The Annals of Statistics
2015, Vol. 43, No. 4, 1774–1800
DOI: 10.1214/15-AOS1324

ON RISK BOUNDS IN ISOTONIC AND OTHER SHAPE RESTRICTED REGRESSION PROBLEMS

BY SABYASACHI CHATTERJEE, ADITYANAND GUNTUBOYINA[1] AND BODHISATTVA SEN[2]

University of Chicago, University of California, Berkeley and Columbia University

We consider the problem of estimating an unknown $\theta \in \mathbb{R}^n$ from noisy observations under the constraint that θ belongs to certain convex polyhedral cones in $\mathbb{R}^n$. Under this setting, we prove bounds for the risk of the least squares estimator (LSE). The obtained risk bound behaves differently depending on the true sequence θ which highlights the adaptive behavior of θ. As special cases of our general result, we derive risk bounds for the LSE in univariate isotonic and convex regression. We study the risk bound in isotonic regression in greater detail: we show that the isotonic LSE converges at a whole range of rates from $\log n/n$ (when θ is constant) to $n^{-2/3}$ (when θ is *uniformly increasing* in a certain sense). We argue that the bound presents a benchmark for the risk of any estimator in isotonic regression by proving nonasymptotic local minimax lower bounds. We prove an analogue of our bound for model misspecification where the true θ is not necessarily nondecreasing.

1. Introduction. Shape constrained regression involves estimating a vector $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ from observations

$$Y_i = \theta_i + \varepsilon_i \qquad \text{for } i = 1, \dots, n, \tag{1}$$

where θ lies in a known convex polyhedral cone $\mathcal{K} \subseteq \mathbb{R}^n$ and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. mean zero errors with finite variance. It may be recalled that convex polyhedral cones are sets of the form

$$\mathcal{K} := \{\theta \in \mathbb{R}^n : A\theta \geq 0\}, \tag{2}$$

where A is a matrix of order $m \times n$ and $\alpha = (\alpha_1, \dots, \alpha_m) \geq 0$ means that $\alpha_i \geq 0$ for each i. Basic properties of convex polyhedral cones can be found, for example, in [26], Chapters 7 and 8.

Received May 2014; revised February 2015.
[1]Supported by NSF Grant DMS-13-09356.
[2]Supported by NSF Grants DMS-11-50435 and AST-1107373.
MSC2010 subject classifications. 62G08, 62C20.
Key words and phrases. Adaptation, convex polyhedral cones, global risk bounds, local minimax bounds, model misspecification, statistical dimension.

In this paper, we focus on such problems when the cone $\mathcal{K}$ is of the special form

$$\mathcal{K}^n_{r,s} := \left\{\theta \in \mathbb{R}^n : \sum_{j=-r}^{s} w_j \theta_{t+j} \geq 0 \text{ for all } 1 + r \leq t \leq n - s\right\}, \tag{3}$$

for some known integers $r \geq 0$ and $s \geq 1$ and nonnegative weights w_j, $-r \leq j \leq s$. Here the integers r and s and the weights w_j, $-r \leq j \leq s$ do not depend on n. Note that when $n < 1 + r + s$, the condition in the definition of $\mathcal{K}^n_{r,s}$ is vacuous so that $\mathcal{K}^n_{r,s} = \mathbb{R}^n$. The dependence of the cone on the weights $\{w_j\}$ is suppressed in the notation $\mathcal{K}^n_{r,s}$.

The following shape constrained regression problems are special instances of our general setup:

(1) When $r = 0, s = 1, w_0 = -1$ and $w_1 = 1$, the cone in (3) consists of all nondecreasing sequences

$$\mathcal{M} := \{\theta \in \mathbb{R}^n : \theta_1 \leq \cdots \leq \theta_n\}.$$

Estimation problem (1) then becomes the well-known isotonic regression problem.

(2) When $r = 1, s = 1, w_{-1} = w_1 = 1$ and $w_0 = -2$, the cone in (3) consists of all convex sequences $\mathcal{C} := \{\theta \in \mathbb{R}^n : 2\theta_i \leq \theta_{i-1} + \theta_{i+1}, i = 2, \ldots, n-1\}$. Then (1) reduces to the usual convex regression problem with equally spaced design points.

(3) k-monotone regression corresponds to $\mathcal{K} := \{\theta \in \mathbb{R}^n : \nabla^k \theta \geq 0\}$ where $\nabla : \mathbb{R}^n \to \mathbb{R}^n$ is given by $\nabla(\theta) := (\theta_2 - \theta_1, \theta_3 - \theta_2, \ldots, \theta_n - \theta_{n-1}, 0)$, and ∇^k represents the k-times composition of ∇. This is also a special case of (3).

Our object of interest in this paper is the least squares estimator (LSE) for θ under the constraint $\theta \in \mathcal{K}$. It is given by $\hat{\theta}(Y; \mathcal{K})$ where $Y = (Y_1, \ldots, Y_n)$ is the observation vector and

$$\hat{\theta}(y; \mathcal{K}) := \operatorname*{argmin}_{\theta \in \mathcal{K}} \|\theta - y\|^2 \qquad \text{for } y \in \mathbb{R}^n, \tag{4}$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^n$. A natural measure of how well $\hat{\theta}(Y; \mathcal{K})$ estimates θ is $\ell^2(\theta, \hat{\theta}(Y; \mathcal{K}))$ where

$$\ell^2(\alpha, \beta) := \frac{1}{n}\|\alpha - \beta\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\alpha_i - \beta_i)^2 \tag{5}$$

with $\alpha = (\alpha_1, \ldots, \alpha_n)$ and $\beta = (\beta_1, \ldots, \beta_n)$. As $\ell^2(\theta, \hat{\theta}(Y; \mathcal{K}))$ is random we study its expectation $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}))$ which is referred to as the risk of the estimator $\hat{\theta}(Y; \mathcal{K})$.

This paper has two aims: (1) For every cone $\mathcal{K}^n_{r,s}$, we prove upper bounds for the risk $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}^n_{r,s}))$ as θ varies in $\mathcal{K}^n_{r,s}$; (2) we isolate the risk bound for the special case of isotonic regression (when $\mathcal{K}^n_{r,s} = \mathcal{M}$) and study its properties in more detail.

1.1. *Upper bounds on the risk of* $\hat{\theta}(Y; \mathcal{K}_{r,s}^n)$. The first part of the paper will be about bounds for the risk $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}_{r,s}^n))$. Our bounds will involve the *statistical dimension* of the cone $\mathcal{K}_{r,s}^n$. For a cone $\mathcal{K}$, defined as in (2), its statistical dimension is given by

$$\delta(\mathcal{K}) := \mathbb{E} D(Z; \mathcal{K}) \qquad \text{where } D(y; \mathcal{K}) := \sum_{i=1}^{n} \frac{\partial}{\partial y_i} \hat{\theta}_i(y; \mathcal{K}) \tag{6}$$

and $Z = (Z_1, \dots, Z_n)$ is a vector whose components are independent standard normal random variables. Note that the quantity $D(y; \mathcal{K})$ is well defined because $\hat{\theta}(y; \mathcal{K})$ is a 1-Lipschitz function of y; see [21]. It was argued in [21] that $D(Y; \mathcal{K})$ provides a measure of the effective dimension of the model. For example, if $\mathcal{K}$ is a linear space of dimension d, then $\hat{\theta}(y; \mathcal{K}) = QY$, where Q is the projection matrix onto $\mathcal{K}$, and $D(y; \mathcal{K}) = \text{trace}(Q) = d$ for all y. It was also shown in [21] that $D(Y; \mathcal{K})$ is the number of distinct values among $\hat{\theta}_1, \dots, \hat{\theta}_n$ for isotonic regression. The term statistical dimension for $\delta(\mathcal{K})$ was first used in [1]; however, the definition of $\delta(\mathcal{K})$ in [1] is different from (6). For connections between the two definitions and more discussion on the notion of statistical dimension, see Section 2.

We are now ready to describe our main result which bounds $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}_{r,s}^n))$ for $\theta \in \mathcal{K}_{r,s}^n$. For each $\theta \in \mathcal{K}_{r,s}^n$, let $k(\theta)$ denote the number of inequalities among $\sum_{j=-r}^{s} w_j \theta_{t+j} \geq 0$ for $1 + r \leq t \leq n - s$ that are strict. In Theorem 2.1, we prove that for every $\theta \in \mathcal{K}_{r,s}^n$,

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}\big(Y; \mathcal{K}_{r,s}^n\big)\big) \leq 6 \inf_{\alpha \in \mathcal{K}_{r,s}^n} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2(1 + k(\alpha))}{n} \delta\big(\mathcal{K}_{r,s}^n\big) \right) \tag{7}$$

under the assumption that $\varepsilon_1, \dots, \varepsilon_n$ are independent normally distributed random variables with mean zero and variance σ^2. This bound behaves differently depending on the form of the true sequence θ and thus describes the adaptive behavior of the LSE; for more details on the inequality, see Section 2. The proof of Theorem 2.1 uses the characterization properties of the projection operator on a closed convex cone. We prove a series of auxiliary results leading to the proof of Theorem 2.1; these results hold for any polyhedral cone $\mathcal{K}$ (not necessarily of the form $\mathcal{K}_{r,s}^n$) and are of independent interest.

1.2. *On risk bounds in isotonic regression.* The second part of the paper is exclusively on isotonic regression. Even in this special case, inequality (7) appears to be new. We provide a reformulation of (7) that bounds the risk of $\hat{\theta}(Y; \mathcal{M})$ using the variation of θ across subsets of $\{1, \dots, n\}$. This results in an inequality that is more interpretable and makes comparison with previous inequalities in isotonic regression more transparent.

To state this bound, we need some notation. Specializing the notation $k(\theta)$ for $\theta \in \mathcal{K}_{r,s}^n$ to the cone $\mathcal{M}$, we get $k(\theta)$ equals the number of inequalities $\theta_i \leq \theta_{i+1}$ for

$i = 1, \dots, n-1$ that are strict. By an abuse of notation, we extend this notation to *interval partitions* of n. An interval partition π of n is a finite sequence of positive integers that sum to n. In combinatorics this is called a composition of n. Let the set of all interval partitions π of n be denoted by Π. Formally, Π can be written as

$$\Pi := \left\{(n_1, n_2, \dots, n_{k+1}) : k \geq 0, n_i \in \mathbb{N} \text{ and } \sum_{i=1}^{k+1} n_i = n\right\}.$$

For each $\pi = (n_1, \dots, n_{k+1}) \in \Pi$, let $k(\pi) := k$.

