Chapter 5

Lower bounds for estimation

5.1 Minimax lower bounds for estimation

After studying the fundamental possibilities and limitations in various testing problems in the white noise model in Chapter 4 we return to estimation problems in this chapter. In Chapter 3 we introduced and studied various regularization methods for nonparametric estimation of the signal f from observations $X = (X_t : t \in [0,1])$ satisfying

$$dX_t = f(t)\,dt + \frac{1}{\sqrt{n}} dW_t$$

We found in particular that under the assumption that f is β-smooth in Hölder or Sobolev sense, there exist estimators that, perhaps up to a logarithmic factor, achieve a rate of convergence of the order $n^{-\beta/(1+2\beta)}$ relative to the pointwise, supremum or L^2-norm. Moreover, such results hold uniformly over Hölder or Sobolev balls. Specifically, we have the upper bounds (3.12), (3.13), (3.14) for appropriately tuned kernel or projection estimators.

The cited rate of convergence results are all risk upper bounds of the form

$$\sup_{f \in \mathcal{F}} \mathrm{E}_f \ell(\hat{f}, f) \leq r_n,$$

where $\hat{f}$ is a specific estimator, $\mathcal{F} \subset L^2[0,1]$ is a class of signals of interest, ℓ is a choice of *loss function*, such as pointwise loss, uniform loss, or squared L^2-loss, and r_n is a sequence converging to 0 as $n \to \infty$. Such results quantify how well a specific estimator performs for a certain class of truths. As we have seen, the bounds typically depend both on the class $\mathcal{F}$ and on the type of loss ℓ that is considered.

In this chapter we consider the converse question. We ask what the best possible performance of *any* estimator is over a given class of signals $\mathcal{F}$. More precisely, we are interested in proving so-called *minimax lower bounds* for estimation of the form

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{E}_f \ell(\hat{f}, f) \geq r_n,$$

where ℓ is a again a loss function and the infimum is over all possible estimators. Such a result quantifies the fundamental difficulty of estimating a signal in the class $\mathcal{F}$, as it implies that no estimator can have a risk of smaller order than r_n, uniformly over the class $\mathcal{F}$. We will study in particular the converse of the results (3.12), (3.13), (3.14) to investigate whether these estimator are rate-optimal in the minimax sense.

5.2 Connection to multiple hypotheses testing problems

The problem of estimating a signal in the white noise model is clearly closely related to the problem of testing multiple hypotheses as studied in Section 4.4. If the class of signals $\mathcal{F}$ of interest is finite the problem is in fact exactly the same. But also if $\mathcal{F}$ is not finite the connection to testing provides a useful approach to proving minimax lower bounds for estimation.

The following proposition make this precise. It assumes that the loss is of the form $\ell = d^p$, for $p > 0$ and d a semimetric on $\mathcal{F} \subset L^2[0,1]$ (i.e. a map $d : \mathcal{F} \times \mathcal{F} \to [0,\infty)$ that is symmetric and that satisfies the triangle inequality, but not necessarily $d(f,g) = 0$ implies $f = g$ for all $f, g \in \mathcal{F}$).

Proposition 5.2.1. *Consider a class $\mathcal{F} \subset L^2[0,1]$ a semimetric d on $\mathcal{F}$ and $p > 0$. If there exist $f_0, f_1, \dots, f_k \in \mathcal{F}$ such that*

(i) for all $i \neq j$, we have $d(f_i, f_j) \geq 2r_n$,

(ii) there exists no consistent test for the multiple hypotheses $H_0 : f = f_0$, $H_1 : f = f_1$, …, $H_k : f = f_k$,

then there exists a constant $c > 0$, independent of n, such that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{E}_f d^p(\hat{f}, f) \geq c r_n^p. \tag{5.1}$$

Proof. First let $\hat{f}$ be an arbitrary $\mathcal{F}$-valued estimator. Consider the $\{0, 1, \dots, k\}$-valued test that selects the f_i that is closest to the estimator with respect to the loss ℓ, i.e.

