Global risk bounds and adaptation in univariate convex regression

Adityanand Guntuboyina and Bodhisattva Sen*
University of California, Berkeley, and Columbia University, New York

May 9, 2013

Abstract

We consider the problem of nonparametric estimation of a convex regression function \( \phi_0 \). We study global risk bounds and adaptation properties of the least squares estimator (LSE) of \( \phi_0 \). Under the natural squared error loss, we show that the risk of the LSE is bounded from above by \( n^{-4/5} \) up to a multiplicative factor that is logarithmic in \( n \). When \( \phi_0 \) is convex and piecewise affine with \( k \) knots, we establish adaptation of the LSE by showing that its risk is bounded from above by \( k^{5/4}/n \) up to logarithmic multiplicative factors. On the other hand, when \( \phi_0 \) has curvature, we show that no estimator can have risk smaller than a constant multiple of \( n^{-4/5} \) in a very strong sense by proving a “local” minimax lower bound. We also study the case of model misspecification where we show that the LSE exhibits the same global behavior provided the loss is measured from the closest convex projection of the true regression function. In addition to the convex LSE, we also provide risk bounds for a natural sieved LSE. In the process of proving our results, we establish some new results on the covering numbers of classes of convex functions which are of independent interest.

1 Introduction

We consider the problem of estimating an unknown convex function \( \phi_0 \) on \([0, 1]\) from observations \((x_1, Y_1), \ldots, (x_n, Y_n)\) drawn according to the model

\[
Y_i = \phi_0(x_i) + \xi_i, \quad \text{for } i = 1, \ldots, n. \tag{1}
\]

*Supported by NSF Grants DMS-1150435 and AST-1107373
where \( x_1, \ldots, x_n \) are fixed points in \([0, 1]\) and \( \xi_1, \ldots, \xi_n \) represent independent mean zero errors. Convex regression is an important problem in the general area of non-parametric estimation under shape constraints. It often arises in applications: typical examples appear in economics (indirect utility, production or cost functions), medicine (dose response experiments) or biology (growth curves).

The most natural and commonly used estimator for \( \phi_0 \) is the full least squares estimator (LSE), \( \hat{\phi}_{ls} \), which is defined as any minimizer of the least squares criterion, i.e.,

\[
\hat{\phi}_{ls} \in \arg\min_{\psi \in \mathcal{C}} \sum_{i=1}^{n} (Y_i - \psi(x_i))^2,
\]

where \( \mathcal{C} \) denotes the set of all real-valued convex functions on \([0, 1]\). \( \hat{\phi}_{ls} \) is not unique even though its values at the data points \( x_1, \ldots, x_n \) are unique. This follows from that fact that \((\hat{\phi}_{ls}(x_1), \ldots, \hat{\phi}_{ls}(x_n)) \in \mathbb{R}^n\) is the projection of \((Y_1, \ldots, Y_n)\) on a closed convex cone. A simple linear interpolation of these values leads to a unique continuous and piecewise linear convex function with possible knots at the data points, which can be treated as the canonical LSE. The canonical LSE can be easily computed by a quadratic programming procedure.

Unlike other methods for function estimation such as those based on kernels which depend on tuning parameters such as smoothing bandwidths, the LSE has the obvious advantage of being completely automated. It was first proposed by Hildreth (1954) for the estimation of production functions and Engel curves. Algorithms for its computation can be found in Dykstra (1983) and Fraser and Massam (1989). The theoretical behavior of the LSE has been investigated by many authors. Its consistency in the supremum norm on compact sets in the interior of the support of the covariate was proved by Hanson and Pledger (1976). Mammen (1991) derived the rate of convergence of the LSE and its derivative at a fixed point, while Groeneboom et al. (2001) proved consistency and derived its asymptotic distribution at a fixed point of positive curvature. Dümbgen et al. (2004) showed that the supremum distance between the LSE and \( \phi_0 \), assuming twice differentiability, on a compact interval in the interior of the support of the design points is of the order \((\log(n)/n)^{2/5}\).

In spite of all the above mentioned work, surprisingly, not much is known about the global risk behavior of the LSE. This is the main focus of this paper. We consider the risk of the LSE under the natural global loss function:

\[
\ell^2(\phi, \psi) := \frac{1}{n} \sum_{i=1}^{n} (\phi(x_i) - \psi(x_i))^2. \tag{2}
\]

We assume, throughout the paper, that, in (1), \( x_1 < x_2 < \cdots < x_n \) are fixed design points in \([0, 1]\) satisfying

\[
c_1 \leq n(x_i - x_{i-1}) \leq c_2, \quad \text{for } i = 2, 3, \ldots, n, \tag{3}
\]

where \( c_1 \) and \( c_2 \) are positive constants, and that \( \xi_1, \ldots, \xi_n \) are independent normally
distributed random variables with mean zero and variance $\sigma^2 > 0$. Our contributions in this paper can be summarized in the following.

1. We establish, for the first time, finite sample risk bounds for the LSE $\hat{\phi}_{ls}$ under the loss $\ell^2$. Specifically, we show that the risk $\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0)$ is bounded from above by $n^{-4/5}$ upto multiplicative factors that are logarithmic in the sample size $n$.

2. We investigate the optimality of the rate $n^{-4/5}$. We show that for convex functions $\phi_0$ having a bounded (from both above and below) curvature on a sub-interval of $[0, 1]$, the rate $n^{-4/5}$ cannot be improved (in a very strong sense) by any other estimator. Specifically we show that a certain “local” minimax risk (see Section 3 for the details), under the loss $\ell^2$, is bounded from below by $n^{-4/5}$. This shows, in particular, that the same holds for the global minimax rate for this problem.

3. An important parametric subclass of $C$ is the class $P_k$ of all piecewise affine convex functions that can be written as the maximum of at most $k$ ($k \geq 1$) affine functions. Indeed, any function in $C$ can be approximated, upto any desired precision, by a member of $P_k$, as $k$ grows. Thus, $P_k$’s offer an attractive alternative to going completely nonparametric when estimating $\phi_0$. Estimation of any $\phi_0 \in P_k$ is a “parametric problem” and a natural estimator is the sieved LSE, $\tilde{\phi}_k$, formally defined by

$$\tilde{\phi}_k \in \arg\min_{\psi \in P_k} \sum_{i=1}^n (Y_i - \psi(x_i))^2.$$  \hspace{1cm} (4)

When the true function $\phi_0$ belongs to $P_k$, we show that the estimator $\tilde{\phi}_k$ converges at the “parametric” rate $k/n$ upto logarithmic factors.

4. We show that the LSE $\hat{\phi}_{ls}$ has remarkable adaption properties. Specifically, when $\phi_0$ belongs to the class $P_k$, we show that $\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0)$ is bounded from above by $k^{5/4}/n$ upto logarithmic factors in $n$. This is extraordinary since the LSE uses no knowledge that $\phi_0 \in P_k$. This result compared with the previous one shows that the LSE only pays a price of at most $k^{1/4}/n$ for the automatic adaptation it enjoys over the classes $P_k$. This is indeed noteworthy, as the estimators $\tilde{\phi}_k$, for $1 < k < n$, cannot be efficiently computed due to the combinatorial nature of the constraint $\psi \in P_k$. Note that it is easy to show that $\tilde{\phi}_k$ coincides with the LSE $\hat{\phi}_{ls}$ when $k \geq n$.

5. We also provide risk bounds in the case of model misspecification where we do not assume that the underlying regression function in (1) is convex. In this case we prove the exact same upper bounds for $\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0)$ where $\phi_0$ now denotes any convex projection (defined in Section 6) of the unknown true regression function. To the best of our knowledge, this is the first result on global risk bounds for the estimation of convex regression functions under model misspecification.
Two special features of our analysis are that: (1) all our risk-bounds are non-asymptotic, and (2) none of our results uses any (explicit) characterization of the LSE (except that it minimizes the least squares criterion) as a result of which our approach can, in principle, be extended to more complex Empirical Risk Minimization (ERM) procedures, including shape restricted function estimation in higher dimensions; see e.g., Seijo and Sen (2011), Seregin and Wellner (2010) and Cule et al. (2010).

The analysis of the risk behavior of \( \hat{\phi}_{ls} \) is complicated mainly due to two facts: (1) \( \hat{\phi}_{ls} \) does not have a closed form expression, and (2) the class \( C \) (over which \( \hat{\phi}_{ls} \) minimizes the least squares criterion) is not totally bounded while in the usual assessment of the performance of least squares or ERM procedures in function estimation problems, it is often assumed that the underlying function class consists of functions that are uniformly bounded. One of the standard approaches to finding risk bounds for ERM procedures, including ours, employs a result due to Van de Geer (2000, pp. 184-185) which reduces the problem to bounding the covering numbers of local balls of the function class around the true function. In our case, these local balls of convex functions are, however, much larger than all other classes of convex functions for which covering numbers are known so far (see Bronshtein (1976), Dryanov (2009) and Guntuboyina and Sen (2013) for existing results on covering numbers for convex functions). This, therefore, necessitated the need for new results on the covering numbers of univariate convex functions, which are also of independent interest. We present these results in Section 7.

Our adaptation behavior of the LSE implies in particular that the LSE converges at different rates depending on the true convex function \( \phi_0 \). We believe that such adaptation is rather unique to problems of shape restricted function estimation and is currently not very well understood. For example, in the related problem of monotone function estimation, which has an enormous literature (see e.g., Grenander (1956), Birgé (1989), Zhang (2002) and the references therein), the only result on adaptive global behavior of the LSE is found in Groeneboom and Pyke (1983); also see van de Geer (1993). This result, however, holds only in an asymptotic sense and only when the true model is a constant function. Results on the pointwise adaptive behavior of the LSE in monotone function estimation are more prevalent and can be found, for example, in Carolan and Dykstra (1999) and Jankowski and Wellner (2012). For convex function estimation, as far as we are aware, adaptation behavior of the LSE has not been studied before. Adaptation behavior for the estimation of a convex function at a single point has been recently studied by Cai and Low (2011) but they focus on different estimators that are based on local averaging techniques.

The rest of the paper is organized as follows. In Section 2 we show that the risk \( \mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0) \) is bounded from above by \( n^{-4/5} \) upto logarithmic factors in \( n \). In Section 3, we show that for convex functions with curvature, the rate \( n^{-4/5} \) cannot be improved by any arbitrary estimator. We establish this by proving a lower bound for the local minimax risk for the estimation of \( \phi_0 \). The adaptation properties of the LSE are studied in Section 4 where it is shown that the LSE is rate adaptive in the sense that if \( \phi_0 \) is assumed to lie in \( P_k \), then the risk of \( \hat{\phi}_{ls} \) is bounded from above by \( k^{5/4}/n \).
(upto logarithmic factors). In Section 5, we study the behavior of \( \tilde{\phi}_k \) when \( \phi_0 \in P_k \). Extensions of our main results for the case of model misspecification are discussed in Section 6. Results on the covering numbers of local balls of convex functions are presented in Section 7. Some auxiliary results about convex functions useful in the proofs of the main results are deferred to Section 8.

2 Worst case risk bound for the LSE

In the following theorem, we show that the risk of the LSE \( \hat{\phi}_{ls} \) under the loss function (2) is bounded from above by \( n^{-4/5} \) upto multiplication by logarithmic factors of \( n \). It turns out that this is the worst possible rate of convergence of the LSE. For certain special convex functions \( \phi_0, \hat{\phi}_{ls} \) converges at much faster rates. We explain this adaptation behavior in Section 4.

Recall that \( \mathcal{C} \) denotes the set of all real-valued convex functions on \([0, 1]\). For \( \phi \in \mathcal{C} \), let \( \mathcal{L}(\phi) \) denote the “distance” of \( \phi \) from affine functions. More precisely,

\[
\mathcal{L}(\phi) := \inf \{ \ell(\phi, \tau) : \tau \text{ is affine on } [0, 1] \}.
\]

Note that \( \mathcal{L}(\phi) = 0 \) when \( \phi \) is affine.

**Theorem 2.1** (Worst case rate). Let \( R := \max(1, \mathcal{L}(\phi_0)) \). There exists a positive constant \( C \) depending only on the ratio \( c_1/c_2 \) such that

\[
E_{\phi_0} \ell^2(\hat{\phi}_{ls}, \phi_0) \leq C \left( \frac{\sigma^2}{n} \right)^{4/5} R^{2/5} \log \frac{n}{2c_1}
\]

whenever

\[
n \geq \max \left( 4c_1, C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2}{R^2} \right).
\]

The proof of the above theorem is based on the following result from Van de Geer (2000, Theorem 9.1) which provides a deviation inequality for the accuracy of least squares estimators in terms of metric entropy properties of the underlying class of functions.

