

Improved Risk Bounds in Isotonic Regression

Sabyasachi Chatterjee, Adityanand Guntuboyina* and Bodhisattva Sen†

Yale University, University of California at Berkeley and Columbia University

November 13, 2013

Abstract

We consider the problem of estimating an unknown non-decreasing sequence θ from finitely many noisy observations. We give an improved global risk upper bound for the isotonic least squares estimator (LSE) in this problem. The obtained risk bound behaves differently depending on the form of the true sequence θ – one gets 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). In particular, when θ has k constant pieces then the risk bound becomes $(k/n) \log(en/k)$. As a consequence, we illustrate the adaptation properties of the LSE. We also prove an analogue of the risk bound for model misspecification, i.e., when θ is not non-decreasing. We also prove local minimax lower bounds for this problem which show that the LSE is nearly optimal in a local nonasymptotic minimax sense.

1 Introduction

We consider the problem of estimating an unknown non-decreasing regression function f_0 from finitely many noisy observations. Specifically, only under the assumption that f_0 is non-decreasing, the goal is to estimate f_0 from data $(x_1, Y_1), \dots, (x_n, Y_n)$ with

$$Y_i = f_0(x_i) + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n, \quad (1)$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d mean zero errors with variance $\sigma^2 > 0$ and $x_1 < \dots < x_n$ are fixed design points. This is one of the canonical problems in the area of shape-constrained nonparametric function estimation.

*Supported by NSF Grant DMS-1309356

†Supported by NSF Grants DMS-1150435 and AST-1107373

The most natural and commonly used estimator for this problem is the monotone least squares estimator (LSE), proposed by [Ayer et al. \(1955\)](#); also see [Grenander \(1956\)](#) for the related problem of estimating a non-increasing density. The LSE is defined as any minimizer of the LS criterion:

$$\hat{f}_{ls} \in \operatorname{argmin}_{g \in \mathcal{C}} \sum_{i=1}^n (Y_i - g(x_i))^2 \quad (2)$$

where $\mathcal{C}$ denotes the set of all real-valued non-decreasing functions on $\mathbb{R}$. The values $\hat{f}_{ls}(x_1), \dots, \hat{f}_{ls}(x_n)$ are unique and can be computed easily using the *pool adjacent violators algorithm*; see [Robertson et al. \(1988, Chapter 1\)](#).

Before describing our results, let us briefly review the existing related theoretical results on monotone function estimation. These results 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 $\hat{f}_{ls}$ is from f_0 .

Results on the local behavior are proved, among others, in [Brunk \(1970\)](#), [Carolan and Dykstra \(1999\)](#), [Cator \(2011\)](#), and [Jankowski and Wellner \(2012\)](#). Under certain regularity conditions on the unknown function f_0 near the interior point x_0 , [Brunk \(1970\)](#) showed that $\hat{f}_{ls}(x_0)$ converges to $f_0(x_0)$ at the rate $n^{-1/3}$ and also characterized the limiting distribution of $n^{1/3}(\hat{f}_{ls}(x_0) - f_0(x_0))$. In the related density estimation problem, [Carolan and Dykstra \(1999\)](#) and [Jankowski and Wellner \(2012\)](#) showed that if the interior point x_0 lies on a flat stretch of the underlying function then the estimator converges to a non-degenerate limit at rate $n^{-1/2}$, and they characterized the limiting distribution. [Cator \(2011\)](#) showed that the rate of convergence of $\hat{f}_{ls}(x_0)$ 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 $\hat{f}_{ls}$ adapts automatically to the unknown function f_0 . [Cator \(2011\)](#) also proved optimality of the LSE for local behavior by proving 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 $\hat{f}_{ls}$ under a global loss function. The most natural and commonly studied global loss function in this setup is

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

for real valued functions f and g . Note that under this loss function, the problem posited in (1) becomes a vector or sequence estimation problem where the goal is to estimate the vector $\theta := (\theta_1, \dots, \theta_n)$ from observations

$$Y_i = \theta_i + \epsilon_i, \quad \text{for } i = 1, 2, \dots, n, \quad (3)$$

under the constraint that the unknown sequence $\theta = (\theta_1, \dots, \theta_n)$ satisfies $\theta_1 \leq \dots \leq \theta_n$. In other words, we assume that θ lies in the closed convex polyhedral cone $\mathcal{M}$ defined as

$$\mathcal{M} := \{t \in \mathbb{R}^n : t_1 \leq t_2 \leq \dots \leq t_n\}.$$

The connection between this sequence estimation problem and the earlier problem of monotone regression is apparent by the identification $\theta_i = f_0(x_i)$. As a result, it is obvious that the exact form of the design points $x_1, \dots, x_n$ is irrelevant in the problem (1) as long as $x_1 < \dots < x_n$.

Henceforth we would be stating our results for sequences although they could be thought of as being generated from a monotone function sampled at certain design points. The LSE, $\hat{\theta}$, of θ is defined as

$$\hat{\theta} := \operatorname{argmin}_{t \in \mathcal{M}} \sum_{i=1}^n (Y_i - t_i)^2 \quad (4)$$

and it is easy to see that it exists and is unique. We study the risk of $\hat{\theta}$ under the loss function $\sum_{i=1}^n (t_i - s_i)^2/n$ which we denote by $\ell^2(t, s)$ by a slight abuse of notation.

The behavior of the LSE, $\hat{\theta}$, under the loss ℓ^2 has been studied in a number of papers including [Van de Geer \(1990, 1993\)](#), [Donoho \(1991\)](#), [Birgé and Massart \(1993\)](#), [Wang \(1996\)](#), [Meyer and Woodroffe \(2000\)](#) and [Zhang \(2002\)](#). The strongest results here are obtained by [Zhang \(2002\)](#) who proved nonasymptotic bounds on the risk $\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta)$ under a minimal set of assumptions on the errors $\epsilon_1, \dots, \epsilon_n$. Among other things, [Zhang \(2002, Theorem 2.2\)](#) showed that

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \lesssim R_Z(n; \theta) \quad (5)$$

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 the variation of the sequence θ . When $V(\theta) = 0$, i.e., when θ is a constant sequence, this inequality implies that

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \lesssim \frac{\sigma^2 \log n}{n}. \quad (6)$$

This result is an example of the adaptation behavior of the LSE in the sense that the risk bound when θ is constant is much smaller than the general upper

bound (5). Similar results (which hold only in an asymptotic sense) on the adaptation of the LSE when θ is constant can be found in [Groeneboom and Pyke \(1983\)](#) and [Van de Geer \(1993\)](#). Other results on the global behavior of the LSE can be found in [Durot \(2002, 2007\)](#) who considered a different global loss function and characterized the limiting distribution of a suitably normalized version.

