1}^{n/2} Q_{i,j}\right| \geq t\right\} \leq 2\exp\left\{-\frac{nt^2}{8C^2\varsigma^2(\log(nk))^2}\right\}.$$

On the other hand, for $j \in [d] \setminus [k]$, we have

$$\begin{aligned}\left|\frac{2}{n}\sum_{i=1}^{n/2} Q_{i,j}\right| &\leq \varsigma\frac{2}{n}\sum_{i=1}^{n/2} z_{i,j} \\ &= \varsigma\left|\frac{2}{n}\sum_{i=1}^{n/2} x_{i,j}\right|.\end{aligned}$$

Standard Gaussian tail bounds then yield

$$\Pr\left\{\left|\frac{2}{n}\sum_{i=1}^{n/2} Q_{i,j}\right| \geq \varsigma t\sqrt{\log(nk)}\right\} \leq 2\exp\left\{-\frac{nt^2}{8}\right\}$$

for each $t \geq 0$. Putting together the pieces with a union bound and choosing constants appropriately, we then have

$$\begin{aligned}\Pr\Bigg\{\left\|\frac{2}{n}\sum_{i=1}^{n/2} Q_i\right\|^2 &\geq \frac{1}{n}\cdot Ck\varsigma^2(\log(nk))^2 \\ &+ \frac{1}{n}\cdot C'(d-k)\varsigma^2\log(nk)\Bigg\} \leq 2dn^{-12}.\end{aligned}$$

It remains to handle the final terms $\{R_i\}_{i=1}^n$. Note that when $j \notin [k]$, we have $R_{i,j} = 0$. It therefore suffices to bound the various $R_{i,j}$ terms when $j \in [k]$. We have

$$\begin{aligned}|R_{i,j}| &= |\mathbb{E}[\max(\Theta^* z_i + b^*)z_{i,j}] \\ &\quad - \mathbb{E}[\max(\Theta^* x_i + b^*)x_{i,j}\mathbf{1}\{\mathcal{E}_i\}] \\ &\quad - \mathbb{E}[\max(\Theta^* x_i + b^*)x_{i,j}\mathbf{1}\{\mathcal{E}_i^c\}]| \\ &= |\mathbb{E}[\max(\Theta^* x_i + b^*)x_{i,j}\mathbf{1}\{\mathcal{E}_i^c\}]|\end{aligned}$$

Expanding this further, we have

$$\begin{aligned}|R_{i,j}| &\leq \mathbb{E}[\max_{\ell\in[k]}(|\langle\theta_\ell^*,\, x_i\rangle| + |b_\ell^*|)|x_{i,j}|\mathbf{1}\{\mathcal{E}_i^c\}] \\ &\leq \mathbb{E}\left[|x_{i,j}|\|x_i\|_\infty(\|\Theta^*\|_{1,\infty} + \|b^*\|_\infty)\mathbf{1}\{\mathcal{E}_i^c\}\right] \\ &= \varsigma\mathbb{E}\left[|x_{i,j}|\|x_i\|_\infty\mathbf{1}\{\mathcal{E}_i^c\}\right] \\ &\leq \varsigma\sum_{\ell=1}^{k}\mathbb{E}\left[|x_{i,j}||x_{i,\ell}|\mathbf{1}\{\mathcal{E}_i^c\}\right].\end{aligned}$$

Note that for a pair (X_1, X_2) of i.i.d. random variables, Jensen's inequality yields the bounds

$$\mathbb{E}[|X_1X_2|\mathbf{1}\{X_1, X_2 \geq \lambda\}] \leq \mathbb{E}[X_1^2\mathbf{1}\{|X_1| \geq \lambda\}]\forall\lambda \geq 0, \text{ and}$$
$$\mathbb{E}[|X_1|\mathbf{1}\{|X_1| \geq \lambda\}] \leq \mathbb{E}[X_1^2\mathbf{1}\{|X_1| \geq \lambda\}]\forall\lambda \geq 1.$$

Furthermore, if X is a standard Gaussian random variable, then a simple calculation (see also Burkardt [85]) yields the bound

$$\mathbb{E}[X^2 \mid |X| \geq \lambda] \leq \frac{1}{2\sqrt{2\pi}}\lambda e^{-\lambda^2/2}, \text{ for all } \lambda \geq \sqrt{2}.$$

Putting together the pieces with $\lambda = 5\sqrt{\log(2nk)}$, we have

$$|R_{i,j}|^2 \leq Ck^2\varsigma^2 \log(nk)(nk)^{-24},$$

and summing over $j \in [k]$ yields the bound

$$\left\| \frac{2}{n} \sum_{i=1}^{n/2} R_i \right\|^2 \leq Ck^2\varsigma^2 \log(nk)(nk)^{-24}.$$

Finally, putting together the pieces with a union bound yields the desired bound on the random variable $\left\| \frac{2}{n} \sum_{i=1}^{n/2} T_i^1 \right\|$.

The second term can be bounded more easily; in particular, on the intersection of the events $\{\mathcal{F}_i\}_{i=1}^n$, we have

$$\begin{aligned} \left\| \frac{2}{n} \sum_{i=1}^{n/2} T_i^2 \right\|^2 &\leq C\sigma^2 \log n \left\| \frac{2}{n} \sum_{i=1}^{n/2} x_i \right\|^2 \\ &\leq C\sigma^2 \frac{(d + \log n) \log n}{n}, \end{aligned}$$

where the final bound holds with probability greater than $1 - cn^{-10}$. Finally, putting the bounds together yields the result.

