15

Minimax lower bounds

In the preceding chapters, we have derived a number of results on the convergence rates of different estimation procedures. In this chapter, we turn to the complementary question: Can we obtain matching lower bounds on estimation rates? This question can be asked both in the context of a specific procedure or algorithm, and in an algorithm-independent sense. We focus on the latter question in this chapter. In particular, our goal is to derive lower bounds on the estimation error achievable by *any procedure*, regardless of its computational complexity and/or storage.

Lower bounds of this type can yield two different but complementary types of insight. A first possibility is that they can establish that known—and possibly polynomial-time—estimators are statistically "optimal", meaning that they have estimation error guarantees that match the lower bounds. In this case, there is little purpose in searching for estimators with lower statistical error, although it might still be interesting to study optimal estimators that enjoy lower computational and/or storage costs, or have other desirable properties such as robustness. A second possibility is that the lower bounds do not match the best known upper bounds. In this case, assuming that the lower bounds are tight, one has a strong motivation to study alternative estimators.

In this chapter, we develop various techniques for establishing such lower bounds. Of particular relevance to our development are the properties of packing sets and metric entropy, as discussed in Chapter 5. In addition, we require some basic aspects of information theory, including entropy and the Kullback–Leibler divergence, as well as other types of divergences between probability measures, which we provide in this chapter.

15.1 Basic framework

Given a class of distributions $\mathcal{P}$, we let θ denote a functional on the space $\mathcal{P}$—that is, a mapping from a distribution $\mathbb{P}$ to a parameter $\theta(\mathbb{P})$ taking values in some space Ω. Our goal is to estimate $\theta(\mathbb{P})$ based on samples drawn from the unknown distribution $\mathbb{P}$.

In certain cases, the quantity $\theta(\mathbb{P})$ uniquely determines the underlying distribution $\mathbb{P}$, meaning that $\theta(\mathbb{P}_0) = \theta(\mathbb{P}_1)$ if and only if $\mathbb{P}_0 = \mathbb{P}_1$. In such cases, we can think of θ as providing a parameterization of the family of distributions. Such classes include most of the usual finite-dimensional parametric classes, as well as certain nonparametric problems, among them nonparametric regression problems. For such classes, we can write $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Omega\}$, as we have done in previous chapters.

In other settings, however, we might be interested in estimating a functional $\mathbb{P} \mapsto \theta(\mathbb{P})$ that does *not* uniquely specify the distribution. For instance, given a class of distributions $\mathcal{P}$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

on the unit interval $[0, 1]$ with differentiable density functions f, we might be interested in estimating the quadratic functional $\mathbb{P} \mapsto \theta(\mathbb{P}) = \int_0^1 (f'(t))^2 \, dt \in \mathbb{R}$. Alternatively, for a class of unimodal density functions f on the unit interval $[0, 1]$, we might be interested in estimating the mode of the density $\theta(\mathbb{P}) = \arg\max_{x \in [0,1]} f(x)$. Thus, the viewpoint of estimating functionals adopted here is considerably more general than a parameterized family of distributions.

15.1.1 Minimax risks

Suppose that we are given a random variable X drawn according to a distribution $\mathbb{P}$ for which $\theta(\mathbb{P}) = \theta^*$. Our goal is to estimate the unknown quantity θ^* on the basis of the data X. An estimator $\widehat{\theta}$ for doing so can be viewed as a measurable function from the domain $\mathcal{X}$ of the random variable X to the parameter space Ω. In order to assess the quality of any estimator, we let $\rho : \Omega \times \Omega \to [0, \infty)$ be a semi-metric,[1] and we consider the quantity $\rho(\widehat{\theta}, \theta^*)$. Here the quantity θ^* is fixed but unknown, whereas the quantity $\widehat{\theta} \equiv \widehat{\theta}(X)$ is a random variable, so that $\rho(\widehat{\theta}, \theta^*)$ is random. By taking expectations over the observable X, we obtain the deterministic quantity $\mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta^*)]$. As the parameter θ^* is varied, we obtain a function, typically referred to as the risk function, associated with the estimator.

The first property to note is that it makes no sense to consider the set of estimators that are good in a pointwise sense. For any *fixed* θ^*, there is always a very good way in which to estimate it: simply ignore the data, and return θ^*. The resulting deterministic estimator has zero risk when evaluated at the fixed θ^*, but of course is likely to behave very poorly for other choices of the parameter. There are various ways in which to circumvent this and related difficulties. The Bayesian approach is to view the unknown parameter θ^* as a random variable; when endowed with some prior distribution, we can then take expectations over the risk function with respect to this prior. A closely related approach is to model the choice of θ^* in an adversarial manner, and to compare estimators based on their worst-case performance. More precisely, for each estimator $\widehat{\theta}$, we compute the worst-case risk $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))]$, and rank estimators according to this ordering. The estimator that is optimal in this sense defines a quantity known as the *minimax risk*—namely,

$$\mathfrak{M}(\theta(\mathcal{P}); \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))], \tag{15.1}$$

where the infimum ranges over all possible estimators, by which we mean measurable functions of the data. When the estimator is based on n i.i.d. samples from $\mathbb{P}$, we use $\mathfrak{M}_n$ to denote the associated minimax risk.

We are often interested in evaluating minimax risks defined not by a norm, but rather by a squared norm. This extension is easily accommodated by letting $\Phi : [0, \infty) \to [0, \infty)$ be an increasing function on the non-negative real line, and then defining a slight generalization of the ρ-minimax risk—namely

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta(\mathbb{P})))]. \tag{15.2}$$

[1] In our usage, a semi-metric satisfies all properties of a metric, except that there may exist pairs $\theta \neq \theta'$ for which $\rho(\theta, \theta') = 0$.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

A particularly common choice is $\Phi(t) = t^2$, which can be used to obtain minimax risks for the mean-squared error associated with ρ.

15.1.2 From estimation to testing

With this set-up, we now turn to the primary goal of this chapter: developing methods for lower bounding the minimax risk. Our first step is to show how lower bounds can be obtained via "reduction" to the problem of obtaining lower bounds for the probability of error in a certain testing problem. We do so by constructing a suitable packing of the parameter space (see Chapter 5 for background on packing numbers and metric entropy).