$$\varphi = \operatorname*{argmin}_i d(\hat{f}, f_i),$$

where in the case of multiple minima we select an arbitrary minimizer. By assumption (i) and the triangle inequality it holds for every i that

$$\mathrm{P}_{f_i}(\varphi \neq i) \leq \mathrm{P}_{f_i}(d(\hat{f}, f_i) \geq r_n),$$

and hence

$$\sup_{f \in \mathcal{F}} \mathrm{P}_f(d(\hat{f}, f) \geq r_n) \geq \max_i \mathrm{P}_{f_i}(\varphi \neq i) \geq \inf_{\varphi} \max_i \mathrm{P}_{f_i}(\varphi \neq i).$$

Then by assumption (ii) it follows that there exists a constant $c > 0$, independent of n, such that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{P}_f(d(\hat{f}, f) \geq r_n) \geq c.$$

The observation that

$$\mathrm{E}_f d^p(\hat{f}, f) \geq \mathrm{E}_f d^p(\hat{f}, f) 1_{d(\hat{f}, f) \geq r_n} \geq r_n^p \mathrm{P}_f(d(\hat{f}, f) \geq r_n)$$

completes the proof. □

Note that the proof of the proposition shows that the power function $x \mapsto x^p$ appearing on both sides of (5.1) may actually be replaced by any increasing function $w : [0, \infty) \to [0, \infty)$ such that $w(0) = 0$. We will however only apply the proposition in the form (5.1), with $p = 1$ or $p = 2$.

If we combine the preceding proposition with the sufficient condition for non-existence of a consistent test given by Proposition 4.4.2 we obtain the following useful result.

Proposition 5.2.2. *Suppose that there exists a finite subset $\mathcal{G}_n \subset \mathcal{F}$ whose elements that are $2r_n$-separated with respect to d, with cardinality $|\mathcal{G}_n|$ and L^2-diameter $diam(\mathcal{G}_n)$. If for some $C > 0$,*

$$e^{\frac{1}{2} n diam^2(\mathcal{G}_n)} \leq C|\mathcal{G}_n|$$

for n large enough, then for some constant $c > 0$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{E}_f d^p(\hat{f}, f) \geq c r_n^p$$

for all $p > 0$.

As a first example we consider classes of signals smoothly parametrized by a Euclidean parameter. As expected the minimax rate is of the order $1/\sqrt{n}$ in that case.

Example 5.2.3 (Smooth parametric models)**.** Suppose $\mathcal{F} = \{f_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is a set with nonempty interior. Assume the model is smooth in the sense that for every $\theta \in \Theta$, there exist constants $0 < c_\theta \leq C_\theta < \infty$ such that

$$c_\theta \|\theta - \psi\| \leq \|f_\theta - f_\psi\|_2 \leq C_\theta \|\theta - \psi\| \tag{5.2}$$

for all ψ in a neighborhood of θ. We are interested in a lower bound for estimators $\hat{\theta}$ relative to the Euclidean distance, so we consider the semimetric $d(f_\theta, f_\psi) = \|\theta - \psi\|$.

Now let θ_0 be the center of a ball that is entirely contained in Θ. Let θ_1 be an element at Euclidean distance $2/(c_{\theta_0}\sqrt{n})$ of θ_0, so that f_{θ_0} and f_{θ_1} are $2/\sqrt{n}$-separated with respect to d. Consider the finite set $\mathcal{G} = \{f_{\theta_0}, f_{\theta_1}\}$. We have $|\mathcal{G}| = 2$ and by the smoothness condition, the L^2-diameter of $\mathcal{G}$ is bounded by $2C_{\theta_0}/(c_{\theta_0}\sqrt{n})$. Hence, the condition of Proposition 5.2.2 is fulfilled for $r_n = 1/\sqrt{n}$ and C large enough. We conclude that for some $c > 0$,

$$\inf_{\hat{f}} \sup_{\theta\in\Theta} \mathrm{E}_\theta \|\hat{\theta} - \theta\|^p \geq \frac{c}{n^{p/2}}$$

for all $p > 0$. So we see that in smooth parametric models it is not possible to construct estimators that have a convergence rate that is better than $1/\sqrt{n}$, uniformly over the parameter space. ■

In parametric models in which the dependence on the parameter is not smooth such as in (5.2) it is possible to have convergence rates different from the usual $1/\sqrt{n}$. Exercise 5.2 considers an example in the context of the white noise model.