2) Proof of Bound (50b)*:* Once again, we decompose the required term as

$$\begin{aligned} \widehat{M}_2 - M_2 = \frac{2}{n} \sum_{i=1}^{n/2} \underbrace{\max(\Theta^* x_i + b^*) \left(x_i x_i^\top - I_d \right)}_{\tau_i^1} \\ + \frac{2}{n} \sum_{i=1}^{n/2} \underbrace{\epsilon_i \left(x_i x_i^\top - I_d \right)}_{\tau_i^2}. \end{aligned}$$

We use the further decomposition

$$\tau_i^1 = \phi_i + \kappa_i + \rho_i$$

where,

$$\begin{aligned} \phi_i &= \max(\Theta^* x_i + b^*) \left(x_i x_i^\top - I_d \right) \\ &\quad - \max(\Theta^* z_i + b^*) \left(z_i z_i^\top - I_d \right), \\ \kappa_i &= \max(\Theta^* z_i + b^*) \left(z_i z_i^\top - I_d \right) \\ &\quad - \mathbb{E}[\max(\Theta^* z_i + b^*) \left(z_i z_i^\top - I_d \right)], \end{aligned}$$

and

$$\begin{aligned} \rho_i &= \mathbb{E}[\max(\Theta^* z_i + b^*) \left(z_i z_i^\top - I_d \right)] \\ &\quad - \mathbb{E}[\max(\Theta^* x_i + b^*) \left(x_i x_i^\top - I_d \right)]. \end{aligned}$$

As before, since $z_i = x_i$ with probability greater than $1 - 2n^{-12}$, the term $\phi_i = 0$ on this event.

Let us further decompose κ_i as

$$\kappa_i = \kappa_i^{(1)} + \kappa_i^{(2)} + I_d \cdot \kappa_i^{(3)},$$

with

$$\begin{aligned} \kappa_i^{(1)} &= \left(\max(\Theta^* z_i + b^*) + \varsigma \sqrt{\log(nk)} \right) z_i z_i^\top \\ &\quad - \mathbb{E}\left[\left(\max(\Theta^* z_i + b^*) + \varsigma \sqrt{\log(nk)} \right) z_i z_i^\top \right], \\ \kappa_i^{(2)} &= \varsigma \sqrt{\log(nk)} \mathbb{E}\left[z_i z_i^\top \right] - I_d, \end{aligned}$$

and

$$\kappa_i^{(3)} = \left(\mathbb{E}[\max(\Theta^* z_i + b^*) - \max(\Theta^* z_i + b^*) \right),$$

so that

$$\begin{aligned} \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i \right\|_{\text{op}} \leq \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i^{(1)} \right\|_{\text{op}} + \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i^{(2)} \right\|_{\text{op}} \\ + \left| \frac{2}{n} \sum_{i=1}^{n/2} \kappa_i^{(3)} \right|. \end{aligned}$$

Since $|\max(\Theta^* z_i + b^*)| \leq C\varsigma\sqrt{\log(nk)}$, the random vector $\sqrt{\max(\Theta^* z_i + b^*) + C\varsigma\sqrt{\log(nk)}} z_i$ is well-defined and bounded; sub-Gaussian concentration bounds [70] can therefore be applied to obtain

$$\begin{aligned} \mathbb{P}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \kappa_i^{(1)} \right\|_{\text{op}} \geq c_1 \varsigma^2 (\log(nk))^2 \left\{ \sqrt{\frac{d}{n}} + \frac{d}{n} + \delta \right\} \right] \\ \leq c_2 \exp\left(-n \min(\delta, \delta^2) \right) \end{aligned}$$

where $\varsigma_1 \log(nk) = \max(\Theta^* z_i + b^*) + \varsigma\sqrt{\log(nk)} \leq 2\varsigma \log(nk)$. Reasoning similarly for the second term, we have

$$\begin{aligned} \mathbb{P}\left[\left\| \frac{1}{n} \sum_{i=1}^{n} \kappa_i^{(2)} \right\|_{\text{op}} \geq c_1 \varsigma^2 (\log(nk))^2 \left\{ \sqrt{\frac{d}{n}} + \frac{d}{n} + \delta \right\} \right] \\ \leq c_2 \exp\left(-n \min(\delta, \delta^2) \right). \end{aligned}$$

Combining these bounds setting $\delta = c_1\sqrt{\frac{d}{n}}$, we have

$$\begin{aligned} \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i^{(1)} \right\|_{\text{op}} + \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i^{(2)} \right\|_{\text{op}} \\ \leq C\varsigma^2 (\log(nk))^2 \left\{ \sqrt{\frac{d}{n}} + \frac{d}{n} \right\} \end{aligned}$$

with probability at least $1 - c\exp(-c'd)$.

The term $\kappa_i^{(3)}$, on the other hand, can be controlled directly via Hoeffding's inequality. Since $\max(\Theta^* z_i + b^*)$ is $C\varsigma\sqrt{\log(nk)}$ sub-Gaussian, we obtain

$$\mathbb{P}\left[\left| \frac{2}{n} \sum_{i=1}^{n/2} \kappa_i^{(3)} \right| \geq \varsigma \sqrt{\log(nk)} t \right] \leq 2 \exp\left\{ -\frac{nt^2}{32} \right\}.$$

Choosing $t = c\sqrt{\frac{d + \log n}{n}}$ and putting together all the pieces, we obtain

$$\begin{aligned} \left\| \frac{2}{n} \sum_{i=1}^{n} \kappa_i \right\|_{\text{op}} \leq C\varsigma^2 (\log(nk))^2 \left\{ \sqrt{\frac{d + \log n}{n}} + \frac{d + \log n}{n} \right\} \\ + c\varsigma \sqrt{\log(nk)} \sqrt{\frac{d}{n}} \end{aligned}$$

with probability at least $1 - cn^{-12}$.