More precisely, suppose that $\{\theta^1, \ldots, \theta^M\}$ is a 2δ-separated set[2] contained in the space $\theta(\mathcal{P})$, meaning a collection of elements $\rho(\theta^j, \theta^k) \geq 2\delta$ for all $j \neq k$. For each θ^j, let us choose some representative distribution $\mathbb{P}_{\theta^j}$—that is, a distribution such that $\theta(\mathbb{P}_{\theta^j}) = \theta^j$—and then consider the M-ary hypothesis testing problem defined by the family of distributions $\{\mathbb{P}_{\theta^j}, \ j = 1, \ldots, M\}$. In particular, we generate a random variable Z by the following procedure:

(1) Sample a random integer J from the uniform distribution over the index set $[M] := \{1, \ldots, M\}$.
(2) Given $J = j$, sample $Z \sim \mathbb{P}_{\theta^j}$.

We let $\mathbb{Q}$ denote the joint distribution of the pair (Z, J) generated by this procedure. Note that the marginal distribution over Z is given by the uniformly weighted mixture distribution $\bar{\mathbb{Q}} := \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$. Given a sample Z from this mixture distribution, we consider the M-ary hypothesis testing problem of determining the randomly chosen index J. A *testing function* for this problem is a mapping $\psi : \mathcal{Z} \to [M]$, and the associated probability of error is given by $\mathbb{Q}[\psi(Z) \neq J]$, where the probability is taken jointly over the pair (Z, J). This error probability may be used to obtain a lower bound on the minimax risk as follows:

Proposition 15.1 (From estimation to testing) *For any increasing function Φ and choice of 2δ-separated set, the minimax risk is lower bounded as*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J], \tag{15.3}$$

where the infimum ranges over test functions.

Note that the right-hand side of the bound (15.3) involves two terms, both of which depend on the choice of δ. By assumption, the function Φ is increasing in δ, so that it is maximized by choosing δ as large as possible. On the other hand, the testing error $\mathbb{Q}[\psi(Z) \neq J]$ is defined in terms of a collection of 2δ-separated distributions. As $\delta \to 0^+$, the underlying testing problem becomes more difficult, and so that, at least in general, we should expect that $\mathbb{Q}[\psi(Z) \neq J]$ grows as δ decreases. If we choose a value δ^* sufficiently small to ensure

[2] Here we enforce only the milder requirement $\rho(\theta^j, \theta^k) \geq 2\delta$, as opposed to the strict inequality required for a packing set. This looser requirement turns out to be convenient in later calculations.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

that this testing error is at least $1/2$, then we may conclude that $\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta^*)$. For a given choice of δ, the other additional degree of freedom is our choice of packing set, and we will see a number of different constructions in the sequel.

We now turn to the proof of the proposition.

Proof For any $\mathbb{P} \in \mathcal{P}$ with parameter $\theta = \theta(\mathbb{P})$, we have

$$\mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta))] \stackrel{\text{(i)}}{\geq} \Phi(\delta)\, \mathbb{P}[\Phi(\rho(\widehat{\theta}, \theta)) \geq \Phi(\delta)] \stackrel{\text{(ii)}}{\geq} \Phi(\delta)\, \mathbb{P}[\rho(\widehat{\theta}, \theta) \geq \delta],$$

where step (i) follows from Markov's inequality, and step (ii) follows from the increasing nature of Φ. Thus, it suffices to lower bound the quantity

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\rho(\widehat{\theta}, \theta(\mathbb{P})) \geq \delta].$$

Recall that $\mathbb{Q}$ denotes the joint distribution over the pair (Z, J) defined by our construction. Note that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\rho(\widehat{\theta}, \theta(\mathbb{P})) \geq \delta] \geq \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}[\rho(\widehat{\theta}, \theta^j) \geq \delta] = \mathbb{Q}[\rho(\widehat{\theta}, \theta^J) \geq \delta],$$

so we have reduced the problem to lower bounding the quantity $\mathbb{Q}[\rho(\widehat{\theta}, \theta^J) \geq \delta]$.

Now observe that any estimator $\widehat{\theta}$ can be used to define a test—namely, via

$$\psi(Z) := \arg\min_{\ell \in [M]} \rho(\theta^\ell, \widehat{\theta}). \tag{15.4}$$

(If there are multiple indices that achieve the minimizing argument, then we break such ties in an arbitrary but well-defined way.) Suppose that the true parameter is θ^j: we then claim that the event $\{\rho(\theta^j, \widehat{\theta}) < \delta\}$ ensures that the test (15.4) is correct. In order to see this implication, note that, for any other index $k \in [M]$, an application of the triangle inequality

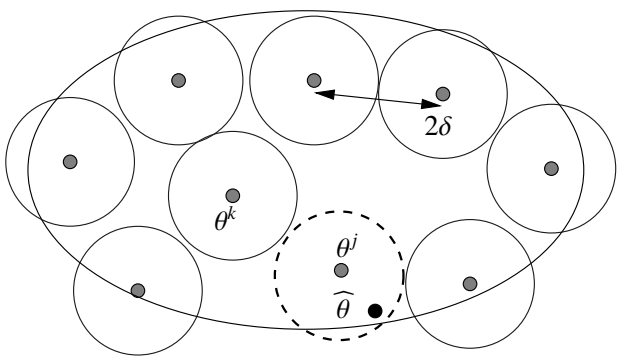

Figure 15.1 Reduction from estimation to testing using a 2δ-separated set in the space Ω in the semi-metric ρ. If an estimator $\widehat{\theta}$ satisfies the bound $\rho(\widehat{\theta}, \theta^j) < \delta$ whenever the true parameter is θ^j, then it can be used to determine the correct index j in the associated testing problem.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

guarantees that

$$\rho(\theta^k, \widehat{\theta}) \geq \underbrace{\rho(\theta^k, \theta^j)}_{\geq 2\delta} - \underbrace{\rho(\theta^j, \widehat{\theta})}_{< \delta} > 2\delta - \delta = \delta,$$

where the lower bound $\rho(\theta^j, \theta^k) \geq 2\delta$ follows by the 2δ-separated nature of our set. Consequently, we have $\rho(\theta^k, \widehat{\theta}) > \rho(\theta^j, \widehat{\theta})$ for all $k \neq j$, so that, by the definition (15.4) of our test, we must have $\psi(Z) = j$. See Figure 15.1 for the geometry of this argument.