An important shortcoming of the risk bound (5) is that it is not optimal in the sense that the risk $\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta)$ can, in general, be much smaller than $R_Z(n; \theta)$. The main result of this paper presents an improvement of (5). To describe our result, let us introduce some notation. We say an *interval partition* π of a positive integer 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

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

For each $\pi \in \Pi$, let $k(\pi)$ denote the length of the associated sequence $\{n_i\}$.

For each $\theta = (\theta_1, \dots, \theta_n) \in \mathcal{M}$, there exist integers k and $n_1, \dots, n_k$ with $n_i \geq 1$ and $n_1 + \dots + n_k = n$ 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 $s_i = n_1 + \dots + n_i$. We refer to the interval partition $\pi_\theta := (n_1, \dots, n_k)$ as the interval partition *generated* by θ . By a slight abuse of notation, we denote $k(\pi_\theta)$ by simply $k(\theta)$. Note that $k(\theta)$ is just the number of distinct values among $\theta_1, \dots, \theta_n$. $k(\theta)$ is a measure of sparsity of the differences $\theta_i - \theta_{i-1}$ for $i = 2, \dots, n$.

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

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

where $s_0 := 0$ and $s_i = n_1 + \dots + n_i$ for $1 \leq i \leq k$. $V_\pi(\theta)$ 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 $V_\pi(\theta) = V(\theta)$.

We are now ready to state our main result. In Theorem 2.1, we prove the risk bound

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq R(n; \theta) \tag{7}$$

where

$$R(n; \theta) := 4 \inf_{\pi \in \Pi} \left(V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \right). \tag{8}$$

Let us explain why our bound (7) is an improvement of (5). Suppose, for example, $\theta_j = I\{j > n/2\}$ (here I denotes the indicator function) and hence

$V(\theta) = 1$. 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 (8).

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

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq \frac{16k(\theta)\sigma^2}{n} \log \frac{en}{k(\theta)}, \quad (9)$$

which is a stronger bound than (5) 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.

We prove some properties of our risk bound $R(n; \theta)$ in Section 3. In Theorem 3.1, 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 . Therefore, our inequality (7) is always only slightly worse off than (5) while being much better in the case of certain sequences θ . We also show in Section 3 that the risk bound $R(n; \theta)$ behaves differently depending on the form of the true sequence θ . This means that the bound (7) 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 $\min_i(\theta_i - \theta_{i-1}) \gtrsim 1/n$). The bound (7) therefore presents a bridge between the two terms in the bound (5). We suspect that such adaptation of the LSE in global estimation is rather unique to shape-constrained estimation. It has also been observed to some degree in Guntuboyina and Sen (2013) in the related convex regression problem.

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 monotone 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 4 by proving lower bounds for the *local minimax risk* near the “true” θ . For $\theta \in \mathcal{M}$, the quantity

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

with

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

will be called the local minimax risk at θ (see Section 4 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 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 4.3 that

$$\mathfrak{R}_n(\theta) \gtrsim \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} \gtrsim \frac{R(n; \theta)}{\log(4n)}; \quad (10)$$

2. when $k(\theta) = k$ and the k values of θ are sufficiently well-separated, we show in Theorem 4.4 that

$$\mathfrak{R}_n(\theta) \gtrsim R(n; \theta) \left(\log \frac{en}{k} \right)^{-2/3}. \quad (11)$$

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 non-asymptotic minimax sense.

We also study the performance of the LSE under model misspecification when the true sequence θ is not necessarily non-decreasing. Here we prove in Theorem 5.1 that $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta}) \leq R(n; \tilde{\theta})$ where $\tilde{\theta}$ denotes the non-decreasing projection of θ (see Section 5 for its definition). This should be contrasted with the risk bound of Zhang (2002) who proved that $\mathbb{E}_\theta \ell^2(\hat{\theta}, \tilde{\theta}) \lesssim R_Z(n; \tilde{\theta})$. As before our risk bound is atmost 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 (58) and Lemma 5.4) and when θ is non-increasing (in which case $k(\tilde{\theta}) = 1$; see Lemma 5.3).

The paper is organized as follows: In Section 2 we state and prove our main upper bound for the risk $\mathbb{E}_\theta \ell^2(\theta, \hat{\theta})$. We investigate the behavior of $R(n; \theta)$ for different values of the true sequence θ and compare it with $R_Z(n; \theta)$ in Section 3. Local minimax lower bounds for $\mathfrak{R}_n(\theta)$ for two different scenarios for θ are proved in Section 4. We study the performance of the LSE under model misspecification in Section 5. Section 6 gives some auxiliary results needed in the proofs of our main results.

Finally, we note that all the results in our paper are non-asymptotic in nature. Moreover, all the proofs make use of very light probabilistic apparatus and are thus accessible to a larger readership.

2 Our risk bound

In the following theorem, we present our risk bound (7); in fact, we prove a slightly stronger inequality than (7). Before stating our main result we introduce some notation. For any sequence $(a_1, a_2, \dots, a_n) \in \mathbb{R}^n$ and any $1 \leq k \leq l \leq n$, let

$$\bar{a}_{k,l} := \frac{1}{l-k+1} \sum_{j=k}^l a_j. \quad (12)$$

We will use this notation mainly when a equals either the sequence Y or θ . Our proof is based on the following explicit representation of the LSE $\hat{\theta}$ (see Chapter 1 of [Robertson et al. \(1988\)](#)):

$$\hat{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{Y}_{k,l}. \quad (13)$$

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

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

where $s_0 = 0$ and $s_i = n_1 + \dots + n_i$ for $1 \leq i \leq k$. 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

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

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 (7) with $D_\pi(\theta)$ replacing $V_\pi(\theta)$ in (7).