It remains to handle the terms $\{\rho_i\}_{i=1}^{n/2}$, and to do so, we use a similar argument to before. We first bound the absolute value

of the (p,q)th entry of each matrix as

$$\begin{aligned}|\rho_i(p,q)| &= |\mathbb{E}[\max(\Theta^* z_i + b^*) z_i z_i^\top(p,q)] \\ &\quad - \mathbb{E}[\max(\Theta^* x_i + b^*) x_i x_i^\top(p,q)\mathbf{1}\{\mathcal{E}_i\}] \\ &\quad + \mathbb{E}[\max(\Theta^* x_i + b^*) x_i x_i^\top(p,q)\mathbf{1}\{\mathcal{E}_i^c\}]| \\ &= |\mathbb{E}[\max(\Theta^* x_i + b^*) x_i x_i^\top(p,q)\mathbf{1}\{\mathcal{E}_i^c\}]|\end{aligned}$$

Expanding this further, we have

$$\begin{aligned}|\rho_i(p,q)| &\le \mathbb{E}[\max_{\ell\in[k]}(|\langle\theta_\ell^*, x_i\rangle| + |b_\ell^*|)|x_{i,p}x_{i,q}|\mathbf{1}\{\mathcal{E}_i^c\}] \\ &\le \varsigma\mathbb{E}\left[|x_{i,p}x_{i,q}|\|x_i\|_\infty \mathbf{1}\{\mathcal{E}_i^c\}\right] \\ &\le \mathbb{E}\left[|x_{i,p}x_{i,q}|\sum_{\ell\in[k]}|x_{i,\ell}|\mathbf{1}\{\mathcal{E}_i^c\}\right].\end{aligned}$$

Also note that $\rho_{p,q} = 0$ unless $p \in [k], q \in [k]$. Hence we finally need to control the terms of the form $\mathbb{E}\left[|X|^3\mathbf{1}\{|X| \ge \lambda\}\right]$ for a standard Gaussian X. Substituting $\lambda = 5\sqrt{\log(nk)}$, a simple calculation of truncated third moment of standard Gaussian ([85]) yields

$$|\rho_i(p,q)| \le \varsigma \log^2(nk)(nk)^{-10},$$

and proceeding as before provides a strictly lower order bound on $\|\rho_i\|_{\text{op}}$ than the remaining terms.

The term τ_i^2 can be bounded more easily. Specifically, on the intersection of the events $\{\mathcal{F}_i\}_{i=1}^{n/2}$, applying [70, Lemma 6.15], we have

$$\begin{aligned}\|\frac{2}{n}\sum_{i=1}^{n/2}\tau_i^2\|_{\text{op}}^2 &\le C\sigma^2\log n\|\frac{2}{n}\sum_{i=1}^{n/2}x_i x_i^\top - I\|_{\text{op}}^2 \\ &\le C\sigma^2\log n\left\{\frac{d+\log n}{n} + \frac{(d+\log n)^2}{n^2}\right\}\end{aligned}$$

where the final bound holds with probability greater than $1-cn^{-12}$. Finally combining all the terms yield the desired result. □

APPENDIX H
BACKGROUND AND TECHNICAL LEMMAS USED IN THE PROOF OF THEOREM 3

In this section, we collect two technical lemmas that were used to prove Theorem 3.

A. Prediction and Estimation Error

Here, we connect the prediction error to the estimation error, which may be of independent interest. Recall our notation dist for the minimum distance between parameters obtainable after relabeling.

Lemma 16: There exists a tuple of universal constants (c_1, c_2) such that for each set of parameters $\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}$:

1) If $n \ge c_1 d$, then we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left(\max_{j\in[k]}\langle\xi_i, \beta_j\rangle - \max_{j\in[k]}\langle\xi_i, \beta_j^*\rangle\right)^2 \\ &\quad\le c_1\mathsf{dist}(\{\beta_j\}_{j=1}^k, \{\beta^*\}_{j=1}^k)\end{aligned}$$

with probability exceeding $1 - c_1\exp(-c_2 n)$.

2) If $n \ge c_1 d\frac{k^3}{\pi_{\min}^3}\log^2(k/\pi_{\min})$, then we have

$$\begin{aligned}&c_2\left(\frac{\pi_{\min}}{k}\right)^3\sum_{j\in[k]}\min_{j'\in[k]}\|\beta_j^* - \beta_{j'}\|^2 \\ &\quad\le \frac{1}{n}\sum_{i=1}^n\left(\max_{j\in[k]}\langle\xi_i, \beta_j\rangle - \max_{j\in[k]}\langle\xi_i, \beta_j^*\rangle\right)^2\end{aligned}$$

with probability exceeding $1 - c_1 k\exp\left(-c_2 n\frac{\pi_{\min}^4}{k^4\log^2(k/\pi_{\min})}\right)$.

Proof: To prove the part 1 of the lemma, we leverage the fact that the max function is 1-Lipschitz with respect to the ℓ_2-norm. Consequently, we obtain

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left(\max_{j\in[k]}\langle\xi_i, \beta_j\rangle - \max_{j\in[k]}\langle\xi_i, \beta_j^*\rangle\right)^2 \\ &\quad\le \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^k\left(\xi_i^\top(\beta_j - \beta_j^*)\right)^2,\end{aligned}$$

where we have ordered the parameters such that $\mathsf{dist}\left(\{\beta_j\}_{j=1}^k, \{\beta_j^*\}_{j=1}^k\right)$ is minimized. We now use the fact that the rows of Ξ are 1-sub-Gaussian (this is restatement of the conclusion of Lemma 13) to complete the proof.