For every $\theta = (\theta_1, \dots, \theta_n) \in \mathcal{M}$, there exist integers $k \geq 0$ and $n_1, \dots, n_{k+1} \geq 1$ with $n_1 + \cdots + n_{k+1} = n$ such that θ is constant on each set $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \dots, k+1$, where $s_0 := 0$ and $s_i = n_1 + \cdots + n_i$. We refer to this interval partition $\pi_\theta := (n_1, \dots, n_{k+1})$ as the interval partition *generated* by θ. Note that $k(\pi_\theta)$ precisely equals $k(\theta)$, the number of inequalities $\theta_i \leq \theta_{i+1}$, for $i = 1, \dots, n-1$, that are strict.

For every $\theta \in \mathcal{M}$ and $\pi := (n_1, \dots, n_{k+1}) \in \Pi$, we define

$$V_\pi(\theta) = \max_{1 \leq i \leq k+1} (\theta_{s_i} - \theta_{s_{i-1}+1}),$$

where $s_0 := 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \leq i \leq k+1$. $V_\pi(\theta)$ can be treated as measure of variation of θ with respect to the partition π. An important property is that $V_{\pi_\theta}(\theta) = 0$ for every $\theta \in \mathcal{M}$. For the trivial partition $\pi = (n)$, it is easy to see that $k(\pi) = 0$ and $V_\pi(\theta) = V(\theta) = \theta_n - \theta_1$.

With this notation, our main result for isotonic regression states that

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}(Y; \mathcal{M})\big) \leq R(n; \theta), \tag{8}$$

where

$$R(n; \theta) = 4 \inf_{\pi \in \Pi} \left(V_\pi^2(\theta) + \frac{4\sigma^2(1 + k(\pi))}{n} \log \frac{en}{1 + k(\pi)} \right).$$

This inequality is very similar to (7); see Remark 3.1 for the connections. The LSE, $\hat{\theta}(Y; \mathcal{M})$, in isotonic regression has the explicit formula (23). This formula is commonly known as the min–max formula; see [24], Chapter 1. Using this formula, we prove inequality (8) in Section 3. We only use the fact that $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. with mean zero and variance σ^2 [normality of $\varepsilon_1, \dots, \varepsilon_n$ is not needed here unlike inequality (7) for which we require normality].

Inequality (8) appears to be new, even though there is a huge literature on univariate isotonic regression. To place this inequality in a proper historical context, we give a brief overview of existing theoretical results on isotonic regression in Section 3. The strongest previous bound on $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{M}))$ is due to Zhang [35], Theorem 2.2, who showed that

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}(Y; \mathcal{M})\big) \lesssim R_Z(n; \theta), \tag{9}$$

where

$$R_Z(n;\theta) := \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} + \frac{\sigma^2 \log n}{n}$$

with

$$V(\theta) := \theta_n - \theta_1.$$

Here, by the symbol $\lesssim$ we mean $\leq$ up to a multiplicative constant. The quantity $V(\theta)$ is known as the variation of the sequence θ.

Our inequality (8) compares favorably with (9) in certain cases. To see this, suppose, for example, that $\theta_j = I\{j > n/2\}$ (here I denotes the indicator function) so that $V(\theta) = 1$. Then $R_Z(n;\theta)$ is essentially $(\sigma^2/n)^{2/3}$ while $R(n;\theta)$ is much smaller because it is at most $(32\sigma^2/n)\log(en/2)$ as can be seen by taking $\pi = \pi_\theta$ in the definition of $R(n;\theta)$ [note that $k(\theta) = 1$].

More generally by taking $\pi = \pi_\theta$ in the infimum of the definition of $R(n;\theta)$, we obtain

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}(Y;\mathcal{M})\big) \leq \frac{16(1+k(\theta))\sigma^2}{n} \log \frac{en}{1+k(\theta)}, \tag{10}$$

which is a stronger bound than (9) when $k(\theta)$ is small. The reader may observe that $k(\theta)$ is small precisely when the differences $\theta_i - \theta_{i-1}$ are sparse. Inequality (8) can be stronger than (9) even in situations when $k(\theta)$ is not small; see Remark 3.3 for an example.

We study properties of $R(n;\theta)$ in Section 4. In Theorem 4.1, we show that $R(n;\theta)$ is bounded from above by a multiple of $R_Z(n;\theta)$ that is at most logarithmic in n. This implies that our inequality (8) is always only slightly worse off than (9) while being much better in the case of certain sequences θ. We also show in Section 4 that the risk bound $R(n;\theta)$ behaves differently, depending on the form of the true sequence θ. This means that bound (8) demonstrates adaptive behavior of the LSE. One gets a whole range of rates from $(\log n)/n$ (when θ is constant) to $n^{-2/3}$ up to logarithmic factors in the worst case [this worst case rate corresponds to the situation where $\min_i(\theta_i - \theta_{i-1}) \gtrsim 1/n$]. Bound (8) therefore presents a bridge between the two terms in the formula for $R_Z(n;\theta)$.

In addition to being an upper bound for the risk of the LSE, we believe that the quantity $R(n;\theta)$ also acts as a benchmark for the risk of any estimator in isotonic regression. By this, we mean that, in a certain sense, no estimator can have risk that is significantly better than $R(n;\theta)$. We substantiate this claim in Section 5 by proving lower bounds for the *local minimax risk* near the "true" θ. For $\theta \in \mathcal{M}$, the quantity

$$\mathfrak{R}_n(\theta) := \inf_{\hat{t}} \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t, \hat{t})$$

with

$$\mathfrak{N}(\theta) := \big\{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \lesssim R(n;\theta)\big\}$$

will be called the local minimax risk at θ; see Section 5 for the rigorous definition of the neighborhood $\mathfrak{N}(\theta)$ where the multiplicative constants hidden by the $\lesssim$ sign are explicitly given. In the above display ℓ_∞ is defined as $\ell_\infty(t,\theta) := \max_i |t_i - \theta_i|$. The infimum here is over all possible estimators $\hat{t}$. $\mathfrak{R}_n(\theta)$ represents the smallest possible (supremum) risk under the knowledge that the true sequence t lies in the neighborhood $\mathfrak{N}(\theta)$. It provides a measure of the difficulty of estimation of θ. Note that the size of the neighborhood $\mathfrak{N}(\theta)$ changes with θ (and with n) and also reflects the difficulty level of the problem.

Under each of the two following setups for θ, and the assumption of normality of the errors, we show that $\mathfrak{R}_n(\theta)$ is bounded from below by $R(n;\theta)$ up to multiplicative logarithmic factors of n. Specifically:

(1) when the increments of θ (defined as $\theta_i - \theta_{i-1}$, for $i = 2,\dots,n$) grow like $1/n$, we prove in Theorem 5.3 that

$$\mathfrak{R}_n(\theta) \gtrsim \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \gtrsim \frac{R(n;\theta)}{\log(4n)}; \tag{11}$$

(2) when $k(\theta) = k$ and the k values of θ are sufficiently well-separated, we show in Theorem 5.4 that

$$\mathfrak{R}_n(\theta) \gtrsim R(n;\theta)\left(\log \frac{en}{k}\right)^{-2/3}. \tag{12}$$

Because $R(n,\theta)$ is an upper bound for the risk of the LSE and also is a local minimax lower bound in the above sense, our results imply that the LSE is near-optimal in a local nonasymptotic minimax sense. Such local minimax bounds are in the spirit of Cator [11] and Cai and Low [9], who worked with the problems of estimating monotone and convex functions respectively at a point. The difference between these works and our own is that we focus on the global estimation problem. In other words, [11] and [9] prove local minimax bounds for the local (pointwise) estimation problem while we prove local minimax bounds for the global estimation problem.

We also study the performance of the LSE in isotonic regression under model misspecification when the true sequence θ is not necessarily nondecreasing. Here we prove in Theorem 6.1 that $\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}(Y;\mathcal{M})) \le R(n;\tilde{\theta})$ where $\tilde{\theta}$ denotes the nondecreasing projection of θ; see Section 6 for its definition. This should be contrasted with the risk bound of Zhang [35] who proved that $\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}(Y;\mathcal{M})) \lesssim R_Z(n;\tilde{\theta})$. As before our risk bound is at most, slightly worse (by a multiplicative logarithmic factor in n) than R_Z, but is much better when $k(\tilde{\theta})$ is small. We describe two situations where $k(\tilde{\theta})$ is small: when θ itself has few constant blocks [see (56) and Lemma 6.4] and when θ is nonincreasing [in which case $k(\tilde{\theta}) = 1$; see Lemma 6.3].

1.3. *Organization of the paper.* The paper is organized as follows: In Section 2 we state and prove our main upper bound for the risk $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}^n_{r,s}))$. In Section 3 we give a direct proof of the risk bound (8) for isotonic regression without assuming normality of the errors. We investigate the behavior of $R(n; \theta)$, the right-hand side of (8), for different values of the true sequence θ and compare it with $R_Z(n; \theta)$, the right-hand of (9), in Section 4. Local minimax lower bounds for isotonic regression are proved in Section 5. We study the performance of the isotonic LSE under model misspecification in Section 6. The supplementary material [13] gives the proofs of some of the results in the paper.

2. A general risk bound for the projection on closed convex polyhedral cones. The goal of this section is to prove inequality (7). Let us first review the well-known characterization of the LSE under the constraint $\theta \in \mathcal{K}$ for an arbitrary convex polyhedral cone $\mathcal{K}$. This LSE is denoted by $\hat{\theta}(Y; \mathcal{K})$ and is defined in (4). The function $y \mapsto \hat{\theta}(y; \mathcal{K})$ is well defined [because for each y and $\mathcal{K}$, the quantity $\hat{\theta}(y; \mathcal{K})$ exists uniquely by the Hilbert projection theorem], nonlinear in y (in general) and can be characterized by (see, e.g., [4], Proposition 2.2.1)

$$\hat{\theta}(y; \mathcal{K}) \in \mathcal{K}, \qquad \langle y - \hat{\theta}(y; \mathcal{K}), \hat{\theta}(y; \mathcal{K}) \rangle = 0 \quad \text{and} \quad \langle y - \hat{\theta}(y; \mathcal{K}), \omega \rangle \le 0 \tag{13}$$

for all $\omega \in \mathcal{K}$.

Inequality (7) involves the notion of statistical dimension [defined in (6)]. The statistical dimension is an important summary parameter for cones, and it has been used in shape-constrained regression [21] and compressed sensing [1, 22]. It is closely related to the Gaussian width of $\mathcal{K}$, which is an important quantity in geometric functional analysis (see, e.g., [32], Chapter 4) and which has also been used to prove recovery bounds in compressed sensing [1, 12, 22, 25, 27]. See [1], Section 10.3, for the precise connection between the statistical dimension and the Gaussian width.

An alternative definition of the statistical dimension $\delta(\mathcal{K})$ of an arbitrary convex polyhedral cone is given by

$$\delta(\mathcal{K}) = \mathbb{E}\|\hat{\theta}(Z; \mathcal{K})\|^2, \tag{14}$$

where $Z = (Z_1, \dots, Z_n)$ is a vector whose components are independent standard normal random variables. The equivalence of (6) and (14) was observed by Meyer and Woodroofe [21], proof of Proposition 2. It is actually an easy consequence of Stein's lemma because the second identity in (13) implies $\mathbb{E}\|\hat{\theta}(Z; \mathcal{K})\|^2 = \mathbb{E}\langle Z, \hat{\theta}(Z; \mathcal{K})\rangle$, and therefore, Stein's lemma on the right-hand side gives the equivalence of (6) and (14).