Theorem 2.1. *For every $\theta \in \mathcal{M}$, the risk of the LSE satisfies the following inequality:*

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq 4 \inf_{\pi \in \Pi} \left(D_\pi^2(\theta) + \frac{4\sigma k(\pi)}{n} \log \frac{en}{k(\pi)} \right). \quad (14)$$

Proof. Fix $1 \leq j \leq n$ and $0 \leq m \leq n-j$. By (13), we have

$$\hat{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{Y}_{k,l} \leq \max_{k \leq j} \bar{Y}_{k,j+m} = \max_{k \leq j} (\bar{\theta}_{k,j+m} + \bar{\epsilon}_{k,j+m})$$

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

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

Taking positive parts, we have

$$\left(\hat{\theta}_j - \theta_j\right)_+ \leq (\bar{\theta}_{j,j+m} - \theta_j) + \max_{k \leq j} (\bar{\epsilon}_{k,j+m})_+.$$

Squaring and taking expectations on both sides, we obtain

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_+^2 \leq \mathbb{E}_\theta \left((\bar{\theta}_{j,j+m} - \theta_j) + \max_{k \leq j} (\bar{\epsilon}_{k,j+m})_+ \right)^2.$$

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

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_+^2 \leq 2 (\bar{\theta}_{j,j+m} - \theta_j)^2 + 2 \mathbb{E} \max_{k \leq j} (\bar{\epsilon}_{k,j+m})_+^2.$$

We observe now that, for fixed integers j and m , the process $\{\bar{\epsilon}_{k,j+m}, k = 1, \dots, j\}$ 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 $\epsilon_1, \dots, \epsilon_{i-1}$ and $\bar{\epsilon}_{i,j+m}$. Therefore, by Doob's inequality for submartingales (see e.g., Theorem 5.4.3 of [Durrett \(2010\)](#)), we have

$$\mathbb{E} \max_{k \leq j} (\bar{\epsilon}_{k,j+m})_+^2 \leq 4 \mathbb{E} (\bar{\epsilon}_{j,j+m})_+^2 \leq 4 \mathbb{E} (\bar{\epsilon}_{j,j+m})^2 \leq \frac{4\sigma^2}{m+1}.$$

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

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_+^2 \leq 2 (\bar{\theta}_{j,j+m} - \theta_j)^2 + \frac{8\sigma^2}{m+1}. \quad (15)$$

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

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)_-^2 \leq 2 (\theta_j - \bar{\theta}_{j-m,j})^2 + \frac{8\sigma^2}{m+1}. \quad (16)$$

Let us now fix $\pi = (n_1, \dots, n_k) \in \Pi$. Let $s_0 := 0$ and $s_i := n_1 + \dots + n_i$ for $1 \leq i \leq k$. 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 $s_{i-1} + 1 \leq j \leq s_i$. We use this choice of $m_1(j)$ in (15) and $m_2(j)$ in (16) to obtain

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \theta_j\right)^2 \leq A_j + B_j$$

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

$$\mathbb{E}_\theta \ell^2(\hat{\theta}, \theta) \leq \frac{1}{n} \sum_{j=1}^n A_j + \frac{1}{n} \sum_{j=1}^n B_j. \quad (17)$$

We shall now prove that

$$\frac{1}{n} \sum_{j=1}^n A_j \leq 2D_\pi^2(\theta) + \frac{8k\sigma^2}{n} \log \frac{en}{k} \quad (18)$$

and

$$\frac{1}{n} \sum_{j=1}^n B_j \leq 2D_\pi^2(\theta) + \frac{8k\sigma^2}{n} \log \frac{en}{k}. \quad (19)$$

We give below the proof of (18) and the proof of (19) 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 $s_{i-1} + 1 \leq j \leq s_i$ and therefore

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

By Lemma 6.2, we get

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

for every $i = 1, \dots, k$. Thus, summing over $i = 1, \dots, k$, 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

$$\sum_{j=1}^n \frac{1}{m_1(j) + 1} = \sum_{i=1}^k \sum_{j=s_{i-1}+1}^{s_i} \frac{1}{s_i - j + 1} = \sum_{i=1}^k \left(1 + \frac{1}{2} + \dots + \frac{1}{n_i} \right). \quad (20)$$

Since the harmonic series $\sum_{l=1}^l 1/l$ is atmost $1 + \log l$ for $l \geq 1$, we obtain

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

where the last inequality is a consequence of the concavity of the logarithm function. We thus obtain

$$\sum_{j=1}^n \frac{1}{m_1(j) + 1} \leq k \log \frac{en}{k}.$$

This proves (18). Combining (18) and (19) proves the theorem. $\square$

Remark 2.1. Theorem 2.1 can be restated in the following way. For each $1 \leq k \leq n$, let $\mathcal{P}_k$ denote the set of all sequences $\alpha \in \mathcal{M}$ with $k(\alpha) \leq k$. With this notation, the statement of Theorem 2.1 can also be expressed as

$$\mathbb{E}_\theta \ell^2(\theta, \hat{\theta}) \leq 4 \min_{1 \leq k \leq n} \left[\inf_{\alpha \in \mathcal{P}_k} \ell^2(\theta, \alpha) + \frac{4\sigma^2 k}{n} \log \frac{en}{k} \right]. \quad (21)$$

This follows from the fact that $\min_{\alpha \in \mathcal{P}_k} \ell^2(\theta, \alpha) = \min_{\pi \in \Pi_k} D_\pi^2(\theta)$ where Π_k is the set of all $\pi \in \Pi$ with $k(\pi) \leq k$.

The bound (21) 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, for example, Barron et al. (1999)) or aggregation (see, for example, Rigollet and Tsybakov (2012)). 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 non-convex 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 the adaptive risk bound $R(n; \theta)$. Such adaptation of the LSE in global estimation is unique to shape-constrained function estimation and has been observed in Guntuboyina and Sen (2013) for the problem of convex regression.

3 The quantity $R(n; \theta)$

In this section, we prove some results about our risk bound $R(n; \theta)$. In the first result below, we prove 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 our risk bound (7) for the LSE is always only slightly worse off than (5) (by a logarithmic multiplicative factor) while being much better when θ is well-approximable by some $\alpha \in \mathcal{M}$ for which $k(\alpha)$ is small.

The proof of the next Theorem relies on Lemma 6.1 which is a result on the approximation of non-decreasing sequences by non-decreasing sequences with fewer pieces. Recall that $V(\theta) := \theta_n - \theta_1$.

Theorem 3.1. For every $\theta \in \mathcal{M}$, we have

$$R(n; \theta) \leq 16 \log(4n) \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} \quad (22)$$

whenever

$$n \geq \max \left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{V(\theta)}{\sigma} \right). \quad (23)$$

Proof. Fix $\theta \in \mathcal{M}$ and suppose (23) holds. Let $V = V(\theta)$. For every fixed $\delta \in [2V/n, V]$, we use Lemma 6.1 to assert the existence of π_δ such that $V_{\pi_\delta}(\theta) \leq \delta$ and

$$k(\pi_\delta) \leq \left\lceil \frac{V}{\delta} \right\rceil \leq \frac{V}{\delta} + 1 \leq \frac{2V}{\delta}.$$

By the definition of $R(n; \theta)$, we have

$$R(n; \theta) \leq 4 \left(V_{\pi_\delta}^2(\theta) + \frac{4\sigma^2 k(\pi_\delta)}{n} \log \frac{en}{k(\pi_\delta)} \right).$$