Therefore, conditioned on $J = j$, the event $\{\rho(\widehat{\theta}, \theta^j) < \delta\}$ is contained within the event $\{\psi(Z) = j\}$, which implies that $\mathbb{P}_{\theta^j}[\rho(\widehat{\theta}, \theta^j) \geq \delta] \geq \mathbb{P}_{\theta^j}[\psi(Z) \neq j]$. Taking averages over the index j, we find that

$$\mathbb{Q}[\rho(\widehat{\theta}, \theta^J) \geq \delta] = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}[\rho(\widehat{\theta}, \theta^j) \geq \delta] \geq \mathbb{Q}[\psi(Z) \neq J].$$

Combined with our earlier argument, we have shown that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \Phi(\delta) \, \mathbb{Q}[\psi(Z) \neq J].$$

Finally, we may take the infimum over all estimators $\widehat{\theta}$ on the left-hand side, and the infimum over the induced set of tests on the right-hand side. The full infimum over all tests can only be smaller, from which the claim follows. □

15.1.3 Some divergence measures

Thus far, we have established a connection between minimax risks and error probabilities in testing problems. Our next step is to develop techniques for lower bounding the error probability, for which we require some background on different types of divergence measures between probability distributions. Three such measures of particular importance are the total variation (TV) distance, the Kullback–Leibler (KL) divergence and the Hellinger distance.

Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions on $\mathcal{X}$ with densities p and q with respect to some underlying base measure ν. Note that there is no loss of generality in assuming the existence of densities, since any pair of distributions have densities with respect to the base measure $\nu = \frac{1}{2}(\mathbb{P} + \mathbb{Q})$. The *total variation (TV) distance* between two distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} := \sup_{A \subseteq \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|. \tag{15.5}$$

In terms of the underlying densities, we have the equivalent definition

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, \nu(dx), \tag{15.6}$$

corresponding to one-half the $L^1(\nu)$-norm between the densities. (See Exercise 3.13 from Chapter 3 for details on this equivalence.) In the sequel, we will see how the total variation distance is closely connected to the Bayes error in binary hypothesis testing.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

A closely related measure of the "distance" between distributions is the *Kullback–Leibler divergence*. When expressed in terms of the densities q and p, it takes the form

$$D(\mathbb{Q} \,\|\, \mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \nu(dx), \tag{15.7}$$

where ν is some underlying base measure defining the densities. Unlike the total variation distance, it is not actually a metric, since, for example, it fails to be symmetric in its arguments in general (i.e., there are pairs for which $D(\mathbb{Q} \,\|\, \mathbb{P}) \neq D(\mathbb{P} \,\|\, \mathbb{Q})$). However, it can be used to upper bound the TV distance, as stated in the following classical result:

Lemma 15.2 (Pinsker–Csiszár–Kullback inequality) *For all distributions $\mathbb{P}$ and $\mathbb{Q}$,*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2} D(\mathbb{Q} \,\|\, \mathbb{P})}. \tag{15.8}$$

Recall that this inequality arose in our study of the concentration of measure phenomenon (Chapter 3). This inequality is also useful here, but instead in the context of establishing minimax lower bounds. See Exercise 15.6 for an outline of the proof of this bound.

A third distance that plays an important role in statistical problems is the *squared Hellinger distance*, given by

$$H^2(\mathbb{P} \,\|\, \mathbb{Q}) := \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \nu(dx). \tag{15.9}$$

It is simply the $L^2(\nu)$-norm between the square-root density functions, and an easy calculation shows that it takes values in the interval $[0, 2]$. When the base measure is clear from the context, we use the notation $H^2(p \,\|\, q)$ and $H^2(\mathbb{P} \,\|\, \mathbb{Q})$ interchangeably.

Like the KL divergence, the Hellinger distance can also be used to upper bound the TV distance:

Lemma 15.3 (Le Cam's inequality) *For all distributions $\mathbb{P}$ and $\mathbb{Q}$,*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{P} \,\|\, \mathbb{Q}) \sqrt{1 - \frac{H^2(\mathbb{P} \,\|\, \mathbb{Q})}{4}}. \tag{15.10}$$

We work through the proof of this inequality in Exercise 15.5.

Let $(\mathbb{P}_1, \ldots, \mathbb{P}_n)$ be a collection of n probability measures, each defined on $\mathcal{X}$, and let $\mathbb{P}^{1:n} = \bigotimes_{i=1}^n \mathbb{P}_i$ be the product measure defined on $\mathcal{X}^n$. If we define another product measure $\mathbb{Q}^{1:n}$ in a similar manner, then it is natural to ask whether the divergence between $\mathbb{P}^{1:n}$ and $\mathbb{Q}^{1:n}$ has a "nice" expression in terms of divergences between the individual pairs.

In this context, the total variation distance behaves badly: in general, it is difficult to

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

express the distance $\|\mathbb{P}^{1:n} - \mathbb{Q}^{1:n}\|_{\mathrm{TV}}$ in terms of the individual distances $\|\mathbb{P}_i - \mathbb{Q}_i\|_{\mathrm{TV}}$. On the other hand, the Kullback–Leibler divergence exhibits a very attractive decoupling property, in that we have

$$D(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = \sum_{i=1}^{n} D(\mathbb{P}_i \,\|\, \mathbb{Q}_i). \tag{15.11a}$$

This property is straightforward to verify from the definition. In the special case of i.i.d. product distributions—meaning that $\mathbb{P}_i = \mathbb{P}_1$ and $\mathbb{Q}_i = \mathbb{Q}_1$ for all i—then we have

$$D(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = nD(\mathbb{P}_1 \,\|\, \mathbb{Q}_1). \tag{15.11b}$$

Although the squared Hellinger distance does not decouple in quite such a simple way, it does have the following property:

$$\tfrac{1}{2}H^2(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = 1 - \prod_{i=1}^{n} \Big(1 - \tfrac{1}{2}H^2(\mathbb{P}_i \,\|\, \mathbb{Q}_i)\Big). \tag{15.12a}$$

Thus, in the i.i.d. case, we have

$$\tfrac{1}{2}H^2(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = 1 - \Big(1 - \tfrac{1}{2}H^2(\mathbb{P}_1 \,\|\, \mathbb{Q}_1)\Big)^n \leq \tfrac{1}{2}nH^2(\mathbb{P}_1 \,\|\, \mathbb{Q}_1). \tag{15.12b}$$

See Exercises 15.3 and 15.7 for verifications of these and related properties, which play an important role in the sequel.

