Spectrahedral Regression

Eliza O'Reilly and Venkat Chandrasekaran

October 27, 2021

Abstract

Convex regression is the problem of fitting a convex function to a data set consisting of input-output pairs. We present a new approach to this problem called spectrahedral regression, in which we fit a spectrahedral function to the data, i.e. a function that is the maximum eigenvalue of an affine matrix expression of the input. This method represents a significant generalization of polyhedral (also called max-affine) regression, in which a polyhedral function (a maximum of a fixed number of affine functions) is fit to the data. We prove bounds on how well spectrahedral functions can approximate arbitrary convex functions via statistical risk analysis. We also analyze an alternating minimization algorithm for the non-convex optimization problem of fitting the best spectrahedral function to a given data set. We show that this algorithm converges geometrically with high probability to a small ball around the optimal parameter given a good initialization. Finally, we demonstrate the utility of our approach with experiments on synthetic data sets as well as real data arising in applications such as economics and engineering design.

Keywords: convex regression, support function estimation, semidefinite programming, approximation of convex bodies.

1 Introduction

The problem of identifying a function that approximates a given dataset of input-output pairs is a central one in data science. In this paper we consider the problem of fitting a convex function to such input-output pairs, a task known as *convex regression.* Concretely, given data $\{x^{(i)}, y^{(i)}\}_{i=1}^{n} \subset \mathbb{R}^d \times \mathbb{R}$, our objective is to identify a convex function $\hat{f}$ such that $\hat{f}(x^{(i)}) \approx y^{(i)}$ for each $i = 1, \ldots, n$. In some applications, one seeks an estimate $\hat{f}$ that is convex and positively homogenous; in such cases, the problem may equivalently be viewed as one of identifying a convex set given (possibly noisy) support function evaluations. Convex reconstructions in such problems are of interest for several reasons. First, prior domain information in the context of a particular application might naturally lead a practitioner to seek convex approximations. One prominent example arises in economics in which the theory of marginal utility implies an underlying convexity relationship. Another important example arises in computed tomography applications in which one has access to support function evaluations of some underlying set, and the goal is to reconstruct the set; here, due to the nature of the data acquisition mechanism, the set may be assumed to be convex without loss of generality. A second reason for preferring a convex reconstruction $\hat{f}$ is computational – in some applications the goal is to subsequently use $\hat{f}$ as an objective or constraint within an optimization formulation. For example, in aircraft design problems, the precise relationship between various attributes of an aircraft is often not known in closed-form, but input-output data are available from simulations; in such cases, identifying a good convex approximation for the input-output relationship is useful for subsequent aircraft design using convex optimization.

A natural first estimator one might write down is:

$$\hat{f}_{\mathrm{LSE}}^{(n)} \in \arg\min_{f:\mathbb{R}^d \to \mathbb{R} \text{ is a convex function}} \quad \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - f(x^{(i)}))^2. \tag{1}$$

There always exists a polyhedral function that attains the minimum in (1), and this function may be computed efficiently via convex quadratic programming [19, 20, 22]. However, this choice suffers from a number of drawbacks. For a large sample size, the quality of the resulting estimate suffers from over-fitting as the complexity of the reconstruction grows with the number of data points. For small sample sizes, the quality of the resulting estimate is often poor due to noise. From a statistical perspective, the estimator may also be suboptimal [15, 16]. For these reasons, it is of interest to regularize the estimator by considering a suitably constrained class of convex functions.

The most popular approach in the literature to penalize the complexity of the reconstruction in (1) is to fit a polyhedral function that is representable as the maximum of at most m affine functions (for a user-specified choice of m) to the given data [9, 10, 12, 18, 25], which is based on the observation that convex functions are suprema of affine functions. However, this approach is inherently restrictive in situations in which the underlying phenomenon is better modeled by a non-polyhedral convex function, which may not be well-approximated by m-polyhedral functions. Further, in settings in which the estimated function is subsequently used within an optimization formulation, the above approach constrains one to using linear-programming (LP) representable functions. See Figure 1 for a demonstration with economic data.

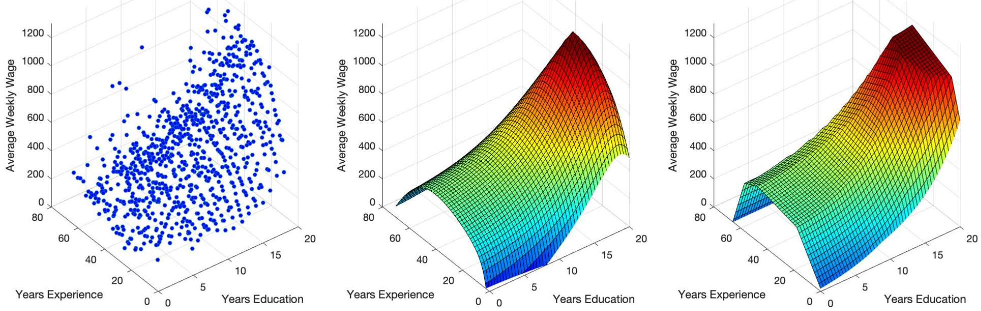

Figure 1: Models for average weekly wage based on years of experience and education using spectrahdedral and polyhedral regression. From left to right: the underlying data set, the spectrahedral ($m = 3$) estimator, and the polyhedral ($m = 6$) estimator. A transformation in the years of education covariate gives a data set that is approximately convex.

To overcome these limitations, we consider fitting spectrahedral functions to data. To define this model class, let $\mathbb{S}_k^m$ denote the set of $m \times m$ real symmetric matrices that are block diagonal with blocks of size at most $k \times k$, with k dividing m.

Definition 1. *Fix positive integers m, k such that k divides m. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called (m, k)-spectrahedral if it can be expressed as follows:*

$$f(x) = \lambda_{\max}\left(\sum_{i=1}^{d} A_i x_i + B\right),$$

where $A_1, \ldots, A_d, B \in \mathbb{S}_k^m$. Here $\lambda_{\max}(\cdot)$ is the largest eigenvalue of a matrix.

An (m, k)-spectrahedral function is convex as it is a composition of a convex function with an affine map. For the case $k = 1$, the matrices $A_1, \ldots, A_d, B$ are all diagonal and we recover the case of m-polyhedral functions. The case $k = 2$ corresponds to second-order-cone-programming (SOCP) representable functions, and the case $k = m$ utilizes the expressive power of semidefinite programming (SDP). In analogy to the enhanced modeling power of SOCP and SDP in comparison to LP, the class of (m, k)-spectrahedral functions is much richer than the set of m-polyhedral functions for general $k > 1$. For instance, when $k = 2$ this class contains the function $f(x) = \|x\|_2$. For estimates that are (m, k)-spectrahedral, subsequently emplying them within optimization formulations yields optimization problems that can be solved via SOCP and SDP.

An (m, k)-spectrahedral function that is positively homogenous (i.e., $B = 0$ in the definition above) is the support function of a convex set that is expressible as the linear image of an (m, k)-spectraplex defined, for positive integers k and m such that k divides m, by

$$\mathcal{S}_{m,k} = \{M \in \mathbb{S}_k^m \mid \operatorname{tr}(M) = 1,\ M \succeq 0\}. \tag{2}$$

We refer to the collection of linear images of $\mathcal{S}_{m,k}$ as (m, k)-*spectratopes*. Again, the case $k = 1$ corresponds to the m-simplex, and the corresponding linear images are m-polytopes. Thus, in the positively homogenous case, our proposal is to identify a linear image of an (m, k)-spectraplex to fit a given set of support function evaluations. We note that the case $k = m$ was recently considered in [24], and we comment in more detail on the comparison between the present paper and [24] in Section 1.2.

1.1 Our Contributions

We consider the following constrained analog of (1):

$$\hat{f}_{m,k}^{(n)} \in \arg\min_{f:\mathbb{R}^d\to\mathbb{R} \text{ is an } (m,k)\text{-spectrahedral function}} \quad \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - f(x^{(i)}))^2. \tag{3}$$

Here the parameters m, k are specified by the user.

First, we investigate in Section 2 the expressive power of (m, k)-spectrahedral functions. Our approach to addressing this question is statistical in nature and it proceeds in two steps. We begin by deriving upper bounds on the error of the constrained estimator (3) (under suitable assumptions on the data $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ supplied to the estimator (3)), which entails computing pseudo-dimension of a set that captures the complexity of the class of spectrahedral functions. As is standard in statistical learning theory, this error decomposes into an estimation error (due to finite sample size) and an approximation error (due to constraining the estimator (3) to a proper subclass of convex functions). We then compare these to known minimax lower bounds on the error of any procedure for identifying a convex function [9, 25]. Combined together, for the case of fixed k (as a function of m) we obtain tight lower bounds on how well an (m, k)-spectrahedral function can approximate a Lipschitz convex function over a compact convex domain, and on how well a linear image of an (m, k)-spectraplex can approximate an arbitrary convex body (see Theorem 8). To the best of our knowledge, such bounds have only been obtained previously in the literature for the case $k = 1$, e.g., how well m-polytopes can approximate arbitrary convex bodies [3, 5].

Second, we investigate in Section 3 the performance of an alternating minimization procedure to solve (3) for a user-specified m, k. This method is a natural generalization of a widely-used approach for fitting m-polyhedral functions, and it was first described in [24] for the case of positively homogenous convex regression with $k = m$. We investigate the convergence properties of this algorithm under the following problem setup. Consider an (m, k)-spectrahedral function $f_* : \mathbb{R}^d \to \mathbb{R}$. Assuming that the covariates $x^{(i)},\ i = 1, \ldots, n$ are i.i.d. Gaussian and each $y^{(i)} = f_*(x^{(i)}) + \varepsilon_i,\ i = 1, \ldots, n$ for i.i.d. Gaussian noise ε_i, we show in Theorem 10 that the alternating minimization algorithm is locally linearly convergent with high probability given sufficiently large n. A key feature of this analysis is that the requirements on the sample size n and the assumptions on the quality of the initial guess are functions of a 'condition number' type quantity associated to f_*, which (roughly speaking) measures how f_* changes if the parameters that describe it are perturbed.

Finally, in Section 4 we give empirical evidence of the utility of our estimator (3) on both synthetic datasets as well as data arising from real-world applications.

1.2 Related Work

There are three broad topics with which our work has a number of connections, and we describe these in detail next.

First, we consider our results in the context of the recent literature in optimization on lift-and-project methods (see the recent survey [6] and the references therein). This body of work has studied the question of the most compact description of a convex body as a linear image of an affine section of a cone, and has provided lower bounds on the sizes of such descriptions for prominent families of cone programs such as LP, SOCP, and SDP. This literature has primarily considered exact descriptions, and there is relatively little work on lower bounds for approximate descriptions (with the exception of the case of polyhedral descriptions). The present paper may be viewed as an approximation-theoretic complement to this body of work, and we obtain tight lower bounds on the expressive power of (m, k)-spectrahedral functions (and on linear images of the (m, k)-spectraplex) for bounded $k > 1$.

Second, recent results provide algorithmic guarantees for the widely used alternating minimization procedure for fitting m-polyhedral functions [7, 8]; this work gives both a local convergence analysis as well as a dimension reduction strategy to restrict the space over which one needs to consider random initializations. In comparison, our results provide only a local convergence analysis, although we do so for a more general alternating minimization procedure that is suitable for fitting general (m, k)-spectrahedral functions. We defer the study of a suitable initialization strategy to future work (see Section 5).

Finally, we note that there is prior work on fitting non-polyhedral functions in the convex regression problem. Specifically, [13] suggests various heuristics to fit a log-sum-exp type function, which may be viewed as a 'soft-max' function. However, these methods do not come with any approximation-theoretic or algorithmic guarantees. In recent work, [24] considered the problem of fitting a convex body given support function evaluations, i.e., the case of positively homogenous convex regression, and proposed reconstructions that are linear images of an (m, m)-spectraplex; in this context, [24] provided an asymptotic statistical analysis of the associated estimator and first described an alternating minimization procedure that generalized the m-polyhedral case, but with no algorithmic guarantees. In comparison to [24], the present paper considers the more general setting of convex regression and also allows for the spectrahedral function to have additional block-diagonal structure, i.e., general (m, k)-spectrahedral reconstructions. Further, we provide algorithmic guarantees in the form of local convergence analysis of the alternating minimization procedure and we provide approximation-theoretic guarantees associated to (m, k)-spectrahedral functions (which rely on finite sample rather than asymptotic statistical analysis).

1.3 Notation

For $\mathcal{A} = (A_1, \ldots, A_d) \in (\mathbb{S}_k^m)^d$, we define for $x \in \mathbb{R}^d$ the linear pencil $\mathcal{A}[x] := \sum_{i=1}^d x_i A_i \in \mathbb{S}_k^m$. The usual vector ℓ_2 norm is denoted $\|\cdot\|_2$ and the sup norm by $\|\cdot\|_\infty$. The matrix Frobenius norm is denoted by $\|\cdot\|_F$, and the matrix operator norm by $\|\cdot\|_{op}$. We denote by $B_d(x, R)$ the ball in $\mathbb{R}^d$ centered at $x \in \mathbb{R}^d$ with radius $R > 0$.