We now proceed to a proof of part 2 of the lemma. Recall the setup of Appendix C along with notation $(\{x_i\}_{i=1}^n, \Theta^*, b^*, \beta^*,)$. Specifically, we have $\beta_j^* = \left(\theta_j^*, b_j^*\right)$ and $(\Theta^*)^\top = [\theta_1^*\ \theta_2^* \ldots \theta_k^*]$. Similarly let $\beta_j = (\theta_j, b_j) \in \mathbb{R}^{d+1}$ and $\Theta^\top = [\theta_1\ \theta_2 \ldots \theta_k]$. In the notation of Section B, we define for each pair (Θ, b), the sets

$$S_j(\Theta, b) = \left\{i \in [n] : \langle x_i, \theta_j\rangle + b_j = \max_{j'\in[k]}(\langle x_i, \theta_{j'}\rangle + b_{j'})\right.$$

for all $j \in [k]$. We use the shorthand $S_j^* = S_j(\Theta^*, b^*)$ and $\widehat{S}_j = S_j(\Theta, b)$ for the rest of the proof. By definition, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\left(\max(\Theta x_i + b) - \max(\Theta^* x_i + b^*)\right)^2 \\ &= \frac{1}{n}\sum_{\substack{\ell\in[k]\\ m\in[k]}}\sum_{i\in S_\ell^*\cap\widehat{S}_m}\left((\langle\theta_\ell^*, x_i\rangle + b_l^*) - (\langle\theta_m, x_i\rangle + b_m)\right)^2 \\ &= \frac{1}{n}\sum_{\substack{\ell\in[k]\\ m\in[k]}}\sum_{i\in S_\ell^*\cap\widehat{S}_m}\left(\langle\beta_\ell^*, \xi_i\rangle - \langle\beta_m, \xi_i\rangle\right)^2 \\ &= \frac{1}{n}\sum_{\substack{\ell\in[k]\\ m\in[k]}}\|\widetilde{\Xi}_{\ell,m}(\beta_\ell^* - \beta_m)\|^2,\end{aligned}$$

where we have let $\widetilde{\Xi}_{\ell,m}$ denote the sub-matrix of Ξ with rows indexed by the set $S_\ell^* \cap \widehat{S}_m$. It is also useful to define the convex sets

$$K_\ell^* := \left\{x \in \mathbb{R}^d : \langle x, \theta_\ell^*\rangle + b_\ell^* = \max_{j'\in[k]}(\langle x, \theta_{j'}^*\rangle + b_{j'}^*)\right\}, \text{ and}$$

$$K_m := \left\{x : \langle x, \theta_m\rangle + b_m = \max_{j'\in[k]}(\langle x, \theta_{j'}\rangle + b_{j'})\right\}$$

for each pair $(\ell, m) \in [k] \times [k]$. By definition, for each $\ell \in [k]$, there exists a corresponding index m_ℓ such that $\mathfrak{vol}(K_\ell^* \cap K_{m_\ell}) \geq \frac{\pi_{\min}}{k}$. Proceeding from above, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\max(\Theta x_i + b) - \max(\Theta^* x_i + b^*)\right)^2 \\
&\geq \frac{1}{n}\sum_{\ell\in[k]}\left\|\widetilde{\Xi}_{\ell,m_\ell}(\beta_\ell^* - \beta_{m_\ell})\right\|^2 \\
&\geq \frac{1}{n}\sum_{\ell\in[k]}\lambda_{\min}\left(\widetilde{\Xi}_{\ell,m_\ell}^\top \widetilde{\Xi}_{\ell,m_\ell}\right)\left\|\beta_\ell^* - \beta_{m_\ell}\right\|^2.
\end{aligned}$$

Finally, applying Lemma 5 in conjunction with the bound $\mathfrak{vol}(K_\ell^* \cap K_{m_\ell}) \geq \frac{\pi_{\min}}{k}$, we obtain that provided $n \geq c_1 d \cdot \frac{k^3}{\pi_{\min}^3}\log^2(k/\pi_{\min})$, we have

$$\lambda_{\min}\left(\widetilde{\Xi}_{\ell,m_\ell}^\top \widetilde{\Xi}_{\ell,m_\ell}\right) \geq \left(\frac{\pi_{\min}}{k}\right)^3 \cdot n$$

with probability exceeding $1 - c_1\exp\left(-c_2 n \frac{\pi_{\min}^4}{k^4\log^2(k/\pi_{\min})}\right)$ for each index $\ell \in [k]$. Taking a union bound over the k indices and combining the pieces completes the proof. □

B. Projection Onto a Finite Collection of Rays

Consider a vector $\theta^* \in \mathbb{R}^n$ observed via the observation model

$$y = \theta^* + \epsilon,$$

where ϵ has independent, zero-mean, σ-sub-Gaussian entries. For a fixed set of M vectors $\{\theta_1, \ldots, \theta_M\}$, denote by $\mathbb{C} := \{c\theta_\ell : c \geq 0, \ell \in [M]\}$ the set of all one-sided rays obtainable with these vectors.

Now consider the projection estimate

$$P_{\mathbb{C}}(y) = \operatorname*{argmin}_{\theta\in\mathbb{C}} \|y - \theta\|^2,$$

which exists since the projection onto each ray exists. The following lemma proves an oracle inequality on the error of such an estimate.

Lemma 17: There are universal constants c, C, c_1 and c_2 such that

$$\begin{aligned}
\Pr\Big\{\|P_{\mathbb{C}}(y) - \theta^*\|^2 \geq c(\min_{\theta\in\mathbb{C}}\|\theta - \theta^*\|^2 + \sigma^2 t(\log M + c_1))\Big\} \\
\leq c_2 e^{-nt(\sqrt{\log M} + c_1)},
\end{aligned}$$

for all $t \geq C\sigma(\sqrt{\log M} + c_1)$.