Note that the function $x \mapsto x \log(en/x)$ is non-decreasing for $x \in (0, n]$ and because $k(\pi_\delta) \leq 2V/\delta \leq n$, we deduce

$$R(n; \theta) \leq 4 \left(\delta^2 + \frac{8V\sigma^2}{n\delta} \log \frac{en\delta}{2V} \right) \leq 4 \left(\delta^2 + \frac{8V\sigma^2}{n\delta} \log \frac{en}{2} \right) \quad (24)$$

where in the last inequality above, we have used $\delta \leq V$. We now make the choice $\delta_0 = (8V\sigma^2/n)^{1/3}$. The constraint $2V/n \leq \delta_0 \leq V$ is guaranteed by the sample size condition in the statement of the theorem. The right hand side of (24) with $\delta = \delta_0$ gives the right hand side of (22), which completes the proof. $\square$

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 3.2. *Suppose $\theta_1 < \theta_2 < \dots < \theta_n$ with*

$$\min_{2 \leq i \leq n} (\theta_i - \theta_{i-1}) \geq \frac{c_1 V(\theta)}{n} \quad (25)$$

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

$$12 \left(\frac{c_1 \sigma^2 V(\theta)}{n} \right)^{2/3} \leq R(n, \theta) \leq 16 \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} \log(4n) \quad (26)$$

provided

$$n \geq \max \left(2, \frac{8\sigma^2}{V^2(\theta)}, \frac{2V(\theta)}{\sigma} \right). \quad (27)$$

Proof. For notational convenience we write V for $V(\theta)$. The upper bound in (26) follows from Theorem 3.1. Note that the upper bound therefore does not need the assumption (25).

For the lower bound in (26), fix $\pi = (n_1, \dots, n_k) \in \Pi$. Let $s_0 = 0$ and $s_i = n_1 + \dots + n_i$ for $1 \leq i \leq k$. By (25), we have

$$(\theta_{s_i} - \theta_{s_{i-1}+1}) \geq \sum_{l=s_{i-1}+2}^{s_i} (\theta_l - \theta_{l-1}) \geq \frac{c_1 V}{n} (n_i - 1)$$

for each $1 \leq i \leq k$. Consequently,

$$V_\pi(\theta) \geq \frac{c_1 V}{n} \max_{1 \leq i \leq k} (n_i - 1) \geq \frac{c_1 V}{nk} \sum_{i=1}^k (n_i - 1) = \frac{c_1 V(n - k)}{nk}.$$

As a result

$$R(n; \theta) \geq 4 \inf_{1 \leq k \leq n} \left(\frac{c_1^2 V^2 (n - k)^2}{n^2 k^2} + \frac{4\sigma^2 k}{n} \log \frac{en}{k} \right).$$

Writing $x = k/n$ and using $\log(en/k) \geq 1$, we get $R(n; \theta) \geq 4 \inf_{0 \leq x \leq 1} h(x)$ where

$$h(x) := \frac{c_1^2 V^2 (1 - x)^2}{n^2 x^2} + 4\sigma^2 x.$$

Now $h(x) \geq 4\sigma^2 x \geq 2\sigma^2$ for $x \in [1/2, 1]$. On the other hand, for $x \in [0, 1/2]$, because $1 - x \geq 1/2$, we get

$$h(x) \geq \inf_{x \geq 0} \left(\frac{c_1^2 V^2}{4n^2 x^2} + 4\sigma^2 x \right),$$

which, by calculus, gives

$$h(x) \geq 3 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3}.$$

Note that

$$3 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3} \leq 2\sigma^2$$

whenever $n \geq 2V/\sigma \geq 3^{3/2} c_1 V / (2^{3/2} \sigma)$. Thus, under this condition, we obtain

$$R(n; \theta) \geq 4 \inf_{0 \leq x \leq 1} h(x) \geq 12 \left(\frac{c_1 \sigma^2 V}{n} \right)^{2/3}$$

which proves the lower bound in (26). The proof is complete. $\square$

Remark 3.1. An important situation where (25) is satisfied is when θ arises from sampling a function on $[0, 1]$ at the points i/n for $i = 1, \dots, n$, 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 3.3. Let $k = k(\theta)$ with

$$\{y : y = \theta_j \text{ for some } j\} = \{\theta_{0,1}, \dots, \theta_{0,k}\}$$

where $\theta_{0,1} < \dots < \theta_{0,k}$. Then

$$\frac{\sigma^2 k}{n} \log \frac{en}{k} \leq R(n; \theta) \leq \frac{16\sigma^2 k}{n} \log \frac{en}{k} \quad (28)$$

provided

$$\min_{2 \leq i \leq k} (\theta_{0,i} - \theta_{0,i-1}) \geq \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \quad (29)$$

Proof. The upper bound in (28) is easily proved by taking $\pi = \pi_\theta$. For the lower bound, let us fix $\pi \in \Pi$. If $n \geq k(\pi) \geq k$, then

$$V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \geq \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \geq \frac{4\sigma^2 k}{n} \log \frac{en}{k} \quad (30)$$

where we have used the fact that $m \mapsto (m/n) \log(en/m)$ is non-decreasing for $1 \leq m \leq n$. On the other hand, if $k(\pi) < k$, then it is easy to see that

$$V_\pi(\theta) \geq \min_{2 \leq i \leq k} (\theta_{0,i} - \theta_{0,i-1})$$

which implies, by (29), that

$$V_\pi^2(\theta) \geq \frac{k\sigma^2}{n} \log \frac{en}{k}.$$

Thus

$$V_\pi^2(\theta) + \frac{4\sigma^2 k(\pi)}{n} \log \frac{en}{k(\pi)} \geq V_\pi^2(\theta) \geq \frac{k\sigma^2}{n} \log \frac{en}{k} \quad (31)$$

for every $\pi \in \Pi$ with $k(\pi) < k$. The inequalities (30) and (31) therefore imply that

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

and this completes the proof. $\square$

4 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 non-asymptotic sense. “Local” here refers to a ball $\{t : \ell_\infty^2(t, \theta) \leq cR(n; \theta)\}$ 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.

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

Lemma 4.1. *The following inequality holds for every $\theta \in \mathcal{M}$ and $c > 0$*

$$\sup_{t \in \mathcal{M} : \ell_\infty^2(t, \theta) \leq cR(n; \theta)} \mathbb{E}_t \ell^2(t, \hat{\theta}) \leq 2(1 + 4c)R(n; \theta). \quad (32)$$

Proof. Inequality (7) gives $\mathbb{E}_t \ell^2(t, \hat{\theta}) \leq R(n; t)$ for every $t \in \mathcal{M}$. Fix $\pi \in \Pi$. By the triangle inequality, we get

$$V_\pi(t) \leq 2\ell_\infty(t, \theta) + V_\pi(\theta).$$

As a result, whenever $\ell_\infty(t, \theta) \leq cR(n; \theta)$, we obtain

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

As a consequence,

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

This proves (32). □

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 (32) 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 (10) and (11). 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 4.1 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 non-asymptotically minimax for such sequences θ . We use the qualifier “approximately” here because of the presence of logarithmic factors on the right hand sides of (10) and (11).