2 Expressiveness of spectrahedral functions via statistical risk bounds

In this section, we first obtain upper bounds on the risk of the (m, k)-spectrahedral estimator in (3) decomposed into the approximation error and estimation error. We then compare this upper bound with known minimax lower bounds on the risk for certain classes of convex functions. This provides lower bounds on the approximation error of (m, k)-spectrahderal functions to these functions classes.

2.1 General Upper Bound on the Risk

To obtain an upper bound on the risk of the estimator (3), we use the general bound obtained in [10, Section 4.1]. To give the statement, consider first the following general framework. Let $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ be observations satisfying

$$y^{(i)} = f_*(x^{(i)}) + \varepsilon_i, \tag{4}$$

for a function $f_* : \mathbb{R}^d \to \mathbb{R}$ contained in some function class $\mathcal{F}$. We assume the errors ε_i are i.i.d. mean zero Gaussians with variance σ^2. Now, let $\{\mathcal{F}_\ell\}_{\ell \in \mathbb{N}}$ be a collection of function classes of growing complexity with m. For each m, define the constrained least squares estimator

$$\hat{f}_m^{(n)} := \text{argmin}_{f \in \mathcal{F}_m} \sum_{i=1}^n (y^{(i)} - f(x^{(i)}))^2.$$

We consider the risk of this estimator in the random design setting[1], where we assume $x^{(1)}, \ldots, x^{(n)}$ are i.i.d. random vectors in $\mathbb{R}^d$ with distribution μ. The risk is then defined by

$$\|\hat{f}_\ell^{(n)} - f_*\|_\mu^2 := \int_{\mathbb{R}^d} (\hat{f}_m^{(n)}(x) - f_*(x))^2 \mathrm{d}\mu(x).$$

Additionally, assume that both f_* and $\mathcal{F}_m$ are uniformly bounded by a positive and finite constant Γ.

As is standard in the theory of empirical processes, the rate is determined by the complexity of the class $\mathcal{F}_m$, which in this case is determined by the pseudo-dimension of the set

$$H_m := \{z \in \mathbb{R}^n : z = (f(x^{(1)}), \ldots, f(x^{(n)})) \text{ for some } f \in \mathcal{F}_m\}. \tag{5}$$

Recall that the pseudo-dimension of subset $B \subset \mathbb{R}^n$, denoted by $\text{Pdim}(B)$, is defined as the maximum cardinality of a subset $\sigma \subseteq \{1, \ldots, n\}$ for which there exists $h \in \mathbb{R}^n$ such that for every $\sigma' \subseteq \sigma$, one can find $a \in A$ with $a_i < h_i$ for $i \in \sigma'$ and $a_i > h_i$ for $i \in \sigma \backslash \sigma'$.

Theorem 4.2 in [10], stated below, provide an upper bound on the risk of $\hat{f}_m^{(n)}$ split into approximation error and estimation error.

Theorem 2. *Let $n \geq 7$. Suppose there is a constant $D_m \geq 1$ such that* $\text{Pdim}(H_m) \leq D_m$*. Then, there exists an absolute constants c such that*

$$\|\hat{f}_m^{(n)} - f_*\|_\mu^2 \leq c \left(\inf_{f \in \mathcal{F}_m} \|f - f_*\|_\mu^2 + \frac{\max\{\sigma^2, \Gamma^2\} D_m \log n}{n} \right). \tag{6}$$

The (m, k)-spectrahedral estimator (3) is a special case of the estimator $\hat{f}_m^{(n)}$ when $\mathcal{F}$ is the class of convex functions $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{F}_m$ is the class of (m, k)-spectrahedral functions as in Definition 1, denoted by $\mathcal{F}_{m,k}$. Since the class is parameterized by $d + 1$ matrices in $\mathbb{S}_k^m$, we define for each $m \in \mathbb{N}$ and $k = 1, \ldots, m$,

$$(\hat{A}_1, \ldots, \hat{A}_d, \hat{B}) \in \text{argmin}_{A_1, \ldots, A_d, B \in \mathbb{S}_k^m} \sum_{j=1}^n \left[y^{(j)} - \lambda_{\max} \left(\sum_{i=1}^d x_i^{(j)} A_i + B \right) \right]^2, \tag{7}$$

[1] One can also consider the risk in the fixed design setting, where one assumes the covariates $\{x^{(i)}\}_{i=1}^n$ are fixed, and risk bounds proved in [10] include this case. The results in this work can be directly extended to this case as well by applying the corresponding results.

and define the (m,k)-spectrahedral estimator of f_* by

$$\hat{f}_{m,k}(x) := \lambda_{\max}\left(\sum_{i=1}^{d} x_i \hat{A}_i + \hat{B}\right).$$

We also define the estimator when $\mathcal{F}$ is the class of support functions of convex bodies (compact and convex subsets) in $\mathbb{R}^d$, denoted by $\mathcal{K}$, and $\mathcal{F}_m$ is the subclass consisting of positively homogeneous (m,k)-spectrahedral functions, or equivalently, support functions of (m,k)-spectratopes. This corresponds to the case when the offset matrix $B = 0$. In this setting, we assume we are given observations $(u^{(1)}, y^{(1)}), \ldots, (u^{(n)}, y^{(n)}) \in \mathbb{S}^{d-1} \times \mathbb{R}$ satisfying

$$y^{(i)} = h_{K_*}(u^{(i)}) + \varepsilon_i,$$

where $h_K(u) := \sup_{x \in K} \langle u, x\rangle$, $u \in \mathbb{S}^{d-1}$, is the support function of a set $K_* \in \mathcal{K}$. We denote the class of (m,k)-spectratopes, or linear images of $\mathcal{S}_{m,k}$ in $\mathbb{R}^d$ by $\mathcal{L}(\mathcal{S}_{m,k})$. To define the (m,k)-spectratope estimator, let

$$(\hat{A}_1, \ldots, \hat{A}_d) \in \operatorname{argmin}_{A_1,\ldots,A_d, \in \mathbb{S}_k^m} \sum_{j=1}^{n} \left[Y_i - \lambda_{\max}\left(\sum_{i=1}^{d} u_i^{(j)} A_i\right)\right]^2,$$

and define

$$\hat{K}_{m,k} = \{z \in \mathbb{R}^d : z = (\langle \hat{A}_1, X\rangle, \ldots, \langle \hat{A}_d, X\rangle) \text{ for some } X \in \mathcal{S}_{m,k}\}.$$

For support function estimation, we notate the risk in terms of the convex bodies. Letting ν denote the probability distribution on $\mathbb{S}^{d-1}$ of $u^{(1)}$, we define the risk

$$\ell_\nu^2(\hat{K}, K) := \int_{\mathbb{S}^{d-1}} (h_{\hat{K}}(u) - h_K(u))^2 \mathrm{d}\nu(u).$$

In the following lemma, we prove an upper bound on the pseudo-dimension of the relevant set (5) needed to apply Theorem 2 for the estimators $\hat{f}_{m,k}$ and $\hat{K}_{m,k}$.

Lemma 3. *For $m, k \in \mathbb{N}$ such that k divides m, define for $x^{(1)}, \ldots, x^{(d)} \in \mathbb{R}^d$,*

$$H_{m,k} := \Big\{ z = \left(\lambda_{\max}\left(\mathcal{A}[x^{(1)}] + B\right), \ldots, \lambda_{\max}\left(\mathcal{A}[x^{(n)}] + B\right)\right) \in \mathbb{R}^n \\ \textit{for some } \mathcal{A} \in (\mathbb{S}_k^m)^d, B \in \mathbb{S}_k^m \Big\},$$

and for $u^{(1)}, \ldots, u^{(d)} \in \mathbb{S}^{d-1}$,

$$\tilde{H}_{m,k} := \left\{ z = \left(\lambda_{\max}\left(\mathcal{A}[u^{(1)}]\right), \ldots, \lambda_{\max}\left(\mathcal{A}[u^{(n)}]\right)\right) \in \mathbb{R}^n \textit{ for some } \mathcal{A} \in (\mathbb{S}_k^m)^d \right\},$$

Then, there exists absolute constant $c_1, c_2 > 0$ such that

$$Pdim(H_{m,k}) \le c_1 km(d+1)\log(c_2 n/k) \quad \textit{and} \quad Pdim(\tilde{H}_{m,k}) \le c_1 kmd\log(c_2 n/k).$$

To prove the lemma, we need the following known result (see for instance, Lemma 2.1 in [1]):

Proposition 4. *Let $p_1, \ldots, p_n$ be fixed polynomials of degree at most m in D variables for $D \le m$. The number of distinct sign vectors $(sgn(p_1(A)), \ldots, sgn(p_n(A)))$ that can be obtained by varying $A \in \mathbb{R}^D$ is at most $2\left(\frac{2enm}{D}\right)^D$.*

Proof. (of Lemma 3) Assume that the pseudo-dimension of $H_{m,k} \subset \mathbb{R}^n$ is ρ. By the definition of pseudo-dimension, the size of the collection of sign vectors

$$\mathcal{G}_{m,k} := \{(\operatorname{sgn}(\lambda_{\max}(\mathcal{A}[x^{(1)}] + B), \ldots, \operatorname{sgn}(\lambda_{\max}(\mathcal{A}[x^{(n)}] + B))) : \mathcal{A} \in (\mathbb{S}_k^m)^d, B \in \mathbb{S}_k^m\}$$

must be at most 2^ρ. For each i,

$$\operatorname{sgn}(\mathcal{A}[x^{(i)}] + B) = \operatorname{sgn}(\min\{p_1(\mathcal{A}, B; x^{(i)}), \ldots, p_m(\mathcal{A}, B; x^{(i)})\}),$$

where $p_\ell(\mathcal{A}, b; x^{(i)}) = \det(-(\mathcal{A}[x^{(i)}] + B)_{\ell:\ell})$ is the determinant of the $\ell \times \ell$ principal submatrix of $-\mathcal{A}[x^{(i)}] - B$. Indeed, $\lambda_{\max}(\mathcal{A}[x^{(i)}] + B) \le 0$ if and only if all of these determinants are non-negative. Thus, the size of $\mathcal{G}_{m,k}$ is the same size of

$$\mathcal{I}_{m,k} := \{(\operatorname{sgn}(p(\mathcal{A}, B; x^{(1)})), \ldots, \operatorname{sgn}(p(\mathcal{A}, B; x^{(n)}))) : \mathcal{A} \in (\mathbb{S}_k^m)^{d+1}\}),$$

where for each i, $p(\mathcal{A}, B; x^{(i)}) := \min\{p_1(\mathcal{A}, B; x^{(i)}), \dots, p_m(\mathcal{A}, B; x^{(i)})\}$ is a piecewise polynomial in $\mathcal{A}$. To bound the size of $\mathcal{I}_{m,k}$, we use the idea from [1]. We can partition $(\mathbb{S}_k^m)^{d+1}$ into at most mn regions over which the vector is coordinate-wise a fixed polynomial. Then we apply Proposition 4.

We have n polynomials of degree at most m in up to $D = (d+1)km$ variables, i.e. the number of degrees of freedom of $d+1$ $m \times m$ k-block matrices. Thus, the number of distinct sign vectors in $\mathcal{I}_m$ satisfies $|\mathcal{I}_m| \le 2mn \left(\frac{2en}{(d+1)k}\right)^{(d+1)km}$. This implies that $2^\rho \le 2mn \left(\frac{2en}{(d+1)k}\right)^{(d+1)km}$, and hence

$$\rho \le \frac{(d+1)km}{\log 2} \log\left(\frac{2en}{(d+1)k}\right) + \frac{\log(2mn)}{\log 2} \le c_1 km(d+1) \log\left(\frac{c_2 n}{k}\right).$$

The second claim follows similarly, where instead $D = dkm$. □

We can now obtain an upper bound on the risk of the estimators $\hat{f}_{m,k}$ and $\hat{K}_{m,k}$. Recall that we assume f_* and functions in $\mathcal{F}_{m,k}$ are uniformly bounded by some $\Gamma \in (0, \infty)$, and for support function estimation we assume K_* and elements of $\mathcal{L}(\mathcal{S}_{m,k})$ are contained in $B_d(0, \Gamma)$.

Theorem 5. *(i) For any convex function $f_* : \mathbb{R}^d \to \mathbb{R}$, there exist absolute constants c and b such that*

$$\|\hat{f}_{m,k} - f_*\|_\mu^2 \le c \left(\inf_{f \in \mathcal{F}_{m,k}} \|f - f_*\|_\mu^2 + \max\{\sigma^2, \Gamma^2\} km(d+1) \frac{\log(bn/k)}{n} \right).$$

(ii) For any convex body K_* in $\mathbb{R}^d$,

$$\ell_\nu^2(\hat{K}_{m,k}, K_*) \le c \left(\inf_{S \in \mathcal{L}(\mathcal{S}_{m,k})} \ell_\nu^2(S, K_*) + \max\{\sigma^2, \Gamma^2\} \frac{mk}{n} d \log(bn) \right)$$

Proof. This result follows from Theorem 2 and Lemma 3. □

2.2 Minimax Rates

The minimax risk for estimating a function in the class $\mathcal{F}$ from $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ in the random design setting is defined by

$$R_\mu(n, \mathcal{F}) := \min_{\hat{f}} \max_{f \in \mathcal{F}} \|\hat{f} - f\|_\mu.$$

In Table 1 we summarize known rates as $n \to \infty$ of this minimax risk for certain sub-classes of convex functions. First consider the class $\mathcal{F}_{m,k}(\Omega)$ of functions in $\mathcal{F}_{m,k}$ with compact and convex domain $\Omega \subset \mathbb{R}^d$. In this case, the approximation error in the risk bound is zero and the rate of convergence is $O\left(\frac{\log n}{n}\right)$. This is the best rate we can achieve when the domain Ω satisfies a certain smoothness assumption (see [10, Theorem 2.6]) and also appealing to the fact that $\mathcal{F}_{m,1} \subseteq \mathcal{F}_{m,k}$. Otherwise, the best lower bound we have is $O\left(\frac{1}{n}\right)$ using standard arguments for parametric estimation.