Proof: We follow the standard technique for bounding the error for non-parametric least squares estimators. From the definition, we have

$$P_{\mathbb{C}}(y) = \operatorname*{argmin}_{\theta\in\mathbb{C}} \|y - \theta\|^2.$$

We substitute the expression for y and obtain

$$P_{\mathbb{C}}(y) = \operatorname*{argmax}_{\theta\in\mathbb{C}}\left[2\langle\epsilon, \theta - \theta^*\rangle - \|\theta - \theta^*\|^2\right].$$

To obtain an upper bound on $\|P_{\mathbb{C}}(y) - \theta^*\|^2$, it is sufficient to control the following quantity (e.g. see [86, Chapter 3], [70, Chapter 13]):

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{C}:\|\theta-\theta^*\|\leq\delta}\langle\epsilon, \theta - \theta^*\rangle\right]$$

for some $\delta > 0$ to be chosen later. Since ϵ is σ-sub-Gaussian, we use Dudley's entropy integral to control the term above. We obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\theta\in\mathbb{C}:\|\theta-\theta^*\|\leq\delta}\langle\epsilon, \theta - \theta^*\rangle\right] \\
&\leq C\sigma\int_0^\delta \sqrt{\log N\left(\varepsilon, \{\theta\in\mathbb{C}, \|\theta - \theta^*\| \leq \delta\}, \ell_2\right)}d\varepsilon,
\end{aligned}$$

where $N(\epsilon, S, \ell_2)$ is the ϵ-covering number of a compact set S in ℓ_2 norm. Note that $\mathbb{C}$ contains scaled versions of M fixed vectors $\{\theta_1, \ldots, \theta_M\}$. For a fixed θ_i, with $i \in [M]$, the covering number $N\left(\varepsilon, \{c\theta_i : c \in \mathbb{R}, \|\theta_i - \theta^*\| \leq \delta\}, \ell_2\right)$ is equivalent to the covering number of a bounded interval (in 1 dimension). Using [87], this is $(1 + \frac{2\delta}{\varepsilon})$. Since there are M such fixed vectors, we obtain

$$N\left(\varepsilon, \{\theta\in\mathbb{C}, \|\theta - \theta^*\| \leq \delta\}, \ell_2\right) \leq C_1 M(1 + \frac{\delta}{\varepsilon}).$$

Substituting, we obtain

$$\mathbb{E}\left[\sup_{\theta\in\mathbb{C}:\|\theta-\theta^*\|\leq\delta}\langle\epsilon, \theta - \theta^*\rangle\right] \leq C\sigma\left(\delta\sqrt{\log M} + C_1\delta\right).$$

Now, the critical inequality ([70, Chapter 13]) takes the form

$$\delta\sigma(\sqrt{\log M} + C_1) \lesssim \delta^2.$$

Hence we can choose $\delta_0 = C_2\sigma(\sqrt{\log M} + C_1)$. Now, for any $t \geq \delta_0$, invoking [70, Theorem 13.13] yields the oracle inequality

$$\begin{aligned}
\|P_{\mathbb{C}}(y) - \theta^*\|^2 &\leq c\left(\|\theta^* - P_{\mathbb{C}}(\theta^*)\|^2 + \sigma^2 t(\log M + c_1)\right) \\
&= c\left(\min_{\theta\in\mathbb{C}}\|\theta - \theta^*\|^2 + \sigma^2 t(\log M + c_1)\right),
\end{aligned}$$

with probability exceeding $1 - c_2 e^{-nt(\sqrt{\log M} + c_1)}$, which proves the lemma. □

APPENDIX I
NP-HARDNESS OF REAL PHASE RETRIEVAL

Our discussion borrows from a similar proof established in [46, Proposition 1] for mixtures of linear regressions. Recall that with n i.i.d observations $\{(x_i, y_i)\}_{i=1}^n$, the max-affine model takes the form

$$y_i = \max_{1\leq j\leq k}(\langle x_i, \theta_j^*\rangle + b_j^*) + \epsilon_i,$$

where $\{\epsilon_i\}_{i=1}^n$ is a sequence of i.i.d zero mean sub-Gaussian noise.

We now consider a special case, where $k = 2$, $b_1^* = b_2^* = 0$, and $\theta_1^* = -\theta_2^*$, corresponding to the real phase retrieval problem. Furthermore, we consider the noiseless case $\epsilon = 0$. Our covariate matrix is given by $X \in \mathbb{R}^{n\times d}$ and the response vector by $y \in \mathbb{R}^n$. We now show that even in

this special case, there is family of instances (X, y) such that solving the least squares problem (5) is NP-hard. In particular, we say that a "solution" to the noiseless phase retrieval problem exists on an instance (X, y) if the least squares objective in equation (5) has minimum value zero.

Proposition 3: Deciding whether a problem instance (X, y) has a solution to the noiseless phase retrieval problem is NP-hard.

Proof: The proof follows from a reduction to the subset-sum problem, the decision version of which is stated as follows: given p numbers $a_1, \ldots, a_p \in \mathbb{R}$, we must decide if there exists a partition $S \subseteq [p]$ such that

$$\sum_{i \in S} a_i = \sum_{j \in S^c} a_j.$$

For each p-dimensional vector a, we design a problem instance (X, y) such that solving the noiseless (real) phase retrieval problem on (X, y) implies deciding on the subset sum problem specified by a.

To accomplish this, take $n = 2p + 1$ and $d = p$, and define the instance

$$X = \begin{bmatrix} I_p \\ I_p \\ 1 \ldots 1 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} a \\ -a \\ 0 \end{bmatrix},$$

where I_p denotes the $p \times p$ identity matrix. By construction, finding a solution to the noiseless (real) phase retrieval problem on this instance corresponds to finding a subset $S \subseteq [2p + 1]$ and a pair of vectors (θ_1^*, θ_2^*) with $\theta_1^* = -\theta_2^*$, such that $X_S \theta_1^* = y_S$, and $X_{S^c} \theta_2^* = y_{S^c}$. Here X_S and y_S are the sub-matrix and sub-vector of X and y respectively restricted to the set S. Note that in general, the set S cannot contain the index i and $p + i$, since they correspond to two mutually exclusive equations. From this observation, we have $\theta_1^*(i) = \{a_i, -a_i\}$, and $\theta_1^*(i) = -\theta_2^*(i)$, where $\theta_1^*(i)$ and $\theta_2^*(i)$ denote the i-th coordinate of θ_1^* and θ_2^*, respectively.