5.3 Minimax risk and Bayes risk

An alternative approach to obtaining minimax lower bounds for estimation starts from the simple but useful observation that the maximum risk of an estimator over a class of functions $\mathcal{F} \subset L^2[0,1]$ is lower bounded by the Bayes risk of that estimator relative to any prior Π on $\mathcal{F}$. Indeed, assuming the appropriate measurability, we have

$$\sup_{f\in\mathcal{F}} \mathrm{E}_f \ell(\hat{f}, f) \geq \int_{\mathcal{F}} \mathrm{E}_f \ell(\hat{f}, f)\, \Pi(df).$$

The estimator minimizing the Bayes risk on the right is, by definition, the Bayes estimator $\tilde{f}$ with respect to the loss ℓ and the prior Π. Hence, we have the bound

$$\inf_{\hat{f}} \sup_{f\in\mathcal{F}} \mathrm{E}_f \ell(\hat{f}, f) \geq \int_{\mathcal{F}} \mathrm{E}_f \ell(\tilde{f}, f)\, \Pi(df) \tag{5.3}$$

for the minimax risk. For every estimator $\hat{f} = \hat{f}(X)$ the Bayes risk can be written as

$$\begin{aligned}\int_{\mathcal{F}} \mathrm{E}_f \ell(\hat{f}(X), f)\, \Pi(df) &= \int_{\mathcal{F}} \int \ell(\hat{f}(x), f)\, P_f(dx)\Pi(df) \\ &= \int \int_{\mathcal{F}} \ell(\hat{f}(x), f)\, \Pi(df \,|\, x) P(dx),\end{aligned}$$

where $\Pi(df \,|\, x)$ is the posterior distribution and P is the marginal distribution of X in the Bayesian setup in which $f \sim \Pi$ and $X \,|\, f \sim \mathrm{P}_f$. The inner integral in the last line of the display typically has a unique minimizer for every fixed

x, which is then necessarily the Bayes estimate relative to the loss ℓ. So in that case the Bayes estimator is given by

$$\tilde{f} = \operatorname*{argmin}_{g \in \mathcal{F}} \int_{\mathcal{F}} \ell(g, f)\, \Pi(df \mid X) \tag{5.4}$$

For certain combinations of loss functions ℓ and priors Π the Bayes estimator can be explicitly determined in this way and the lower bound (5.3) can be used to obtain a bound for the minimax risk. We can for instance give an alternative proof of Proposition 5.2.2 using this approach.

Alternative proof of Proposition 5.2.2. Consider the loss function $\ell(f, g) = 1_{d(f,g) \geq r_n}$. As before, observe that for the risk of the estimator $\hat{f}$ we have

$$\mathrm{E}_f d^p(\hat{f}, f) \geq r_n^p \mathrm{E}_f \ell(\hat{f}, f).$$

Hence in view of (5.3) is suffices to show that for a particular prior Π on $\mathcal{F}$, it holds that the Bayes risk

$$\int_{\mathcal{F}} \mathrm{E}_f \ell(\tilde{f}, f)\, \Pi(df) \tag{5.5}$$

is bounded away from 0, with $\tilde{f}$ the Bayes estimator for the loss ℓ.

Denote the functions in $\mathcal{G}_n \subset \mathcal{F}$ by $f_1, f_2, \ldots, f_k$, with $k = |\mathcal{G}_n|$. We consider a discrete prior Π on $\mathcal{F}$ that assigns mass π_j to the function f_j for $j = 1, 2, \ldots, k$. The posterior is then a discrete measure on the same points, with f_j receiving posterior mass proportional to $\pi_j p_{f_j}(X)$. Moreover, we see from (5.4) that the Bayes estimator $\tilde{f}$ for the loss ℓ is the (random) function in $\mathcal{G}_n$ which receives the most posterior mass (check!). Since the f_j are $2r_n$-separated, it follows that the Bayes risk (5.5) is given by

$$\sum \pi_j \mathrm{P}_{f_j}(d(\tilde{f}, f_j) \geq r_n) = \sum \pi_j \mathrm{P}_{f_j}\Big(\max_i \pi_i p_{f_i}(X) > \pi_j p_{f_j}(X) \Big).$$