Additionally we consider two non-parametric sub-classes of convex functions. First is Lipschitz convex regression, where we assume the true function f_* belongs to the class $\mathcal{C}_L(\Omega)$ of L-Lipschitz convex functions with convex and compact full-dimensional support $\Omega \subset \mathbb{R}^d$. Second is support function estimation, where we assume the true function is the support function of a set K belonging to the collection $\mathcal{K}(\Gamma)$ of convex and compact subsets of $\mathbb{R}^d$ contained in the ball $B_d(0, \Gamma)$ for some finite $\Gamma > 0$. In both settings, the usual LSE over the whole class is minimax sub-optimal [15, 16], necessitating a regularized LSE to obtain the minimax rate.

Table 1: Minimax Rates for Sub-Classes of Convex Functions

$\mathcal{F}$	$\mathcal{F}_{m,k}(\Omega)$, for Ω smooth [10]	$\mathcal{C}_L(\Omega)$ [25]	$\mathcal{K}(\Gamma)$ [9]
$R_\mu(n, \mathcal{F})$	$\frac{\log n}{n}$	$n^{-\frac{4}{d+4}}$	$n^{-\frac{4}{d+3}}$

2.3 Approximation Rates

For Lipschitz convex regression, Lemma 4.1 in [25] implies the following: for $f_* \in \mathcal{C}_L(\Omega)$,

$$\inf_{f \in \mathcal{F}_{m,1}} \|f - f_*\|_\mu \le \inf_{f \in \mathcal{F}_{m,1}} \|f - f_*\|_\infty \le c_{d,\Omega,L} m^{-2/d}. \tag{8}$$

For support function estimation, let $d_H(S, K) := \|h_S - h_K\|_\infty$ denote the Hausdorff distance between any S and K in $\mathcal{K}$. A classical result of Bronshtein (see Section 4.1 in [3]) implies that

$$\inf_{S \in \mathcal{L}(\mathcal{S}_{m,1})} \ell_\nu(S, K) \leq \inf_{\mathcal{L}(\mathcal{S}_{m,1})} d_H(S, K) \leq c_{d,\Gamma} m^{-2/(d-1)}, \tag{9}$$

This result is also the core of the proof of (8).

We first show that inserting (8) and (9) into Theorem 5 and optimizing over m gives general upper bounds on the risk for our (m, k)-spectrahedral estimators. These rates match the minimax rate up to logarithmic factors for fixed $k > 0$, and even when k is allowed to depend logarithmically on m.

Corollary 6. *Suppose $k_m = f(m)$ for a non-decreasing and differentiable function $f : \mathbb{R} \to (0, m]$.*

(a) (Lipschitz convex regression) Suppose $f_* \in \mathcal{C}_L(\Omega)$ and define the function

$$g(m) := f'(m) m^{\frac{2d+4}{d}} + f(m) m^{\frac{d+4}{d}}.$$

Then, for $\alpha_n = g^{-1}\left(\frac{2n}{d(d+1)\max\{\sigma^2,\Gamma^2\}\log(bn)}\right)$,

$$\inf_{m \geq 1} \|\hat{f}_{m,k_m} - f_*\|_\mu^2 \leq c_{d,\Omega,\Gamma}\left(\alpha_n^{-4/d} + \max\{\sigma^2, \Gamma^2\}(d+1)\alpha_n f(\alpha_n)\frac{\log(bn)}{n}\right), \tag{10}$$

(b) (Support function estimation) Suppose $K_* \in \mathcal{K}(\Gamma)$ and define the function

$$g(m) := f'(m) m^{\frac{2(d+1)}{d-1}} + f(m) m^{\frac{d+3}{d-1}}.$$

Then, for $\alpha_n = g^{-1}\left(\frac{2n}{(d-1)d\max\{\sigma^2,\Gamma^2\}\log(bn)}\right)$,

$$\inf_{m \geq 1} \ell_\nu^2(\hat{K}_{m,k_m}, K_*) \leq c_{d,\Gamma}\left(\alpha_n^{-\frac{4}{d-1}} + \max\{\sigma^2, \Gamma^2\}(d+1)\alpha_n f(\alpha_n)\frac{\log(bn)}{n}\right). \tag{11}$$

We now provide two specific examples for particular functions f:

(i) If $f(m) = km^r$ for fixed $k > 0$ and $r \in [0, 1]$, then

$$\inf_{m \geq 1} \|\hat{f}_{m,k_m} - f_*\|_\mu^2 \leq O\left(n^{-\frac{4}{(r+1)d+4}} \log(bn)^{\frac{4}{(r+1)d+4}}\right),$$

and

$$\inf_{m \geq 1} \ell_\nu^2(\hat{K}_{m,k_m}, K_*) \leq O\left(n^{-\frac{4}{(r+1)(d-1)+4}} \log(bn)^{\frac{4}{(r+1)(d-1)+4}}\right).$$

(ii) If $f(m) = \log m$, then $\alpha_n = O\left(n^{\frac{d}{d+4}} \log(n)^{-\frac{2d}{d+4}}\right)$, and

$$\inf_{m \geq 1} \|\hat{f}_{m,k_m} - f_*\|_\mu^2 \leq O\left(n^{-\frac{4}{d+4}} \log(n)^{\frac{8}{d+4}}\right),$$

and

$$\inf_{m \geq 1} \ell_\nu^2(\hat{K}_{m,k_m}, K_*) \leq O\left(n^{-\frac{4}{d+3}} \log(n)^{\frac{8}{d+3}}\right).$$

Indeedn, the inverse of $h(x) := x^a \log(x)$ is $h^{-1}(x) = \left(\frac{ax}{W(ax)}\right)^{1/a}$, where W is the Lambert W function. The bound then follows from the fact that W satisfies $\log W(x) = \log x - W(x)$ and as $x \to \infty$, $W(x) \sim \log(x)$.

Remark 7. *For the case $k = 1$, Corollary 6 recovers the results in [9] and [10] showing these estimators obtain the minimax rate (up to logarithmic factors) for the relevant class of functions.*

Proof. We prove equation (10), and the second statement follows by a similar argument. By Theorem 6 and (8),

$$\|\hat{f}_{m,k_m} - f_*\|_\mu^2 \leq c_{d,\Omega,L}\left(m^{-4/d} + \frac{\max\{\sigma^2, \Gamma^2\}(d+1)f(m)m}{n}\log(bn)\right).$$

The $m_\star$ that minimizes the expression in the parentheses above satisfies

$$0 = -\frac{4}{d}(m_\star)^{-\frac{4}{d}-1} + \frac{\max\{\sigma^2, \Gamma^2\}(d+1)\log(bn)}{n}\left(f'(m_\star)m_\star + f(m_\star)\right),$$

or equivalently,

$$\frac{4n}{d(d+1)\max\{\sigma^2, \Gamma^2\}\log(bn)} = f'(m_\star)m_\star^{\frac{2d+4}{d}} + f(m_\star)m_\star^{\frac{d+4}{d}} = g(m_\star).$$

Then, $m_\star = g^{-1}\left(\frac{4n}{d(d+1)\max\{\sigma^2,\Gamma^2\}\log(bn)}\right)$ and plugging back into the upper bound gives the result. □

As stated previously, an important observation from Corollary 6 is that when $k_m = k$ is a fixed constant that does not depend on m, the risk bounds for an optimal choice $m_\star$ match (up to logarithmic factors) the minimax lower bounds of the classes $\mathcal{C}_L(\Omega)$ and $\mathcal{C}(\Gamma)$. This indicates that the approximation error for the classes $\mathcal{F}_{m,k}$ and $\mathcal{S}_{m,k}$ for fixed k cannot be improved from what was used in the proof. Indeed, this statistical risk analysis provides the following main result of this section: approximation rate lower bounds for the parametric classes $\mathcal{F}_{m,k}$ and $\mathcal{S}_{m,k}$.

Theorem 8. *Suppose there exists an absolute constant $c > 0$ and $t \in [0,1]$ such that $k_m \leq cm^t$ for all m large enough. Let $f_* \in \mathcal{C}_L(\Omega)$. For all $\varepsilon > 0$, for all m large enough,*

$$\inf_{f \in \mathcal{F}_{m,k_m}} \|f - f_*\|_\infty \geq c_{d,L,\Omega} m^{-2(1+t)/d - \varepsilon}.$$

Also, let $K_* \in \mathcal{K}(\Gamma)$. For all $\varepsilon > 0$, for all m large enough,

$$\inf_{S \in \mathcal{S}_{m,k_m}} d_H(S, K_*) \geq c_{d,\Gamma} m^{-2(1+t)/(d-1) - \varepsilon}.$$

Remark 9. *For constant k (i.e. $t = 0$), Theorem 8 implies*

$$\inf_{f \in \mathcal{F}_{m,k}} \|f - f_*\|_\infty = \tilde{O}(n^{-2/d}) \;\; \textit{and} \;\; \inf_{S \in \mathcal{S}_{m,k}} d_H(S, K_*) = \tilde{O}(n^{-2/(d-1)}),$$

where the $\tilde{O}$ notation ignores polylogarithmic factors.

Proof. We argue by contradiction. Suppose that for all $m > 0$,

$$\inf_{f \in \mathcal{F}_{m,k}} \|f - f_*\|_\mu^2 \leq c_1 m^{-r},$$

for some constant c_1 (that may depend on L and Ω) and fixed $r > \frac{4}{d}(1+t)$. Then by Theorem 5, there exist constants c_2, b such that

$$n^{-4/(d+4)} \leq c_2 \inf_{m>0} \left(m^{-r} + \max\{\sigma^2, \Gamma^2\} m^{t+1} (d+1) \frac{\log(bn/k)}{n} \right).$$

The infimum on the right side is achieved at $m_\star = \left(\frac{rn}{\max\{\sigma^2,\Gamma^2\}k(d+1)\log(bn/k)} \right)^{\frac{1}{t+r+1}}$, and thus

$$n^{-4/(d+4)} \leq c_2 n^{-\frac{r}{t+r+1}} \log(bn/k)^{\frac{r}{t+r+1}} (\max\{\sigma^2, \Gamma^2\} k(d+1))^{\frac{r}{t+r+1}} \left[r^{\frac{-r}{t+r+1}} + r^{\frac{t+1}{t+r+1}} \right].$$

For this inequality to hold for all n, it must be that $r \leq \frac{4}{d}(1+t)$, a contradiction. The second statement is proved similarly. □

3 Computational Guarantees

3.1 Alternating Minimization Algorithm

We now describe an alternating minimization algorithm to solve the non-convex optimization problem (3). Let $\xi_i = (x^{(i)}, 1) \in \mathbb{R}^{d+1}$ for each $i = 1, \ldots, n$ and let $\mathcal{A}_* \in (\mathbb{S}_k^m)^{d+1}$ be the true underlying parameters. That is, for each $i = 1, \ldots, n$, we observe

$$y_i = \lambda_{\max}(\mathcal{A}_*[\xi^{(i)}]) + \varepsilon_i.$$

We assume the ε_i's are i.i.d. mean zero Gaussian noise with variance σ^2.

One step of the algorithm starts with a fixed parameter $\mathcal{A} \in (\mathbb{S}_k^m)^{d+1}$. Then, compute the maximizing eigenvector $u^{(i)} \in \mathbb{S}^{m-1}$, $i = 1, \ldots, n$, such that for $U^{(i)} = u^{(i)}(u^{(i)})^T$, $\langle U^{(i)}, \mathcal{A}[\xi^{(i)}] \rangle = \lambda_{\max}\left(\mathcal{A}[\xi^{(i)}]\right)$. With the $U^{(i)}$'s fixed, update $\mathcal{A}$ by solving the linear least squares problem:

$$\mathcal{A}^+ \in \operatorname{argmin}_{\mathcal{A} \in (\mathbb{S}_k^m)^{d+1}} \frac{1}{n} \sum_{i=1}^n \left(y^{(i)} - \langle U^{(i)}, \mathcal{A}[\xi^{(i)}] \rangle \right)^2, \tag{12}$$

where $\langle U^{(i)}, \mathcal{A}[\xi^{(i)}] \rangle = \langle \mathcal{A}, \xi^{(i)} \otimes U^{(i)} \rangle = \sum_{j=1}^d \langle A_j, \xi_j^{(i)} U^{(i)} \rangle$.