Let \( \mathcal{E} \) denote an arbitrary class of functions on \([0, 1]\) containing the true function \( \phi_0 \) and consider the least squares (or ERM) estimator:

\[
\hat{\phi}(\mathcal{E}) \in \arg\min_{\psi \in \mathcal{E}} \sum_{i=1}^{n} (Y_i - \psi(x_i))^2.
\]

In other words, \( \hat{\phi}(\mathcal{E}) \) is any minimizer of the least squares criterion over the class \( \mathcal{E} \). Observe that \( \hat{\phi}(\mathcal{C}) = \hat{\phi}_{ls} \). For \( r > 0 \), let

\[
S(\phi_0, r, \mathcal{E}) := \{ \psi \in \mathcal{E} : \ell^2(\phi_0, \psi) \leq r^2 \}.
\]
For simplicity of notation, when $E = C$, we denote $S(\phi_0, r, E) = S(\phi_0, r, C)$ by just $S(\phi_0, r)$.

The notion of covering numbers will be used quite often in the sequel. For $\epsilon > 0$ and a subset $S$ of functions, the $\epsilon$-covering number of $S$ under the metric $\ell$, denoted by $M(\epsilon, S, \ell)$, is defined as the smallest number of closed balls of radius $\epsilon$ whose union contains $S$.

The next result is taken from Van de Geer (2000, Theorem 9.1).

**Theorem 2.2** (Van de Geer). Suppose $H$ is a function on $(0, \infty)$ such that

$$H(r) \geq \int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r, E), \ell)} \, d\epsilon$$

for every $r > 0$ and such that $H(r)/r^2$ is decreasing on $(0, \infty)$. Then there exists a universal constant $C$ such that

$$\mathbb{P}_{\phi_0} \left( \ell^2(\hat{\phi}(E), \phi_0) > \delta \right) \leq C \sum_{s \geq 0} \exp \left( -\frac{n2^{2s}\delta}{C^2\sigma^2} \right)$$

for every $\delta > 0$ satisfying $\sqrt{n}\delta \geq C\sigma H(\sqrt{\delta})$.

Our main step in the application of Theorem 2.2 for the proof of Theorem 2.1 is the following result which bounds the entropy integral $\int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r, E), \ell)} \, d\epsilon$.

**Theorem 2.3.** Suppose $n \geq 4c_1$. There exists a constant $K$ depending only on the ratio $c_1/c_2$ such that

$$\int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r, E), \ell)} \, d\epsilon \leq K \left( \log \frac{n}{2c_1} \right)^{5/8} r^{3/4} \left( r^2 + \mathcal{L}^2(\phi_0) \right)^{1/8}$$

for every $\phi_0 \in C$ and $r > 0$.

We prove Theorem 2.3 in Section 7.1 in a series of steps some of which can be viewed as improvements of existing results for covering numbers of convex functions. Below, we provide the proof of Theorem 2.1 using Theorems 2.2 and 2.3.

**Proof of Theorem 2.1.** Let us define

$$\delta_0 := A \left( \frac{\sigma^2}{n} \right)^{4/5} R^{2/5} \log \frac{n}{2c_1}$$

where $A$ is a constant whose value will be specified shortly. Observe that $\delta_0 \leq R^2$ whenever

$$n \geq A^{5/4} \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2}{R^2}.$$
Theorem 2.3 asserts the existence of a positive constant \( K \) depending only on the ratio \( c_1/c_2 \) such that (note that \( \mathcal{L}(\phi_0) \leq R \))

\[
\int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r), \ell)} d\epsilon \leq K \left( \log \frac{n}{2c_1} \right)^{5/8} (r^2 + R^2)^{1/8} r^{3/4}
\]

(6)

for every \( r > 0 \) whenever \( n \geq 4c_1 \).

Suppose now that

\[
n \geq \max \left( 4c_1, A^{5/4} \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2}{R^2} \right)
\]

(7)

so that \( \delta_0 \leq R^2 \) and the inequality (6) holds for every \( r > 0 \). Let \( H(r) \) denote the right hand side of (6). It is clear that \( H(r)/r^2 \) is decreasing on \((0, \infty)\). As a result, a condition of the form \( \sqrt{n} \delta \geq C \sigma H(\sqrt{\delta}) \) for some positive constant \( C \) holds for every \( \delta \geq \delta_0 \) provided it holds for \( \delta = \delta_0 \). Clearly

\[
\frac{H(\sqrt{\delta_0})}{\delta_0} = K \left( \log \frac{n}{2c_1} \right)^{5/8} \delta_0^{-5/8} (\delta_0 + R^2)^{1/8}.
\]

Assuming that (7) holds and noting then that \( \delta_0 \leq R^2 \), we get

\[
\frac{H(\sqrt{\delta_0})}{\delta_0} \leq 2^{1/8} K \left( \log \frac{n}{2c_1} \right)^{5/8} \delta_0^{-5/8} R^{1/4} = 2^{1/8} KA^{-5/8} \sqrt{n}/\sigma.
\]

We shall now use Theorem 2.2 for \( \mathcal{E} = C \). Let \( C \) be the constant given by Theorem 2.2. By the above inequality, the condition \( \sqrt{n} \delta \geq C \sigma H(\sqrt{\delta}) \) holds for each \( \delta \geq \delta_0 \) provided \( A = 2^{1/5}(CK)^{8/5} \). Thus by Theorem 2.2, we obtain

\[
\mathbb{P}_{\phi_0} \left( \ell^2(\hat{\phi}_{ts}, \phi_0) > \delta \right) \leq C \sum_{s \geq 0} \exp \left( -\frac{n2^{4s} \delta}{C^2 \sigma^2} \right)
\]

(8)

for all \( \delta \geq \delta_0 \) whenever \( n \) satisfies (7). Using the expression for \( \delta_0 \) and (7), we get for \( \delta \geq \delta_0 \),

\[
\frac{n \delta}{\sigma^2} \geq \frac{n \delta_0}{\sigma^2} = A \left( \frac{n}{\sigma^2} \right)^{1/5} R^{2/5} \log \frac{n}{2c_1} \geq A^{5/4} \left( \log \frac{n}{2c_1} \right)^{5/4}.
\]

(9)

We thus have

\[
\mathbb{P}_{\phi_0} \ell^2(\hat{\phi}_{ts}, \phi_0) > \delta \right) \leq C_1 \exp \left( -\frac{n \delta}{C_1^2 \sigma^2} \right)
\]

for some constant \( C_1 \) (depending only on \( C \) and \( A = 2^{1/5}(CK)^{8/5} \) provided \( n \) satisfies (7). Integrating both sides of this inequality with respect to \( \delta \) (and using (9) again), we obtain the risk bound

\[
\mathbb{E}_{\phi_0} \ell^2(\hat{\phi}_{ts}, \phi_0) \leq C_2 \delta_0 = C_2 \left( \frac{\sigma^2}{n} \right)^{4/5} AR^{2/5} \log \frac{n}{2c_1}
\]

for some positive constant \( C_2 \) depending only on \( C \) and \( K \). Because \( C \) is an absolute constant and \( K \) only depends on the ratio \( c_1/c_2 \), the proof is complete by an appropriate renaming of the constant \( C \).
3 Non-adaptable convex functions

We show in this section that, for a reasonably large class of convex functions, the rate $n^{-4/5}$ cannot be improved by any estimator. This is the class of convex functions whose curvature is bounded (from both above and below) on a sub-interval of $[0, 1]$. More precisely, for a subinterval $[a, b]$ of $[0, 1]$ and positive real numbers $\kappa_1 < \kappa_2$, we define $\mathcal{K} := \mathcal{K}(a, b, \kappa_1, \kappa_2)$ to be the class of all convex functions $\phi$ on $[0, 1]$ which are twice differentiable on $[a,b]$ and which satisfy $\kappa_1 \leq \phi''(x) \leq \kappa_2$ for all $x \in [a, b]$.

For every function $\phi_0 \in \mathcal{K}$, let us define the local neighborhood $N(\phi_0)$ of $\phi_0$ in $C$ by

$$N(\phi_0) := \left\{ \phi \in C : \sup_{x \in [0,1]} |\phi(x) - \phi_0(x)| \leq \left( \frac{\kappa_2 c_2^2}{32} \right)^{1/5} \left( \frac{\sigma^2}{n} \right)^{2/5} \right\}.$$ 

Recall that the constant $c_1$ is defined in (3). We define the local minimax risk of $\phi_0 \in \mathcal{K}$ to be

$$\mathcal{R}_n(\phi_0) := \inf_{\hat{\phi}} \sup_{\phi \in N(\phi_0)} \mathbb{E}_\phi \ell^2(\phi, \hat{\phi}), \quad (10)$$

the infimum above being over all possible estimators $\hat{\phi}$. $\mathcal{R}_n(\phi_0)$ represents the smallest possible risk under the knowledge that the unknown convex function $\phi$ lies in the local neighborhood $N(\phi_0)$ of $\phi_0$.

In the next theorem, we shall show that the local minimax risk of every function $\phi_0 \in \mathcal{K}$ is bounded from below by a constant multiple of $n^{-4/5}$. Observe that the $L^2$ diameter of $N(\phi_0)$ defined as $\sup_{\phi_1, \phi_2 \in N(\phi_0)} \ell^2(\phi_1, \phi_2)$ is bounded from above by $n^{-4/5}$ upto multiplicative factors that are independent of $n$. Therefore, the supremum risk over $N(\phi_0)$ of any reasonable estimator is bounded from above by $n^{-4/5}$ upto multiplicative factors. The next theorem shows that if $\phi_0 \in \mathcal{K}$, then the supremum risk of every estimator is also bounded from below by $n^{-4/5}$ upto multiplicative factors. Therefore, one cannot estimate $\phi_0$ at a rate faster than $n^{-4/5}$.

**Theorem 3.1** (Lower bound). For every $\phi_0 \in \mathcal{K}(a, b, \kappa_1, \kappa_2)$, we have

$$\mathcal{R}_n(\phi_0) \geq \frac{\kappa_1^2}{4096c_2} \left( \frac{\sqrt{c_1}}{\kappa_2} \right)^{8/5} (b-a) \left( \frac{\sigma^2}{n} \right)^{4/5} \quad (11)$$

provided $n^2 \geq (2c_2)^{5/2} \kappa_2/(\sigma \sqrt{c_1})$.

Prototypical examples of functions in $\mathcal{K}$ include power functions $x^k$ for $k \geq 2$ and the above theorem implies that every estimator has rate at least $n^{-4/5}$ for all these functions. Note that the LSE has the rate $n^{-4/5}$ upto logarithmic factors of $n$ for all functions $\phi_0$. In particular, the LSE is rate optimal (upto logarithmic factors) for all functions in $\mathcal{K}$.

Prominent examples of functions not in the class $\mathcal{K}$ include the piecewise affine convex functions. For these functions, faster rates are possible. We shall show in the next
section that the LSE converges at the parametric rate (upto logarithmic factors) for these functions.

Our proof of Theorem 3.1 is based on the application of Assouad’s lemma, the following version of which is a consequence of Lemma 24.3 of van der Vaart (1998, pp. 347). We start by introducing some notation. Let \( P_\phi \) denote the joint distribution of the observations \((x_1, Y_1), \ldots, (x_n, Y_n)\) when the true convex function equals \( \phi \). For two probability measures \( P \) and \( Q \) having densities \( p \) and \( q \) with respect to a common measure \( \mu \), the total variation distance, \( \|P - Q\|_{TV} \), is defined as \( \int (|p - q|/2) d\mu \) and the Kullback-Leibler divergence, \( D(P\|Q) \), is defined as \( \int p \log(p/q) d\mu \). Pinsker’s inequality asserts
\[
D(P\|Q) \geq 2\|P - Q\|_{TV}^2 \tag{12}
\]
for all probability measures \( P \) and \( Q \).

**Lemma 3.2 (Assouad).** Let \( m \) be a positive integer and suppose that, for each \( \tau \in \{0,1\}^m \), there is an associated convex function \( \phi_\tau \) in \( N(\phi_0) \). Then the following inequality holds:
\[
\mathcal{R}_n(\phi_0) \geq \frac{m}{8} \min_{\tau \neq \tau'} \frac{\mathcal{L}^2(\phi_\tau, \phi_{\tau'})}{\Upsilon(\tau, \tau')} \min_{\tau, \tau'} \left( 1 - \frac{\|P_{\phi_\tau} - P_{\phi_{\tau'}}\|_{TV} }{\|\phi_{\phi_{\tau'}}\|_{TV} } \right),
\]
where \( \Upsilon(\tau, \tau') := \sum_i \{ \tau_i \neq \tau'_i \} \).