Now we consider the particular prior weights

$$\pi_1 = \frac{1}{1 + (k-1)e^{-\frac{1}{2}nD^2}},$$
$$\pi_2 = \cdots = \pi_k = \frac{e^{-\frac{1}{2}nD^2}}{1 + (k-1)e^{-\frac{1}{2}nD^2}},$$

where $D = \max_{i,j} \|f_i - f_j\|$ is the L^2-diameter of $\mathcal{G}_n$. For this choice is it easily verified that the Bayes risk is bounded away from 0 under the conditions of the proposition (Exercise 5.3). □

We note that this proof only depends on the specific aspects of the white noise model in a few places. The crucial elements that make the proof work

are first of all the existence a "large" set of well-separated elements f_j in the parameter space that are "difficult" to estimate and second the ability to appropriately bound the tail probability of the likelihood ratio $(p_{f_i}/p_{f_j})(X)$ under the measure P_{f_j}. Using the same approach we can also obtain minimax lower bounds in other statistical settings, as we illustrate in Section 5.5.

5.4 Lower bounds for estimating smooth functions

In this section we use the developed machinery to prove the converse of the results (3.12), (3.13) and (3.14). The theorems below imply that the rates obtained in these results are optimal. Appropriately tuned kernel estimators are rate-optimal with respect to pointwise and uniform risk over Hölder balls and an appropriately tuned projection estimator is rate-optimal with respect to the L^2-norm, uniformly over a Sobolev ball with a given regularity.

Theorem 5.4.1 (Minimax lower bound over Hölder balls for pointwise risk)**.** *Let $t \in (0,1)$ be fixed. For $\beta, R, p > 0$ we have*

$$\inf_{\hat{f}} \sup_{f \in C_R^\beta[0,1]} \mathrm{E}_f |\hat{f}(t) - f(t)|^p \gtrsim n^{-p\beta/(1+2\beta)},$$

where the infimum is over all estimators.

Proof. Consider the semimetric $d(f,g) = |f(t) - g(t)|$ on $C_R^\beta[0,1]$. We know from (a minor adaptation of) Example 4.2.4 that for every $\sigma > 0$ there exists a function $f_\sigma \in C_R^\beta[0,1]$ with $f(t)$ of the order σ^β and squared L^2-norm of the order $\sigma^{1+2\beta}$. Now for $\sigma = n^{-1/(1+2\beta)}$ consider the set $\mathcal{G} = \{0, f_\sigma\}$. We have $d(0, f_\sigma) = |f_\sigma(t)| \sim n^{-\beta/(1+2\beta)}$, so the functions in $\mathcal{G}$ are $2cn^{-\beta/(1+2\beta)}$-separated with respect to d, for some constant $c > 0$. Since the number of elements of $\mathcal{G}$ is 2 and the L^2-diameter is of the order $1/\sqrt{n}$, an application of Proposition 5.2.2 proves the statement of the theorem. □

A similar argument proves the following minimax lower bound relative to the uniform distance.

Theorem 5.4.2 (Minimax lower bound over Hölder balls for uniform risk)**.** *Let $0 < u < v < 1$ be fixed. For $\beta, R, p > 0$ we have*

$$\inf_{\hat{f}} \sup_{f \in C_R^\beta[0,1]} \mathrm{E}_f \sup_{u<t<v} |\hat{f}(t) - f(t)|^p \gtrsim \Big(\frac{n}{\log n}\Big)^{-p\beta/(1+2\beta)},$$

where the infimum is over all estimators.

Proof. Exercise 5.4. □

To prove a minimax lower bound that complements the upper bound (3.14) we want to follow the same strategy as above, using Proposition 5.2.2 in combination with the examples from Chapter 4 of functions in a Sobolev ball that are difficult to distinguish. A complication is that in this case the functions that are for instance constructed in Example 4.4.3 do not fulfil the separation condition required by Proposition 5.2.2. To remedy this we will consider smaller collections of functions, obtained by deleting functions from the class constructed in Example 4.4.3 that are too close together in L^2-sense.

