Submitted to Bernoulli

On the risk of convex-constrained least squares estimators under misspecification

BILLY FANG[1,*] and ADITYANAND GUNTUBOYINA[1,**]

[1]*Dept. of Statistics, University of California, Berkeley CA 94720*
E-mail: [*]blfang@berkeley.edu; [**]aditya@stat.berkeley.edu

We consider the problem of estimating the mean of a noisy vector. When the mean lies in a convex constraint set, the least squares projection of the random vector onto the set is a natural estimator. Properties of the risk of this estimator, such as its asymptotic behavior as the noise tends to zero, have been well studied. We instead study the behavior of this estimator under misspecification, that is, without the assumption that the mean lies in the constraint set. For appropriately defined notions of risk in the misspecified setting, we prove a generalization of a low noise characterization of the risk due to Oymak and Hassibi [8] in the case of a polyhedral constraint set. An interesting consequence of our results is that the risk can be much smaller in the misspecified setting than in the well-specified setting. We also discuss consequences of our result for isotonic regression.

Keywords: convex constraint, isotonic regression, least squares, misspecification, statistical dimension.

1. Introduction

In many statistical problems, it is common to model the observations $y_1, \ldots, y_n \in \mathbb{R}$ as $y_i = \theta_i^* + \sigma z_i$ where $\theta_1^*, \ldots, \theta_n^*$ are unknown parameters of interest, $z_1, \ldots, z_n$ represent noise or error variables that have mean zero, and $\sigma > 0$ denotes a scale parameter. In vector notation, this is equivalent to writing

$$Y = \theta^* + \sigma Z,$$

where $Y := (y_1, \ldots, y_n)$, $\theta^* := (\theta_1^*, \ldots, \theta_n^*)$, and $Z := (z_1, \ldots, z_n)$. A common instance of this model is the Gaussian sequence model, where the $z_1, \ldots, z_n$ are independent standard Gaussian random variables, in which case the model can be written as $Y \sim N(\theta^*, \sigma^2 I_n)$, where I_n is the $n \times n$ identity matrix.

A standard method of estimating θ^* from the observation vector Y is to fix a closed convex set $\mathcal{C}$ of $\mathbb{R}^n$ and use the least squares estimator under the constraint given by $\theta \in \mathcal{C}$. Specifically, the least squares projection is

$$\Pi_{\mathcal{C}}(x) := \operatorname*{argmin}_{\theta \in \mathcal{C}} \|x - \theta\|^2,$$

(where $\|\cdot\|$ denotes the standard Euclidean norm in $\mathbb{R}^n$), and one estimates θ^* by

$$\hat{\theta}(Y) := \Pi_{\mathcal{C}}(Y).$$

When $\mathcal{C}$ is taken to be $\{X\beta : \|\beta\|_1 \leq R\}$ for some deterministic $n \times p$ matrix X and $R > 0$, this estimator becomes LASSO in the constrained form as originally proposed by Tibshirani [11]. When $\mathcal{C}$ is taken to be $\{X\beta : \min_j \beta_j \geq 0\}$, this estimator becomes nonnegative least squares. Note that shape restricted regression estimators are special cases of nonnegative least squares for appropriate choices of X (see, for example, Groeneboom and Jongbloed [4]). Also, note that both sets $\{X\beta : \|\beta\|_1 \leq R\}$ and $\{X\beta : \min_j \beta_j \geq 0\}$ are examples of polyhedral sets. Therefore in most applications, the constraint set $\mathcal{C}$ is polyhedral.

There exist many results in the literature studying the accuracy of $\hat{\theta}(Y)$ as an estimator for θ^*. Most of these results make the assumption that $\theta^* \in \mathcal{C}$. In this paper, we shall refer to this assumption as the *well-specified* assumption. Essentially, the constraint set $\mathcal{C}$ can be taken to be a part of the model specification, and the assumption $\theta^* \in \mathcal{C}$ means that the true mean vector θ^* satisfies the model assumptions, i.e. the model is well-specified.

Under the well-specified assumption, it is reasonable and common to measure the accuracy of $\hat{\theta}(Y)$ via its risk under squared Euclidean distance. More precisely, the risk of $\hat{\theta}(Y)$ is defined by

$$R(\hat{\theta}, \theta^*) := \mathbb{E}_{\theta^*} \|\hat{\theta}(Y) - \theta^*\|^2$$

where $\mathbb{E}_{\theta^*}$ refers to expectation taken with respect to the noise Z in the model $Y = \theta^* + \sigma Z$.

Many results on $R(\hat{\theta}, \theta^*)$ in the well-specified setting are available in the literature. Of all the available results, let us isolate two results from Oymak and Hassibi [8] because of their generality. In the setting where $Z \sim N(0, I_n)$, Oymak and Hassibi [8] first proved the upper bound

$$\frac{1}{\sigma^2} R(\hat{\theta}, \theta^*) \leq \delta(T_{\mathcal{C}}(\theta^*)), \tag{1}$$

where $T_{\mathcal{C}}(\theta^*)$ denotes the *tangent cone* of $\mathcal{C}$ at θ^*, defined by

$$T_{\mathcal{C}}(\theta^*) = \mathrm{cl}\left\{\alpha(\theta - \theta^*) : \alpha \geq 0, \theta \in \mathcal{C}\right\}, \tag{2}$$

("cl" denotes closure), and where $\delta\left(T_{\mathcal{C}}(\theta^*)\right)$ denotes the *statistical dimension* of the cone $T_{\mathcal{C}}(\theta^*)$. In general, the statistical dimension of a closed cone $T \subseteq \mathbb{R}^n$ is defined as

$$\delta(T) := \mathbb{E}\|\Pi_T(Z)\|^2, \tag{3}$$

where the expectation is with respect to $Z \sim N(0, I_n)$. Many properties of the statistical dimension are covered by Amelunxen et al. [2].

In the case when the constraint set $\mathcal{C}$ is a subspace, the estimator $\hat{\theta}(Y)$ is linear and, in this case, it is easy to see that $\delta(T_{\mathcal{C}}(\theta^*))$ is simply the dimension of $\mathcal{C}$, so that inequality (1) becomes an equality. For general closed convex sets, it is therefore reasonable to ask how tight inequality (1) is. It is not hard to construct examples of $\mathcal{C}$ and $\theta^* \in \mathcal{C}$ where inequality (1) is loose for fixed $\sigma > 0$. However Oymak and Hassibi [8] proved remarkably that the upper bound in (1) is tight in the limit as $\sigma \downarrow 0$ (we shall refer to this in the sequel as the *low σ limit*); that is, when $Z \sim N(0, I_n)$,

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} R(\hat{\theta}, \theta^*) = \delta(T_{\mathcal{C}}(\theta^*)). \tag{4}$$

In summary, Oymak and Hassibi [8] proved that $\sigma^2 \delta(T_{\mathcal{C}}(\theta^*))$ is a nice formula for the risk of $\hat{\theta}(Y)$ that is, in general, an upper bound which is tight in the low σ limit.

We remark that although Oymak and Hassibi [8] state the results (1) and (4) for the specific case $Z \sim N(0, I_n)$, their proof automatically extends to the more general setting where Z is an arbitrary zero mean random vector with $\mathbb{E}\|Z\|^2 < \infty$ (the components $Z_1, \ldots, Z_n$ of Z can be arbitrarily dependent), provided we generalize the definition (3) of statistical dimension by taking the expectation with respect to Z, without assuming Z is standard Gaussian. We refer to this modification of the definition (3) as the *generalized statistical dimension* of the cone T. As a slight abuse of notation, we use the same notation $\delta(\cdot)$ for this more general concept, with the understanding that the expectation in the definition is with respect to the distribution of Z. By dropping the Gaussian assumption, the generalized statistical dimension loses much of the interpretability and nice geometric properties of the usual statistical dimension [2], but still serves as an abstract notion of the size of a cone T with respect to a distribution Z.

This paper deals with the behavior of the estimator $\hat{\theta}(Y)$ when the assumption $\theta^* \in \mathcal{C}$ is violated. We shall refer to the situation when $\theta^* \notin \mathcal{C}$ as the *misspecified* setting. Note that, in practice, one can never know if the unknown θ^* truly lies in $\mathcal{C}$. It is therefore necessary to study the behavior of $\hat{\theta}(Y)$ under misspecification.

For the misspecified setting, one must first note that it is no longer reasonable to measure the performance of $\hat{\theta}(Y)$ by the risk $R(\hat{\theta}, \theta^*)$, simply because $\hat{\theta}(Y)$ is constrained to be in $\mathcal{C}$ and hence cannot be expected to be close to θ^* which is essentially unconstrained. There are two natural notions of accuracy of $\hat{\theta}(Y)$ in the misspecified setting, which we call the *misspecified risk* and the *excess risk*. The misspecified risk is defined as

$$M(\hat{\theta}, \theta^*) := \mathbb{E}_{\theta^*}\|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2, \tag{5}$$

and the excess risk is defined as

$$E(\hat{\theta}, \theta^*) := \mathbb{E}_{\theta^*}\|\hat{\theta}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2. \tag{6}$$

The misspecified risk, $M(\hat{\theta}, \theta^*)$, is motivated by the observation that, in the misspecifed case, the estimator $\hat{\theta}(Y)$ is really estimating $\Pi_{\mathcal{C}}(\theta^*)$ so it is natural to measure its squared distance from $\Pi_{\mathcal{C}}(\theta^*)$. On the other

hand, the excess risk, $E(\hat{\theta}, \theta^*)$, measures the squared distance of the estimator from θ^* relative to the squared distance of $\Pi_{\mathcal{C}}(\theta^*)$ from θ^*. We refer the reader to Bellec [3] and Section 2 for some background and basic properties on these notions of accuracy under misspecification. For example, it can be shown that $M(\hat{\theta}, \theta^*)$ is always less than or equal to $E(\hat{\theta}, \theta^*)$ (see (12)). It is easy to see that both of these risk measures equal $R(\hat{\theta}, \theta^*)$ in the well-specifed case i.e.,

$$R(\hat{\theta}, \theta^*) = M(\hat{\theta}, \theta^*) = E(\hat{\theta}, \theta^*), \qquad \text{when } \theta^* \in \mathcal{C}.$$

An analogue to inequality (1) for the case of misspecification has been proved by Bellec [3, Corollary 2.2], who showed that

$$\frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) \le \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) \le \delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))). \tag{7}$$

Again, although this was originally stated for $Z \sim N(0, I_n)$, it holds for arbitrary zero mean random vectors Z with $\mathbb{E}\|Z\|^2 < \infty$. Note the similarity between the right-hand sides of the inequalities (1) and (7). The only difference is that the tangent cone at θ^* is replaced by the tangent cone at $\Pi_{\mathcal{C}}(\theta^*)$ in the case of misspecification. Moreover, in the well-specified setting, the above inequality (7) reduces to (1).

It is now very natural to ask if the second inequality in (7) is tight in the low σ limit. One might guess that this should be the case given the result (4) for the well-specified setting. However, it turns out that (7) is not sharp in the low σ limit. The main contribution of this paper is to provide an exact formula for the low σ limit of $M(\hat{\theta}, \theta^*)$ and $E(\hat{\theta}, \theta^*)$ when $\mathcal{C}$ is polyhedral. Specifically, in Theorem 3.1, we prove that if the noise Z is zero mean with $\mathbb{E}\|Z\|^2 < \infty$ and if $\mathcal{C}$ is polyhedral, then

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) = \delta\big(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}\big), \tag{8}$$

where $v^{\perp} := \{u \in \mathbb{R}^n : \langle u, v \rangle = 0\}$ for vectors $v \in \mathbb{R}^n$. As we remarked earlier, in most applications, the constraint set $\mathcal{C}$ is polyhedral.

Because the set $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ is a subset of $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$, the right hand side of (8) is never larger than $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$. Under the assumption that the polyhedron $\mathcal{C}$ has a *nonempty interior*, along with a mild condition on the noise Z, it can be proved that the right hand side of (8) is *strictly* smaller than $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$ when $\theta^* \notin \mathcal{C}$ (an even stronger statement is proved in Lemma 3.4), which then implies that $\lim_{\sigma \downarrow 0} \sigma^{-2} M(\hat{\theta}, \theta^*) < \lim_{\sigma \downarrow 0} \sigma^{-2} R(\hat{\theta}, \Pi_{\mathcal{C}}(\theta^*))$. This inequality is more interpretable in the following form:

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\theta^*} \|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 < \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\Pi_{\mathcal{C}}(\theta^*)} \|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \qquad \text{whenever } \theta^* \notin \mathcal{C}. \tag{9}$$

(Note that the nonempty interior assumption is essential; for instance when $\mathcal{C}$ is a proper subspace of $\mathbb{R}^n$, the inequality (9) becomes equality.) Inequality (9) can be qualitatively understood as follows. The left hand side above corresponds to misspecification where the data are generated from $\theta^* \notin \mathcal{C}$ while the right hand side corresponds to the well-specified setting where the data are generated from $\Pi_{\mathcal{C}}(\theta^*)$. Note that in both cases, the estimator $\hat{\theta}(Y)$ is really estimating $\Pi_{\mathcal{C}}(\theta^*)$ so it is natural to compare the squared expected distance to $\Pi_{\mathcal{C}}(\theta^*)$ in both situations. The interesting aspect is that (in the low σ limit) the expected squared distance is smaller in the misspecified setting compared to the well-specified setting. To the best of our knowledge, this fact has not been noted in the literature previously at this level of generality, although it has been noticed in certain instances of specific problems such as isotonic density estimation and regression (see, for example, Jankowski [6]).

Our main result, Theorem 3.1, is stated and proved in Section 3 where some intuition is also provided for the exact form of the low σ limit in misspecification. The low σ limit can be explicitly computed in certain specific situations. In Section 4, we specialize to the Gaussian model $Z \sim N(0, I_n)$ and study in detail the examples when $\mathcal{C}$ is the nonnegative orthant and when $\mathcal{C}$ is the monotone cone (this latter case corresponds to isotonic regression).

In Section 5, we explore issues naturally related to Theorem 3.1. In Section 5.1, we consider the situation when $\mathcal{C}$ is not polyhedral. It seems hard to characterize the low σ misspecification limits in this case but it is possible to compute them when $\mathcal{C}$ is the unit ball. It is interesting to note that the low σ limits of $M(\hat{\theta}, \theta^*)$ and $E(\hat{\theta}, \theta^*)$ are different in this case (in sharp contrast to the polyhedral situation). In Section 5.2, we deal

with the risks when σ is large. Under some conditions, it is possible to write a formula for the large σ limits of $M(\hat{\theta}, \theta^*)$ and $E(\hat{\theta}, \theta^*)$; see Proposition 5.3. In Section 5.3, we deal with the maximum normalized risks:

$$\sup_{\sigma>0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) \quad \text{and} \quad \sup_{\sigma>0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*). \tag{10}$$

In the well-specified setting, inequalities (1) and (4) together imply that the maximum normalized risk equals $\delta(T_{\mathcal{C}}(\theta^*))$. However in the misspecified setting, the quantities (10) lie between $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp})$ and $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$. It seems hard to write down an exact formula for the quantities (10) but we present some simulation evidence in Section 5.3 to argue that they can be strictly between $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp})$ and $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$.

We conclude with an appendix that contains technical lemmas and proofs of the various intermediate results throughout the paper.

2. Background and Notation

In this short section, we shall set up some notation and also recollect standard results in convex analysis that will be used in the remainder of the paper.

For $x \in \mathbb{R}^n$ and $r > 0$, we denote by $B_r(x) \coloneqq \{u \in \mathbb{R}^n : \|u - x\| \le r\}$ the closed ball of radius r centered at x. For $v \in \mathbb{R}^n$, let $v^{\perp} \coloneqq \{u \in \mathbb{R}^n : \langle u, v\rangle = 0\}$ denote the hyperplane with normal vector v. For $\theta_0 \in \mathcal{C}$, let $F_{\mathcal{C}}(\theta_0) \coloneqq \{\theta - \theta_0 : \theta \in \mathcal{C}\}$ be the result of re-centering the set $\mathcal{C}$ about θ_0. Also recall the definition of the tangent cone (2) and note that $T_{\mathcal{C}}(\theta_0) = \mathrm{cl}\{\alpha x : x \in F_{\mathcal{C}}(\theta_0), \alpha > 0\}$.

If A is an $m \times n$ matrix and $J \subseteq \{1, \ldots, m\}$, we let a_j denote the jth row of A, and let A_J denote the matrix obtained by combining the rows of A indexed by J.

A *polyhedron* refers to a set of the form $\{x \in \mathbb{R}^n : Ax \le b\}$ for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^n$ where the inequality $\le$ is interpreted coordinate-wise, i.e. $\langle a_j, x\rangle \le b_j$ for $j = 1, \ldots, m$. We will assume that no two pairs (a_j, b_j) and (a_k, b_k) are scalar multiples of each other. A *polyhedral cone* is a set of the form $\{x \in \mathbb{R}^n : Ax \le 0\}$ for some $A \in \mathbb{R}^{m \times n}$. Again, we will assume that no two rows of A are scalar multiples of each other. A *face* of a polyhedron refers to any subset obtained by setting some of the polyhedron's linear inequality constraints to equality instead.