As a consequence, if θ_1^* (and $\theta_2^* = -\theta_1^*$) satisfies the first $2p$ equations in this system, then the final equation demands that

$$\sum_{i \in S} \theta_1^*(i) = 0 = \sum_{j \in S^c} \theta_2^*(j).$$

By construction, note that this is accomplished if and only if

$$\sum_{i \in S} a_i = \sum_{j \in S^c} a_j,$$

and so a solution to the noiseless (real) phase retrieval problem on (X, y) yields a solution to the subset-sum problem, as desired. □

ACKNOWLEDGMENT

The authors thank Bodhi Sen for helpful discussions, and the anonymous reviewers for valuable comments that improved the scope and presentation of the article.

REFERENCES

[1] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, "Max-affine regression with universal parameter estimation for small-ball designs," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2706–2710.

[2] R. W. Harrison, "Phase problem in crystallography," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 10, no. 5, pp. 1046–1055, May 1993.

[3] C. Fienup and J. Dainty, "Phase retrieval and image reconstruction for astronomy," *Image Recovery, Theory Appl.*, vol. 231, p. 275, Jan. 1987.

[4] A. Chai, M. Moscoso, and G. Papanicolaou, "Array imaging using intensity-only measurements," *Inverse Problems*, vol. 27, no. 1, 2010, Art. no. 015005.

[5] F. Fogel, I. Waldspurger, and A. d'Aspremont, "Phase retrieval for imaging problems," *Math. Program. Comput.*, vol. 8, no. 3, pp. 311–335, Sep. 2016.

[6] E. Bronshtein, "ε-entropy of convex sets and functions," *Siberian Math. J.*, vol. 17, no. 3, pp. 393–398, 1976.

[7] A. Guntuboyina and B. Sen, "Covering numbers for convex functions," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 1957–1965, Apr. 2013.

[8] F. Gao and J. A. Wellner, "Entropy of convex functions on $\mathbb{R}^d$," *Constructive Approximation*, vol. 46, no. 3, pp. 565–592, 2017.

[9] Q. Han and J. A. Wellner, "Multivariate convex regression: Global risk bounds and adaptation," 2016, *arXiv:1601.06844*.

[10] A. Magnani and S. P. Boyd, "Convex piecewise-linear fitting," *Optim. Eng.*, vol. 10, no. 1, pp. 1–17, 2009.

[11] L. A. Hannah and D. B. Dunson, "Multivariate convex regression with adaptive partitioning," *J. Mach. Learn. Res.*, vol. 14, no. 3, pp. 3261–3294, 2013.

[12] G. Balázs, "Convex regression: Theory, practice, and applications," Ph.D. dissertation, Dept. Comput. Sci., Univ. Alberta, Edmonton, AB, Canada, 2016.

[13] J. L. Prince and A. S. Willsky, "Reconstructing convex sets from support line measurements," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 4, pp. 377–389, Apr. 1990.

[14] J. Gregor and F. R. Rannou, "Three-dimensional support function estimation and application for projection magnetic resonance imaging," *Int. J. Imag. Syst. Technol.*, vol. 12, no. 1, pp. 43–50, 2002.

[15] R. J. Gardner, *Geometric Tomography* (Encyclopedia of Mathematics and its Applications), 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2006.

[16] A. Guntuboyina, "Optimal rates of convergence for convex set estimation from support functions," *Ann. Statist.*, vol. 40, no. 1, pp. 385–411, Feb. 2012.

[17] Y. S. Soh and V. Chandrasekaran, "Fitting tractable convex sets to support function evaluations," 2019, *arXiv:1903.04194*.

[18] S. A. van de Geer, *Regression Analysis and Empirical Processes CWI Tract. Stichting Mathematisch Centrum, Centrum Voor Wiskunde en Informatica*, vol. 45. Amsterdam, The Netherlands: Rijksuniv, 1988. [Online]. Available: https://books.google.com/books/about/Regression_Analysis_and_Empirical_Proces.html?id=8vkkygEACAAJ

[19] E. Seijo and B. Sen, "Nonparametric least squares estimation of a multivariate convex regression function," *Ann. Statist.*, vol. 39, no. 3, pp. 1633–1657, Jun. 2011.

[20] E. Lim and P. W. Glynn, "Consistency of multidimensional convex regression," *Oper. Res.*, vol. 60, no. 1, pp. 196–208, Feb. 2012.

[21] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen, "A computational framework for multivariate convex regression and its variants," *J. Amer. Stat. Assoc.*, vol. 114, no. 525, pp. 318–331, Jan. 2019.

[22] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2796–2804.

[23] I. Waldspurger, "Phase retrieval with random Gaussian sensing vectors by alternating projections," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3301–3312, May 2018.

[24] T. Zhang, "Phase retrieval using alternating minimization in a batch setting," *Appl. Comput. Harmon. Anal.*, vol. 49, no. 1, pp. 279–295, Jul. 2020.

[25] S. Balakrishnan, M. J. Wainwright, and B. Yu, "Statistical guarantees for the EM algorithm: From population to sample-based analysis," *Ann. Statist.*, vol. 45, no. 1, pp. 77–120, 2017.

[26] R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, no. 2, pp. 237–246, 1972.

[27] J. R. Fienup, "Phase retrieval algorithms: A comparison," *Appl. Opt.*, vol. 21, no. 15, pp. 2758–2769, Aug. 1982.

[28] E. M. L. Beale and R. J. A. Little, "Missing values in multivariate analysis," *J. Roy. Stat. Soc. B, Methodol.*, vol. 37, no. 1, pp. 129–145, 1975.