We make here the assumption that the errors $\epsilon_1, \dots, \epsilon_n$ 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 [Van der Vaart \(1998, pp. 347\)](#).

Lemma 4.2 (Assouad). *Let m be a positive integer and suppose that, for each $\tau \in \{0, 1\}^m$, there is an associated non-decreasing sequence θ^τ in $N(\theta)$, where $N(\theta)$ is a neighborhood of θ . Then the following inequality holds:*

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

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

The two inequalities (10) and (11) are proved in the next two subsections.

4.1 Uniform increments

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

$$\frac{c_1 V(\theta)}{n} \leq \theta_i - \theta_{i-1} \leq \frac{c_2 V(\theta)}{n} \quad \text{for } i = 2, \dots, n \quad (33)$$

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 (33) means that the increments of θ are in a sense uniform. An important example in which (33) 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 (33).

Theorem 4.3. Suppose θ satisfies (33) 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 $\mathfrak{R}_n(\theta) := \inf_t \sup_{t \in \mathfrak{N}(\theta)} \mathbb{E}_t \ell^2(t, \hat{t})$ satisfies the following inequality

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 3^{2/3}}{256 c_2^{4/3}} \left(\frac{\sigma^2 V(\theta)}{n} \right)^{2/3} \geq \frac{c_1^2 3^{2/3}}{4096 c_2^{4/3} \log(4n)} \frac{R(n; \theta)}{12} \quad (34)$$

provided

$$n \geq \max \left(2, \frac{24\sigma^2}{V^2(\theta)}, \frac{2c_2 V(\theta)}{\sigma} \right). \quad (35)$$

Proof. Let $V = V(\theta)$. Fix an integer $1 \leq k \leq n$ and let $m := \lfloor n/k \rfloor$ where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x . For each $\tau = (\tau_1, \dots, \tau_m) \in \{0, 1\}^m$, define $\theta^\tau \in \mathcal{M}$ by

$$\theta_j^\tau = \theta_j - \sum_{i=1}^m \tau_i (\theta_j - \theta_{(i-1)k+1}) I \{ (i-1)k + 1 \leq j \leq ik \}$$

for $1 \leq j \leq n$. In other words, when j lies in the interval $[(i-1)k + 1, ik]$, the value θ_j^τ equals θ_j if $\tau_i = 0$ and $\theta_{(i-1)k+1}$ if $\tau_i = 1$. The monotonicity of $\{\theta_i^\tau\}$ therefore follows easily from the monotonicity of $\{\theta_i\}$.

We apply Assouad's lemma to these functions θ^τ . Clearly for every $\tau, \tau' \in \{0, 1\}^m$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) = \frac{1}{n} \sum_{i=1}^m I \{ \tau_i \neq \tau'_i \} \sum_{j=(i-1)k+1}^{ik} (\theta_j - \theta_{(i-1)k+1})^2.$$

For a fixed $1 \leq i \leq m$ and $(i-1)k + 1 \leq j \leq ik$, as

$$\theta_j - \theta_{(i-1)k+1} = \sum_{l=(i-1)k+2}^j (\theta_l - \theta_{l-1}),$$

assumption (33) gives

$$\theta_j - \theta_{(i-1)k+1} \geq \frac{c_1 V}{n} (j - (i-1)k - 1).$$

Therefore

$$\sum_{j=(i-1)k+1}^{ik} (\theta_j - \theta_{(i-1)k+1})^2 \geq \frac{c_1^2 V^2}{n^2} \sum_{i=1}^{k-1} i^2 = \frac{c_1^2 V^2}{6n^2} k(k-1)(2k-1)$$

for every $1 \leq i \leq m$ and $(i-1)k+1 \leq j \leq ik$. Consequently, for every $\tau, \tau' \in \{0, 1\}^m$, we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) \geq \frac{c_1^2 V^2}{6n^3} k(k-1)(2k-1) \Upsilon(\tau, \tau'). \quad (36)$$

Similarly, by using the upper bound in (33), we obtain

$$\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{c_2^2 V^2}{6n^3} k(k-1)(2k-1) \Upsilon(\tau, \tau'). \quad (37)$$

Also,

$$\begin{aligned} \max_{1 \leq j \leq n} |\theta_j^\tau - \theta_j| &= \max_{1 \leq i \leq m} \max_{(i-1)k+1 \leq j \leq ik} |\theta_j - \theta_{(i-1)k+1}| \\ &\leq \max_{1 \leq i \leq m} \max_{(i-1)k+1 \leq j \leq ik} \frac{c_2 V}{n} (j - (i-1)k - 1). \end{aligned}$$

Consequently,

$$\ell_\infty(\theta^\tau, \theta) = \max_{1 \leq j \leq n} |\theta_j^\tau - \theta_j| \leq \frac{c_2 V k}{n}. \quad (38)$$

By Pinsker's inequality,

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}^2 \leq \frac{1}{2} D(\mathbb{P}_{\theta^\tau} \|\mathbb{P}_{\theta^{\tau'}}) = \frac{n}{4\sigma^2} \ell^2(\theta^\tau, \theta^{\tau'})$$

and when $\Upsilon(\tau, \tau') = 1$, we have from (37) that

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}^2 \leq \frac{c_2^2 V^2}{24\sigma^2 n^2} k(k-1)(2k-1) \leq \frac{c_2^2 V^2 k^3}{12\sigma^2 n^2}. \quad (39)$$

Thus, by Lemma 4.2, (36) and (39), we get

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 V^2 m}{48n^3} k(k-1)(2k-1) \left[1 - \frac{k^{3/2} V c_2}{\sigma n \sqrt{12}} \right] \quad (40)$$

provided

$$\theta^\tau \in \mathfrak{N}(\theta) \quad \text{for every } \tau \in \{0, 1\}^m. \quad (41)$$

We now make the choice

$$k = \left(\frac{\sqrt{3} \sigma n}{c_2 V} \right)^{2/3}.$$

It then follows from (38) that

$$\ell_\infty^2(\theta^\tau, \theta) \leq \frac{c_2^2 V^2 k^2}{n^2} = (3c_2)^{2/3} \left(\frac{\sigma^2 V}{n} \right)^{2/3}.$$

We now use Theorem 3.2 to show (41). Note first that (35) is a stronger condition than (27) because $c_2 \geq 1$. We can thus use the first inequality in (26) to get

$$\ell_\infty^2(\theta^\tau, \theta) \leq \left(\frac{3c_2}{c_1} \right)^{2/3} \frac{R(n; \theta)}{12}$$

which proves (41).

