The Annals of Statistics
2018, Vol. 46, No. 2, 745–780
https://doi.org/10.1214/17-AOS1566

SHARP ORACLE INEQUALITIES FOR LEAST SQUARES ESTIMATORS IN SHAPE RESTRICTED REGRESSION[1]

BY PIERRE C. BELLEC

ENSAE, UMR CNRS 9194 and Rutgers University

The performance of Least Squares (LS) estimators is studied in shape-constrained regression models under Gaussian and sub-Gaussian noise. General bounds on the performance of LS estimators over closed convex sets are provided. These results have the form of sharp oracle inequalities that account for the model misspecification error. In the presence of misspecification, these bounds imply that the LS estimator estimates the projection of the true parameter at the same rate as in the well-specified case.

In isotonic and unimodal regression, the LS estimator achieves the nonparametric rate $n^{-2/3}$ as well as a parametric rate of order k/n up to logarithmic factors, where k is the number of constant pieces of the true parameter. In univariate convex regression, the LS estimator satisfies an adaptive risk bound of order q/n up to logarithmic factors, where q is the number of affine pieces of the true regression function. This adaptive risk bound holds for any collection of design points. While Guntuboyina and Sen [*Probab. Theory Related Fields* **163** (2015) 379–411] established that the nonparametric rate of convex regression is of order $n^{-4/5}$ for equispaced design points, we show that the nonparametric rate of convex regression can be as slow as $n^{-2/3}$ for some worst-case design points. This phenomenon can be explained as follows: Although convexity brings more structure than unimodality, for some worst-case design points this extra structure is uninformative and the nonparametric rates of unimodal regression and convex regression are both $n^{-2/3}$. Higher order cones, such as the cone of β-monotone sequences, are also studied.

1. Introduction. This paper studies shape-constrained regression models, which includes isotonic, convex and unimodal regression. These regression models are nonparametric but enjoy a canonical estimator, namely, the *shape-constrained* Least-Squares (LS) estimator. This canonical estimator requires no tuning parameter, which is in contrast to most other nonparametric models where the choice of the tuning parameter can be a challenging issue. We study the performance of the shape-constrained LS estimator in univariate fixed-design regression under the squared loss. This problem was studied in, among others, Meyer and

Received October 2015; revised March 2017.

[1] This work was supported by GENES and by the French National Research Agency (ANR) under the Grants IPANEMA (ANR-13-BSH1-0004-02) and Labex ECODEC (ANR-11-LABEX-0047).

MSC2010 subject classifications. 62G08, 62C20.

Key words and phrases. Shape restricted regression, convexity, minimax rates, Gaussian width, concentration.

Woodroofe [18], Zhang [25], Chatterjee [9], Guntuboyina and Sen [16] and Chatterjee et al. [11].

Assume that we have the observations

$$Y_i = \mu_i + \xi_i, \qquad i = 1, \dots, n, \tag{1.1}$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T \in \mathbb{R}^n$ is unknown, $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^T$ is a noise vector with n-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ where $\sigma > 0$ and $I_{n\times n}$ is the $n \times n$ identity matrix. The vector $\mathbf{y} = (Y_1, \dots, Y_n)^T$ is observed and the goal is to estimate $\boldsymbol{\mu}$. The estimation error is measured with the scaled norm $\|\cdot\|$ defined by

$$\|\boldsymbol{u}\|^2 = \frac{1}{n}\sum_{i=1}^n u_i^2, \qquad \boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n. \tag{1.2}$$

The error of an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ is given by $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. Denote by $\mathbb{E}_{\boldsymbol{\mu}}$ and $\mathbb{P}_{\boldsymbol{\mu}}$ the expectation and the probability with respect to the distribution of the random variable $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi}$.

The components $\mu_1, \dots, \mu_n$ of $\boldsymbol{\mu}$ can be interpreted as the values of an unknown regression function $f : \mathbb{R} \to \mathbb{R}$ at given, deterministic design points $x_1 < \cdots < x_n$. We observe $Y_i = f(x_i) + \xi_i$ for $i = 1, \dots, n$ and our goal is to construct estimates $\hat{\mu}_i$ that are close to $f(x_i)$. The shape constraint is a nonparametric class of functions, for instance, the class of nondecreasing functions or the class of convex functions. For any design points $x_1 < \cdots < x_n$, define the sets

$$\mathcal{S}_n^{\uparrow} := \{\boldsymbol{u} = (f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n \text{ for some nondecreasing } f\}, \tag{1.3}$$

$$K^C_{x_1,\dots,x_n} := \{\boldsymbol{u} = (f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n \text{for some convex } f\}. \tag{1.4}$$

The above notation emphasizes that the set $\mathcal{S}_n^{\uparrow}$ does not depend on a particular collection of design points $x_1, \dots, x_n$, whereas the set $K^C_{x_1,\dots,x_n}$ does depend on the design points $x_1, \dots, x_n$. The isotonic regression problem studies the class of nondecreasing functions and the set $\mathcal{S}_n^{\uparrow}$, equivalently defined as

$$\mathcal{S}_n^{\uparrow} := \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : u_i \le u_{i+1}, i = 1, \dots, n-1\}.$$

The convex regression problem studies the class of convex functions and the set $K^C_{x_1,\dots,x_n}$. If $x_1 < \cdots < x_n$ are equispaced design points in $\mathbb{R}$, that is, $x_i = (i-1)(x_2 - x_1) + x_1$, $i = 2, \dots, n$, then the set $K^C_{x_1,\dots,x_n}$ is equal to

$$\mathcal{S}_n^{\mathrm{C}} := \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : 2u_i \le u_{i+1} + u_{i-1}, i = 2, \dots, n-1\}. \tag{1.5}$$

This paper studies the Least Squares (LS) estimator in shape restricted regression under model misspecification. The LS estimator over a nonempty closed set $K \subset \mathbb{R}^n$ is defined by

$$\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) \in \operatorname*{argmin}_{\boldsymbol{u}\in K} \|\mathbf{y} - \boldsymbol{u}\|^2.$$

If K is a closed set, there exists at least one solution to this minimization problem and $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ denotes any such solution. The sets $\mathcal{S}_n^{\uparrow}$, $K_{x_1,\dots,x_n}^C$ are closed and convex and we will study the performance of the isotonic LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ and the convex LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K_{x_1,\dots,x_n}^C)$.

The techniques developed in the present paper deepens our understanding of two problems in shape restricted regression: (a) the role of the design points introduced in the next subsection, and (b), the consequences of model misspecification, introduced in Section 1.2. A detailed summary of our contributions is given in Section 1.3.

1.1. *On the design points in univariate shape-constrained regression models.* It is clear that the design points play no particular role in isotonic regression since the set (1.3) is the same for any collection of design points $x_1 < \cdots < x_n$. However, the role of the design points is not clear in convex regression. Although the convex LS estimator for equispaced design points is well studied in the literature [10, 16], little is known about its performance if the design points are not equispaced. This raises the following question:

- What is the performance of the convex LS estimator if the design points are allowed to be arbitrarily close or arbitrarily far from each other?

We now review the literature on the isotonic and convex LS estimators. The following quantities will be useful. First, define the total variation

$$V(\boldsymbol{\theta}) := \max_{i=1,\dots,n} \theta_i - \min_{i=1,\dots,n} \theta_i \qquad \text{for all } \boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^T \in \mathbb{R}^n. \tag{1.6}$$

If $\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathcal{S}_n^{\uparrow}$, its total variation is simply $V(\boldsymbol{u}) = u_n - u_1$. For $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$, let $k(\boldsymbol{u}) \geq 1$ be the minimal integer k such that $\boldsymbol{u}$ is piecewise constant with k pieces. For $\boldsymbol{u} \in K_{x_1,\dots,x_n}^C$, let $q(\boldsymbol{u}) \geq 1$ be the minimal integer q such that $\boldsymbol{u}$ is piecewise affine with q pieces [cf. Section 1.4 below for formal definitions of $k(\cdot)$ and $q(\cdot)$].

Previous results on the performance of the LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ can be found in [9, 11, 18, 25]. Two types of risk bounds or oracle inequalities have been obtained so far. If, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^T \in \mathcal{S}_n^{\uparrow}$, it is known [9, 11, 18, 25] that for some absolute constant $c > 0$,

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \frac{c\sigma^2 \log(en)}{n} + c\sigma^2 \left(\frac{V(\boldsymbol{\mu})}{\sigma n}\right)^{2/3} \tag{1.7}$$

and $c \leq 12.3$; cf. [25]. If $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, the following oracle inequality was proved in [11]:

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq 6 \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{1.8}$$

The risk bounds (1.7) and (1.8) hold under the assumption that $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$, which does not allow for any model misspecification. We will see in Section 3 that the misspecified case $\boldsymbol{\mu} \notin \mathcal{S}_n^{\uparrow}$ can be handled provided that an error term of the form $\inf_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2$ is added to the right hand side of (1.7). The risk bound (1.7) implies that the LS estimator achieves the rate $n^{-2/3}$ while (1.8) yields a parametric rate (up to logarithmic factors) if $\boldsymbol{\mu}$ is well approximated by a piecewise constant sequence with not too many pieces. Let us note that the bound (1.8) can be used to obtain that $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ converges at the rate $n^{-2/3}$ up to logarithmic factors, thanks to the approximation argument given in [5], Lemma 2.

Mimimax lower bounds that match (1.7) and (1.8) up to logarithmic factors have been obtained in [5, 11]. If $D > 0$ is a fixed parameter and $\log(en)^3\sigma^2 \leq nD^2$, the bound (1.7) yields the rate $(D\sigma^2)^{2/3}n^{-2/3}$ for the risk of $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$. By the lower bound [5], Corollary 5, this rate is minimax optimal over the class $\{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow} : V(\boldsymbol{u}) \leq D\}$ if $\log(en)^3\sigma^2 \leq nD^2$. Proposition 4 in [5] shows that there exist absolute constants $c, c' > 0$ such that, for any estimator $\hat{\boldsymbol{\mu}}$,

$$\sup_{\boldsymbol{\mu}\in\mathcal{S}_n^{\uparrow}:k(\boldsymbol{\mu})\leq k} \mathbb{P}_{\boldsymbol{\mu}}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq c\sigma^2 k/n) \geq c'. \tag{1.9}$$

Together, (1.8) and (1.9) establish that for any $k = 1, \dots, n$, the minimax rate over the class $\{\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow} : k(\boldsymbol{\mu}) \leq k\}$ is of order $\sigma^2 k/n$ up to logarithmic factors.

The performance of the convex LS estimator with equispaced design points has been recently studied in [11, 16], where it was proved that if $\boldsymbol{\mu} \in \mathcal{S}_n^{\text{C}}$, the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\text{C}})$ satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C\left(\min_{\boldsymbol{u}\in\mathcal{S}_n^{\text{C}}}\left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 q(\boldsymbol{u})}{n}\left(\log\frac{en}{q(\boldsymbol{u})}\right)^{5/4}\right)\right) \tag{1.10}$$

for some absolute constant $C > 1$. Guntuboyina and Sen [16] showed that the LS estimator achieves the nonparametric rate $n^{-4/5}$ up to a logarithmic factor and Chatterjee [10] later proved that the logarithmic was not necessary: The LS estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\text{C}})$ satisfies [10, 16]

$$\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \frac{C\sigma^2(1 + V(\boldsymbol{u})/\sigma)^{2/5}}{n^{4/5}} \tag{1.11}$$

for some absolute constant $C > 0$. The bound (1.10) yields an almost parametric rate if $\boldsymbol{\mu}$ can be well approximated by a piecewise affine sequence with not too many pieces. If $\bar{V} > 0$ is a fixed parameter and $\bar{V} \geq \sigma$, the bound (1.11) yields the rate $(\bar{V}\sigma^4)^{2/5}n^{-4/5}$, which is minimax optimal over the class $\{\boldsymbol{\mu} \in \mathcal{S}_n^{\text{C}} : V(\boldsymbol{\mu}) \leq \bar{V}\}$ [10, 16].

The performance of the LS estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K^C_{x_1,\dots,x_n})$ is also studied in [16] in the case where the design points are almost equispaced: The bounds (1.10) and (1.11) both hold if $\mathcal{S}_n^{\mathrm{C}}$ is replaced with $K_{x_1,\dots,x_n}^{C}$ and if $C > 0$ is a constant that depends on the ratio

$$\frac{\max_{i=2,\dots,n}(x_i - x_{i-1})}{\min_{i=2,\dots,n}(x_i - x_{i-1})}, \tag{1.12}$$

and this constant C becomes arbitrarily large as this ratio tends to infinity.

Inequalities (1.11) and (1.10) provide an accurate picture of the performance of the LS estimator for equispaced (or almost equispaced) design points. However, little is known on the behavior of the convex LS estimator for design points that are not equispaced, and some natural questions arise:

- Does the adaptive risk bound (1.10) still hold for nonequispaced design points?
- How is the bound (1.11) impacted for nonequispaced design points? Is the nonparametric rate still of order $n^{-4/5}$ if the design points are allowed to be arbitrarily close to each other?

These questions will be answered in Section 4. Section 4.1 shows that the adaptive risk bound (1.10) holds irrespective of the design points, and Section 4.2 shows that the nonparametric rate of the convex LS estimator can be as slow as $n^{-2/3}$ for some worst-case design points.

1.2. *Accounting for model misspecification.* Let K be a subset of $\mathbb{R}^n$. If the unknown regression vector $\boldsymbol{\mu}$ lies in K, we say that the model is well-specified. If $\boldsymbol{\mu} \in K$, an estimator $\hat{\boldsymbol{\mu}}$ enjoys good performance if the squared error $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ is small, either in expectation or with high probability. If $\boldsymbol{\mu} \notin K$, we say that the model is misspecified. In that case, several natural quantities are of interest to assess the performance of an estimator $\hat{\boldsymbol{\mu}}$. This includes the regret of order 1 and the regret of order 2 of an estimator $\hat{\boldsymbol{\mu}}$ which are defined by

$$R_2(\hat{\boldsymbol{\mu}}) := \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \min_{\boldsymbol{u}\in K} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2, \quad R_1(\hat{\boldsymbol{\mu}}) := \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in K} \|\boldsymbol{u} - \boldsymbol{\mu}\|. \tag{1.13}$$

If the set K is closed and convex and $\hat{\boldsymbol{\mu}}$ is valued in K, another quantity of interest is the estimation error with respect to the projection of $\boldsymbol{\mu}$ onto K:

$$\|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|^2. \tag{1.14}$$

Estimation of $\Pi_K(\boldsymbol{\mu})$ by the LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ and upper bounds on the quantity (1.14) have been considered in [25], Section 4, for $K = \mathcal{S}_n^{\uparrow}$, and in [16], Section 6, for $K = \mathcal{S}_n^{\mathrm{C}}$. Misspecification bounds usually take the form of oracle inequalities, that is, bounds such as

$$\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le C \min_{\boldsymbol{u}\in K} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + r_n, \tag{1.15}$$

where $C \ge 1$ is called the leading constant, and r_n is a small quantity called the remainder term. If $C = 1$ in (1.15), we say that the oracle inequality is *sharp* or *exact*. Such inequalities were derived in [5] in isotonic and convex regression for an estimator $\hat{\boldsymbol{\mu}}$ that cannot be computed in practice.

Although model misspecification was studied in [5, 11, 16, 25], a clear comparison of the quantities (1.13)–(1.14) is lacking. A goal of the present paper is to provide a comparison of the quantities (1.13)–(1.14) as well as general tools to bound from above these quantities in the form of sharp oracle inequalities; cf. Section 2. Section 2.3 will highlight advantages of sharp oracle inequalities, that is, oracle inequalities such as (1.15) with leading constant $C = 1$.

1.3. *Organization of the paper.* Section 1.4 defines our notation and Section 1.5 recalls properties of closed convex sets and closed convex cones. The contributions of the paper are organized as follows.

In Section 2, we establish general tools to derive sharp oracle inequalities for the LS estimator over a closed convex set K; cf. Corollary 2.2 and Theorem 2.3. These results are generalized to the nonconvex case in Remarks 2.1 and 2.4. Section 2.3 compares different quantities that represent the estimation error when the model is misspecified.