[29] H. O. Hartley, "Maximum likelihood estimation from incomplete data," *Biometrics*, vol. 14, no. 2, pp. 174–194, 1958.

[30] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 8, no. 1, pp. 129–151, 1996.

[31] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, no. 2, pp. 181–214, 1994.

[32] A. T. Chaganty and P. Liang, "Spectral experts for estimating mixtures of linear regressions," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1040–1048.

[33] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.

[34] P. Tseng, "An analysis of the EM algorithm and entropy-like proximal point methods," *Math. Oper. Res.*, vol. 29, no. 1, pp. 27–44, Feb. 2004.

[35] S. Chrétien and A. O. Hero, "On EM algorithms and their proximal generalizations," *ESAIM, Probab. Statist.*, vol. 12, pp. 308–326, Apr. 2008.

[36] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1260–1268.

[37] K. Zhong, P. Jain, and I. S. Dhillon, "Mixed linear regression with multiple components," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2190–2198.

[38] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan, "Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4116–4124.

[39] C. Daskalakis, C. Tzamos, and M. Zampetakis, "Ten steps of EM suffice for mixtures of two Gaussians," in *Proc. 30th Annu. Conf. Learn. Theory*, 2017, pp. 704–710.

[40] J. Xu, D. J. Hsu, and A. Maleki, "Global analysis of expectation maximization for mixtures of two Gaussians," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2676–2684.

[41] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis, "Global convergence of EM algorithm for mixtures of two component linear regression," 2018, *arXiv:1810.05752*.

[42] D. Hsu and S. M. Kakade, "Learning mixtures of spherical gaussians: Moment methods and spectral decompositions," in *Proc. 4th Conf. Innov. Theor. Comput. Sci. (ITCS)*, 2013, pp. 11–20.

[43] X. Yi, C. Caramanis, and S. Sanghavi, "Solving a mixture of many random linear equations by tensor decomposition and alternating minimization," 2016, *arXiv:1608.05749*.

[44] Y. Shuo Tan and R. Vershynin, "Phase retrieval via randomized Kaczmarz: Theoretical guarantees," 2017, *arXiv:1706.09993*.

[45] G. M. Tallis, "The moment generating function of the truncated multinormal distribution," *J. Roy. Statist. Soc. B, Methodol.*, vol. 23, no. 1, pp. 223–229, 1961.

[46] X. Yi, C. Caramanis, and S. Sanghavi, "Alternating minimization for mixed linear regression," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 613–621.

[47] Y. Shen and S. Sanghavi, "Iterative least trimmed squares for mixed linear regression," 2019, *arXiv:1902.03653*.

[48] G. Lecué and S. Mendelson, "Minimax rate of convergence and the performance of ERM in phase recovery," 2013, *arXiv:1311.5024*.

[49] Y. S. Soh, "Fitting convex sets to data: Algorithms and applications," Ph.D. dissertation, California Inst. Technol., Pasadena, CA, USA, 2019.

[50] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, Jun. 2010.

[51] H. Sedghi, M. Janzamin, and A. Anandkumar, "Provable tensor methods for learning mixtures of generalized linear models," *Proc. Mach. Learn. Res.*, vol. 51, pp. 1223–1231, May 2014.

[52] K. Zhong, Z. Song, P. Jain, P. L. Bartlett, and I. S. Dhillon, "Recovery guarantees for one-hidden-layer neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 4140–4149.

[53] Y. Chen and E. Candes, "Solving random quadratic systems of equations is nearly as easy as solving linear systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 739–747.

[54] A. Daniely, S. Sabato, and S. S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 485–493.

[55] D. Babichev and F. Bach, "Slice inverse regression with score functions," *Electron. J. Statist.*, vol. 12, no. 1, pp. 1507–1543, Jan. 2018.

[56] S. Goel, V. Kanade, A. Klivans, and J. Thaler, "Reliably learning the ReLU in polynomial time," in *Proc. Conf. Learn. Theory*, 2017, pp. 1004–1042.

[57] A. Guntuboyina and B. Sen, "Nonparametric shape-restricted regression," *Stat. Sci.*, vol. 33, no. 4, pp. 568–594, Nov. 2018.

[58] A. Guntuboyina and B. Sen, "Global risk bounds and adaptation in univariate convex regression," *Probab. Theory Rel. Fields*, vol. 163, nos. 1–2, pp. 379–411, Oct. 2015.

[59] P. C. Bellec, "Sharp Oracle inequalities for least squares estimators in shape restricted regression," *Ann. Statist.*, vol. 46, no. 2, pp. 745–780, Apr. 2018.

[60] A. M. Medina and M. Mohri, "Learning theory and algorithms for revenue optimization in second price auctions with reserve," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 262–270.

[61] J. Morgenstern and T. Roughgarden, "Learning simple auctions," in *Proc. Conf. Learn. Theory*, 2016, pp. 1298–1318.

[62] K.-C. Li, "Sliced inverse regression for dimension reduction," *J. Amer. Stat. Assoc.*, vol. 86, no. 414, pp. 316–327, 1991.

[63] J. L. Horowitz, *Semiparametric Nonparametric Methods Econometrics*, vol. 12. New York, NY, USA: Springer, 2009.

[64] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.

[65] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.

[66] T. Zhang, "Phase retrieval by alternating minimization with random initialization," 2018, *arXiv:1812.01255*.

[67] T. Cai, X. Li, and Z. Ma, "Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow," *Ann. Statist.*, vol. 44, no. 5, pp. 2221–2251, 2016.

[68] M. Rudelson and R. Vershynin, "Hanson-wright inequality and sub-Gaussian concentration," *Electron. Commun. Probab.*, vol. 18, pp. 1–9, Jan. 2013.

[69] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," 2010, *arXiv:1011.3027*.

[70] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[71] M. Kanter and H. Proppe, "Reduction of variance for Gaussian densities via restriction to convex sets," *J. Multivariate Anal.*, vol. 7, no. 1, pp. 74–81, Mar. 1977.