15.2 Binary testing and Le Cam's method

The simplest type of testing problem, known as a binary hypothesis test, involves only two distributions. In this section, we describe the connection between binary testing and the total variation norm, and use it to develop various lower bounds, culminating in a general technique known as Le Cam's method.

15.2.1 Bayes error and total variation distance

In a binary testing problem with equally weighted hypotheses, we observe a random variable Z drawn according to the mixture distribution $\bar{\mathbb{Q}} := \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. For a given decision rule $\psi : \mathcal{Z} \to \{0, 1\}$, the associated probability of error is given by

$$\mathbb{Q}[\psi(Z) \neq J] = \tfrac{1}{2}\mathbb{P}_0[\psi(Z) \neq 0] + \tfrac{1}{2}\mathbb{P}_1[\psi(Z) \neq 1].$$

If we take the infimum of this error probability over all decision rules, we obtain a quantity known as the *Bayes risk* for the problem. In the binary case, the Bayes risk can actually be expressed explicitly in terms of the total variation distance $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}$, as previously defined in equation (15.5)—more precisely, we have

$$\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] = \tfrac{1}{2}\Big\{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\Big\}. \tag{15.13}$$

Note that the worst-case value of the Bayes risk is one-half, achieved when $\mathbb{P}_1 = \mathbb{P}_0$, so that the hypotheses are completely indistinguishable. At the other extreme, the best-case Bayes

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

risk is zero, achieved when $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}} = 1$. This latter equality occurs, for instance, when $\mathbb{P}_0$ and $\mathbb{P}_1$ have disjoint supports.

In order to verify the equivalence (15.13), note that there is a one-to-one correspondence between decision rules ψ and measurable partitions (A, A^c) of the space $\mathcal{X}$; more precisely, any decision rule ψ is uniquely determined by the set $A = \{x \in \mathcal{X} \mid \psi(x) = 1\}$. Thus, we have

$$\sup_{\psi} \mathbb{Q}[\psi(Z) = J] = \sup_{A \subseteq \mathcal{X}} \left\{ \tfrac{1}{2}\mathbb{P}_1(A) + \tfrac{1}{2}\mathbb{P}_0(A^c) \right\} = \tfrac{1}{2} \sup_{A \subseteq \mathcal{X}} \{\mathbb{P}_1(A) - \mathbb{P}_0(A)\} + \tfrac{1}{2}.$$

Since $\sup_\psi \mathbb{Q}[\psi(Z) = J] = 1 - \inf_\psi \mathbb{Q}[\psi(Z) \neq J]$, the claim (15.13) then follows from the definition (15.5) of the total variation distance.

The representation (15.13), in conjunction with Proposition 15.1, provides one avenue for deriving lower bounds. In particular, for any pair of distributions $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ such that $\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$, we have

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2} \{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\}. \tag{15.14}$$

Let us illustrate the use of this simple lower bound with some examples.

Example 15.4 (Gaussian location family) For a fixed variance σ^2, let $\mathbb{P}_\theta$ be the distribution of a $\mathcal{N}(\theta, \sigma^2)$ variable; letting the mean θ vary over the real line defines the Gaussian location family $\{\mathbb{P}_\theta, \theta \in \mathbb{R}\}$. Here we consider the problem of estimating θ under either the absolute error $|\widehat{\theta} - \theta|$ or the squared error $(\widehat{\theta} - \theta)^2$ using a collection $Z = (Y_1, \ldots, Y_n)$ of n i.i.d. samples drawn from a $\mathcal{N}(\theta, \sigma^2)$ distribution. We use $\mathbb{P}_\theta^n$ to denote this product distribution.

Let us apply the two-point Le Cam bound (15.14) with the distributions $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$. We set $\theta = 2\delta$, for some δ to be chosen later in the proof, which ensures that the two means are 2δ-separated. In order to apply the two-point Le Cam bound, we need to bound the total variation distance $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$. From the second-moment bound in Exercise 15.10(b), we have

$$\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \tfrac{1}{4}\left\{ e^{n\theta^2/\sigma^2} - 1 \right\} = \tfrac{1}{4}\left\{ e^{4n\delta^2/\sigma^2} - 1 \right\}. \tag{15.15}$$

Setting $\delta = \frac{1}{2}\frac{\sigma}{\sqrt{n}}$ thus yields

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\widehat{\theta} - \theta|] \geq \frac{\delta}{2}\left\{ 1 - \frac{1}{2}\sqrt{e-1} \right\} \geq \frac{\delta}{6} = \frac{1}{12}\frac{\sigma}{\sqrt{n}} \tag{15.16a}$$

and

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] \geq \frac{\delta^2}{2}\left\{ 1 - \frac{1}{2}\sqrt{e-1} \right\} \geq \frac{\delta^2}{6} = \frac{1}{24}\frac{\sigma^2}{n}. \tag{15.16b}$$

Although the pre-factors $1/12$ and $1/24$ are not optimal, the scalings $\sigma/\sqrt{n}$ and σ^2/n are sharp. For instance, the sample mean $\widetilde{\theta}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ satisfies the bounds

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\widetilde{\theta}_n - \theta|] = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\widetilde{\theta}_n - \theta)^2] = \frac{\sigma^2}{n}.$$

In Exercise 15.8, we explore an alternative approach, based on using the Pinsker–Csiszár–Kullback inequality from Lemma 15.2 to upper bound the TV distance in terms of the KL divergence. This approach yields a result with sharper constants. ♣