In our application of Assouad’s lemma, we shall make use of Lemma 8.1 (stated and proved in Section 8) which bounds the distance between functions in \( \mathcal{R} \) and their piecewise linear interpolants. We are now ready to prove Theorem 3.1.

**Proof of Theorem 3.1.** Fix \( m \geq 1 \) and let \( t_i = a + (b - a)i/m \) for \( i = 0, \ldots, m \). For each \( i = 1, \ldots, m \), let \( \alpha_i \) define the linear interpolant of the points \((t_{i-1}, \phi_0(t_{i-1}))\) and \((t_i, \phi_0(t_i))\) i.e.,
\[
\alpha_i(x) := \phi_0(t_{i-1}) + \frac{\phi_0(t_i) - \phi_0(t_{i-1})}{t_i - t_{i-1}}(x - t_{i-1}) \quad \text{for } x \in [0, 1].
\]

By error estimates for linear interpolation (see e.g., Chapter 3 of Atkinson (1989)), for every \( x \in [t_{i-1}, t_i] \), there exists a point \( t_x \in [t_{i-1}, t_i] \) for which
\[
|\phi_0(x) - \alpha_i(x)| = (x - t_{i-1})(t_i - x) \frac{\phi''_0(t_x)}{2}
\]
which implies, by our assumption on \( \phi_0 \), that
\[
|\phi_0(x) - \alpha_i(x)| \leq (x - t_{i-1})(t_i - x) \frac{\kappa_2}{2} \quad \text{for every } x \in [t_{i-1}, t_i]. \tag{13}
\]
By convexity of \( \phi_0 \), it is obvious that \( \alpha_i(x) \geq \phi_0(x) \) for \( x \in [t_{i-1}, t_i] \) and \( \alpha_i(x) \leq \phi_0(x) \) for \( x \notin [t_{i-1}, t_i] \).
Now for each $\tau \in \{0, 1\}^m$, let us define
\[
\phi_\tau(x) := \max_{i: \tau_i = 1} \left( \phi_0(x), \max_{i: \tau_i = 0} \alpha_i(x) \right) \quad \text{for } x \in [0, 1].
\]
The functions $\phi_\tau$ are clearly convex because they equal the pointwise maximum of convex functions. Moreover, for $x \in [t_{i-1}, t_i]$, we have
\[
\phi_\tau(x) = \begin{cases} 
\alpha_i(x) & \text{if } \tau_i = 1 \\
\phi_0(x) & \text{if } \tau_i = 0
\end{cases}
\]
We apply Assouad’s lemma to these functions $\phi_\tau$. Clearly, for every $\tau, \tau' \in \{0, 1\}^m$, we have
\[
\ell^2(\phi_\tau, \phi_{\tau'}) = \sum_{i: \tau_i \neq \tau'_i} \ell^2(\phi_0, \max(\phi_0, \alpha_i)) \geq \Upsilon(\tau, \tau') \min_{1 \leq i \leq m} \ell^2(\phi_0, \max(\phi_0, \alpha_i))
\]
where $\Upsilon(\tau, \tau') := \sum_i \{\tau_i \neq \tau'_i\}$. Also, by inequality (12), we get
\[
\|P_{\phi_\tau} - P_{\phi_{\tau'}}\|_TV \leq \frac{1}{2} D(P_{\phi_\tau} \| P_{\phi_{\tau'}}).
\]
By the Gaussian assumption and independence of the errors, the Kullback-Leibler divergence $D(P_{\phi_\tau} \| P_{\phi_{\tau'}})$ can be easily calculated to be $n\ell^2(\phi_\tau, \phi_{\tau'})/(2\sigma)$. We therefore obtain
\[
\|P_{\phi_\tau} - P_{\phi_{\tau'}}\|_TV \leq \frac{\sqrt{n}}{2\sigma} \ell(\phi_\tau, \phi_{\tau'}). \quad \text{Also when } \Upsilon(\tau, \tau') = 1, \text{ it is clear that}
\]
\[
\ell^2(\phi_\tau, \phi_{\tau'}) \leq \max_{1 \leq i \leq m} \ell^2(\phi_0, \max(\phi_0, \alpha_i)).
\]
Thus by the application of Assouad’s lemma, we obtain the following lower bound for $\mathcal{R}_n(\phi_0)$:
\[
\frac{m}{8} \min_{1 \leq i \leq m} \ell^2(\phi_0, \max(\phi_0, \alpha_i)) \left\{ 1 - \frac{\sqrt{n}}{2\sigma} \sqrt{\max_{1 \leq i \leq m} \ell^2(\phi_0, \max(\phi_0, \alpha_i))} \right\} \quad \text{for each } \tau \in \{0, 1\}^m.
\]
(14) provided
\[
\phi_\tau \in N(\phi_0) \quad \text{for each } \tau \in \{0, 1\}^m.
\]
(15) We now use Lemma 8.1 to bound the quantity $\ell^2(\phi_0, \max(\phi_0, \alpha_i))$ from both above and below. Since $\alpha_i$ is the linear interpolant of $(t_{i-1}, \phi_0(t_{i-1}))$ and $(t_i, \phi_0(t_i))$, we use Lemma 8.1 with $a = t_{i-1}$ and $b = t_i$ to assert
\[
\frac{\kappa_2^2(t_i - t_{i-1})^5}{4096c_2} \leq \ell^2(\phi_0, \max(\phi_0, \alpha_i)) \leq \frac{\kappa_2^2(t_i - t_{i-1})^5}{32c_1}
\]
provided $n \geq 4c_2/(t_i - t_{i-1})$. Because $t_i - t_{i-1} = (b - a)/m$, we get
\[
\frac{\kappa_1^2(b - a)^5}{4096c_2m^5} \leq \ell^2(\phi_0, \max(\phi_0, \alpha_i)) \leq \frac{\kappa_2^2(b - a)^5}{32c_1m^5}
\]

provided \( n \geq 4c_2m/(b-a) \). Using this in (14), we deduce that

\[
\mathfrak{R}_n(\phi_0) \geq \frac{m}{8} \frac{\kappa_1^2 (b-a)^5}{4096m^5c_2} \left( 1 - \frac{\sqrt{n\kappa_2}}{2\sigma} \sqrt{\frac{(b-a)^5}{m^532c_1}} \right).
\]

The choice

\[
m = \frac{b-a}{\kappa_1} = \left( \frac{\sqrt{\kappa_2}}{\sigma \sqrt{32c_1}} \right)^{2/5}
\]

leads to (11). The constraint \( n \geq 4c_2m/(b-a) \) translates to

\[
n^2 \geq (2c_2)^{5/2}\kappa_2/(\sigma \sqrt{c_1}).
\]

We just need to verify (15). For this, note that

\[
sup_{x \in [0,1]} |\phi_\tau(x) - \phi_0(x)| = \max_{1 \leq i \leq m} \sup_{x \in [t_{i-1},t_i]} |\phi_0(x) - \max(\phi_0(x), \alpha_i(x))|
\]

and by (13), we get

\[
sup_{x \in [0,1]} |\phi_\tau(x) - \phi_0(x)| \leq \max_{1 \leq i \leq m} \frac{(t_i - t_{i-1})^2}{8} \frac{(b-a)^2\kappa_2}{8m^2} = \frac{(b-a)^2\kappa_2}{8m^2}.
\]

Finally the choice (16) leads to

\[
sup_{x \in [0,1]} |\phi_\tau(x) - \phi_0(x)| \leq \left( \frac{\kappa_1^2 c_1^2}{32} \right)^{1/5} \left( \frac{\sigma^2}{n} \right)^{2/5}
\]

which ensures (15). The proof is complete. \(\Box\)

4 Adaptation properties of the LSE

As we have seen in the last section, no estimator can converge faster to \( \phi_0 \) than the rate \( n^{-4/5} \) for \( \phi_0 \in \mathfrak{R} \). The LSE attains this rate \( n^{-4/5} \), up to logarithmic factors, as proved in Theorem 2.1. But for functions outside this class, faster rates are possible. The piecewise affine convex functions are an important class of functions that lie outside the class \( \mathfrak{R} \). We shall show in this section that, for these functions, the LSE converges at the parametric rate. In other words, the LSE automatically adapts to these functions. There is a small cost for this adaptation that we shall explain in the next section.

This adaptation behavior is proved in the following theorem. The notion of piecewise affine convex functions is relevant here. A convex function \( \alpha \) on \([0,1]\) is said to be piecewise affine if there exists an integer \( k \) and points \( 0 = t_0 < t_1 < \cdots < t_k = 1 \) such that \( \alpha \) is affine on each of the \( k \) intervals \([t_{i-1},t_i]\) for \( i = 1, \ldots, k \). We define \( k(\alpha) \) to be the smallest such \( k \). Let \( \mathcal{P}_k \) denote the collection of all piecewise affine convex functions with \( k(\alpha) \leq k \) and let \( \mathcal{P} \) denote the collection of all piecewise affine convex functions on \([0,1]\).
Theorem 4.1. Let $\alpha$ be a piecewise affine convex function on $[0,1]$ for which
\[
\ell^2(\phi_0, \alpha) \leq \frac{a\sigma^2k^{5/4}(\alpha)}{n}
\text{ for some positive constant } a.
\]
(17)

Then there exists a positive constant $C$ depending only on $a$ and the ratio $c_1/c_2$ such that
\[
E \phi_0 \ell^2(\phi_0, \hat{\phi}_{ls}) \leq C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2k^{5/4}(\alpha)}{n}
\]
(18)
whenever $n \geq 4c_1$.

Before giving the proof of the above result, let us first explain below that an immediate consequence of the above theorem is that when $\phi_0$ is a piecewise affine convex function, the LSE converges at the parametric rate upto logarithmic factors.

Example 4.2 (Rate for piecewise affine convex functions). Suppose $\phi_0$ is a piecewise affine convex function on $[0,1]$. Then (17) is clearly satisfied for every $a > 0$ with $\alpha = \phi_0$. We thus have the risk bound
\[
E \phi_0 \ell^2(\phi_0, \hat{\phi}_{ls}) \leq C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2k^{5/4}(\alpha)}{n}
\]
whenever $n \geq 4c_1$. This is the parametric rate $1/n$ upto logarithmic factors.

Remark 4.1. It is worth pointing out that the above risk bound depends on the piecewise affine convex function $\phi_0$ only through $k(\phi_0)$, the number of affine segments of $\phi_0$.

Remark 4.2. In order to construct estimators that adapt over a collection of model classes, one usually considers estimators that minimize a combination of a goodness of fit criterion over the individual model classes and a penalty that penalizes the complexity of the model classes. While the theory for such penalization estimators is well-developed (see e.g., Barron et al. (1999)), these estimators are, in general, computationally expensive and also rely on certain tuning parameters which might be difficult to choose in practice. In contrast, the convex LSE is very easy to compute, does not depend on any tuning parameter and, remarkably, it adapts over the classes $\mathcal{P}_k$ as $k$ varies.

Our results in Sections 2 and 3 together imply that the minimax rate for this problem is, upto logarithmic factors, $n^{-4/5}$. From the classical theory of nonparametric statistics, it follows that $n^{-4/5}$ is the same rate that one obtains for the estimation of twice differentiable functions on an interval. This equivalence of the two rates leads to the belief that convex function estimation is similar to the estimation of twice differentiable functions. Further credibility to this belief is lent by the deep theorem of Alexandrov (1939) which states that, in an appropriate sense, all convex functions are twice differentiable. Theorem 4.1 presents an important aspect where
convex function estimation differs from the estimation of twice differentiable functions. For the problem of estimating twice differentiable functions, the local minimax rate, $\mathcal{R}_n(\phi_0)$ (see (10)), is of the order $n^{-4/5}$ for all $\phi_0$, including, for example, affine functions. But Theorem 4.1 and our results of the previous section demonstrate that, for convex function estimation, $\mathcal{R}_n(\phi_0)$ changes with $\phi_0$. The reason for this difference in rates is that the class of convex functions, $\mathcal{C}$, is locally non-uniform in the sense that the local neighborhoods around certain convex functions (e.g., affine functions) are much sparser than local neighborhoods around other convex functions. On the other hand, in the class of twice differentiable functions, all local neighborhoods are, in some sense, equally sized.