In the remainder of this section, we shall collect some standard results above convex projections that will be used in the paper. These results can be found in a standard reference such as [5]. Recall that $\Pi_{\mathcal{C}}(x)$ denotes the projection of a vector $x \in \mathbb{R}^n$ on a closed convex set $\mathcal{C}$. It is well known that $\Pi_{\mathcal{C}}(x)$ is the unique vector in $\mathcal{C}$ satisfying the optimality condition

$$\langle z - \Pi_{\mathcal{C}}(x), x - \Pi_{\mathcal{C}}(x)\rangle \le 0, \qquad \forall z \in \mathcal{C}. \tag{11}$$

Consequently, we have the following Pythagorean inequality

$$\|z - x\|^2 = \|z - \Pi_{\mathcal{C}}(x)\|^2 + \|\Pi_{\mathcal{C}}(x) - x\|^2 + 2\langle z - \Pi_{\mathcal{C}}(x), \Pi_{\mathcal{C}}(x) - x\rangle \ge \|z - \Pi_{\mathcal{C}}(x)\|^2 + \|\Pi_{\mathcal{C}}(x) - x\|^2.$$

Plugging in $z = \Pi_{\mathcal{C}}(y)$ and $x = \theta^*$ shows that the misspecified error is upper bounded by the excess error, that is,

$$\|\Pi_{\mathcal{C}}(y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \le \|\Pi_{\mathcal{C}}(y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2, \qquad \forall y \in \mathbb{R}^n. \tag{12}$$

If instead we plug in $z = \Pi_{\mathcal{C}}(\theta^*)$ to (11), we have $\langle \Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(\theta^*), x - \Pi_{\mathcal{C}}(x)\rangle \ge 0$, which implies

$$\begin{aligned}
\|\Pi_{\mathcal{C}}(x) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2 &= -\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(\theta^*)\|^2 + 2\langle \Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(\theta^*), \Pi_{\mathcal{C}}(x) - \theta^*\rangle \\
&\le -\|\Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(\theta^*)\|^2 + 2\langle \Pi_{\mathcal{C}}(x) - \Pi_{\mathcal{C}}(\theta^*), x - \theta^*\rangle \\
&\le \|x - \theta^*\|^2.
\end{aligned}$$

Combining this with (12), we see that for $Y = \theta^* + \sigma Z$ we have

$$0 \le \|\Pi_{\mathcal{C}}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \le \|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2 \le \sigma^2 \|Z\|^2. \tag{13}$$

In the special case where $\mathcal{C}$ is a cone, the optimality condition (11) implies that $\Pi_{\mathcal{C}}(x)$ is the unique vector in $\mathcal{C}$ satisfying

$$\langle \Pi_{\mathcal{C}}(x), x - \Pi_{\mathcal{C}}(x)\rangle = 0, \quad \text{and} \quad \langle z, x - \Pi_{\mathcal{C}}(x)\rangle \le 0, \quad \forall z \in \mathcal{C}. \tag{14}$$

3. Main theorem: low noise limit for polyhedra

Our main result below provides a precise characterization of the low σ limits of the risks (5) and (6) (normalized by σ^2) in the misspecified setting (i.e., when $\theta^* \notin \mathcal{C}$) for polyhedral $\mathcal{C}$. An implication of this result is that the low σ limit can be much smaller than the upper bound (7) of Bellec [3].

Theorem 3.1 (Low noise limit of risk for polyhedra)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex set, and let $Y = \theta^* + \sigma Z$ where $\theta^* \in \mathbb{R}^n$ is not necessarily in $\mathcal{C}$, and Z is zero mean with $\mathbb{E}\|Z\|^2 < \infty$. Suppose the following "locally polyhedral" condition holds.*

$$\begin{aligned} &T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \text{ is a polyhedral cone, and} \\ &T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap B_{r^*}(0) = F_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap B_{r^*}(0) \text{ for some } r^* > 0. \end{aligned} \tag{15}$$

Then,

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) = \delta\big(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}\big). \tag{16}$$

Note again that $\delta(\cdot)$ denotes the generalized statistical dimension induced by the noise Z, and reduces to the usual statistical dimension [2] when $Z \sim N(0, I_n)$.

We remark that from the proof of Theorem 3.1 below, one can show the following stronger non-asymptotic result. Let $\mathcal{K} \coloneqq \Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ so that the low noise limit (16) is $\delta(\mathcal{K})$. Then for any $\sigma > 0$,

$$\begin{aligned} \left| \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) - \delta(\mathcal{K}) \right| &\le 2\mathbb{E}\big[\|Z\|^2 \mathbf{1}_{\{\|Z\| > r^*/\sigma\}}\big] \\ \left| \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) - \delta(\mathcal{K}) \right| &\le 2\mathbb{E}\big[\|Z\|^2 \mathbf{1}_{\{\|Z\| > r^*/\sigma\}}\big] \end{aligned} \tag{17}$$

Note that the error term on the right-hand side (17) tends to zero as $\sigma \downarrow 0$, due to $\mathbb{E}\|Z\|^2 < \infty$ and the dominated convergence theorem, which recovers the result (16) of the theorem.

We remark that the "locally polyhedral" condition (15) essentially states that $\mathcal{C}$ looks like a polyhedron in a neighborhood around $\Pi_{\mathcal{C}}(\theta^*)$. As established in the following lemma, it automatically holds if $\mathcal{C}$ is a polyhedron, so one can replace any mention of condition (15) with "$\mathcal{C}$ is a polyhedron" for the sake of readability. We provide some remarks on the case when $\mathcal{C}$ is not polyhedral in Section 5.1.

Lemma 3.2. *Let $\mathcal{C}$ be a polyhedron. Then the locally polyhedral condition* (15) *holds for any $\theta^* \in \mathbb{R}^n$.*

Next, the following lemma establishes that the set $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ that appears in the limit (16) is a face of the tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$.

Lemma 3.3. *Let $\theta^* \in \mathbb{R}^n$ and let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex set satisfying the locally polyhedral condition* (15)*. Let $A \in \mathbb{R}^{m \times n}$ be such that $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) = \{u : Au \le 0\}$. Then there exists some subset $J \subseteq \{1, \dots, m\}$ such that*

$$T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp} = \{u : A_J u = 0, A_{J^c} u \le 0\}.$$

Thus, $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ is a face of $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$.

Both the above lemmas are proved in Appendix A.

If $\theta^* \in \mathcal{C}$ then we have $\Pi_{\mathcal{C}}(\theta^*) = \theta^*$, and Theorem 3.1 reduces to the result (4) of Oymak and Hassibi [8]: the excess risk and the misspecified risk become the same, and the common limit is the statistical dimension of $T_{\mathcal{C}}(\theta^*)$. We must remark here that the result of Oymak and Hassibi [8] holds for non-polyhedral $\mathcal{C}$ as well. We discuss the non-polyhedral setting further in Section 5.1.

Theorem 3.1 states that in the misspecified case $\theta^* \notin \mathcal{C}$, the low sigma limit still involves the tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$, but one needs to intersect it with the hyperplane $(\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ before taking the statistical dimension. Due to the optimality condition (11) characterizing $\Pi_{\mathcal{C}}$, the tangent cone lies entirely on one side of the hyperplane, so the hyperplane does not intersect the interior of the tangent cone. Therefore, the interior

of the tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$ does not contribute to the low σ limit of the risk under misspecification. This makes sense because when $\theta^* \notin \mathcal{C}$ and σ is small, the observation vector Y is outside $\mathcal{C}$ with high probability so that $\hat{\theta}(Y)$ lies on the boundary of $\mathcal{C}$. In fact, in the proof of the theorem we essentially show that for any $\sigma > 0$, under the event $\{\|Z\| \leq r^*/\sigma\}$ (which has probability tending to 1 as $\sigma \downarrow 0$) we have

$$\Pi_{\mathcal{C}}(Y) = \Pi_{\Pi_{\mathcal{C}}(\theta^*)+T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))\cap(\theta^*-\Pi_{\mathcal{C}}(\theta^*))^{\perp}}(Y).$$

In general, the intersection $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ can be anything from $\{0\}$ to the full tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$ and so the low sigma limit can be anything between 0 and $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$. The case when the limit equals zero corresponds to the situation where θ^* lies in the interior of the preimage of $\Pi_{\mathcal{C}}(\theta^*)$ under the map $\Pi_{\mathcal{C}}$ so that every point in some neighborhood of θ^* is projected onto the same point $\Pi_{\mathcal{C}}(\theta^*)$ (see Figure 1c for an example).

The following lemma (proved in Appendix A), provides mild conditions under which the intersection $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$ has strictly smaller generalized statistical dimension than the full tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$.

Lemma 3.4. *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a polyhedron with nonempty interior. Then*

$$\sup_{\theta^* \notin \mathcal{C}: \Pi_{\mathcal{C}}(\theta^*) = \theta_0} \delta\big(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}\big) < \delta(T_{\mathcal{C}}(\theta_0)).$$

for every $\theta_0 \in \mathcal{C}$, provided the random vector Z has nonzero probability of lying in the interior of $T_{\mathcal{C}}(\theta_0)$.

As mentioned already, Lemma 3.4 combined with the main result Theorem 3.1 implies the risk gap (9). In summary, under the nonempty interior assumption, if we think of the low σ limit as a function of θ^*, we see that as θ^* approaches $\mathcal{C}$ from the outside there is a "jump" when θ^* enters $\mathcal{C}$. This "jump" phenomenon is not unique to the polyhedral case. In Section 5.1 we discuss a non-polyhedral example that also exhibits this jump phenomenon.

Theorem 3.1 suggests something that may seem nonintuitive: if $\theta^* \notin \mathcal{C}$ and we use the estimator $\hat{\theta}(Y) = \Pi_{\mathcal{C}}(Y)$, the risk when $Y = \theta^* + \sigma Z$ is smaller than the risk when $Y = \Pi_{\mathcal{C}}(\theta^*) + \sigma Z$. As mentioned already, in the case $Y = \theta^* + \sigma Z$ the estimator is actually estimating $\Pi_{\mathcal{C}}(\theta^*)$, not θ^*. Moreover, the risks (5) and (6) measure error relative to $\Pi_{\mathcal{C}}(\theta^*)$ rather than to θ^*. Furthermore, the intuition is that in the low σ limit, the estimator $\hat{\theta}(Y)$ in the misspecified setting is a projection onto a much smaller set than in the well-specified setting (essentially, a face of a tangent cone instead of the full tangent cone), so more of the original noise in Y is eliminated. This qualitatively explains why having Y generated from θ^* outside $\mathcal{C}$ allows the estimator to estimate $\Pi_{\mathcal{C}}(\theta^*)$ better than if Y were generated from $\Pi_{\mathcal{C}}(\theta^*)$ instead.

Finally, we observe that in the misspecified setting, there is a gap between Bellec's upper bound $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$ (7) and the low σ risk limit, unlike in the well-specified setting where the result (4) implies that the normalized risk increases to the upper bound in the low σ limit. The upper bound, which is constant in σ, can become very loose as $\sigma \downarrow 0$. However, in Section 5.3 we shown a few examples where the normalized risk is close to the upper bound for some σ, as well as examples where the normalized risk remains much smaller than the upper bound for all $\sigma > 0$.

3.1. Proof of Theorem 3.1

We establish one key lemma (proved in Appendix A) before proving Theorem 3.1. It is a deterministic result that contains the core of the argument: roughly, if we have a polyhedral cone $\mathcal{T}$ and any $\theta^* \in \mathbb{R}^n$ satisfying $\Pi_{\mathcal{T}}(\theta^*) = 0$, then any point u sufficiently near θ^* will have its projection $\Pi_{\mathcal{T}}(u)$ lying in the hyperplane with normal direction θ^*.

Lemma 3.5 (Key lemma)**.** *Fix $\theta^* \in \mathbb{R}^n$, and let $\mathcal{T}$ be a closed convex set such that the re-centered set $\{\theta - \Pi_{\mathcal{T}}(\theta^*) : \theta \in \mathcal{T}\}$ is a polyhedral cone. Then there exists $r > 0$ such that*

$$\Pi_{\mathcal{T}}(u) - \Pi_{\mathcal{T}}(\theta^*) \in (\theta^* - \Pi_{\mathcal{T}}(\theta^*))^{\perp}, \qquad \forall u \in B_r(\theta^*). \tag{18}$$

With this lemma, along with some standard results collected in Section 2, we can proceed with proving Theorem 3.1.

Proof of Theorem 3.1. We first prove

$$\lim_{\sigma\downarrow 0}\frac{1}{\sigma^2}M(\hat{\theta},\theta^*)=\delta\big(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))\cap(\theta^*-\Pi_{\mathcal{C}}(\theta^*))^{\perp}\big). \tag{19}$$

For any $r>0$ we can write

$$\frac{1}{\sigma^2}M(\hat{\theta},\theta^*)=\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|\Pi_{\mathcal{C}}(Y)-\Pi_{\mathcal{C}}(\theta^*)\|^2\mathbf{1}_{\{Y\in B_r(\theta^*)\}}\big]+\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|\Pi_{\mathcal{C}}(Y)-\Pi_{\mathcal{C}}(\theta^*)\|^2\mathbf{1}_{\{Y\notin B_r(\theta^*)\}}\big]. \tag{20}$$

We claim the second term on the right-hand side vanishes as $\sigma\downarrow 0$ (regardless of the value of $r>0$). Since the projection $\Pi_{\mathcal{C}}$ is non-expansive [5],

$$0\le\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|\Pi_{\mathcal{C}}(Y)-\Pi_{\mathcal{C}}(\theta^*)\|^2\mathbf{1}_{\{Y\notin B_r(\theta^*)\}}\big]\le\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|Y-\theta^*\|^2\mathbf{1}_{\{Y\notin B_r(\theta^*)\}}\big]=\mathbb{E}\big[\|Z\|^2\mathbf{1}_{\{\|Z\|>r/\sigma\}}\big].$$

Then, the dominated convergence theorem implies the right-hand side tends to zero as $\sigma\downarrow 0$, because $\mathbb{E}\|Z\|^2<\infty$ and the random variable $\|Z\|^2\mathbf{1}_{\{\sigma\|Z\|>r\}}$ converges to zero pointwise.

Thus, it remains to show

$$\lim_{\sigma\downarrow 0}\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|\Pi_{\mathcal{C}}(Y)-\Pi_{\mathcal{C}}(\theta^*)\|^2\mathbf{1}_{\{Y\in B_r(\theta^*)\}}\big]=\delta\big(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))\cap(\theta^*-\Pi_{\mathcal{C}}(\theta^*))^{\perp}\big) \tag{21}$$

for some $r>0$.

We define the re-centered tangent cone

$$\mathcal{T}:=\{\Pi_{\mathcal{C}}(\theta^*)+u:u\in T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))\}.$$

We claim there exists some $r>0$ such that

$$\Pi_{\mathcal{C}}(u)=\Pi_{\mathcal{T}}(u),\qquad\forall u\in B_r(\theta^*). \tag{22}$$

Indeed, note that the locally polyhedral condition (15) implies the existence of some $r^*>0$ such that

$$\mathcal{C}\cap B_{r^*}(\Pi_{\mathcal{C}}(\theta^*))=\mathcal{T}\cap B_{r^*}(\Pi_{\mathcal{C}}(\theta^*)) \tag{23}$$

Since both projections $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{T}}$ are continuous [5] at θ^*, there exists some $r>0$ such that the image of $B_r(\theta^*)$ under both projections lies in $B_{r^*}(\Pi_{\mathcal{C}}(\theta^*))$. In fact, since the projections are each non-expansive [5], we may take $r=r^*$. Thus the local equality (22) of the projections follows from the locally polyhedral condition (23).

By combining this argument with Lemma 3.5, we have shown there exists some $r>0$ that satisfies not only (22), but also (18). With this value of r, the equality (22) implies that replacing each instance of $\mathcal{C}$ with $\mathcal{T}$ in (21) does not change either side, since $\Pi_{\mathcal{C}}(Y)=\Pi_{\mathcal{T}}(Y)$, $\Pi_{\mathcal{C}}(\theta^*)=\Pi_{\mathcal{T}}(\theta^*)$, and

$$T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))=T_{\mathcal{C}\cap B_{r^*}(\Pi_{\mathcal{C}}(\theta^*))}(\Pi_{\mathcal{C}}(\theta^*))=T_{\mathcal{T}\cap B_{r^*}(\Pi_{\mathcal{C}}(\theta^*))}(\Pi_{\mathcal{C}}(\theta^*))=T_{\mathcal{T}}(\Pi_{\mathcal{T}}(\theta^*)),$$

by the equality (23) and the definition of the tangent cone. Thus it remains to prove

$$\lim_{\sigma\downarrow 0}\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\big[\|\Pi_{\mathcal{T}}(Y)-\Pi_{\mathcal{T}}(\theta^*)\|^2\mathbf{1}_{\{Y\in B_r(\theta^*)\}}\big]=\delta(\mathcal{K}), \tag{24}$$

where $\mathcal{K}:=T_{\mathcal{T}}(\theta^*)\cap(\theta^*-\Pi_{\mathcal{T}}(\theta^*))^{\perp}$.

Since r satisfies (18), some re-centering yields

$$\Pi_{\mathcal{T}}(Y)-\Pi_{\mathcal{T}}(\theta^*)=\Pi_{T_{\mathcal{T}}(\theta^*)}(Y-\Pi_{\mathcal{T}}(\theta^*))=\Pi_{\mathcal{K}}(Y-\Pi_{\mathcal{T}}(\theta^*)) \tag{25}$$

in the event $\{Y\in B_r(\theta^*)\}$.

For $W := (\theta^* - \Pi_{\mathcal{T}}(\theta^*))^\perp$, we claim

$$\Pi_{\mathcal{K}} = \Pi_{\mathcal{K}} \circ \Pi_W.$$

In fact this holds for any subspace W and closed convex $\mathcal{K} \subseteq W$, by the Pythagorean theorem:

$$\Pi_{\mathcal{K}}(x) = \operatorname*{argmin}_{u \in \mathcal{K}} \|x-u\|^2 = \operatorname*{argmin}_{u \in \mathcal{K}} \big\{ \|x - \Pi_W(x)\|^2 + \|\Pi_W(x) - u\|^2 \big\} = \Pi_{\mathcal{K}}(\Pi_W(x)).$$

Applying this to (25) yields

$$\begin{aligned}
\Pi_{\mathcal{T}}(Y) - \Pi_{\mathcal{T}}(\theta^*) &= \Pi_{\mathcal{K}}(Y - \Pi_{\mathcal{T}}(\theta^*)) \\
&= \Pi_{\mathcal{K}}(\Pi_W(\theta^* + \sigma Z - \Pi_{\mathcal{T}}(\theta^*))) \\
&= \Pi_{\mathcal{K}}(\Pi_W(\sigma Z)) && \Pi_W \text{ is linear, } \Pi_W(\theta^* - \Pi_{\mathcal{T}}(\theta^*)) = 0 \\
&= \Pi_{\mathcal{K}}(\sigma Z) = \sigma \Pi_{\mathcal{K}}(Z) && \mathcal{K} \text{ is a cone}
\end{aligned}$$

in the event $\{Y \in B_r(\theta^*)\}$. By plugging this into the left-hand side of equation (24), we have

$$\lim_{\sigma \downarrow 0} \mathbb{E}_{\theta^*} \big[\|\Pi_{\mathcal{K}}(Z)\|^2 \mathbf{1}_{\{Y \in B_r(\theta^*)\}} \big] = \mathbb{E}\|\Pi_{\mathcal{K}}(Z)\|^2 = \delta(\mathcal{K}),$$

where the first equality follows by dominated convergence ($\|\Pi_{\mathcal{K}}(Z)\|^2 \le \|Z\|^2$ and $\mathbb{E}\|Z\|^2 < \infty$). This verifies the desired equality (24) and concludes the proof of the first low σ limit (19).