Also, with our choice of k , it is easy to check that $2 \leq k \leq n/2$ because (35) implies that

$$n \geq \max \left(\frac{2\sqrt{2}c_2V}{\sqrt{3}\sigma}, \frac{24\sigma^2}{c_2^2V^2} \right).$$

Inequality (40) with our choice of k gives (note that $k(k-1)(2k-1) \geq 3k^3/4$) that

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 V^2 m k^3}{128 n^3} = \frac{3c_1^2 \sigma^2 m}{128 c_2^2 n}.$$

Because $k \leq n/2$, we have $m = \lfloor n/k \rfloor \geq n/(2k)$. Thus

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^2 V^2 m k^3}{128 n^3} \geq \frac{c_1^2 V^2 k^2}{256 n^2} = \frac{c_1^2 3^{2/3}}{256 c_2^{4/3}} \left(\frac{\sigma^2 V}{n} \right)^{2/3}.$$

This proves the first inequality in (34). The second inequality in (34) follows from Theorem 3.1. The proof is complete. $\square$

4.2 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 (33). Specifically, we assume that $k(\theta) = k$ and that the k values of θ are sufficiently well-separated and prove inequality (11).

Let $k = k(\theta)$. There exist integers $n_1, \dots, n_k$ with $n_i \geq 1$ and $n_1 + \dots + n_k = n$ 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 $s_i := n_1 + \dots + n_i$. 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 $\theta_{0,1} < \dots < \theta_{0,k}$.

Theorem 4.4. *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*

$$\min_{2 \leq i \leq k} (\theta_{0,i} - \theta_{0,i-1}) \geq \sqrt{\frac{k\sigma^2}{n} \log \frac{en}{k}}. \quad (42)$$

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

$$\mathfrak{R}_n(\theta) \geq \frac{c_1^{7/3}}{2^{31/3} c_2^2} R(n; \theta) \left(\log \frac{en}{k} \right)^{-2/3} \quad (43)$$

provided

$$\frac{n}{k} \geq \max \left(\left(\frac{4}{c_1^2} \log \frac{en}{k} \right)^{1/3}, \exp \left(\frac{1 - 4c_1}{4c_1} \right) \right). \quad (44)$$

Proof. For notational convenience, we write

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

First note that under assumption (42), Theorem 3.3 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 $\{u : s_{i-1} + 1 \leq u \leq s_i\}$ where it takes the value $\theta_{0,i}$, it follows that $\ell_\infty(\theta^\tau, \theta) \leq \beta_n$. This implies that $\theta^\tau \in \mathfrak{N}(\theta)$ for every τ as $\beta_n^2 \leq R(n; \theta)$.

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

$$\ell^2(\theta^\tau, \theta^{\tau'}) = \frac{1}{n} \sum_{i=1}^k \sum_{v=1}^{m_i} \frac{l\beta_n^2}{m_i^2} I\{\tau_{iv} \neq \tau'_{iv}\} = \frac{l\beta_n^2}{n} \sum_{i=1}^k \frac{\Upsilon(\tau_i, \tau'_i)}{m_i^2}. \quad (45)$$

Because

$$m_i \leq \frac{n_i}{l} \leq \frac{c_2 n}{kl} \quad \text{for each } 1 \leq i \leq k$$

we have

$$\ell^2(\theta^\tau, \theta^{\tau'}) \geq \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3} \sum_{i=1}^k \Upsilon(\tau_i, \tau'_i) = \frac{k^2 l^3 \beta_n^2}{c_2^2 n^3} \Upsilon(\tau, \tau'). \quad (46)$$

Also, from (45), we get

$$\ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{l\beta_n^2}{n(\min_{1 \leq i \leq k} m_i^2)} \quad \text{when } \Upsilon(\tau, \tau') = 1. \quad (47)$$

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

$$\min_{1 \leq i \leq k} m_i \geq \frac{c_1 n}{2kl}. \quad (48)$$

Combining the above inequality with (47), we deduce

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

This and Pinsker's inequality give

$$\|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}^2 \leq \frac{1}{2} D(\mathbb{P}_{\theta^\tau} \|\mathbb{P}_{\theta^{\tau'}}) = \frac{n}{4\sigma^2} \ell^2(\theta^\tau, \theta^{\tau'}) \leq \frac{k^2 l^3 \beta_n^2}{c_1^2 n^2 \sigma^2} \quad (49)$$

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

The inequalities (46) and (49) in conjunction with Assouad's lemma give

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

Because of (48), we get $M = \sum_i m_i \geq k \min_i m_i \geq c_1 n/(2l)$ and thus

$$\mathfrak{R}_n(\theta) \geq \frac{c_1 k^2 l^2 \beta_n^2}{16c_2^2 n^2} \left(1 - \frac{k\beta_n l^{3/2}}{c_1 n \sigma}\right). \quad (50)$$

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

$$l = \left(\frac{c_1 n \sigma}{2k\beta_n}\right)^{2/3}. \quad (51)$$

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

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

This gives rise to two lower bounds for n which are collected in (44).

As a consequence of (51), we get that $l^{3/2} \leq c_1 n \sigma / (2k\beta_n)$, which ensures that the term inside the parantheses on the right hand side of (50) is atleast $1/2$. This gives

$$\mathfrak{R}_n(\theta) \geq \frac{c_1 k^2 l^2 \beta_n^2}{32c_2^2 n^2} \geq \frac{c_1^{7/3}}{2^{19/3} c_2^2} \frac{k\sigma^2}{n} \left(\log \frac{en}{k}\right)^{1/3}. \quad (52)$$

To complete the proof, we use Theorem 3.3. Specifically, the second inequality in (28) gives

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

The proof is complete by combining the above inequality with (52). $\square$

5 Risk bound under model misspecification

We consider model (3) where now the true sequence θ is not necessarily assumed to be in $\mathcal{M}$. We study the behavior of the LSE $\hat{\theta}$ defined exactly as in (4). The goal of this section is to prove an inequality analogous to (7) for model misspecification. It turns out here that the LSE is really estimating the non-decreasing 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 Robertson et al. (1988, Chapter 1), it follows that

$$\tilde{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{\theta}_{k,l} \quad \text{for } 1 \leq j \leq n, \quad (53)$$

where $\bar{\theta}_{k,l}$ is as defined in (12).