Rates in between the parametric rate and the worst case rate $n^{-4/5}$ are possible depending on how well $\phi_0$ can be approximated by piecewise affine functions. The proof of Theorem 4.1 is given next. It is based on Theorem 2.2 and the following entropy integral bound whose proof is given in Section 7.2.

**Theorem 4.3 (Entropy integral bound).** Suppose $n \geq 4c_1$. Then there exists a constant $K$ depending only on the ratio $c_1/c_2$ such that

$$\int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r), \ell)} d\epsilon \leq K \left( \log \frac{n}{2c_1} \right)^{5/8} r^{3/4} \inf_{\alpha \in \mathcal{P}} \left[ k^{5/8}(\alpha) \left( r^2 + \ell^2(\phi_0, \alpha) \right)^{1/8} \right]$$

for every $\phi_0 \in \mathcal{C}$ and $r > 0$.

**Proof of Theorem 4.1.** Fix a piecewise affine convex function $\alpha$ for which (17) is satisfied for a positive constant $a$. We assume that $n \geq 4c_1$ throughout. Let us define

$$\delta_0 := A \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2 k^{5/8}(\alpha)}{n}$$

for a constant $A$ whose value will be specified shortly. The condition (17) and $n \geq 4c_1$ imply together that

$$\ell^2(\phi_0, \alpha) \leq \delta_0 \frac{a}{A} (\log 2)^{-5/4} \quad (19)$$

Theorem 4.3 shows the existence of a positive constant $K$ depending only on the ratio $c_1/c_2$ such that

$$\int_0^r \sqrt{\log M(\epsilon, S(\phi_0, r), \ell)} d\epsilon \leq K \left( \log \frac{n}{2c_1} \right)^{5/8} k^{5/8}(\alpha) r^{3/4} \left( r^2 + \ell^2(\phi_0, \alpha) \right)^{1/8}$$

for every $r > 0$. Let $H(r)$ denote the right hand side above. It is clear that $H(r)/r^2$ is decreasing on $(0, \infty)$. As a result, a condition of the form $\sqrt{n}\delta \geq C\sigma H(\sqrt{\delta})$ for some positive constant $C$ holds for every $\delta \geq \delta_0$ provided it holds for $\delta = \delta_0$. Because of (19), it can be easily checked that

$$H(\sqrt{\delta_0}) \leq K \left( \log \frac{n}{2c_1} \right)^{5/8} k^{5/8}(\alpha) \sqrt{\delta_0} \left( 1 + \frac{a}{A} (\log 2)^{-5/4} \right)^{1/8}.$$
Consequently, we have
\[ \frac{H(\sqrt{\delta_0})}{\delta_0} \leq \frac{K}{\sqrt{A}} \left( 1 + \frac{a}{A} (\log 2)^{-5/4} \right)^{1/8} \frac{\sqrt{n}}{\sigma}. \]  
(20)

We shall now use Theorem 2.2 with \( E = C \) and \( \phi = \phi_0 \). Let \( C \) be the positive constant given by Theorem 2.2. By inequality (20), we can clearly choose \( A \) depending only on \( K, C \) and \( a \) so that \( \sqrt{n}\delta_0 \geq C\sigma H(\sqrt{\delta_0}) \). Because \( H(r)/r^2 \) is a decreasing function of \( r \), this choice of \( A \) also ensures that \( \sqrt{n}\delta \geq C\sigma H(\sqrt{\delta}) \) for every \( \delta \geq \delta_0 \). Thus by Theorem 2.2, we obtain
\[ P_{\phi_0} \left( \ell^2(\hat{\phi}_{ls}, \phi_0) > \delta \right) \leq C \sum_{s \geq 0} \exp \left( -\frac{n2^{2s}\delta}{C^2\sigma^2} \right) \quad \text{for all } \delta \geq \delta_0. \]  
(21)

Note further that
\[ \frac{n\delta_0}{\sigma^2} \geq A(\log 2)^{5/4} \]  
(22)

which implies that the sum on the right hand side of (21) is dominated by the first term. We thus have
\[ P_{\phi_0} \left( \ell^2(\hat{\phi}_{ls}, \phi_0) > \delta \right) \leq C_1 \exp \left( -\frac{n\delta}{C_1\sigma^2} \right) \quad \text{for all } \delta \geq \delta_0. \]

for a constant \( C_1 \) depending upon only \( C \) and \( A \). The required risk bound (18) is now derived by integrating both sides of the above inequality with respect to \( \delta \) (and using (22) in the process). \( \square \)

5 Risk bounds for \( \bar{\phi}_k \)

This section is about the sieved LSE \( \bar{\phi}_k \) defined in (4). Even though it might be sensible to use this estimator when one has the knowledge that \( \phi_0 \in \mathcal{P}_k \), it should be noted that, in practice, \( \bar{\phi}_k \) is an infeasible estimator because one does not know which \( k \) to use and also because it is not easy to compute because of the combinatorial nature of the minimization that it is based on.

In this section, we show that when \( \phi_0 \in \mathcal{P}_k \), then the risk of \( \bar{\phi}_k \) is bounded from above by \( k/n \) upto logarithmic factors in \( n \). In the last section, we saw that when \( \phi_0 \in \mathcal{P}_k \), the LSE has risk bounded from above by \( k^{5/4}/n \) upto logarithmic factors in \( n \). Thus, the LSE only pays a price of at most a factor of \( k^{1/4}/n \) in risk for the adaptive and the computational advantages that it enjoys over the classes \( \mathcal{P}_k \).

**Theorem 5.1.** There exist positive constants \( C \) and \( b \) such that
\[ E_{\phi_0} \ell^2(\phi_0, \bar{\phi}_k) \leq C \frac{\sigma^2 k}{n} \log^2(bn) \]
for every \( k \leq n \) and \( \phi_0 \in \mathcal{P}_k \).
Remark 5.1. Unlike all other risk upper bounds in this paper, Theorem 5.1 holds with no assumptions whatsoever on the design points \( x_1, \ldots, x_n \).

The proof is again based on an application of Theorem 2.2 with \( \mathcal{E} = \mathcal{P}_k \). The main work is done in the following result whose proof is given in Section 7.3. For simplicity of notation, we denote \( S(\phi, r, \mathcal{P}_k) \) by \( S_k(\phi, r) \).

**Theorem 5.2.** There exists two universal positive constants \( K \) and \( b \) such that

\[
\int_0^r \sqrt{\log M(\epsilon, S_k(\phi_0, r), \ell)} d\epsilon \leq K r \sqrt{k \log (bn)}
\]

for every \( 1 \leq k \leq n \), \( \phi_0 \in \mathcal{C} \) and \( r > 0 \).

We next prove Theorem 5.1 using Theorems 2.2 and 5.2.

**Proof of Theorem 5.1.** We apply Theorem 2.2 with \( \mathcal{E} := \mathcal{P}_k \). Note that the estimator \( \hat{\phi}(\mathcal{E}) \) is precisely \( \hat{\phi}_k \).

By Theorem 5.2, the function \( H \) appearing in Theorem 2.2 can be taken to be \( H(r) := K r \sqrt{k \log (bn)} \) for two positive constants \( K \) and \( b \). The condition \( \sqrt{n} \delta \geq C \sigma H(\sqrt{\delta}) \) is then easily seen to be equivalent to \( \delta \geq \delta_0 \) where

\[
\delta_0 := K^2 C^2 n \sigma^2 \left( \log (bn) \right)^2.
\]

Theorem 2.2 therefore asserts the existence of a positive constant \( C \) such that

\[
P_{\phi_0} \left( \ell^2(\phi_0, \hat{\phi}_k) > \delta \right) \leq C \sum_{s \geq 0} \exp \left( -\frac{n^2 2^s \delta^2}{C^2 \sigma^2} \right) \quad \text{for all } \delta \geq \delta_0.
\]

We now proceed to complete the proof just like the proof of Theorem 4.1.

6 Model misspecification

In this section, we evaluate the performance of the convex LSE \( \hat{\phi}_{ls} \) in the case when the unknown regression function (to be denoted by \( f_0 \)) is not necessarily convex. Specifically, suppose that \( f_0 \) is an unknown function on \([0, 1]\) that is not necessarily convex. We consider observations \((x_1, Y_1), \ldots, (x_n, Y_n)\) from the model:

\[
Y_i = f_0(x_i) + \xi_i, \quad \text{for } i = 1, \ldots, n,
\]

where \( x_1 < \cdots < x_n \) are fixed design points in \([0, 1]\) and \( \xi_1, \ldots, \xi_n \) are independent normal variables with zero mean and variance \( \sigma^2 \).
The convex LSE $\hat{\phi}_{ls}$ is defined in the same way as before as any convex function that minimizes the sum of squares criterion. Since the true function $f_0$ is not necessarily convex, it turns out that the LSE is really estimating the convex projections of $f_0$. Any convex function $\phi_0$ on $[0, 1]$ that minimizes $\ell^2(f_0, \phi)$ over $\phi \in C$ is a convex projection of $f_0$ i.e.,

$$\phi_0 \in \arg\min_{\phi \in C} \sum_{i=1}^{n} (f_0(x_i) - \phi(x_i))^2.$$ 

Convex projections are not unique. However, because $\{(\phi(x_1), \ldots, \phi(x_n)) : \phi \in C\}$ is a convex closed subset of $\mathbb{R}^n$, it follows (see, for example Stark and Yang (1988, Chapter 2)) that the vector $(\phi_0(x_1), \ldots, \phi_0(x_n))$ is unique for every convex projection $\phi_0$ and, moreover, we have the inequality:

$$\ell^2(f_0, \phi) \geq \ell^2(f_0, \phi_0) + \ell^2(\phi_0, \phi) \quad \text{for every } \phi \in C. \quad (23)$$

In this section, we shall prove risk upper bounds for $\mathbb{E}_{f_0} \ell^2(f_0, \phi_0)$. These bounds are analogous to Theorems 2.1 and 4.1. Our main technical tool is the following theorem which is very similar to Theorem 2.2. Its proof proceeds in the same way as the proof of Theorem 2.2 (see Van de Geer (2000, proof of Theorem 9.1)) except that one needs to crucially use inequality (23). We provide below a sketch of its proof.

**Theorem 6.1.** Let $\phi_0$ denote any convex projection of $f_0$. Suppose $H$ is a function on $(0, \infty)$ such that

$$H(r) \geq \int_{0}^{r} \sqrt{\log M(\epsilon, S(\phi_0, r))} d\epsilon \quad \text{for every } r > 0$$

and such that $H(r)/r^2$ is decreasing on $(0, \infty)$. Then there exists a universal constant $C$ such that

$$\mathbb{P}_{f_0} \left( \ell^2(\hat{\phi}_{ls}, \phi_0) > \delta \right) \leq C \sum_{s \geq 0} \exp \left( -\frac{n2^s\delta}{C^2\sigma^2} \right)$$

for every $\delta > 0$ satisfying $\sqrt{n}\delta \geq C\sigma H(\sqrt{\delta})$.

**Proof.** Because $\phi_0$ is convex, we have, by the definition of $\hat{\phi}_{ls}$, that

$$\frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{\phi}_{ls}(x_i) \right)^2 \leq \frac{1}{n} \sum_{i=1}^{n} (Y_i - \phi_0(x_i))^2.$$

Writing $Y_i = f_0(x_i) + \xi_i$ and simplifying the above expression, we get

$$\ell^2(f_0, \hat{\phi}_{ls}) - \ell^2(f_0, \phi_0) \leq \frac{2}{n} \sum_{i=1}^{n} \xi_i \left( \hat{\phi}_{ls}(x_i) - \phi_0(x_i) \right).$$

Inequality (23) applied with $\phi = \hat{\phi}_{ls}$ gives

$$\ell^2(\hat{\phi}_{ls}, \phi_0) \leq \ell^2(f_0, \hat{\phi}_{ls}) - \ell^2(f_0, \phi_0).$$
Combining the above two inequalities, we obtain

\[ \ell^2(\hat{\phi}_{ls}, \phi_0) \leq \frac{2}{n} \sum_{i=1}^{n} \xi_i \left( \hat{\phi}_{ls}(x_i) - \phi_0(x_i) \right). \]

This is of the same form as the “basic inequality” of Van de Geer (2000, pp. 148). From here, the proof proceeds just as the proof of Theorem 9.1 in Van de Geer (2000).

We are now ready to present our risk bounds for \( \mathbb{E} \ell^2(\hat{\phi}_{ls}, \phi_0) \). We omit the proofs of these results because they are similar to the proofs in the well-specified case.