The following result will ensure that we will still have a sufficiently large class of separated "difficult" functions. The lemma is a so-called *packing bound* for the space $\{-1,1\}^k$ of k-dimensional vectors of signs, endowed with the Hamming distance. In general, an *ε-packing* of a metric space is a collection of disjoint balls of radius ε in that space. The associated *ε-packing number* is the maximum cardinality of an ε-packing, i.e. the maximum number of disjoint balls that fit into the space. This number somehow measures how "massive", or "complex" the space is.

Lemma 5.4.3 (Varshamov-Gilbert packing bound for vectors of signs)**.** *Let $k \geq 8$ be a natural number. There exists a subset of $\{-1,1\}^k$ whose elements are $k/8$-separated for the Hamming distance and whose cardinality is larger than a^k for a universal constant $a > 1$.*

Proof. Let $S = \{-1,1\}^k$ and let d_{ham} be the Hamming distance on S. Fix a point $s_1 \in S$ and consider the set

$$S_1 = \{s \in S : d_{\text{ham}}(s, s_1) > k/8\}.$$

Next, fix a point $s_2 \in S_1$ and define

$$S_2 = \{s \in S_1 : d_{\text{ham}}(s, s_2) > k/8\}.$$

This procedure is repeated m times, until the set S_m is empty.

By construction the points $s_1, s_2, \ldots, s_m$ are $k/8$-separated with respect to d_{ham}. To determine the number m, let $n_j =$ the number of points that are removed in step j to construct the set S_j from S_{j-1}. Clearly $n_1+n_2+\cdots+n_m = 2^k$. Moreover, every n_j is bounded by the maximal number of points in a Hamming ball of radius $k/8$, hence

$$n_j \leq \sum_{i \leq k/8} \binom{k}{i}.$$

Together, we obtain the bound

$$2^k \leq m \sum_{i \leq k/8} \binom{k}{i}.$$

The last inequality can be written as $\mathrm{P}(X \leq k/8) \geq 1/m$, where X is a binomial random variable with parameters $1/2$ and k. By Hoeffding's inequality,

$$\mathrm{P}(X \leq k/8) \leq e^{-\frac{9}{32}k} \tag{5.6}$$

(see Exercise 5.5). We conclude that $m \geq a^k$, with $a = \exp(9/32) \approx 1.32$. □

We can now prove the minimax lower bound that complements the upper bound (3.14), showing that an appropriately tuned projection estimator is indeed rate-optimal with respect to the L^2-norm, uniformly over a Sobolev ball with a given regularity.

Theorem 5.4.4 (Minimax lower bound over Sobolev balls for L^2-risk)**.** *For all $\beta, R, p > 0$ we have*

$$\inf_{\hat{f}} \sup_{f \in H^\beta_R[0,1]} \mathrm{E}_f \|\hat{f} - f\|_2^p \gtrsim n^{-p\beta/(1+2\beta)},$$

where the infimum is over all estimators.

Proof. In Example 4.4.3 we constructed for every $\varepsilon > 0$ a set of 2^k functions f_θ, $\theta \in \{-1,1\}^k$, for $k = n^{-1/(1+2\beta)}$, satisfying

$$\|f_\theta - f_\psi\|_2^2 = 4\varepsilon^2 k^{-1-2\beta} d_{\mathrm{ham}}(\theta, \psi)$$

and

$$\|f_\theta\|_{H^\beta}^2 = \text{const.} \times \varepsilon^2.$$

By Lemma 5.4.3 there we can find a subset $\Theta \subset \{-1,1\}^k$ that is $k/8$-separated for the Hamming distance and whose cardinality is of the order at least a^k for some $a > 1$. Then for $\varepsilon > 0$ small enough the fucntions in the set $\mathcal{G}_n = \{f_\theta : \theta \in \Theta\}$ are contained in $H^\beta_R[0,1]$, are $2r_n$-separated for r_n a multiple of $n^{-\beta/(1+2\beta)}$, and the L^2-diameter of $\mathcal{G}_n$ is bounded by a constant times $\varepsilon n^{-\beta/(1+2\beta)}$ (check!). It follows that all conditions of Proposition 5.2.2 are fulfilled if $\varepsilon > 0$ is small enough. □

5.5 Minimax lower bounds in other statistical settings

The ideas and techniques we have developed in this chapter are not only useful for deriving minimax lower bounds in the specific context of the white noise model. One relatively straightforward example is considered in Exercise 5.6, which asks to derive a minimax lower bound for a nonparametric regression model. As another illustration, we consider the problem of estimating a sparse vector in the normal means model in this section.