We now prove the other equality

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*).$$

We claim

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\theta^*} \big[\big(\|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2 \big) \mathbf{1}_{\{Y \notin B_r(\theta^*)\}} \big] = 0 \tag{26}$$

for any $r > 0$. Applying some basic properties (13) of the projection $\Pi_{\mathcal{C}}$ yields

$$0 \le \frac{1}{\sigma^2} \mathbb{E}_{\theta^*} \big[\big(\|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2 \big) \mathbf{1}_{\{Y \notin B_r(\theta^*)\}} \big] \le \mathbb{E}\big[\|Z\|^2 \mathbf{1}_{\{\|Z\| \ge r/\sigma\}} \big],$$

so applying the dominated convergence theorem as before leads to the limit (26).

Thus, it suffices to prove

$$\begin{aligned}
&\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\theta^*} \big[\|\Pi_{\mathcal{C}}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \mathbf{1}_{\{Y \in B_r(\theta^*)\}} \big] \\
&= \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} \mathbb{E}_{\theta^*} \big[\big(\|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2 \big) \mathbf{1}_{\{Y \in B_r(\theta^*)\}} \big]
\end{aligned} \tag{27}$$

for some $r > 0$. We choose r as before so that (18) and (22) both hold. By the same reasoning as before, we can replace each instance of $\mathcal{C}$ with $\mathcal{T}$ without changing anything. Furthermore, the condition (18) implies we have $\langle \Pi_{\mathcal{T}}(Y) - \Pi_{\mathcal{T}}(\theta^*), \theta^* - \Pi_{\mathcal{T}}(\theta^*) \rangle = 0$ in the event $\{Y \in B_r(\theta^*)\}$, so the Pythagorean inequality (12) becomes equality:

$$\|\Pi_{\mathcal{T}}(Y) - \Pi_{\mathcal{T}}(\theta^*)\|^2 \mathbf{1}_{\{Y \in B_r(\theta^*)\}} = \big(\|\Pi_{\mathcal{T}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{T}}(\theta^*) - \theta^*\|^2 \big) \mathbf{1}_{\{Y \in B_r(\theta^*)\}}.$$

Therefore the equality (27) holds, which concludes the proof of Theorem 3.1. □

4. Examples

In this section, we assume the Gaussian noise model $Z \sim N(0, I_n)$, or equivalently $Y \sim N(\theta^*, \sigma^2 I_n)$. Thus, $\delta(\cdot)$ denotes the usual statistical dimension [2], where Z in the definition (3) is a standard Gaussian vector.

4.1. Nonnegative orthant

We now apply Theorem 3.1 to the *nonnegative orthant* $\mathbb{R}^n_+ \coloneqq \{u \in \mathbb{R}^n : u_i \geq 0, \forall i\}$. In Figure 1 we provide visualizations of the geometry of the main theorem when applied to this constraint set.

Corollary 4.1 (Nonnegative orthant)**.** *Let $Y \sim N(\theta^*, \sigma^2 I)$ where $\theta^* \in \mathbb{R}^n$. Let $n_+ \coloneqq \sum_{i=1}^n \mathbf{1}_{\{\theta_i^* > 0\}}$ and $n_0 \coloneqq \sum_{i=1}^n \mathbf{1}_{\{\theta_i^* = 0\}}$ denote the number of positive components and number of zero components of θ^* respectively. Then the normalized excess risk* (6) *and normalized mispecified risk* (5) *of the least squares estimator $\hat{\theta}(Y) \coloneqq \Pi_{\mathbb{R}^n_+}(Y)$ with respect to $\mathbb{R}^n_+$ both tend to*

$$\frac{n_0}{2} + n_+$$

as $\sigma \downarrow 0$.

Proof. By Theorem 3.1, it suffices to prove that the statistical dimension term in (16) is $\frac{n_0}{2} + n_+$. Note that for $y \in \mathbb{R}^n$, $\Pi_{\mathbb{R}^n_+}(y) = \max\{y, 0\}$ is obtained by taking the component-wise maximum of y with 0. Consequently,

$$T_{\mathbb{R}^n_+}(\Pi_{\mathbb{R}^n_+}(\theta^*)) = \{u \in \mathbb{R}^n : u_i \geq 0 \text{ if } (\Pi_{\mathbb{R}^n_+}(\theta^*))_i = 0\} = \{u \in \mathbb{R}^n : u_i \geq 0 \text{ if } \theta_i^* \leq 0\}.$$

Also,

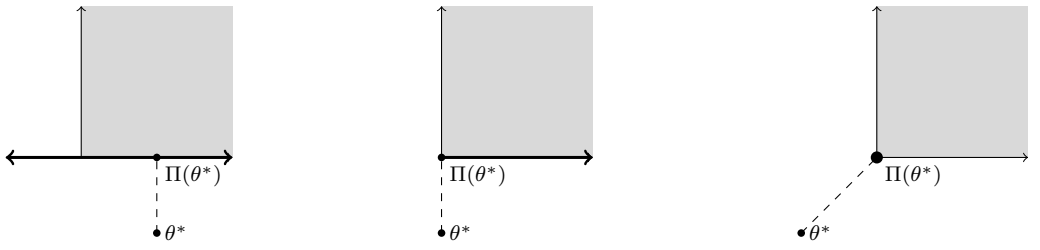
$$(\theta^* - \Pi_{\mathbb{R}^n_+}(\theta^*))^\perp = \left\{ u \in \mathbb{R}^n : \sum_{i:\theta_i^* < 0} \theta_i^* u_i = 0 \right\}$$

The intersection is thus

$$T_{\mathbb{R}^n_+}(\Pi_{\mathbb{R}^n_+}(\theta^*)) \cap (\theta^* - \Pi_{\mathbb{R}^n_+}(\theta^*))^\perp = \left\{ u \in \mathbb{R}^n : \begin{array}{ll} u_i \geq 0 & \text{if } \theta_i^* = 0 \\ u_i = 0 & \text{if } \theta_i^* < 0 \end{array} \right\} \cong \mathbb{R}^{n_+} \times \mathbb{R}^{n_0}_+ \times \{0\}^{n - n_+ - n_0}.$$

The result follows by noting $\delta(\mathbb{R}) = 1$ and $\delta(\mathbb{R}_+) = 1/2$ and by using the fact that $\delta(T_1 \times T_2) = \delta(T_1) + \delta(T_2)$ for any two cones T_1 and T_2 [2]. □

Remark 4.2. *For $\theta^* \in \mathbb{R}^n$ let n_+ and n_0 be as defined in Corollary 4.1. Then the low σ limit for the corresponding well-specified problem $Y \sim N(\Pi_{\mathbb{R}^n_+}(\theta^*), \sigma^2 I)$ is $\frac{n - n_+}{2} + n_+$ since all negative components of θ^* are sent to zero by $\Pi_{\mathbb{R}^n_+}$. This is larger than the low σ limit for the misspecified problem $Y \sim N(\theta^*, \sigma^2 I)$ because $n - n_+ \geq n_0$, with strict inequality if $\theta^* \notin \mathbb{R}^n_+$.*



(a) $\theta = (1, -1)$; $\delta = 1$ (b) $\theta = (0, -1)$; $\delta = 1/2$ (c) $\theta = (-1, -1)$; $\delta = 0$

Figure 1: $\mathbb{R}^2_+$ is marked by the gray area. The intersection $T_{\mathbb{R}^2_+}(\Pi_{\mathbb{R}^2_+}(\theta^*)) \cap (\theta^* - \Pi_{\mathbb{R}^2_+}(\theta^*))^\perp$ [translated to be centered at $\Pi_{\mathbb{R}^2_+}(\theta^*)$] is marked by the bold lines in the first two examples, and the bold point in the third example. Each sub-caption states the statistical dimension $\delta = \delta(T_{\mathbb{R}^2_+}(\Pi_{\mathbb{R}^2_+}(\theta^*)) \cap (\theta^* - \Pi_{\mathbb{R}^2_+}(\theta^*))^\perp)$.

4.2. Consequences for isotonic regression

This section details interesting consequences of Theorem 3.1 for isotonic regression under misspecification. Let

$$\mathcal{S}^n := \{u \in \mathbb{R}^n : u_1 \le \cdots \le u_n\}$$

be the *monotone cone*. We call elements of $\mathcal{S}^n$ *nondecreasing*.

By a *block*, we refer to a set of the form $\{k, k+1, \ldots, l\}$ for two nonnegative integers $k \le l$. Consider a partition of $\{1, \ldots, n\}$ into blocks $I_1, \ldots, I_m$ listed in increasing order (i.e., the maximum entry of I_i is strictly smaller than the minimum entry of I_j for $i < j$). Let $|I_j|$ denote the cardinality of I_j and note that $\sum_{j=1}^m |I_j| = n$ as $I_1, \ldots, I_m$ form a partition of $\{1, \ldots, n\}$. Let $\mathcal{S}_{|I_1|,\ldots,|I_m|}$ denote the induced *block monotone cone* defined as

$$\mathcal{S}_{|I_1|,\ldots,|I_m|} := \{u \in \mathcal{S}^n : u \text{ is constant on each of the blocks } I_1, \ldots, I_m\} \tag{28}$$

For example,

$$\mathcal{S}_{2,3,2} = \{u \in \mathbb{R}^{2+3+2} : u_1 = u_2 \le u_3 = u_4 = u_5 \le u_6 = u_7\}.$$

Theorem 3.1 implies the following result, which we prove in Section B.3.

Proposition 4.3 (Isotonic regression)**.** *Let $Y \sim N(\theta^*, \sigma^2 I)$ where $\theta^* \in \mathbb{R}^n$. Let $(J_1, \ldots, J_K)$ be the partition of $\{1, \ldots, n\}$ into blocks such that $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant on each J_k with respective values $\mu_1 < \cdots < \mu_K$. For each $k \in \{1, \ldots, K\}$, there exists a unique finest partition $(I_1^k, \ldots, I_{m_k}^k)$ of J_k into blocks such that for all $j \in \{1, \ldots, m_k\}$, the mean of the components of θ^* on each I_j^k equals μ_k; that is,*

$$\frac{1}{|I_j^k|} \sum_{i \in I_j^k} \theta_i^* = \mu_k, \qquad 1 \le j \le m_k. \tag{29}$$

Then the common low σ limit of the normalized excess risk (6) *and normalized misspecified risk* (5) *of the isotonic least squares estimator $\hat{\theta}(Y) := \Pi_{\mathcal{S}^n}(Y)$ equals*

$$\sum_{k=1}^K \delta\Big(\mathcal{S}_{|I_1^k|,\ldots,|I_{m_k}^k|}\Big). \tag{30}$$

It is clear from the above proposition that the low σ behavior of the isotonic estimator under misspecification crucially depends on the statistical dimension of the block monotone cone $\mathcal{S}_{|I_1^k|,\ldots,|I_{m_k}^k|}$. [We remark again that throughout this section we only deal with the usual statistical dimension, where the noise Z in the definition (3) is standard Gaussian.] Here, we provide two simple properties of the block monotone cone (28), each of which implies that when the block sizes are equal, the statistical dimension is simply that of $\mathcal{S}^{m_k}$. The first result provides a direct connection to weighted isotonic regression.

Lemma 4.4 (Weighted isotonic regression)**.** *Let $z \in \mathbb{R}^n$ and let $I_1, \ldots, I_m$ be a partition of $\{1, \ldots, n\}$ into blocks. Let $\bar{z}_{I_j} := \frac{1}{|I_j|} \sum_{i \in I_j} z_i$. Then $\Pi_{\mathcal{S}_{|I_1|,\ldots,|I_m|}}(y)$ is the vector that is constant on the blocks $I_1, \ldots, I_m$ with constant values $x_1^*, \ldots, x_m^*$, where $x^* = (x_1^*, \ldots, x_m^*)$ is*

$$x^* = \operatorname*{argmin}_{x \in \mathcal{S}^m} \sum_{j=1}^m |I_j| (x_i - \bar{z}_{I_j})^2.$$

In other words, the values on the constant blocks of $\Pi_{\mathcal{S}_{|I_1|,\ldots,|I_m|}}(z)$ can be found by weighted isotonic regression of $(\bar{z}_{I_1}, \ldots, \bar{z}_{I_m}) \in \mathbb{R}^m$ with weights $|I_1|, \ldots, |I_m|$.

Consequently, when $|I_1| = \cdots = |I_m|$, the statistical dimension of the block monotone cone is

$$\delta(\mathcal{S}_{|I_1|,\ldots,|I_m|}) = \sum_{j=1}^m \frac{1}{j}.$$

The next lemma shows $\mathcal{S}_{|I_1|,\ldots,|I_m|}$ is isometric to a particular cone in the lower-dimensional space $\mathbb{R}^m$.

Lemma 4.5 (Block monotone cone isometry)**.** *The block monotone cone $\mathcal{S}_{|I_1|,\ldots,|I_m|} \subseteq \mathbb{R}^n$ is isometric to*

$$\left\{ v \in \mathbb{R}^m : \frac{v_1}{\sqrt{|I_1|}} \leq \cdots \leq \frac{v_m}{\sqrt{|I_m|}} \right\} \subseteq \mathbb{R}^m, \tag{31}$$

and thus both sets have the same statistical dimension. In particular, if $|I_1| = \cdots = |I_m|$, then the statistical dimension of the block monotone cone is

$$\delta(\mathcal{S}_{|I_1|,\ldots,|I_m|}) = \sum_{j=1}^{m} \frac{1}{j}.$$

Both lemmas are proved in Section B.1. Note that for the case $|I_1| = \cdots = |I_m| = 1$, both lemmas reduce to the statement of the statistical dimension of the monotone cone $\mathcal{S}^n$ [2, Eq. D.12]. More generally, when the m blocks have equal size, the statistical dimension of the associated block monotone cone is the same as that of the monotone cone $\mathcal{S}^m$. In Section B.2, we discuss what Lemma 4.5 suggests for the completely general case when the block sizes are arbitrary.

By combining either of these two lemmas with Proposition 4.3, we immediately obtain an explicit expression for the low σ limits in a special case. For $m \geq 1$, we denote the harmonic number $\sum_{j=1}^{m}(1/j)$ by H_m.

Corollary 4.6 (Isotonic regression with equal sub-block sizes)**.** *Consider the setting of Proposition 4.3. In the special case where*

$$|I_1^k| = \cdots = |I_{m_k}^k| \qquad \text{for each } k \in \{1, \ldots, K\}, \tag{32}$$

the common low σ limit has the following explicit expression:

$$\sum_{k=1}^{K} H_{m_k} = \sum_{k=1}^{K} \sum_{j=1}^{m_k} \frac{1}{j}.$$

See the examples to follow (as well as Section B.2) for further discussion about how the statistical dimension of $\mathcal{S}_{|I_1^k|,\ldots,|I_{m_k}^k|}$ behaves in general, when the special condition (32) does not hold.

In Table 1, we demonstrate how to apply this theorem to various cases of θ^*. In the "partition of θ^*" column, we use square brackets to partition the components of θ^* into K blocks according to the constant pieces $\mu_1 < \cdots < \mu_K$ of $\Pi_{\mathcal{S}^n}(\theta^*)$, and then within the kth group use parentheses to further partition the components into m_k sub-blocks each with common mean μ_k.

θ^*	$\Pi_{\mathcal{S}^n}(\theta^*)$	partition of θ^*	$m_1, \ldots, m_K$	$\sum_{k=1}^{K} H_{m_k}$
$(0,0,0,0,0,0)$	$(0,0,0,0,0,0)$	$[(0),(0),(0),(0),(0),(0)]$	6	$H_6 = 2.45$
$(1,-1,1,-1,1,-1)$	$(0,0,0,0,0,0)$	$[(1,-1),(1,-1),(1,-1)]$	3	$H_3 = 1.8\bar{3}$
$(5,3,1,-1,-3,-5)$	$(0,0,0,0,0,0)$	$[(5,3,1,-1,-3,-5)]$	1	$H_1 = 1$
$(-1,-1,-1,-1,2,2)$	$(-1,-1,-1,-1,2,2)$	$[(-1),(-1),(-1),(-1)],[(2),(2)]$	4, 2	$H_4 + H_2 = 3.58\bar{3}$
$(0,-2,1,-3,2,2)$	$(-1,-1,-1,-1,2,2)$	$[(0,-2),(1,-3)],[(2),(2)]$	2, 2	$H_2 + H_2 = 3$
$(0,0,-2,-2,3,1)$	$(-1,-1,-1,-1,2,2)$	$[(0,0,-2,-2)],[(3,1)]$	1, 1	$H_1 + H_1 = 2$

Table 1. Examples of how to compute the limit in Proposition 4.3 in the special case (32).

We now discuss in detail what Proposition 4.3 states for certain cases of θ^*.

1. **Comparison with well-specified case.** In the well-specified case where $\theta^* \in \mathcal{S}^n$, we have $\theta_j^* = \mu_k$ for all $j \in J_k$ and $k \in \{1, \ldots, K\}$, so the finest partition of each J_k is the partition into singleton sets.