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Mean-squared error decaying as n^{-1} is typical for parametric problems with a certain type of regularity, of which the Gaussian location model is the archetypal example. For other "non-regular" problems, faster rates become possible, and the minimax lower bounds take a different form. The following example provides one illustration of this phenomenon:

Example 15.5 (Uniform location family) Let us consider the uniform location family, in which, for each $\theta \in \mathbb{R}$, the distribution $\mathbb{U}_\theta$ is uniform over the interval $[\theta, \theta + 1]$. We let $\mathbb{U}_\theta^n$ denote the product distribution of n i.i.d. samples from $\mathbb{U}_\theta$. In this case, it is not possible to use Lemma 15.2 to control the total variation norm, since the Kullback–Leibler divergence between $\mathbb{U}_\theta$ and $\mathbb{U}_{\theta'}$ is infinite whenever $\theta \neq \theta'$. Accordingly, we need to use an alternative distance measure: in this example, we illustrate the use of the Hellinger distance (see equation (15.9)).

Given a pair $\theta, \theta' \in \mathbb{R}$, let us compute the Hellinger distance between $\mathbb{U}_\theta$ and $\mathbb{U}_{\theta'}$. By symmetry, it suffices to consider the case $\theta' > \theta$. If $\theta' > \theta + 1$, then we have $H^2(\mathbb{U}_\theta \,||\, \mathbb{U}_{\theta'}) = 2$. Otherwise, when $\theta' \in (\theta, \theta + 1]$, we have

$$H^2(\mathbb{U}_\theta \,||\, \mathbb{U}_{\theta'}) = \int_\theta^{\theta'} dt + \int_{\theta+1}^{\theta'+1} dt = 2\,|\theta' - \theta|.$$

Consequently, if we take a pair θ, θ' such that $|\theta' - \theta| = 2\delta := \frac{1}{4n}$, then the relation (15.12b) guarantees that

$$\frac{1}{2} H^2(\mathbb{U}_\theta^n \,||\, \mathbb{U}_{\theta'}^n) \leq \frac{n}{2}\, 2\,|\theta' - \theta| = \frac{1}{4}.$$

In conjunction with Lemma 15.3, we find that

$$\|\mathbb{U}_\theta^n - \mathbb{U}_{\theta'}^n\|_{\mathrm{TV}}^2 \leq H^2(\mathbb{U}_\theta^n \,||\, \mathbb{U}_{\theta'}^n) \leq \tfrac{1}{2}.$$

From the lower bound (15.14) with $\Phi(t) = t^2$, we conclude that, for the uniform location family, the minimax risk is lower bounded as

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] \geq \frac{(1 - \frac{1}{\sqrt{2}})}{128} \frac{1}{n^2}.$$

The significant aspect of this lower bound is the faster n^{-2} rate, which should be contrasted with the n^{-1} rate in the regular situation. In fact, this n^{-2} rate is optimal for the uniform location model, achieved for instance by the estimator $\widetilde{\theta} = \min\{Y_1, \ldots, Y_n\}$; see Exercise 15.9 for details. ♣

Le Cam's method is also useful for various nonparametric problems, for instance those in which our goal is to estimate some functional $\theta \colon \mathscr{F} \to \mathbb{R}$ defined on a class of densities $\mathscr{F}$. For instance, a standard example is the problem of estimating a density at a point, say $x = 0$, in which case $\theta(f) := f(0)$ is known as an evaluation functional.

An important quantity in the Le Cam approach to such problems is the Lipschitz constant of the functional θ with respect to the Hellinger norm, given by

$$\omega(\epsilon; \theta, \mathscr{F}) := \sup_{f, g \in \mathscr{F}} \left\{ |\theta(f) - \theta(g)| \mid H^2(f \,||\, g) \leq \epsilon^2 \right\}. \tag{15.17}$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Here we use $H^2(f \,\|\, g)$ to mean the squared Hellinger distance between the distributions associated with the densities f and g. Note that the quantity ω measures the size of the fluctuations of $\theta(f)$ when f is perturbed in a Hellinger neighborhood of radius ϵ. The following corollary reveals the importance of this Lipschitz constant (15.17):

Corollary 15.6 (Le Cam for functionals) *For any increasing function Φ on the non-negative real line and any functional $\theta\colon \mathscr{F} \to \mathbb{R}$, we have*

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{F}} \mathbb{E}\Big[\Phi\big(\widehat{\theta} - \theta(f)\big)\Big] \geq \frac{1}{4}\Phi\Big(\frac{1}{2}\,\omega\Big(\frac{1}{2\sqrt{n}}; \theta, \mathscr{F}\Big)\Big). \tag{15.18}$$

Proof We adopt the shorthand $\omega(t) \equiv \omega(t; \theta, \mathscr{F})$ throughout the proof. Setting $\epsilon^2 = \frac{1}{4n}$, choose a pair f, g that achieve[3] the supremum defining $\omega(1/(2\sqrt{n}))$. By a combination of Le Cam's inequality (Lemma 15.3) and the decoupling property (15.12b) for the Hellinger distance, we have

$$\|\mathbb{P}_f^n - \mathbb{P}_g^n\|_{\mathrm{TV}}^2 \leq H^2(\mathbb{P}_f^n \,\|\, \mathbb{P}_g^n) \leq nH^2(\mathbb{P}_f \,\|\, \mathbb{P}_g) \leq \tfrac{1}{4}.$$

Consequently, Le Cam's bound (15.14) with $\delta = \frac{1}{2}\omega(\frac{1}{2\sqrt{n}})$ implies that

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{F}} \mathbb{E}\Big[\Phi\big(\widehat{\theta} - \theta(f)\big)\Big] \geq \frac{1}{4}\Phi\Big(\frac{1}{2}\,\omega\Big(\frac{1}{2\sqrt{n}}\Big)\Big),$$

as claimed. □

The elegance of Corollary 15.6 is in that it reduces the calculation of lower bounds to a geometric object—namely, the Lipschitz constant (15.17). Some concrete examples are helpful to illustrate the basic ideas.