In Section 3, we apply results of Section 2 to the isotonic LS estimator. We obtain an adaptive risk bound that is tight with sharp numerical constants.

Section 4 studies the role of the design points in univariate convex regression and answers questions 1, 2 and 3 raised in the Introduction:

- On the one hand, the adaptive risk bound (1.10) holds for any collection of design points.
- On the other hand, although the nonparametric rate is of order $n^{-4/5}$ for equispaced design points, this rate can be as slow as $n^{-2/3}$ for some worst-case design points that are studied in Section 4.2.

Section 5 illustrates the results of Section 2.3 for $K = \mathcal{S}_n^{\uparrow}$ and $K = \mathcal{S}_n^{\mathrm{C}}$: If $\boldsymbol{\mu} \notin K$ then the LS estimator consistently estimates the projection of the true parameter $\boldsymbol{\mu}$ onto K at the same rate as in the well-specified case. An extension to sub-Gaussian noise is given in Section 6. Some proofs are delayed to Appendices A and B. An outcome of the proof techniques used to study convex regression is the oracle inequalities satisfied by the unimodal LS estimator in Appendix C. The supplementary material [3] contains generalizations of the results in isotonic and convex regression to higher order cones. Our main results are summarized in Table 1.

Finally, although the focus of the present paper is on shape-constrained models, Remark 2.2 provides general oracle inequalities for some penalized estimators.

1.4. *Notation.* We consider the observations (1.1). We also use the notation $\boldsymbol{g} := (1/\sigma)\boldsymbol{\xi}$ so that $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\xi} = \boldsymbol{\mu} + \sigma\boldsymbol{g}$ and $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. The scaled norm $\|\cdot\|$ is defined in (1.2). Let also $|\cdot|_\infty$ be the infinity norm and $|\cdot|_2$ be the Euclidean norm, so that $\frac{1}{n}|\cdot|_2^2 = \|\cdot\|^2$. For any vector $\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n$ and any set

TABLE 1

Rates of estimation of the LS estimator for several shape constraints, for $\sigma = 1$ and $V(\boldsymbol{\mu}) \leq 1$. The rates of estimation of $\boldsymbol{\mu}$ in the well-specified case are the same as the rates of estimation of $\Pi_K(\boldsymbol{\mu})$ in the misspecified case. In unimodal regression, the misspecified rate is undefined because of the nonconvexity of the set $\mathcal{U}$

		Well-specified rate	**Misspecified rate**
$\boldsymbol{K}$		$\mathbb{E}\\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\boldsymbol{K}) - \boldsymbol{\mu}\\|^2$ **if** $\boldsymbol{\mu} \in K$	$\mathbb{E}\\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\boldsymbol{K}) - \boldsymbol{\Pi_K(\mu)}\\|^2$
$\mathcal{S}_n^{\uparrow}$	Isotonic regression	$\frac{k(\boldsymbol{\mu})}{n} \wedge \frac{1}{n^{2/3}}$	$\frac{k(\Pi_K(\boldsymbol{\mu}))}{n} \wedge \frac{1}{n^{2/3}}$
$\mathcal{S}_n^{\mathrm{C}}$	Convex regression with equispaced design	$\frac{q(\boldsymbol{\mu})}{n} \wedge \frac{1}{n^{4/5}}$	$\frac{q(\Pi_K(\boldsymbol{\mu}))}{n} \wedge \frac{1}{n^{4/5}}$
$K^C_{x_1,\dots,x_n}$	Convex regression for worst-case design	$\frac{q(\boldsymbol{\mu})}{n} \wedge \frac{1}{n^{2/3}}$	$\frac{q(\Pi_K(\boldsymbol{\mu}))}{n} \wedge \frac{1}{n^{2/3}}$
$\mathcal{U}$	Unimodal regression	$\frac{k(\boldsymbol{\mu})}{n} \wedge \frac{1}{n^{2/3}}$	Undefined

$T \subset \{1, \dots, n\}$ of the form $T = \{t_1, \dots, t_p\}$ with $1 \leq t_1 < t_2 < \cdots < t_p \leq n$, define $\boldsymbol{u}_T = (u_{t_1}, \dots, t_{t_p})^T$, the restriction of $\boldsymbol{u}$ to the set T.

The total variation $V(\boldsymbol{u})$ of any $\boldsymbol{u} \in \mathbb{R}^n$ is defined in (1.6). If $\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathcal{S}_n^{\uparrow}$, its total variation is simply $V(\boldsymbol{u}) = u_n - u_1$.

Let $x_1 < \cdots < x_n$ be real numbers that will be called the design points. The sets $\mathcal{S}_n^{\uparrow}$, $\mathcal{S}_n^{\mathrm{C}}$ and $K^C_{x_1,\dots,x_n}$ are defined in (1.3), (1.4) and (1.5). They are closed convex subsets of $\mathbb{R}^n$. An equivalent definition of the set (1.4) is

$$K^C_{x_1,\dots,x_n} := \left\{\boldsymbol{u} \in \mathbb{R}^n : \frac{u_i - u_{i-1}}{x_i - x_{i-1}} \leq \frac{u_{i+1} - u_i}{x_{i+1} - x_i}, i = 2, \dots, n-1\right\}. \tag{1.16}$$

For any $\boldsymbol{u} = (u_1, \dots, u_n)^T \in K^C_{x_1,\dots,x_n}$, we say that $\boldsymbol{u}$ is piecewise affine with k pieces if there exist real numbers $a_1, \dots, a_k$ and a partition $(T_1, \dots, T_k)$ of $\{1, \dots, n\}$ such that

$$u_i = a_j(x_i - x_l) + u_l, \qquad i, l \in T_j, j = 1, \dots, k.$$

If $\boldsymbol{u} = (f(x_1), \dots, f(x_n))^T$ for some convex function $f : \mathbb{R} \to \mathbb{R}$ and f is a piecewise affine function with k pieces, then $\boldsymbol{u}$ is piecewise affine with k pieces. For any $\boldsymbol{u} \in K^C_{x_1,\dots,x_n}$, let $q(\boldsymbol{u}) \geq 1$ be the smallest integer q such that the sequence $\boldsymbol{u}$ is piecewise affine with q pieces. The quantity $q(\boldsymbol{u}) \geq 1$ satisfies

$$q(\boldsymbol{u}) - 1 \leq \left|\left\{i = 2, \dots, n-1 : \frac{u_i - u_{i-1}}{x_i - x_{i-1}} < \frac{u_{i+1} - u_i}{x_{i+1} - x_i}\right\}\right|.$$

Let $m = 1, \dots, n$. A sequence $\boldsymbol{u} \in \mathbb{R}^n$ is unimodal with mode at position m if and only if $\boldsymbol{u}_{\{1,\dots,m\}}$ is nonincreasing and $\boldsymbol{u}_{\{m,\dots,n\}}$ is nondecreasing. Define the convex set

$$K_m := \{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : u_1 \geq \dots \geq u_m \leq u_{m-1} \leq \dots \leq u_n\}. \tag{1.17}$$

The convex set K_m is the set of all unimodal sequences with mode at position m and

$$\mathcal{U} := \bigcup_{m=1,\dots,n} K_m \tag{1.18}$$

is the set of all unimodal sequences. The set $\mathcal{U}$ is nonconvex for $n \geq 3$. For all $\boldsymbol{u} \in \mathcal{U}$, let $k(\boldsymbol{u})$ be the smallest integer k such that $\boldsymbol{u}$ is piecewise constant with k pieces. Formally, $k(\boldsymbol{u})$ is the smallest integer k such that there exists a partition $(T_1, \dots, T_k)$ of $\{1, \dots, n\}$ such that for all $l = 1, \dots, k$, the sequence $\boldsymbol{u}_{T_l}$ is constant and the set T_l is convex in the sense that if $a, b \in T_l$ then T_l contains all integers between a and b. If $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$, then $k(\boldsymbol{u}) \geq 1$ is the integer such that $\boldsymbol{u}$ has $k(\boldsymbol{u}) - 1$ jumps, that is, $k(\boldsymbol{u}) - 1$ is the number of inequalities $u_i \leq u_{i+1}$ that are strict for $i = 1, \dots, n-1$.

The unimodal LS estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{U})$ enjoys properties similar to that of the isotonic LS estimator: the rate of estimation is parametric if the true regression vector $\boldsymbol{\mu}$ is unimodal with few constant pieces, and the nonparametric rate is of order $n^{-2/3}$; cf. [13, 14] and Appendix C below.

1.5. *Preliminary properties of closed convex sets.* We recall here several properties of convex sets that are used throughout the paper. Given a closed convex set $K \subset \mathbb{R}^n$, denote by $\Pi_K : \mathbb{R}^n \to K$ the projection onto K. For all $\mathbf{y} \in \mathbb{R}^n$, $\Pi_K(\mathbf{y})$ is the unique vector in K such that

$$(\boldsymbol{u} - \Pi_K(\mathbf{y}))^T (\mathbf{y} - \Pi_K(\mathbf{y})) \leq 0, \qquad \boldsymbol{u} \in K. \tag{1.19}$$

Inequality (1.19) can be rewritten as follows:

$$\|\Pi_K(\mathbf{y}) - \mathbf{y}\|^2 + \|\boldsymbol{u} - \Pi_K(\mathbf{y})\|^2 \leq \|\boldsymbol{u} - \mathbf{y}\|^2, \qquad \mathbf{y} \in \mathbb{R}^n, \boldsymbol{u} \in K, \tag{1.20}$$

which is a consequence of the cosine theorem. The LS estimator over K is exactly the projection of $\mathbf{y}$ onto K, that is, $\hat{\boldsymbol{\mu}}^{\text{LS}}(K) = \Pi_K(\mathbf{y})$. In this case, (1.20) yields that for all $\boldsymbol{u} \in K$,

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \mathbf{y}\|^2 \leq \|\boldsymbol{u} - \mathbf{y}\|^2 - \|\boldsymbol{u} - \hat{\boldsymbol{\mu}}^{\text{LS}}(K)\|^2. \tag{1.21}$$

Inequality (1.21) can be interpreted in terms of strong convexity: the LS estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ solves an optimization problem where the function to minimize is strongly convex with respect to the norm $\|\cdot\|$. Strong convexity grants inequality (1.21), which is stronger than the inequality

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{V}) - \mathbf{y}\|^2 \leq \|\boldsymbol{u} - \mathbf{y}\|^2 \qquad \text{for all } \boldsymbol{u} \in \mathcal{V}, \tag{1.22}$$

which holds for any closed set $\mathcal{V} \subset \mathbb{R}^n$.

Now, assume that K is a closed convex cone. In this case, (1.19) is equivalent to the statement that for all $\mathbf{y} \in \mathbb{R}^n$, $\Pi_K(\mathbf{y})$ is the unique vector in K such that

$$\Pi_K(\mathbf{y})^T\mathbf{y} = |\Pi_K(\mathbf{y})|_2^2 \quad \text{and} \quad \forall \boldsymbol{\theta} \in K, \qquad \boldsymbol{\theta}^T\mathbf{y} \le \boldsymbol{\theta}^T\Pi_K(\mathbf{y}). \tag{1.23}$$

The property (1.23) readily implies that for any $\boldsymbol{v} \in \mathbb{R}^n$ we have

$$|\Pi_K(\boldsymbol{v})|_2 = \sup_{\boldsymbol{\theta}\in K:|\boldsymbol{\theta}|_2\le 1} \boldsymbol{v}^T\boldsymbol{\theta}. \tag{1.24}$$

Define the statistical dimension of the cone K by

$$\delta(K) := \mathbb{E}\big[|\Pi_K(\boldsymbol{g})|_2^2\big] = \mathbb{E}\big[\boldsymbol{g}^T\Pi_K(\boldsymbol{g})\big] = \mathbb{E}\Big[\Big(\sup_{\boldsymbol{\theta}\in K:|\boldsymbol{\theta}|_2\le 1} \boldsymbol{g}^T\boldsymbol{\theta}\Big)^2\Big], \tag{1.25}$$

where $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. The Gaussian width of a closed convex cone K is defined by

$$w(K) = \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(\mathbf{0},I_{n\times n})}\Big[\sup_{\boldsymbol{v}\in K:|\boldsymbol{v}|_2=1} \boldsymbol{g}^T\boldsymbol{v}\Big].$$

The quantities $\delta(K)$ and $w(K)$ are closely related as they satisfy $w^2(K) \le \delta(K) \le w^2(K) + 1$ for any closed convex cone K [1], Proposition 10.2. The following properties of the statistical dimension are useful for our purpose. If $K \subset \mathbb{R}^q$, $C \subset \mathbb{R}^p$ are two closed convex cones, then $K \times C$ is a closed convex cone in $\mathbb{R}^{q+p}$ and

$$\delta(K \times C) = \delta(K) + \delta(C). \tag{1.26}$$

The statistical dimension $\delta(\cdot)$ is monotone in the following sense: If K, L are two closed convex cones in $\mathbb{R}^n$, then

$$K \subset L \quad \Rightarrow \quad \delta(K) \le \delta(L). \tag{1.27}$$

We refer the reader to [1], Proposition 3.1, for straightforward proofs of the equivalence between the definitions (1.25) and the properties (1.26), (1.27) and (1.24). An exact formula is available for the statistical dimension of $\mathcal{S}_n^{\uparrow}$. Namely, it is proved in [1], (D.12), that

$$\delta(\mathcal{S}_n^{\uparrow}) = \sum_{k=1}^{n} \frac{1}{k}, \tag{1.28}$$

and this formula readily implies that

$$\log(n) \le \delta(\mathcal{S}_n^{\uparrow}) \le \log(en). \tag{1.29}$$

The following upper bound on the statistical dimension of the cone $K_{x_1,\dots,x_n}^C$ is derived in [16]:

$$\delta(K_{x_1,\dots,x_n}^C) \le c(\log(en))^{5/4}, \tag{1.30}$$

for some constant $c > 0$ that depends on the ratio (1.12). In Theorem 4.1, we derive a tighter bound independent of the design points.

2. General tools to derive sharp oracle inequalities. In this section, we develop general tools to derive sharp oracle inequalities for the LS estimator over a closed convex set. Generalizations to nonconvex sets are given in Remarks 2.1 and 2.4.

2.1. *Statistical dimension of the tangent cone.* Let $\boldsymbol{\mu} \in \mathbb{R}^n$, let K be a closed convex subset of $\mathbb{R}^n$ and let $\boldsymbol{u} \in \mathbb{R}^n$. Define the tangent cone at $\boldsymbol{u}$ by

$$\mathcal{T}_{K,\boldsymbol{u}} := \text{closure}\{t(\boldsymbol{v} - \boldsymbol{u}) : t \geq 0, \boldsymbol{v} \in K\}.$$

If K is a closed convex cone, then $\mathcal{T}_{K,\boldsymbol{u}} = \{\boldsymbol{v} - t\boldsymbol{u} \mid \boldsymbol{v} \in K, t \geq 0\}$.