Recall that in the normal means model we observe a vector Y satisfying $Y = \theta + \varepsilon$, where $\theta \in \mathbb{R}^n$ is the parameter of interest and $\varepsilon \sim N_n(0, I)$. We have seen in Example 3.4.5 and Exercise 3.18 that if the vector θ is assumed to be s-sparse and the degree of sparsity satisfies $s = o(n)$, then there exist estimators which have rate of convergence $\sqrt{s \log(n/s)}$ with respect to the Euclidean distance. Both the ℓ^0-penalized MLE and the lasso estimator achieve this rate, provided they are appropriately tuned. The theorem below gives a minimax lower bound which complements these results.

In the proof of the lower bound we use the following lemma, which gives a packing bound for the space of s-sparse vectors in $\mathbb{R}^n$. In the proof we use, as before, the notation $\|x\|_0 = |\{i : x_i \neq 0\}|$ for the ℓ^0-norm of a vector x, which is its number of non-zero coordinates. A vector x is s-sparse if $\|x\|_0 \leq s$.

Lemma 5.5.1 (Packing bound for sparse vectors)**.**

(i) Let $s \leq n$ be natural numbers and $p \in (0, 1)$. There exists a set of s-sparse vectors in the unit ball of $\mathbb{R}^n$ whose elements are p-separated for the Euclidean distance and whose cardinality is larger than $\exp(c_p s \log(n/C_p s))$ for constants

$$c_p = \frac{1 - p^2}{4}, \quad C_p = (1 - p^2)\Big(\frac{e}{p^2}\Big)^{2p^2/(1-p^2)}.$$

(ii) Let $r \in (0, 1)$. There exist constants $\varepsilon, c > 0$ such that if s and n are natural numbers satisfying $s \leq rn$, then there exists a set of s-sparse vectors in the unit ball of $\mathbb{R}^n$ whose elements are ε-separated for the Euclidean distance and whose cardinality is larger than $n \vee \exp(cs \log(n/s))$.

Proof. (i). Let S be the set of s-sparse vectors whose non-zero coordinates are all equal $1/\sqrt{s}$. Note that these vectors belong to the unit ball of $\mathbb{R}^n$. Let X and Y be independent random vectors in $\mathbb{R}^n$, uniformly distributed on S. Since $\|X - Y\|^2 = \|X - Y\|_0/s$, we have

$$\mathrm{P}(\|X - Y\| \leq p) = \mathrm{P}(\|X - Y\|_0 \leq p^2 s).$$

Now given X, we have (check!)

$$\begin{aligned}\mathrm{P}(\|X-Y\|_0 \le p^2 s \,|\, X) &= \frac{|\{y \in S : \|y-X\|_0 \le p^2 s\}|}{|S|} \\ &= \binom{n}{s}^{-1}\binom{s}{(1-p^2)s}\binom{n-(1-p^2)s}{p^2 s} \\ &= \binom{n}{(1-p^2)s}^{-1}\binom{s}{p^2 s}^2.\end{aligned}$$

Using the bounds $(n/k)^k \le \binom{n}{k} \le (ne/k)^k$ for binomial coefficients it follows that

$$\mathrm{P}(\|X-Y\|_0 \le p^2 s) \le \Big(\frac{e}{p^2}\Big)^{2p^2 s}\Big(\frac{(1-p^2)s}{n}\Big)^{(1-p^2)s} = \Big(C_p \frac{s}{n}\Big)^{(1-p^2)s}.$$

It follows that if $X_1, \dots, X_N$ are N independent vectors that are uniformly distributed on S, then

$$\mathrm{P}\Big(\|X_i - X_j\|_2 \le p \text{ for some } i \ne j\Big) \le N^2 e^{-(1-p^2)s \log \frac{n}{C_p s}}.$$

For $N = e^{c_p s \log \frac{n}{C_p s}}$ this is strictly less than 1, so that

$$\mathrm{P}\Big(\|X_i - X_j\|_2 > p \text{ for all } i \ne j\Big) > 0.$$

Since the distribution of $(X_1, \dots, X_N)$ is discrete, this implies that there exists at least one realization consisting of p-separated vectors.