Then $m_k = |J_k|$ for each k, and moreover $|I_j^k| = 1$ for all valid k and j. Thus, Proposition 4.3 implies that both low σ limits are

$$\sum_{k=1}^{K} H_{|J_k|} \coloneqq \sum_{k=1}^{K} \sum_{j=1}^{|J_k|} \frac{1}{j},$$

This is precisely the upper bound (7) for the monotone cone as computed by Bellec [3, Prop. 3.1], so we recover the low σ limit (4). Computations for the well-specified examples $\theta^* = (0,0,0,0,0,0)$ and $\theta^* = (-1,-1,-1,-1,2,2)$ appear in Table 1.

Now, consider the misspecified problem $Y \sim N(\theta^*, \sigma^2 I_n)$ with $\theta^* \notin \mathcal{S}^n$, and compare the statement of Proposition 4.3 with the corresponding statemetn for the well-specified problem $Y \sim N(\Pi_{\mathcal{S}^n}(\theta^*), \sigma^2 I)$. In both cases, the partition of $\{1,\dots,n\}$ into $(J_1,\dots,J_K)$ is the same. However, we showed above that in the well-specified problem, the sub-partition of each J_k consists of singletons, whereas for the misspecified problem we may get nontrivial partitions $(I_1^k,\dots,I_{m_k}^k)$. Noting the inclusion $\mathcal{S}_{|I_1^k|,\dots,|I_{m_k}^k|} \subseteq \mathcal{S}^{|J_k|}$ for each k and comparing (30) for the two cases yields

$$\sum_{k=1}^{K} \delta(\mathcal{S}_{|I_1^k|,\dots,|I_{m_k}^k|}) \le \sum_{k=1}^{K} \delta(\mathcal{S}^{|J_k|}),$$

which shows that in general the misspecified low σ limit is smaller than the corresponding well-specified limit.

2. **Decreasing sequences.** Suppose θ^* is nonincreasing and nonconstant i.e., $\theta^* \in (-\mathcal{S}^n) \setminus \mathcal{S}^n$. Then $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant (see [10] for various properties of $\Pi_{\mathcal{S}^n}$), so $K = 1$ and $\mu_1 = \frac{1}{n}\sum_{i=1}^n \theta_i^*$. We also claim $m_1 = 1$. Indeed, if $m_1 > 1$ then there exists some $j < n$ such that $\mu = \frac{1}{j}\sum_{i=1}^j \theta_i^* = \frac{1}{n-j}\sum_{i=j+1}^n \theta_i^*$. However, the fact that θ^* is nonincreasing and nonconstant implies $\frac{1}{j}\sum_{i=1}^j \theta_i^* > \frac{1}{n-j}\sum_{i=j+1}^n \theta_i^*$, a contradiction. Thus, Proposition 4.3 implies that both low σ limits are

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) = 1. \tag{33}$$

(In fact, by combining the above argument with the proof of Proposition 4.3, we have shown that the intersection $T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^\perp$ is simply the subspace of constant sequences.) On the other hand, since $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant, the low σ limit in the well-specified setting $Y \sim N(\Pi_{\mathcal{S}^n}(\theta^*), \sigma^2 I_n)$ is $\sum_{j=1}^n \frac{1}{j} \asymp \log n$, which is much larger.

The logarithmic term appears here in the well-specified case due to the well-known spiking effect of isotonic regression (documented, for example, by Pal [9], Wu et al. [12], Zhang [13]). Indeed, the isotonic estimator is inconsistent near the end points which leads to the logarithm term in the risk. However, in the misspecified case when θ^* is nonincreasing and nonconstant, a combination of the proof of Theorem 3.1 (in particular Lemma 3.5) with the fact that $T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^\perp$ is the subspace of all constant sequences implies $\hat{\theta}(Y)$ is a constant sequence with probability increasing to 1 as $\sigma \downarrow 0$, in which case the constant value must be the sample mean $\bar{Y} \coloneqq \frac{1}{n}\sum_{i=1}^n Y_i$. Alternatively, one can rephrase the geometric argument in Lemma 3.5 more simply in this example; when σ is small, Y is near θ^* and thus is also nondecreasing with high probability, in which case $\hat{\theta}(Y)$ is constant, due to the properties of the projection $\Pi_{\mathcal{S}^n}$. Hence, in this situation the estimator is essentially a linear projection onto a one-dimensional subspace (hence the low noise limit of 1) and does not suffer from any spiking at the endpoints, and consequently there are no logarithmic terms in the risk in the misspecified case in the low sigma limit.

Computations for the specific example when $\theta^* = (5,3,1,-1,-3,-5)$ appear in Table 1.

3. In the first half of Table 1 we consider three choices for θ^* that project to $\Pi_{\mathcal{S}^n}(\theta^*) = (0,0,0,0,0,0)$. Here $K = 1$ and the sub-block sizes $|I_1^1|,\dots,|I_{m_1}^1|$ are equal in each case (namely, the common block size is 1, 2, and 6 respectively), so we are in the special case (32). Thus, the limit is $\sum_{j=1}^{m_1} \frac{1}{j}$ where m_1 is the number of sub-blocks. We see that for the misspecified θ^* the low σ limits are smaller.

One can heuristically interpret Theorem 3.1 for the example $\theta^* = (1,-1,1,-1,1,-1)$ as follows. With probability increasing to 1 as $\sigma \downarrow 0$, the estimator $\hat{\theta}(Y)$ is nondecreasing and piecewise constant on three equally sized blocks, so the low σ limit is the same as if we were estimating $(0,0,0)$ in $\mathcal{S}^3$.

4. Similarly in the second half of Table 1 we consider three θ^* that project to $\Pi_{\mathcal{S}^n}(\theta^*) = (-1,-1,-1,-1,2,2)$. Here $K = 2$ but, since the low σ limit decomposes, we can simply consider each constant piece separately. Again, we see that the more sub-blocks I_i^j, the higher the statistical dimension, with the well-specified case having the most sub-blocks (all singletons).
5. **Beyond equi-sized sub-blocks.** The concrete examples we have considered so far have been in the special case (32). In a few other cases we can still provide the low σ limit. (See also Section B.2 for further discussion.)
   (a) If $K = 1$ and $m_1 = 2$, then the low σ limit is $\delta(S_{|I_1^1|,|I_2^1|})$. By Lemma 4.5, this is the same as the statistical dimension of the half space $\{u \in \mathbb{R}^2 : u_1/\sqrt{|I_1|} \le u_2/\sqrt{|I_2|}\}$, which is 1.5. However, when $m_1 > 2$, it is difficult to compute $\delta(S_{|I_1^1|,\dots|I_{m_1}^1|})$ unless we are in the special case $|I_1^1| = \cdots = |I_{m_1}^1|$.
   (b) In some other extreme cases we can get an approximation. For example, if
   $$\theta^* = (0, \underbrace{1,\dots,1}_{(n-2)/2}, \underbrace{-1,\dots,-1}_{(n-2)/2}, 0),$$
   then $\Pi_{\mathcal{S}^n}(\theta^*) = (0,\dots,0)$, so the low σ limit is $\delta(\mathcal{S}_{1,n-2,1})$. Lemma 4.5 shows that this is the same as the statistical dimension of $\{u \in \mathbb{R}^3 : u_1 \le u_2/\sqrt{n-2} \le u_3\}$. As $n \to \infty$ tends to this set tends to $\{u \in \mathbb{R}^3 : u_1 \le 0 \le u_3\}$ which has statistical dimension $1 + \frac{1}{2} + \frac{1}{2} = 2$. Thus $\delta(\mathcal{S}_{1,n-2,1}) \to 2$ as $n \to \infty$. We used simulations to verify that the low σ limit is indeed near 2 even for $n = 20$.

4.3. An example in convex regression

In the previous section, we showed (33) that when θ^* is nonincreasing and nonconstant, the low noise limits for isotonic regression are 1. In this section we prove the analogous result in the setting of convex regression.

We fix design points $x_1 < \cdots < x_n$. In this section we use the shorthand $x \coloneqq (x_1,\dots,x_n)$. The relevant cone for convex regression on this design is

$$\mathcal{K}_x \coloneqq \left\{\theta \in \mathbb{R}^n : \frac{\theta_i - \theta_{i-1}}{x_i - x_{i-1}} \le \frac{\theta_{i+1} - \theta_i}{x_{i+1} - x_i}, \quad i = 2,\dots,n-1\right\}. \tag{34}$$

The constraints naturally arise when considering $\theta_i = f(x_i)$ for some convex function f.

Unlike in the case of isotonic regression, the statistical dimension of tangent cones of $\mathcal{K}_x$ is not known exactly. However, upper bounds are available and the best known upper bound is due to Bellec [3, Prop. 4.2] who proved that for every $\theta \in \mathcal{K}_x$,

$$\delta(T_{\mathcal{K}_x}(\theta)) \le 8q \log \frac{en}{q}, \tag{35}$$

where q is the number of affine pieces in θ (i.e., one plus the number of inequalities in the definition (34) of $\mathcal{K}_x$ that hold with strict inequality).

Because finding an exact formula for $\delta(T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*)))$ seems intractable, it is unlikely that an exact formula for the statistical dimension of $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{K}_x}(\theta^*))^\perp$ can be found. As a result, in contrast to the case of isotonic regression, it might not be possible to precisely determine the limits in (16) for most $\theta^* \in \mathbb{R}^n$ in convex regression. However in the special case where θ^* is concave and nonlinear, it is indeed possible to characterize the limits in (16). This is the content of the next result.

Proposition 4.7. *Suppose $\theta^* \in (-\mathcal{K}_x) \setminus \mathcal{K}_x$ is concave and nonlinear. Then the common low σ limit of the normalized excess risk* (6) *and normalized misspecified risk* (5) *of the convex least squares estimator $\hat\theta(Y) \coloneqq \Pi_{\mathcal{K}_x}(Y)$ is* 2.

We prove this result in Appendix C. The intuition behind this result is completely analogous to the aforementioned example in isotonic regression. Here, when σ is small, Y is concave and nonlinear with high probability, and gets projected onto the subspace $\operatorname{span}\{(1,\dots,1), x\}$ of affine vectors on x, which has dimension 2.

Note that for the setting of Proposition 4.7, the analogous well-specified problem is $Y \sim N(\Pi_{\mathcal{K}_x}(\theta^*), \sigma I_n)$, which by the upper bound (7) (and the bound (35)) has low noise limit $\delta(T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*))) \le 8\log(en)$. So, in this situation, the misspecified low noise limits, 2, are much smaller than the corresponding well-specified limit, $\asymp \log n$ (assuming the upper bound (35) is tight). Recall the analogous situation (33) for decreasing θ^* in isotonic regression, where the misspecified and well-specified limits are 1 and $\asymp \log n$ respectively.

5. Further discussion

5.1. Generalizing Theorem 3.1 to the non-polyhedral case

Note that Theorem 3.1 requires the condition (15) i.e., that $\mathcal{C}$ is locally a polyhedron near $\Pi_{\mathcal{C}}(\theta^*)$. Here we comment on the situation when $\mathcal{C}$ is non-polyhedral. Although non-polyhedral convex sets can be approximated by polyhedra, the low σ limit magnifies the local geometry of the set and ignores the goodness of such an approximation. As a stark counterexample, consider any closed convex $\mathcal{C} \subseteq \mathbb{R}^2$ with nonempty interior, and let $Z \sim N(0, I_n)$. For any polygon in $\mathbb{R}^2$, Theorem 3.1 implies that the low σ limits are either 0, 1/2, or 1 because in $\mathbb{R}^2$ the intersection of a convex cone with a line intersecting the origin is either the origin, a ray, or a line. Thus, for a sequence of polygons approximating $\mathcal{C}$ the sequence of corresponding low σ limits need not even have a limit, never mind the matter of two different sequences of polygonal approximations having a common limit. Therefore, the low σ limit for general $\mathcal{C}$ cannot be found using a polyhedral approximation.

In order to understand how the low σ limits behave for general $\mathcal{C}$, we consider the following specific example. Let $\mathcal{C} \coloneqq \{\theta \in \mathbb{R}^n : \|\theta\| \le 1\}$ be the unit ball so that $\Pi_{\mathcal{C}}(x) = \frac{x}{\max\{\|x\|,1\}}$. Also let $\theta^* \coloneqq (r, 0, \ldots, 0)$ for some $r > 1$ so that $\Pi_{\mathcal{C}}(\theta^*) = (1, 0, \ldots, 0)$. By rotational symmetry of $\mathcal{C}$, the case of any general $\theta^* \notin \mathcal{C}$ can be reduced to this case.

In the corresponding well-specified case $Y \sim N(\Pi_{\mathcal{C}}(\theta^*), \sigma^2 I_n))$, the result (4) of Oymak and Hassibi [8] implies that the normalized misspecified risk (5) and the normalized excess risk (6) are equal in the low σ limit with common value

$$\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))) = n - \frac{1}{2},$$

since the tangent cone is the half space $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) = \{x \in \mathbb{R}^n : x_1 \le 0\}$.

However, in the misspecified case, we observe some new phenomena that do not occur for polyhedra.

Proposition 5.1 (Low noise limits for the ball)**.** *Let $\mathcal{C} \coloneqq \{\theta \in \mathbb{R}^n : \|\theta\|_2 \le 1\}$, $\theta^* \notin \mathcal{C}$, and $Y \sim N(\theta^*, \sigma^2 I_n)$. For the estimator $\hat{\theta}(Y) = \Pi_{\mathcal{C}}(Y)$, we have*

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \frac{n-1}{\|\theta^*\|^2}, \tag{36a}$$

$$\lim_{\sigma \downarrow 0} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) = \frac{n-1}{\|\theta^*\|}. \tag{36b}$$

The proof involves direct computation and appears in Appendix D.

We now highlight some of the interesting behavior. In the polyhedral case, both limits were equal; in the proof of Theorem 3.1 (in particular Lemma 3.5) we showed that with probability increasing to 1 (in the low σ limit), Y would be projected onto the hyperplane $(\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp$, producing the orthogonality required for the Pythagorean inequality (12) to become an equality. In the general case, the Pythagorean inequality is not tight, and we explicitly see from this example that even in the low noise limit the the excess risk can be strictly larger than the misspecified risk.

Note that in contrast to the corresponding well-specified case $Y \sim N(\Pi_{\mathcal{C}}(\theta^*), \sigma^2 I_n)$ which has limit $n - \frac{1}{2}$, the misspecified limits $\frac{n-1}{\|\theta^*\|^2}$ and $\frac{n-1}{\|\theta^*\|}$ both tend to $n-1$ as $\|\theta^*\| \downarrow 1$, so there is a "jump" in the limits between the misspecified and well-specified setting. This is also a feature of Theorem 3.1 when the polyhedron $\mathcal{C}$ has nonempty interior, as we discussed earlier (see Lemma 3.4).

This example shows that Theorem 3.1 does not hold for nonpolyhedral constraint sets $\mathcal{C}$, as the two normalized risks are not equal in this particular example of the unit ball, and moreover neither limit equals

$$\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp) = \delta(\{u : \langle u, \theta^* \rangle \le 0\} \cap (\theta^*)^\perp) = \delta((\theta^*)^\perp) = n - 1.$$

The intuition for Theorem 3.1 is that, in the polyhedral case, the projections of Y largely end up in some face of the polyhedron $\mathcal{C}$, which can be approximated by a lower-dimensional cone, for which the statistical dimension is well defined. When $\mathcal{C}$ is not polyhedral, the generalization of this "face" is hard to conceptualize and is likely not well approximated by a cone, so a statistical dimension can not be even applied. Indeed, for general $\mathcal{C}$ such as the ball, tangent cones are extremely poor approximations for the set. Contrary to this drawback, the result (4) of Oymak and Hassibi [8] shows that tangent cones are good enough for the well-specified setting. However for the misspecified setting, we expect that any general result for the low σ limits does not involve a statistical dimension of some cone, since the surface of $\mathcal{C}$ is the essential object of interest and cannot be approximated by some cone except in special settings like the polyhedral case.

As mentioned already, Theorem 3.1 shows that in the misspecified setting, the upper bound (7), which holds for all σ, is not tight in the low σ limit. One might ask whether a better upper bound for all σ can be achieved, but Figure 2 shows that for some values of σ the risks can be close to the upper bound, represented by the solid horizontal line. We observed this behavior in other examples (see also Figure 5): the risks can be close to the upper bound for some moderate values of σ, and then converge to the strictly smaller low σ limit. Replacing the upper bound (7), which is constant in σ, with a σ-dependent upper bound would be an interesting result, but it would have to be extremely dependent on the geometry of the set $\mathcal{C}$. In the following sections we further discuss the normalized risks as a function of σ.

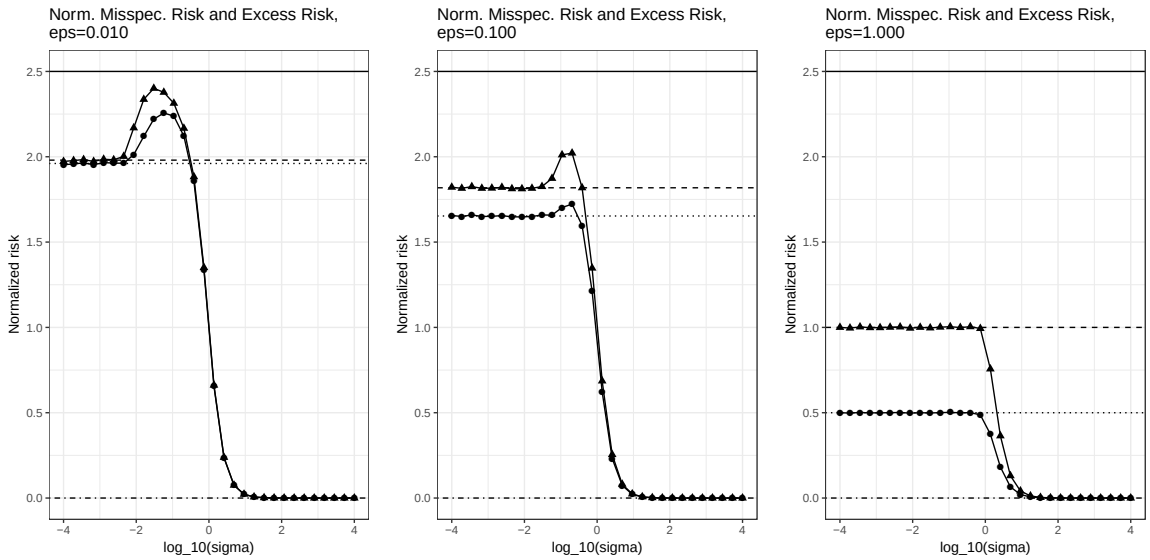


Figure 2: Empirical estimates of the normalized misspecified risk (•) and normalized excess risk (▲) plotted against $\log_{10}(\sigma)$, for the ball $\mathcal{C} = \{\theta \in \mathbb{R}^n : \|\theta\| \le 1\}$ in the case $n = 3$ with $\theta^* = (1+\epsilon, 0, 0)$ and $\epsilon \in \{0.01, 0.1, 1\}$. The solid horizontal line represents the upper bound $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))) = n - \frac{1}{2} = 2.5$ guaranteed by (7). The dotted lines and dashed lines are the predicted low σ limits $\frac{n-1}{(1+\epsilon)^2}$ and $\frac{n-1}{1+\epsilon}$ respectively. The dash-dot line is the high σ limit 0.