PROPOSITION 2.1. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, let K be a closed convex subset of $\mathbb{R}^n$ and let $\boldsymbol{u} \in K$. Then if $\boldsymbol{g} = (1/\sigma)\boldsymbol{\xi}$, we have almost surely*

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}\|^2 - \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 &\leq \frac{\sigma^2}{n}\Big(\sup_{\boldsymbol{\theta} \in \mathcal{T}_{K,\boldsymbol{u}} : |\boldsymbol{\theta}|_2^2 \leq 1} \boldsymbol{\theta}^T \boldsymbol{g}\Big)^2 \\ &= \frac{\sigma^2}{n}\big|\Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{g})\big|_2^2.\end{aligned} \tag{2.1}$$

PROOF. Let $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\text{LS}}(K)$. Then (1.21) yields

$$\begin{aligned}|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 - |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 &\leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \\ &= 2\boldsymbol{\xi}^T\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2,\end{aligned} \tag{2.2}$$

where $\hat{\boldsymbol{\theta}}$ is defined by $\hat{\boldsymbol{\theta}} = (1/|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2)(\hat{\boldsymbol{\mu}} - \boldsymbol{u})$ if $\hat{\boldsymbol{\mu}} \neq \boldsymbol{u}$ and $\hat{\boldsymbol{\theta}} = \mathbf{0}$ otherwise. By construction we have $\hat{\boldsymbol{\theta}} \in \mathcal{T}_{K,\boldsymbol{u}}$ and $|\hat{\boldsymbol{\theta}}|_2^2 \leq 1$. Using the simple inequality $2ab - b^2 \leq a^2$ with $a = \sup_{\boldsymbol{\theta} \in \mathcal{T}_{K,\boldsymbol{u}} : |\boldsymbol{\theta}|_2 \leq 1} \boldsymbol{\xi}^T\boldsymbol{\theta}$ and $b = |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2$, we obtain

$$\begin{aligned}|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 - |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 &\leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \\ &\leq \Big(\sup_{\boldsymbol{\theta} \in \mathcal{T}_{K,\boldsymbol{u}} : |\boldsymbol{\theta}|_2^2 \leq 1} \boldsymbol{\theta}^T\boldsymbol{\xi}\Big)^2.\end{aligned}$$

The equality (1.24) completes the proof. Alternatively, one can observe that $\hat{\boldsymbol{\mu}} - \boldsymbol{u} \in \mathcal{T}_{K,\boldsymbol{u}}$ and thus

$$\begin{aligned}2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 &= |\boldsymbol{\xi}|_2^2 - |\boldsymbol{\xi} - (\hat{\boldsymbol{\mu}} - \boldsymbol{v})|_2^2 \\ &\leq |\boldsymbol{\xi}|_2^2 - |\boldsymbol{\xi} - \Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{\xi})|_2^2 \\ &= \big|\Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{\xi})\big|_2^2,\end{aligned}$$

where the last equality is a consequence of the Pythagorean theorem. □

By definition of the statistical dimension, $\delta(\mathcal{T}_{K,\boldsymbol{u}}) := \mathbb{E}|\Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{g})|_2^2$ so that (2.1) readily yields the sharp oracle inequality in expectation (2.4) below. Bounds with high probability are obtained as follows. Let $L \subset \mathbb{R}^n$ be a closed convex cone. By (1.24), we have $|\Pi_L(\boldsymbol{g})|_2 = \sup_{\boldsymbol{x}\in L:|\boldsymbol{x}|_2\le 1} \boldsymbol{x}^T\boldsymbol{g}$. Thus, by the concentration of suprema of Gaussian processes ([6], Theorem 5.8), we have

$$\mathbb{P}\big(|\Pi_L(\boldsymbol{g})|_2 > \mathbb{E}|\Pi_L(\boldsymbol{g})|_2 + \sqrt{2x}\big) \le e^{-x},$$

and by Jensen's inequality we have $(\mathbb{E}|\Pi_L(\boldsymbol{g})|_2)^2 \le \delta(L)$. Combining these two bounds, we obtain

$$\mathbb{P}\big(|\Pi_L(\boldsymbol{g})|_2 \le \delta(L)^{1/2} + \sqrt{2x}\big) \ge 1 - e^{-x}. \tag{2.3}$$

From (2.1), applying this concentration inequality to the cone $L = \mathcal{T}_{K,\boldsymbol{u}}$ yields (2.5) below.

COROLLARY 2.2. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$, let K be a closed convex subset of $\mathbb{R}^n$. If $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_{n\times n})$ then*

$$\mathbb{E}\big[\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2\big] \le \min_{\boldsymbol{u}\in K}\left[\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n}\delta(\mathcal{T}_{K,\boldsymbol{u}})\right]. \tag{2.4}$$

Furthermore, for all $x > 0$ with probability at least $1 - e^{-x}$ we have

$$\begin{aligned}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2 &\le \min_{\boldsymbol{u}\in K}\left[\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n}\big(\delta(\mathcal{T}_{K,\boldsymbol{u}})^{1/2} + \sqrt{2x}\big)^2\right] \\ &\le \min_{\boldsymbol{u}\in K}\left[\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2}{n}\big(2\delta(\mathcal{T}_{K,\boldsymbol{u}}) + 4x\big)\right].\end{aligned} \tag{2.5}$$

In the well-specified case, an upper bound similar to (2.4) was derived in [7, 19]. Oymak and Hassibi [19] also proved a worst-case lower bound that matches the upper bound.

The surveys [1, 8] provide general recipes to bound from above the statistical dimension of cones of several types. For instance, the statistical dimension of $\mathcal{S}_n^{\uparrow}$ is given by the exact formula (1.28). Bounds on the statistical dimension of a closed convex cone K can be obtained using the inequality $\delta(K) \le w(K)^2 + 1$ and by bounding from above the Gaussian width $w(K)$ using Dudley integral bound and metric entropy results. This technique is used in [16] to derive the bound (1.30).

REMARK 2.1 (Tangent cones and nonconvex LS estimators). A result similar to Proposition 2.1 and Corollary 2.2 holds for closed nonconvex sets. Let $\mathcal{V} \subset \mathbb{R}^n$ be a closed nonconvex set. The LS estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{V})$ satisfies almost surely

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \le \min_{\boldsymbol{u}\in\mathcal{V}}\Big[|\boldsymbol{u} - \boldsymbol{\mu}|_2 + 2\Big(\sup_{\boldsymbol{\theta}\in\mathcal{T}(\boldsymbol{u}):|\boldsymbol{\theta}|_2\le 1} \boldsymbol{\theta}^T\boldsymbol{\xi}\Big)\Big], \tag{2.6}$$

where the set $\mathcal{T}(\boldsymbol{u})$ is defined by $\mathcal{T}(\boldsymbol{u}) := \{t(\boldsymbol{v} - \boldsymbol{u}) \mid t \ge 0, \boldsymbol{v} \in \mathcal{V}\}$.

PROOF OF (2.6). Let $\boldsymbol{u}$ be a minimizer of the right hand side of the previous display. Let $\hat{R} := |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2$ and $R := |\boldsymbol{u} - \boldsymbol{\mu}|_2$. Inequality (1.22) can be rewritten as $\hat{R}^2 - R^2 \leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u})$ and thus

$$\hat{R} - R = \frac{\hat{R}^2 - R^2}{\hat{R} + R} \leq \frac{2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u})}{\hat{R} + R} = 2\boldsymbol{\xi}^T\hat{\boldsymbol{\theta}},$$

where $\hat{\boldsymbol{\theta}} := (1/(\hat{R} + R))(\hat{\boldsymbol{\mu}} - \boldsymbol{u})$. The vector $\hat{\boldsymbol{\theta}}$ belongs to the set $\mathcal{T}(\boldsymbol{u})$ and $\hat{\boldsymbol{\theta}}$ has norm at most one because of the triangle inequality $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 \leq \hat{R} + R$. We have proved (2.6). □

Note that because of nonconvexity, it may be impossible to compute the nonconvex LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{V})$ in practice. Performance bounds for an iterative algorithm that approximate a nonconvex LS estimator are given in [20], Theorem 2.6.

The oracle inequality of (2.6) is sharp (i.e., it has leading constant 1), but it is an oracle inequality with respect to the loss $\|\cdot\|$ rather than to the squared loss $\|\cdot\|^2$. An oracle inequality with respect to the loss $\|\cdot\|^2$ is stronger than an oracle inequality with respect to the loss $\|\cdot\|$. Indeed, if an estimator $\hat{\boldsymbol{\mu}}$ satisfies $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in E} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + r_n$ for some set E and some $r_n > 0$, then the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$ yields $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \min_{\boldsymbol{u}\in E} \|\boldsymbol{u} - \boldsymbol{\mu}\| + r_n^{1/2}$. Thus, the oracle inequality (2.6) is weaker than the oracle inequality obtained in the convex case studied in Proposition 2.1.

Let $M > 1$ be an integer. Consider the special case of a union of closed convex sets $\mathcal{V} = K_1 \cup \cdots \cup K_M$ where $K_j \subset \mathbb{R}^n$ is a closed convex set for all $j = 1, \dots, M$. Then the LS estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{V})$ satisfies

$$(2.7)\quad |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \leq \min_{\boldsymbol{u}\in\mathcal{V}}\Big[|\boldsymbol{u} - \boldsymbol{\mu}|_2 + 2\sigma\Big(\max_{j=1,\dots,M} \delta(\mathcal{T}_{K_j,\boldsymbol{u}})^{1/2} + \sqrt{2(x + \log M)}\Big)\Big]$$

with probability at least $1 - e^{-x}$.

PROOF OF (2.7). In this case, inequality (2.6) yields that almost surely

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2 + 2 \max_{j=1,\dots,M}\Big(\sup_{\boldsymbol{\theta}\in\mathcal{T}_{K_j,\boldsymbol{u}}:|\boldsymbol{\theta}|_2\leq 1} \boldsymbol{\theta}^T\boldsymbol{\xi}\Big).$$

Applying the concentration inequality (2.3) to $L = \mathcal{T}_{K_j,\boldsymbol{u}}$ for all $j = 1, \dots, M$ and using the union bound completes the proof of (2.7) □

An application of (2.7) to unimodal regression is given in Appendix C.

REMARK 2.2 (Proximal operator). A result similar to Corollary 2.2 holds for estimators defined by the proximal mapping $\hat{\boldsymbol{\mu}} = \mathrm{argmin}_{\boldsymbol{v}\in\mathbb{R}^n} H(\boldsymbol{v})$ where $H(\boldsymbol{v}) = |\boldsymbol{v} - \mathbf{y}|_2^2 + 2\gamma(\boldsymbol{v})$ and $\gamma : \mathbb{R}^n \to [0, +\infty]$ is a proper convex function. The function $\boldsymbol{v} \to |\boldsymbol{v} - \mathbf{y}|_2^2$ is 1-strongly convex. The function H is the sum of the convex function $\gamma(\cdot)$ and a 1-strongly convex function, thus H is also 1-strongly convex, that is,

$$H(\hat{\boldsymbol{\mu}}) \leq H(\boldsymbol{u}) + \mathbf{d}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \qquad \text{for all } \boldsymbol{u} \in \mathbb{R}^n, \tag{2.8}$$

where $\mathbf{d}$ is any vector in the subdifferential of H at $\hat{\boldsymbol{\mu}}$. Let $\boldsymbol{u} \in \mathbb{R}^n$ be deterministic. As $\hat{\boldsymbol{\mu}}$ is a minimizer of $H(\cdot)$, one can take $\mathbf{d} = \mathbf{0}$ in (2.8). This can be rewritten as

$$\mathcal{E} := |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 - |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 \leq 2\big(\gamma(\boldsymbol{u}) - \gamma(\hat{\boldsymbol{\mu}}) - \boldsymbol{\xi}^T(\boldsymbol{u} - \hat{\boldsymbol{\mu}})\big) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2.$$

By monotonicity of the subdifferential, for all $\mathbf{h}$ in the subdifferential of $\gamma(\cdot)$ at $\boldsymbol{u}$ we have $\gamma(\boldsymbol{u}) - \gamma(\hat{\boldsymbol{\mu}}) \leq \mathbf{h}^T(\boldsymbol{u} - \hat{\boldsymbol{\mu}})$. Using the Cauchy–Schwarz inequality and the elementary inequality $2ab - b^2 \leq a^2$ yields

$$\begin{aligned}\mathcal{E} &\leq 2(\mathbf{h} - \boldsymbol{\xi})^T(\boldsymbol{u} - \hat{\boldsymbol{\mu}}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \\ &\leq 2|\mathbf{h} - \boldsymbol{\xi}|_2 |\boldsymbol{u} - \hat{\boldsymbol{\mu}}|_2 - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \\ &\leq |\mathbf{h} - \boldsymbol{\xi}|_2^2.\end{aligned}$$

By taking the infimum over all $\mathbf{h}$ in the subdifferential of $\gamma(\cdot)$ at $\boldsymbol{u}$ we have established that

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 + \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2^2 \qquad \text{almost surely,} \quad \text{and} \tag{2.9}$$

$$\mathbb{E}_{\boldsymbol{\mu}} |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 + \mathbb{E}_{\boldsymbol{\mu}} \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2^2. \tag{2.10}$$

Here, the remainder term is the squared distance from the noise vector $\boldsymbol{\xi}$ to the subdifferential of γ at $\boldsymbol{u}$. This extends the result of [7, 19] to the misspecified case. A high-probability bound analogous to (2.10) can be obtained as follows. Since the subdifferential $\partial\gamma(\boldsymbol{u})$ is a closed convex set, the map $\boldsymbol{\xi} \to \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2$ is 1-Lipschitz and we have

$$\mathbb{P}\Big(\inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2 \leq \mathbb{E} \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2 + \sigma\sqrt{2x}\Big) \geq 1 - e^{-x}$$

for any $x > 0$; cf. [6], Theorem 5.6. This yields the high-probability bound

$$\begin{aligned}\mathbb{P}\Big(|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 + \Big(\mathbb{E} \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{\xi} - \mathbf{h}|_2 + \sigma\sqrt{2x}\Big)^2\Big) \\ \geq 1 - e^{-x}.\end{aligned}$$

These results can be extended to sub-Gaussian noise, see Section 6.

Interestingly, it is possible to deduce (2.4) from (2.10). To see this, let K be a closed convex set and let γ be the indicator of K, that is, $\gamma(\boldsymbol{v}) = 0$ if $\boldsymbol{v} \in K$ and $+\infty$ otherwise. In this case, the subdifferential of γ at $\boldsymbol{u}$ is the normal cone

$$\mathcal{N}_{K,\boldsymbol{u}} := \{\mathbf{h} \in \mathbb{R}^n : (\boldsymbol{v} - \boldsymbol{u})^T \mathbf{h} \leq 0 \text{ for all } \boldsymbol{v} \in K\}.$$

To deduce (2.4) from (2.10), it is enough to prove that $\mathbf{h} = \boldsymbol{\xi} - \Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{\xi})$ belongs to the normal cone $\mathcal{N}_{K,\boldsymbol{u}}$. Let $\boldsymbol{\pi} = \Pi_{\mathcal{T}_{K,\boldsymbol{u}}}(\boldsymbol{\xi})$. For any $\boldsymbol{v} \in K$ and $t > 0$, by definition of the tangent cone, $t(\boldsymbol{v} - \boldsymbol{u})$ belongs to $\mathcal{T}_{K,\boldsymbol{u}}$. By the characterization (1.19) we thus have

$$\begin{aligned} 0 &\geq (1/t)(\boldsymbol{\xi} - \boldsymbol{\pi})^T)\big(t(\boldsymbol{v} - \boldsymbol{u}) - \boldsymbol{\pi}\big) \\ &\geq (\boldsymbol{\xi} - \boldsymbol{\pi})^T(\boldsymbol{v} - \boldsymbol{u}) - (1/t)(\boldsymbol{\xi} - \boldsymbol{\pi})^T\boldsymbol{\pi}. \end{aligned}$$

Letting t go to infinity, we obtain that $0 \geq (\boldsymbol{\xi} - \boldsymbol{\pi})^T(\boldsymbol{v} - \boldsymbol{u})$ for all $\boldsymbol{v} \in K$, that is, $\mathbf{h} = \boldsymbol{\xi} - \pi$ belongs to the normal cone $\mathcal{N}_{K,\boldsymbol{u}}$.

2.2. *Localized Gaussian widths.* In this section, we develop yet another technique to derive sharp oracle inequalities for LS estimators over closed convex sets. This technique is associated with localized Gaussian widths rather than statistical dimensions of tangent cones. The result is given in Theorem 2.3 below. Recently, other general methods have been proposed [11, 21, 24], but these methods did not provide oracle inequalities with leading constant 1.