Example 15.7 (Pointwise estimation of Lipschitz densities) Let us consider the family of densities on $[-\frac{1}{2}, \frac{1}{2}]$ that are bounded uniformly away from zero, and are Lipschitz with constant one—that is, $|f(x) - f(y)| \leq |x - y|$ for all $x, y \in [-\frac{1}{2}, \frac{1}{2}]$. Suppose that our goal is to estimate the linear functional $f \mapsto \theta(f) := f(0)$. In order to apply Corollary 15.6, it suffices to lower bound $\omega(\frac{1}{2\sqrt{n}}; \theta, \mathscr{F})$ and we can do so by choosing a pair $f_0, g \in \mathscr{F}$ with $H^2(f_0 \,\|\, g) = \frac{1}{4n}$, and then evaluating the difference $|\theta(f_0) - \theta(g)|$. Let $f_0 \equiv 1$ be the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$. For a parameter $\delta \in (0, \frac{1}{6}]$ to be chosen, consider the function

$$\phi(x) = \begin{cases} \delta - |x| & \text{for } |x| \leq \delta, \\ |x - 2\delta| - \delta & \text{for } x \in [\delta, 3\delta], \\ 0 & \text{otherwise.} \end{cases} \tag{15.19}$$

See Figure 15.2 for an illustration. By construction, the function ϕ is 1-Lipschitz, uniformly

[3] If the supremum is not achieved, then we can choose a pair that approximate it to any desired accuracy, and repeat the argument.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

bounded with $\|\phi\|_\infty = \delta \leq \frac{1}{6}$, and integrates to zero—that is, $\int_{-1/2}^{1/2} \phi(x)\,dx = 0$. Consequently, the perturbed function $g := f_0 + \phi$ is a density function belonging to our class, and by construction, we have the equality $|\theta(f_0) - \theta(g)| = \delta$.

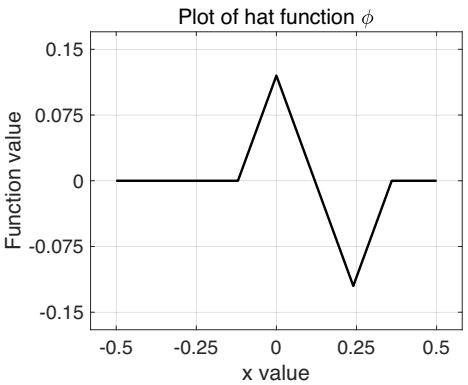

Figure 15.2 Illustration of the hat function ϕ from equation (15.19) for $\delta = 0.12$. It is 1-Lipschitz, uniformly bounded as $\|\phi\|_\infty \leq \delta$, and it integrates to zero.

It remains to control the squared Hellinger distance. By definition, we have

$$\tfrac{1}{2}H^2(f_0 \,\|\, g) = 1 - \int_{-1/2}^{1/2} \sqrt{1 + \phi(t)}\,dt.$$

Define the function $\Psi(u) = \sqrt{1+u}$, and note that $\sup_{u \in \mathbb{R}} |\Psi''(u)| \leq \frac{1}{4}$. Consequently, by a Taylor-series expansion, we have

$$\tfrac{1}{2}H^2(f_0 \,\|\, g) = \int_{-1/2}^{1/2} \{\Psi(0) - \Psi(\phi(t))\}\,dt \leq \int_{-1/2}^{1/2} \left\{-\Psi'(0)\phi(t) + \tfrac{1}{8}\phi^2(t)\right\} dt. \tag{15.20}$$

Observe that

$$\int_{-1/2}^{1/2} \phi(t)\,dt = 0 \quad \text{and} \quad \int_{-1/2}^{1/2} \phi^2(t)\,dt = 4\int_0^\delta (\delta - x)^2\,dx = \tfrac{4}{3}\delta^3.$$

Combined with our Taylor-series bound (15.20), we find that

$$H^2(f_0 \,\|\, g) \leq \tfrac{2}{8} \cdot \tfrac{4}{3}\,\delta^3 = \tfrac{1}{3}\delta^3.$$

Consequently, setting $\delta^3 = \frac{3}{4n}$ ensures that $H^2(f_0 \,\|\, g) \leq \frac{1}{4n}$. Putting together the pieces, Corollary 15.6 with $\Phi(t) = t^2$ implies that

$$\inf_{\widehat{\theta}} \sup_{f \in \mathscr{F}} \mathbb{E}\Big[\big(\widehat{\theta} - f(0)\big)^2\Big] \geq \frac{1}{16}\omega^2\Big(\frac{1}{2\sqrt{n}}\Big) \gtrsim n^{-2/3}.$$

This $n^{-2/3}$ lower bound for the Lipschitz family can be achieved by various estimators, so that we have derived a sharp lower bound. ♣

We now turn to the use of the two-class lower bound for a nonlinear functional in a non-parametric problem. Although the resulting bound is non-trivial, it is *not* a sharp result—unlike in the previous examples. Later, we will develop Le Cam's refinement of the two-

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

class approach so as to obtain sharp rates.

Example 15.8 (Lower bounds for quadratic functionals) Given positive constants $c_0 < 1 < c_1$ and $c_2 > 1$, consider the class of twice-differentiable density functions

$$\mathscr{F}_2([0,1]) := \left\{ f \colon [0,1] \to [c_0, c_1] \,\middle|\, \|f''\|_\infty \leq c_2 \text{ and } \int_0^1 f(x)\,dx = 1 \right\} \tag{15.21}$$

that are uniformly bounded above and below, and have a uniformly bounded second derivative. Consider the quadratic functional $f \mapsto \theta(f) := \int_0^1 (f'(x))^2\,dx$. Note that $\theta(f)$ provides a measure of the "smoothness" of the density: it is zero for the uniform density, and becomes large for densities with more erratic behavior. Estimation of such quadratic functionals arises in a variety of applications; see the bibliographic section for further discussion.

We again use Corollary 15.6 to derive a lower bound. Let f_0 denote the uniform distribution on $[0,1]$, which clearly belongs to $\mathscr{F}_2$. As in Example 15.7, we construct a perturbation g of f_0 such that $H^2(f_0 \,\|\, g) = \frac{1}{4n}$; Corollary 15.6 then gives a minimax lower bound of the order $(\theta(f_0) - \theta(g))^2$.