We are now ready to prove our main result, Theorem 2.1, which gives inequality (7). This theorem applies to any cone of the form (3). For the proof of Theorem 2.1, we state certain auxiliary results (Lemmas 2.4, 2.5 and 2.6), whose proofs can be found in the supplementary material [13]. These supplementary results hold for any polyhedral cone (2).

THEOREM 2.1. *Fix $n \geq 1$, $r \geq 0$ and $s \geq 1$. Consider the problem of estimating $\theta \in \mathcal{K}_{r,s}^n$ from (1) for independent $N(0, \sigma^2)$ errors $\varepsilon_1, \dots, \varepsilon_n$. Then inequality (7) holds for every $\theta \in \mathcal{K}_{r,s}^n$ with $k(\theta)$ denoting the number of inequalities among $\sum_{j=-r}^{s} w_j \theta_{t+j} \geq 0$, for $1 + r \leq t \leq n - s$, that are strict.*

Before we prove Theorem 2.1 the following remarks are in order.

REMARK 2.1 (Stronger version). From the proof of Theorem 2.1, it will be clear that the risk of the LSE satisfies a stronger inequality than (7). For $\alpha \in \mathcal{K}_{r,s}^n$ with $k(\alpha) = k$, let $1 + r \leq t_1 < \cdots < t_k \leq n - s$ denote the values of t for which the inequalities $\sum_{j=-r}^{s} w_j \alpha_{t+j} \geq 0$ are strict. Let

$$\tau(\alpha) := \delta\big(\mathcal{K}_{r,s}^{t_1 - 1 + s}\big) + \delta\big(\mathcal{K}_{r,s}^{t_2 - t_1}\big) + \cdots + \delta\big(\mathcal{K}_{r,s}^{t_k - t_{k-1}}\big) + \delta\big(\mathcal{K}_{r,s}^{n - t_k - s + 1}\big). \tag{15}$$

The proof of Theorem 2.1 will imply that

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}\big(Y; \mathcal{K}_{r,s}^n\big)\big) \leq 6 \inf_{\alpha \in \mathcal{K}_{r,s}^n} \bigg(\ell^2(\theta, \alpha) + \frac{\sigma^2}{n} \tau(\alpha)\bigg). \tag{16}$$

The observation that $\delta(\mathcal{K}_{r,s}^n)$ is increasing in n (note that the weights w_j, $-r \leq j \leq s$, do not depend on n) implies that $\tau(\alpha) \leq (1 + k(\alpha))\delta(\mathcal{K}_{r,s}^n)$ for all $\alpha \in \mathcal{K}_{r,s}^n$, and hence inequality (16) is stronger than (7).

REMARK 2.2 (Connection to the facial structure of $\mathcal{K}_{r,s}^n$). Every convex polyhedral cone (2) has a well-defined facial structure. Indeed, a standard result (see, e.g., [26], Section 8.3) states that a subset F of a convex polyhedral cone $\mathcal{K}$, as defined in (2), is a face if and only if F is nonempty and $F = \{\theta \in \mathcal{K} : \tilde{A}\theta = 0\}$ for some $\tilde{m} \times n$ matrix $\tilde{A}$ whose rows are a subset of the rows of A. The dimension of F equals $n - \rho(\tilde{A})$ where $\rho(\tilde{A})$ denotes the rank of $\tilde{A}$. It is then clear that if $\theta \in \mathcal{K}_{r,s}^n$ is in a low-dimensional face of $\mathcal{K}_{r,s}^n$, then $k(\theta)$ must be small. Now if $\delta(\mathcal{K}_{r,s}^n)$ is at most logarithmic in n (which is indeed the case for the case of isotonic and convex regression; see Examples 2.2 and 2.3), then bound (7) implies that the risk of the LSE is bounded from above by the parametric rate σ^2/n (up to multiplicative logarithmic factors in n) provided θ is in a low-dimensional face of $\mathcal{K}_{r,s}^n$. Therefore, the LSE automatically adapts to vectors in low-dimensional faces of $\mathcal{K}_{r,s}^n$. For general θ, the risk is bounded from above by a combination of how close θ is to a k-dimensional face of $\mathcal{K}_{r,s}^n$ and $\sigma^2 \delta(\mathcal{K}_{r,s}^n)(1 + k)/n$ as k varies.

EXAMPLE 2.2 (Isotonic regression). Isotonic regression corresponds to $r = 0$, $s = 1$, $w_0 = -1$ and $w_1 = 1$ so that $\mathcal{K}_{r,s}^n$ becomes $\mathcal{M}$. It turns out that the statistical dimension of this cone satisfies

$$\delta(\mathcal{M}) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} \qquad \text{for every } n \geq 1, \tag{17}$$

which immediately implies that $\delta(\mathcal{M}) \le \log(en)$. This can be proved using symmetry arguments formalized in the theory of finite reflection groups; see [1], Appendix C.4, where the proof of (17) is sketched.

Now let $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathcal{M}$ with $k(\alpha) = k$. Then there exist integers $n_1, \dots, n_{k+1} \ge 1$ with $n_1 + \cdots + n_{k+1} = n$ such that α is constant on each set $\{j : s_{i-1} + 1 \le j \le s_i\}$, for $i = 1, \dots, k+1$, where $s_0 := 0$ and $s_i = n_1 + \cdots + n_i$. It is easy to check then that the quantity $\tau(\alpha)$ defined in (15) equals

$$\tau(\alpha) = \delta(\mathcal{M}^{n_1}) + \cdots + \delta(\mathcal{M}^{n_{k+1}}),$$

where $\mathcal{M}^i := \{(\theta_1, \dots, \theta_i) \in \mathbb{R}^i : \theta_1 \le \cdots \le \theta_i\}$ is the monotone cone in $\mathbb{R}^i$. Inequality (17) then gives

$$\tau(\alpha) \le \sum_{i=1}^{k+1} \log(en_i) \le (k+1)\log\left(\frac{en}{k+1}\right)$$

because of the concavity of $x \mapsto \log x$. Inequality (16) therefore gives

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{M})) \le 6 \inf_{\alpha \in \mathcal{M}} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2(k(\alpha)+1)}{n} \log \frac{en}{k(\alpha)+1}\right). \tag{18}$$

This inequality is closely connected to (8), as we describe in detail in Remark 3.1. Note that we require normality of $\varepsilon_1, \dots, \varepsilon_n$. In Section 3, we prove an inequality which gives a variant of inequality (18) with different multiplicative constants but without the assumption of normality.

Inequality (18) can be restated in the following way. For each $0 \le k \le n-1$, let $\mathcal{P}_k$ denote the set of all sequences $\alpha \in \mathcal{M}$ with $k(\alpha) \le k$. With this notation, inequality (18) can be rewritten as

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{M})) \le 6 \min_{0 \le k \le n-1} \left[\inf_{\alpha \in \mathcal{P}_k} \ell^2(\theta, \alpha) + \frac{\sigma^2(k+1)}{n} \log \frac{en}{k+1}\right]. \tag{19}$$

Bound (19) reflects adaptation of the LSE with respect to the classes $\mathcal{P}_k$. Such risk bounds are usually provable for estimators based on empirical model selection criteria (see, e.g., [3]) or aggregation; see, for example, [23]. Specializing to the present situation, in order to adapt over $\mathcal{P}_k$ as k varies, one constructs LSEs over each $\mathcal{P}_k$ and then either selects one estimator from this collection by an empirical model selection criterion or aggregates these estimators with data-dependent weights. In this particular situation, such estimators are very difficult to compute as minimizing the LS criterion over $\mathcal{P}_k$ is a nonconvex optimization problem. In contrast, the LSE can be easily computed by a convex optimization problem. It is remarkable that the LSE, which is constructed with no explicit model selection criterion in mind, achieves adaptive risk bound (18). In the next example, we illustrate this adaptation for the LSE in convex regression.

EXAMPLE 2.3 (Convex regression). Convex regression with equispaced design points corresponds to $\mathcal{K}_{r,s}^n$ with $r = s = 1$, $w_{-1} = w_1 = 1$ and $w_0 = -2$. It turns out that the statistical dimension of this cone satisfies

$$\delta(\mathcal{K}_{1,1}^n) \le C\big(\log(en)\big)^{5/4} \qquad \text{for all } n \ge 1, \tag{20}$$

where C is a universal positive constant. This is proved in [19], Theorem 3.1, via metric entropy results for classes of convex functions.

Let $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathcal{K}_{-1,1}^n$ with $k(\alpha) = k$. Let $t_1, \dots, t_k$ denote the values of t where the inequality $2\theta_t \le \theta_{t-1} + \theta_{t+1}$ is strict. With $n_1 := t_1$, $n_i := t_i - t_{i-1}$ for $i = 2, \dots, k$ and $n_{k+1} = n - t_k$, we have, from (15) and (20),

$$\tau(\alpha) = \sum_{i=1}^{k+1} \delta(\mathcal{K}_{1,1}^{n_i}) \le C \sum_{i=1}^{k+1} (\log en_i)^{5/4}.$$

Using the fact that $x \mapsto (\log x)^{5/4}$ is concave for $x \ge e$, we have (note that $\sum_i n_i = n$)

$$\tau(\alpha) \le (k+1)\left(\log \frac{en}{k+1}\right)^{5/4}.$$

Inequality (16) then becomes

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}_{1,1}^n)) \le C \inf_{\alpha \in \mathcal{K}_{1,1}^n} \left(\ell^2(\theta, \alpha) + \frac{\sigma^2(k(\alpha)+1)}{n}\left(\log \frac{en}{k(\alpha)+1}\right)^{5/4}\right).$$

Note that the quantity $1 + k(\alpha)$ can be interpreted as the number of affine pieces of the convex sequence α. This risk bound is the analogue of inequality (18) for convex regression, and it highlights the adaptation of the convex LSE to piecewise affine convex functions. A weaker version of this inequality appeared in [19], Theorem 2.3.

2.1. *Proof of Theorem* 2.1. We now prove Theorem 2.1. We shall first state some general results (Lemmas 2.4, 2.5 and 2.6), whose proofs can be found in the supplementary material [13], for the risk of $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K}))$, which hold for every $\mathcal{K}$ of the form (2). Theorem 2.1 will then be proved by specializing these results for $\mathcal{K} = \mathcal{K}_{r,s}^n$.