THEOREM 2.3. *Let K be a closed convex subset of $\mathbb{R}^n$, let $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that for some $\boldsymbol{u} \in K$, there exists $t_*(\boldsymbol{u}) > 0$ such that*

$$\mathbb{E} \sup_{\boldsymbol{v}\in K:|\boldsymbol{v}-\boldsymbol{u}|_2 \leq t_*(\boldsymbol{u})} \boldsymbol{\xi}^T(\boldsymbol{v} - \boldsymbol{u}) \leq \frac{t_*(\boldsymbol{u})^2}{2}. \tag{2.11}$$

Then for any $x > 0$, with probability greater than $1 - e^{-x}$,

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2 - \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 \leq \frac{(t_*(\boldsymbol{u}) + \sigma\sqrt{2x})^2}{n} \leq \frac{2t_*^2(\boldsymbol{u}) + 4\sigma^2 x}{n}. \tag{2.12}$$

Furthermore, $\mathbb{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \boldsymbol{\mu}\|^2 \leq \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{2t_*^2(\boldsymbol{u}) + 4\sigma^2}{n}$.

PROOF. The proof of Theorem 2.3 is related to the isomorphic method [2]. Let $t = t_*(\boldsymbol{u})$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ for brevity. The concentration inequality for suprema of Gaussian processes [6], Theorem 5.8, yields that on an event $\Omega(x)$ of probability greater than $1 - e^{-x}$,

$$Z := \sup_{\boldsymbol{v}\in K:|\boldsymbol{v}-\boldsymbol{u}|_2 \leq t} \boldsymbol{\xi}^T(\boldsymbol{v} - \boldsymbol{u}) \leq \mathbb{E}[Z] + t\sigma\sqrt{2x} \leq t^2/2 + t\sigma\sqrt{2x}.$$

On the one hand, if $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 \leq t$, then by (2.2) on $\Omega(x)$ we have

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 - |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 \leq 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) - |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2^2 \leq 2Z \leq t^2 + 2t\sigma\sqrt{2x} \leq (t + \sigma\sqrt{2x})^2.$$

On the other hand, if $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 > t$, then $\alpha := t/|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2$ belongs to $(0, 1)$. If $\boldsymbol{v} = \alpha\hat{\boldsymbol{\mu}} + (1 - \alpha)\boldsymbol{u}$ then $\alpha(\hat{\boldsymbol{\mu}} - \boldsymbol{u}) = \boldsymbol{v} - \boldsymbol{u}$, by convexity of K we have $\boldsymbol{v} \in K$ and by definition of α, the equality $|\boldsymbol{v}-\boldsymbol{u}|_2 = t$ holds. On $\Omega(x)$,

$$2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}}-\boldsymbol{u}) - |\hat{\boldsymbol{\mu}}-\boldsymbol{u}|_2^2 = (2/\alpha)\boldsymbol{\xi}^T(\boldsymbol{v}-\boldsymbol{u}) - t^2/\alpha^2 \leq (2/\alpha)Z - t^2/\alpha^2$$
$$= (2t/\alpha)(Z/t) - t^2/\alpha^2 \leq (Z/t)^2 \leq (t+\sigma\sqrt{2x})^2,$$

where we used $2ab - b^2 \leq a^2$ with $b = t/\alpha$ and $a = Z/t$. Thus (2.12) holds on $\Omega(x)$ for both cases $|\hat{\boldsymbol{\mu}}-\boldsymbol{u}|_2 \leq t$ and $|\hat{\boldsymbol{\mu}}-\boldsymbol{u}|_2 > t$. Finally, the elementary inequality $(u+v)^2 \leq 2u^2 + 2v^2$ yields that $(t+\sigma\sqrt{2x})^2 \leq 2t^2 + 4\sigma^2 x$.

Integration of (2.12) yields the bound in expectation. □

Note that condition (2.11) does not depend on the true vector $\boldsymbol{\mu}$, but only depends on the vector $\boldsymbol{u}$ that appears on the right-hand side of the oracle inequality. The left-hand side of (2.11) is the Gaussian width of K localized around $\boldsymbol{u}$. This differs from the recent analysis of [9] where the Gaussian width localized around $\boldsymbol{\mu}$ is studied. An advantage of considering the Gaussian width localized around $\boldsymbol{u}$ is that the resulting oracle inequality (2.12) is sharp, that is, with leading constant 1. Chatterjee [9] proved that the Gaussian width localized around $\boldsymbol{\mu}$ characterizes a deterministic quantity $t_{\boldsymbol{\mu}}$ such that $|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2$ concentrates around $t_{\boldsymbol{\mu}}$. This result from [9] grants both an upper bound and a lower bound on $|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2$, but it does not imply nor is implied by a sharp oracle inequality such as (2.12) above. Thus, the result of [9] is of a different nature than (2.12).

A strategy to find a quantity t_* that satisfies (2.11) is to use metric entropy results together with Dudley integral bound, although Dudley integral bound may not be tight ([6], Section 13.1, Exercises 13.4 and 13.5).

REMARK 2.3. A referee pointed out the following argument to derive oracle inequalities in deviation from a bound in expectation. It is observed in [23] that if K is closed and convex, the function $\boldsymbol{\xi} \to |\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2$ is 1-Lipschitz and by the concentration of a Lipschitz function of a standard normal random variable [6], Theorem 5.6, we have

$$|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2 \leq \mathbb{E}|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2 + \sigma\sqrt{2x}$$

with probability at least $1 - e^{-x}$. This holds for *any* closed convex set K. In the well-specified setting, this means that one can always obtain risk bounds in deviation from a bound on the expected error $\mathbb{E}|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2$. In the misspecified setting, assume that $\hat{\boldsymbol{\mu}}^{\text{LS}}(K)$ satisfy a sharp oracle inequality in expectation of the form $\mathbb{E}|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u}-\boldsymbol{\mu}|_2^2 + r_n$ for some $\boldsymbol{u} \in K$ and some quantity r_n. Then the elementary inequality $(a+b)^2 \leq (1+\epsilon)x^2 + (1+1/\epsilon)y^2$ yields that

$$|\hat{\boldsymbol{\mu}}^{\text{LS}}(K) - \boldsymbol{\mu}|_2^2 \leq (1+\epsilon)|\boldsymbol{u}-\boldsymbol{\mu}|_2^2 + (1+1/\epsilon)r_n$$

holds with probability at least $1 - e^{-x}$ for any $\epsilon > 0$. Thus, thanks to the Lipschitz concentration argument from [23], an oracle inequality in expectation with leading constant 1 implies an oracle inequality in deviation with leading constant $(1+\epsilon)$.

REMARK 2.4 (Nonconvex analog of Theorem 2.3). It is possible to generalize the previous result to the LS estimator over a nonconvex set $\mathcal{V}$.

PROPOSITION 2.4. *Let $\mathcal{V}$ be a closed subset of $\mathbb{R}^n$, let $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and that for some $\boldsymbol{u} \in \mathcal{V}$, there exists $t_*(\boldsymbol{u}) > 0$ such that*

$$\mathbb{E}\Big(\sup_{\alpha\in[0,1]}\Big[\sup_{\boldsymbol{v}\in\mathcal{V}:|\boldsymbol{v}-\boldsymbol{u}|_2\le t_*(\boldsymbol{u})} \alpha\boldsymbol{\xi}^T(\boldsymbol{v}-\boldsymbol{u})\Big]\Big) \le t_*(\boldsymbol{u})^2/2.$$

Then for any $x > 0$, with probability greater than $1 - e^{-x}$,

$$|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{V}) - \boldsymbol{\mu}|_2 \le |\boldsymbol{u} - \boldsymbol{\mu}|_2 + t_*(\boldsymbol{u}) + 2\sigma\sqrt{2x}.$$

PROOF. Let $t = t_*(\boldsymbol{u})$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{V})$ for brevity. Let $\hat{R} := |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2$ and $R := |\boldsymbol{u} - \boldsymbol{\mu}|_2$. If $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 \le t$ then the claim is trivial because of the triangle inequality. Thus, we only treat the case $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 > t$. Inequality (1.22) can be rewritten as $\hat{R}^2 - R^2 \le 2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u})$ and thus

$$\hat{R} - R = \frac{\hat{R}^2 - R^2}{\hat{R} + R} \le \frac{2\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u})}{\hat{R} + R} = (2/t)\hat{\alpha}\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u}),$$

where $\hat{\alpha} = t/(R + \hat{R})$. As we treat the case $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 > t$, the triangle inequality $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 \le \hat{R} + R$ yields that $\hat{\alpha} \in [0, 1]$. Thus the right-hand side of the previous display is bounded from above by $(2/t)Z_t$ where

$$Z_t := \sup_{\alpha\in[0,1]}\Big[\sup_{\boldsymbol{v}\in\mathcal{V}:|\boldsymbol{v}-\boldsymbol{u}|_2\le t} \alpha\boldsymbol{\xi}^T(\boldsymbol{v}-\boldsymbol{u})\Big]. \tag{2.13}$$

The concentration of a supremum of a Gaussian process yields that with probability at least $1 - e^{-x}$, we have $Z_t \le \mathbb{E}[Z_t] + t\sigma\sqrt{2x} \le t^2/2 + t\sigma\sqrt{2x}$ and the proof is complete. □

2.3. *Estimation of the projection of the true parameter.* Let K be a subset of $\mathbb{R}^n$ that represents the underlying regression model. Model misspecification allows that the true parameter $\boldsymbol{\mu}$ does not belong to K. Let $\hat{\boldsymbol{\mu}}$ be an estimator of $\boldsymbol{\mu}$. This section compares different ways to measure the error of the estimator $\hat{\boldsymbol{\mu}}$ under misspecification. The proposition below compares the regret of order 1 [cf. (1.13)], the regret of order 2 [cf. (1.13)] as well as the estimation error of $\Pi_K(\boldsymbol{\mu})$ [cf. (1.14)].

If $\hat{\boldsymbol{\mu}}$ is valued in K, it is clear that $R_1(\hat{\boldsymbol{\mu}}) \ge 0$, $R_2(\hat{\boldsymbol{\mu}}) \ge 0$ and that $R_1(\hat{\boldsymbol{\mu}})^2 \le R_2(\hat{\boldsymbol{\mu}})$ by using the elementary inequality $(a - b)^2 \le |a^2 - b^2|$ for all $a, b \ge 0$.

If K is closed and convex, then the triangle inequality yields

$$\begin{aligned} R_1(\hat{\boldsymbol{\mu}}) &= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \min_{\boldsymbol{u}\in K}\|\boldsymbol{u} - \boldsymbol{\mu}\| \\ &= \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \|\Pi_K(\boldsymbol{\mu}) - \boldsymbol{\mu}\| \\ &\le \|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|. \end{aligned}$$

Furthermore, if $\hat{\boldsymbol{\mu}}$ is valued in K and K is closed and convex, then (1.20) with $\mathbf{y}$ replaced by $\boldsymbol{\mu}$ and $\boldsymbol{u}$ replaced by $\hat{\boldsymbol{\mu}}$ can be rewritten as

$$\|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|^2 \leq \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\boldsymbol{\mu} - \Pi_K(\boldsymbol{\mu})\|^2 \quad \text{for all } \hat{\boldsymbol{\mu}} \in K.$$

Thus, if K is convex, for any estimator $\hat{\boldsymbol{\mu}}$ valued in K we have $R_1^2(\hat{\boldsymbol{\mu}}) \leq \|\hat{\boldsymbol{\mu}} - \Pi_E(\boldsymbol{\mu})\|^2 \leq R_2(\hat{\boldsymbol{\mu}})$. The following proposition sums up the relationship between the quantity (1.14) and the regrets of order 1 and 2 in the case of a closed convex set K.

PROPOSITION 2.5 (Misspecification inequalities). *Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $K \subset \mathbb{R}^n$ be a closed convex set. Then $\min_{\boldsymbol{u} \in K} \|\boldsymbol{u} - \boldsymbol{\mu}\| = \|\Pi_K(\boldsymbol{\mu}) - \boldsymbol{\mu}\|$ and for any $\hat{\boldsymbol{\mu}} \in K$, the following holds almost surely*:

$$\begin{aligned}(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| - \|\Pi_K(\boldsymbol{\mu}) - \boldsymbol{\mu}\|)^2 &\leq \|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|^2 \\ &\leq \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 - \|\Pi_K(\boldsymbol{\mu}) - \boldsymbol{\mu}\|^2.\end{aligned}$$

Estimation of $\Pi_K(\boldsymbol{\mu})$ by the LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K)$ has been considered for instance in [25], Section 4, for $K = \mathcal{S}_n^{\uparrow}$, and in [16], Section 6, for $K = \mathcal{S}_n^{\mathrm{C}}$. Proposition 2.5 above shows that for any quantity r_n and any estimator $\hat{\boldsymbol{\mu}}$ valued in a closed convex set K, we have

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \|\Pi_K(\boldsymbol{\mu}) - \boldsymbol{\mu}\|^2 + r_n \quad \text{implies} \quad \|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|^2 \leq r_n,$$

that is, a sharp oracle inequality with leading constant 1 automatically implies an upper bound on the estimation error $\|\hat{\boldsymbol{\mu}} - \Pi_K(\boldsymbol{\mu})\|^2$. Finally, the following corollary is a consequence of Proposition 2.5, Proposition 2.1 and Theorem 2.3.

COROLLARY 2.6. *Let K be a closed convex set and let $\boldsymbol{\mu} \in \mathbb{R}^n$. Then, almost surely,*

$$\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \Pi_K(\boldsymbol{\mu})\| \leq (\sigma/\sqrt{n})|\Pi_{\mathcal{T}_{K,\Pi_K(\boldsymbol{\mu})}}(\boldsymbol{g})|_2,$$

where $\boldsymbol{g} = (1/\sigma)\boldsymbol{\xi}$. Furthermore, if $t_* > 0$ is such that

$$\mathbb{E}\Big[\sup_{\boldsymbol{v} \in K : |\boldsymbol{v} - \Pi_K(\boldsymbol{\mu})|_2 \leq t_*} \boldsymbol{\xi}^T(\boldsymbol{v} - \Pi_K(\boldsymbol{\mu}))\Big] \leq \frac{t_*^2}{2},$$

then for all $x > 0$, with probability at least $1 - e^{-x}$ we have

$$|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \Pi_K(\boldsymbol{\mu})|_2 \leq t_* + \sigma\sqrt{2x}.$$

These results highlight a major advantage of oracle inequalities with leading constant 1 over oracle inequalities with leading constant strictly greater than 1. Indeed, oracle inequalities with leading constant 1 yield an upper bound on the estimation error $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K) - \Pi_K(\boldsymbol{\mu})\|$ for any closed convex set K. The results of this section will be applied to $K = \mathcal{S}_n^{\uparrow}$ and $K = \mathcal{S}_n^{\mathrm{C}}$ in Section 5.

3. Sharp bounds in isotonic regression. We study in this section the performance of $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ using the general tools developed in the previous section. We first apply Corollary 2.2. To do so, we need to bound from above the statistical dimension of the tangent cone $\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}}$. In fact, it is possible to characterize the tangent cone $\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}}$ and to obtain a closed formula for its statistical dimension.

PROPOSITION 3.1. *Let $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ and let $k = k(\boldsymbol{u})$. Let $(T_1, \dots, T_k)$ be a partition of $\{1, \dots, n\}$ such that $\boldsymbol{u}$ is constant on each T_j, $j = 1, \dots, k$. Then*

$$\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}} = \mathcal{S}^{\uparrow}{}_{|T_1|} \times \cdots \times \mathcal{S}^{\uparrow}{}_{|T_k|}.$$

PROOF. Let $\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}} = \mathcal{T}$ for brevity. If $\boldsymbol{u}$ is constant, then it is clear that $\mathcal{T} = \mathcal{S}_n^{\uparrow}$ so we assume that $\boldsymbol{u}$ has at least one jump, that is, $k(\boldsymbol{u}) \geq 2$. As $\mathcal{S}_n^{\uparrow}$ is a cone we have $\mathcal{T} = \{\boldsymbol{v} - t\boldsymbol{u} \mid t \geq 0, \boldsymbol{v} \in \mathcal{S}_n^{\uparrow}\}$. Thus, the inclusion $\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}} \subset \mathcal{S}^{\uparrow}{}_{|T_1|} \times \cdots \times \mathcal{S}^{\uparrow}{}_{|T_k|}$ is straightforward. For the reverse inclusion, we use the following argument based on [12], Remark 4.1. Let $\boldsymbol{x} \in \mathcal{S}^{\uparrow}{}_{|T_1|} \times \cdots \times \mathcal{S}^{\uparrow}{}_{|T_k|}$ and let $\varepsilon > 0$ be the minimal jump of the sequence $\boldsymbol{u}$, that is, $\varepsilon = \min_{i=1,\dots,n-1: u_{i+1} > u_i}(u_{i+1} - u_i)$. If $t = |\boldsymbol{x}|_\infty/(4\varepsilon)$, then the vector $\boldsymbol{v} := t\boldsymbol{u} + \boldsymbol{x}$ belongs to $\mathcal{S}_n^{\uparrow}$, which completes the proof. □

Using (1.26) and (1.29), we obtain $\delta(\mathcal{T}_{\mathcal{S}_n^{\uparrow},\boldsymbol{u}}) = \sum_{j=1}^{k(\boldsymbol{u})} \sum_{t=1}^{|T_j|} \frac{1}{t} \leq \sum_{j=1}^{k(\boldsymbol{u})} \log(e \times |T_j|)$. Following [11], Example 2.2, this quantity is bounded from above by $k(\boldsymbol{u}) \log(en/k(\boldsymbol{u}))$ by Jensen's inequality. Applying Corollary 2.2, leads to the following result.