5.2. High noise limit

Although not interesting in its own right, the high noise limit of the normalized risks can help characterize the maximum risks (10) as we discuss in Section 5.3 below. Proofs for this section appear in Appendix E.

For a closed convex set $\mathcal{C}$ we define the *core cone*

$$K_{\mathcal{C}} := \bigcap_{\theta \in \mathcal{C}} T_{\mathcal{C}}(\theta). \tag{37}$$

Recall the notation for the re-centered set $F_{\mathcal{C}}(\theta_0) = \{\theta - \theta_0 : \theta \in \mathcal{C}\}$ where $\theta_0 \in \mathcal{C}$. For a vector $v \in \mathbb{R}^n$ we let $\mathbb{R}_+ v := \{\alpha u : \alpha \ge 0\}$. We have the following equivalent characterizations of the core cone.

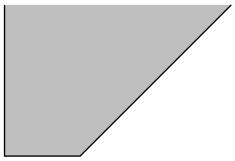
Lemma 5.2 (Characterizations of the core cone)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed convex set. For any $\theta_0 \in \mathcal{C}$,*

$$K_{\mathcal{C}} \overset{(i)}{=} \{v : \mathbb{R}_+ v \subseteq F_{\mathcal{C}}(\theta_0)\} \overset{(ii)}{=} \bigcap_{\sigma>0} \frac{F_{\mathcal{C}}(\theta_0)}{\sigma}.$$

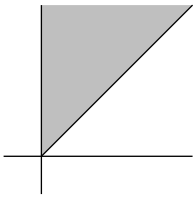
Additionally, the inclusion $K_{\mathcal{C}} \subseteq T_{\mathcal{C}}(\theta)$ holds for any $\theta \in \mathcal{C}$. If furthermore $F_{\mathcal{C}}(\theta_0)$ is a cone, then the equality $K_{\mathcal{C}} = T_{\mathcal{C}}(\theta)$ holds if and only if $\theta_0 - (\theta - \theta_0) \in \mathcal{C}$; in particular, taking $\theta = \theta_0$ shows that $K_{\mathcal{C}} = T_{\mathcal{C}}(\theta_0) = F_{\mathcal{C}}(\theta_0)$.

Thus, up to a translation, the core cone can either be viewed as the result of shrinking $\mathcal{C}$ radially toward $\theta_0 \in \mathcal{C}$, or as the largest cone centered at $\theta_0 \in \mathcal{C}$ that is contained in $\mathcal{C}$. An interesting point is that $\theta_0 \in \mathcal{C}$ can be chosen arbitrarily. An example is given in Figure 3.

Furthermore, in the case when $\mathcal{C}$ is a cone, the core cone $K_{\mathcal{C}}$ simply equals this cone $\mathcal{C}$, and we can characterize which tangent cones are the "smallest" in the sense that they equal the intersection (37) of all tangent cones.



(a) $\mathcal{C} = \{(x, y) : x \geq 0, y \geq \max(x, 1)\}$



(b) $K_{\mathcal{C}} = \{(x, y) : 0 \leq x \leq y\}$

Figure 3: Example of a closed convex set $\mathcal{C}$ and its core cone $K_{\mathcal{C}}$. By the definition (37) and Lemma 5.2, we can equivalently think of the core cone as 1) the intersection of all tangent cones of $\mathcal{C}$, 2) the result of radially shrinking $\mathcal{C}$ toward any given point $\theta_0 \in \mathcal{C}$, or 3) the collection of all rays that lie entirely in $\mathcal{C}$ with a given initial point $\theta_0 \in \mathcal{C}$.

The following result shows that under a boundedness condition, the core cone characterizes both high σ limits.

Proposition 5.3 (High noise limit)**.** *Let $\mathcal{C}$ be a closed convex set. Let $\theta^* \in \mathbb{R}^n$ and $Y := \theta^* + \sigma Z$ where Z is a zero mean random vector with $\mathbb{E}\|Z\|^2 < \infty$. If the condition*

$$\sup_{x \in \mathbb{R}^n} \left(\|\Pi_{F_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))}(x)\|^2 - \|\Pi_{K_{\mathcal{C}}}(x)\|^2 \right) < \infty. \tag{38}$$

holds, then

$$\lim_{\sigma \to \infty} \frac{1}{\sigma^2} M(\hat{\theta}, \theta^*) = \lim_{\sigma \to \infty} \frac{1}{\sigma^2} E(\hat{\theta}, \theta^*) = \delta(K_{\mathcal{C}}).$$

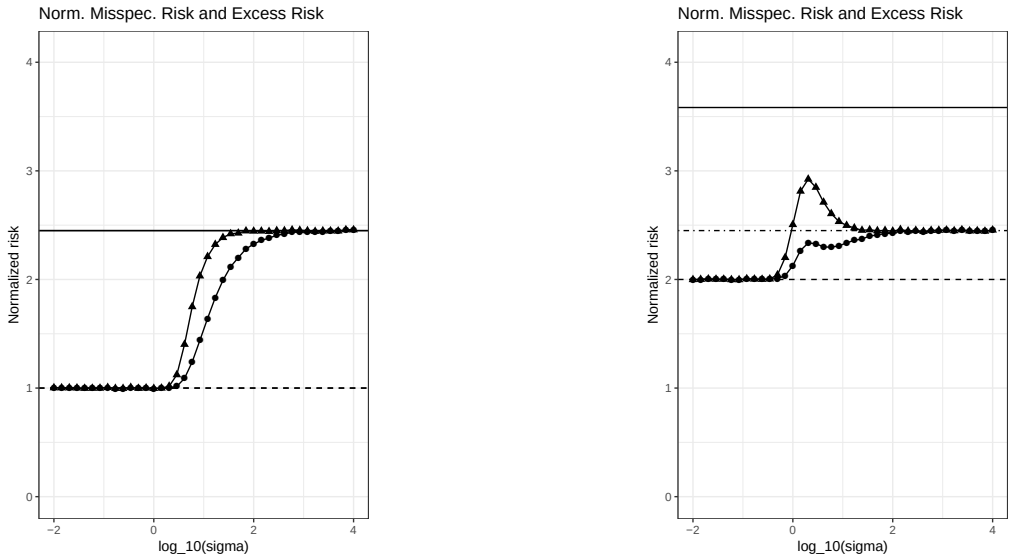
The main hurdle in applying Proposition 5.3 is verifying the condition (38). The following result covers two cases where it is easy to verify the condition.

Corollary 5.4 (Orthant and bounded sets)**.** *Let $\theta^* \in \mathbb{R}^n$ and $Y \sim N(\theta^*, \sigma^2 I_n)$.*

- *If $\mathcal{C} = \mathbb{R}^n_+$ is the nonnegative orthant, then the high σ limits are $\delta(\mathbb{R}^n_+) = n/2$.*
- *Let $\mathcal{C}$ be a closed convex set. $K_{\mathcal{C}} = \{0\}$ if and only if $\mathcal{C}$ is bounded, in which case both high σ limits are* 0*.*

Figure 2 and Figure 5 illustrate the result of this corollary.

Verifying the condition (38) for more general $\mathcal{C}$ is more difficult. We believe it might hold for polyhedral cones with any θ^*, in which case Proposition 5.3 would imply that the high σ limits are $\delta(\mathcal{C})$. For example, we are unable to verify that the monotone cone $\mathcal{C} = \mathcal{S}^n$ satisfies the condition (38), but our simulations in Figure 4 agree with the implication of Proposition 5.3 that the high noise limit is $\delta(\mathcal{S}^n) = H_n$. An interesting feature of the examples presented thus far is that the high σ limits (including the veracity of (38)) do not depend on θ^*.



(a) $\theta^* = (5, 3, 1-1, -3, -5)$ (b) $\theta^* = (0, 0, -2, -2, 3, 1)$

Figure 4: Empirical estimates of the normalized misspecified risk (•) and normalized excess risk (▲) plotted against $\log_{10}(\sigma)$, for the monotone cone $\mathcal{C} = \mathcal{S}^n$ in the case $n = 6$, for two different values of θ^*. The solid horizontal lines represent the upper bound $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$ guaranteed by (7), which are $\delta(\mathcal{S}^6) = H_6 = 2.45$ and $\delta(\mathcal{S}^4 \times \mathcal{S}^2) = H_4 + H_2 = 3.58\overline{3}$ respectively. The dashed lines are the predicted low σ limits, which are 1 and 2 respectively. (See Section 4.2 for details on the computations.) The dash-dot line is the predicted high σ limit $\delta(\mathcal{S}^6) = H_6 = 2.45$ (see discussion below Corollary 5.4).

Remark 5.5. *More generally, suppose $\mathcal{C}$ is a general cone. By applying Lemma 5.2 with $\theta_0 = 0$ and $\theta = \Pi_{\mathcal{C}}(\theta^*)$, we observe that the core cone $K_{\mathcal{C}}$ is $\mathcal{C}$, and moreover $\mathcal{C} \subseteq T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$, with equality if and only if $-\Pi_{\mathcal{C}}(\theta^*) \in \mathcal{C}$. Thus, if the condition* (38) *holds, then Proposition 5.3 implies the high σ limits are $\delta(\mathcal{C})$, and moreover Lemma 5.2 implies that these limits equal Bellec's upper bound* (7)*, $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$, if and only if θ^* satisfies $-\Pi_{\mathcal{C}}(\theta^*) \in \mathcal{C}$.*

However, the condition (38) does not hold for all $\mathcal{C}$. One can verify numerically that the epigraph $\mathcal{C} := \{u \in \mathbb{R}^2 : u_2 \geq u_1^2\}$, whose core cone is $K_{\mathcal{C}} = \{(0, u_2) : u_2 \geq 0\}$, does not satisfy (38). Simulations also show that the high σ limits are larger than $\delta(K_{\mathcal{C}}) = 1/2$. In general, it is unclear exactly when the core cone does or does not characterize the high σ limits.

5.3. Maximum normalized risk

Our low and high σ limit results Theorem 3.1 and Proposition 5.3 provides an incomplete characterization of the maximum normalized risks (10). As mentioned already in (7), $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$ is an upper bound for both suprema.

In the well-specified case $\theta^* \in \mathcal{C}$, both suprema reduce to the usual normalized risk $\sigma^{-2}R(\hat{\theta}, \theta^*)$; moreover the upper bound becomes $\delta(T_{\mathcal{C}}(\theta^*))$, and is attained as $\sigma \downarrow 0$ by the result (4) of Oymak and Hassibi [8].

However, in the misspecified case we have shown in Theorem 3.1 that in general the low σ limit does not attain the upper bound (7). Moreover, simulations show that in some cases even the suprema do not attain the upper bound; see Figure 2 and Figure 5. We see that for some cases the suprema are close to the upper bound, but for others it is much smaller.

Of course, if one can show that either the low σ limit or the high σ limit is equal to the upper bound $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)))$, then we know the upper bound is attained either as $\sigma \downarrow 0$ or $\sigma \to \infty$ respectively. However, in the settings of Theorem 3.1 and Proposition 5.3, this seldom happens. As discussed already, if $\mathcal{C}$ is polyhedral with nonempty interior, then the low σ limit is strictly smaller than the upper bound. If Proposition 5.3

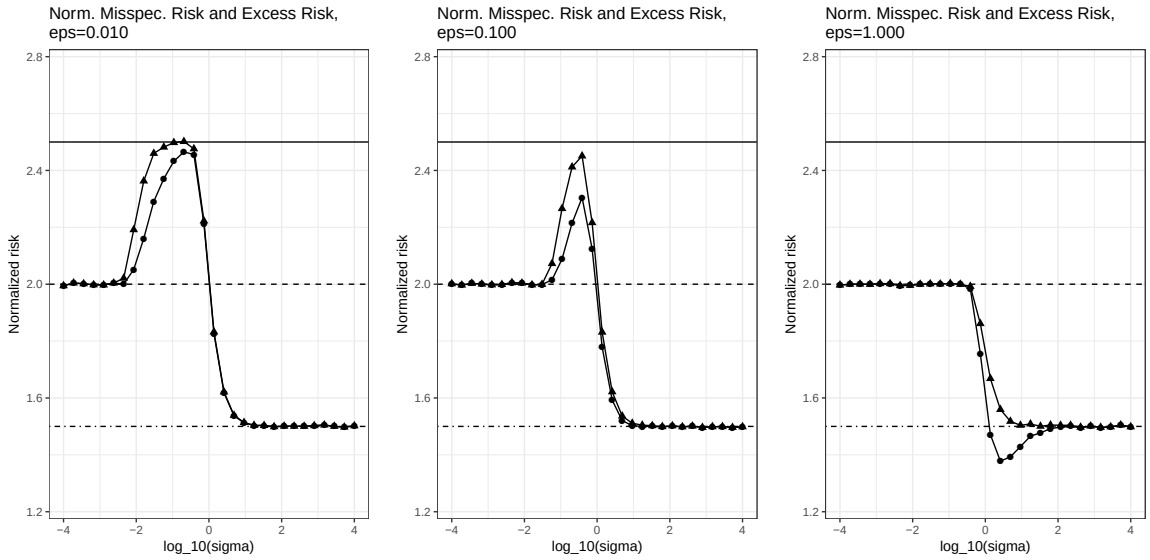


Figure 5: Empirical estimates of the normalized misspecified risk (•) and normalized excess risk (▲) plotted against $\log_{10}(\sigma)$, for the orthant $\mathcal{C} := \mathbb{R}^3_+$ and $\theta^* = (1, 1, -\epsilon)$ with $\epsilon \in \{0.01, 0.1, 1\}$. The solid horizontal line represents the upper bound $\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))) = n - \frac{1}{2}$ guaranteed by (7). The dashed line is the common low σ limit $n - 1$ (see Corollary 4.1). The dash-dot line is the high σ limit $\delta(\mathbb{R}^n_+) = 3/2$.

applies, then $K_{\mathcal{C}} = \bigcap_{\theta \in \mathcal{C}} T_{\mathcal{C}}(\theta) \subseteq T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$ shows that the high σ limit is typically strictly smaller than the upper bound; for the special case where $\mathcal{C}$ is a cone, see Remark 5.5 for a necessary and sufficient condition for the high σ limit to equal the upper bound.

Thus in most cases the suprema are attained at some moderate values of σ, but it is difficult to provide a characterization of these maximizing values σ, as well as the value of the suprema and whether they are close to the upper bound or not. The plots suggest that as θ^* gets closer to $\mathcal{C}$, the suprema get closer to the upper bound as well.

Appendix A: Proofs of lemmas in Section 3

The next lemma is a technical device for representing the largest face of a polyhedral cone that lies in a particular hyperplane. It is useful for proving Lemma 3.3 and Lemma 3.5.

Lemma A.1 (Largest face in hyperplane)**.** *Let $\mathcal{K} = \{u : Au \leq 0\} \subseteq \mathbb{R}^n$ be a polyhedral cone, where $A \in \mathbb{R}^{m \times n}$ has distinct rows. For each $y \in \mathbb{R}^n$, consider the subsets $J \subseteq \{1, \ldots, m\}$ satisfying*

$$\{u : A_J u = 0\} \subseteq (y - \Pi_{\mathcal{K}}(y))^{\perp}. \tag{39}$$

We let J_y denote the smallest such subset.

This subset J_y characterizes a face of $\mathcal{K}$ in the following way.

$$\mathcal{K} \cap (y - \Pi_{\mathcal{K}}(y))^{\perp} = \{u : A_{J_y} u = 0, A_{J_y^c} u \leq 0\}. \tag{40}$$

Proof. The optimality condition for a projection onto a cone (14) implies $\langle y - \Pi_{\mathcal{K}}(y), u \rangle \leq 0$ for all $u \in \mathcal{K}$. If $\mathcal{K}$ contains both u and $-u$, then this implies $u \in (y - \Pi_{\mathcal{K}}(y))^{\perp}$. Thus for $J = \{1, \ldots, m\}$, (39) holds because $\{u : A_J u = 0\} \subseteq \mathcal{K}$. This shows the existence of subsets J that satisfy (39).

Next, note that if J and J' both satisfy (39), then $J \cap J'$ does as well, because

$$\{u : A_{J \cap J'} u = 0\} = \{u + v : A_J u = A_{J'} v = 0\} \subseteq (y - \Pi_{\mathcal{K}}(y))^{\perp}.$$

So, letting J_y be the intersection of all J satisfying (39) yields the unique subset of minimal size.

The $\supseteq$ inclusion in (40) follows immediately from $\{u : A_{J_y}u = 0\} \subseteq (y - \Pi_{\mathcal{K}}(y))^\perp$. For the other inclusion, suppose $v \in \mathcal{K} \cap (y - \Pi_{\mathcal{K}}(y))^\perp$. Then $Av \leq 0$, so it remains to verify $A_{J_y}v = 0$. That is, if $J \subseteq \{1, \dots, m\}$ denotes the indices j for which $\langle a_j, v\rangle = 0$, we want to show $J_y \subseteq J$; furthermore, this reduces to showing J satisfies (39), by minimality of J_y.