Algorithm 1 Alternating Minimization for Spectrahedral Regression

```
**Input**: Collection of inputs and outputs $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$; initialization $\mathcal{A} \in (\mathbb{S}_k^m)^{d+1}$
**Algorithm**: Repeat until convergence
    **Step 1**: Update optimal eigenvector $u^{(i)} \leftarrow \lambda_{\max}(\mathcal{A}[\xi^{(i)}])$
    **Step 2**: Update $\mathcal{A}$ by solving (12). $\mathcal{A}^+ \leftarrow (\Xi_{\mathcal{A}}^T \Xi_{\mathcal{A}})^{-1} \Xi_{\mathcal{A}}^T y$, where $\Xi_{\mathcal{A}}^T = (\xi^{(1)} \otimes U^{(1)} | \cdots | \xi^{(n)} \otimes U^{(n)}) \in \mathbb{R}^{(d+1)m^2 \times n}$.
**Output**: Final iterate $\mathcal{A}$
```

3.2 Convergence Guarantee

The following result shows that under certain conditions, this alternating minimization procedure converges geometrically to a small ball around the true parameters given a good initialization. To state the initialization condition in the result, we define for $\mathcal{A} \in (\mathbb{S}_k^m)^d$ the similarity transformation $\mathcal{O}(\mathcal{A}) = (OA_1O^T, \ldots, OA_dO^T)$ for an orthogonal $m \times m$ matrix O. Note that the eigenvalues of $\mathcal{A}[x]$ for $x \in \mathbb{R}^d$ are invariant under any $\mathcal{O}$. In the following we only consider the setting where $k = m$, and denote $\mathbb{S}^m := \mathbb{S}_m^m$.

The proof of the following result appears after the statement and it depends on multiple lemmas that we state and prove in the appendix.

Theorem 10. *Assume X is a standard Gaussian random vector in $\mathbb{R}^d$ and let $\xi = (X, 1) \in \mathbb{R}^{d+1}$. Also suppose that the true parameter $\mathcal{A}_* \in (\mathbb{S}^m)^{d+1}$ satisfies the following spectral condition:*

$$\inf_{u \in \mathbb{S}^{d-1}} \lambda_1(\mathcal{A}_*[u]) - \lambda_2(\mathcal{A}_*[u]) := \kappa > 0, \tag{13}$$

where $\lambda_1 := \lambda_{\max}$ and λ_2 is the second largest eigenvalue. Let $\tau \in (0, 1)$. There exist constants c_i, $i = 1, \ldots, 5$ such that if the initial parameter choice $\mathcal{A}^{(0)}$ satisfies

$$\|\mathcal{A}^{(0)} - \mathcal{O}(\mathcal{A}_*)\|_F^2 \leq \frac{\kappa^2}{16(d+1)m^2} \left(\frac{1-\tau}{1+\tau} \right),$$

for some similarity transformation $\mathcal{O}$ and

$$n \geq c_1(d+1) \max \left\{ \tau^{-2} m^{10}, \frac{1}{\kappa^2} \left(\frac{1+\tau}{1-\tau} \right) \frac{m^6(d+1)\sigma^2 \log(n)^2}{(1-\tau)} \right\},$$

then the error at iteration t satisfies

$$\|\mathcal{A}^{(t)} - \mathcal{O}(\mathcal{A}_*)\|_F^2 \leq \left(\frac{3}{4} \right)^t \|\mathcal{A}^{(0)} - \mathcal{O}(\mathcal{A}_*)\|_F^2 + \frac{c_3 m^4 (d+1)\sigma^2 \log(n)^2}{n(1-\tau)},$$

with probability greater than $1 - 4e^{-c_4 \tau^2 n / m^{10}} - n^{-c_5 m^2 (d+1)}$.

Remark 11. *The assumption* (13) *is not always satisfied, but we provide some examples where it is. First, consider the case where $d = 2$, A_1 and A_2 are non-commutative matrices, and $A_3 = 0$. Let $a_{ij} = (a_{ij}^{(1)}, a_{ij}^{(2)}) \in \mathbb{R}^2$. In this case the eigengap is*

$$\lambda_1(u_1A_1 + u_2A_2) - \lambda_2(u_1A_1 + u_2A_2) = \langle u, a_{11} - a_{22} \rangle^2 + 4\langle u, a_{12} \rangle^2.$$

This follows from the computation of eigenvalues using the characteristic polynomial and the quadratic formula. We see that for any A_1 and A_2 such that $a_{12} \notin \{0, a_{11} - a_{22}\}$, the eigengap is a strictly positive number. For example, if $a_{11} - a_{22} = e_1$ and $2a_{12} = e_2$, then

$$\lambda_1(u_1A_1 + u_2A_2) - \lambda_2(u_1A_1 + u_2A_2) = \langle u, e_1 \rangle^2 + \langle u, e_2 \rangle^2 = u_1^2 + u_2^2 = 1.$$

Another set of parameters $\mathcal{A}_*$ that satisfy condition (13) *is when $\lambda_{\max}(\mathcal{A}_*[x]) = \|x\|_2$. In fact, for any spectrahedral function $f(x) = \lambda_{\max}(\mathcal{A}_*[x])$ that is differentiable for all x, $\mathcal{A}_*$ must necessarily satisfy* (13).

Proof. First, given assumption (13), we show that for n large enough, for all parameters $\mathcal{A}$ satisfying for some similarity transform $\mathcal{O}$,

$$\|\mathcal{A} - \mathcal{O}(\mathcal{A}_*)\|_F^2 \leq \frac{\kappa^2}{16(d+1)m^2} \left(\frac{1-\tau}{1+\tau} \right), \tag{14}$$

the parameter $\mathcal{A}^+$ obtained after applying one iteration of the algorithm satisfies

$$\|\mathcal{A}^+ - \mathcal{O}(\mathcal{A}_*)\|_F^2 \leq \frac{3}{4} \|\mathcal{A} - \mathcal{O}(\mathcal{A}^*)\|_F^2 + O\left(\frac{\log n}{n} \right) \tag{15}$$

with high probability.

Let $U^{(i)} = u^{(i)}(u^{(i)})^T$ be such that $\lambda_{\max}(\mathcal{A}[\xi^{(i)}]) = \langle U^{(i)}, \mathcal{A}[\xi^{(i)}]\rangle$. The update $\mathcal{A}^+$ then equals

$$\mathcal{A}^+ = (\Xi_{\mathcal{A}}^T \Xi_{\mathcal{A}})^{-1}\Xi_{\mathcal{A}}^T y,$$

where $\Xi_{\mathcal{A}}^T = (\xi^{(1)} \otimes U^{(1)}| \cdots |\xi^{(n)} \otimes U^{(n)}) \in \mathbb{R}^{(d+1)m\times m\times n}$. Note that $(\Xi_{\mathcal{A}}\mathcal{A})_i = \langle U^{(i)}, \mathcal{A}[\xi^{(i)}]\rangle$. Throughout the rest of the proof, we sometime abuse notation and consider the Kronecker product $\xi \otimes U$ for $\xi \in \mathbb{R}^{d+1}$ and $U \in \mathbb{R}^{m\times m}$ to be the vector $\mathrm{Vec}(\xi \otimes U) \in \mathbb{R}^{(d+1)m^2}$.

By the invariance $\lambda_{\max}(\mathcal{A}[x]) = \lambda_{\max}(\mathcal{O}(\mathcal{A})[x])$ for all $\mathcal{O}$, without loss of generality we can assume in the following that $\mathcal{A}_* = \mathcal{O}(\mathcal{A}_*)$ for the transformation $\mathcal{O}$ satisfying assumption (14). Let $y_* \in \mathbb{R}^n$ and $u_*^{(i)} \in \mathbb{S}^{d-1}$ be such that for $U_*^{(i)} := u_*^{(i)}\left(u_*^{(i)}\right)^T$,

$$y_i^* = \langle U_*^{(i)}, \mathcal{A}_*[\xi^{(i)}]\rangle = \lambda_{\max}(\mathcal{A}_*[\xi^{(i)}]).$$

Also denote by $P_{\Xi_{\mathcal{A}}} = \Xi_{\mathcal{A}}(\Xi_{\mathcal{A}}^T\Xi_{\mathcal{A}})^{-1}\Xi_{\mathcal{A}}^T$ the orthogonal projection matrix onto the span of the columns of $\Xi_{\mathcal{A}}$. Then, we have the following deterministic upper bound:

$$\begin{aligned}\|\Xi_{\mathcal{A}}(\mathcal{A}^+ - \mathcal{A}_*)\|^2 &= \|P_{\Xi_{\mathcal{A}}} y - \Xi_{\mathcal{A}}\mathcal{A}_*\|^2 = \|P_{\Xi_{\mathcal{A}}} y^* + P_{\Xi_{\mathcal{A}}}\varepsilon - \Xi_{\mathcal{A}}\mathcal{A}_*\|^2 \\ &\leq 2\|P_{\Xi_{\mathcal{A}}}(y^* - \Xi_{\mathcal{A}}\mathcal{A}_*)\|^2 + 2\|P_{\Xi_{\mathcal{A}}}\varepsilon\|^2 \\ &\leq 2\sum_{i=1}^n \left(\langle U_*^{(i)}, \mathcal{A}_*[\xi^{(i)}]\rangle - \langle U^{(i)}, \mathcal{A}_*[\xi^{(i)}]\rangle\right)^2 + 2\|P_{\Xi_{\mathcal{A}}}\varepsilon\|^2.\end{aligned}$$

Now, since $\langle U^{(i)} - U_*^{(i)}, \mathcal{A}[\xi^{(i)}]\rangle \geq 0$,

$$\begin{aligned}\left(\langle U_*^{(i)}, \mathcal{A}_*[\xi^{(i)}]\rangle - \langle U^{(i)}, \mathcal{A}_*[\xi^{(i)}]\rangle\right)^2 &\leq \left(\langle U_*^{(i)} - U^{(i)}, \mathcal{A}_*'\xi^{(i)}\rangle + \langle U^{(i)} - U_*^{(i)}, \mathcal{A}[\xi^{(i)}]\rangle\right)^2 \\ &= \left\langle \mathcal{A} - \mathcal{A}_*, \xi^{(i)} \otimes (U^{(i)} - U_*^{(i)})\right\rangle^2.\end{aligned}$$

We also have the lower bound $\|\Xi_{\mathcal{A}}(\mathcal{A}^+ - \mathcal{A}_*)\|^2 \geq \lambda_{\min}(\Xi_{\mathcal{A}}^T\Xi_U)\|\mathcal{A}^+ - \mathcal{A}_*\|^2$. Thus,

$$\begin{aligned}\|\mathcal{A}^+ - \mathcal{A}_*\|^2 &\leq \frac{2}{\lambda_{\min}(\Xi_{\mathcal{A}}^T\Xi_{\mathcal{A}})}\left[\|\Xi_{\mathcal{A}-\mathcal{A}_*}(\mathcal{A} - \mathcal{A}_*)\|_2^2 + \|P_{\Xi_{\mathcal{A}}}\varepsilon\|^2\right] \\ &\leq \frac{2}{\lambda_{\min}(\Xi_{\mathcal{A}}^T\Xi_{\mathcal{A}})}\left[\lambda_{\max}(\Xi_{\mathcal{A}-\mathcal{A}_*}^T\Xi_{\mathcal{A}-\mathcal{A}_*})\|\mathcal{A} - \mathcal{A}_*\|^2 + \|P_{\Xi_{\mathcal{A}}}\varepsilon\|^2\right]. \end{aligned} \tag{16}$$

where $\Xi_{\mathcal{A}-\mathcal{A}_*} = \left(\xi^{(1)} \otimes (U^{(1)} - U_*^{(1)})| \cdots |\xi^{(n)} \otimes (U^{(n)} - U_*^{(n)})\right)$.