The next result gives the worst case upper bound for the risk. Its proof is very similar to the proof of Theorem 2.1 with the only difference being the use of Theorem 6.1 in place of Theorem 2.2.

**Theorem 6.2** (Worst case). Let \( \phi_0 \) denote any convex projection of \( f_0 \) and let \( R := \max(1, \mathcal{L}(\phi_0)) \). There exists a positive constant \( C \) depending only on the ratio \( c_1/c_2 \) such that

\[ \mathbb{E} \ell^2(\hat{\phi}_{ls}, \phi_0) \leq C \left( \frac{\sigma^2}{n} \right)^{4/5} R^{2/5} \log \frac{n}{2c_1} \]

whenever

\[ n \geq \max \left( 4c_1, C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2}{R^2} \right). \]

When \( \phi_0 \) can be well-approximated by a piecewise affine function having not too many knots, then \( \hat{\phi}_{ls} \) converges to \( \phi_0 \) at a much faster rate. This is shown in the following theorem whose proof is exactly similar to that of Theorem 4.1, the only difference being the use of Theorem 6.1 in place of Theorem 2.2.

**Theorem 6.3.** Let \( \phi_0 \) be any convex projection of \( f_0 \) and suppose that there exists a piecewise affine function \( \alpha \) on \([0, 1]\) such that

\[ \ell^2(\phi_0, \alpha) \leq \frac{a\sigma^2 k^{5/4}(\alpha)}{n} \]

for some positive constant \( a \).

Then there exists a positive constant \( C \) depending only on \( a \) and the ratio \( c_1/c_2 \) such that

\[ \mathbb{E} \ell^2(\hat{\phi}_{ls}, \phi_0) \leq C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2 k^{5/4}(\alpha)}{n} \]

whenever \( n \geq 4c_1 \).

**Remark 6.1.** Suppose \( \phi_0 \) is any convex projection of \( f_0 \) that lies in \( \mathcal{P}_k \) for some \( k \geq 1 \). Then the above result is applicable with \( \alpha = \phi_0 \) and we obtain

\[ \mathbb{E} \ell^2(\hat{\phi}_{ls}, \phi_0) \leq C \left( \log \frac{n}{2c_1} \right)^{5/4} \frac{\sigma^2 k^{5/4}}{n} \]

for every \( n \geq 4c_1 \).
An illuminating example of the above corollary occurs when $f_0$ is a concave function. In this case, we show in Lemma 8.6 (stated and proved in Section 8) that $\phi_0$ can be taken to be an affine function, i.e., $\phi_0 \in P_1$. As a result, we assert that if $f_0$ is concave, then $\hat{\phi}_{ls}$ converges to any convex projection at the parametric rate upto a logarithmic factor of $n$.

7 Metric entropy calculations

The goal of this section is to prove Theorems 2.3, 4.3 and 5.2 which were crucially used in our main risk upper bounds.

7.1 Proof of Theorem 2.3

Theorem 2.3 follows from Theorem 4.3, which is proved in the next subsection. Indeed, we just take the upper bound given by Theorem 2.3 and replace the infimum over $\alpha \in P$ to just the infimum over affine functions (note that $k(\alpha) = 1$ for affine functions) to obtain Theorem 4.3.

7.2 Proof of Theorem 4.3

The proof of Theorem 4.3 requires good upper bounds on the subclass $S(\phi_0, r)$ of convex functions (defined in (5) for $E = C$) which we shall derive below in a series of steps. The first result on covering numbers for convex functions was proved in Bronshtein (1976) who considered classes of convex functions that were uniformly bounded and uniformly Lipschitz. The convex functions in the class $S(\phi_0, r)$ are neither uniformly bounded nor uniformly Lipschitz with constants that are independent of $n$ and hence the results of Bronshtein (1976) cannot be directly applied to yield upper bounds for $S(\phi_0, r)$.

Bronshtein’s result was extended by Dryanov (2009) in the univariate case (and by Guntuboyina and Sen (2013) in the general multivariate case) to yield the following result which deals with the class of all uniformly bounded convex functions with no Lipschitz constraint. For $a < b$ and $B > 0$, let $\mathcal{C}([a, b], B)$ denote the class of all convex functions $f$ on $[a, b]$ which satisfy $|f(x)| \leq B$ for all $x \in [a, b]$.

**Theorem 7.1.** There exist positive constants $c$ and $\epsilon_0$ such that for every $B > 0$ and $b > a$, we have

$$
\log M (\mathcal{C}([a, b], B), \epsilon, L_2) \leq c \left( \frac{\epsilon}{B(b-a)^{1/2}} \right)^{-1/2}
$$
for every \( \epsilon \leq \epsilon_0 B(b - a)^{1/2} \). Here \( L_2 \) denotes the metric:

\[
L_2(f, g) := \left( \int_a^b |f(x) - g(x)|^2 dx \right)^{1/2}.
\]

The convex functions in the class \( S(\phi_0, r) \) are, unfortunately, not uniformly bounded by a constant that is independent of \( n \). Therefore, Theorem 7.1 cannot also be directly used to prove good upper bounds on the covering numbers of \( S(\phi_0, r) \). Our first result here is to extend Theorem 7.1 by replacing the uniform boundedness constraint by a weaker integral constraint. This is the content of the following theorem. For \( a < b \) and \( B > 0 \), let \( \mathcal{I}([a, b], B) \) denote the class of all real-valued convex functions \( f \) on \([a, b]\) for which \( \int_a^b f^2(x) dx \leq B^2 \).

**Theorem 7.2.** There exist constants \( c \) and \( \epsilon_0 \) such that

\[
\log M(\epsilon, \mathcal{I}([0, 1], B), L_2[\eta, 1 - \eta]) \leq c \left( \log \frac{1}{2\eta} \right)^{5/4} \left( \epsilon \frac{1}{B} \right)^{-1/2}
\]

for all

\[
\epsilon \leq B \epsilon_0 \left( \log \frac{1}{2\eta} \right)^{1/2} \quad \text{and} \quad \eta \leq 1/4.
\]

Here, by \( L_2[\eta, 1 - \eta] \), we mean the metric where the distance between \( f \) and \( g \) is given by

\[
\left( \int_{\eta}^{1-\eta} (f(x) - g(x))^2 dx \right)^{1/2}.
\]

The proof of this theorem is provided below. The main idea of the proof is the observation that functions in \( \mathcal{I}([a, b], B) \) become uniformly bounded on subintervals of \([a, b]\) that are sufficiently far away from the boundary points. This is made rigorous in Lemma 8.3. On such subintervals, we use Theorem 7.1 to bound the covering numbers. Theorem 7.2 is then proved by putting together these different covering numbers as shown below.

**Proof of Theorem 7.2.** By a trivial scaling argument, we can assume without loss of generality that \( B = 1 \). We write

\[
\int_{\eta}^{1/2} (\phi(x) - f(x))^2 dx = \sum_{i=0}^{l} \int_{\eta_i}^{\eta_{i+1}} (\phi(x) - f(x))^2 dx
\]

for a finite sequence \( \eta_0 = \eta < \eta_1 < \eta_2 < \cdots < \eta_l < 1/2 \leq \eta_{l+1} \). Now, by Lemma 8.3, the restriction of the functions \( \phi \in \mathcal{I}([0, 1], 1) \) to \([\eta_i, \eta_{i+1}]\) are convex and uniformly bounded by \( 2\sqrt{3\eta_i}^{-1/2} \). Therefore, by Theorem 7.1, there exist positive constants \( c \)
and $\epsilon_0$ such that we can cover the functions in $\mathcal{F}([0, 1], 1)$ in the $L_2[\eta_i, \eta_{i+1}]$ metric to within $\alpha_i$ by a finite set having cardinality at most

$$
\exp \left[ c \left( \frac{\alpha_i \sqrt{\eta_i}}{\sqrt{\eta_{i+1} - \eta_i}} \right)^{-1/2} \right]
$$

provided $\alpha_i \sqrt{\eta_i} \leq \epsilon_0 \sqrt{\eta_{i+1} - \eta_i}$.

We, therefore, get a cover for functions in $\mathcal{F}([0, 1], 1)$ in the $L_2[\eta, 1/2]$ metric of size less than or equal to $\left( \sum_{i=0}^{l} \alpha_i^2 \right)^{1/2}$ and cardinality at most

$$
\exp \left[ c \sum_{i=0}^{l} \left( \frac{\alpha_i \sqrt{\eta_i}}{\sqrt{\eta_{i+1} - \eta_i}} \right)^{-1/2} \right]
$$

provided $\max_{0 \leq i \leq l} \frac{\alpha_i \sqrt{\eta_i}}{\sqrt{\eta_{i+1} - \eta_i}} \leq \epsilon_0$.

Taking $\eta_i = 2^i \eta$ and $\alpha_i = \epsilon (l + 1)^{-1/2}$, we get that

$$
\log M(\epsilon, \mathcal{F}([0, 1], 1), L_2[\eta, 1/2]) \leq c \epsilon^{-1/2} (l + 1)^{5/4}
$$

provided $\epsilon \leq \sqrt{l + 1} \epsilon_0$.

Because $\eta_l \leq 1/2$, we get that $l \leq -\log(2\eta)/\log 2$. Also since $\eta \leq 1/4$, we have $-\log(2\eta)/\log 2 \geq 1$ and $l + 1 \leq -2 \log(2\eta)/\log 2$. Further, we have $l + 1 \geq -\log(2\eta)$ because $\eta_{i+1} \geq 1/2$. As a result, we deduce

$$
\log M(\epsilon, \mathcal{F}([0, 1], 1), L_2[\eta, 1/2]) \leq c \epsilon^{-1/2} \left( \log \frac{1}{2\eta} \right)^{5/4}
$$

provided $\epsilon \leq \epsilon_0 \left( \log \frac{1}{2\eta} \right)^{1/2}$. By symmetry, the same upper bound will also hold for $\log M(\epsilon, \mathcal{F}([0, 1], 1), L_2[1/2, 1 - \eta])$. The proof is completed by putting these two bounds together.

We are now ready to prove covering number bounds for the class $S(\phi_0, r)$. The simplest case corresponds to $\phi_0$ being affine and is proved in the next theorem. Its proof is based on Theorem 7.2. We would also need to switch between the pseudometric $\ell$ and the $L_2[\eta, 1 - \eta]$ metric. This will be made convenient by the use of Lemma 8.4.

**Theorem 7.3 (Affine case).** There exist positive constants $c$ and $\epsilon_0$ depending only on the ratio $c_1/c_2$ such that for every affine function $\tau$ on $[0, 1]$, we have

$$
\log M(\epsilon, S(\tau, r), \ell) \leq c \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{\epsilon}{r} \right)^{-1/2}
$$

whenever $r > 0, n \geq 4c_1$ and $\epsilon \leq \epsilon_0 r$.

**Proof.** By the translation invariance of the Euclidean distance and the fact that $\tau$ is affine (which implies that $\phi - \tau$ is convex for all $\phi \in \mathcal{C}$), we have

$$
M(\epsilon, S(\tau, r), \ell) = M(\epsilon, S(0, r), \ell),
$$

20
where the zero on the right hand side above refers to the zero function. By an 
elementary scaling argument, it follows that

\[ M(\epsilon, S(0, r), \ell) = M(\epsilon/r, S(0, 1), \ell). \]

We, therefore, only need to prove the theorem for \( \tau \equiv 0 \) and \( r = 1 \). For ease of 
notation, let us denote \( S(0, 1) \) by \( S \).

Because \( x_i - x_{i-1} \geq c_1/n \) for all \( i = 2, \ldots, n \), we have \( x_2, \ldots, x_{n-1} \in [c_1/n, 1-(c_1/n)] \).

We shall first prove an upper bound for \( \log M(\epsilon, S, \ell_1) \) where

\[ \ell_1^2(\phi, \psi) := \frac{1}{n-2} \sum_{i=2}^{n-1} (\phi(x_i) - \psi(x_i))^2. \]

For each function \( \phi \in S \), let \( \tilde{\phi} \) be the convex function on \([x_2, x_{n-1}]\) defined by

\[ \tilde{\phi}(x) := \frac{x_{i+1} - x}{x_{i+1} - x_i} \phi(x_i) + \frac{x - x_i}{x_{i+1} - x_i} \phi(x_{i+1}) \quad \text{for} \quad x_i \leq x \leq x_{i+1} \]

where \( i = 2, \ldots, n-2 \). Also let \( \tilde{S} := \{ \tilde{\phi} : \phi \in S \} \).