We begin by recalling a result of Meyer and Woodroofe [21] who related the risk of $\hat{\theta}(Y; \mathcal{K})$ to the function $D(\cdot; \mathcal{K})$. Specifically, [21], Proposition 2, proved that

$$\mathbb{E}_0 \ell^2(0, \hat{\theta}(Y; \mathcal{K})) = \frac{\sigma^2 \delta(\mathcal{K})}{n} = \frac{\sigma^2}{n} \mathbb{E}_0 D(Y; \mathcal{K})$$

and

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{K})) \le \frac{\sigma^2}{n} \mathbb{E}_\theta D(Y; \mathcal{K}) \qquad \text{for every } \theta \in \mathcal{K}. \tag{21}$$

These can be proved via Stein's lemma; see [21], proof of Proposition 2. It might be helpful to observe here that the function $D(y;\mathcal{K})$ satisfies $D(ty;\mathcal{K}) = D(y;\mathcal{K})$ for every $t \in \mathbb{R}$, and this is a consequence of the fact that $\hat{\theta}(ty;\mathcal{K}) = t\hat{\theta}(y;\mathcal{K})$ and the characterization (13).

Our first lemma below states that the risk of the LSE is equal to $\sigma^2\delta(\mathcal{K})/n$ for all θ belonging to the lineality space $\mathcal{L} := \{\theta \in \mathbb{R}^n : A\theta = 0\}$ of $\mathcal{K}$. The lineality space $\mathcal{L}$ will be crucial in the proof of Theorem 2.1. The lineality space of the cone for isotonic regression is the set of all constant sequences. The lineality space of the cone for convex regression is the set of all affine sequences. Also, we say that two convex polyhedral cones $\mathcal{K}_1$ and $\mathcal{K}_2$ are orthogonal if $\langle \gamma_1, \gamma_2 \rangle = 0$ for all $\gamma_1 \in \mathcal{K}_1$ and $\gamma_2 \in \mathcal{K}_2$.

LEMMA 2.4. *For every $\theta \in \mathbb{R}^n$ with $\theta = \gamma_1 + \gamma_2$ for some $\gamma_1 \in \mathcal{L}$ and $\gamma_2 \perp \mathcal{K}$ (i.e., $\langle \gamma_2, \omega \rangle = 0$ for all $\omega \in \mathcal{K}$), we have $\mathbb{E}_\theta D(Y;\mathcal{K}) = \delta(\mathcal{K})$.*

LEMMA 2.5. *Let $\mathcal{K}$ be an arbitrary convex polyhedral cone. Suppose $\mathcal{K}_1, \dots, \mathcal{K}_l$ are orthogonal polyhedral cones with lineality spaces $\mathcal{L}_1, \dots, \mathcal{L}_l$ such that $\mathcal{K} \subseteq \mathcal{K}_1 + \cdots + \mathcal{K}_l$. Then*

$$\mathbb{E}_\theta D(Y;\mathcal{K}) \le 2\big(\delta(\mathcal{K}_1) + \cdots + \delta(\mathcal{K}_l)\big) \qquad \textit{for every } \theta \in \mathcal{K} \cap (\mathcal{L}_1 + \cdots + \mathcal{L}_l).$$

The next lemma allows us to bound the risk of the LSE at θ by a combination of the risk at α and the distance between θ and α.

LEMMA 2.6. *The risk of the LSE satisfies the following inequality*:

$$\mathbb{E}_\theta \ell^2\big(\theta, \hat{\theta}(Y;\mathcal{K})\big) \le 3 \inf_{\alpha \in \mathcal{K}} \big[2\ell^2(\theta,\alpha) + \mathbb{E}_\alpha \ell^2\big(\alpha, \hat{\theta}(Y;\mathcal{K})\big)\big] \qquad \textit{for every } \theta \in \mathcal{K}.$$

We are now ready to prove Theorem 2.1.

PROOF OF THEOREM 2.1. By Lemma 2.6, it is enough to prove that

$$\mathbb{E}_\alpha \ell^2\big(\alpha, \hat{\theta}(Y;\mathcal{K}_{r,s}^n)\big) \le 2\big(1 + k(\alpha)\big)\frac{\sigma^2\delta(\mathcal{K}_{r,s}^n)}{n} \qquad \text{for every } \alpha \in \mathcal{K}_{r,s}^n.$$

Fix $\alpha \in \mathcal{K}_{r,s}^n$, and let $k = k(\alpha)$, which means that k of the inequalities $\sum_{j=-r}^{s} w_j \alpha_{t+j} \ge 0$ for $1+r \le t \le n-s$ are strict. Let $1+r \le t_1 < \cdots < t_k \le n-s$ denote the indices of the inequalities that are strict. We partition the set $\{1, \dots, n\}$ into $k+1$ disjoint sets $E_0, \dots, E_k$ where

$$E_0 := \{1, \dots, t_1 - 1 + s\}, \qquad E_k := \{t_k + s, \dots, n\}$$

and

$$E_i := \{t_i + s, \dots, t_{i+1} - 1 + s\} \qquad \text{for } 1 \le i \le k-1.$$

Also for each $0 \le i \le k$, let

$$F_i := \{t \in \mathbb{Z} : t - r \in E_i \text{ and } t + s \in E_i\}.$$

We now apply Lemma 2.5 with

$$\mathcal{K}_i := \left\{\theta \in \mathbb{R}^n : \theta_j = 0 \text{ for } j \notin E_i \text{ and } \sum_{j=-r}^{s} w_j \theta_{t+j} \ge 0 \text{ for } t \in F_i\right\}$$

for $i = 0, \dots, k$. The lineality space of $\mathcal{K}_i$ is, by definition,

$$\mathcal{L}_i = \left\{\theta \in \mathbb{R}^n : \theta_j = 0 \text{ for } j \notin E_i \text{ and } \sum_{j=-r}^{s} w_j \theta_{t+j} = 0 \text{ for } t \in F_i\right\}.$$

$\mathcal{K}_0, \dots, \mathcal{K}_k$ are orthogonal convex polyhedral cones because $E_0, \dots, E_k$ are disjoint. Also $\mathcal{K} \subseteq \mathcal{K}_0 + \cdots + \mathcal{K}_k$ because every $\theta \in \mathcal{K}$ can be written as $\theta = \sum_{i=0}^{k} \theta^{(i)}$ where $\theta_j^{(i)} := \theta_j I\{j \in E_i\}$ (it is easy to check that $\theta^{(i)} \in \mathcal{K}_i$ for each i). Further, note that $\alpha \in \mathcal{L}_0 + \cdots + \mathcal{L}_k$ since $\alpha^{(i)} \in \mathcal{L}_i$ for every i. Lemma 2.5 thus gives $\mathbb{E}_\alpha D(Y; \mathcal{K}) \le 2 \sum_{i=0}^{k} \delta(\mathcal{K}_i)$. Inequality (21) then implies that

$$\mathbb{E}_\alpha \ell^2\big(\alpha, \hat{\theta}(Y; \mathcal{K})\big) \le \frac{2\sigma^2}{n} \sum_{i=0}^{k} \delta(\mathcal{K}_i).$$

It is now easy to check that $\delta(\mathcal{K}_i) = \delta(\mathcal{K}_{r,s}^{|E_i|})$ for each i which proves (16). The proof of (7) is now complete by the observation $\delta(\mathcal{K}_{r,s}^{|E_i|}) \le \delta(\mathcal{K}_{r,s}^{n})$ as $|E_i| \le n$. □

3. Risk bound in isotonic regression. In this section, we provide a proof of inequality (8) using an explicit formula of the LSE in isotonic regression. Our proof does not require normality of $\varepsilon_1, \dots, \varepsilon_n$. We also explain the similarities between inequalities (7) and (8). Before we get to inequality (8), however, we give a brief overview of existing theoretical results in isotonic regression.

Usually isotonic regression is posed as a function estimation problem in which the unknown object of interest is a nondecreasing function f_0, and one observes data from model (1) with $\theta_i := f_0(x_i), i = 1, \dots, n$, where $x_1 < \cdots < x_n$ are fixed design points. The most natural and commonly used estimator for this problem is the monotone LSE defined as any nondecreasing function $\hat{f}_{\text{ls}}$ on $\mathbb{R}$ for which $(\hat{f}_{\text{ls}}(x_1), \dots, \hat{f}_{\text{ls}}(x_n)) = \hat{\theta}(Y; \mathcal{M})$. This estimator was proposed by [7] and [2]; also see [16] for the related problem of estimating a nonincreasing density. Note that $\hat{\theta}(Y; \mathcal{M})$ can be computed easily using the *pool adjacent violators algorithm*; see [24], Chapter 1.

Existing theoretical results on isotonic regression can be grouped into two categories: (1) results on the behavior of the LSE at an interior point (which is sometimes known as local behavior), and (2) results on the behavior of a global loss function measuring how far $\hat{f}_{\text{ls}}$ is from f_0.

Results on the local behavior are proved, among others, in [8, 10, 11, 17, 18, 20, 34]. Under certain regularity conditions on the unknown function f_0 near the interior point x_0, it was proved in [8] that $\hat{f}_{\mathrm{ls}}(x_0)$ converges to $f_0(x_0)$ at the rate $n^{-1/3}$ and also characterized the limiting distribution of $n^{1/3}(\hat{f}_{\mathrm{ls}}(x_0) - f_0(x_0))$. In the related (nonincreasing) density estimation problem, the authors of [10, 18, 20] demonstrated that if the interior point x_0 lies on a flat stretch of the underlying function, then the LSE (which is also the nonparametric maximum likelihood estimator, usually known as the Grenander estimator) converges to a nondegenerate limit at rate $n^{-1/2}$, and they characterized the limiting distribution. In [11], Cator demonstrated that the rate of convergence of $\hat{f}_{\mathrm{ls}}(x_0)$ to $f_0(x_0)$ depends on the local behavior of f_0 near x_0, and explicitly described this rate for each f_0. In this sense, the LSE $\hat{f}_{\mathrm{ls}}$ adapts automatically to the unknown function f_0. In [11], it was also proved that the LSE is optimal for local behavior by establishing a local asymptotic minimax lower bound.

Often in monotone regression, the interest is in the estimation of the entire function f_0, as opposed to just its value at one fixed point. In this sense, it is more appropriate to study the behavior of $\hat{f}_{\mathrm{ls}}$ under a global loss function. The most natural and commonly studied global loss function in this setup is

$$L(f, g) := \frac{1}{n}\sum_{i=1}^{n}\big(f(x_i) - g(x_i)\big)^2 = \ell^2(\check{f}, \check{g}),$$

where $\check{f} := (f(x_1), \dots, f(x_n))$ and $\check{g} := (g(x_1), \dots, g(x_n))$. Note that under this loss function, the function estimation problem becomes exactly the same as the sequence estimation problem described in (1), where the goal is to estimate the vector $\theta := (\theta_1, \dots, \theta_n)$ under the constraint $\theta \in \mathcal{M}$ and the loss function (5). The behavior of $\hat{\theta}(Y; \mathcal{M})$, under the loss ℓ^2, has been studied in a number of papers including [6, 14, 21, 29, 30, 33, 35]. If one looks at the related (nonincreasing) density estimation problem, Birgé [5] developed nonasymptotic risk bounds for the Grenander estimator, measured with the L_1-loss, whereas Van de Geer [30] has results on the Hellinger distance. As mentioned in the Introduction, the strongest existing bound on $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}(Y; \mathcal{M}))$ is due to [35], Theorem 2.2. We recalled this inequality in (9) and compared it to our bound (8) in some situations.