Lemmas 12 and 13 then imply the following. For $\tau \in (0,1)$, there exist absolute constants c_1, c_2 such that if $n \geq c_1\tau^{-2}(d+1)m^{10}$, then with probability greater than $1 - 2e^{-c_2\tau^2 n/m^{10}}$,

$$\lambda_{\max}(\Xi_{\mathcal{A}-\mathcal{A}_*}^T\Xi_{\mathcal{A}-\mathcal{A}_*}) \leq n\lambda_{\max}(\mathbb{E}[(\xi \otimes (U - U_*))(\xi \otimes (U - U_*))^T])\,(1+\tau) \tag{17}$$

for all $\mathcal{A}$ satisfying assumption (14). Since $\lambda_{\max}$ is convex function, Jensen's inequality implies

$$\lambda_{\max}(\mathbb{E}[(\xi \otimes (U - U_*))(\xi \otimes (U - U_*))^T]) \leq \mathbb{E}[\|\xi \otimes (U - U_*)\|^2].$$

Then, by the definition of the Kronecker product,

$$\|\xi \otimes (U - U_*)\|^2 = \sum_{i=1}^d \sum_{j,k=1}^m \xi_i^2 (U - U_*)_{jk}^2 = \|\xi\|_2^2\|U - U_*\|_F^2.$$

Next note that $\|U - U_*\|_F^2 \leq 2\|u - u_*\|_2^2$, where $U = uu^T$, $U_* = u_*u_*^T$ and $u, u_* \in \mathbb{S}^{d-1}$. Then, by a variation of the Davis-Kahan Theorem (Corollary 3 in [28]),

$$\|u - u_*\|_2 \leq \frac{2^{3/2}\|(\mathcal{A} - \mathcal{A}_*)'\xi\|_{op}}{\lambda_1(\mathcal{A}_*'\xi) - \lambda_2(\mathcal{A}_*'\xi)} \leq 2^{3/2}\kappa^{-1}\|\mathcal{A} - \mathcal{A}_*\|_F.$$

Putting the bounds together and using assumption (14),

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[(\xi \otimes (U - U_*))(\xi \otimes (U - U_*))^T]) &\leq 2^{5/2}\kappa^{-2}\|\mathcal{A} - \mathcal{A}_*\|_F^2\mathbb{E}[\|\xi\|_2^2] \\ &\leq 6\kappa^{-2}(d+1)\|\mathcal{A} - \mathcal{A}_*\|_F^2 \leq \frac{3}{8m^2}\left(\frac{1-\tau}{1+\tau}\right). \end{aligned} \tag{18}$$

Plugging the bound (18) into (17) gives

$$\lambda_{\max}(\Xi_{\mathcal{A}-\mathcal{A}_*}^T \Xi_{\mathcal{A}-\mathcal{A}_*}) \leq \frac{3n}{8m^2}(1-\tau)\,. \tag{19}$$

Also by Lemmas 12 and 13 if $n \geq c_1\tau^{-2}(d+1)m^{10}$, then with probability greater that $1-2e^{-c_2\tau^2 n/m^{10}}$,

$$\lambda_{\min}(\Xi_{\mathcal{A}}^T \Xi_{\mathcal{A}}) \geq n\lambda_{\max}(\mathbb{E}[(\xi\otimes U)(\xi\otimes U)^T])\,(1-\tau) \tag{20}$$

for all $\mathcal{A}$ satisfying (14). We then have the following lower bound:

$$\begin{aligned}\lambda_{\max}(\mathbb{E}[(\xi\otimes U)(\xi\otimes U)^T]) &\geq \frac{1}{(d+1)m^2}\mathrm{Tr}\left[\mathbb{E}(\xi\otimes U)(\xi\otimes U)^T\right]\\ &= \frac{1}{(d+1)m^2}\sum_{i=1}^{d+1}\sum_{j,k=1}^{m}\mathbb{E}[\xi_i^2(u_ju_k)^2] = \frac{1}{(d+1)m^2}\sum_{i=1}^{d+1}\mathbb{E}[\xi_i^2] = m^{-2}.\end{aligned} \tag{21}$$

Plugging the bound (21) into (20) gives

$$\lambda_{\min}(\Xi_{\mathcal{A}}^T \Xi_{\mathcal{A}}) \geq nm^{-2}\,(1-\tau)\,, \tag{22}$$

and finally combining (19) and (22) with (16) implies

$$\|\mathcal{A}^+ - \mathcal{A}^*\|_F^2 \leq \frac{3}{4}\|\mathcal{A}-\mathcal{A}^*\|_F^2 + \frac{m^2\|P_{\Xi_{\mathcal{A}}}\varepsilon\|_2^2}{n(1-\tau)}.$$

It remains to bound the error term. For this, we apply Lemma 14, which says that there exist constants $c_3, c_4 > 0$ such that

$$\|P\varepsilon\|_2^2 \leq c_3\log(n)^2\sigma^2 m^2(d+1)$$

for all $\mathcal{A}$ satisfying (14) with probability greater than $1-e^{-c_4(d+1)m^2\log(n)}$.

This implies that for $n \geq c_1\tau^{-2}(d+1)m^{10}$, with probability $1-4e^{-c_2\tau^2 n/m^{10}} - n^{-c_4 m^2(d+1)}$,

$$\|\mathcal{A}^+ - \mathcal{A}^*\|_F^2 \leq \frac{3}{4}\|\mathcal{A}-\mathcal{A}^*\|_F^2 + \frac{c_3 m^4(d+1)\sigma^2\log(n)^2}{n(1-\tau)}.$$

We now show that given the above upper bound, $\mathcal{A}^+$ also satisfies (14). Indeed, for

$$n \geq 4\cdot 16m^2(d+1)\frac{1}{\kappa^2}\left(\frac{1+\tau}{1-\tau}\right)\frac{c_3 m^4(d+1)\sigma^2\log(n)^2}{(1-\tau)},$$

we have

$$\frac{c_3 m^4(d+1)\sigma^2\log(n)^2}{n(1-\tau)} \leq \frac{\kappa^2}{4}\cdot\frac{1}{16m^2(d+1)}\left(\frac{1-\tau}{1+\tau}\right)$$

and thus

$$\|\mathcal{A}^+ - \mathcal{A}_*\|_F^2 \leq \frac{3}{4}\|\mathcal{A}-\mathcal{A}_*\|_F^2 + \frac{c_3 m^4(d+1)\sigma^2\log(n)^2}{n(1-\tau)} \leq \frac{\kappa^2}{16m^2(d+1)}\left(\frac{1-\tau}{1+\tau}\right).$$

The final conclusion follows from the fact that after t iterations, applying the bound (15) t times gives

$$\begin{aligned}\|\mathcal{A}^{(t)} - \mathcal{A}_*\|_F^2 &\leq \left(\frac{3}{4}\right)^t\|\mathcal{A}^{(0)}-\mathcal{A}_*\|_F^2 + \frac{c_3 m^4(d+1)\sigma^2\log(n)^2}{n(1-\tau)}\sum_{k=0}^{\infty}\left(\frac{3}{4}\right)^k\\ &\leq \left(\frac{3}{4}\right)^t\|\mathcal{A}^{(0)}-\mathcal{A}_*\|_F^2 + \frac{c_4 m^4(d+1)\sigma^2\log(n)^2}{n(1-\tau)},\end{aligned}$$

and all t bounds hold simultaneously with probability at least $1-4e^{-c_2\tau^{-2}n/m^{10}} - n^{-c_4 m^2(d+1)}$. □

4 Numerical Experiments

In this section, we empirically compare spectrahedral and polyhedral regression for estimating a convex function from data. More specifically, we compare (m,m)-spectrahedral estimators to $m(m+1)/2$-polyhedral estimators, both of which have $m(m+1)/2$ degrees of freedom per dimension. For each experiment, we apply the alternating minimzation algorithm with multiple random initializations, and the solution that minimizes the least squared error is selected. We adapted the code [23] for support function estimation used in [24] for spectrahedral regression.

4.1 Synthetic Regression Problems

The first experiments use synthetically generated data from a known convex function, one from a spectrahedral function and another from a convex function that is neither polyhedral nor spectrahedral. In both problems below, the root-mean-squared error (RMSE) is obtained by first obtaining estimators form 200 noisy training data points and then evaluating the RMSE of the estimators on 200 test points generated from the true function. We ran the alternating minimization algorithm with 50 random initializations for 200 steps or until convergence, and chose the best estimator.

First, we consider n i.i.d. data points distributed as (X, Y), where $X \in \mathbb{R}^2$ is uniformly distributed in $[-1, 1]^2$, and

$$Y = \sqrt{X_1^2 + X_2^2} + \varepsilon, \tag{23}$$

where $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. In Figure 2, we have plotted polyhedral and spectrahedral estimators obtained from $n = 20, 50$ and 200 data points. The RMSE for both models is given in Table 2. The function $y = \|x\|_2$ is a spectrahedral function, and the spectrahedral estimator performs better than the polyhedral estimator as expected.

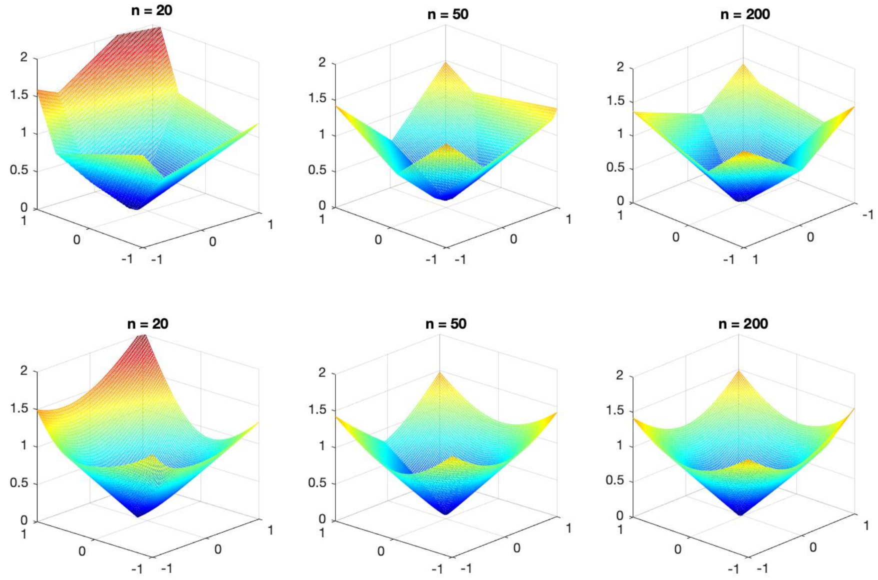

Figure 2: Polyhedral ($m = 6$) and Spectrahedral ($m = 3$) reconstructions of the convex function $y = \|x\|_2$ from $n = 20$, 50, and 200 data points from model (23).

Second, we consider n i.i.d. data points generated as $(X, Y) \in \mathbb{R} \times \mathbb{R}$, where $X \sim \mathcal{N}(0, 1)$ and

$$Y = \exp(bX) + \varepsilon, \tag{24}$$

where $b = 1.1394$ and $\varepsilon \sim \mathcal{N}(0, 0.1^2)$. The underlying convex function is neither polyhedral nor spectrahedral, but the spectrahedral estimator better captures the smoothness of the function as illustrated in Figure 3. The spectrahedral estimator also outperforms the polyhedral estimator with respect to the RMSE when comparing the model fitted to the training data set to the test data set, see Table 2.

4.2 Predicting Average Weekly Wages

The first experiment we perform on real data is predicting average weekly wages based on years of education and experience. This data set is also studied in [12]. The data set is from the 1988 Current Population Survey (CPS) and can be obtained as the data set ex1029 in the Sleuth2 package in R. It consists of 25,361 records of weekly wages for full-time, adult, male workers for 1987, along with years experience and years of education. It is reasonable to expect that wages are concave with respect the years experience. Indeed, at first wages increase with more experience, but with a decreasing return each year until a peak of earnings is reached, and then they begin to decline. Wages are also expected to increase as the number of years of education increases,

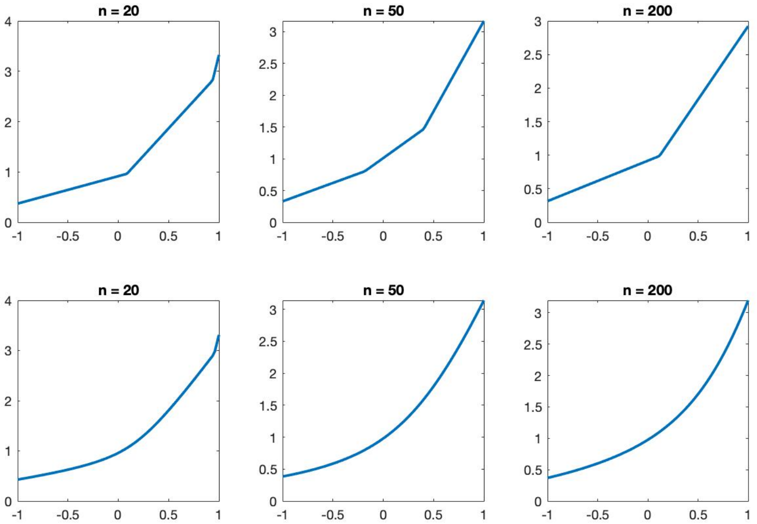

Figure 3: Polyhedral ($m = 6$) and Spectrahedral ($m = 3$) reconstructions of the convex function $y = \exp(\langle x, b \rangle)$ from $n = 20$, 50, and 200 noisy data points from model (24).

Table 2: RSME for polyhedral and spectrahedral estimators of $y = \exp(bx)$ from model (24) as m increases.

Model	$m(m+1)/2$	Spectrahedral	Polyhedral
(23)	3	0.0183	0.1332
	6	0.0207	0.0416
	10	0.0243	0.0362
(24)	3	0.1098	0.2281
	6	0.0902	0.1153
	10	0.0793	0.0902

but not in a concave way. However, as in [12], we use the transformation $1.2^{\text{years education}}$ to obtain a concave relationship. We used polyhedral and spectrahedral regression to fit convex functions to this data set, as illustrated in Figure 1. We also estimated the RMSE for different values of $m(m+1)/2$ (the degrees of freedom per dimension) through hold-out validation with 20% of the data points, see Table 3. This generalization error is smaller for the spectrahedral estimator than the polyhedral estimator in each case.

4.3 Convex Approximation in Engineering Applications

In the following two examples, we consider applications of convex regression in engineering applications where the goal is to subsequently use the convex estimator as an objective or constraint in an optimization problem. Polyhedral regression returns a convex function compatible with a linear program, and using spectrahedral regression provides an estimator compatible with semidefinite programming.

4.3.1 Aircraft Data

In this experiment, we consider the XFOIL aircraft design problem studied in [13]. The profile drag on an airplane wing is described by a coefficient CD that is a function of the Reynolds number (Re), wing thickness ratio (τ), and lift coefficient (CL). There is not an analytical expression for this relationship, but it can be simulated using XFOIL [4]. For a fixed τ, after a logarithmic transformation, the data set can be approximated well by a convex function. We fit both spectrahedral and polyhedral functions to this data set, and the best fits for the whole data set appears in Figure 4 for models with 6 degrees for freedom per dimension. Then, we performed hold-out validation, training on 80% of the data and testing on the remaining 20%. The RMSE is given in Table 3, where we observe that the spectrahedral estimator achieves a smaller error than polyhedral regression.