Any u satisfying $A_J u = 0$ can be rewritten as $u = v + w$ for some w also satisfying $A_J w = 0$. There exists some $c > 0$ such that both $v + cw$ and $v - cw$ are in $\mathcal{K}$ because all the linear constraints outside of J are strict inequalities at v. Then, the optimality condition for the projection onto a cone, yields $\langle v + cw, y - \Pi_{\mathcal{K}}(y)\rangle \leq 0$ and $\langle v - cw, y - \Pi_{\mathcal{K}}(y)\rangle \leq 0$. Since $v \in (y - \Pi_{\mathcal{K}}(y))^\perp$, this yields $w \in (y - \Pi_{\mathcal{K}}(y))^\perp$ and thus $u \in (y - \Pi_{\mathcal{K}}(y))^\perp$, which verifies that J satisfies (39). □

Proof of Lemma 3.2. By definition there exist an integer m, matrix $A \in \mathbb{R}^{m\times n}$, and vector $b \in \mathbb{R}^m$ such that $\mathcal{C} \coloneqq \{u \in \mathbb{R}^n : Au \leq b\}$. Fix $\theta^* \in \mathbb{R}^n$ and let $\theta_0 \coloneqq \Pi_{\mathcal{C}}(\theta^*)$. We will show

$$T_{\mathcal{C}}(\theta_0) = \{u : A_J u \leq 0\},$$

where $J = \{j : \langle a_j, \theta_0\rangle = b_j\}$. Then $T_{\mathcal{C}}(\theta_0)$ is a polyhedral cone.

If $u \in T_{\mathcal{C}}(\theta_0)$ then for some $r^* > 0$ we have $\theta_0 + r^*u \in \mathcal{C}$. Thus, $b_J \geq A_J(\theta_0 + ru) = b_J + r^*A_J u$ which implies $A_J u \leq 0$.

Conversely, suppose u satisfies $A_J u \leq 0$. Choose $r^* > 0$ so that $r^*\langle a_j, u\rangle \leq b_j - \langle a_j, \theta_0\rangle$ for all $j \notin J$. This is possible because $b_j > \langle a_j, \theta_0\rangle$ for each $j \notin J$. Then $\theta_0 + r^*u \in \mathcal{C}$ so $u \in T_{\mathcal{C}}(\theta_0)$.

Finally, we need to prove the second part of the locally polyhedral condition (15), which will follow if we show $T_{\mathcal{C}}(\theta_0) \cap B_{r^*}(0) \subseteq F_{\mathcal{C}}(\theta_0)$ for some $r^* > 0$. If $u \in T_{\mathcal{C}}(\theta_0)$ then $A_J u \leq 0 = b_J - A_J\theta_0$, so it suffices to find some r^* such that $A_{J^c}u \leq b_{J^c} - A_{J^c}\theta_0$ for any $u \in B_{r^*}(0)$. For each $j \notin J$, we have $\langle a_j, \theta_0\rangle < b_j$ so there exists some $r^* > 0$ such that all $\theta \in B_{r^*}(\theta_0)$ satisfy $\langle a_j, \theta\rangle < b_j$ for all $j \notin J$. Taking $u = \theta - \theta_0$ concludes the proof. □

Proof of Lemma 3.3. Let $\mathcal{T} \coloneqq \{u + \Pi_{\mathcal{C}}(\theta^*) : u \in T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))\}$. Using the locally polyhedral condition (15) and continuity [5] of $\Pi_{\mathcal{C}}$ and $\Pi_{\mathcal{T}}$, we have $\Pi_{\mathcal{T}}(\theta^*) = \Pi_{\mathcal{C}}(\theta^*)$ (e.g., see the verification of (25)), and thus translating yields $\Pi_{T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))}(\theta^* - \Pi_{\mathcal{C}}(\theta^*)) = 0$. Applying Lemma A.1 with $\mathcal{K} = T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$, $y = \theta^* - \Pi_{\mathcal{C}}(\theta^*)$, and $\Pi_{\mathcal{K}}(y) = 0$ concludes the proof. □

Proof of Lemma 3.4. Fix $\theta_0 \in \mathcal{C}$. For any $\theta^* \notin \mathcal{C}$ such that $\Pi_{\mathcal{C}}(\theta^*) = \theta_0$, Lemma 3.2 implies the locally polyhedral condition (15) holds, and thus Lemma 3.3 establishes that $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp$ is a face of the tangent cone $T_{\mathcal{C}}(\theta_0)$.

Since the tangent cone has finitely many faces, the supremum is actually a maximum over the statistical dimensions of finitely many such lower-dimensional faces. Thus it remains to show

$$\delta(T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp) < \delta(T_{\mathcal{C}}(\theta_0))$$

for each $\theta^* \notin \mathcal{C}$ such that $\Pi_{\mathcal{C}}(\theta^*) = \theta_0$.

The set $(\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp$ is a hyperplane (not all of $\mathbb{R}^n$) because $\theta^* \notin \mathcal{C}$. Using the fact that the tangent cone $T_{\mathcal{C}}(\theta_0)$ has nonempty interior (because it contains the translation $F_{\mathcal{C}}(\theta_0)$ of $\mathcal{C}$), we see that the intersection $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^\perp$ is a face that that lies in a strictly lower-dimensional subspace of $\mathbb{R}^n$, and is therefore strictly smaller than the full cone $T_{\mathcal{C}}(\theta_0)$. Thus, we just need to show $\delta(T') < \delta(T)$ for any polyhedral cone T with nonempty interior in $\mathbb{R}^n$, and any face T' of T that lies in a strictly lower-dimensional subspace of $\mathbb{R}^n$.

For a point $x \in \mathbb{R}^n$ and a set $S \subseteq \mathbb{R}^n$ let $d(x, S) \coloneqq \inf_{\theta\in S}\|x - \theta\|$. Note that the Moreau decomposition for cones [2, Sec. B] implies $\|\Pi_{\mathcal{K}}(x)\| = d(x, \mathcal{K}^\circ)$ for any $x \in \mathbb{R}^n$ and any cone $\mathcal{K}$, where $\mathcal{K}^\circ \coloneqq \{u \in \mathbb{R}^n : \langle u, \theta\rangle \leq 0, \forall\theta \in \mathcal{K}\}$ denotes the polar cone of $\mathcal{K}$. Since $T^\circ \subseteq (T')^\circ$, we have

$$d(x, (T')^\circ) \leq d(x, T^\circ), \qquad \forall x \in \mathbb{R}^n.$$

Thus, if we show the random vector Z has nonzero probability of being in the set

$$\mathcal{A} \coloneqq \{x \in \mathbb{R}^n : d(x, (T')^\circ) < d(x, T^\circ)\} = \{x \in \mathbb{R}^n : \|\Pi_{T'}(x)\| < \|\Pi_T(x)\|\},$$

then we immediately have the desired strict inequality

$$\delta(T') = \mathbb{E}d(Z, (T')^\circ) < \mathbb{E}d(Z, T^\circ) = \delta(T).$$

To prove the above claim that $\mathbb{P}(Z \in \mathcal{A}) > 0$, we show below that the interior of T is contained in $\mathcal{A}$; then our assumption on Z will conclude the proof.

Let x be in the interior of T. Then $x \in T \setminus T'$. Moreover, if we let U be the smallest linear subspace of $\mathbb{R}^n$ containing T', then $x \notin U$ as well. Note the the Pythagorean theorem implies

$$\|\Pi_T(x)\|^2 = \|x\|^2 = \|\Pi_U(x)\|^2 + \|x - \Pi_U(x)\|^2 > \|\Pi_U(x)\|^2. \tag{41}$$

We also have

$$\Pi_{T'}(x) = \operatorname*{argmin}_{\theta \in T'} \|\theta - x\|^2 = \operatorname*{argmin}_{\theta \in T'} \big\{ \|\theta - \Pi_U(x)\|^2 + \|\Pi_U(x) - x\|^2 \big\} = \Pi_{T'}(\Pi_U(x)),$$

so combining this with the above inequality (41) and the optimality condition (14) for the projection of $\Pi_U(x)$ onto the cone T', we have

$$\|\Pi_{T'}(x)\|^2 = \|\Pi_{T'}(\Pi_U(x))\|^2 = \|\Pi_U(x)\|^2 - \|\Pi_U(x) - \Pi_{T'}(\Pi_U(x))\|^2 \le \|\Pi_U(x)\|^2 < \|\Pi_T(x)\|^2,$$

and thus $x \in \mathcal{A}$. □

Proof of Lemma 3.5. The lemma holds immediately if $\theta^* \in \mathcal{T}$, so we assume $\theta^* \notin \mathcal{T}$.

By translating, we may without loss of generality assume $\Pi_{\mathcal{T}}(\theta^*) = 0$ so that the cone is centered at 0 and can be written as $\mathcal{T} = \{u : Au \le 0\}$ for some number of constraints m and some matrix $A \in \mathbb{R}^{m \times n}$. The objective then reduces to

$$\Pi_{\mathcal{T}}(y) \in (\theta^*)^\perp, \qquad \text{for all } y \in B_r(\theta^*).$$

For any $y \in \mathbb{R}^n$ let $J_y \subseteq \{1, \dots, m\}$ be as defined in Lemma A.1 for our polyhedral cone $\mathcal{T}$; it characterizes the largest face of $\mathcal{T}$ that lies in $(\theta^*)^\perp$. We claim there exists $r > 0$ such that

$$\{u : A_{J_y} u = 0\} \subseteq (\theta^*)^\perp, \quad \forall y \in B_r(\theta^*). \tag{42}$$

If not, then there exists a sequence of points $y_k \notin \mathcal{T}$ converging to θ^* such that $\{u : A_{J_{y_k}} u = 0\} \not\subseteq (\theta^*)^\perp$ for all k. Since there are finitely many distinct subsets J_{y_k}, we may take a subsequence and without loss of generality assume it is common subset $J = J_{y_k}$ for all k, and $\{u : A_J u = 0\} \not\subseteq (\theta^*)^\perp$. By the definition (39) of J_{y_k}, any u satisfying $A_J u = 0$ also satisfies $\langle y_k - \Pi_{\mathcal{T}}(y_k), u\rangle = 0$. By continuity of $\Pi_{\mathcal{T}}$ and taking $k \to \infty$, we have $\langle \theta^*, u\rangle = 0$ as well, a contradiction.

Finally, since the optimality condition (11) for $\Pi_{\mathcal{T}}$ implies $\langle \Pi_{\mathcal{T}}(y), y - \Pi_{\mathcal{T}}(y)\rangle = 0$ for any $y \in \mathbb{R}^n$, (40) implies $\Pi_{\mathcal{T}}(y) \in \{u : A_{J_y} u = 0\}$. Combining this with (42) concludes the proof. □

Appendix B: Proofs for Section 4.2 (isotonic regression)

B.1. Proofs of block monotone cone lemmas

Proof of Lemma 4.4. The first claim follows from decomposing the squared Euclidean distance into blocks.

$$\begin{aligned}
\min_{v \in \mathcal{S}_{|I_1|,\dots,|I_m|}} \|v - z\|^2 &= \min_{x \in \mathcal{S}^m} \sum_{j=1}^m \sum_{i \in I_j} (x_j - z_i)^2 \\
&= \min_{x \in \mathcal{S}^m} \sum_{j=1}^m \sum_{i \in I_j} ((x_j - \bar{z}_{I_j})^2 + (\bar{z}_{I_j} - y_i)^2) \\
&= \sum_{j=1}^m \sum_{i \in I_j} (z_i - \bar{z}_{I_j})^2 + \min_{x \in \mathcal{S}^m} \sum_{j=1}^m |I_j| (x_j - \bar{z}_{I_j})^2.
\end{aligned}$$

Let Z and Z' be standard Gaussian in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. If $|I_1| = \cdots = |I_m| = r$, then the first claim implies

$$\delta(\mathcal{S}_{|I_1|,\ldots,|I_m|}) \coloneqq \mathbb{E}\|\Pi_{\mathcal{S}_{|I_1|,\ldots,|I_m|}}(Z)\|^2 \overset{(i)}{=} r\mathbb{E}\|\Pi_{\mathcal{S}^m}(Z'/\sqrt{r})\|^2 \overset{(ii)}{=} \mathbb{E}\|\Pi_{\mathcal{S}^m}(Z')\|^2 =: \delta(\mathcal{S}^m) = \sum_{j=1}^{m} \frac{1}{j},$$

where (i) is due to $Z'/\sqrt{r} \overset{d}{=} (\bar{Z}_{I_1}, \ldots, \bar{Z}_{I_m})$, and (ii) is due to $\Pi_{\mathcal{C}}(cx) = c\Pi_{\mathcal{C}}(x)$ for a cone $\mathcal{C}$ and $c > 0$ (e.g., [3, Sec. 1.6]). The statistical dimension of $\mathcal{S}^m$ is proved by Amelunxen et al. [2, Sec. D.4]. □

Proof of Lemma 4.5. We use two useful properties of the statistical dimension of any cone $\mathcal{C}$ [2, Prop. 3.1].

- Rotational invariance: for any orthogonal transformation Q, we have $\delta(Q\mathcal{C}) = \delta(\mathcal{C})$.
- Invariance under embedding: $\delta(\mathcal{C} \times \{0\}^k) = \delta(\mathcal{C})$.

Thus it suffices to provide an orthogonal transformation Q such that $Q\mathcal{S}_{|I_1|,\ldots,|I_m|}$ is an embedding of the cone (31) into $\mathbb{R}^n$.

Let e_i denote the ith standard basis vector in $\mathbb{R}^n$. Let the last element of each block be denoted $k_j \coloneqq \max I_j$ for $1 \le j \le m$, with $k_0 = 0$ for convenience. The block monotone cone $\mathcal{S}_{|I_1|,\ldots,|I_m|}$ is defined by the following constraints for $u \in \mathbb{R}^n$.

$$\langle e_i - e_{i+1}, u\rangle \le 0, \qquad i \in \{k_1, \ldots, k_m\} \tag{43a}$$

$$\langle e_i - e_{i+1}, u\rangle = 0, \qquad i \in \{1, \ldots, n-1\} \setminus \{k_1, \ldots, k_m\} \tag{43b}$$

Let us focus on an arbitrary block I_j. Consider the $|I_j| \times |I_j|$ matrix

$$\tilde{A}_j = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}$$

Because $\tilde{A}_j$ is full rank, the QR decomposition implies there exists an $|I_j| \times |I_j|$ orthogonal matrix $\tilde{Q}_j$ such that $\tilde{R}_j \coloneqq \tilde{Q}_j\tilde{A}_j$ is upper triangular with positive diagonal entries, and this decomposition is unique.

The block diagonal matrix Q with blocks $\tilde{Q}_1, \ldots, \tilde{Q}_m$ is an $n \times n$ orthogonal matrix. Let A and R also be block diagonal, each constructed similarly using the $\tilde{A}_j$ and the $\tilde{R}_j$ respectively, so that $U = QA$. We consider $Q\mathcal{S}_{|I_1|,\ldots,|I_m|}$. We use the fact that if $v = Qu$ then $\langle b, u\rangle \le 0 \iff \langle Qb, v\rangle \le 0$ to rewrite the constraints (43a) and (43b). The following hold for each $j = 1, \ldots, m$.

- Note that the ith column of A is $a_i = e_i - e_{i+1}$ when $k_{j-1} < i < k_j$. For these i, the equality constraints (43b) after the transformation become $0 = \langle Q(e_i - e_{i+1}), v\rangle = \langle r_i, v\rangle$ where r_i is the ith column of R. Since $\tilde{R}_j$ is upper triangular with nonzero diagonal entries (because $\tilde{A}_j$ is full rank), induction on $i = k_{j-1} + 1, \ldots, k_j - 1$ implies
$$v_i = 0, \qquad k_{j-1} < i < k_j.$$
- When $j < m$, we have $e_{k_j} = a_k$ and $e_{k_j+1} = a_{k_j} + a_{k_j+1} + \cdots + a_{k_{j+1}}$ Thus for $j < m$ the inequality constraint $\langle e_{k_j} - e_{k_j+1}, u\rangle \le 0$ becomes
$$0 \ge \langle Q(e_{k_j} - e_{k_j+1}), v\rangle = \langle r_{k_j} - r_{k_j+1} - r_{k_j+2} - \cdots - r_{k_{j+1}}, v\rangle = \langle r_{k_j} - r_{k_{j+1}}, v\rangle,$$
where the last equality is due to $\langle r_i, v\rangle = 0$ for $k_j < i < k_{j+1}$, by the previous point. Since $\tilde{R}_j$ and $\tilde{R}_{j+1}$ are each upper triangular, the inequality reduces to $r_{k_j,k_j}v_{k_j} \le r_{k_{j+1},k_{j+1}}v_{k_{j+1}}$, where $r_{k,k}$ denotes the kth diagonal entry of R. Lemma B.1 (proved below) computes these diagonal elements and yields
$$\frac{v_{k_j}}{\sqrt{|I_j|}} \le \frac{v_{k_{j+1}}}{\sqrt{|I_{j+1}|}}.$$

Therefore we have shown that $Q\mathcal{S}_{|I_1|,\ldots,|I_m|}$ consists of all vectors satisfying

$$\frac{v_{k_1}}{\sqrt{|I_1|}} \le \frac{v_{k_2}}{\sqrt{|I_2|}} \le \cdots \le \frac{v_{k_m}}{\sqrt{|I_m|}}, \text{ and } v_i = 0, \forall i \in \{1,\ldots,n\} \setminus \{k_1,\ldots,k_m\}.$$

We have thus verified the claim that $Q\mathcal{S}_{|I_1|,\ldots,|I_m|}$ is an embedding of (31) into $\mathbb{R}^n$.