In the following theorem, we prove inequality (8) using explicit characterization of $\hat{\theta}(Y; \mathcal{M})$ without requiring normality of $\varepsilon_1, \dots, \varepsilon_n$. In fact, we prove an inequality that is slightly stronger than (8).

We need the following notation. For simplicity, we use $\hat{\theta}$ for $\hat{\theta}(Y; \mathcal{M})$ and $\hat{\theta}_j$ for the components of $\hat{\theta}(Y; \mathcal{M})$. For any sequence $(a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ and any $1 \le u \le v \le n$, let

$$\bar{a}_{u,v} := \frac{1}{v - u + 1}\sum_{j=u}^{v} a_j. \tag{22}$$

We will use this notation mainly when a equals Y, θ or ε. Our proof uses ideas similar to those in [35], Section 2 (see Remark 3.2 for details about the connections with [35], Section 2) and is based on the following explicit representation of the LSE $\hat{\theta}$ (see [24], Chapter 1):

$$\hat{\theta}_j = \min_{v \geq j} \max_{u \leq j} \bar{Y}_{u,v}. \tag{23}$$

For $x \in \mathbb{R}$, we write $x_+ := \max\{0, x\}$ and $x_- := -\min\{0, x\}$. For $\theta \in \mathcal{M}$ and $\pi = (n_1, \dots, n_{k+1}) \in \Pi$, let

$$D_\pi(\theta) = \left(\frac{1}{n} \sum_{i=1}^{k+1} \sum_{j=s_{i-1}+1}^{s_i} (\theta_j - \bar{\theta}_{s_{i-1}+1, s_i})^2\right)^{1/2},$$

where $s_0 = 0$ and $s_i = n_1 + \cdots + n_i$, for $1 \leq i \leq k+1$. Like $V_\pi(\theta)$, this quantity $D_\pi(\theta)$ can also be treated as a measure of the variation of θ with respect to π. This measure also satisfies $D_{\pi_\theta}(\theta) = 0$ for every $\theta \in \mathcal{M}$. Moreover,

$$D_\pi(\theta) \leq V_\pi(\theta) \qquad \text{for every } \theta \in \mathcal{M} \text{ and } \pi \in \Pi.$$

When $\pi = (n)$ is the trivial partition, $D_\pi(\theta)$ turns out to be just the standard deviation of θ. In general, $D_\pi^2(\theta)$ is analogous to the within group sum of squares term in ANOVA with the blocks of π being the groups. Below, we prove a stronger version of (8) with $D_\pi(\theta)$ replacing $V_\pi(\theta)$ in the definition of $R(n; \theta)$.

THEOREM 3.1. *Suppose $Y_1, \dots, Y_n$ are observations from model (1) with $\varepsilon_1, \dots, \varepsilon_n$ being i.i.d. with mean zero and variance σ^2. For every $\theta \in \mathcal{M}$, the risk of $\hat{\theta} = \hat{\theta}(Y; \mathcal{M})$ satisfies the following inequality*:

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq 4 \inf_{\pi \in \Pi} \left(D_\pi^2(\theta) + \frac{4\sigma^2(1 + k(\pi))}{n} \log \frac{en}{1 + k(\pi)} \right). \tag{24}$$

PROOF. Fix $1 \leq j \leq n$ and $0 \leq m \leq n - j$. By (23), we have

$$\hat{\theta}_j = \min_{v \geq j} \max_{u \leq j} \bar{Y}_{u,v} \leq \max_{u \leq j} \bar{Y}_{u,j+m} = \max_{u \leq j} (\bar{\theta}_{u,j+m} + \bar{\varepsilon}_{u,j+m}),$$

where, in the last equality, we used $\bar{Y}_{u,v} = \bar{\theta}_{u,v} + \bar{\varepsilon}_{u,v}$. By the monotonicity of θ, we have $\bar{\theta}_{u,j+m} \leq \bar{\theta}_{j,j+m}$ for all $u \leq j$. Therefore, for every $\theta \in \mathcal{M}$, we get

$$\hat{\theta}_j - \theta_j \leq (\bar{\theta}_{j,j+m} - \theta_j) + \max_{u \leq j} \bar{\varepsilon}_{u,j+m}.$$

Taking positive parts, we have

$$(\hat{\theta}_j - \theta_j)_+ \leq (\bar{\theta}_{j,j+m} - \theta_j) + \max_{u \leq j} (\bar{\varepsilon}_{u,j+m})_+.$$

Squaring and taking expectations on both sides, we obtain

$$\mathbb{E}_\theta (\hat{\theta}_j - \theta_j)_+^2 \leq \mathbb{E}_\theta \Big((\bar{\theta}_{j,j+m} - \theta_j) + \max_{u \leq j} (\bar{\varepsilon}_{u,j+m})_+ \Big)^2.$$

Using the elementary inequality $(a+b)^2 \le 2a^2 + 2b^2$ we get

$$\mathbb{E}_\theta(\hat{\theta}_j - \theta_j)_+^2 \le 2(\bar{\theta}_{j,j+m} - \theta_j)^2 + 2\mathbb{E}\max_{u \le j}(\bar{\varepsilon}_{u,j+m})_+^2.$$

We observe now that, for fixed integers j and m, the process $\{\bar{\varepsilon}_{u,j+m}, u = 1, \dots, j\}$ is a martingale with respect to the filtration $\mathcal{F}_1, \dots, \mathcal{F}_j$ where $\mathcal{F}_i$ is the sigma-field generated by the random variables $\varepsilon_1, \dots, \varepsilon_{i-1}$ and $\bar{\varepsilon}_{i,j+m}$. Therefore, by Doob's inequality for submartingales (see, e.g., Theorem 5.4.3 of [15]), we have

$$\mathbb{E}\max_{u \le j}(\bar{\varepsilon}_{u,j+m})_+^2 \le 4\mathbb{E}(\bar{\varepsilon}_{j,j+m})_+^2 \le 4\mathbb{E}(\bar{\varepsilon}_{j,j+m})^2 \le \frac{4\sigma^2}{m+1}.$$

So using the above result we get the following pointwise upper bound for the positive part of the risk:

$$\mathbb{E}_\theta(\hat{\theta}_j - \theta_j)_+^2 \le 2(\bar{\theta}_{j,j+m} - \theta_j)^2 + \frac{8\sigma^2}{m+1}. \tag{25}$$

Note that the above upper bound holds for any arbitrary m, $0 \le m \le n-j$. By a similar argument we can get the following pointwise upper bound for the negative part of risk which now holds for any m, $0 \le m \le j$:

$$\mathbb{E}_\theta(\hat{\theta}_j - \theta_j)_-^2 \le 2(\theta_j - \bar{\theta}_{j-m,j})^2 + \frac{8\sigma^2}{m+1}. \tag{26}$$

Let us now fix $\pi = (n_1, \dots, n_{k+1}) \in \Pi$. Let $s_0 := 0$ and $s_i := n_1 + \cdots + n_i$ for $1 \le i \le k+1$. For each $j = 1, \dots, n$, we define two integers $m_1(j)$ and $m_2(j)$ in the following way: $m_1(j) = s_i - j$ and $m_2(j) = j - 1 - s_{i-1}$ when $s_{i-1} + 1 \le j \le s_i$. We use this choice of $m_1(j)$ in (25) and $m_2(j)$ in (26) to obtain $\mathbb{E}_\theta(\hat{\theta}_j - \theta_j)^2 \le A_j + B_j$ where

$$A_j := 2(\bar{\theta}_{j,j+m_1(j)} - \theta_j)^2 + \frac{8\sigma^2}{m_1(j)+1}$$

and

$$B_j := 2(\theta_j - \bar{\theta}_{j-m_2(j),j})^2 + \frac{8\sigma^2}{m_2(j)+1}.$$

This results in the risk bound

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \le \frac{1}{n}\sum_{j=1}^n A_j + \frac{1}{n}\sum_{j=1}^n B_j.$$

We shall now prove that

$$\frac{1}{n}\sum_{j=1}^n A_j \le 2D_\pi^2(\theta) + \frac{8(k+1)\sigma^2}{n}\log\frac{en}{k+1} \tag{27}$$

and

$$\frac{1}{n}\sum_{j=1}^{n} B_j \le 2D_\pi^2(\theta) + \frac{8(k+1)\sigma^2}{n}\log\frac{en}{k+1}. \tag{28}$$

We give below the proof of (27), and the proof of (28) is nearly identical. Using the form of A_j, we break up $\frac{1}{n}\sum_{j=1}^{n} A_j$ into two terms. For the first term, note that $j + m_1(j) = s_i$, for $s_{i-1}+1 \le j \le s_i$ and therefore

$$\sum_{j=1}^{n}(\bar{\theta}_{j,j+m_1(j)} - \theta_j)^2 = \sum_{i=1}^{k+1}\sum_{j=s_{i-1}+1}^{s_i}(\bar{\theta}_{j,s_i} - \theta_j)^2.$$

By Lemma 11.2 in the supplementary material [13], we get

$$\sum_{j=s_{i-1}+1}^{s_i}(\bar{\theta}_{j,s_i} - \theta_j)^2 \le \sum_{j=s_{i-1}+1}^{s_i}(\bar{\theta}_{s_{i-1}+1,s_i} - \theta_j)^2$$

for every $i = 1, \dots, k+1$. Thus summing over $i = 1, \dots, k+1$, and multiplying by $2/n$ proves that the first term in $\frac{1}{n}\sum_{j=1}^{n} A_j$ is bounded from above by $2D_\pi^2(\theta)$. To bound the second term, we write

$$\sum_{j=1}^{n}\frac{1}{m_1(j)+1} = \sum_{i=1}^{k+1}\sum_{j=s_{i-1}+1}^{s_i}\frac{1}{s_i - j + 1} = \sum_{i=1}^{k+1}\left(1 + \frac{1}{2} + \cdots + \frac{1}{n_i}\right). \tag{29}$$

Since the harmonic series $\sum_{i=1}^{l} 1/l$ is at most $1 + \log l$ for $l \ge 1$, we obtain

$$\sum_{j=1}^{n}\frac{1}{m_1(j)+1} \le k+1+\sum_{i=1}^{k+1}\log n_i \le k+1+(k+1)\log\left(\frac{\sum_i n_i}{k+1}\right),$$

where the last inequality is a consequence of the concavity of the logarithm function. This proves (27) because $\sum_i n_i = n$. Combining (27) and (28) proves the theorem. □

REMARK 3.1. For each $\pi = (n_1, \dots, n_{k+1}) \in \Pi$, let $\mathcal{M}_\pi$ denote the set of all $\alpha \in \mathcal{M}$ such that α is constant on each set $\{j : s_{i-1}+1 \le j \le s_i\}$ for $i = 1, \dots, k+1$. Then it is easy to see that

$$\inf_{\alpha \in \mathcal{M}_\pi} \ell^2(\theta, \alpha) = D_\pi^2(\theta).$$

Using this, it is easy to see that inequality (24) is equivalent to

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \le 4 \inf_{\alpha \in \mathcal{M}}\left(\ell^2(\theta, \alpha) + \frac{4\sigma^2(1+k(\alpha))}{n}\log\frac{en}{1+k(\alpha)}\right). \tag{30}$$

Inequality (24) therefore differs from inequality (18) only by its multiplicative constants. It should be noted that we proved (18) assuming normality of $\varepsilon_1, \dots, \varepsilon_n$

while (24) was proved without using normality. Inequality (8) is slightly weaker than (24) because $D_\pi(\theta) \le V_\pi(\theta)$. We still work with (8) in isotonic regression as opposed to (24) because it is easier to compare (8) to existing inequalities, and also, as we shall show in Section 5, inequality (8) is nearly optimal.