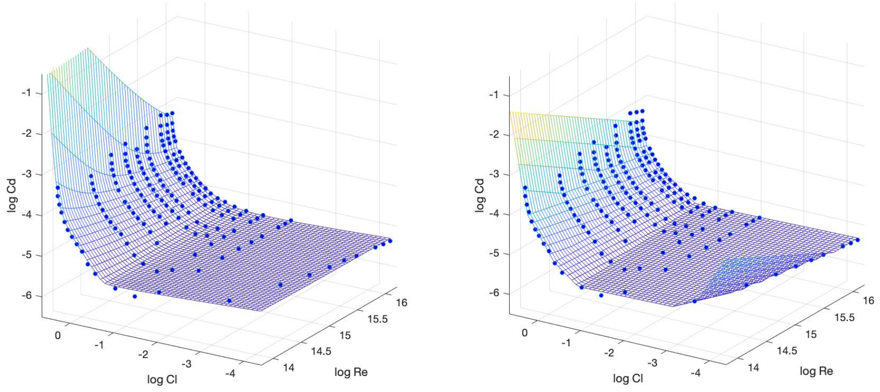

Figure 4: Spectrahedral ($m = 3$) and Polyhedral ($m = 6$) estimators of the log of drag coefficient vs log of Reynolds number and lift coefficient for a fixed thickness ratio $\tau = 8\%$.

Table 3: RSME for polyhedral and spectrahedral estimators for real data and engineering experiments.

Application	$m(m+1)/2$	Spectrahedral	Polyhedral
Average Weekly Wages	3	142.1166	145.5803
	6	140.1173	141.4989
	10	140.0352	141.9851
Aircraft Profile Drag	3	0.086	0.0895
	6	0.0576	0.0709
	10	0.0452	0.0515
Circuit Design	3	0.0085	0.02
	6	0.0072	0.012
	10	0.0072	0.0088

4.3.2 Power Modeling For Circuit Design

A circuit is an interconnected collection of electrical components including batteries, resistors, inductors, capacitors, logical gates, and transistors. In circuit design, the goal is to optimize over variables such as devices, gates, threshold, and power supply voltages in order to minimize circuit delay or physical area. The power dissipated, P, is a function of gate supply V_{dd} and threshold voltages V_{th}. The following model, see [13] and [11], can be used to study this relationship:

$$P = V_{dd}^2 + 30V_{dd}e^{-(V_{th}-0.06V_{dd})/0.039}.$$

We generate n i.i.d. data points as in [11] as follows. For each input-output pair, first sample $u = (V_{dd}, V_{th})$ uniformly over the domain $1.0 \leq V_{dd} \leq 2.0$ and $0.2 \leq V_{th} \leq 0.4$ and compute $P(u)$. Then, apply the transformation $(x, y) = (\log u, \log P(u))$. We fit this collection of transformed data points using polyhedral and spectrahedral regression, and the estimators for $n = 20, 50$, and 200 are illustrated in Figure 5. We also perform hold-out validation with 20% of the data for the case $n = 200$ and the RMSE appears in in Table 3. By this measure, the spectrahedral estimator performs much better than the polyhedral estimator in this application.

5 Discussion and Future Work

In this work, we have introduced spectrahedral regression as a new method for estimating a convex function from noisy measurements. Spectrahedral estimators are appealing from a qualitative and quantitative perspective and we have shown they hold advantages over the usual LSE methods as well as polyhedral estimators when the underlying convex function is non-polyhedral. Our theoretical results and numerical experiments call for further study of the expressivity of this model class and its computational advantages. We now describe a few directions of future research.

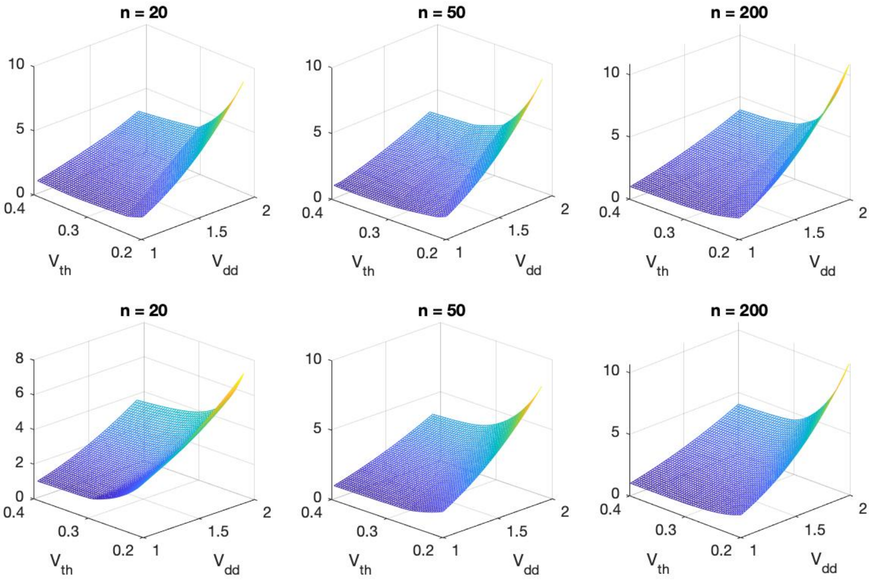

Figure 5: Polyhedral ($m = 6$) and Spectrahedral ($m = 3$) estimators of $n = 20, 50$, and 200 transformed data points generated from the power dissipation model.

An interesting open question is to obtain the approximation rate for (m, k)-spectrahedral functions to the class of Lipschitz convex functions and (m, k)- spectratopes to the class of convex bodies for general k. There is extensive literature on this approximation question for polytopes (see, for instance, [3, 5]), and we have obtained matching bounds (up to logarithmic factors) for fixed $k > 1$. For k depending on m, and in particular in the case $k = m$, the literature is more limited, one example is [2]. Progress in this direction would complete our understanding of the expressive power of the model presented here and have important consequences for how well semidefinite programming can approximate a general convex program.

We have also proved computational guarantees for a natural alternating minimization algorithm for spectrahedral regression. However, this convergence guarantee depends on a good initialization. In practice, running the algorithm with multiple random initializations and taking the estimator with the smallest error works well, but it would be very interesting to extend the results on initialization in [7] to the spectrahedral case. Another line of future work is to extend other methods to solve the non-convex optimization (3) in the max-affine case such as the adaptive partitioning method in [12] and the method proposed in [25]. These algorithms also lack theoretical guarantees and it would be interesting to obtain conditions under which these methods obtain good estimates of the true parameter.

Acknowledgments

E. O. was supported by NSF MSPRF Award 2002255 with additional funding from ONR Award N00014-18-1-2363. V. C. was supported in part by National Science Foundation grant CCF-1637598, in part by National Science Foundation grant DMS-2113724, and in part by AFOSR grant FA9550-20-1-0320.

References

[1] P. L. Bartlett, V. Maiorov, and R. Meir, *Almost linear VC-dimension bounds for piecewise polynomial networks*, Neural Computation, 10 (1998), pp. 2159–2173.

[2] A. I. Barvinok, *Approximations of convex bodies by polytopes and by projections of spectrahedra*, arXiv:1204.0471, (2012).

[3] E. M. Bronshtein, *Approximation of convex sets by polytopes*, Journal of Mathematical Sciences, 153 (2008).

[4] M. Drela, *Xfoil subsonic airfoil development system. Open source software available at http://web.mit.edu/drela/Public/web/xfoil/*, 2000.

[5] R. M. Dudley, *Metric entropy of some classes of sets with differentiable boundaries*, Journal of Approximation Theory, 10 (1974), pp. 227–236.

[6] H. Fawzi, J. Gouveia, P. A. Parrilo, J. Saunderson, and R. R. Thomas, *Lifting for simplicity: Concise descriptions of convex sets*, arXiv:2002.09788, (2020).

[7] A. Ghosh, A. Guntuboyina, A. Pananjady, and K. Ramchandran, *Max-affine regression I: Parameter estimation for Gaussian designs*. `https://tinyurl.com/spyjmgv`, 2019.

[8] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran, *Max-affine regression with universal parameter estimation for small-ball designs*, in 2020 IEEE International Symposium on Information Theory (ISIT), 2020, pp. 2706–2710.

[9] A. Guntuboyina, *Optimal rates of convergence for convex set estimation from support functions*, The Annals of Statistics, 40 (2012), pp. 385–411.

[10] Q. Han and J. A. Wellner, *Multivariate convex regression: global risk bounds and adaptation.*, arXiv:1601.06844, (2016).

[11] L. A. Hannah and D. B. Dunson, *Ensemble methods for convex regression with applications to geometric programming based circuit design*, Proceedings of the 29th International Conference on Machine Learning, (2012).

[12] L. A. Hannah and D. B. Dunson, *Multivariate convex regression with adaptive partitioning*, Journal of Machine Learning Research, 14 (2013), pp. 3261–3294.

[13] W. W. Hoburg, P. G. Kirschen, and P. Abbeel, *Fitting geometric programming models to data*, Optimization and Engineering, (2015).

[14] Y. Klochkov and N. Zhivotovskiy, *Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method*, Electronic Journal of Probability, 25 (2020), pp. 1–30.

[15] G. Kur, F. Gao, A. Guntuboyina, and B. Sen, *Convex regression in multidimensions: Suboptimality of least squares estimators*, arXiv:2006.02044, (2020).

[16] G. Kur, A. Rakhlin, and A. Guntuboyina, *On suboptimality of least squares with application to estimation of convex bodies*, in Proceedings of Thirty Third Conference on Learning Theory, J. Abernethy and S. Agarwal, eds., vol. 125 of Proceedings of Machine Learning Research, PMLR, July 2020, pp. 2406–2424.

[17] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes.*, Springer-Verlag Berlin Heidelberg, 2013.

[18] A. Magnani and S. P. Boyd, *Convex piecewise-linear fitting*, Optimization and Engineering volume, 10 (2009), p. 1–17.

[19] J. Prince and A. Willsky, *Reconstructing convex sets from support line measurements*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 377–389.

[20] J. Prince and A. Willsky, *Convex set reconstruction using shape information*, CVGIP: Graphical Models and Image Processing, 53 (1991), pp. 413–427.

[21] M. Rudelson and R. Vershynin, *Hanson-Wright inequality and sub-gaussian concentration*, Electronic Communications in Probability, 18 (2013), pp. 1 – 9.

[22] E. Seijo and B. Sen, *Nonparametric least squares estimation of a multivariate convex regression function*, The Annals of Statistics, 39 (2011), pp. 1633 – 1657.

[23] Y. S. Soh, *Code for "Fitting tractable convex sets to support function evaluations"*. `https://github.com/yssoh/cvxreg`, 2019.

[24] Y. S. Soh and V. Chandrasekaran, *Fitting tractable convex sets to support function evaluations*, Discrete & Computational Geometry, (2021), pp. 1–42.

[25] G. B. A. G. C. Szepesvári, *Near-optimal max-affine estimators for convex regression*, Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, 38 (2015).

[26] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing, Cambridge University Press, 2012, pp. 210–268.

[27] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, 2020.

[28] T. W. Y. Yu and R. J. Samworth, *A useful variant of the Davis—Kahan theorem for statisticians*, Biometrika, 102 (2015), pp. 315–323.

A Lemmas for the Proof of Theorem 10

We first give a few definitions that are needed in following lemmas. A random vector $\xi \in \mathbb{R}^d$ is sub-Gaussian with parameter η if $\mathbb{E}[X] = 0$ and for each $u \in \mathbb{S}^{d-1}$, $\mathbb{E}\left[e^{\lambda\langle u, X\rangle}\right] \leq e^{\lambda^2\eta^2/2}, \quad$ for all $\lambda \in \mathbb{R}$. The sub-Gaussian norm of a random variable X, denoted by $\|X\|_{\psi_2}$, is defined as

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}.$$

For $\xi \in \mathbb{R}^d$, the sub-Gaussian norm is defined as $\|\xi\|_{\psi_2} := \sup_{u \in \mathbb{S}^{d-1}} \|\langle \xi, u\rangle\|_{\psi_2}$. The sub-exponential norm of X, denoted by $\|X\|_{\psi_1}$, is defined as

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\},$$

and the sub-exponential norm of a vector is defined similarly.

We also recall that the covering number of a Euclidean ball satisfies

$$\mathcal{N}(B_q(z, R), \|\cdot\|_2, \varepsilon) \leq (1 + 2R/\varepsilon)^q \tag{25}$$

for $\varepsilon \leq 2r$ by a standard volume argument.

The proofs rely on uniform spectral concentration bounds of a sample covariance matrix, which follow from Bernstein's inequality and Dudley's inequality. A general reference for the ideas in the lemmas below is [27].