By Lemma 8.4 and the assumption that \( x_i - x_{i-1} \geq c_1/n \) for all \( i \), we get that

\[ \ell_1^2(\phi, \psi) \leq \frac{6}{c_1} \int_{x_2}^{x_{n-1}} \left( \tilde{\phi}(x) - \tilde{\psi}(x) \right)^2 dx \]

for every pair of functions \( \phi \) and \( \psi \) in \( S \). Letting \( \delta := c_1\epsilon/12 \) this inequality implies that

\[ M(\epsilon, S, \ell_1) \leq M(\delta, \tilde{S}, L_2[x_2, x_{n-1}]). \]

Again by Lemma 8.4 and the assumption \( x_i - x_{i-1} \leq c_2/n \), we have that

\[ \int_{x_1}^{x_n} \phi^2(x) dx \leq \frac{c_2}{n} \sum_{i=1}^{n} \phi^2(x_i) \leq c_2 \quad \text{for every} \quad \phi \in S. \]

As a result, we have that \( \tilde{S} \subseteq J([x_1, x_n], c_2) \). Further, because \( x_2 \geq x_1 + c_1/n \) and \( x_{n-1} \leq x_n - c_1/n \), we get that

\[ M(\delta, \tilde{S}, L_2[x_2, x_{n-1}]) \leq M(\delta, J([x_1, x_n], c_2), L_2[x_1 + \eta, x_n - \eta]) \]

where \( \eta := c_1/n \). Because \( [x_1, x_n] \subseteq [0, 1] \), it follows trivially that

\[ M(\delta, J([x_1, x_n], c_2), L_2[x_1 + \eta, x_n - \eta]) \leq M(\delta, J([0, 1], c_2), L_2[\eta, 1 - \eta]). \]

Thus, by Theorem 7.2, we assert the existence of two positive constants \( c \) and \( \epsilon_0 \) such that

\[ \log M(\epsilon, S, \ell_1) \leq c \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{c_1\epsilon}{c_2} \right)^{-1/2} \]

(24)
provided

\[ n \geq 4c_1 \quad \text{and} \quad \epsilon \leq \frac{c_0 c_2}{c_1} \left( \log \frac{n}{2c_1} \right)^{1/2} . \]

Now for every pair of functions \( \phi \) and \( \psi \) in \( S \), we have

\[ \ell^2(\psi, \phi) \leq \ell^2_1(\psi, \phi) + \frac{1}{n} \sum_{i \in \{1, n\}} (\phi(x_i) - \psi(x_i))^2 . \]

We make the simple observation that \((\phi(x_1), \phi(x_n))\) lies in the closed ball of radius \( \sqrt{n} \) in \( \mathbb{R}^2 \) denoted by \( B_2(0, \sqrt{n}) \). As a result, we have

\[ M(\epsilon, S, \ell) \leq M\left( \frac{\epsilon}{\sqrt{2}}, S, \ell_1 \right) M\left( \sqrt{\frac{2}{\epsilon}}, B_2(0, \sqrt{n}) \right) \leq \left( 1 + \frac{2\sqrt{2}}{\epsilon} \right)^2 M\left( \frac{\epsilon}{\sqrt{2}}, S, \ell_1 \right) \]

where the covering number of \( B_2(0, \sqrt{n}) \) is in the usual Euclidean metric. Using (24), we get

\[ \log M(\epsilon, S, \ell) \leq 2 \log \left( 1 + \frac{2\sqrt{2}}{\epsilon} \right) + c \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{c_1 \epsilon}{\sqrt{2c_2}} \right)^{-1/2} \]

provided

\[ n \geq 4c_1 \quad \text{and} \quad \epsilon \leq \frac{c_0 c_2}{c_1} \left( \log \frac{n}{2c_1} \right)^{1/2} . \]

Let \( u := \sqrt{2} \log \frac{2\epsilon_0 c_2}{c_1} \) and note that (25) holds whenever \( n \geq 4c_1 \) and \( \epsilon \leq u \). Further, when \( \epsilon \leq u \), we have

\[ \log \left( 1 + \frac{2\sqrt{2}}{\epsilon} \right) \leq \log \left( \frac{u + 2\sqrt{2}}{\epsilon} \right) \leq 2\epsilon^{-1/2} \sqrt{u + 2\sqrt{2}} \]

where we have used the inequality \( \log x \leq 2\sqrt{x} \) for \( x > 0 \). As a result, we have

\[ \log M(\epsilon, S, \ell) \leq \epsilon^{-1/2} \left( 4 \sqrt{u + 2\sqrt{2}} + c \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{c_1 \epsilon}{\sqrt{2c_2}} \right)^{-1/2} \right) \]

whenever \( n \geq 4c_1 \) and \( \epsilon \leq u \). The proof is therefore completed upon renaming the constants \( c_0 \) and \( c \) appropriately. \( \square \)

We next consider the case of convex piecewise affine functions \( \phi_0 \).

**Theorem 7.4** (Piecewise affine case). Let \( \alpha \) be a convex function on \([0, 1]\) that is affine on each of \( k \) subintervals of \([0, 1]\). Then there exist points \( c \) and \( \epsilon_0 \) depending only on the ratio \( c_1/c_2 \) such that

\[ \log M(\epsilon, S(\alpha, r), \ell) \leq ck^{5/4} \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{\epsilon}{r} \right)^{-1/2} \]

provided \( n \geq 4c_1 \) and \( \epsilon \leq r\epsilon_0 \sqrt{k} \).
Proof. Suppose that \( \alpha \) is affine on each of the \( k \) intervals \( I_i = (t_{i-1}, t_i) \) for \( i = 2, \ldots, k \), where \( 0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1 \), and \( I_1 = [0, t_1] \). Then there exist \( k \) affine functions \( \tau_1, \ldots, \tau_k \) on \( [0, 1] \) such that \( \alpha(x) = \tau_i(x) \) for \( x \in I_i \) for every \( i = 1, \ldots, k \).

For every pair of functions \( f \) and \( g \) on \( [0, 1] \), we have the trivial identity: \( \ell^2(f, g) = \sum_{i=1}^k \ell_i^2(f, g) \) where
\[
\ell_i^2(f, g) := \frac{1}{n} \sum_{j \in I_i} (f(x_j) - g(x_j))^2.
\]

As a result, we clearly have
\[
M(\epsilon, S(\alpha, r), \ell) \leq \prod_{i=1}^k M(\epsilon/\sqrt{k}, S(\alpha, r), \ell_i)
\]

Since the pseudometric \( \ell_i \) only depends on those \( x_j \) which lie in \( I_i \) and because \( \alpha \) equals the affine function \( \tau_i \) on \( I_i \), we have
\[
M(\epsilon/\sqrt{k}, S(\alpha, r), \ell_i) = M(\epsilon/\sqrt{k}, S(\tau_i, r), \ell_i) \leq M(\epsilon/\sqrt{k}, S(\tau_i, r), \ell).
\]

We deduce therefore that
\[
M(\epsilon, S(\alpha, r), \ell) \leq \prod_{i=1}^k M(\epsilon/\sqrt{k}, S(\tau_i, r), \ell).
\]

Now by Theorem 7.3, there exist positive constants \( c \) and \( \epsilon_0 \) depending only on the ratio \( c_1/c_2 \) such that
\[
\log M(\epsilon/\sqrt{k}, S(\tau_i, r), \ell) \leq c \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{\epsilon}{\sqrt{kr}} \right)^{-1/2}
\]
whenever \( n \geq 4c_1 \) and \( \epsilon \leq \epsilon_0 r^{1/2} k \). The proof follows by combining (26) and (27).

We now have all the ingredients for the proof of Theorem 4.3. Putting these together and using the simple integration Lemma 8.5 completes the proof as shown below.

Proof of Theorem 4.3. Suppose \( n \geq 4c_1 \). Fix functions \( \phi_0 \in \mathcal{C} \) and \( \alpha \in \mathcal{P} \). The simple inequality \((a + b)^2 \leq 2a^2 + 2b^2\) gives
\[
\ell^2(\phi, \alpha) \leq 2\ell^2(\phi, \phi_0) + 2\ell^2(\phi_0, \alpha).
\]

Therefore \( \ell^2(\phi, \alpha) \leq 2r^2 + 2\ell^2(\phi_0, \alpha) \) for every \( \phi \in S(\phi_0, r) \). Hence
\[
M(\epsilon, S(\phi_0, r), \ell) \leq M \left( \epsilon, S \left( \alpha, \sqrt{2r^2 + \ell^2(\phi_0, \alpha)} \right), \ell \right).
\]

We now use Theorem 7.4 to claim the existence of two positive constants \( c \) and \( \epsilon_0 \) depending only on the ratio \( c_1/c_2 \) such that
\[
\log M(\epsilon, S(\phi_0, r), \ell) \leq c k^{5/4}(\alpha) \left( \log \frac{n}{2c_1} \right)^{5/4} \left( \frac{\epsilon}{\sqrt{r^2 + \ell^2(\phi_0, \alpha)}} \right)^{-1/2}
\]

23
provided $\epsilon \leq \epsilon_0 \sqrt{k(\alpha) \sqrt{r^2 + \ell^2(\phi_0, \alpha)}}$. We now complete the proof by applying Lemma 8.5 with

$$A = ck^{5/4}(\alpha) \left( \log \frac{n}{2c_1} \right)^{5/4}, \quad B = \sqrt{r^2 + \ell^2(\phi_0, \alpha)} \quad \text{and} \quad u = \epsilon_0 \sqrt{k(\alpha)}.$$

\[ \square \]

7.3 Proof of Theorem 5.2

The main ingredient in the proof of Theorem 5.2 is the following lemma.

Lemma 7.5. Fix $k \geq 1$ and $\phi_0 \in \mathcal{C}$. The following inequality holds for every $r > 0$ and $\epsilon > 0$:

$$M(\epsilon, S_k(\phi_0, r); \ell) \leq \left( 4 + \frac{2B}{\epsilon} \sqrt{n} \right)^{b_1 k \log(b_2 k)},$$

where $b_1$ and $b_2$ are universal positive constants.

For the proof of this lemma, we use available techniques for bounding covering numbers using combinatorial notions of dimension. Specifically, we use the notion of pseudodimension, introduced by Pollard (1990, Chapter 4) as a generalization of the Vapnik-Červonenkis dimension to classes of real valued functions. The pseudodimension of a subset $A$ of $\mathbb{R}^n$ is defined as the maximum cardinality of a subset $\sigma \subseteq \{1, \ldots, n\}$ for which there exists $(h_1, \ldots, h_n) \in \mathbb{R}^n$ such that: for every $\sigma' \subseteq \sigma$, one can find $(a_1, \ldots, a_n) \in A$ with $a_i < h_i$ for $i \in \sigma'$ and $a_i > h_i$ for $i \in \sigma \setminus \sigma'$. The following theorem is a special case of results in Pollard (1990, Chapter 4) and gives an upper bound for the covering number (with respect to the Euclidean metric) of a subset of $\mathbb{R}^n$ in terms of its pseudodimension. Stronger results of this kind have been proved by Mendelson and Vershynin (2003) and the following theorem is also a special case of Mendelson and Vershynin (2003, Theorem 1).

Theorem 7.6. Let $A$ be a subset of $\mathbb{R}^n$ with $\max_i |a_i| \leq B$ for all $a = (a_1, \ldots, a_n) \in A$. If the pseudodimension of $A$ is at most $V$, then

$$M(t, A) \leq \left( 4 + \frac{2B}{t} \sqrt{n} \right)^{bV} \quad \text{for every} \ t > 0,$$

where $b$ is a universal positive constant. Here the covering number $M(t, A)$ is measured in the usual Euclidean metric on $\mathbb{R}^n$.

Proof of Lemma 7.5. Fix $k \geq 1$ and $\phi_0 \in \mathcal{C}$. Let $x_0 := (\phi_0(x_1), \ldots, \phi_0(x_n))$ and

$$H_k := \{ x \in \mathbb{R}^n : x = (\phi(x_1), \ldots, \phi(x_n)) \text{ for some } \phi \in \mathcal{P}_k \}.$$

Clearly

$$M(\epsilon, S_k(\phi_0, r); \ell) = M(B_{\alpha}(0, \sqrt{n}r) \cap H_k - x_0, \sqrt{n}\epsilon),$$

24
where \( H_k - x_0 := \{ x - x_0 : x \in H_k \} \). The covering number on the right hand side above is with respect to the usual Euclidean metric on \( \mathbb{R}^n \).