REMARK 3.2. Bounding the infimum in the right-hand side of (24) by taking $\pi = \pi_\theta$ and letting $k(\theta) = k$, we obtain

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \le \frac{16\sigma^2(1+k)}{n} \log \frac{en}{1+k}. \tag{31}$$

This inequality might be implicit in the arguments of [35], Section 2. It might be possible to prove (31) by applying [35], Theorem 2.1, to each of the $k+1$ constant pieces of θ and by bounding the resulting quantities via arguments in [35], proof of Theorem 2.2.

REMARK 3.3. Note that $k(\theta)$ does not have to be small for (8) to be an improvement of (9). One only needs that $V_\pi(\theta)$ be small for some partition π with small $k(\pi)$. Equivalently, from (30), one needs that $\ell^2(\theta, \alpha)$ is small for some $\alpha \in \mathcal{M}$ with small $k(\alpha)$. This is illustrated below.

Let $\{a_j, j \ge 1\}$ be an arbitrary countable subset of $[0, 1]$, and let $\{p_j, j \ge 1\}$ denote any probability sequence, that is, $p_j \ge 0, \sum_j p_j = 1$. Fix $n \ge 1$, and let $\theta_i := \sum_{j : a_j \le i/n} p_j$ for $i = 1, \dots, n$. We will argue below that, for many choices of $\{p_j, j \ge 1\}$, inequality (30) gives a faster rate of convergence than $n^{-2/3}$, even though $k(\theta)$ can be as large as n.

Indeed, fix $1 \le k \le n$, and define $\alpha_i = \sum_{j \le k : a_j \le i/n} p_j$. It is then clear that $k(\alpha) \le k$. Also for each $1 \le i \le n$, we have

$$0 \le \theta_i - \alpha_i = \sum_{j > k : a_j \le i/n} p_j \le \sum_{j > k} p_j.$$

This implies that $\ell(\theta, \alpha) \le \sum_{j : j > k} p_j$. Thus inequality (30) gives

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \le 4 \inf_k \left[\left(\sum_{j : j > k} p_j \right)^2 + \frac{4\sigma^2(1+k)}{n} \log \frac{en}{1+k} \right]. \tag{32}$$

When $\sum_{j : j > k} p_j = o(k^{-1})$, it can be checked that the bound above is faster than $n^{-2/3}$. This happens, for instance, when $p_j \propto j^{-a}$ for $a \ge 3$. In fact, when $p_j = 2^{-j}$, (32) gives the parametric rate up to logarithmic factors.

However, when, say $p_j \propto j^{-a}$ for $1 < a \le 2$, (32) does not give a rate that is faster than $n^{-2/3}$. It might be possible here to use a different approximation vector $\alpha \in \mathcal{M}$ which would still yield a rate of $o(n^{-2/3})$, but we do not have a proof of this. A result from the literature that is relevant here is [35], inequality (2.10). For the vectors θ considered above (and in certain more general situations), this inequality gives an asymptotic bound of $o(n^{-2/3})$ (without quantifying the exact order) for the risk for all choices of $\{a_j\}$ and $\{p_j\}$, even for $p_j \propto j^{-a}$, $1 < a \le 2$.

EXAMPLE 3.2. We prove in Theorem 4.1 in the next section that the bound given by Theorem 3.1 is always smaller than a logarithmic multiplicative factor of the usual cube root rate of convergence for every $\theta \in \mathcal{M}$ with $V(\theta) > 0$. Here, we shall demonstrate this in the special case of the sequence $\theta = (1/n, 2/n, \dots, 1)$ where the bound in (24) can be calculated exactly. Indeed, if $\pi = (n_1, \dots, n_k)$ with $n_i \geq 1$ and $\sum_{i=1}^k n_i = n$, direct calculation gives

$$D_\pi^2(\theta) = \frac{1}{12n^3}\left(\sum_{i=1}^k n_i^3 - n\right).$$

Now Hölder's inequality gives $n = \sum_{i=1}^k n_i \leq (\sum_{i=1}^k n_i^3)^{1/3} k^{2/3}$ which means that $\sum_{i=1}^k n_i^3 \geq n^3/k^2$. Therefore, for every fixed $k \in \{1, \dots, n\}$ such that n/k is an integer, $D_\pi^2(\theta)$ is minimized over all partitions π with $k(\pi) = k$ when $n_1 = n_2 = \cdots = n_k = n/k$. This gives $\inf_{\pi : k(\pi)=k} D^2(\pi) = (k^{-2} - n^{-2})/12$. As a consequence, Theorem 3.1 yields the bound

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq \frac{1}{3} \inf_{k : n/k \in \mathbb{Z}} \left(\frac{1}{k^2} - \frac{1}{n^2} + \frac{48\sigma^2 k}{n} \log(en/k)\right).$$

Now with the choice $k \sim (n/\sigma^2)^{1/3}$, we get the cube root rate for $\hat{\theta}$ up to logarithmic multiplicative factors in n. We generalize this to arbitrary $\theta \in \mathcal{M}$ with $V(\theta) > 0$ in Theorem 4.1.

4. The quantity $R(n; \theta)$. In this section, we state some results about the quantity $R(n; \theta)$ appearing in our risk bound (8). Recall also the quantity $R_Z(n\theta)$ that appears in (9). The first result of this section states that $R(n; \theta)$ is always bounded from above by $R_Z(n; \theta)$ up to a logarithmic multiplicative factor in n. This implies that (8) is always only slightly worse off than (9) (by a logarithmic multiplicative factor) while being much better when θ is well-approximated by some $\alpha \in \mathcal{M}$ for which $k(\alpha)$ is small. Recall that $V(\theta) := \theta_n - \theta_1$. The proofs of all the results in this section can be found in the supplementary material [13].

THEOREM 4.1. *For every $\theta \in \mathcal{M}$, we have*

$$R(n; \theta) \leq 16 \log(4n) \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \quad (33)$$

whenever

$$n \geq \max\left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{V(\theta)}{\sigma}\right). \quad (34)$$

In the next result, we characterize $R(n; \theta)$ for certain strictly increasing sequences θ where we show that it is essentially of the order $(\sigma^2 V(\theta)/n)^{2/3}$. In some sense, $R(n; \theta)$ is maximized for these strictly increasing sequences. The prototypical sequence we have in mind here is $\theta_i = i/n$ for $1 \leq i \leq n$.

THEOREM 4.2. *Suppose* $\theta_1 < \theta_2 < \cdots < \theta_n$ *with*

$$\min_{2 \le i \le n} (\theta_i - \theta_{i-1}) \ge \frac{c_1 V(\theta)}{n} \tag{35}$$

for a positive constant $c_1 \le 1$. *Then we have*

$$12\left(\frac{c_1 \sigma^2 V(\theta)}{n}\right)^{2/3} \le R(n, \theta) \le 16\left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \log(4n) \tag{36}$$

provided

$$n \ge \max\left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{2V(\theta)}{\sigma}\right). \tag{37}$$

REMARK 4.1. An important situation where (35) is satisfied is when θ arises from sampling a function on [0, 1] at the points i/n for $i = 1, \ldots, n$, assuming that the derivative of the function is bounded from below by a positive constant.

Next we describe sequences θ for which $R(n;\theta)$ is $(k(\theta)\sigma^2/n)\log(en/k(\theta))$, up to multiplicative factors. For these sequences our risk bound is potentially far superior to $R_Z(n;\theta)$.

THEOREM 4.3. *Let* $k = k(\theta)$ *with* $\{y : y = \theta_j \text{ for some } j\} = \{\theta_{0,1}, \ldots, \theta_{0,k}\}$ *where* $\theta_{0,1} < \cdots < \theta_{0,k}$. *Then*

$$\frac{\sigma^2 k}{n} \log \frac{en}{k} \le R(n;\theta) \le \frac{16\sigma^2 k}{n} \log \frac{en}{k} \tag{38}$$

provided

$$\min_{2 \le i \le k} (\theta_{0,i} - \theta_{0,i-1}) \ge \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \tag{39}$$

5. Local minimax optimality of the LSE. In this section, we establish an optimality property of the LSE. Specifically, we show that $\hat{\theta}$ is locally minimax optimal in a nonasymptotic sense. "Local" here refers to a ball $\{t : \ell_\infty^2(t,\theta) \le cR(n;\theta)\}$ around the true parameter θ for a positive constant c. The reason we focus on local minimaxity, as opposed to the more traditional notion of global minimaxity, is that the rate $R(n;\theta)$ changes with θ. Note that, moreover, lower bounds on the global minimax risk follow from our local minimax lower bounds. Such an optimality theory based on local minimaxity has been pioneered by Cai and Low [9] and Cator [11] for the problem of estimating a convex or monotone function at a point.

We start by proving an upper bound for the local supremum risk of $\hat{\theta}$. Recall that $\ell_\infty(t,\theta) := \max_{1 \le i \le n} |t_i - \theta_i|$.