Lemma 12. *Let ξ be an η-sub-Gaussian r.v. in $\mathbb{R}^d$ and $\mathcal{A}_* \in (\mathbb{S}^m)^d$ be such that*

$$\inf_{u \in \mathbb{S}^{d-1}} \lambda_1(\mathcal{A}_*[u]) - \lambda_2(\mathcal{A}_*[u]) := \kappa > 0.$$

Define the set $B(\mathcal{A}_*, \kappa/4) := \{\mathcal{A} \in (\mathbb{S}^m)^d : \|\mathcal{A} - \mathcal{A}_*\|_F \leq \frac{\kappa}{4}\}$. For each $\mathcal{A} \in (\mathbb{S}^m)^d$, define $U_{\mathcal{A}}$ to be the rank one matrix such that

$$\langle \xi \otimes U_{\mathcal{A}}, \mathcal{A}\rangle = \langle U_{\mathcal{A}}, \mathcal{A}\xi\rangle = \lambda_{\max}(\mathcal{A}\xi).$$

Then, $\|\xi \otimes U_{\mathcal{A}}\|_{\psi_2} \leq \eta m^2$ and for all $\mathcal{A}_1, \mathcal{A}_2 \in B(\mathcal{A}_*, r)$,

$$\|\xi \otimes U_{\mathcal{A}_1} - \xi \otimes U_{\mathcal{A}_2}\|_{\psi_2} \leq \frac{8m^2}{\kappa}\|\mathcal{A}_1 - \mathcal{A}_2\|_F.$$

Proof. For the first claim, recall that $\xi \otimes U_{\mathcal{A}}$ is sub-Gaussian if $\langle \xi \otimes U_{\mathcal{A}}, v\rangle$ is sub-Gaussian for every $v \in \mathbb{S}^{dm^2-1}$. Indeed, we first see that for $v \in \mathbb{S}^{dm^2-1}$,

$$|\langle \xi \otimes U, v\rangle| = \left|\sum_{i=1}^{d}\sum_{k=1}^{m^2} \xi_i U_k v_{ik}\right| \leq \sum_{k=1}^{m^2} |\langle \xi, v^{(k)}\rangle||U_k| \leq \sum_{k=1}^{m^2} |\langle \xi, v^{(k)}\rangle|,$$

where $v^{(k)} = (v_{1k}, \ldots, v_{dk}) \in \mathbb{R}^d$. Then, by the triangle inequality,

$$\|\langle \xi \otimes U, v\rangle\|_{\psi_2} \leq \left\|\sum_{k=1}^{m^2} |\langle \xi, v^{(k)}\rangle|\right\|_{\psi_2} \leq \sum_{k=1}^{m^2} \left\||\langle \xi, v^{(k)}\rangle|\right\|_{\psi_2} \leq \eta \sum_{k=1}^{m^2} \|v^{(k)}\|_2 \leq \eta m^2.$$

For the second claim, first note that for all $\mathcal{A} \in B(\mathcal{A}_*, \kappa/4)$, Weyl's inequality implies

$$\lambda_1(\mathcal{A}[u]) - \lambda_2(\mathcal{A}[u]) \geq \lambda_1(\mathcal{A}_*[u]) - \lambda_2(\mathcal{A}_*[u]) - 2\|\mathcal{A} - \mathcal{A}_*\|_{op} \geq \frac{\kappa}{2} > 0.$$

Then, observe that $\|U_1 - U_2\|_F^2 \leq 2\|u_1 - u_2\|_2^2$, where $U_1 = u_1 u_1^T$, $U_2 = u_2 u_2^T$ and $u_1, u_2 \in \mathbb{S}^{m-1}$. By a variation of the Davis-Kahan Theorem (Corollary 3 in [28]),

$$\|U_1 - U_2\|_F \leq \sqrt{2}\|u_1 - u_2\|_2 \leq \frac{4\|(\mathcal{A}_1 - \mathcal{A}_2)[\xi]\|_{op}}{\lambda_1(\mathcal{A}_2[\xi]) - \lambda_2(\mathcal{A}_2[\xi])} \leq \frac{8}{\kappa}\|\mathcal{A}_1 - \mathcal{A}_2\|_F.$$

This implies that

$$\|\xi \otimes (U_1 - U_2)\|_{\psi_2} \leq \frac{8\|\mathcal{A}_1 - \mathcal{A}_2\|_F}{\kappa}\left\|\xi \otimes \frac{U_1 - U_2}{\|U_1 - U_2\|_F}\right\|_{\psi_2} \leq \frac{8\eta m^2}{\kappa}\|\mathcal{A}_1 - \mathcal{A}_2\|_F,$$

□

Lemma 13. *Define $B_q(z, R) := \{x \in \mathbb{R}^q : \|x - z\|_2 \leq R\}$ for $R > 0$ and $z \in \mathbb{R}^q$. Let $\{\xi_a\}_{a \in B_q(z,R)}$ be stochastic process in $\mathbb{R}^d$ such that* *(i)* $\|\xi_a\|_{\psi_2} \le \eta$;

(ii) for all $a_1,\, a_2 \in B_q(z,R)$, $\|\xi_{a_1} - \xi_{a_2}\|_{\psi_2} \le K\|a_1 - a_2\|_2$.

Define $\Xi_a \in \mathbb{R}^{N\times d}$ *to be the matrix with* N *i.i.d. rows in* $\mathbb{R}^d$ *distributed as* ξ_a*, and let* $\Sigma_a := \mathbb{E}[\xi_a\xi_a^T]$*. Fix* $\tau \in (0,1)$*. Then, there exist absolute constants* $c_0, c_1, c_2 > 0$ *such that if* $n \ge c_0\tau^{-2}K^2\eta^4R^2\max\{q,d\}$*,*

$$\mathbb{P}\left(\sup_{a\in B_q(z,R)}\left\|\frac{1}{n}\Xi_a^T\Xi_a - \Sigma_a\right\|_{op} \ge \tau\|\Sigma_a\|_{op}\right) \le e^{-c_1 n\tau^2/K^2\eta^4R^2}.$$

This implies that with probability greater than $1 - e^{-c_1 n\tau^2/K^2\eta^4R^2}$*,*

$$\lambda_{\max}(\Sigma_a)(1-\tau) \le \inf_{a\in B_q(z,R)}\frac{\lambda_{\min}\left(\Xi_a^T\Xi_a\right)}{n} \le \sup_{a\in B_q(z,R)}\frac{\lambda_{\max}\left(\Xi_a^T\Xi_a\right)}{n} \le \lambda_{\max}(\Sigma_a)(1+\tau).$$

Proof. First suppose that for all a, $\Sigma_a = I$, i.e. that is ξ_a is isotropic. For the general case, since $\Sigma_a^{-1/2}\xi_a$ is isotropic, the conclusion follows from the fact that $\left\|\frac{1}{n}\Xi_a^T\Xi_a - \Sigma_a\right\|_{op} \le \|\Sigma_a\|_{op}\left\|\frac{1}{n}\Sigma_a^{-1}\Xi_a^T\Xi_a - I\right\|_{op}$.

We first show that for any $x \in \mathbb{S}^{d-1}$, the stochastic process $X_a := \frac{1}{\sqrt{n}}\|\Xi_a x\| - 1$ has sub-Gaussian increments $\|X_{a_1} - X_{a_2}\|_{\psi_2} = \frac{1}{\sqrt{n}}\left\|\|\Xi_{a_1}x\|_2 - \|\Xi_{a_2}x\|_2\right\|_{\psi_2}$.

Case 1: $s \in [0, 4K\sqrt{n}]$. We first see that

$$\begin{aligned}
&\mathbb{P}\left(\left|\|\Xi_{a_1}x\|_2 - \|\Xi_{a_2}x\|_2\right| \ge s\|a_1 - a_2\|_2\right)\\
&\qquad = \mathbb{P}\left(\frac{\left|\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2\right|}{\|a_1 - a_2\|} \ge s(\|\Xi_{a_1}x\|_2 + \|\Xi_{a_2}x\|_2)\right)\\
&\qquad \le \mathbb{P}\left(\frac{\left|\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2\right|}{\|a_1 - a_2\|} \ge s\|\Xi_{a_1}x\|_2\right)\\
&\qquad \le \mathbb{P}\left(\frac{\left|\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2\right|}{\|a_1 - a_2\|} \ge \frac{s\sqrt{n}}{2}\right) + \mathbb{P}\left(\|\Xi_{a_1}x\|_2 \le \frac{\sqrt{n}}{2}\right)\\
&\qquad \le \mathbb{P}\left(\frac{\left|\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2\right|}{\|a_1 - a_2\|} \ge \frac{s\sqrt{n}}{2}\right) + \mathbb{P}\left(\left|\|\Xi_{a_1}x\|_2 - \sqrt{n}\right| \ge \frac{s}{8K}\right). \qquad (26)
\end{aligned}$$

Then, note that

$$\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2 = \sum_{i=1}^n \langle \xi_{a_1}^{(i)} - \xi_{a_2}^{(i)}, x\rangle\langle \xi_{a_1}^{(i)} + \xi_{a_2}^{(i)}, x\rangle,$$

and by Lemma 2.7.7 in [27],

$$\begin{aligned}
\|\langle \xi_{a_1}^{(i)} - \xi_{a_2}^{(i)}, x\rangle\langle \xi_{a_1}^{(i)} + \xi_{a_2}^{(i)}, x\rangle\|_{\psi_1} &\le \|\langle \xi_{a_1}^{(i)} - \xi_{a_2}^{(i)}, x\rangle\|_{\psi_2}\|\langle \xi_{a_1}^{(i)} + \xi_{a_2}^{(i)}, x\rangle\|_{\psi_2}\\
&\le 2\eta K\|a_1 - a_2\|_2.
\end{aligned}$$

Each term in the sum also has zero mean. Indeed,

$$\mathbb{E}[\langle \xi_{a_1}^{(i)} - \xi_{a_2}^{(i)}, x\rangle\langle \xi_{a_1}^{(i)} + \xi_{a_2}^{(i)}, x\rangle] = \mathbb{E}[\langle \xi_{a_1}^{(i)}, x\rangle^2 - \langle \xi_{a_2}^{(i)}, x\rangle^2] = 0.$$

Applying Bernstein's inequality (Corollary 2.8.3 in [27]) gives for all $t \ge 0$,

$$\mathbb{P}\left(\frac{\|\Xi_{a_1}x\|_2^2 - \|\Xi_{a_2}x\|_2^2}{\|a_1 - a_2\|} \ge t\right) \le 2e^{-c_1\min\left\{\frac{t^2}{4\eta^2K^2n}, \frac{t}{2\eta K}\right\}}.$$

For the second tail probability in (26), Theorem 3.1.1 in [27] implies

$$\mathbb{P}\left(\left|\|\Xi_{a_1}x\|_2 - \sqrt{n}\right| \ge t\right) \le 2e^{-\frac{c_2t^2}{\eta^4}},$$

where we have used that ξ_a is isotropic. Thus, since $s < 4K\sqrt{n}$ and $\eta \ge 1$,

$$\mathbb{P}\left(\frac{\left|\|\Xi_{a_1}x\|_2 - \|\Xi_{a_2}x\|_2\right|}{\|a_1 - a_2\|_2} \ge s\right) \le 2e^{-c_1\min\{\frac{s^2}{16\eta^2K^2}, \frac{s\sqrt{n}}{4\eta K}\}} + 2e^{-\frac{c_2s^2}{64\eta^4K^2}} \le 4e^{-\frac{c_3s^2}{\eta^4K^2}}.$$

Case 2: $s \geq 4\eta K\sqrt{n}$. By the triangle inequality,

$$\begin{aligned}
&\mathbb{P}\left(\frac{|\|\Xi_{a_1}x\|_2 - \|\Xi_{a_2}x\|_2|}{\|a_1 - a_2\|} \geq s\right) \leq \mathbb{P}\left(\frac{\|(\Xi_{a_1} - \Xi_{a_2})x\|^2}{\|a_1 - a_2\|^2} \geq s^2\right) \\
&= \mathbb{P}\left(\frac{\|(\Xi_{a_1} - \Xi_{a_2})x\|^2}{\|a_1 - a_2\|^2} - n\frac{\mathbb{E}[\langle \xi_{a_1} - \xi_{a_2}, x\rangle^2]}{\|a_1 - a_2\}^2} \geq s^2 - n\frac{\mathbb{E}[\langle \xi_{a_1} - \xi_{a_2}, x\rangle^2]}{\|a_1 - a_2\|^2}\right) \\
&\leq \mathbb{P}\left(\frac{\|(\Xi_{a_1} - \Xi_{a_2})x\|^2}{\|a_1 - a_2\|^2} - n\frac{\mathbb{E}[\langle \xi_{a_1} - \xi_{a_2}, x\rangle^2]}{\|a_1 - a_2\|^2} \geq s^2 - 4K^2 n\right) \\
&\leq \mathbb{P}\left(\left|\frac{\|(\Xi_{a_1} - \Xi_{a_2})x\|^2}{\|a_1 - a_2\|^2} - n\frac{\mathbb{E}[\langle \xi_{a_1} - \xi_{a_2}, x\rangle^2]}{\|a_1 - a_2\|^2}\right| \geq \frac{3s^2}{4}\right).
\end{aligned}$$

where for the second to last inequality we have used that

$$\mathbb{E}[\langle \xi_{a_1} - \xi_{a_2}, x\rangle^2] \leq 4\|\xi_{a_1} - \xi_{a_2}\|_{\psi_2}^2 \leq 4K^2\|a_1 - a_2\|_2^2,$$

and the last inequality follows from the lower bound on t and the fact that $\eta \geq 1$. By Bernstein's inequality again (Corollary 2.8.3 in [27]) and the lower bound on t,

$$\mathbb{P}\left(\frac{|\|\Xi_{a_1}x\|_2 - \|\Xi_{a_2}x\|_2|}{\|a_1 - a_2\|} \geq s\right) \leq 2e^{-c_4 \min\left\{\frac{s^4}{nK^4}, \frac{s^2}{K^2}\right\}} \leq 2e^{-\frac{c_4 s^2}{K^2}}$$