[72] S. S. Vempala, "Learning convex concepts from Gaussian distributions with PCA," in *Proc. IEEE 51st Annu. Symp. Found. Comput. Sci.*, Oct. 2010, pp. 124–130.

[73] M. Ledoux, *The Concentration Measure Phenomenon*, no. 89. Providence, RI, USA: AMS, 2001.

[74] B. James, K. James, and Y. Qi, "Limit distribution of the sum and maximum from multivariate Gaussian sequences," *J. Multivariate Anal.*, vol. 98, no. 3, pp. 517–532, Mar. 2007.

[75] Y. Yu, T. Wang, and R. J. Samworth, "A useful variant of the Davis–Kahan theorem for statisticians," *Biometrika*, vol. 102, no. 2, pp. 315–323, 2015.

[76] A. F. Siegel, "A surprising covariance involving the minimum of multivariate normal variables," *J. Amer. Stat. Assoc.*, vol. 88, no. 421, pp. 77–80, Mar. 1993.

[77] J. S. Liu, "Siegel's formula via Stein's identities," *Statist. Probab. Lett.*, vol. 21, no. 3, pp. 247–251, Oct. 1994.

[78] B. Manjunath and S. Wilhelm, "Moments calculation for the double truncated multivariate normal density," 2012, *arXiv.1206.5387*.

[79] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Series in Statistics). New York, NY, USA: Springer, 2009. [Online]. Available: https://link.springer.com/book/10.1007/b13794#about

[80] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.

[81] V. N. Vapnik and A. Y. Chervonenkis, "The uniform convergence of frequencies of the appearance of events to their probabilities," in *Doklady Akademii Nauk*, vol. 181, no. 4. Moscow, Russia: Russian Academy of Sciences, 1968, pp. 781–783.

[82] M. Csikos, A. Kupavskii, and N. H. Mustafa, "Optimal bounds on the VC-dimension," 2018, *arXiv:1807.07924*.

[83] G. Hargé, "A convex/log-concave correlation inequality for Gaussian measure and an application to abstract Wiener spaces," *Probab. Theory Rel. Fields*, vol. 130, no. 3, pp. 415–440, Nov. 2004.

[84] Y. Hu, "Itô-Wiener chaos expansion with exact residual and correlation, variance inequalities," *J. Theor. Probab.*, vol. 10, no. 4, pp. 835–848, 1997.

[85] J. Burkardt. (2014). *The Truncated Normal Distribution*. [Online]. Available: https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf
[86] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes* (Springer Series in Statistics). New York, NY, USA: Springer-Verlag, 1996, doi: 10.1007/978-1-4757-2545-2.
[87] R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47. Cambridge, U.K.: Cambridge Univ. Press, 2018.

Avishek Ghosh received the bachelor's degree from the Department of Electronics and Telecommunication Engineering, Jadavpur University, the master's degree from the Department of Electrical Communication Engineering, Indian Institute of Science (IISc), Bengaluru, and the Ph.D. degree from the Department of Electrical Engineering and Computer Sciences (EECS), UC Berkeley, advised by Prof. Kannan Ramchandran and Prof. Aditya Guntuboyina. He is currently an HDSI (Data Science) Post-Doctoral Fellow with the University of California, San Diego. His research interests are broadly in theoretical machine learning, including federated learning and multi-agent reinforcement/bandit learning. In particular, he is interested in theoretically understanding challenges in multi-agent systems and competition/collaboration across agents.

Ashwin Pananjady received the B.Tech. degree in electrical engineering from IIT Madras in 2014 and the Ph.D. degree in electrical engineering and computer sciences (EECS) from the University of California at Berkeley in 2020. He is currently an Assistant Professor with the Georgia Institute of Technology with a joint appointment between the School of Industrial and Systems Engineering, and the School of Electrical and Computer Engineering. His research in high-dimensional statistics, information theory, and optimization has won the Inaugural Lawrence D. Brown Ph.D. Student Award from the Institute of Mathematical Statistics, the David J. Sakrison Memorial Prize (EECS Dissertation Award, UC Berkeley), and the Simons-Berkeley Research Fellowship in Probability, Geometry and Computation in High Dimensions.

Adityanand Guntuboyina received the B.Stat. and M.Stat. degrees from the Indian Statistical Institute in 2004 and 2006, respectively, and the Ph.D. degree in statistics from Yale University in 2011. He was a Post-Doctoral Research Associate with the Wharton Statistics Department, University of Pennsylvania, from July 2012 to December 2012. He is currently an Associate Professor in statistics with the University of California at Berkeley. His research interests include nonparametric function estimation, Bayesian and empirical Bayesian methodology, empirical processes, statistical learning theory, and information theory. He received the NSF CAREER Award in 2017.

Kannan Ramchandran (Fellow, IEEE) is currently a Professor of electrical engineering and computer science with UC Berkeley, where he has been since 1999. Prior to that, he was on the faculty of the University of Illinois at Urbana-Champaign from 1993 to 1999. He has published extensively and holds more than 12 patents. His current research interests are in machine learning, statistical signal processing and coding, information theory, and distributed cloud computation. He was a recipient of the 2017 IEEE Kobayashi Computers and Communications Award for his pioneering contributions to the theory and practice of distributed storage codes and distributed compression. He has received several awards for his research and teaching, including two IEEE Information Theory Society and Communication Society joint best paper awards in 2012 and 2020, the IEEE Communication Society Data Storage Best Paper Award in 2010, two best paper awards from the IEEE Signal Processing Society in 1993 and 1999, the Hank Magnusky Scholar Award at Illinois in 1998, the Okawa Foundation Award for Outstanding Research at Berkeley in 2001, and the EECS Departmental Outstanding Teaching Award at Berkeley in 2009.