In order to construct the perturbation, let $\phi \colon [0,1] \to \mathbb{R}$ be a fixed twice-differentiable function that is uniformly bounded as $\|\phi\|_\infty \leq \frac{1}{2}$, and such that

$$\int_0^1 \phi(x)\,dx = 0 \quad \text{and} \quad b_\ell := \int_0^1 (\phi^{(\ell)}(x))^2\,dx > 0 \qquad \text{for } \ell = 0, 1. \tag{15.22}$$

Now divide the unit interval $[0,1]$ into m sub-intervals $[x_j, x_{j+1}]$, with $x_j = \frac{j}{m}$ for $j = 0, \ldots, m-1$. For a suitably small constant $C > 0$, define the shifted and rescaled functions

$$\phi_j(x) := \begin{cases} \frac{C}{m^2}\phi(m(x - x_j)) & \text{if } x \in [x_j, x_{j+1}], \\ 0 & \text{otherwise.} \end{cases} \tag{15.23}$$

We then consider the density $g(x) := 1 + \sum_{j=1}^m \phi_j(x)$. It can be seen that $g \in \mathscr{F}_2$ as long as the constant C is chosen sufficiently small. See Figure 15.3 for an illustration of this construction.

Let us now control the Hellinger distance. Following the same Taylor-series argument as in Example 15.7, we have

$$\begin{aligned} \frac{1}{2}H^2(f_0 \,\|\, g) \;=\; 1 - \int_0^1 \sqrt{1 + \sum_{j=1}^m \phi_j(x)}\,dx &\leq \frac{1}{8}\int_0^1 \Big(\sum_{j=1}^m \phi_j(x)\Big)^2 dx \\ &= \frac{1}{8}\sum_{j=1}^m \int_0^1 \phi_j^2(x)\,dx \\ &= c\,b_0 \frac{1}{m^4}, \end{aligned}$$

where $c > 0$ is a universal constant. Consequently, the choice $m^4 := 2c\,b_0\,n$ ensures that $H^2(f_0 \,\|\, g) \leq \frac{1}{n}$, as required for applying Corollary 15.6.

It remains to evaluate the difference $\theta(f_0)$ and $\theta(g)$. On one hand, we have $\theta(f_0) = 0$,

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

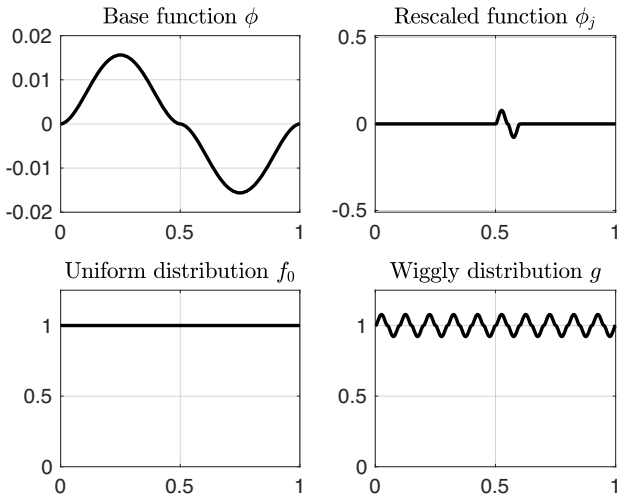

Figure 15.3 Illustration of the construction of the density g. Upper left: an example of a base function ϕ. Upper right: function ϕ_j is a rescaled and shifted version of ϕ. Lower left: original uniform distribution. Lower right: final density g is the superposition of the uniform density f_0 with the sum of the shifted functions $\{\phi_j\}_{j=1}^m$.

whereas on the other hand, we have

$$\theta(g) = \int_0^1 \left(\sum_{j=1}^m \phi_j'(x) \right)^2 dx = m \int_0^1 (\phi_j'(x))^2 \, dx = \frac{C^2 b_1}{m^2}.$$

Recalling the specified choice of m, we see that $|\theta(g) - \theta(f_0)| \geq \frac{K}{\sqrt{n}}$ for some universal constant K independent of n. Consequently, Corollary 15.6 with $\Phi(t) = t$ implies that

$$\sup_{f \in \mathscr{F}_2} \mathbb{E}[|\widehat{\theta}(f) - \theta(f)|] \gtrsim n^{-1/2}. \tag{15.24}$$

This lower bound, while valid, is *not optimal*—there is no estimator that can achieve error of the order of $n^{-1/2}$ uniformly over $\mathscr{F}_2$. Indeed, we will see that the minimax risk scales as $n^{-4/9}$, but proving this optimal lower bound requires an extension of the basic two-point technique, as we describe in the next section. ♣

15.2.2 Le Cam's convex hull method

Our discussion up until this point has focused on lower bounds obtained by single pairs of hypotheses. As we have seen, the difficulty of the testing problem is controlled by the total variation distance between the two distributions. Le Cam's method is an elegant generalization of this idea, one which allows us to take the convex hulls of two classes of distributions. In many cases, the separation in total variation norm as measured over the convex hulls is much smaller than the pointwise separation between two classes, and so leads to better lower bounds.

More concretely, consider two subsets $\mathcal{P}_0$ and $\mathcal{P}_1$ of $\mathcal{P}$ that are 2δ-separated, in the sense

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

that

$$\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta \qquad \text{for all } \mathbb{P}_0 \in \mathcal{P}_0 \text{ and } \mathbb{P}_1 \in \mathcal{P}_1. \tag{15.25}$$

Lemma 15.9 (Le Cam) *For any 2δ-separated classes of distributions $\mathcal{P}_0$ and $\mathcal{P}_1$ contained within $\mathcal{P}$, any estimator $\widehat{\theta}$ has worst-case risk at least*

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\left[\rho(\widehat{\theta}, \theta(\mathbb{P}))\right] \geq \frac{\delta}{2} \sup_{\substack{\mathbb{P}_0 \in \text{conv}(\mathcal{P}_0) \\ \mathbb{P}_1 \in \text{conv}(\mathcal{P}_1)}} \left\{1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}}\right\}. \tag{15.26}$$