Proposition 2.5.2 in [27] then implies that for all $t \geq 0$,

$$\|X_{a_1} - X_{a_2}\|_{\psi_2} \leq \frac{K\eta^2}{\sqrt{n}}\|a_1 - a_2\|_2.$$

Now, by Theorem 8.1.6 in [27] and (25), for $\delta > 0$ and $n \geq q\delta^{-2}$,

$$\sup_{a \in B_q(z,R)} \left|\frac{1}{\sqrt{n}}\|\Xi_a x\| - 1\right| \leq \frac{c_5 K\eta^2}{\sqrt{n}}\left(c_6 R\sqrt{q} + 2R\delta\sqrt{n}\right) \leq c_7 K\eta^2 R\delta, \tag{27}$$

with probability greater than $1 - 2e^{-\delta^2 n}$. Now, let $\tau \in (0,1)$. By the inequality $|z^2 - 1| \leq 3\max\{|z-1|, |z-1|^2\}$ for all $z \geq 0$,

$$\mathbb{P}\left(\sup_{a \in B_q(z,R)} \left|\frac{1}{n}\|\Xi_a x\|_2^2 - 1\right| \geq \frac{\tau}{2}\right) \leq \mathbb{P}\left(\sup_{a \in B_q(z,R)} \left|\frac{1}{\sqrt{n}}\|\Xi_a x\|_2 - 1\right| \geq \frac{\tau}{6}\right).$$

Letting $\delta = \frac{\tau}{6c_7 K\eta^2 R}$ in (27) gives the following. For $n \geq c_8 qK^2\eta^4 R^2\tau^{-2}$,

$$\mathbb{P}\left(\sup_{a \in B_q(z,R)} \left|\frac{1}{n}\|\Xi_a x\|_2^2 - 1\right| \geq \frac{\tau}{2}\right) \leq 2e^{-n\tau^2/c_8 K^2\eta^4 R^2}.$$

Finally, by Lemma 5.3 in [26],

$$\sup_{a \in B_q(r)} \|\frac{1}{n}\Xi_a^T\Xi_a - I\|_{op} \leq 2\max_{x \in \mathcal{N}} \sup_{a \in B_q(r)} \left|\frac{1}{n}\|\Xi_a x\|_2^2 - 1\right|,$$

where $\mathcal{N}$ is a $\frac{1}{4}$-net of the unit sphere $\mathbb{S}^{d-1}$. Lemma 5.4 in [26] implies $|\mathcal{N}| \leq 9^d$. Applying the union bound then gives, for $n \geq c_8 qK^2\eta^4 R^2\tau^{-2}$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{a \in B_q(z,R)} \|\frac{1}{n}\Xi_a^T\Xi_a - I\|_{op} \geq \tau\right) \leq \mathbb{P}\left(\max_{x \in \mathcal{N}} \sup_{a \in B_q(r)} \left|\frac{1}{n}\|\Xi_a x\|_2^2 - \|x\|_2^2\right| \geq \frac{\tau}{2}\right) \\
&\leq |\mathcal{N}|\mathbb{P}\left(\sup_{a \in B_q(z,R)} \left|\frac{1}{n}\|\Xi_a x\|_2^2 - 1\right| \geq \frac{\tau}{2}\right) \leq 2 \cdot 9^d e^{-n\tau^2/c_8 K^2\eta^4 R^2},
\end{aligned}$$

Thus, there exist absolute constants b_1, b_2 such that for $n \geq b_1\tau^{-2}K^2\eta^4 R^2\max\{q, d\}$,

$$\mathbb{P}\left(\sup_{a \in B_q(z,R)} \left\|\frac{1}{n}\Xi_a^T\Xi_a - I\right\|_{op} \geq \tau\right) \leq 2 \cdot e^{-b_2 n\tau^2/K^2\eta^4 R^2}.$$

□

Lemma 14. *Consider the setting of Theorem 10. Let $B(\mathcal{A}_*, \kappa/4) := \{\mathcal{A} \in (\mathbb{S}^m)^d : \|\mathcal{A} - \mathcal{A}_*\|_F^2 \leq \kappa/4\}$ and define $\mathcal{P} := \{P_{\Xi_{\mathcal{A}}} : \mathcal{A} \in \mathcal{B}(\mathcal{A}_*, \kappa/4)\}$. Then, there exist absolute constants c_0, c_1, and c_2 such that for $n \geq c_0$,*

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2 \geq c_1 \log(n)^2 \sigma^2 m^2 (d+1)\right) \leq \exp\left\{-c_2(d+1)m^2\log(n)\right\}.$$

Proof. First, note that for all $P \in \mathcal{P}$, $\|P\|_F^2 = (d+1)m^2$ and

$$\mathbb{E}[\|P\varepsilon\|_2^2] = \sum \mathbb{E}[\langle v_i, \varepsilon\rangle^2] = \sum \|v_i\|_2^2 \sigma^2 = (d+1)m^2\sigma^2. \tag{28}$$

Then,

$$\sup_{P\in\mathcal{P}} \left(\|P\varepsilon\|^2 - \mathbb{E}[\|P\varepsilon\|^2]\right) - \mathbb{E}\sup_{P\in\mathcal{P}} \left(\|P\varepsilon\|^2 - \mathbb{E}[\|P\varepsilon\|^2]\right) = \sup_{P\in\mathcal{P}} \|P\varepsilon\|^2 - \mathbb{E}\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2.$$

Now, recall that $M := \|\max_{i=1,\ldots,n} \varepsilon_i\|_{\psi_2} \leq c_0\sigma\sqrt{\log n}$ for an absolute constant c_0 [17]. Applying Theorem 1.1 in [14] to the family of matrices $\{P^TP : P \in \mathcal{P}\}$ gives: For $t \geq \max\{c_0\sigma\sqrt{\log(n)}\mathbb{E}\left[\sup_{P\in\mathcal{P}} \|P\varepsilon\|_2\right], c_0^2\sigma^2\log(n)\}$,

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2 - \mathbb{E}\left[\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2\right] \geq t\right) \leq e^{-\frac{c_2}{\sigma^2\log(n)}\min\left\{\frac{t^2}{\mathbb{E}\left[\sup_{P\in\mathcal{P}}\|P\varepsilon\|\right]^2}, t\right\}}. \tag{29}$$

Also by (28), $\mathbb{E}\left[\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2\right] \geq (d+1)m^2\sigma^2$ and thus,

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2 - (d+1)m^2\sigma_\varepsilon^2 \geq t\right) \leq e^{-\frac{c_2}{\sigma^2\log(n)}\min\left\{\frac{t^2}{\mathbb{E}\left[\sup_{P\in\mathcal{P}}\|P\varepsilon\|\right]^2}, t\right\}}.$$

Letting $t = c_3\sigma^2\log(n)^2(d+1)m^2$ for a constant $c_3 > 0$ large enough,

$$\begin{aligned}\mathbb{P}\left(\sup_{P\in\mathcal{P}} \|P\varepsilon\|^2 \geq (c_3\log(n)^2 + 1)\sigma^2 m^2 (d+1)\right)& \\ \leq e^{-c_4(d+1)m^2\log(n)\min\left\{\frac{\log(n)^2(d+1)m^2\sigma^2}{\mathbb{E}\left[\sup_{P\in\mathcal{P}}\|P\varepsilon\|\right]^2}, 1\right\}}&.\end{aligned} \tag{30}$$

We now upper bound $\mathbb{E}\left[\sup_{P\in\mathcal{P}} \|P\varepsilon\|\right]$. For $P = P_{\Xi_{\mathcal{A}}} \in \mathcal{P}$,

$$\|P\varepsilon\|_2 = \|\Xi_{\mathcal{A}}(\Xi_{\mathcal{A}}^T\Xi_{\mathcal{A}})^{-1}\Xi_{\mathcal{A}}^T\varepsilon\|_2 \leq \frac{\|\Xi_{\mathcal{A}}\|_2}{\|\Xi_{\mathcal{A}}^T\Xi_{\mathcal{A}}\|_2}\|\Xi_{\mathcal{A}}^T\varepsilon\|_2 = \frac{\|\Xi_{\mathcal{A}}^T\varepsilon\|_2}{\|\Xi_{\mathcal{A}}^T\|_2} = \frac{\|\Xi_{\mathcal{A}}^T\varepsilon\|_2}{\sqrt{\sum_{i=1}^n \|\xi^{(i)}\|_2^2}}.$$

Define the stochastic process $X_{\mathcal{A}} := \frac{\|\Xi_{\mathcal{A}}^T\varepsilon\|_2}{\sum_{i=1}^n \|\xi^{(i)}\|_2}$. Then, for $\mathcal{A}$ and $\mathcal{B}$ in $(\mathbb{S}^m)^d$,

$$\begin{aligned}\|\Xi_{\mathcal{A}} - \Xi_{\mathcal{B}}\|_F^2 &= \sum_{i=1}^n\sum_{j=1}^d\sum_{k=1}^{m^2} |\xi_j^{(i)}U_k^{(i)} - \xi_j^{(i)}V_k^{(i)}|^2 = \sum_{i=1}^n \|\xi^{(i)}\|_2^2\left[2 - 2\langle U^{(i)}, V^{(i)}\rangle\right] \\ &= 2\sum_{i=1}^n \|\xi^{(i)}\|_2^2\left[1 - \langle u^{(i)}, v^{(i)}\rangle^2\right] = 2\sum_{i=1}^n \|\xi^{(i)}\|_2^2 \sin\Theta(u^{(i)}, v^{(i)})^2.\end{aligned}$$

By a variant of the Davis-Kahan Theorem (Corollary 3 in [28]) and (13),

$$\sin\Theta(u^{(i)}, v^{(i)}) \leq \frac{2\|\mathcal{A}[\xi^{(i)}] - \mathcal{B}[\xi^{(i)}]\|_{op}}{\lambda_1(\mathcal{A}[\xi^{(i)}]) - \lambda_2(\mathcal{A}[\xi^{(i)}])} \leq \frac{4}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}.$$

Then, $\frac{\|\Xi_{\mathcal{A}} - \Xi_{\mathcal{B}}\|_F}{\sqrt{\sum_{i=1}^n \|\xi^{(i)}\|_2^2}} \leq \frac{8}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}$, and by the Hanson-Wright Inequality [21, Theorem 2.1], since ε is independent of the $\xi^{(i)}$'s, there is a constant c_4 such that

$$\begin{aligned}\mathbb{P}\left(|X_{\mathcal{A}} - X_{\mathcal{B}}| \geq t + \frac{8\sigma}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}\right) &\leq \mathbb{P}\left(\frac{\|(\Xi_{\mathcal{A}}^T - \Xi_{\mathcal{B}}^T)\varepsilon\|_2}{\sqrt{\sum_{i=1}^n \|\xi^{(i)}\|_2^2}} \geq t + \frac{8\sigma}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}\right) \\ &\leq \mathbb{P}\left(\left|\frac{\|(\Xi_{\mathcal{A}}^T - \Xi_{\mathcal{B}}^T)\varepsilon\|_2}{\sqrt{\sum_{i=1}^n \|\xi^{(i)}\|_2^2}} - \frac{\sigma\|\Xi_{\mathcal{A}} - \Xi_{\mathcal{B}}\|_F}{\sqrt{\sum_{i=1}^n \|\xi^{(i)}\|_2^2}}\right| \geq t\right) \leq 2\exp\left\{-\frac{c_4\kappa^2t^2}{64\sigma^4\|\mathcal{A} - \mathcal{B}\|_{op}^2}\right\}.\end{aligned}$$

Thus, $\{X_{\mathcal{A}}\}_{\mathcal{A}}$ has sub-gaussian increments and there is a constant c_5 such that

$$\|X_{\mathcal{A}} - X_{\mathcal{B}}\|_{\psi_2} \leq \|X_{\mathcal{A}} - X_{\mathcal{B}} - \frac{8\sigma}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}\|_{\psi_2} + \frac{8\sigma}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op} \leq \frac{c_5\sigma}{\kappa}\|\mathcal{A} - \mathcal{B}\|_{op}.$$

Then, by Theorem 8.1.6 in [27] and (25),

$$\mathbb{E}\left[\sup_{P\in\mathcal{P}}\|P\varepsilon\|\right] \le \mathbb{E}\left[|X_{\mathcal{A}}|\right] + \mathbb{E}\left[\sup_{\mathcal{A}\in B(\mathcal{A}_*,\kappa/4)}|X_{\mathcal{A}}| - \mathbb{E}[|X_{\mathcal{A}}|]\right] \le c_8 m\sigma\sqrt{d+1}.$$

Combining this bound with (30) gives

$$\mathbb{P}\left(\sup_{P\in\mathcal{P}}\|P\varepsilon\|^2 \ge (c_3\log(n)^2+1)m^2(d+1)\right) \le e^{-c_9(d+1)\sigma^2 m^2\log(n)\min\left(\frac{\log(n)^2}{c_8^2},1\right)}.$$

Taking $n \ge e^{-c_8^2}$ completes the proof. □