We now show that the pseudodimension of \( H_k \), which clearly equals the pseudodimension of \( H_k - x_0 \), is less than or equal to \( 2\kappa k \log(\zeta k) \) where \( \zeta = 2/\log 2 \). An application of Theorem 7.6 would then complete the proof. Note that the quantity \( B \) in the statement of Theorem 7.6 can be taken to be \( \sqrt{nr} \) because \( |a_i| \leq \sqrt{nr} \) for every \((a_1, \ldots, a_n) \in B_{n}(0, \sqrt{nr})\).

Every \( \phi \) in \( P_k \) can be written as \( \phi(x) = \max_{1 \leq i \leq k} \tau_i(x) \) where \( \tau_1, \ldots, \tau_k \) are affine on \([0, 1]\). As a result, \( H_1 \) is a linear subspace with dimension at most 2 which implies (see Pollard (1990, pp. 15)) that the pseudodimension of \( H_1 \) is at most 2. The pseudodimension of \( H_k \), which consists of coordinatewise maxima of at most \( k \) points in \( H_1 \), can then be bounded from above following the argument in Pollard (1990, Proof of Lemma (5.1)). Indeed, that argument shows that the pseudodimension of \( H_k \) is bounded from above by the smallest positive integer \( m \) for which

\[
\left( m_0 \right) + \left( m_1 \right) + \left( m_2 \right) < 2^{m/k}. \tag{28}
\]

If \( B \) denotes a binomial random variable with parameters \( m \) and \( 1/2 \), then the left hand side of the above inequality equals \( 2^m \Pr\{B \geq m - 2\} \) and can be bounded, for every \( a > 0 \), by

\[
2^m \Pr\{B \geq m - 2\} \leq 2^m \Pr\{a^B \geq a^{m-2}\} \leq 2^m a^{2-m} \mathbb{E} a^B = (1 + a)^m a^{2-m}. \tag{29}
\]

If we now choose \( a = \left(2^{1/(2k)} - 1\right)^{-1} \), then \((1 + a)/a = 2^{1/(2k)}\) and also, applying the inequality \( x - 1 > \log x \) to \( x = 2^{1/(2k)} \), we get that \( a < 2k/(\log 2) \). Therefore, from (29), we have

\[
\left( m_0 \right) + \left( m_1 \right) + \left( m_2 \right) = 2^m \Pr\{B \geq m - 2\} < \left( \frac{2k}{\log 2} \right)^2 2^{m/(2k)}. \]

The following inequality therefore ensures that \( m \) satisfies (28):

\[
\left( \frac{2k}{\log 2} \right)^2 2^{m/(2k)} \leq 2^{m/k} \text{ or, equivalently, } m \geq 2\kappa k \log(\zeta k),
\]

where \( \zeta = 2/(\log 2) \). It follows, therefore, that the pseudodimension of \( H_k \) is at most \( 2\kappa k \log(\zeta k) \). The proof of Lemma 7.5 is complete.

We are now ready to prove Theorem 5.2.

**Proof of Theorem 5.2.** Fix \( k \geq 1, \phi_0 \in \mathcal{C} \) and \( r > 0 \). By Lemma 7.5, we have

\[
\int_0^r \sqrt{\log M(\epsilon, S_k(\phi_0, r), \ell)} d\epsilon \leq \sqrt{b_1 k \log(b_2 k)} \int_0^r \sqrt{\log \left( 4 + \frac{2\sqrt{nr}}{\epsilon} \right)} d\epsilon
\]
The integral on the right hand side above equals, by the change of variable $r = \epsilon x$,

$$r \int_{1}^{\infty} \frac{\sqrt{\log(4 + 2x\sqrt{n})}}{x^2} dx$$

which is easily seen to be bounded from above by $K_1 r \sqrt{\log(4 + 2\sqrt{n})}$ for some universal positive constant $K_1$. The proof is complete because $\log(b_2 k) \log(4 + 2\sqrt{n}) \leq \log^2(bn)$ for a positive constant $b$ whenever $1 \leq k \leq n$. □

8 Some auxiliary results

Lemma 8.1. Fix a function $\phi_0 \in \mathcal{K}(a, b, \kappa_1, \kappa_2)$. Let $\alpha$ denote the linear interpolant of the points $(a, \phi_0(a))$ and $(b, \phi_0(b))$ i.e.,

$$\alpha(x) := \phi_0(a) + \frac{\phi_0(b) - \phi_0(a)}{b - a} (x - a) \quad \text{for } x \in [0, 1].$$

For every $n \geq 4c_2/(b - a)$, we have

$$\frac{\kappa_1^2(b - a)^5}{4096c_2} \leq \ell^2(\phi_0, \max(\phi_0, \alpha)) \leq \frac{\kappa_2^2(b - a)^5}{32c_1}. \quad (30)$$

Proof of Lemma 8.1. By convexity of $\phi_0$, it is obvious that $\alpha(x) \geq \phi_0(x)$ for $x \in [a, b]$ and $\alpha(x) \leq \phi_0(x)$ for $x \notin [a, b]$. We therefore have

$$\ell^2(\phi_0, \max(\phi_0, \alpha)) = \frac{1}{n} \sum_{i=1}^{n} (\alpha(x_i) - \phi_0(x_i))^2 I \{x_i \in [a, b]\}, \quad (31)$$

where $I$ denotes the indicator function. By standard error estimates for linear interpolation, for every $x \in [a, b]$, there exists a point $t_x \in [a, b]$ for which

$$|\phi_0(x) - \alpha(x)| = (x - a)(b - x) \frac{\phi_0''(t_x)}{2}$$

which implies, by our assumption on $\phi_0$, that

$$(x - a)(b - x) \frac{\kappa_1}{2} \leq |\phi_0(x) - \alpha(x)| \leq (x - a)(b - x) \frac{\kappa_2}{2}. \quad (32)$$

We first prove the lower bound in (30). From (31) and (32), we get

$$\ell^2(\phi_0, \max(\phi_0, \alpha)) \geq \frac{\kappa_1^2}{4n} \sum_{i=1}^{n} (x_i - a)^2(b - x_i)^2 I \{x_i \in [a, b]\} \geq \frac{\kappa_2^2}{4n} \sum_{i=1}^{n} (x_i - a)^2(b - x_i)^2 I \{(3a + b)/4, (a + 3b)/4\}.$$
Clearly \((x - a)(b - x) \geq (b - a)^2/16\) for every \(x \in [(3a + b)/4, (a + 3b)/4]\) and hence,

\[
\ell^2(\phi_0, \max(\phi_0, \alpha)) \geq \frac{\kappa_1^2}{1024} \frac{(b - a)^4}{n} \sum_{i=1}^{n} I \{x_i \in [(3a + b)/4, (a + 3b)/4]\}. \]

We now use Lemma 8.2 (to get a lower bound on the number of points \(x_1, \ldots, x_n\) that are contained in the interval \([(3a + b)/4, (a + 3b)/4]\)) to obtain

\[
\ell^2(\phi_0, \max(\phi_0, \alpha)) \geq \frac{\kappa_1^2}{1024} \frac{(b - a)^4}{n} \left( \frac{n(b - a)}{2c_2} - 1 \right). \]

The condition \(n \geq 4c_2/(b - a)\) now implies that

\[
\frac{n(b - a)}{2c_2} - 1 \geq \frac{n(b - a)}{4c_2}
\]

which completes the proof of the lower bound in (30).

To prove the upper bound in (30), we again use (31) and (32) to write

\[
\ell^2(\phi_0, \max(\phi_0, \alpha)) \leq \frac{\kappa_2^2}{4n} \sum_{i=1}^{n} (x_i - a)^2(b - x_i)^2 I \{x_i \in [a, b]\}. \]

Because \((x - a)(b - x) \leq (b - a)^2/4\) for all \(x \in [a, b]\), we obtain

\[
\ell^2(\phi_0, \max(\phi_0, \alpha)) \leq \frac{\kappa_2^2}{64} \frac{(b - a)^4}{n} \sum_{i=1}^{n} I \{x_i \in [a, b]\}. \]

We use Lemma 8.2 (to obtain an upper bound on the number of points \(x_1, \ldots, x_n\) that are contained in \([a, b]\)) to get

\[
\ell^2(\phi_0, \max(\phi_0, \alpha)) \leq \frac{\kappa_2^2}{64} \frac{(b - a)^4}{n} \left( \frac{n(b - a)}{c_1} + 1 \right) \]

Because \(n \geq 4c_2/(b - a)\) and \(c_2 \geq c_1\), we have

\[
\frac{n(b - a)}{c_1} + 1 \leq \frac{2n(b - a)}{c_1}
\]

and this completes the proof. \(\square\)

**Lemma 8.2.** Let \(x_1 < \cdots < x_n\) be fixed points in \([0, 1]\) satisfying \(c_1 \leq n(x_i - x_{i-1}) \leq c_2\) for all \(2 \leq i \leq n\). Let \([a, b]\) be a subinterval of \([0, 1]\) that contains \(m\) of the \(n\) real numbers \(x_1, \ldots, x_n\). Then

\[
\frac{n(b - a)}{c_2} - 1 \leq m \leq \frac{n(b - a)}{c_1} + 1. \tag{33}
\]
Proof. Let \( x_0 := \max (x_1 - c_2/n, 0) \) and \( x_{n+1} := \min (x_n + c_2/n, 1) \). Let 
\[
\{x_1, \ldots, x_n\} \cap [a,b] = \{x_{k+1}, \ldots, x_{k+m}\}
\]
for some \( 0 \leq k \leq n - m \). Clearly 
\[
b - a \geq x_{k+m} - x_{k+1} = \sum_{i=k+2}^{k+m} (x_i - x_{i-1}) \geq \frac{c_1(m-1)}{n}
\]
which gives the upper bound in (33). On the other hand,
\[
b - a \leq x_{k+m+1} - x_k = \sum_{i=k+1}^{k+m+1} (x_i - x_{i-1}) \leq \frac{c_2(m+1)}{n}
\]
which gives the lower bound in (33). The proof is complete. \( \square \)

Lemma 8.3. Let \( \phi \) be a convex function on \([0,1]\) for which 
\[
\int_0^1 |\phi(x)|^p \, dx \leq 1 \text{ for a fixed } p \geq 1.
\]
Then \( |\phi(y)| \leq 2(1 + p)^{1/p} \max \left( y^{-1/p}, (1-y)^{-1/p} \right) \) for all \( y \in (0,1) \).

Proof. Enough to prove the theorem for \( 0 < y < 1/2 \).

Suppose \( \phi(y) > y^{-1/p} \). Then, by convexity of \( \phi \), the condition \( \phi(x) > \phi(y) \) must hold either for all \( x \in (0,y) \) or for all \( x \in (y,1) \). Therefore,
\[
1 \geq \int_0^1 |\phi(x)|^p \, dx \geq \phi(y)^p \min(y,1-y) \geq \phi(y)^p y
\]
which gives a contradiction. Therefore \( \phi(y) \leq y^{-1/p} \).

Suppose, if possible, that \( \phi(y) < - cy^{-1/p} \) for some \( c > 1 \). We consider the following cases separately.

Case (i): Assume \( \phi(0) < - cy^{-1/p} \). In this case, by convexity of \( \phi \), it follows that 
\( \phi(x) < - cy^{-1/p} \) for all \( x \in [0,y] \). Therefore \( |\phi(x)| > cy^{-1/p} \) and thus
\[
1 \geq \int_0^1 |\phi(x)|^p \, dx \geq \int_0^y \frac{c^p}{y} \, dx = c^p.
\]
This contradicts \( c > 1 \).

Case (ii): Here \( \phi(0) \geq - cy^{-1/p} \). We now consider the following two subcases:

1. \( \phi(0) \leq 0 \). Then \( \phi(x) \leq 0 \) for all \( x \in [0,y] \). For each \( 0 \leq x \leq y \), we have, by convexity,
\[
\phi(x) \leq \left(1 - \frac{x}{y}\right) \phi(0) + \frac{x}{y} \phi(y) \leq \frac{x}{y} \phi(y).
\]
Thus \( y \phi(x) \leq x \phi(y) \leq 0 \) for each \( 0 \leq x \leq y \). As a result,
\[
y^p |\phi(x)|^p \geq x^p |\phi(y)|^p \quad \text{for } 0 \leq x \leq y.
\]
Integrating both sides from \( x = 0 \) to \( x = y \), we obtain
\[
y^p \int_0^y |\phi(x)|^p dx \geq |\phi(y)|^p \frac{y^{p+1}}{p+1}
\]
which implies that \( |\phi(y)|^p \leq (p + 1)/y \), i.e., \( |\phi(y)| \leq (1 + p)^{1/p}y^{-1/p} \) which is a contradiction if \( c > (1 + p)^{1/p} \).