THEOREM 3.2. *For all $n \geq 2$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$,*

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{3.1}$$

Furthermore, for any $x > 0$ we have with probability greater than $1 - \exp(-x)$,

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{2\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right) + \frac{4\sigma^2 x}{n}. \tag{3.2}$$

Let us discuss some features of Theorem 3.2 that are new. First, the estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ satisfies oracle inequalities both in deviation with exponential probability bounds and in expectation; cf. (3.2) and (3.1), respectively. Previously known oracle inequalities of this type for the LS estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ were only proved in expectation.

Second, both (3.1) and (3.2) are sharp oracle inequalities, that is, with leading constant 1. Although sharp oracle inequalities were obtained using aggregation methods [5], this is the first known sharp oracle inequality for the LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$.

Third, the assumption $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ is not needed, as opposed to the result of [11].

Last, the constant 1 in front of $\frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})}$ in (3.1) is optimal for the LS estimator. To see this, assume that there exists an absolute constant $c < 1$ such that for all $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ and $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$,

$$\mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}} \left(\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{c\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})} \right). \tag{3.3}$$

Set $\boldsymbol{\mu} = 0$. Thanks to (1.29), the left-hand side of the above display is bounded from below by $\sigma^2 \log(n)/n$ while while the right-hand side is equal to $c\sigma^2 \log(e \times n)/n$. Thus, it is impossible to improve the constant in front of $\frac{\sigma^2 k(\boldsymbol{u})}{n} \log \frac{en}{k(\boldsymbol{u})}$ for the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. However, it is still possible that for another estimator $\hat{\boldsymbol{\mu}}$, (3.3) holds with $c < 1$ or without the logarithmic factor. We do not know whether such an estimator exists.

We now highlight the adaptive behavior of the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$. Let $\boldsymbol{u}^* \in \mathcal{S}_n^{\uparrow}$ be a minimizer of the right-hand side of (3.1). Let $k = k(\boldsymbol{u}^*)$ and let $(T_1, \dots, T_k)$ be a partition of $\{1, \dots, n\}$ such that $\boldsymbol{u}^*$ is constant on all T_j, $j = 1, \dots, k$. Given $T_1, \dots, T_k$, consider the piecewise constant oracle

$$\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}} \in \operatorname*{argmin}_{\boldsymbol{u} \in W_{T_1,\dots,T_k}} \|\mathbf{y} - \boldsymbol{u}\|^2,$$

where $W_{T_1,\dots,T_k}$ is the linear subspace of all sequences that are constant on all T_j, $j = 1, \dots, k$. This subspace has dimension k, so the estimator $\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}}$ satisfies

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}} - \boldsymbol{\mu}\|^2 &= \min_{\boldsymbol{u} \in W_{T_1,\dots,T_k}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n} \\ &\leq \|\boldsymbol{u}^* - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k}{n}. \end{aligned}$$

Furthermore, if $\boldsymbol{\mu} \in \mathcal{S}_n^{\uparrow}$ then $\min_{\boldsymbol{u} \in W_{T_1,\dots,T_k}} \|\boldsymbol{u} - \boldsymbol{\mu}\|^2 = \|\boldsymbol{u}^* - \boldsymbol{\mu}\|^2$ and if R^* is the expected prediction error of the oracle, that is, $R^* = \mathbb{E}\|\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}} - \boldsymbol{\mu}\|^2$, then $R^* = \frac{\sigma^2 k}{n} + \|\boldsymbol{u}^* - \boldsymbol{\mu}\|^2$. By Theorem 3.2, we obtain

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 &\leq \|\boldsymbol{u}^* - \boldsymbol{\mu}\|^2 + \frac{\sigma^2 k \log(en/k)}{n} \\ &= R^* + \frac{\sigma^2 k \log(n/k)}{n} \\ &\leq R^* \log(en/k). \end{aligned}$$

Thus, (3.1) can be interpreted in the sense that without the knowledge of $T_1, \dots, T_k$, the performance of $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ is similar to that of $\hat{\boldsymbol{\mu}}^{\mathrm{ORACLE}}$ up to the logarithmic factor $\log(en/k)$. Of course, the knowledge of $T_1, \dots, T_k$ is not accessible in practice, so $\hat{\boldsymbol{\mu}}^{\text{ORACLE}}$ is an oracle that can only serve as a benchmark. This adaptive behavior of $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ was observed in [11].

We now apply Theorem 2.3. Using Dudley integral bound and the entropy bounds from [15], Chatterjee [9] established that there exists an absolute constant $c > 0$ such that if $\boldsymbol{u} \in \mathcal{S}_n^{\uparrow}$ and $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$, then

$$\begin{aligned} &\mathbb{E} \sup_{\boldsymbol{\theta}\in\mathcal{S}_n^{\uparrow}:|\boldsymbol{\theta}-\boldsymbol{u}|_2\le t} \boldsymbol{\xi}^T(\boldsymbol{\theta}-\boldsymbol{u}) \le \frac{t^2}{16} \\ &\qquad \text{for all } t \ge t_*(\boldsymbol{u}) := c\sigma\big(1+V(\boldsymbol{u})/\sigma\big)^{1/3} n^{1/6}. \end{aligned} \tag{3.4}$$

The constant 16 is arbitrary and could be replaced by any positive numerical constant provided that the numerical constant c is modified accordingly. This bound combined with Theorem 2.3 yields the following.

COROLLARY 3.3. *There exists an absolute constant $c > 0$ such that the following holds. Let $n \ge 2$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$. Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow}) - \boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u}\in\mathcal{S}_n^{\uparrow}}\left[\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{2c\sigma^2(1+V(\boldsymbol{u})/\sigma)^{2/3}}{n^{2/3}}\right] + \frac{4\sigma^2 x}{n}. \tag{3.5}$$

As this is an application of Theorem 2.3, the corresponding bound in expectation also holds. The novelty of Corollary 3.3 is that the leading constant is 1. Although model misspecification was considered in [16, 25], no oracle inequalities were obtained. As we will see in Section 5, oracle inequalities with leading constant 1 such as (3.5) yield that under misspecification, the LS estimator $\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\uparrow})$ consistently estimates $\Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu})$ at the rate $n^{-2/3}$.

4. Convex regression and arbitrary design points.

The goal of this section is to study univariate convex regression for nonequispaced design points.

4.1. *Parametric rate for any design if $\boldsymbol{\mu}$ has few affine pieces.*

We now present a new argument to bound from above the statistical dimension of the cone of convex sequences.

THEOREM 4.1. *Let $n \ge 3$. Let $x_1 < \cdots < x_n$ be real numbers and consider the cone $K^C_{x_1,\dots,x_n}$ defined in (1.16). Let $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. Then*

$$\delta(K^C_{x_1,\dots,x_n}) = \mathbb{E}[|\Pi_{K^C_{x_1,\dots,x_n}}(\boldsymbol{g})|_2^2] \le 8\log(en). \tag{4.1}$$

PROOF. Let $K = K^C_{x_1,\dots,x_n}$ for brevity. A convex sequence $\boldsymbol{u} = (u_1,\dots,u_n) \in K$ is first nonincreasing and then nondecreasing, that is, there exists $m \in \{1,\dots,n\}$ such that $u_1 \geq u_2 \geq \cdots \geq u_m \leq u_{m+1} \leq \cdots \leq u_n$, hence the sequence $\boldsymbol{u}$ is unimodal. Thus, if $K_m, m = 1,\dots,n$ are the sets defined in (1.17), then $K \subset \mathcal{U} = \bigcup_{m=1,\dots,n} K_m$ where the sets $K_m, m = 1,\dots,n$ are defined in (1.17). Using (1.26), (1.27) and (1.29) we obtain

$$\delta(K_m) \leq \delta\big(\mathcal{S}^{\downarrow}_m \times \mathcal{S}^{\uparrow}_{n-m}\big) \leq \log(em) + \log\big(e(n-m)\big) \leq 2\log(en).$$

By (1.24), almost surely we have

$$\begin{aligned} 0 \leq \big|\Pi_K(\boldsymbol{g})\big|_2 &= \sup_{\boldsymbol{u}\in K:|\boldsymbol{u}|_2\leq 1} \boldsymbol{g}^T\boldsymbol{u} \\ &\leq \max_{m=1,\dots,n} \sup_{\boldsymbol{u}\in K_m:|\boldsymbol{u}|_2\leq 1} \boldsymbol{g}^T\boldsymbol{u} = \max_{m=1,\dots,n} \big|\Pi_{K_m}(\boldsymbol{g})\big|_2. \end{aligned}$$

Using (2.3) and the union bound, for all $x > 0$, we have with probability at least $1 - e^{-x}$ the inequality $|\Pi_K(\boldsymbol{g})|_2^2 \leq \max_{m=1,\dots,n} \delta(K_m)^{1/2} + \sqrt{2(x + \log n)}$. As $(a+b)^2 \leq 2a^2 + 2b^2$, on the same event of probability at least $1 - e^{-x}$,

$$\begin{aligned} \big|\Pi_K(\boldsymbol{g})\big|_2^2 &\leq 2 \max_{m=1,\dots,n} \delta(K_m) + 4(x + \log n) \\ &\leq 4\log(en) + 4(x + \log n). \end{aligned}$$

Integration of this probability bound completes the proof. □

Remarkably, this bound on the statistical dimension does not depend on the design points $x_1,\dots,x_n$. Furthermore, the bound (4.1) improves upon (1.30) as the exponent $5/4$ is reduced to 1.

PROPOSITION 4.2. *Let $n \geq 3$, and let $\boldsymbol{u}$ be an element of the cone $K^C_{x_1,\dots,x_n}$ defined in* (1.16). *The statistical dimension of the tangent cone at $\boldsymbol{u}$ satisfies*

$$\delta(\mathcal{T}_{K^C_{x_1,\dots,x_n},\boldsymbol{u}}) \leq 8q(\boldsymbol{u})\log\bigg(\frac{en}{q(\boldsymbol{u})}\bigg).$$

PROOF. Let $q = q(\boldsymbol{u})$. Let $(T_1,\dots,T_q)$ be a partition of $\{1,\dots,n\}$ such that $\boldsymbol{u}$ is affine on each T_j, $j = 1,\dots,q$. Let $\boldsymbol{x} \in K^C_{x_1,\dots,x_n}$. A convex sequence minus an affine sequence is convex, thus for all $j = 1,\dots,q$, $(\boldsymbol{x} - \boldsymbol{u})_{T_j}$ is convex in the sense that it belongs to $K^C_{x_i:i\in T_j}$. Thus,

$$\mathcal{T}_{K^C_{x_1,\dots,x_n},\boldsymbol{u}} \subset \mathcal{C} := K^C_{x_i:i\in T_1} \times K^C_{x_i:i\in T_2} \times \cdots \times K^C_{x_i:i\in T_q}.$$

Using (1.27), (1.26), Theorem 4.1 and Jensen's inequality, we have

$$\begin{aligned}\delta(\mathcal{T}_{K^C_{x_1,\dots,x_n},\boldsymbol{u}}) \le \delta(\mathcal{C}) &\le \sum_{j=1}^{q} 8\log\bigl(e|T_j|\bigr)\\ &\le 8q\log\Biggl(\frac{e}{q}\sum_{j=1}^{q}|T_j|\Biggr) = 8q\log\biggl(\frac{en}{q}\biggr).\end{aligned}$$

□

Combining Corollary 2.2 and Proposition 4.2 yields the following.

THEOREM 4.3. *Let $n \ge 3$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Let $x_1 < \cdots < x_n$ be real numbers. Then for any $x > 0$, the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K^C_{x_1,\dots,x_n})$ satisfies*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \le \min_{\boldsymbol{u}\in K^C_{x_1,\dots,x_n}} \biggl(\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{16\sigma^2 q(\boldsymbol{u})}{n}\log\frac{en}{q(\boldsymbol{u})}\biggr) + \frac{4\sigma^2 x}{n}$$

with probability greater than $1 - \exp(-x)$.

As this is an application of Corollary 2.2, the corresponding bound in expectation also holds. Theorem 4.3 does not depend on the design points $x_1, \dots, x_n$. In particular, Theorem 4.3 and the corresponding result in expectation hold for nonequispaced design points and design points that can be arbitrarily close to each other. This improves upon the oracle inequality (1.10) proved in [11, 16] where C is strictly greater than 1 and depends on the design points through the ratio (1.12).

For all $q \ge 2$, define the linear function $\Delta_q : \mathbb{R}^q \to \mathbb{R}^{q-1}$ by

$$\Delta_q(\boldsymbol{u}) = (u_2 - u_1, u_3 - u_2, \dots, u_q - u_{q-1})^T$$

for all $\boldsymbol{u} = (u_1, \dots, u_q)^T$, so that $\mathcal{S}_n^{\uparrow} = \{\boldsymbol{u} \in \mathbb{R}^n : \Delta_n \boldsymbol{u} \ge \mathbf{0}\}$ and $\mathcal{S}_n^{\mathrm{C}} = \{\boldsymbol{u} \in \mathbb{R}^n : \Delta_{n-1}\Delta_n \boldsymbol{u} \ge \mathbf{0}\}$. It is possible to define higher-order cones as follows. For $q \ge 3$, define $\Delta_q^2 : \mathbb{R}^q \to \mathbb{R}^{q-2}$ by $\Delta_q^2 = \Delta_{q-1} \circ \Delta_q$ and for all $\beta = 1, \dots, q-1$ define $\Delta_q^\beta : \mathbb{R}^q \to \mathbb{R}^{q-\beta}$ by

$$\Delta_q^\beta = \Delta_{q-\beta+1} \circ \cdots \circ \Delta_{q-1} \circ \Delta_q.$$

For any positive integer $\beta < n$, define the cone

$$\mathcal{S}_n^{[\beta]} := \bigl\{\boldsymbol{u} = (u_1, \dots, u_n)^T \in \mathbb{R}^n : \Delta_n^\beta \boldsymbol{u} \ge \mathbf{0} = (0, \dots, 0)^T\bigr\}.$$

In particular, $\mathcal{S}_n^{[1]} = \mathcal{S}_n^{\uparrow}$ is the cone of nondecreasing sequences and $\mathcal{S}_n^{[2]} = \mathcal{S}_n^{\mathrm{C}}$ is the cone of convex sequences. Theorems 1, 2 and 3 in the supplementary material [3] generalize Theorems 3.2, 4.1 and 4.3 to the cones $\mathcal{S}^{[\beta]}$ for $\beta \ge 3$.

We now turn to the nonparametric rate of convex regression. The next section shows that, unlike to adaptive bound of Theorem 4.3 which holds irrespective of design points, the nonparametric rate of convex regression varies substantially based on the design points $x_1, \dots, x_n$.