LEMMA 5.1. *The following inequality holds for every $\theta \in \mathcal{M}$ and $c > 0$:*

$$\sup_{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \le cR(n;\theta)} \mathbb{E}_t \ell^2(t, \hat{\theta}) \le 2(1+4c) R(n;\theta). \tag{40}$$

PROOF. Inequality (8) gives $\mathbb{E}_t \ell^2(t, \hat{\theta}) \le R(n;t)$ for every $t \in \mathcal{M}$. Fix $\pi \in \Pi$. By the triangle inequality, we get $V_\pi(t) \le 2\ell_\infty(t,\theta) + V_\pi(\theta)$. As a result, whenever $\ell_\infty(t,\theta) \le cR(n;\theta)$, we obtain

$$V_\pi^2(t) \le 2V_\pi^2(\theta) + 8\ell_\infty^2(t,\theta) \le 2V_\pi^2(\theta) + 8cR(n;\theta).$$

As a consequence,

$$\begin{aligned}\mathbb{E}_t \ell^2(t,\hat{\theta}) \le R(n;t) &\le \inf_{\pi \in \Pi} \left(2V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{n}{k(\pi)} \right) + 8cR(n;\theta) \\ &\le 2R(n;\theta) + 8cR(n;\theta).\end{aligned}$$

This proves (40). □

We now show that $R(n;\theta)$, up to logarithmic factors in n, is a lower bound for the local minimax risk at θ, defined as the infimum of the right-hand side of (40) over all possible estimators $\hat{\theta}$. We prove this under each of the assumptions (1) and (2) (stated in the Introduction) on θ. Specifically, we prove the two inequalities (11) and (12). These results mean that, when θ satisfies either of the two assumptions (1) or (2), no estimator can have a supremum risk significantly better than $R(n;\theta)$ in the local neighborhood $\{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \lesssim R(n;\theta)\}$. On the other hand, Lemma 5.1 states that the supremum risk of the LSE over the same local neighborhood is bounded from above by a constant multiple of $R(n;\theta)$. Putting these two results together, we deduce that the LSE is approximately locally nonasymptotically minimax for such sequences θ. We use the qualifier "approximately" here because of the presence of logarithmic factors on the right-hand sides of (11) and (12).

We make here the assumption that the errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and normally distributed with mean zero and variance σ^2. For each $\theta \in \mathcal{M}$, let $\mathbb{P}_\theta$ denote the joint distribution of the data $Y_1, \dots, Y_n$ when the true sequence equals θ. As a consequence of the normality of the errors, we have

$$D(\mathbb{P}_\theta \| \mathbb{P}_t) = \frac{n}{2\sigma^2} \ell^2(t,\theta),$$

where $D(P\|Q)$ denotes the Kullback–Leibler divergence between the probability measures P and Q. Our main tool for the proofs is Assouad's lemma, the following version of which is a consequence of Lemma 24.3 of [31], page 347.

LEMMA 5.2 (Assouad). *Let m be a positive integer and suppose that, for each $\tau \in \{0,1\}^m$, there is an associated nondecreasing sequence θ^τ in $N(\theta)$, where $N(\theta)$ is a neighborhood of θ. Then the following inequality holds*:

$$\inf_{\hat{t}} \sup_{t \in N(\theta)} \mathbb{E}_t \ell^2(t, \hat{t}) \geq \frac{m}{8} \min_{\tau \neq \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{\Upsilon(\tau, \tau')} \min_{\Upsilon(\tau,\tau')=1} \big(1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{\mathrm{TV}}\big),$$

where $\Upsilon(\tau, \tau') := \sum_i I\{\tau_i \neq \tau'_i\}$ denotes the Hamming distance between τ and τ' and $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance between probability measures. The infimum here is over all possible estimators $\hat{t}$.

Inequalities (11) and (12) are proved in the next two subsections.

5.1. *Uniform increments.* In this section, we assume that θ is a strictly increasing sequence with $V(\theta) = \theta_n - \theta_1 > 0$ and that

$$\frac{c_1 V(\theta)}{n} \leq \theta_i - \theta_{i-1} \leq \frac{c_2 V(\theta)}{n} \qquad \text{for } i = 2, \dots, n \tag{41}$$

for some $c_1 \in (0, 1]$ and $c_2 \geq 1$. Because $V(\theta) = \sum_{i=2}^n (\theta_i - \theta_{i-1})$, assumption (41) means that the increments of θ are in a sense uniform. An important example in which (41) is satisfied is when $\theta_i = f_0(i/n)$ for some function f_0 on $[0, 1]$ whose derivative is uniformly bounded from above and below by positive constants.

In the next theorem, we prove that the local minimax risk at θ is bounded from below by $R(n;\theta)$ (up to logarithmic multiplicative factors) when θ satisfies (41).

THEOREM 5.3. *Suppose θ satisfies* (41), *and let*

$$\mathfrak{N}(\theta) := \left\{ t \in \mathcal{M} : \ell_\infty^2(t, \theta) \leq \left(\frac{3c_2}{c_1}\right)^{2/3} \frac{R(n;\theta)}{12} \right\}.$$

Then the local minimax risk $\mathfrak{R}_n(\theta) := \inf_{\hat{t}} \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t, \hat{t})$ satisfies the following inequality:

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 3^{2/3}}{256 c_2^{4/3}} \left(\frac{\sigma^2 V(\theta)}{n}\right)^{2/3} \geq \frac{c_1^2 3^{2/3}}{4096 c_2^{4/3}} \frac{R(n;\theta)}{\log(4n)}, \tag{42}$$

provided

$$n \geq \max\left(2, \frac{24\sigma^2}{V^2(\theta)}, \frac{2c_2 V(\theta)}{\sigma}\right). \tag{43}$$

Theorem 5.3 is closely connected to minimax lower bounds for Lipschitz classes of functions. Indeed, using the notation $v_f := (f(1/n), \dots, f(1))$ for functions f on $[0, 1]$, it can be argued that

$$\{t = \theta + v_f : \|f'\|_\infty \leq c_1 V(\theta), \|f\|_\infty \leq c'_1 n^{-1/3}\}$$

is a subset of $\mathfrak{N}(\theta)$ for appropriate positive constants c_1 and c_1'. Lower bound (42) then follows from usual lower bounds for Lipschitz classes which are outlined, for example, in [28], Chapter 2. A direct proof of Theorem 5.3 is included in the supplementary material [13].

5.2. *Piecewise constant*. Here, we again show that the local minimax risk at θ is bounded from below by $R(n;\theta)$ (up to logarithmic multiplicative factors). The difference from the previous section is that we work under a different assumption from (41). Specifically, we assume that $k(\theta) = k$ and that the k values of θ are sufficiently well separated and prove inequality (12).

Let $k = k(\theta)$. There exist integers $n_1, \dots, n_k$ with $n_i \geq 1$ and $n_1 + \cdots + n_k = n$ such that θ is constant on each set $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \dots, k$ where $s_0 := 0$ and $s_i := n_1 + \cdots + n_i$. Also, let the values of θ on the sets $\{j : s_{i-1} + 1 \leq j \leq s_i\}$ for $i = 1, \dots, k$ be denoted by $\theta_{0,1} < \cdots < \theta_{0,k}$.

THEOREM 5.4. *Suppose $c_1 n/k \leq n_i \leq c_2 n/k$ for all $1 \leq i \leq k$ for some $c_1 \in (0,1]$ and $c_2 \geq 1$ and that*

$$\min_{2\leq i\leq k} (\theta_{0,i} - \theta_{0,i-1}) \geq \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \tag{44}$$

Then, with $\mathfrak{N}(\theta)$ defined as $\{t \in \mathcal{M} : \ell_\infty^2(t,\theta) \leq R(n;\theta)\}$, the local minimax risk, $\mathfrak{R}_n(\theta) = \inf_{\hat{t}} \sup_{t\in\mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t,\hat{t})$, satisfies

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^{7/3}}{2^{31/3} c_2^2} R(n;\theta) \left(\log \frac{en}{k}\right)^{-2/3},$$

provided

$$\frac{n}{k} \geq \max\left(\left(\frac{4}{c_1^2} \log \frac{en}{k}\right)^{1/3}, \exp\left(\frac{1-4c_1}{4c_1}\right)\right). \tag{45}$$

PROOF. For notational convenience, we write

$$\beta_n^2 := \frac{k\sigma^2}{n} \log \frac{en}{k}.$$

First note that under assumption (44), Theorem 4.3 implies that $\beta_n^2 \leq R(n;\theta)$.

Let $1 \leq l \leq \min_{1\leq i\leq k} n_i$ be a positive integer whose value will be specified later, and let $m_i := \lfloor n_i/l \rfloor$ for $i = 1, \dots, k$. We also write M for $\sum_{i=1}^k m_i$.

The elements of the finite set $\{0,1\}^M$ will be represented as $\tau = (\tau_1, \dots, \tau_k)$ where $\tau_i = (\tau_{i1}, \dots, \tau_{im_i}) \in \{0,1\}^{m_i}$. For each $\tau \in \{0,1\}^M$, we specify $\theta^\tau \in \mathcal{M}$ in the following way. For $s_{i-1} + 1 \leq u \leq s_i$, the quantity θ_u^τ is defined as

$$\theta_{0,i} + \frac{\beta_n}{m_i} \sum_{v=1}^{m_i} (v - \tau_{iv}) I\{(v-1)l + 1 \leq u - s_{i-1} \leq vl\}$$
$$+ \beta_n I\{s_{i-1} + m_i l + 1 \leq u \leq s_i\}.$$

Because θ is constant on the set $\{u : s_{i-1} + 1 \le u \le s_i\}$ where it takes the value $\theta_{0,i}$, it follows that $\ell_\infty(\theta^\tau, \theta) \le \beta_n$. This implies that $\theta^\tau \in \mathfrak{N}(\theta)$ for every τ as $\beta_n^2 \le R(n; \theta)$.

Also, because of the assumption $\min_{2 \le i \le k}(\theta_{0,i} - \theta_{0,i-1}) \ge \beta_n$, it is evident that each θ^τ is nondecreasing. We will apply Assouad's lemma to θ^τ, $\tau \in \{0, 1\}^M$. For $\tau, \tau' \in \{0, 1\}^M$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) = \frac{1}{n} \sum_{i=1}^{k} \sum_{v=1}^{m_i} \frac{l\beta_n^2}{m_i^2} I\{\tau_{iv} \ne \tau'_{iv}\} = \frac{l\beta_n^2}{n} \sum_{i=1}^{k} \frac{\Upsilon(\tau_i, \tau'_i)}{m_i^2}. \tag{46}$$

Because

$$m_i \le \frac{n_i}{l} \le \frac{c_2 n}{kl} \qquad \text{for each } 1 \le i \le k,$$

we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) \ge \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3} \sum_{i=1}^{k} \Upsilon(\tau_i, \tau'_i) = \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3} \Upsilon(\tau, \tau'). \tag{47}$$

Also, from (46), we get

$$\ell^2(\theta^\tau, \theta^{\tau'}) \le \frac{l\beta_n^2}{n(\min_{1 \le i \le k} m_i^2)} \qquad \text{when } \Upsilon(\tau, \tau') = 1. \tag{48}$$

The quantity $\min_i m_i^2$ can be easily bounded from below by noting that $n_i/l < m_i + 1 \le 2m_i$ and that $n_i \ge c_1 n/k$. This gives

$$\min_{1 \le i \le k} m_i \ge \frac{c_1 n}{2kl}. \tag{49}$$

Combining the above inequality with (48), we deduce

$$\ell^2(\theta^\tau, \theta^{\tau'}) \le \frac{4k^2 l^3 \beta_n^2}{c_1^2 n^3} \qquad \text{whenever } \Upsilon(\tau, \tau') = 1.$$

This and Pinsker's inequality give

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{\mathrm{TV}}^2 \le \frac{1}{2} D(\mathbb{P}_{\theta^\tau} \| \mathbb{P}_{\theta^{\tau'}}) = \frac{n}{4\sigma^2} \ell^2(\theta^\tau, \theta^{\tau'}) \le \frac{k^2 l^3 \beta_n^2}{c_1^2 n^2 \sigma^2} \tag{50}$$

whenever $\Upsilon(\tau, \tau') = 1$.