2. \( \phi(0) > 0 \). Let \( z \in (0, y) \) be such that \( \phi(z) = 0 \). For \( x < z \), we can write, by convexity,
\[
0 = \phi(z) \leq \frac{y - z}{y - x} \phi(x) + \frac{z - x}{y - x} \phi(y)
\]
which implies that
\[
0 > \phi(y) \geq \frac{y - z}{x - z} \phi(x).
\]
As a result, \( |z - x|^p |\phi(y)|^p \leq |y - z|^p |\phi(x)|^p \) for \( 0 < x < z \). Integrating both sides from \( x = 0 \) to \( x = z \), we get
\[
|\phi(y)|^p \frac{z^{p+1}}{p+1} \leq |y - z|^p \int_0^z |\phi(x)|^p dx.
\]
(34)

For \( z < x < y \), again, by convexity, we write
\[
\phi(x) \leq \frac{x - z}{y - z} \phi(y) + \frac{y - x}{y - z} \phi(z) = \frac{x - z}{y - z} \phi(y) \leq 0.
\]
As a result, \( |y - z|^p |\phi(x)|^p \geq |x - z|^p |\phi(y)|^p \). Integrating from \( x = z \) to \( x = y \), we get
\[
|\phi(y)|^p \frac{(y - z)^{p+1}}{p+1} \leq |y - z|^p \int_z^y |\phi(x)|^p dx.
\]
(35)

Adding the two inequalities (34) and (35), we obtain
\[
\frac{|\phi(y)|^p}{p+1} \left( z^{p+1} + (y - z)^{p+1} \right) \leq |y - z|^p \int_0^y |\phi(x)|^p dx < y^p.
\]

Now
\[
z^{p+1} + (y - z)^{p+1} \geq \min_{0 < u < y} \left( u^{p+1} + (y - u)^{p+1} \right) = 2^{-p}y^{p+1}.
\]
Combining, we obtain
\[
|\phi(y)| < 2(1 + p)^{1/p}y^{-1/p}
\]
which results in a contradiction if \( c \geq 2(1 + p)^{1/p}y^{-1/p} \).
Lemma 8.4 (Interpolation Lemma). Fix $x_1 < x_2 < \cdots < x_n$ and suppose that $c_1 \leq n(x_i - x_{i-1}) \leq c_2$ for all $2 \leq i \leq n$. For every function $f$ on $[x_1, x_n]$, associate another function $\tilde{f}$ on $[x_1, x_n]$ by

$$\tilde{f}(x) := \frac{x_{i+1} - x}{x_{i+1} - x_i} f(x_i) + \frac{x - x_i}{x_{i+1} - x_i} f(x_{i+1}) \quad \text{for } x_i \leq x \leq x_{i+1}$$

where $i = 1, \ldots, n-1$. Then for every pair of functions $f$ and $g$ on $[x_1, x_n]$, we have

$$\frac{1}{c_2} \int_{x_1}^{x_n} \left( \tilde{f}(x) - \tilde{g}(x) \right)^2 \, dx \leq \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \leq \frac{6}{c_1} \int_{x_1}^{x_n} \left( \tilde{f}(x) - \tilde{g}(x) \right)^2 \, dx.$$

Proof. It is elementary to check that for every $1 \leq i \leq n-1$, we have

$$\int_{x_i}^{x_{i+1}} \left( \tilde{f}(x) - \tilde{g}(x) \right)^2 \, dx = \frac{x_{i+1} - x_i}{3} (\alpha^2 + \beta^2 + \alpha \beta)$$

where $\alpha := f(x_i) - g(x_i)$ and $\beta = f(x_{i+1}) - g(x_{i+1})$. Using the inequalities

$$-\frac{\alpha^2 - \beta^2}{2} \leq \alpha \beta \leq \frac{\alpha^2 + \beta^2}{2},$$

we obtain

$$\frac{c_1(\alpha^2 + \beta^2)}{6n} \leq (x_{i+1} - x_i) \frac{\alpha^2 + \beta^2}{6} \leq \int_{x_i}^{x_{i+1}} \left( \tilde{f}(x) - \tilde{g}(x) \right)^2 \, dx \leq (x_{i+1} - x_i) \frac{\alpha^2 + \beta^2}{2} \leq \frac{c_2(\alpha^2 + \beta^2)}{2n}.$$

Adding these inequalities from $i = 1$ to $i = n-1$, we deduce

$$\frac{c_1}{6n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \leq \int_{x_1}^{x_n} \left( \tilde{f}(x) - \tilde{g}(x) \right)^2 \, dx \leq \frac{c_2}{n} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2$$

which yields the desired result. \qed

Remark 8.1. Observe that if $f$ is a convex function on $[a, b]$, then $\tilde{f}$ is also convex on $[a, b]$.

Lemma 8.5 (A simple integration lemma). Suppose $f : (0, \infty) \to (0, \infty)$ is a non-increasing function satisfying

$$f(\epsilon) \leq A \left( \frac{\epsilon}{B} \right)^{-1/2} \quad \text{whenever } \epsilon \leq uB$$

for some positive constants $A, B$ and $u$. Then for every $r \leq B$, we have

$$\int_{0}^{r} \sqrt{f(\epsilon)} \, d\epsilon \leq \sqrt{AB^{1/4}r^{3/4}} \left( \frac{4}{3} + u^{-1/4} \right).$$
Proof. If \( r \leq uB \), then clearly
\[
\int_0^r \sqrt{f(\epsilon)} \, d\epsilon \leq \int_0^r \sqrt{A} \epsilon^{-1/4} B^{1/4} \, d\epsilon = \frac{4}{3} \sqrt{AB^{1/4}} r^{3/4}.
\]
If \( r > uB \), then because \( f \) is non-increasing, we have
\[
\int_0^r \sqrt{f(\epsilon)} \, d\epsilon \leq \int_0^{uB} \sqrt{A} \epsilon^{-1/4} B^{1/4} \, d\epsilon + \sqrt{f(AB)} (r - uB)
\]
\[
= \frac{4}{3} \sqrt{AB^{1/4}} (uB)^{3/4} + \sqrt{f(AB)} (r - uB)
\]
\[
\leq \frac{4}{3} \sqrt{AB^{1/4}} r^{3/4} + r \sqrt{f(AB)}
\]
\[
\leq \frac{4}{3} \sqrt{AB^{1/4}} r^{3/4} + \sqrt{Au^{-1/4}} r.
\]
Because \( r \leq B \), we have \( r \leq B^{1/4} r^{3/4} \). Using this in the second term above completes the proof. \( \square \)

Lemma 8.6. The set of all convex projections of a concave function \( f_0 \) includes an affine function.

Proof. We prove this result by the method of contradiction. Suppose that there is no convex projection that is affine. Let \( \phi_0 \) be the continuous piecewise affine convex projection of \( f_0 \). For a function \( g : [0, 1] \rightarrow \mathbb{R} \) we define \( g(0+) := \lim_{x \rightarrow 0^+} g(x) \) and \( g(1-) := \lim_{x \rightarrow 1^-} g(x) \). This notation is necessary as \( f_0 \) need not be continuous at the boundary points \( \{0, 1\} \).

Case (i): Suppose that \( f_0(0+) \geq \phi_0(0) \) and \( f_0(1-) \geq \phi_0(1) \). Then the affine function \( \tilde{\phi}_0 \) obtained by joining \((0, \phi_0(0))\) and \((1, \phi_0(1))\), i.e., \( \tilde{\phi}_0(x) = (1 - x) \phi_0(0) + x \phi_0(1) \), for \( x \in [0, 1] \), lies in-between \( \phi_0 \) and \( f_0 \) (as \( f_0 \) is concave) and \( \ell^2(\phi_0, f_0) \geq \ell^2(\tilde{\phi}_0, f_0) \), giving rise to a contradiction.

Case (ii): Suppose that \( f_0(0+) < \phi_0(0) \) and \( f_0(1-) \geq \phi_0(1) \). Then there is a point \( u \in (0, 1) \) such that \( f_0(u) = \phi_0(u) \). Let us define \( \tilde{\phi} \) to be the affine function joining \((u, \phi_0(u))\) and \((1, \phi_0(1))\). Again, \( \tilde{\phi}_0 \) lies in-between \( \phi_0 \) and \( f_0 \) and \( \ell^2(\phi_0, f_0) \geq \ell^2(\tilde{\phi}_0, f_0) \), thus giving rise to a contradiction.

Case (iii): Suppose that \( f_0(0+) \geq \phi_0(0) \) and \( f_0(1-) < \phi_0(1) \). A similar analysis as in (ii) by looking at the affine function obtained by joining \((0, \phi_0(0))\) and \((v, \phi_0(v))\) where \( \phi_0(v) = f_0(v) \), \( v \in (0, 1) \), gives a contradiction.

Case (iv): Suppose that \( f_0(0+) < \phi_0(0) \) and \( f_0(1-) < \phi_0(1) \). Suppose that there are two points \( u_0, u_1 \in (0, 1) \) such that \( f_0(u_i) = \phi_0(u_i) \), for \( i = 1, 2 \). Then define \( \tilde{\phi} \) to be the affine function joining \((u_0, \phi_0(u_0))\) and \((u_1, \phi_0(u_1))\). Again, \( \tilde{\phi}_0 \) lies in-between \( \phi_0 \) and \( f_0 \) and \( \ell^2(\phi_0, f_0) \geq \ell^2(\tilde{\phi}_0, f_0) \), thus giving rise to a contradiction. Suppose that \( f_0 \) and \( \phi_0 \) touch at just one point \( v \in (0, 1) \). Then defining \( \tilde{\phi}_0 \) to be the affine function that passes through \((v, \phi_0(v))\) and is a sub-gradient to both \( \phi_0 \) and \( f_0 \) at \( v \) yields
a contradiction. If \( f_0 \) and \( \phi_0 \) do not touch at all then defining \( \tilde{\phi}_0 \) to be any affine function lying between \( \phi_0 \) and \( f_0 \) shows that \( \ell^2(\phi_0, f_0) \geq \ell^2(\tilde{\phi}_0, f_0) \). This completes the proof.

**Remark 8.2.** Note that if \( n > 2 \), the convex projection of a concave \( f_0 \) is in fact unique on \((0, 1)\) and affine.

## References

Alexandrov, A. D. (1939). Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. *Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser.*, 6:3–35.

Atkinson, K. E. (1989). *An introduction to numerical analysis*. John Wiley & Sons Inc., New York, second edition.

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. (1989). The Grenander estimator: a nonasymptotic approach. *Ann. Statist.*, 17(4):1532–1549.

Bronshtein, E. M. (1976). \( \epsilon \)-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17:393–398.

Cai, T. and Low, M. (2011). A framework for estimation of convex functions. available at http://www-stat.wharton.upenn.edu/~tcai/.

Carolan, C. and Dykstra, R. (1999). Asymptotic behavior of the Grenander estimator at density flat regions. *Canad. J. Statist.*, 27(3):557–566.

Cule, M. L., Samworth, R. J., and Stewart, M. I. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density (with discussion). *Journal of the Royal Statistical Society, Series B*, 72:545–600.

Dryanov, D. (2009). Kolmogorov entropy for classes of convex functions. *Constructive Approximation*, 30:137–153.

Dümbgen, L., Freitag, S., and Jongbloed, G. (2004). Consistency of concave regression with an application to current-status data. *Math. Methods Statist.*, 13(1):69–81.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.*, 78(384):837–842.

Fraser, D. A. S. and Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to concave regression. *Scand. J. Statist.*, 16(1):65–74.
Grenander, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957).

Groeneboom, P., Jongbloed, G., and Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698.

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). Covering numbers for convex functions. *IEEE Trans. Inf. Th.*, 59(4):1957–1965.

Hanson, D. L. and Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.*, 4(6):1038–1050.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.*, 49:598–619.

Jankowski, H. K. and Wellner, J. A. (2012). Convergence of linear functionals of the grenander estimator under misspecification. available at http://http://arxiv.org/abs/1207.6614.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.*, 19(2):741–759.

Mendelson, S. and Vershynin, R. (2003). Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Hayward, CA.

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

Seregin, A. and Wellner, J. A. (2010). Nonparametric estimation of multivariate convex-transformed densities. *Annals of Statistics*, 38:3751–3781.

Stark, H. and Yang, Y. (1988). *Vector space projections*. Wiley, New York.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21(1):14–44.

Van de Geer, S. (2000). *Applications of Empirical Process Theory*. Cambridge University Press.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Zhang, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.*, 30(2):528–555.