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 $s_i = n_1 + \dots + n_i$ for $1 \leq i \leq k$. It is easy to check that $S_\pi(t) \leq V_\pi(t)$ for every $t \in \mathcal{M}$. The following is the main result of this section.

Theorem 5.1. *For every $\theta \in \mathbb{R}^n$, the LSE satisfies*

$$\mathbb{E}_\theta \ell^2(\tilde{\theta}, \hat{\theta}) \leq 4 \inf_{\pi \in \Pi} \left(S_\pi^2(\tilde{\theta}) + \frac{4\sigma k(\pi)}{n} \log \frac{en}{k(\pi)} \right) \leq R(n; \tilde{\theta}). \quad (54)$$

Proof. We only need to prove the first inequality in (54). The second inequality follows from the fact that $S_\pi(\tilde{\theta}) \leq V_\pi(\tilde{\theta})$.

Fix $1 \leq j \leq n$ and $0 \leq m \leq n - j$. Recall the notation in (12). Let

$$l_j := \max \left\{ l : j + m \leq l \leq n \text{ and } \max_{k \leq j+m} \bar{\theta}_{k,l} = \tilde{\theta}_{j+m} \right\}.$$

Then l_j is well defined by the definition of $\tilde{\theta}$ in (53). We now write

$$\hat{\theta}_j = \min_{l \geq j} \max_{k \leq j} \bar{Y}_{k,l} \leq \max_{k \leq j} \bar{Y}_{k,l_j} = \max_{k \leq j} \bar{\theta}_{k,l_j} + \max_{k \leq j} \bar{\epsilon}_{k,l_j}.$$

As a result,

$$\hat{\theta}_j \leq \max_{k \leq j+m} \bar{\theta}_{k,l_j} + \max_{k \leq j} \bar{\epsilon}_{k,l_j} = \tilde{\theta}_{j+m} + \max_{k \leq j} \bar{\epsilon}_{k,l_j}.$$

This implies that

$$\left(\hat{\theta}_j - \tilde{\theta}_j \right)_+ \leq (\tilde{\theta}_{j+m} - \tilde{\theta}_j) + \max_{k \leq j} (\bar{\epsilon}_{k,l_j})_+.$$

Now arguing as in the proof of Theorem 2.1 and noting that $l_j - j \geq m$, we get

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \tilde{\theta}_j \right)_+^2 \leq 2 \left(\tilde{\theta}_{j+m} - \tilde{\theta}_j \right)^2 + \frac{8\sigma^2}{m+1}.$$

Also we have the corresponding inequality for the negative part:

$$\mathbb{E}_\theta \left(\hat{\theta}_j - \tilde{\theta}_j \right)_-^2 \leq 2 \left(\tilde{\theta}_{j+m} - \tilde{\theta}_j \right)^2 + \frac{8\sigma^2}{m+1}.$$

Now fix an interval partition π . For each $1 \leq j \leq n$, we define two integers $m_1(j)$ and $m_2(j)$ in the following way. For $s_{i-1} + 1 \leq j \leq s_i$, let $m_1(j) = s_i - j$ and $m_2(j) = j - 1 - s_{i-1}$. This results in the risk bound

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

where

$$A_j := 2 \left(\tilde{\theta}_{j+m_1(j)} - \tilde{\theta}_j \right)^2 + \frac{8\sigma^2}{m_1(j) + 1}. \quad (55)$$

and

$$B_j := 2 \left(\tilde{\theta}_j - \tilde{\theta}_{j-m_2(j)} \right)^2 + \frac{8\sigma^2}{m_2(j) + 1}.$$

We shall now prove that

$$\frac{1}{n} \sum_{j=1}^n A_j \leq 2S_\pi^2(\tilde{\theta}) + \frac{8k\sigma^2}{n} \log \frac{en}{k} \quad (56)$$

and

$$\frac{1}{n} \sum_{j=1}^n B_j \leq 2S_\pi^2(\tilde{\theta}) + \frac{8k\sigma^2}{n} \log \frac{en}{k}. \quad (57)$$

We give the argument for (56) below. The argument for (57) follows similarly. From (55), $\sum_{j=1}^n A_j/n$ can be broken into two terms. For the first term note that $j + m_1(j) = s_i$ for $s_{i-1} + 1 \leq j \leq s_i$, and therefore

$$2 \sum_{j=1}^n \left(\tilde{\theta}_{j+m_1(j)} - \tilde{\theta}_j \right)^2 = 2 \sum_{i=1}^k \sum_{j=s_{i-1}+1}^{s_i} \left(\tilde{\theta}_{s_i} - \tilde{\theta}_j \right)^2 = 2S_\pi^2(\tilde{\theta}).$$

For the second term, arguing exactly as in (20) we get the upper bound $(8k\sigma^2/n) \log(en/k)$. This proves (56) which completes the proof. $\square$

By taking π in the infimum in the upper bound of (54) to be the interval partition generated by $\tilde{\theta}$, we obtain the following result which is the analogue of (9) for model misspecification.

Corollary 5.2. *For every arbitrary sequence θ of length n (not necessarily non-decreasing),*

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

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 non-increasing, i.e., $\theta_1 \geq \theta_2 \geq \dots \geq \theta_n$. This implies that the LSE converges to $\tilde{\theta}$ at the rate $\sigma^2 \log(en)/n$ when θ is non-increasing.

Lemma 5.3. *$k(\tilde{\theta}) = 1$ if θ is non-increasing.*

Proof. By (53), $\tilde{\theta}_1 = \min_{l \geq 1} \bar{\theta}_{1,l}$. Because θ is non-increasing, we get therefore that $\tilde{\theta}_1 = \bar{\theta}_{1,n}$. Similarly $\tilde{\theta}_n = \max_{k \leq n} \bar{\theta}_{k,n} = \bar{\theta}_{1,n}$. Since $\tilde{\theta}$ has to be non-decreasing, it follows that $\tilde{\theta}_j = \bar{\theta}_{1,n}$ for all j . The proof is complete. $\square$

To state our next result, let

$$b(t) := \sum_{i=1}^{n-1} I\{t_i \neq t_{i+1}\} + 1 \quad \text{for } t \in \mathbb{R}^n. \quad (58)$$

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

Lemma 5.4. *For any sequence $\theta \in \mathbb{R}^n$ we have $k(\tilde{\theta}) \leq b(\theta)$.*