(ii). Follows from part (ii), see Exercise 5.7. □

Theorem 5.5.2 (Minimax lower bound for sparse signals in the normal means model)**.** *Let $r \in (0,1)$ and let s, n be natural numbers such that $n \ge 2$ and $s \le rn$. Then for the normal means model we have, for every $p > 0$,*

$$\inf_{\hat\theta} \sup_{\theta : \|\theta\|_0 \le s} \mathrm{E}_\theta \|\hat\theta - \theta\|^p \gtrsim \Big(s \log \frac{n}{s}\Big)^{p/2},$$

where the infimum is over all estimators.

Proof. Fix $\delta > 0$. By multiplying the vectors exhibited in part (ii) of the preceding lemma by δ we see that there exist $N \ge 2 \vee \exp(cs \log(n/s))$ s-sparse vectors $\theta_1, \theta_2, \dots, \theta_N$ with L^2-norm δ which are $2\varepsilon\delta$-separated for some $\varepsilon, c > 0$. Let Π be a discrete prior on the set of s-sparse vectors which assigns prior probability π_j to θ_j for every j. Then by following the line of reasoning in the alternative proof of Proposition 5.2.2 in Section 5.3, we see that to lower

bound the minimax risk by a constant times a particular choice of δ, it suffices to show that for that δ and for an appropriate choice Π,

$$\sum \pi_j \mathrm{P}_{\theta_j}\Big(\max_i \pi_i p_{\theta_i}(Y) > \pi_j p_{\theta_j}(Y)\Big) \tag{5.7}$$

is bounded away from 0, where $p_\theta(Y)$ is the likelihood in the model. As before we use the simple lower bound

$$\mathrm{P}_{\theta_j}\Big(\max_i \pi_i p_{\theta_i}(Y) > \pi_j p_{\theta_j}(Y)\Big) \geq \mathrm{P}_{\theta_j}\Big(\frac{p_{\theta_1}}{p_{\theta_j}}(Y) > \frac{\pi_j}{\pi_1}\Big).$$

Observe that for every j, the likelihood ratio satisfies

$$\frac{p_{\theta_1}}{p_{\theta_j}}(Y) = e^{\langle \theta_1-\theta_j, Y-\theta_j\rangle - \frac{1}{2}\|\theta_1-\theta_j\|^2} =_d e^{\|\theta_1-\theta_j\|Z - \frac{1}{2}\|\theta_1-\theta_j\|^2},$$

where Z is standard normal under P_{θ_j} (check!). It follows that the sum (5.7) is bounded from below by

$$\sum_{j\neq 1} \pi_j\Big(1 - \Phi\Big(\frac{\log \pi_j/\pi_1 + \frac{1}{2}\|\theta_1-\theta_j\|^2}{\|\theta_1-\theta_j\|}\Big)\Big).$$

For all j we have that $\|\theta_1 - \theta_j\| \leq 2\delta$, by the triangle inequality. Hence if we consider the prior given by

$$\pi_1 = \frac{1}{1+(N-1)e^{-2\delta^2}},$$
$$\pi_2 = \cdots = p_k = \frac{e^{-2\delta^2}}{1+(N-1)e^{-2\delta^2}},$$

then (5.7) is further lower bounded by $1/2$ times

$$\frac{(N-1)e^{-2\delta^2}}{1+(N-1)e^{-2\delta^2}}.$$

Since $N \geq 2$, this is bounded away from 0 as soon as $Ne^{-2\delta^2}$ is bounded away from 0. Using the other lower bound on N we see that is the case for the choice

$$\delta = \sqrt{\frac{1}{2}cs\log\frac{n}{s}}.$$

This completes the proof of the theorem. □

5.6 Exercises

Exercise 5.1 (Minimax risk over large classes of signals)**.**

(i) Prove that for every $p > 0$,

$$\inf_{\hat{f}} \sup_{f \in L^2[0,1]} \mathrm{E}_f \|\hat{f} - f\|_2^p \to \infty.$$