4.2. *Worst-case design points in convex regression and the rate $n^{-2/3}$.* The nonparametric rate for estimation of convex sequences is of order $n^{-4/5}$ for equispaced design points. This was established in [16] using metric entropy bounds and an extra logarithmic factor present in [16] was later removed in [10]. It is proved in [10], (3.3), that

$$\mathbb{E} \sup_{\boldsymbol{\theta}\in\mathcal{S}_n^{\uparrow}:|\boldsymbol{\theta}-\boldsymbol{u}|_2\leq t} \boldsymbol{\xi}^T(\boldsymbol{\theta}-\boldsymbol{u}) \leq \frac{t^2}{2} \qquad \text{for all } t \geq c\sigma\left(1+\frac{V(\boldsymbol{u})}{\sigma}\right)^{1/5} n^{1/10}.$$

We can combine this bound on the localized Gaussian width with Theorem 2.3 to obtain the following sharp oracle inequality.

COROLLARY 4.4. *There exists an absolute constant $c > 0$ such that the following holds. Let $n \geq 3$ and $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$. Then for any $x > 0$, with probability greater than $1 - \exp(-x)$,*

$$\|\hat{\boldsymbol{\mu}}^{\text{LS}}(\mathcal{S}_n^{\text{C}}) - \boldsymbol{\mu}\|^2 \leq \min_{\boldsymbol{u}\in\mathcal{S}_n^{\text{C}}}\left[\|\boldsymbol{u}-\boldsymbol{\mu}\|^2 + \frac{2c\sigma^2(1+V(\boldsymbol{u})/\sigma)^{2/5}}{n^{4/5}}\right] + \frac{4\sigma^2 x}{n}.$$

As this is an application of Theorem 2.3, the corresponding bound in expectation also holds. It is natural to ask whether the nonparametric rate $n^{-4/5}$ can be achieved by the LS estimator for any collection design points. The following result provides a negative answer: There exist design points such that no estimator can achieve a better rate than $n^{-2/3}$. This rate $n^{-2/3}$ is substantially slower than the nonparametric rate $n^{-4/5}$ achieved by the LS estimator in convex regression with equispaced design points.

THEOREM 4.5. *Let $V \geq \sigma/\sqrt{n}$. There exists design points $x_1 < \cdots < x_n$ that depend on V such that, for any estimator $\hat{\boldsymbol{\mu}}$,*

$$\sup_{\boldsymbol{\mu}\in K^C_{x_1,\dots,x_n}\cap\mathcal{S}_n^{\uparrow}:\mu_n-\mu_1\leq 2V} \mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\|^2 \geq C\sigma^2\max\left[\left(\frac{V^2}{\sigma^2 n^2}\right)^{1/3}, \frac{1}{n}\right]\right) \geq c,$$

where $c, C > 0$ are absolute constants.

PROOF. It was proved in [5], Proposition 4 and Corollary 5, that for some integer $M \geq 2$, there exist $\boldsymbol{\mu}_0, \dots, \boldsymbol{\mu}_M \in \mathcal{S}_n^{\uparrow}$ such that $V(\boldsymbol{\mu}_j) \leq V$ and

$$\text{(4.2)} \quad \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\| \geq \frac{C\sigma^2}{4}\max\left[\left(\frac{V^2}{\sigma^2 n^2}\right)^{1/3}, \frac{1}{n}\right], \qquad \frac{n}{2\sigma^2}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_0\| \leq \frac{\log M}{16}$$

for all distinct $j, k \in \{0, \dots, M\}$ and some absolute constant $C > 0$. The quantity $\frac{n}{2\sigma^2}\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_0\|$ is the Kullback–Leibler divergence from $\mathcal{N}(\boldsymbol{\mu}_j, \sigma^2 I_{n\times n})$ to $\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2 I_{n\times n})$.

Define $\boldsymbol{v} = (v_1, \dots, v_n)^T$ by $v_i = iV/n$ for all $i = 1, \dots, n$ so that $V(\boldsymbol{v}) \leq V$ and $\boldsymbol{v}$ is strictly increasing. We define $\boldsymbol{u}^0, \dots, \boldsymbol{u}^M$ by $\boldsymbol{u}^j = \boldsymbol{\mu}_j + \boldsymbol{v}$ so that $\boldsymbol{u}^0, \dots, \boldsymbol{u}^M$ are strictly increasing. Furthermore, since $\boldsymbol{\mu}_j - \boldsymbol{\mu}_k = \boldsymbol{u}^j - \boldsymbol{u}^k$ it is clear that (4.2) still holds if $\boldsymbol{\mu}_j, \boldsymbol{\mu}_k$ are replaced by $\boldsymbol{u}^j, \boldsymbol{u}^k$. Applying [22], Theorem 2.7, yields that for any estimator $\hat{\boldsymbol{\mu}}$,

$$\sup_{j=0,\dots,M} \mathbb{P}_{\boldsymbol{u}_j}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \geq C\sigma^2 \max\left[\left(\frac{V^2}{\sigma^2 n^2}\right)^{1/3}, \frac{1}{n}\right]\right) \geq c,$$

where $c, C > 0$ are absolute constants.

Let $\epsilon := \frac{1}{2} \wedge \min_{j=1,\dots,M} \min_{i=2,\dots,n-1} \frac{u_{i+1}^j - u_i^j}{u_i^j - u_{i-1}^j}$. Since the sequences $\boldsymbol{u}^0, \dots, \boldsymbol{u}^M$ are strictly increasing we have $\epsilon > 0$. Define the design points $x_1 < \cdots < x_n$ by $x_i = -\epsilon^i$ for all $i = 1, \dots, n$. Then for all $j = 0, \dots, M$ we have

$$\frac{x_{i+1} - x_i}{x_i - x_{i-1}} = \epsilon \leq \frac{u_{i+1}^j - u_i^j}{u_i^j - u_{i-1}^j}$$

for all $i = 2, \dots, n-1$, and by (1.16) this implies that $\boldsymbol{u}^j \in K_{x_1,\dots,x_n}^C$. It remains to show that $V(\boldsymbol{u}^j) \leq 2V$, which is a consequence of $V(\boldsymbol{\mu}_j) \leq V$ and $V(\boldsymbol{v}) \leq V$. □

This result emphasizes the importance of the design points in univariate convex regression. For equispaced design points the rate is of order $n^{-4/5}$, but for nonequispaced design points the rate can be substantially slower. Inspection of the proof reveals that the design points of Theorem 4.5 are contained in $[-1, 0)$ and they concentrate around the boundary at 0. If the practitioner can choose the design points, then design points that concentrate around the boundaries should be avoided.

Any convex function is unimodal so that the inclusion $K_{x_1,\dots,x_n}^C \subset \mathcal{U}$ holds for any design points $x_1 < \cdots < x_n$. Intuitively, this inclusion means that convexity brings more structure than unimodality. Theorem 4.6 below shows that the convex LS enjoys essentially the same risk bounds and oracle inequalities as those satisfied by the unimodal LS estimator in Appendix C.

THEOREM 4.6. *Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $x_1 < \cdots < x_n$ be any real numbers. Then for all $x > 0$, with probability at least $1 - 2e^{-x}$, the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K_{x_1,\dots,x_n}^C)$ satisfies*

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|$$
$$\leq \min_{\boldsymbol{u} \in \mathcal{U}}\left[\|\boldsymbol{u} - \boldsymbol{\mu}\| \vee \frac{\sigma}{\sqrt{n}}\left(2\sqrt{(k(\boldsymbol{u}) + 1)\log\left(\frac{en}{k(\boldsymbol{u}) + 1}\right)} + 3\sqrt{2(x + \log n)}\right)\right].$$

The proof of Theorem 4.6 is given in Appendix B. Theorem 4.6 can be readily used to prove that the convex LS estimator achieves the rate $n^{-2/3}$ (up to logarithmic factors) for any collection of design points. We proceed as follows. Let $V = V(\boldsymbol{\mu})$ be the total variation of $\boldsymbol{\mu}$, where $\boldsymbol{\mu} \in K^C_{x_1,\dots,x_n}$ is an unknown convex sequence. Let $k = 1, \dots, n$ be an integer that will be specified later. The approximation argument developed in [5], Lemma 2, for isotonic regression can be trivially extended to unimodal sequence: There exists a unimodal sequence $\boldsymbol{u}$ with at most $2k$ constant pieces and such that $|\boldsymbol{\mu} - \boldsymbol{u}|_\infty \leq V/(2k)$. Theorem 4.6 with $x = \log n$ yields that, with probability at least $1 - 2/n$, we have

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \left(\frac{V}{2k}\right) \vee \frac{\sigma}{\sqrt{n}}\left(2\sqrt{(2k+1)\log\left(\frac{en}{2k+1}\right)} + 6\sqrt{\log n}\right).$$

If $V\sqrt{n} \geq \sigma$, then choose $k \geq 1$ as an integer such that $k \leq (V^2 n/(\sigma^2))^{1/3} < 2k$. With this choice of k, bounding from above the right-hand side of the previous display yields that, with probability at least $1 - 2/n$, we have

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq \left(\frac{\sigma^{2/3} V^{1/3}}{n^{1/3}}\right) \vee \left(\frac{2\sqrt{3}\sigma^{2/3} V^{1/3}}{n^{1/3}}\sqrt{\log(en)} + \frac{6\sigma\sqrt{\log n}}{\sqrt{n}}\right).$$

If $V\sqrt{n} < \sigma$, then we choose $k = 1$, and in this case we obtain $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \leq (2\sqrt{3} + 6)\sigma\sqrt{\log(n)/n}$. Thus, we obtain the estimation rate $n^{-2/3}$ for the squared loss $\|\cdot\|^2$, up to logarithmic factors.

The following result provides an alternate proof that the estimation rate of convex regression for any collection is design point is of order $n^{-2/3}$. It is an application of Theorem 2.3.

THEOREM 4.7. *There exists an absolute constant $c > 0$ such that the following holds. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and let $x_1 < \cdots < x_n$ be any real numbers. Then for all $x > 0$, with probability at least $1 - 2e^{-x}$, the estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K^C_{x_1,\dots,x_n})$ satisfies*

$$\begin{aligned}
&\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \\
&\quad \leq \min_{\boldsymbol{u} \in \mathcal{K}^C_{x_1,\dots,x_n}} \left[\|\boldsymbol{u} - \boldsymbol{\mu}\|^2 + 2\max\left(c\sigma^2\left(\frac{\sigma + V(\boldsymbol{u})}{\sigma n}\right)^{2/3}, 64\sigma^2\log(n)\right)\right] \\
&\qquad + \frac{4\sigma^2 x}{n}.
\end{aligned}$$

The proof of Theorem 4.7 is given in Appendix B. The novelty of Theorem 4.7 is that the rate $n^{-2/3}$ is achieved by the convex LS estimator for *any* collection of univariate design points. This rate $n^{-2/3}$ is substantially slower than the rate $n^{-4/5}$ of convex regression with equispaced design points.

Together, Theorems 4.7 and 4.5 establish that this rate is minimal over all design points and all univariate convex functions with bounded total variation. To make this precise, define the minimax quantity

$$\mathfrak{R}(V) := \sup_{x_1 < \cdots < x_n} \inf_{\hat{\boldsymbol{\mu}}} \sup_{\boldsymbol{\mu} \in K^C_{x_1,\dots,x_n} : V(\boldsymbol{\mu}) \le V} \mathbb{E}_{\boldsymbol{\mu}}[\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2], \qquad V > 0,$$

where the first supremum is taken over all $x_1, \dots, x_n \in \mathbb{R}$ such that $x_1 < \cdots < x_n$ and the infimum is taken over all estimators that may depend on $x_1, \dots, x_n$ (for instance, the convex LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K^C_{x_1,\dots,x_n})$ depends on the design points). The quantity $\mathfrak{R}(V)$ represents the minimax risk over all possible univariate design points, and over all convex sequences. By taking $\boldsymbol{u} = \boldsymbol{\mu}$ in Theorem 4.7 and by integration, we obtain that

$$\begin{aligned} \mathfrak{R}(V) &\le \sup_{x_1 < \cdots < x_n} \sup_{\boldsymbol{\mu} \in K^C_{x_1,\dots,x_n} : V(\boldsymbol{\mu}) \le V} \mathbb{E}_{\boldsymbol{\mu}}[\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(K^C_{x_1,\dots,x_n}) - \boldsymbol{\mu}\|^2] \\ &\le C\sigma^2 \left(\frac{\sigma + V}{\sigma n}\right)^{2/3} \end{aligned}$$

for some absolute constant $C > 0$. On the other hand, Theorem 4.5 and Markov inequality yield that $\mathfrak{R}(V) \ge C'\sigma^2(\frac{V}{\sigma n})^{2/3}$ for some absolute constant $C' > 0$ provided that $V \ge \sigma$. This establishes that the nonparametric rate of univariate convex regression over all possible design points is of order $n^{-2/3}$. This rate is substantially slower than the rate $n^{-4/5}$ observed by [16] for equispaced design points. In summary, there is no hope to achieve the nonparametric rate $n^{-4/5}$ for any univariate design points.

As a convex function is unimodal, the inclusion $K^C_{x_1,\dots,x_n} \subset \mathcal{U}$ holds. The convex constraints that define $K^C_{x_1,\dots,x_n}$ are more restrictive than the unimodal constraint, that is, convexity brings more structure than unimodality. For equispaced design points, the extra structure brought by convexity yields a nonparametric rate of order $n^{-4/5}$ which is faster than the unimodal nonparametric rate $n^{-2/3}$; cf. Appendix C. However, for some worst-case design points, this extra structure is uninformative from a statistical standpoint: The nonparametric rates of convex and unimodal regression are of the same order $n^{-2/3}$.

5. Estimation of $\boldsymbol{\Pi}_K(\boldsymbol{\mu})$ in isotonic and convex regression.

In this section, we exhibit some consequences of Section 2.3 in isotonic and convex regression. If $K = \mathcal{S}_n^{\uparrow}$, Proposition 2.5, Corollary 2.2 and Theorem 3.2 with $\boldsymbol{u} = \Pi_K(\boldsymbol{\mu})$ imply the following. If $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ and $\boldsymbol{\pi} = \Pi_{\mathcal{S}_n^{\uparrow}}(\boldsymbol{\mu})$, then

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\pi}\| \le \sigma\sqrt{\frac{k(\boldsymbol{\pi})}{n}\log\left(\frac{en}{k(\boldsymbol{\pi})}\right)} + \sigma\sqrt{\frac{2x}{n}}\right) \ge 1 - e^{-x},$$

while Proposition 2.5, Theorem 2.3 and Corollary 3.3 yield

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\pi}\|^2 \leq 2c\sigma^2\left(\frac{\sigma + V(\boldsymbol{\pi})}{\sigma n}\right)^{2/3} + \frac{4\sigma^2 x}{n}\right) \geq 1 - e^{-x},$$

for any $\boldsymbol{\mu} \in \mathbb{R}^n$, where $c > 0$ is an absolute constant. That is, in the misspecified case, the LS estimator $\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\uparrow})$ estimates $\boldsymbol{\pi}$ at the parametric rate if $\boldsymbol{\pi}$ has few constant pieces, and at the nonparametric rate $n^{-2/3}$ otherwise.

Similar conclusions can be drawn in convex regression from Theorem 4.3 and Corollary 4.4. If $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{S}_n^{\mathrm{C}})$ and $\boldsymbol{\pi} = \Pi_{\mathcal{S}_n^{\mathrm{C}}}(\boldsymbol{\mu})$ we have

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\pi}\|^2 \leq \frac{16\sigma^2 q(\boldsymbol{\pi})}{n} \log \frac{en}{q(\boldsymbol{\pi})} + \frac{4\sigma^2 x}{n}\right) \geq 1 - e^{-x},$$

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\pi}\|^2 \leq \frac{C\sigma(1 + V(\boldsymbol{\pi})/\sigma)^{2/5}\log(en)}{n^{4/5}} + \frac{16\sigma^2 x}{n}\right) \geq 1 - e^{-x}.$$

In other words, under misspecification, the LS estimator consistently estimates the projection $\Pi_K(\boldsymbol{\mu})$ and the rate of estimation of $\Pi_K(\boldsymbol{\mu})$ under misspecification is at least as fast as the rate of estimation of $\boldsymbol{\mu}$ when the model is well specified.