Proof. As $\tilde{\theta}$ is always non-decreasing it is enough to show that $b(\tilde{\theta}) \leq b(\theta)$. We will show that $I(\tilde{\theta}_i \neq \tilde{\theta}_{i+1}) \leq I(\theta_i \neq \theta_{i+1})$ for every $i \in \{1, \dots, n-1\}$ using the method of contradiction. Fix $i \in \{1, \dots, n-1\}$ and assume that $\theta_i = \theta_{i+1}$ and $\tilde{\theta}_i \neq \tilde{\theta}_{i+1}$. Let $\theta_i = \theta_{i+1} = c$. Define another sequence $t \in \mathbb{R}^n$ such that for $j = 1, \dots, n$,

$$t_j = \left(\frac{\tilde{\theta}_i + \tilde{\theta}_{i+1}}{2} \right) \cdot I\{i \leq j \leq i+1\} + \tilde{\theta}_j \cdot I\{i \leq j \leq i+1\}^c.$$

We note that t is non-decreasing as $\tilde{\theta}$ is non-decreasing. Now,

$$\begin{aligned} n[\ell^2(\theta, \tilde{\theta}) - \ell^2(\theta, t)] &= \sum_{j=i}^{i+1} [(\theta_j - \tilde{\theta}_j)^2 - (\theta_j - t_j)^2] \\ &= \sum_{j=i}^{i+1} [(\tilde{\theta}_j - c)^2 - (t_j - c)^2] = \sum_{j=i}^{i+1} (\tilde{\theta}_j - t_j)^2 > 0. \end{aligned}$$

But this is a contradiction as $\tilde{\theta}$ is the non-decreasing projection of θ . $\square$

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 $(16b(\theta)\sigma^2/n) \log(en/b(\theta))$.

6 Some auxiliary results

Lemma 6.1. *For every $\theta \in \mathcal{M}$ and $\delta > 0$, there exists an interval partition π with*

$$V_\pi(\theta) \leq \delta \text{ and } k(\pi) \leq \left\lceil \frac{V(\theta)}{\delta} \right\rceil \quad (59)$$

where $\lceil x \rceil$ denotes the smallest integer that is larger than or equal to x .

Proof. Fix $\theta \in \mathcal{M}$ and $\delta > 0$. Write V for $V(\theta)$. Let $s_0 := 0$ and we recursively define s_i by

$$s_i := \sup \{s_{i-1} + 1 \leq s \leq n : \theta_s - \theta_{s_{i-1}+1} \leq \delta\}$$

while $s_{i-1} < n$. This construction will result in integers $s_0 = 0 < s_1 < \dots < s_m = n$ having the property that $\theta_{s_i+1} - \theta_{s_{i-1}+1} > \delta$ for every $i = 1, \dots, m$. As a result, we obtain

$$V \geq (\theta_{s_{m-1}+1} - \theta_1) > (m-1)\delta$$

which is the same as $m \leq \lceil V/\delta \rceil$. Let $\pi = (n_1, \dots, n_m)$ where $n_i = s_i - s_{i-1}$. Clearly $k(\pi) = m \leq \lceil V/\delta \rceil$. The definition of s_i ensures that $\theta_{s_i} - \theta_{s_{i-1}+1} \leq \delta$ for every i which implies that $V_\pi(\theta) \leq \delta$. The proof is complete. $\square$

Lemma 6.2. *The following inequality holds for every finite sequence $a_1, \dots, a_k$ of real numbers:*

$$\sum_{j=1}^k (a_j - \bar{a}_{j,k})^2 \leq \sum_{j=1}^k (a_j - \bar{a})^2 \quad (60)$$

where $\bar{a}_{j,k}$ denotes the mean of $a_j, \dots, a_k$ and $\bar{a}$ is the mean of the entire sequence $a_1, \dots, a_k$.

Proof. We shall prove this by induction on k . Note that the inequality is trivial for $k = 1$ and $k = 2$. Suppose we have proved it for all $k \leq m$ and for all sequences $a_1, \dots, a_k$. We now consider $k = m+1$ and a finite sequence $a_1, \dots, a_{m+1}$ and prove that

$$\sum_{j=1}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \sum_{j=1}^{m+1} (a_j - \bar{a})^2. \quad (61)$$

Note that the first term (for $j = 1$) on both sides of (61) equal each other because $\bar{a}_{1,m+1} = \bar{a}$ and can therefore be cancelled. After cancellation, note

that the right hand side depends on a_1 while the left hand side does not. As a result, (61) holds for all sequences $a_1, \dots, a_{m+1}$ if and only if

$$\sum_{j=2}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \inf_{a_1 \in \mathbb{R}} \sum_{j=2}^{m+1} (a_j - \bar{a})^2.$$

Clearly the right hand side above is minimized when we choose a_1 so that $\bar{a} = \bar{a}_{2,m+1}$. Thus (61) holds if and only if

$$\sum_{j=2}^{m+1} (a_j - \bar{a}_{j,m+1})^2 \leq \sum_{j=2}^{m+1} (a_j - \bar{a}_{2,m+1})^2.$$

But this is precisely the inequality (60) for $m = k$ and the sequence $a_2, \dots, a_{m+1}$ which is true by the induction hypothesis. This completes the proof. $\square$

References

- 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.
- Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalisation. *Probability Theory and Related Fields*, 113:301–413.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150.
- Brunk, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*, pages 177–197. Cambridge Univ. Press, London.
- Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.*, 27(3):557–566.
- Cator, E. (2011). Adaptivity and optimality of the monotone least-squares estimator. *Bernoulli*, 17:714–735.
- Donoho, D. (1991). Gelfand n-widths and the method of least squares. Technical report, University of California, Berkeley. Department of Statistics, Technical report.
- Durot, C. (2002). Sharp asymptotics for isotonic regression. *Probability Theory and Related Fields*, 122:222–240.
- Durot, C. (2007). On the l_p error of monotonicity constrained estimators. *Annals of Statistics*, 35:1080–1104.

- Durrett, R. (2010). *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition.
- Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).
- Groeneboom, P. and Pyke, R. (1983). Asymptotic normality of statistics based on the convex minorants of empirical distribution functions. *Ann. Probab.*, 11(2):328–345.
- Guntuboyina, A. and Sen, B. (2013). Global risk bounds and adaptation in univariate convex regression. *Under review*, available at <http://arxiv.org/abs/1305.1648>.
- Jankowski, H. K. and Wellner, J. A. (2012). Convergence of linear functionals of the grenander estimator under misspecification. available at <http://arxiv.org/abs/1207.6614>.
- Meyer, M. and Woodroffe, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.*, 28(4):1083–1104.
- Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.
- Van de Geer, S. (1990). Estimating a regression function. *Annals of Statistics*, 18:907–924.
- Van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44.
- Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.
- Wang, Y. (1996). The l_2 risk of an isotonic estimate. *Comm. Statist. Theory Methods*, 25:281–294.
- Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555.