(ii) Let $L_1^2[0,1]$ be the unit ball in $L^2[0,1]$, i.e. $L_1^2[0,1] = \{f \in L^2[0,1] : \|f\|_2 \leq 1\}$. Prove that for every $p > 0$, there exists a constant $c > 0$ such that

$$\inf_{\hat{f}} \sup_{f \in L_1^2[0,1]} \mathrm{E}_f \|\hat{f} - f\|_2^p \geq c.$$

(iii) Recall that a subset of a metric space is called totally bounded if for all $\varepsilon > 0$ it can be covered by finitely many balls of radius ε.

Let $\mathcal{F} \subset L^2[0,1]$ be a set of that is bounded but not totally bounded. Prove that also in this case it holds that for every $p > 0$, there exists a constant $c > 0$ such that

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{E}_f \|\hat{f} - f\|_2^p \geq c.$$

What do these statements mean in terms of convergence rates?

Exercise 5.2 (Minimax estimation in a non-regular parametric model)**.** Consider the (parametric) model

$$dX_t = \operatorname{sign}(t - \theta)\, dt + \frac{1}{\sqrt{n}}\, dW_t,$$

where $\operatorname{sign}(x) = 1_{x \geq 0} - 1_{x < 0}$ and $\theta \in (0,1)$ is an unknown parameter.

(i) Prove the minimax lower bound

$$\inf_{\hat{\theta}} \sup_{\theta \in (0,1)} \mathrm{E}_\theta |\hat{\theta} - \theta| \gtrsim \frac{1}{n}.$$

(ii) Describe the MLE $\hat{\theta}_{\mathrm{MLE}}$ in this model and show that it achieves the rate $1/n$, uniformly over $(0,1)$, i.e.

$$\sup_{\theta \in (0,1)} \mathrm{E}_\theta |\hat{\theta}_{\mathrm{MLE}} - \theta| \lesssim \frac{1}{n}.$$

(iii) Find the limiting distribution of $n(\hat{\theta}_{\mathrm{MLE}} - \theta)$.

Exercise 5.3 (Alternative proof of Proposition 5.2.2)**.** Complete the alternative proof of Proposition 5.2.2 given in Section 5.3.

Exercise 5.4 (Proof of Theorem 5.4.2)**.** Give a proof of Theorem 5.4.2.

Exercise 5.5 (Special case of Hoeffding's inequality)**.** Hoeffding's inequality implies that if $X_1, \ldots, X_k$ are independent Rademacher variables (± 1 with probability $1/2$) and $a_1, \ldots, a_j$ are real numbers, then for all $x > 0$,

$$\mathrm{P}\Big(\sum_{j=1}^{k} a_j X_j \geq x\Big) \leq e^{-\frac{1}{2}\frac{x^2}{\sum a_j^2}}.$$

(i) Prove this inequality. (Hint: first show that for every $\lambda > 0$ we have $\mathrm{E}\exp(\lambda X_j) \leq \exp(\lambda^2/2)$.)

(ii) Use the inequality to prove (5.6).

Exercise 5.6 (Minimax lower bound for nonparametric regression)**.** Consider the fixed design regression model in which we observe $Y_1, \ldots, Y_n$ satisfying the relation

$$Y_i = f(i/n) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent standard normal variables and $f : [0,1] \to \mathbb{R}$ is the function of interest. In this statistical setting, derive a (non-trivial) minimax lower bound for the for pointwise risk $\mathrm{E}_f(\hat{f}(t) - f(t))^2$ over the Hölder ball $C_1^\beta[0,1]$ for $\beta > 0$, where t is a fixed point in $(0,1)$.

Exercise 5.7 (ε-separated sparse vectors)**.** Verify the details of part (i) of the proof of Lemma 5.5.1 and provide a proof of part (ii).