When the blocks all have equal size r, the cone (31) becomes the monotone cone $\mathcal{S}^m$, whose statistical dimension is $\sum_{j=1}^m \frac{1}{j}$ [2, Sec. D.4]. □

Lemma B.1. *Consider the $n \times n$ matrix*

$$A = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}.$$

There exists a unique orthogonal matrix Q and a unique upper triangular matrix R with positive diagonal entries such that $A = QR$. The bottom-right entry of R is $r_{n,n} = 1/\sqrt{n}$.

Proof. Let q_i be the ith column of Q. The last column q_n is orthogonal to the span of the first $n-1$ columns of A, so q_n is either $(1,\ldots,1)/\sqrt{n}$ or its negative. The positivity constraint on the diagonal entries of R implies the former, and thus $r_{n,n} = \langle q_n, e_n \rangle = 1/\sqrt{n}$. □

B.2. Statistical dimension of the block monotone cone in general

In Lemma 4.5 we provided an expression for the statistical dimension of the block monotone cone $\mathcal{S}_{|I_1|,\ldots,|I_m|}$ when the block sizes were equal. In general, the statistical dimension can be higher or lower than $\sum_{j=1}^m \frac{1}{j}$. Consider the following examples for $m = 3$.

Lemma 4.5 implies $\mathcal{S}_{n-2,1,1}$ has the same statistical dimension as $\{v \in \mathbb{R}^3 : v_1/\sqrt{n-2} \le v_2 \le v_3\}$. As $n \to \infty$ this latter cone approaches $\{v \in \mathbb{R}^3 : 0 \le v_2 \le v_3\}$ which has statistical dimension $1 + \left(\frac{1}{8} \cdot 2 + \frac{1}{2} \cdot 1\right) = \frac{7}{4} = 1.75$, which is smaller than $\sum_{j=1}^3 \frac{1}{j} = \frac{11}{6} = 1.8\bar{3}$.

On the other hand, $\mathcal{S}_{1,n-2,1}$ has the same statistical dimension as $\{v \in \mathbb{R}^3 : v_1 \le v_2/\sqrt{n-2} \le v_3\}$. As $n \to \infty$ this latter cone approaches $\{v \in \mathbb{R}^3 : v_1 \le 0, v_3 \ge 0\}$ which has statistical dimension $1 + \frac{1}{2} + \frac{1}{2} = 2$, which is larger than $1.8\bar{3}$.

We suspect that the approach used to prove the statistical dimension of $\mathcal{S}^n$ [2, Sec. D.4], which uses the theory of finite reflection groups, cannot be generalized for $\mathcal{S}_{|I_1|,\ldots,|I_m|}$, due to the asymmetry of (31). However, using a result of Klivans and Swartz [7], it is possible to show that the average statistical dimension among all block monotone cones with a given [unordered] set of m block sizes is H_m [1, Prop. 6.6].

B.3. Proof of Proposition 4.3

When applying Theorem 3.1, it is useful to characterize $\mathcal{S}^n$ and its tangent cones using conic generators. If $T \subseteq \mathbb{R}^n$ is a cone and there exist $x_1,\ldots,x_p \in T$ such that

$$T = \left\{ \sum_{i=1}^p \alpha_i x_i : \alpha_i \ge 0, \forall i \right\},$$

then we call $x_1,\ldots,x_p$ the *conic generators* of T, and write

$$T = \text{cone}\{x_1,\ldots,x_p\}.$$

Lemma B.2. *Let $\theta^* \in \mathbb{R}^n$ and let $\mathcal{C} \subseteq \mathbb{R}^n$ be closed and convex. If the tangent cone $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$ is generated by $x_1, \dots, x_p \in \mathbb{R}^n$, i.e. $T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) = \operatorname{cone}\{x_1, \dots, x_p\}$, then*

$$T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp} = \operatorname{cone}(\{x_1, \dots, x_p\} \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}).$$

Proof of Lemma B.2. The inclusion $\supset$ is immediate, so it remains to prove the inclusion $\subseteq$. Note that the optimality condition (11) implies $\langle \theta^* - \Pi_{\mathcal{C}}(\theta^*), x\rangle \leq 0$ for any $x \in T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$. In particular, if $v \in T_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$, then v can be written as the conical combination $v = \sum_{i=1}^p \alpha_i x_i$ with $\alpha_i \geq 0$, and we have

$$0 = \langle \theta^* - \Pi_{\mathcal{C}}(\theta^*), v\rangle = \sum_{i=1}^{p} \alpha_i \underbrace{\langle \theta^* - \Pi_{\mathcal{C}}(\theta^*), x_i\rangle}_{\leq 0}.$$

Thus, if a generator x_i is not in the hyperplane $(\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$, then $\alpha_i = 0$, so x_i does not contribute in the conical combination of v. Thus, v can be written as a conical combination of generators in $(\theta^* - \Pi_{\mathcal{C}}(\theta^*))^{\perp}$. □

We are now ready to prove Proposition 4.3.

Proof of Proposition 4.3. By Theorem 3.1, it suffices to prove that the statistical dimension term is $\sum_{k=1}^K \delta\Big(\mathcal{S}_{|I_1^k|,\dots,|I_{m_k}^k|}\Big)$.

For $p \geq 1$ let

$$M_p \coloneqq \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & 1 & \cdots & 1 \\ & 1 & \cdots & 1 \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix} \in \mathbb{R}^{(p+1)\times p}.$$

The rows of M_p are the conic generators of $\mathcal{S}^p$.

Suppose first that $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant, so that $K = 1$ and $J_1 = \{1, \dots, n\}$. Then $\Pi_{\mathcal{S}^n}(\theta^*) = (\mu_1, \mu_1, \dots, \mu_1)$ where $\mu_1 \coloneqq \frac{1}{n}\sum_{i=1}^n \theta_i^*$; this follows directly by minimizing $\sum_{i=1}^n (\theta_i^* - \mu_1)^2$ with respect to μ_1.

The finest partition $(I_1^1, \dots, I_{m_1}^1)$ of J_1 into blocks satisfying (29) can be constructed greedily as follows. Begin populating I_1^1 with the elements of $\{1, \dots, n\}$ in order, stopping as soon as the mean of the elements of I_1^1 is μ_1. Then begin populating I_2^1 with the remaining elements in order, again stopping when the mean of the elements in I_2^1 is μ_1. Continue in this manner until the last element n is placed in a subset $I_{m_1}^1$. The mean of the elements of this last subset $I_{m_1}^1$ is μ_1 as well, since the mean of all components of θ^* is μ_1. Thus this partition satisfies (29). To establish uniqueness, note that if some other partition of J_1 satisfies (29), then our partition $(I_1^1, \dots, I_{m_1}^1)$ must be a refinement, due to the greedy construction.

Because $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant, the tangent cone there is $T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) = \mathcal{S}^n$ [3, Prop. 3.1], which is generated by the rows of M_n. In order to use Lemma B.2, we need to determine which rows of M_n are in the hyperplane $(\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^{\perp}$. We already know the mean of the components of $\theta^* - \Pi_{\mathcal{S}^n}(\theta^*)$ is zero, so the first two rows are in the hyperplane.

We claim that exactly $m_1 - 1$ of the remaining $n - 1$ rows of M_n also lie in the hyperplane. Explicitly, if $(I_1^1, \dots, I_{m_1}^1)$ is without loss of generality assumed to be sorted in increasing order, then the remaining rows of M_n that lie in the hyperplane are the indicator vectors for

$$\bigcup_{j=u}^{m_k} I_j^1, \qquad 2 \leq u \leq m_k. \tag{44}$$

No other rows of M_n can be in the hyperplane, else there would exist a finer partition of J_1.

So, Lemma B.2 implies $T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^{\perp}$ is the cone generated by $(-1, \dots, -1)$, $(1, \dots, 1)$, and the indicator vectors of the subsets (44), otherwise known as the cone of nondecreasing vectors that are piecewise constant on the blocks $I_1^1, \dots, I_{m_1}^1$. Its statistical dimension is denoted by $\delta(\mathcal{S}_{|I_1^1|,\dots,|I_{m_1}^1|})$. This concludes the proof in the case when $\Pi_{\mathcal{S}^n}(\theta^*)$ is constant.

We now turn to the general case where $\Pi_{\mathcal{S}^n}(\theta^*)$ is piecewise constant with values $\mu_1 < \cdots < \mu_K$ on $J_1, \ldots, J_K$ respectively. We claim

$$\mu_k = \frac{1}{|J_k|} \sum_{i \in J_k} \theta_i^*. \tag{45}$$

Since $\mathcal{S}^n$ is a cone, the projection satisfies $\langle \theta^* - \Pi_{\mathcal{S}^n}(\theta^*), x \rangle \le 0$ for all $x \in \mathcal{S}^n$, with equality if $x = \Pi_{\mathcal{C}}(\theta^*)$ (e.g., [3, Sec. 1.6]). Letting $x_1, \ldots, x_{n+1}$ be the conic generators of $\mathcal{S}^n$ (the rows of M_n), we have $\Pi_{\mathcal{C}}(\theta^*) = \sum_{i=1}^{n+1} \alpha_i x_i$ for some coefficients $\alpha_i \ge 0$. Then,

$$0 = \langle \theta^* - \Pi_{\mathcal{S}^n}(\theta^*), \Pi_{\mathcal{C}}(\theta^*) \rangle = \sum_{i=1}^{n+1} \alpha_i \underbrace{\langle \theta^* - \Pi_{\mathcal{S}^n}(\theta^*), x_i \rangle}_{\le 0},$$

which implies $\langle \theta^* - \Pi_{\mathcal{S}^n}(\theta^*), x_i \rangle = 0$ if $\alpha_i > 0$. Consequently, if $\Pi_{\mathcal{S}^n}(\theta^*)$ changes value from component $j-1$ to j, then $\sum_{i=j}^{n} [\theta_i^* - (\Pi_{\mathcal{S}^n}(\theta^*))_i] = 0$. Thus (45) holds.

By Proposition 3.1 of [3], the tangent cone is

$$T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) = \mathcal{S}^{n_1} \times \cdots \times \mathcal{S}^{n_K},$$

which is generated by the rows of the block diagonal matrix

$$A \coloneqq \begin{bmatrix} M_{n_1} & & \\ & \ddots & \\ & & M_{n_K} \end{bmatrix}.$$

To find which rows of A are in the hyperplane $(\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^\perp$, we can treat each block M_{n_k} separately and repeat the above argument. Doing so shows that $T_{\mathcal{S}^n}(\Pi_{\mathcal{S}^n}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{S}^n}(\theta^*))^\perp$ is the cone of vectors that are piecewise constant on $(I_1^1, \ldots, I_{m_1}^1, \ldots, I_1^K, \ldots, I_{m_K}^K)$ and are increasing within each of the blocks $(J_1, \ldots, J_K)$. The statistical dimension of this cone is $\sum_{k=1}^{K} \delta(\mathcal{S}_{|I_1^k|, \ldots, |I_{m_k}^k|})$. □

Appendix C: Proofs for Section 4.3 (convex regression)

Recall we fix a design $x_1 < \cdots < x_n$ and use the shorthand $x \coloneqq (x_1, \ldots, x_n)$.

We collect a few technical lemmas before proving Proposition 4.7.

Lemma C.1 (Conic generators for $\mathcal{K}_x$)**.** *A set of conic generators for $\mathcal{K}_x$ are $\pm(1, \ldots, 1)$, $\pm x$, and $g^{(2)}, \ldots, g^{(n-1)}$ where $g_i^{(j)} \coloneqq \max\{x_i - x_j, 0\}$.*

Proof. If $\theta \in \mathcal{K}_x$ then

$$\theta = \left(\theta_2 - \frac{\theta_2 - \theta_1}{x_2 - x_1} x_2\right)(1, \ldots, 1) + \frac{\theta_2 - \theta_1}{x_2 - x_1} x + \sum_{j=2}^{n-1} \left(\frac{\theta_{i+1} - \theta_i}{x_{i+1} - x_i} - \frac{\theta_i - \theta_{i-1}}{x_i - x_{i-1}}\right) g^{(j)}.$$

Note that the coefficients for the $g^{(j)}$ are nonnegative by the definition of $\mathcal{K}_x$. □

Lemma C.2. *If $\theta_0 \in \operatorname{span}\{(1, \ldots, 1), x\}$, then $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta_0)) = \mathcal{K}_x$.*

Proof. If $\theta \in \mathcal{K}_x$, then for any $\alpha > 0$, we have $\alpha(\theta - \theta_0) \in \mathcal{K}_x$, since subtracting an affine function from a convex function yields a convex function, and scaling by a positive constant does not affect convexity. Since $\mathcal{K}_x$ is closed, we have $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta_0)) \subseteq \mathcal{K}_x$.

Conversely, if $v \in \mathcal{K}_x$, then $\theta_0 + v \in \mathcal{K}_x$ since adding an affine function to a convex function yields a convex function. Thus $v \in T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta_0))$ □

Lemma C.3. *Suppose $\theta^* \in -\mathcal{K}_x$ is concave. Then $\Pi_{\mathcal{K}_x}(\theta^*)$ is the projection of θ^* on* $\mathrm{span}\{(1,\dots,1), x\}$*, i.e. it is obtained by a simple linear regression of θ^* on x.*

Proof. By definition, there exists a concave function $f : [x_1, x_n] \to \mathbb{R}$ such that $\theta_i^* = f(x_i)$ for all i. For any $\theta \in \mathcal{K}_x$, there exists a convex function $g : [x_1, x_n] \to \mathbb{R}$ such that $\theta_i = g(x_i)$ for all i.

If g is not affine on $[x_1, x_n]$, then it intersects f in exactly 0, 1, or 2 points. In any of these cases, there exists an affine function h such that $\min\{f, g\} \le h \le \max\{f, g\}$ on $[x_1, x_2]$. (One can note that such an h must touch all intersection points of f and g, if there are any.)

If we let θ' be defined by $\theta'_i := h(x_i)$, then by construction we have $\|\theta' - \theta^*\|^2 < \|\theta - \theta^*\|^2$. (The strict inequality appears since equality would only occur if f and g are both affine, but we assumed g is not affine.) Therefore, $\Pi_{\mathcal{K}_x}(\theta^*)$ must be linear, and the result of the lemma follows immediately. □

We are now ready to prove the proposition.

Proof of Proposition 4.7. By Lemma C.3, $\Pi_{\mathcal{K}_x}(\theta^*)$ is obtained by a linear regression of θ^* on x. Thus $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*)) = \mathcal{K}_x$ by Lemma C.2.

By Lemma B.2, the intersection $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{K}_x}(\theta^*))^\perp = \mathcal{K}_x \cap (\theta^* - \Pi_{\mathcal{K}_x}(\theta^*))^\perp$ is the cone generated by the conic generators of $\mathcal{K}_x$ that are orthogonal to $\theta^* - \Pi_{\mathcal{K}_x}(\theta^*)$.

We already know $\pm(1,\dots,1)$ and $\pm x$ are orthogonal to $\theta^* - \Pi_{\mathcal{K}_x}(\theta^*)$, by Lemma C.3. We claim each $g^{(j)}$ for $j = 2, \dots, n-2$ is not orthogonal to $\theta^* - \Pi_{\mathcal{K}_x}(\theta^*)$. Otherwise, we would have $\langle g^{(j)} - \Pi_{\mathcal{K}_x}(\theta^*), \theta^* - \Pi_{\mathcal{K}_x}(\theta^*)\rangle = 0$ and then

$$\|g^{(j)} - \theta^*\|^2 = \|g^{(j)} - \Pi_{\mathcal{K}_x}(\theta^*)\|^2 + \|\Pi_{\mathcal{K}_x}(\theta^*) - \theta^*\|^2 < \|\Pi_{\mathcal{K}_x}(\theta^*) - \theta^*\|^2,$$

where $g^{(j)} \neq \Pi_{\mathcal{K}_x}(\theta^*)$ by Lemma C.3. This contradicts the optimality of $\Pi_{\mathcal{K}_x}(\theta^*)$.

As mentioned above, we therefore have $T_{\mathcal{K}_x}(\Pi_{\mathcal{K}_x}(\theta^*)) \cap (\theta^* - \Pi_{\mathcal{K}_x}(\theta^*))^\perp = \mathrm{span}\{(1,\dots,1), x\}$, which has [statistical] dimension 2. Applying Theorem 3.1 concludes the proof of the proposition.

□

Appendix D: Proof of Proposition 5.1

Let $r := \|\theta^*\|$. By rotating the problem, we may without loss of generality assume $\theta^* = (r, 0, \dots, 0)$.

Let $E := \{Y \in B_{(r-1)/2}(\theta^*)\}$. Then we have $E \subseteq \{Y \notin \mathcal{C}\}$, so under the event E we have $\hat{\theta}(Y) = Y/\|Y\|$. Noting $\|Y\|^2 = \|\theta^* + \sigma Z\|^2 = r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2$, we have

$$\frac{1}{\sigma^2}\|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 = \frac{1}{\sigma^2}\left(\frac{r + \sigma Z_1}{\sqrt{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2}} - 1\right)^2 + \frac{\sum_{i=2}^n Z_i^2}{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2}.$$

The second term converges to $r^{-2}\sum_{i=2}^n Z_i^2$ as $\sigma \downarrow 0$. We show the first term vanishes as $\sigma \downarrow 0$. Defining $g(\sigma) := \|\theta^* + \sigma Z\|$, we have

$$\begin{aligned} g(\sigma) &= \sqrt{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2} \\ g'(\sigma) &= \frac{rZ_1 + \sigma\|Z\|^2}{g(\sigma)} \\ g''(\sigma) &= \frac{\|Z\|^2}{g(\sigma)} - \frac{(rZ_1 + \sigma\|Z\|^2)g'(\sigma)}{g(\sigma)^2} \end{aligned}$$

Moreover we have $g(0) = r$, $g'(0) = Z_1$, and $g''(0) = (\|Z\|^2 - Z_1^2)/r$. Then by L'Hôpital's rule,

$$\begin{aligned} &\lim_{\sigma\downarrow 0}\frac{1}{\sigma}\left(\frac{r + \sigma Z_1}{\sqrt{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2}} - 1\right) \\ &= \lim_{\sigma\downarrow 0}\frac{r + \sigma Z_1 - g(\sigma)}{\sigma g(\sigma)} = \lim_{\sigma\downarrow 0}\frac{Z_1 - g'(\sigma)}{g(\sigma) + \sigma g'(\sigma)} = \frac{Z_1 - Z_1}{r + 0} = 0. \end{aligned}$$

Note $\mathbf{1}_E \to 1$ almost surely as $\sigma \downarrow 0$. Thus, $\sigma^{-2}\|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \mathbf{1}_E \to r^{-2}\sum_{i=2}^n Z_i^2$ almost surely. By the upper bound (7) we may use the dominated convergence theorem to get

$$\lim_{\sigma\downarrow 0} \frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\Big[\|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \mathbf{1}_E\Big] = \frac{1}{r^2}\sum_{i=2}^n \mathbb{E}Z_i^2 = \frac{n-1}{r^2}.$$

To conclude the proof of the first limit (36a), note that

$$\lim_{\sigma\downarrow 0} \frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\Big[\|\hat{\theta}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \mathbf{1}_{E^c}\Big] = 0,$$

which holds by the argument used in the proof of Theorem 3.1 (e.g., see the second term in (20)).