Inequalities (47) and (50) in conjunction with Assouad's lemma give

$$\mathfrak{R}_n(\theta) \ge \frac{M k^2 l^3 \beta_n^2}{8 c_2^2 n^3} \left(1 - \frac{k\beta_n l^{3/2}}{c_1 n \sigma}\right).$$

Because of (49), we get $M = \sum_i m_i \ge k \min_i m_i \ge c_1 n/(2l)$, and thus

$$\mathfrak{R}_n(\theta) \ge \frac{c_1 k^2 l^2 \beta_n^2}{16 c_2^2 n^2} \left(1 - \frac{k\beta_n l^{3/2}}{c_1 n \sigma}\right). \tag{51}$$

The value of the integer l will now be specified. We take

$$l = \left(\frac{c_1 n\sigma}{2k\beta_n}\right)^{2/3}. \tag{52}$$

Because $\min_i n_i \ge c_1 n/k$, we can ensure that $1 \le l \le \min_i n_i$ by requiring that

$$1 \le \left(\frac{c_1 n\sigma}{2k\beta_n}\right)^{2/3} \le \frac{c_1 n}{k}.$$

This gives rise to two lower bounds for n which are collected in (45).

As a consequence of (52), we get that $l^{3/2} \le c_1 n\sigma/(2k\beta_n)$, which ensures that the term inside the parentheses on the right-hand side of (51) is at least $1/2$. This gives

$$\mathfrak{R}_n(\theta) \ge \frac{c_1 k^2 l^2 \beta_n^2}{32 c_2^2 n^2} \ge \frac{c_1^{7/3}}{2^{19/3} c_2^2} \frac{k\sigma^2}{n}\left(\log\frac{en}{k}\right)^{1/3}. \tag{53}$$

To complete the proof, we use Theorem 4.3. Specifically, the second inequality in (38) gives

$$\frac{k\sigma^2}{n} \ge \frac{R(n;\theta)}{16}\left(\log\frac{en}{k}\right)^{-1}.$$

The proof is complete by combining the above inequality with (53). $\square$

6. Risk bound under model misspecification. We consider isotonic regression under the misspecified setting where the true sequence is not necessarily nondecreasing. Specifically, consider model (1) where now the true sequence θ is not necessarily assumed to be in $\mathcal{M}$. We study the behavior of the LSE $\hat{\theta} = \hat{\theta}(Y; \mathcal{M})$. The goal of this section is to prove an inequality analogous to (8) for model misspecification. It turns out here that the LSE is really estimating the nondecreasing projection of θ on $\mathcal{M}$ defined as $\tilde{\theta} \in \mathcal{M}$ that minimizes $\ell^2(t,\theta)$ over $t \in \mathcal{M}$. From [24], Chapter 1, it follows that

$$\tilde{\theta}_j = \min_{l \ge j} \max_{k \le j} \bar{\theta}_{k,l} \qquad \text{for } 1 \le j \le n, \tag{54}$$

where $\bar{\theta}_{k,l}$ is as defined in (22).

We define another measure of variation for $t \in \mathcal{M}$ with respect to an interval partition $\pi = (n_1, \dots, n_k)$:

$$S_\pi(t) = \left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=s_{i-1}+1}^{s_i}(t_{s_i} - t_j)^2\right)^{1/2},$$

where $s_0 = 0$ and $s_i = n_1 + \cdots + n_i$ for $1 \le i \le k$. It is easy to check that $S_\pi(t) \le V_\pi(t)$ for every $t \in \mathcal{M}$. The following is the main result of this section. The proofs of all the results in this section can be found in the supplementary material [13].

THEOREM 6.1. *For every $\theta \in \mathbb{R}^n$, the LSE satisfies*

$$\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}) \le 4 \inf_{\pi \in \Pi} \left(S_\pi^2(\tilde{\theta}) + \frac{4\sigma k(\pi)}{n} \log \frac{en}{k(\pi)} \right) \le R(n; \tilde{\theta}). \tag{55}$$

REMARK 6.1. By Theorem 4.1, the quantity $R(n; \tilde{\theta})$ is bounded from above by $(\sigma^2 V(\tilde{\theta})/n)^{2/3}$ up to a logarithmic multiplicative factor in n. Therefore, Theorem 6.1 implies that the LSE $\hat{\theta}$ converges to the projection of θ onto the space of monotone vectors at least the $n^{-2/3}$ rate, up to a logarithmic factor in n. The convergence rate will be much faster if $k(\tilde{\theta})$ is small or if $\tilde{\theta}$ is well approximated by a monotone vector α with small $k(\alpha)$.

By taking π in the infimum in the upper bound of (55) to be the interval partition generated by $\tilde{\theta}$, we obtain the following result which is the analogue of (10) for model misspecification.

COROLLARY 6.2. *For every arbitrary sequence θ of length n (not necessarily nondecreasing),*

$$\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}) \le \frac{16\sigma^2 k(\tilde{\theta})}{n} \log \frac{en}{k(\tilde{\theta})}.$$

In the next pair of results, we prove two upper bounds on $k(\tilde{\theta})$. The first result shows that $k(\tilde{\theta}) = 1$ (i.e., $\tilde{\theta}$ is constant) when θ is nonincreasing, that is, $\theta_1 \ge \theta_2 \ge \cdots \ge \theta_n$. This implies that the LSE converges to $\tilde{\theta}$ at the rate $\sigma^2 \log(en)/n$ when θ is nonincreasing.

LEMMA 6.3. *$k(\tilde{\theta}) = 1$ if θ is nonincreasing.*

To state our next result, let

$$b(t) := \sum_{i=1}^{n-1} I\{t_i \ne t_{i+1}\} + 1 \qquad \text{for } t \in \mathbb{R}^n. \tag{56}$$

$b(t)$ can be interpreted as the number of constant blocks of t. For example, when $n = 5$ and $t = (0, 0, 1, 1, 1, 0)$, $b(t) = 3$. Observe that $b(t) = k(t)$ for $t \in \mathcal{M}$.

LEMMA 6.4. *For any sequence $\theta \in \mathbb{R}^n$, we have $k(\tilde{\theta}) \le b(\theta)$.*

As a consequence of the above lemma, we obtain that for every $\theta \in \mathbb{R}^n$, the quantity $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta})$ is bounded from above by $(16 b(\theta)\sigma^2/n) \log(en/b(\theta))$.

Acknowledgements. We thank the Associate Editor and the anonymous referees for constructive suggestions that significantly improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "On risk bounds in isotonic and other shape restricted regression problems" (DOI: 10.1214/15-AOS1324SUPP; .pdf). In the supplementary paper [13] we provide the proofs of Lemmas 2.4, 2.5, 2.6, 6.3 and 6.4 and Theorems 4.1, 4.2, 4.3, 5.3 and 6.1. We also state and prove Lemma 11.1, which is used in the proof of Theorem 4.1, and Lemma 11.2, which is used in the proof of Theorem 3.1.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2013). Living on the edge: A geometric theory of phase transitions in convex optimization. Available at arXiv:1303.6672.

[2] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641–647. MR0073895

[3] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[4] BERTSEKAS, D. P. (2003). *Convex Analysis and Optimization*. Athena Scientific, Belmont, MA. MR2184037

[5] BIRGÉ, L. (1989). The Grenander estimator: A nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298

[6] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

[7] BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26** 607–616. MR0073894

[8] BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* (*Proc. Sympos., Indiana Univ., Bloomington, Ind.*, 1969) 177–197. Cambridge Univ. Press, London. MR0277070

[9] CAI, T. and LOW, M. (2011). A framework for estimation of convex functions. Available at http://www-stat.wharton.upenn.edu/~tcai/.

[10] CAROLAN, C. and DYKSTRA, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.* **27** 557–566. MR1745821

[11] CATOR, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli* **17** 714–735. MR2787612

[12] 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

[13] CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). Supplement to "On risk bounds in isotonic and other shape restricted regression problems." DOI:10.1214/15-AOS1324SUPP.

[14] DONOHO, D. (1991). Gel'fand n-widths and the method of least squares. Technical report, Dept. Statistics, Univ. California, Berkeley, CA.

[15] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th ed. Cambridge Univ. Press, Cambridge. MR2722836

[16] GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

[17] GROENEBOOM, P. (1983). The concave majorant of Brownian motion. *Ann. Probab.* **11** 1016–1027. MR0714964

[18] GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). 539–555. Wadsworth, Belmont, CA. MR0822052

[19] GUNTUBOYINA, A. and SEN, B. (2014). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* 1–33.

[20] JANKOWSKI, H. (2014). Convergence of linear functionals of the Grenander estimator under misspecification. *Ann. Statist.* **42** 625–653. MR3210981

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

[22] OYMAK, S. and HASSIBI, B. (2013). Sharp MSE bounds for proximal denoising. Available at arXiv:1305.2714.

[23] RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134

[24] ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester. MR0961262

[25] RUDELSON, M. and VERSHYNIN, R. (2006). Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. In 40*th Annual Conference on Information Sciences and Systems* 207–212. IEEE, New York.

[26] SCHRIJVER, A. (1986). *Theory of Linear and Integer Programming*. Wiley, Chichester. MR0874114

[27] STOJNIC, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. Available at arXiv:0907.3666.

[28] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[29] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. MR1056343

[30] VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. MR1212164

[31] VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[32] VERSHYNIN, R. (2014). Lectures in geometric functional analysis. Available at www-personal.umich.edu/~romanv.

[33] WANG, Y. (1996). The L_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods* **25** 281–294. MR1379445

[34] WRIGHT, F. T. (1981). The asymptotic behavior of monotone regression estimates. *Ann. Statist.* **9** 443–448. MR0606630

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

S. CHATTERJEE
DEPARTMENT OF STATISTICS
UNIVERSITY OF CHICAGO
5734 S. UNIVERSITY AVENUE
CHICAGO, ILLINOIS 60637
USA
E-MAIL: sabyasachi@galton.uchicago.edu

A. GUNTUBOYINA
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
423 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: aditya@stat.berkeley.edu

B. SEN
DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
1255 AMSTERDAM AVENUE
NEW YORK, NEW YORK 10027
USA
E-MAIL: bodhi@stat.columbia.edu