6. Extension to sub-Gaussian noise. We explain in this section that if the noise random vector has independent sub-Gaussian components, then the general results given in Corollary 2.2 and Theorem 2.3 still hold. The argument below relies on the well-studied contraction principle ([17], Section 4.2) and was used for a similar purpose in [4].

We use below the following elementary fact about the composition of convex functions. Let $g : \mathbb{R}^n \to [0, +\infty)$ and $f : [0, +\infty) \to [0, +\infty)$ be two functions. If the functions f, g are both convex and if f is nondecreasing then $F = f \circ g$ is also convex. In particular, if $f(t) = t^2$ or $f(t) = e^{\lambda t}$ and $g(\boldsymbol{v}) = \sup_{\boldsymbol{\theta} \in T} \boldsymbol{v}^T\boldsymbol{\theta}$ for a set T with $\{\mathbf{0}\} \subset T \subset \mathbb{R}^n$, then $F = f \circ g$ is convex.

PROPOSITION 6.1. *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. The conclusions of Corollary 2.2 still hold if the assumption $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ is replaced by*

$$\mathbb{E}F(\boldsymbol{\xi}) \leq \mathbb{E}F(\sigma\boldsymbol{x}) \qquad \textit{for all convex function } F : \mathbb{R}^n \to [0, +\infty). \tag{6.1}$$

PROOF. Let $L = \mathcal{T}_{K,\boldsymbol{u}}$ where $K, \boldsymbol{u}$ are defined in Corollary 2.2. To prove (2.4) from the almost-sure inequality (2.1), we only need $\mathbb{E}|\Pi_L(\boldsymbol{\xi})|_2^2 \leq \sigma^2\delta(L)$, which is granted by (6.1) for the convex function F defined by $F(\boldsymbol{v}) = |\Pi_L(\boldsymbol{v})|_2^2 = (\sup_{\boldsymbol{\theta}\in L:|\boldsymbol{\theta}|_2\leq 1} \boldsymbol{v}^T\boldsymbol{\theta})^2$; cf. (1.24) for the last equality.

To prove (2.5) from (6.1), it is sufficient to prove the concentration inequality (2.3). Let $\lambda \geq 0$. Applying (6.1) to the convex function F defined by $F(\boldsymbol{v}) = \exp(\lambda|\Pi_L(\boldsymbol{v})|_2)$, we have

$$\log \mathbb{E}F(\boldsymbol{\xi}) \leq \log \mathbb{E}F(\sigma\boldsymbol{x}) \leq \lambda\mathbb{E}|\Pi_L(\sigma\boldsymbol{x})|_2 + \sigma^2\lambda^2/2 \leq \lambda\sigma\delta(L)^{1/2} + \sigma^2\lambda^2/2,$$

where we used [6], Theorems 5.5 and 5.8, and Jensen's inequality for the two last inequalities. A Chernoff bound yields (2.3) for any random vector $\boldsymbol{\xi}$ that satisfy (6.1). □

PROPOSITION 6.2 (Sub-Gaussian analog of Remark 2.2). *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$. Consider the setting of Remark 2.2. If we drop the assumption $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$ and instead assume that (6.1) holds, then we have*

$$\mathbb{E}_{\boldsymbol{\mu}} |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 + \mathbb{E}_{\boldsymbol{\mu}} \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\sigma \boldsymbol{x} - \mathbf{h}|_2^2$$

and

$$|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2^2 \leq |\boldsymbol{u} - \boldsymbol{\mu}|_2^2 + \Big(\mathbb{E} \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\sigma \boldsymbol{x} - \mathbf{h}|_2 + \sigma\sqrt{2x}\Big)^2$$

with probability at least $1 - e^{-x}$.

PROOF. We start from (2.9). The map $d : \boldsymbol{a} \to \inf_{\mathbf{h} \in \partial\gamma(\boldsymbol{u})} |\boldsymbol{a} - \mathbf{h}|_2$ is the distance to the set $\partial\gamma(\boldsymbol{u})$, hence it is convex. The function $t \to t^2$ is convex increasing on $[0, +\infty)$, thus the function $F : \boldsymbol{a} \to d(\boldsymbol{a})^2$ is also convex. Applying (6.1) yields that $\mathbb{E}[d(\boldsymbol{\xi})^2] \leq \mathbb{E}[d(\sigma\boldsymbol{x})^2]$ which completes the proof of the oracle inequality in expectation.

For the high-probability bound, define the function $F(\boldsymbol{a}) = \exp(\lambda d(\boldsymbol{a}))$. The function F is convex and the function d is 1-Lipschitz. By (6.1) and [6], Theorems 5.5 and 5.8, we have

$$\log \mathbb{E} F(\boldsymbol{\xi}) \leq \log \mathbb{E} F(\sigma \boldsymbol{x}) \leq \lambda \mathbb{E}\big[d(\sigma \boldsymbol{x})\big] + \sigma^2\lambda^2/2.$$

A Chernoff bound completes the proof. □

PROPOSITION 6.3 (Sub-Gaussian analog of Theorem 2.3). *Let K be a closed convex subset of $\mathbb{R}^n$, let $\boldsymbol{\mu} \in \mathbb{R}^n$. Assume that the noise random vector $\boldsymbol{\xi}$ satisfy (6.1) and that for some $\boldsymbol{u} \in K$, there exists $t_*(\boldsymbol{u}) > 0$ such that*

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})}\Big[\sup_{\boldsymbol{v} \in K : |\boldsymbol{v} - \boldsymbol{u}|_2 \leq t_*(\boldsymbol{u})} (\sigma \boldsymbol{x})^T(\boldsymbol{v} - \boldsymbol{u})\Big] \leq \frac{t_*(\boldsymbol{u})^2}{2}.$$

Then for any $x > 0$, (2.12) holds with probability greater than $1 - e^{-x}$.

PROOF. Let $t = t_*(\boldsymbol{u})$. The conclusions of Theorem 2.3 hold on the event

$$\Big\{\sup_{\boldsymbol{v} \in K : |\boldsymbol{v} - \boldsymbol{u}|_2 \leq t} \boldsymbol{\xi}^T(\boldsymbol{v} - \boldsymbol{u}) \leq t^2/2 + t\sigma\sqrt{2x}\Big\}. \tag{6.2}$$

Let $\lambda \geq 0$. Define the functions g and F by $g(\boldsymbol{v}) = \sup_{\boldsymbol{\theta} \in K : |\boldsymbol{\theta} - \boldsymbol{u}|_2 \leq t} \boldsymbol{v}^T(\boldsymbol{\theta} - \boldsymbol{u})$ and $F(\boldsymbol{v}) = \exp(\lambda g(\boldsymbol{v}))$. By (6.1) and [6], Theorems 5.5, we have

$$\log \mathbb{E} F(\boldsymbol{\xi}) \leq \log \mathbb{E} F(\sigma \boldsymbol{x}) \leq \lambda \mathbb{E}\big[g(\sigma \boldsymbol{x})\big] + \frac{\sigma^2\lambda^2 t^2}{2} \leq \frac{\lambda t^2 + \sigma^2\lambda^2 t^2}{2}.$$

A Chernoff bound yields that the event (6.2) has probability at least $1-e^{-x}$, which completes the proof. □

Thus, Corollary 2.2 and Theorem 2.3 both hold for any random vector $\boldsymbol{\xi}$ that satisfies (6.1). By the contraction argument [17], Lemmas 4.4 and 4.6, we explain below that any random vector with independent sub-Gaussian components satisfy (6.1).

PROPOSITION 6.4. *Let $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, I_{n\times n})$ and let $x_1, \dots, x_n$ be the components of $\boldsymbol{x}$. Assume that the components $\xi_1, \dots, \xi_n$ of $\boldsymbol{\xi}$ are symmetric, independent and satisfy*

$$\mathbb{P}(|\xi_i| > t) \le 8\mathbb{P}\left(\frac{\sigma|x_i|}{8} > t\right) \qquad \forall t > 0, i = 1, \dots, n. \tag{6.3}$$

Then (6.1) *holds.*

In other words, (6.1) is satisfied if the tail of ξ_i is bounded from above by 8 times the tail of a normal random variable $\mathcal{N}(0, \sigma/8)$. The constant 8 above is arbitrary, it could be replaced by any large enough numerical constant. In fact, the property (6.3) holds for every sub-Gaussian random variables: If the ξ_i satisfies $\mathbb{E}[e^{(8\xi_i/\sigma)^2}] \le e$ then (6.3) holds; cf. [4], Lemma H.1, where a similar contraction argument is used. As our framework and notation are different than the setting of [17], we provide below a concise proof of this argument.

PROOF OF PROPOSITION 6.4. Let $B_1, \dots, B_n$ be i.i.d. Bernoulli random variables with $\mathbb{P}(B_i = 1) = 1/8 = 1 - \mathbb{P}(B_i = 0)$, independent of $\boldsymbol{x}$ and $\boldsymbol{\xi}$. For all $i = 1, \dots, n$ we have $\mathbb{P}(|B_i\xi_i| > t) \le \mathbb{P}((\sigma/8)|x_i| > t)$ so that it is possible, using inverse distribution function, to define the random variables B_i, ξ_i, x_i on a large enough probability space such that $|B_i\xi_i| \le (\sigma/8)|x_i|$ almost surely. Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. Rademacher random variables with $\mathbb{P}(\epsilon_i = \pm 1) = 1/2$, independent of all other random variables. Let $\mathbb{E}_\epsilon$ be the conditional expectation with respect to $\epsilon_1, \dots, \epsilon_n$ conditionally on $(B_i, \xi_i, x_i)_{i=1,\dots,n}$. For all $\boldsymbol{\alpha} \in [-1, +1]^n$, define $G(\boldsymbol{\alpha}) = \mathbb{E}_\epsilon F((\sigma/8)\epsilon_1\alpha_1 x_1, \dots, (\sigma/8)\epsilon_n\alpha_n x_n)$. Then we have

$$\begin{aligned} \mathbb{E}_\epsilon F(\epsilon_1 B_1\xi_1, \dots, \epsilon_n B_n\xi_n) &\le \sup_{\boldsymbol{\alpha}\in[-1,1]^n} G(\boldsymbol{\alpha}) \\ &= \sup_{\boldsymbol{\alpha}\in\{-1,1\}^n} G(\boldsymbol{\alpha}) \\ &= G(1, 1, \dots, 1). \end{aligned}$$

The first equality holds because $G(\cdot)$ is convex and a convex function a maximized at an extreme point of $[-1, +1]^n$, that is, at some $\boldsymbol{\alpha}^* \in \{-1, 1\}^n$. The second inequality holds because ϵ_i has the same distribution as $\alpha_i^*\epsilon_i$.

As $1/8 = \mathbb{E}B_i$ for all $i = 1, \dots, n$, by Jensen's inequality with respect to $B_1, \dots, B_n$ we have

$$\begin{aligned}\mathbb{E}F\left(\frac{\boldsymbol{\xi}}{8}\right) &= \mathbb{E}F\left(\frac{\epsilon_1\xi_1}{8}, \dots, \frac{\epsilon_n\xi_n}{8}\right)\\ &\le \mathbb{E}F(\epsilon_1 B_1\xi_1, \dots, \epsilon_n B_n\xi_n)\\ &\le \mathbb{E}G(1, \dots, 1)\\ &= \mathbb{E}F\left(\frac{\sigma \boldsymbol{x}}{8}\right),\end{aligned}$$

where the first and last equalities are consequences of the fact that for all $i = 1, \dots, n$, $\xi_i\epsilon_i$ is equal in distribution to ξ_i and $x_i\epsilon_i$ is equal in distribution to x_i. □

Finally, the requirement that the components of $\boldsymbol{\xi}$ are symmetric can be relaxed by adding symmetrization step [17], Lemma 6.3.

APPENDIX A: PROOF PRELIMINARIES

Recall that the sets $K_1, \dots, K_n, \mathcal{U}$ are defined in (1.17) and (1.18).

LEMMA A.1. *Let $\boldsymbol{u} \in \mathcal{U}$ be a unimodal sequence. For all $m = 1, \dots, n$, the statistical dimension of the tangent cone of K_m at $\boldsymbol{u}$ satisfies*

$$\delta(\mathcal{T}_{K_m,\boldsymbol{u}}) \le \big(k(\boldsymbol{u}) + 1\big) \log\left(\frac{en}{k(\boldsymbol{u}) + 1}\right).$$

PROOF. Let $m = 1, \dots, n$, $k = k(\boldsymbol{u})$ and let $(T_1, \dots, T_k)$ be a partition of $\{1, \dots, n\}$ such that $\boldsymbol{u}$ is constant on each T_l and T_l is convex for all $l = 1, \dots, k$. Let $l^* \in \{1, \dots, k\}$ be the unique integer such that $m \in T_{l^*}$, and let $T^* = T_{l^*}$. Let $\boldsymbol{v} \in K_m$. Then for all $l < l^*$, the sequence $(\boldsymbol{v} - \boldsymbol{u})_{T_l}$ is nonincreasing and for all $l > l^*$, the sequence $(\boldsymbol{v} - \boldsymbol{u})_{T_l}$ is nondecreasing. Furthermore, if $A = T^* \cap \{1, \dots, m\}$ and $B = T^* \cap \{m + 1, \dots, n\}$, the sequence $(\boldsymbol{v} - \boldsymbol{u})_A$ is nonincreasing and the sequence $(\boldsymbol{v} - \boldsymbol{u})_B$ is nondecreasing. We have proved the inclusion

$$\mathcal{T}_{K_m,\boldsymbol{u}} \subset \mathcal{C} := \mathcal{S}^{\downarrow}_{|T_1|} \times \cdots \times \mathcal{S}^{\downarrow}_{|T_{l^*-1}|} \times \mathcal{S}^{\downarrow}_{|A|} \times \mathcal{S}^{\uparrow}_{|B|} \times \mathcal{S}^{\uparrow}_{|T_{l^*+1}|} \times \cdots \times \mathcal{S}^{\uparrow}_{|T_k|},$$

where for all integer $q \ge 1$, $\mathcal{S}^{\uparrow}_q$ is the cone of nondecreasing sequences in $\mathbb{R}^q$ and $\mathcal{S}^{\downarrow}_q$ is the cone of nonincreasing sequences in $\mathbb{R}^q$. Using (1.27), (1.26) and (1.29), we obtain

$$\delta(\mathcal{T}_{K_m,\boldsymbol{u}}) \le \delta(\mathcal{C}) \le \log\big(e|A|\big) + \log\big(e|B|\big) + \sum_{l=1,\dots,k:l\ne l^*} \log\big(e|T_l|\big).$$

Using Jensen's inequality with the fact that $|A| + |B| + \sum_{l=1,\dots,k:l\ne l^*} |T_l| = n$, we obtain $\delta(\mathcal{T}_{K_m,\boldsymbol{u}}) \le (k + 1) \log \frac{en}{k+1}$. □

LEMMA A.2. *Let $\boldsymbol{u} \in \mathcal{U}$ and let Z_t be the random variable* (2.13) *with $\mathcal{V} = \mathcal{U}$. Then if $t_*(\boldsymbol{u})$ is defined as the right-hand side of* (3.4) *and $t = \max(t_*(\boldsymbol{u}), 8\sigma\sqrt{\log n})$, we have $\mathbb{E}[Z_t] \le t^2/2$.*

PROOF. Let $\lambda > 0$ be a constant that will be specified later. Let also $Z_{t,m}$ be the random variable (2.13) with $\mathcal{V} = K_m$. Then clearly $Z_t = \max_{m=1,\dots,n} Z_{t,m}$. By [6], Theorem 5.5, we have

$$\begin{aligned}\mathbb{E}\exp(\lambda Z_t) &\le \sum_{m=1}^{n} \mathbb{E}\exp(\lambda Z_{t,m}) \\ &\le \sum_{m=1}^{n} \exp\bigl(\lambda\mathbb{E}[Z_{t,m}] + \lambda^2 t^2\sigma^2/2\bigr).\end{aligned} \tag{A.1}$$