A similar proof holds for the second limit (36b). Let E and $g(\sigma)$ be the same as before. Then

$$\begin{aligned}
&\frac{1}{\sigma^2}\Big(\|\hat{\theta}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2\Big) \\
&= \frac{1}{\sigma^2}\left(\frac{r + \sigma Z_1}{\sqrt{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2}} - r\right)^2 + \frac{\sum_{i=2}^n Z_i^2}{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2} - \frac{(r-1)^2}{\sigma^2} \\
&= \frac{1}{\sigma^2}\left[\left(\frac{r+\sigma Z_1}{g(\sigma)} - r\right)^2 - (r-1)^2\right] + \frac{\sum_{i=2}^n Z_i^2}{r^2 + 2\sigma r Z_1 + \sigma^2\|Z\|^2}.
\end{aligned}$$

Again, the second term tends to $r^{-2}\sum_{i=2}^n Z_i^2$ as $\sigma \downarrow 0$. To handle the first term we use L'Hôpital's rule again. Let

$$\begin{aligned}
h(\sigma) &:= \frac{r+\sigma Z_1}{g(\sigma)} - r \\
h'(\sigma) &= \frac{Z_1}{g(\sigma)} - \frac{(r+\sigma Z_1)g'(\sigma)}{g(\sigma)^2} \\
h''(\sigma) &= -\frac{Z_1 g'(\sigma)}{g(\sigma)^2} + 2\frac{(r+\sigma Z_1)g'(\sigma)^2}{g(\sigma)^3} - \frac{Z_1 g'(\sigma) + (r+\sigma Z_1)g''(\sigma)}{g(\sigma)^2}
\end{aligned}$$

Recalling the limits $g(0) = r$, $g'(0) = Z_1$, and $g''(0) = (\|Z\|^2 - Z_1^2)/r$, we have $h(\sigma) \to -(r-1)$, $h'(\sigma) \to 0$, and

$$h''(0) = -\frac{Z_1^2}{r^2} + 2\frac{rZ_1^2}{r^3} - \frac{Z_1^2 + \|Z\|^2 - Z_1^2}{r^2} = \frac{Z_1^2 - \|Z\|^2}{r^2}.$$

Then, L'Hôpital's rule allows us to compute the limit of the first term.

$$\begin{aligned}
&\lim_{\sigma\downarrow 0}\frac{1}{\sigma^2}\left[\left(\frac{r+\sigma Z_1}{g(\sigma)} - r\right)^2 - (r-1)^2\right] \\
&= \lim_{\sigma\downarrow 0}\frac{h(\sigma)^2 - (r-1)^2}{\sigma^2} = \lim_{\sigma\downarrow 0}\frac{h(\sigma)h'(\sigma)}{\sigma} = \lim_{\sigma\downarrow 0}(h'(\sigma)^2 + h(\sigma)h''(\sigma)) = \frac{(r-1)(\|Z\|^2 - Z_1^2)}{r^2}.
\end{aligned}$$

Combining terms yields

$$\frac{1}{\sigma^2}\Big(\|\hat{\theta}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2\Big)\mathbf{1}_E \to \frac{(r-1)(\|Z\|^2 - Z_1^2) + \sum_{i=2}^n Z_i^2}{r^2} = \frac{\sum_{i=2}^n Z_i^2}{r},$$

so again by dominated convergence with the upper bound (7), we have

$$\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\Big[\Big(\|\hat{\theta}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2\Big)\mathbf{1}_E\Big] \to \frac{n-1}{r}.$$

To conclude the proof of (36b), note that

$$\frac{1}{\sigma^2}\mathbb{E}_{\theta^*}\Big[\Big(\|\hat{\theta}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2\Big)\mathbf{1}_{E^c}\Big] \to 0,$$

which was proved in the proof of Theorem 3.1 (see (26)).

Appendix E: Proofs for Section 5.2

The following lemma shows that the left-hand side of (38) is nonnegative.

Lemma E.1. *For any $\theta_0 \in \mathcal{C}$,*

$$\|\Pi_{F_{\mathcal{C}}(\theta_0)}(x)\|^2 \geq \|\Pi_{K_{\mathcal{C}}}(x)\|^2.$$

Proof of Lemma E.1. Because $K_{\mathcal{C}}$ is a cone, we have $\langle x, \Pi_{K_{\mathcal{C}}}(x)\rangle = \|\Pi_{K_{\mathcal{C}}}(x)\|^2$. Since $K_{\mathcal{C}} \subseteq F_{\mathcal{C}}(\theta_0)$, the optimality condition for $\Pi_{F_{\mathcal{C}}(\theta_0)}(x)$ implies $\langle x - \Pi_{F_{\mathcal{C}}(\theta_0)}(x), \Pi_{K_{\mathcal{C}}}(x)\rangle \leq 0$ and thus

$$\|\Pi_{K_{\mathcal{C}}}(x)\|^2 \leq \langle \Pi_{F_{\mathcal{C}}(\theta_0)}(x), \Pi_{K_{\mathcal{C}}}(x)\rangle \leq \|\Pi_{F_{\mathcal{C}}(\theta_0)}(x)\|\|\Pi_{K_{\mathcal{C}}}(x)\|.$$

Thus $\|\Pi_{F_{\mathcal{C}}(\theta_0)}(x)\| \geq \|\Pi_{K_{\mathcal{C}}}(x)\|$ and $M_{\theta_0} \geq 0$. □

Proof of Lemma 5.2. We first prove the equalities (i) and (ii).

(i) Let $v \in \{u : \mathbb{R}_+ u \subseteq F_{\mathcal{C}}(\theta_0)\}$ and let $\theta \in \mathcal{C}$. For any $c > 0$ we have $\theta_0 + cv \in \mathcal{C}$, and convexity implies $\theta + \alpha(\theta_0 + cv - \theta) \in \mathcal{C}$ for all $\alpha \in [0,1]$. For large c we have $\|\theta_0 + cv - \theta\| > 1$ and thus $\theta + \frac{\theta_0 + cv - \theta}{\|\theta_0 + cv - \theta\|} \in \mathcal{C}$. Taking $c \to \infty$ and using the fact that $\mathcal{C}$ is closed yields $\theta + \frac{v}{\|v\|} \in \mathcal{C}$ and thus $v \in T_{\mathcal{C}}(\theta)$. Since θ was arbitrary, we have $v \in K_{\mathcal{C}}$.
Conversely, suppose $v \in K_{\mathcal{C}}$. Let $c^* \coloneqq \sup\{c > 0 : \theta_0 + cv \in \mathcal{C}\}$. The supremum is over a nonempty set because $v \in T_{\mathcal{C}}(\theta_0)$. Suppose for sake of contradiction that $c^* < \infty$. Since $\mathcal{C}$ is closed, $\theta_0 + c^* v \in \mathcal{C}$. Thus $v \in T_{\mathcal{C}}(\theta_0 + c^* v)$ which implies $\theta_0 + (c^* + \alpha)v \in \mathcal{C}$ for some $\alpha > 0$, contradicting the definition of c^*. Thus $c^* = \infty$ and $\theta_0 + cv \in \mathcal{C}$ for all $c > 0$.

(ii) Both sides can be expressed as the set of $v \in \mathbb{R}^n$ satisfying $\theta_0 + \sigma v \in \mathcal{C}$ for all $\sigma > 0$.

We now prove the second part of the lemma. The definition (37) implies $K_{\mathcal{C}} \subseteq T_{\mathcal{C}}(\theta)$ for any $\theta \in \mathcal{C}$.

Now, assume $F_{\mathcal{C}}(\theta_0)$ is a cone. If the reverse inclusion $T_{\mathcal{C}}(\theta) \subseteq F_{\mathcal{C}}(\theta)$ holds, then $\theta_0 - \theta \in T_{\mathcal{C}}(\theta) = F_{\mathcal{C}}(\theta)$ so $\theta_0 - (\theta - \theta_0) \in \mathcal{C}$. Conversely, suppose $\theta_0 - (\theta - \theta_0) \in \mathcal{C}$. If $v \in T_{\mathcal{C}}(\theta)$, then $\theta + cv \in \mathcal{C}$ for some $c > 0$. By convexity, $\theta_0 + cv/2 \in \mathcal{C}$, so $v \in F_{\mathcal{C}}(\theta_0)$. Thus $T_{\mathcal{C}}(\theta) \subseteq F_{\mathcal{C}}(\theta)$. □

Proof of Proposition 5.3. We use Y instead of $\theta^* + \sigma Z$ throughout the proof, but note that Y depends on σ.

Without loss of generality we can translate the problem so that $\Pi_{\mathcal{C}}(\theta^*) = 0$.

In view of (13), we may use the dominated convergence theorem on $\sigma^{-2}\|\Pi_{\mathcal{C}}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2$, so

$$\begin{aligned}
&\lim_{\sigma\to\infty} \frac{1}{\sigma^2}\mathbb{E}\|\Pi_{\mathcal{C}}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 \\
&= \mathbb{E} \lim_{\sigma\to\infty} \frac{1}{\sigma^2}\|\Pi_{\mathcal{C}}(Y) - \Pi_{\mathcal{C}}(\theta^*)\|^2 && \text{dom. conv. with } \mathbb{E}\|Z\|^2 \\
&= \mathbb{E} \lim_{\sigma\to\infty} \frac{1}{\sigma^2}\|\Pi_{\mathcal{C}}(Y)\|^2 \\
&\overset{(i)}{=} \mathbb{E}\|\Pi_{K_{\mathcal{C}}}(Z)\|^2 = \delta(K_{\mathcal{C}}),
\end{aligned}$$

where we verify the equality (i) below.

Similarly, (13) allows us to use the dominated convergence theorem again for the excess risk.

$$\begin{aligned}
&\lim_{\sigma\to\infty} \frac{1}{\sigma^2}\big(\mathbb{E}\|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\Pi_{\mathcal{C}}(\theta^*) - \theta^*\|^2\big) \\
&= \mathbb{E} \lim_{\sigma\to\infty} \frac{1}{\sigma^2}\big(\|\Pi_{\mathcal{C}}(Y) - \theta^*\|^2 - \|\theta^*\|\big) && \text{dom. conv. with } \mathbb{E}\|Z\|^2 \\
&= \mathbb{E} \lim_{\sigma\to\infty} \frac{1}{\sigma^2}\big(\|\Pi_{\mathcal{C}}(Y)\|^2 - 2\langle \Pi_{\mathcal{C}}(Y), \theta^*\rangle\big) \\
&\overset{(ii)}{=} \mathbb{E}\|\Pi_{K_{\mathcal{C}}}(Z)\|^2 = \delta(K_{\mathcal{C}}).
\end{aligned}$$

It remains to verify (i) and (ii).

(i)

$$\begin{aligned}
&\left|\frac{1}{\sigma^2}\|\Pi_{\mathcal{C}}(Y)\|^2 - \|\Pi_{K_{\mathcal{C}}}(Z)\|^2\right| \\
&\le \frac{1}{\sigma^2}\left|\|\Pi_{\mathcal{C}}(Y)\|^2 - \|\Pi_{K_{\mathcal{C}}}(Y)\|^2\right| + \left|\frac{1}{\sigma^2}\|\Pi_{K_{\mathcal{C}}}(Y)\|^2 - \|\Pi_{K_{\mathcal{C}}}(Z)\|^2\right| \\
&\le \frac{c}{\sigma^2} + \left|\|\Pi_{K_{\mathcal{C}}}(\theta^*/\sigma + Z)\|^2 - \|\Pi_{K_{\mathcal{C}}}(Z)\|^2\right| && \text{Lemma E.1; } K_{\mathcal{C}} \text{ is a cone} \\
&\overset{\sigma\to\infty}{\longrightarrow} 0. && x \mapsto \|\Pi_{K_{\mathcal{C}}}(x)\|^2 \text{ is continuous}
\end{aligned}$$

(ii) We already showed $\|\Pi_{\mathcal{C}}(Y)\|^2/\sigma^2 \to \|\Pi_{K_{\mathcal{C}}}(Z)\|^2$, so it suffices to show the cross term vanishes. Indeed, we have $\|\Pi_{\mathcal{C}}(Y)\|/\sigma \to \|\Pi_{K_{\mathcal{C}}}(Z)\|$, so

$$\frac{1}{\sigma^2}|\langle \Pi_{\mathcal{C}}(Y), \theta^*\rangle| \le \frac{1}{\sigma^2}\|\Pi_{\mathcal{C}}(Y)\|\|\theta^*\| \overset{\sigma\to\infty}{\longrightarrow} 0.$$

□

Proof of Corollary 5.4. We begin with the first claim. Since $\mathcal{C} = \mathbb{R}^n_+$ is a cone, we have $K_{\mathcal{C}} = \mathbb{R}^n_+$. Provided we verify (38), the result follows from Proposition 5.3. Let $\theta \coloneqq \Pi_{\mathcal{C}}(\theta^*)$ and fix $x \in \mathbb{R}^n$. Then some casework yields

$$\|\Pi_{F_{\mathcal{C}}(\theta)}(x)\|^2 - \|\Pi_{K_{\mathcal{C}}}(x)\|^2 = \sum_{i=1}^n \max\{x_i, -\theta_i\}^2 - \sum_{i=1}^n \max\{x_i, 0\}^2 \le \sum_{i=1}^n \theta_i^2 = \|\theta\|^2 =: c.$$

We now turn to the second claim. If $\mathcal{C}$ is bounded, then by Lemma 5.2, $K_{\mathcal{C}} = \{u : \mathbb{R}_+ u \subseteq F_{\mathcal{C}}(\theta_0)\} = \{0\}$ for any $\theta_0 \in \mathcal{C}$.

Conversely, suppose $\mathcal{C}$ is unbounded and fix $\theta_0 \in \mathcal{C}$. Let

$$U_r \coloneqq \{v \in S^{n-1} : \theta_0 + cv \notin \mathcal{C} \text{ for some } c \in (0, r)\}.$$

This set is open: if (v_n) is a sequence in U_r^c converging to v, then the fact that $\mathcal{C}$ is closed implies $\theta_0 + rv_n \in \mathcal{C}$ for all n, and consequently $\theta_0 + rv \in \mathcal{C}$ and finally $v \in U_r^c$.

If $\bigcup_{r>0} U_r$ is an open cover of the compact set S^{n-1}, then $S^{n-1} \subseteq U_r$ for some $r > 0$, which implies $\mathcal{C} \subseteq B_r(\theta_0)$, a contradiction. Thus, some direction $v \in S^{n-1}$ does not lie in $\bigcup_{r>0} U_r$, i.e., $\theta_0 + cv \in \mathcal{C}$ for all $c \ge 0$. This implies $cv \in K_{\mathcal{C}}$ for all $c \ge 0$.

We now apply Proposition 5.3. If $\mathcal{C}$ is bounded, then so is $F_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$. Choosing c large enough so that $F_{\mathcal{C}}(\Pi_{\mathcal{C}}(\theta^*))$ lies in the ball of radius c suffices to satisfy (38). Then Proposition 5.3 implies that the high σ limits are $\delta(K_{\mathcal{C}}) = 0$. □

Acknowledgments

We thank the reviewers for their helpful comments and suggestions. We also thank Dennis Amelunxen for an informative email correspondence and to Bodhisattva Sen for helpful discussions.

References

[1] Amelunxen, D. and M. Lotz (2015). Intrinsic volumes of polyhedral cones: a combinatorial perspective. *arXiv preprint arXiv:1512.06033*.

[2] Amelunxen, D., M. Lotz, M. B. McCoy, and J. A. Tropp (2014). Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, iau005.

[3] Bellec, P. C. (2015). Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*.

[4] Groeneboom, P. and G. Jongbloed (2014). *Nonparametric Estimation under Shape Constraints: Estimators, Algorithms and Asymptotics*, Volume 38. Cambridge University Press.
[5] Hiriart-Urruty, J.-B. and C. Lemaréchal (2012). *Fundamentals of convex analysis*. Springer Science & Business Media.
[6] Jankowski, H. (2014). Convergence of linear functionals of the Grenander estimator under misspecification. *Ann. Statist. 42*(2), 625–653.
[7] Klivans, C. J. and E. Swartz (2011). Projection volumes of hyperplane arrangements. *Discrete & Computational Geometry 46*(3), 417.
[8] Oymak, S. and B. Hassibi (2013). Sharp mse bounds for proximal denoising. *Foundations of Computational Mathematics*, 1–65.
[9] Pal, J. K. (2008). Spiking problem in monotone regression: Penalized residual sum of squares. *Statistics & Probability Letters 78*(12), 1548–1556.
[10] Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Chichester: John Wiley & Sons Ltd.
[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
[12] Wu, J., M. C. Meyer, and J. D. Opsomer (2015). Penalized isotonic regression. *Journal of Statistical Planning and Inference 161*, 12–24.
[13] Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist. 30*(2), 528–555.