Let $m = 1,\dots,n$. We now bound $\mathbb{E}[Z_{t,m}]$ from above. Assume without loss of generality that the mode of $\boldsymbol{u}$ is after m, that is, $\boldsymbol{u} \in K_j$ for some $j \ge m$. Let $T := \{1,\dots,m\}$, $E := \{m+1,\dots,j-1\}$ and $S := \{j,\dots,n\}$. Then for all $\boldsymbol{v} \in K_m$, by definition of T, E and S we have

$$\boldsymbol{v}_T \in \mathcal{S}^{\downarrow}_{|T|}, \qquad \boldsymbol{u}_T \in \mathcal{S}^{\downarrow}_{|T|}, \qquad (\boldsymbol{v}-\boldsymbol{u})_E \in \mathcal{S}^{\uparrow}_{|E|}, \qquad \boldsymbol{v}_S \in \mathcal{S}^{\uparrow}_{|S|}, \qquad \boldsymbol{u}_S \in \mathcal{S}^{\uparrow}_{|S|}.$$

Thus, the quantity $\mathbb{E}[Z_{t,m}]$ is bounded from above by

$$\mathbb{E} \sup_{\alpha\in[0,1]} \sup_{\boldsymbol{x}\in\mathcal{S}^{\downarrow}_{|T|}:|\boldsymbol{x}-\boldsymbol{u}_T|_2\le t} \alpha\boldsymbol{\xi}_T^T(\boldsymbol{x}-\boldsymbol{u}_T) \tag{A.2}$$

$$+\mathbb{E} \sup_{\alpha\in[0,1]} \sup_{\boldsymbol{x}\in\mathcal{S}^{\uparrow}_{|E|}:|\boldsymbol{x}|_2\le t} \alpha\boldsymbol{\xi}_E^T\boldsymbol{x} \tag{A.3}$$

$$+\mathbb{E} \sup_{\alpha\in[0,1]} \sup_{\boldsymbol{x}\in\mathcal{S}^{\uparrow}_{|S|}:|\boldsymbol{x}-\boldsymbol{u}_S|_2\le t} \alpha\boldsymbol{\xi}_S^T(\boldsymbol{x}-\boldsymbol{u}_S). \tag{A.4}$$

The set $S^{\uparrow}_{|E|}$ is a cone so the supremum $\sup_{\alpha\in[0,1]}$ can be dropped from the second term (A.3). For the first term (A.2), we have $\alpha(\boldsymbol{x}-\boldsymbol{u}_T) = (\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{u}_T) - \boldsymbol{u}_T = \boldsymbol{x}' - \boldsymbol{u}_T$ and $\boldsymbol{x}' \in \mathcal{S}^{\downarrow}_{|T|}$ by convexity. Thus the supremum $\sup_{\alpha\in[0,1]}$ can be dropped. The same argument allows us to drop $\sup_{\alpha\in[0,1]}$ in (A.4). Using (3.4) and $V(\boldsymbol{u}) \ge \max(V(\boldsymbol{u}_T), V(\boldsymbol{u}_S))$ we get that $\mathbb{E}[Z_{t,m}] \le 3t^2/16$ for all $m = 1,\dots,n$. We plug this bound into (A.1) with $\lambda = 1/(8\sigma^2)$ to obtain

$$\begin{aligned}\mathbb{E}e^{\lambda Z_t} &\le \sum_{m=1}^{n} e^{\lambda\mathbb{E}[Z_{t,m}]+\lambda^2t^2\sigma^2/2} \\ &\le ne^{\lambda(3t^2/16+t^2/16)} \\ &= e^{\lambda(t^2/4+8\sigma^2\log n)}.\end{aligned}$$

Jensen's inequality implies $\mathbb{E}[Z_t] \le t^2/4 + 8\sigma^2 \log(n)$. As $t \ge 8\sigma\sqrt{\log n}$, we have established that $\mathbb{E}[Z_t] \le t^2/2$. □

APPENDIX B: PROOFS FOR CONVEX REGRESSION

PROOF OF THEOREM 4.6. Let $\boldsymbol{u} \in \mathcal{U}$ be a unimodal sequence. Define the random variable

$$G := \boldsymbol{\xi}^T(\boldsymbol{u} - \boldsymbol{\mu})/|\boldsymbol{u} - \boldsymbol{\mu}|_2 \qquad \text{if } \boldsymbol{u} \neq \boldsymbol{\mu} \quad \text{and} \quad G = 0 \qquad \text{otherwise.}$$

Let $\mathcal{T}_{K_m,\boldsymbol{u}}$, $m = 1, \dots, n$ be the tangent cones from Lemma A.1. Define the random variable $Y_{\boldsymbol{u}}$ by

$$Y_{\boldsymbol{u}} := \sup_{\boldsymbol{v} \in \bigcup_{m=1,\dots,n} \mathcal{T}_{K_m,\boldsymbol{u}} : |\boldsymbol{v}|_2 \le 1} \boldsymbol{\xi}^T \boldsymbol{v},$$

Let $\tilde{R} := |\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}|_2$ and $R := |\boldsymbol{u} - \boldsymbol{\mu}|_2$. We first prove that, almost surely,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\| \le \max\left(\|\boldsymbol{u} - \boldsymbol{\mu}\|, \frac{2Y_{\boldsymbol{u}} + G}{\sqrt{n}} \right) \tag{B.1}$$

It is enough to prove that $\tilde{R} > R$ implies $\tilde{R} \le 2Y_{\boldsymbol{u}} + G$. Assume that $\tilde{R} > R$. Inequality (2.2) yields $\tilde{R}^2 \le \boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ and thus

$$\begin{aligned} \tilde{R}^2 &\le \boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \\ &= \frac{\boldsymbol{\xi}^T(\hat{\boldsymbol{\mu}} - \boldsymbol{u})}{|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2} |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 + G|\boldsymbol{u} - \boldsymbol{\mu}|_2 \\ &\le Y_{\boldsymbol{u}} |\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 + GR, \end{aligned}$$

where we used that $\hat{\boldsymbol{\mu}} \in \mathcal{U}$ since a convex sequence is unimodal. We have $R < \tilde{R}$ and by the triangle inequality, $|\hat{\boldsymbol{\mu}} - \boldsymbol{u}|_2 \le \tilde{R} + R < 2\tilde{R}$, which proves that $\tilde{R}^2 \le 2Y_{\boldsymbol{u}}\tilde{R} + G\tilde{R}$. Dividing by $\tilde{R}$ completes the proof of (B.1).

We now prove the oracle inequality. Since G is centered Gaussian with variance at most σ^2 we have

$$\mathbb{P}\left(G > \sigma\sqrt{2(x + \log n)}\right) \le e^{-x}/n \le e^{-x}. \tag{B.2}$$

Furthermore, using (1.24), we have $Y_{\boldsymbol{u}} = \max_{m=1,\dots,n} |\Pi_{\mathcal{T}_{K_m,\boldsymbol{u}}}(\boldsymbol{\xi})|_2$. We apply (2.3) to $L = \mathcal{T}_{K_m,\boldsymbol{u}}$ for all $m = 1, \dots, n$ and the union bound to obtain

$$\mathbb{P}\left(Y_{\boldsymbol{u}} \le \sigma \max_{m=1,\dots,n} \delta(\mathcal{T}_{K_m,\boldsymbol{u}})^{1/2} + \sigma\sqrt{2(x + \log n)}\right) \ge 1 - e^{-x}. \tag{B.3}$$

Lemma A.1 provides an upper bound on $\max_{m=1,\dots,n} \delta(\mathcal{T}_{K_m,\boldsymbol{u}})^{1/2}$. We complete the proof by combining (B.2) and (B.3) with the union bound. □

PROOF OF THEOREM 4.7. Let $\boldsymbol{u}$ be a minimizer of the right-hand side of the oracle inequality from Theorem 4.7. Let t and Z_t be defined in Lemma A.2. As $K^C_{x_1,\dots,x_n} \subset \mathcal{U}$, Lemma A.2 yields that

$$\mathbb{E} \sup_{\boldsymbol{v} \in K^C_{x_1,\dots,x_n} : |\boldsymbol{v}-\boldsymbol{u}|_2 \le t} \boldsymbol{\xi}^T(\boldsymbol{v}-\boldsymbol{u}) \le \mathbb{E}[Z_t] \le t^2/2. \tag{B.4}$$

Applying Theorem 2.3 completes the proof. □

APPENDIX C: PERFORMANCE OF THE UNIMODAL LS ESTIMATOR

The results below show that the unimodal LS estimator enjoys the same performance as the convex LS estimator in Theorems 4.6 and 4.7. The proof ingredients are also the same: Theorems 4.6 and 4.7 rely on Lemmas A.1 and A.2, and we explain below that these two Lemmas can be used to study the performance of the unimodal LS estimator. Note that the oracle inequalities (C.1)–(C.2) below first appeared in [14]; our initial result on unimodal regression only featured the risk bound (C.3) below and no oracle inequalities.

Assume in this paragraph that $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{n\times n})$. First, for any $x > 0$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$ we have

$$\begin{aligned} &\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\| \\ &\quad \le \min_{\boldsymbol{u}\in\mathcal{U}}\Bigg[\|\boldsymbol{u}-\boldsymbol{\mu}\| + \frac{2\sigma}{\sqrt{n}}\Bigg(\sqrt{(k(\boldsymbol{u})+1)\log\bigg(\frac{en}{k(\boldsymbol{u})+1}\bigg)} \\ &\qquad + \sqrt{2(x+\log n)}\Bigg)\Bigg] \end{aligned} \tag{C.1}$$

with probability at least $1 - e^{-x}$. Inequality (C.1) is a direct application of (2.7) and Lemma A.1. Second, there exists an absolute constant $c > 0$ such that for any $x > 0$ and any $\boldsymbol{\mu} \in \mathbb{R}^n$ we have

$$\begin{aligned} &\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\| \\ &\quad \le \min_{\boldsymbol{u}\in\mathcal{U}}\Bigg[\|\boldsymbol{u}-\boldsymbol{\mu}\| + \max\Bigg(c\sigma\bigg(\frac{\sigma+V(\boldsymbol{u})}{\sigma n}\bigg)^{1/3}, 8\sigma\sqrt{\frac{\log n}{n}}\Bigg)\Bigg] \\ &\qquad + \frac{2\sigma\sqrt{2x}}{\sqrt{n}} \end{aligned} \tag{C.2}$$

with probability at least $1 - e^{-x}$. Inequality (C.2) is a direct application of Proposition 2.4 and Lemma A.2. Our proofs of Theorems 4.6 and 4.7 in convex regression and (C.1)–(C.2) in unimodal regression are similar and the rates are also the same up to numerical constants. This resemblance is due to the fact that convex and unimodal regression become essentially the same problem for the worst-case design points studied in Section 4.2.

The performance of the unimodal LS estimator is also studied in [13] and [14]. Chatterjee and Lafferty [13] initially obtained inequality (C.2) in the well-specified case ($\boldsymbol{\mu} = \boldsymbol{u}$), as well as an adaptive risk bound of the form $\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\|^2 \leq \frac{C\sigma^2}{n}(k(\boldsymbol{\mu})\log(en))^{3/2}$ with high probability. An intermediary arXiv revision of the present article proved the following well-specified version of (C.1), which showed that the exponent $3/2$ could be reduced to 1. If $\boldsymbol{\mu} \in \mathcal{U}$, then

$$\begin{aligned} &\|\hat{\boldsymbol{\mu}}^{\mathrm{LS}}(\mathcal{U}) - \boldsymbol{\mu}\| \\ &\quad \leq \frac{2\sigma}{\sqrt{n}}\left(\sqrt{(k(\boldsymbol{\mu})+1)\log\left(\frac{en}{k(\boldsymbol{\mu})+1}\right)} + \sqrt{2(x+\log n)}\right) \end{aligned} \tag{C.3}$$

holds with probability at least $1 - e^{-x}$. During the writing of the arXiv revision of the present article in which (C.3) was introduced, we became aware of a similar result by Flammation et al. [14] obtained independently in the context of statistical seriation. Interestingly, (C.3) and the result of Flammation et al. [14] were proved using different techniques. Inequality (C.3) is an outcome of the concentration inequality (2.3) and of upper bounds on the statistical dimension of tangent cones, while Flammation et al. [14] prove an oracle inequality using metric entropy bounds and the variational representation studied in [9, 13]. An advantage of the proof presented here is that the numerical constants of (C.3) are explicit and reasonably small.

Acknowledgments. The author would like to thank the Associate Editor and three anonymous referees for comments that substantially improved the paper. The author also thanks Alexandre Tsybakov for helpful comments and Philippe Rigollet for informing him of the unimodal regression result of Flamarion et al. [14].

SUPPLEMENTARY MATERIAL

Supplement to "Sharp oracle inequalities for Least Squares estimators in shape restricted regression" (DOI: 10.1214/17-AOS1566SUPP; .pdf). The supplementary material contains generalizations of the results in isotonic and convex regression to higher order cones. Theorems 1, 2 and 3 in the supplementary material generalize Theorems 3.2, 4.1 and 4.3 to the cones $\mathcal{S}^{[\beta]}$ for $\beta \geq 3$.

REFERENCES

[1] Amelunxen, D., Lotz, M., McCoy, M. B. and Tropp, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453

[2] BARTLETT, P. L. and MENDELSON, S. (2006). Empirical minimization. *Probab. Theory Related Fields* **135** 311–334. MR2240689

[3] BELLEC, P. C. (2018). Supplement to "Sharp oracle inequalities for Least Squares estimators in shape restricted regression." DOI:10.1214/17-AOS1566SUPP.

[4] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2016). Slope meets lasso: Improved oracle bounds and optimality. ArXiv preprint. Available at arXiv:1605.08651.

[5] BELLEC, P. C. and TSYBAKOV, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* **16** 1879–1892. MR3417801

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193

[7] CHANDRASEKARAN, V. and JORDAN, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proc. Natl. Acad. Sci. USA* **110** E1181–E1190. MR3047651

[8] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found. Comput. Math.* **12** 805–849. MR2989474

[9] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. MR3269982

[10] CHATTERJEE, S. (2016). An improved global risk bound in concave regression. *Electron. J. Stat.* **10** 1608–1629. MR3522655

[11] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878

[12] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On matrix estimation under monotonicity constraints. ArXiv preprint. Available at arXiv:1506.03430.

[13] CHATTERJEE, S. and LAFFERTY, J. (2017). Adaptive risk bounds in unimodal regression. *Bernoulli*. To appear. Available at arXiv:1512.02956.

[14] FLAMMARION, N., MAO, C. and RIGOLLET, P. (2016). Optimal rates of statistical seriation. ArXiv preprint. Available at arXiv:1607.02435.

[15] GAO, F. and WELLNER, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.* **98** 1751–1764. MR2392431

[16] GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. MR3405621

[17] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015

[18] MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. MR1810920

[19] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131

[20] OYMAK, S., RECHT, B. and SOLTANOLKOTABI, M. (2015). Sharp time–data tradeoffs for linear inverse problems. ArXiv preprint. Available at arXiv:1507.04793.

[21] PLAN, Y., VERSHYNIN, R. and YUDOVINA, E. (2017). High-dimensional estimation with geometric constraints. *Inf. Inference* **6** 1–40. MR3636866

[22] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. MR2724359

[23] VAN DE GEER, S. and WAINWRIGHT, M. (2015). On concentration for (regularized) empirical risk minimization. ArXiv preprint. Available at arXiv:1512.00677.

[24] VERSHYNIN, R. (2015). Estimation in high dimensions: A geometric perspective. In *Sampling Theory*, *A Renaissance*. *Appl. Numer. Harmon. Anal.* 3–66. Birkhäuser, Basel. MR3467418

[25] ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. MR1902898

ENSAE
3 avenue Pierre Larousse
92245 Malakoff Cedex
France
and
Department of Statistics & Biostatistics
Rutgers University
501 Hill Center, Busch Campus
110 Frelinghuysen Road
Piscataway, New Jersey 08854
USA
E-mail: pcb71@stat.rutgers.edu