Proof For any estimator $\widehat{\theta}$, let us define the random variables

$$V_j(\widehat{\theta}) = \frac{1}{2\delta} \inf_{\mathbb{P}_j \in \mathcal{P}_j} \rho(\widehat{\theta}, \theta(\mathbb{P}_j)), \qquad \text{for } j = 0, 1.$$

We then have

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] &\geq \tfrac{1}{2}\left\{\mathbb{E}_{\mathbb{P}_0}[\rho(\widehat{\theta}, \theta(\mathbb{P}_0))] + \mathbb{E}_{\mathbb{P}_1}[\rho(\widehat{\theta}, \theta(\mathbb{P}_1))]\right\} \\ &\geq \delta\left\{\mathbb{E}_{\mathbb{P}_0}[V_0(\widehat{\theta})] + \mathbb{E}_{\mathbb{P}_1}[V_1(\widehat{\theta})]\right\}. \end{aligned}$$

Since the right-hand side is linear in $\mathbb{P}_0$ and $\mathbb{P}_1$, we can take suprema over the convex hulls, and thus obtain the lower bound

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \geq \delta \sup_{\substack{\mathbb{P}_0 \in \text{conv}(\mathcal{P}_0) \\ \mathbb{P}_1 \in \text{conv}(\mathcal{P}_1)}} \left\{\mathbb{E}_{\mathbb{P}_0}[V_0(\widehat{\theta})] + \mathbb{E}_{\mathbb{P}_1}[V_1(\widehat{\theta})]\right\}.$$

By the triangle inequality, we have

$$\rho(\widehat{\theta}, \theta(\mathbb{P}_0)) + \rho(\widehat{\theta}, \theta(\mathbb{P}_1)) \geq \rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta.$$

Taking infima over $\mathbb{P}_j \in \mathcal{P}_j$ for each $j = 0, 1$, we obtain

$$\inf_{\mathbb{P}_0 \in \mathcal{P}_0} \rho(\widehat{\theta}, \theta(\mathbb{P}_0)) + \inf_{\mathbb{P}_1 \in \mathcal{P}_1} \rho(\widehat{\theta}, \theta(\mathbb{P}_1)) \geq 2\delta,$$

which is equivalent to $V_0(\widehat{\theta}) + V_1(\widehat{\theta}) \geq 1$. Since $V_j(\widehat{\theta}) \geq 0$ for $j = 0, 1$, the variational representation of the TV distance (see Exercise 15.1) implies that, for any $\mathbb{P}_j \in \text{conv}(\mathcal{P}_j)$, we have

$$\mathbb{E}_{\mathbb{P}_0}[V_0(\widehat{\theta})] + \mathbb{E}_{\mathbb{P}_1}[V_1(\widehat{\theta})] \geq 1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}},$$

which completes the proof. □

In order to see how taking the convex hulls can decrease the total variation norm, it is instructive to return to the Gaussian location model previously introduced in Example 15.4:

Example 15.10 (Sharpened bounds for Gaussian location family) In Example 15.4, we used a two-point form of Le Cam's method to prove a lower bound on mean estimation in the Gaussian location family. A key step was to upper bound the TV distance $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\text{TV}}$ between the n-fold product distributions based on the Gaussian models $\mathcal{N}(\theta, \sigma^2)$ and $\mathcal{N}(0, \sigma^2)$,

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

respectively. Here let us show how the convex hull version of Le Cam's method can be used to sharpen this step, so as obtain a bound with tighter constants. In particular, setting $\theta = 2\delta$ as before, consider the two families $\mathcal{P}_0 = \{\mathbb{P}_0^n\}$ and $\mathcal{P}_1 = \{\mathbb{P}_\theta^n, \mathbb{P}_{-\theta}^n\}$. Note that the mixture distribution $\bar{\mathbb{P}} := \frac{1}{2}\mathbb{P}_\theta^n + \frac{1}{2}\mathbb{P}_{-\theta}^n$ belongs to $\operatorname{conv}(\mathcal{P}_1)$. From the second-moment bound explored in Exercise 15.10(c), we have

$$\|\bar{\mathbb{P}} - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \tfrac{1}{4}\Big\{e^{\frac{1}{2}(\frac{\sqrt{n}\theta}{\sigma})^4} - 1\Big\} = \tfrac{1}{4}\Big\{e^{\frac{1}{2}(\frac{2\sqrt{n}\delta}{\sigma})^4} - 1\Big\}. \tag{15.27}$$

Setting $\delta = \frac{\sigma t}{2\sqrt{n}}$ for some parameter $t > 0$ to be chosen, the convex hull Le Cam bound (15.26) yields

$$\min_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\widehat{\theta} - \theta|] \geq \frac{\sigma}{4\sqrt{n}} \sup_{t>0} \Big\{t\big(1 - \frac{1}{2}\sqrt{e^{\frac{1}{2}t^4} - 1}\big)\Big\} \geq \frac{3}{20}\frac{\sigma}{\sqrt{n}}.$$

This bound is an improvement over our original bound (15.16a) from Example 15.4, which has the pre-factor of $\frac{1}{12} \approx 0.08$, as opposed to $\frac{3}{20} = 0.15$ obtained from this analysis. Thus, even though we used the same base separation δ, our use of mixture distributions reduced the TV distance—compare the bounds (15.27) and (15.15)—thereby leading to a sharper result. ♣

In the previous example, the gains from extending to the convex hull are only in terms of the constant pre-factors. Let us now turn to an example in which the gain is more substantial. Recall Example 15.8 in which we investigated the problem of estimating the quadratic functional $f \mapsto \theta(f) = \int_0^1 (f'(x))^2\,dx$ over the class $\mathscr{F}_2$ from equation (15.21). Let us now demonstrate how the use of Le Cam's method in its full convex hull form allows for the derivation of an optimal lower bound for the minimax risk.

Example 15.11 (Optimal bounds for quadratic functionals) For each binary vector $\alpha \in \{-1, +1\}^m$, define the distribution $\mathbb{P}_\alpha$ with density given by

$$f_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j \phi_j(x).$$

Note that the perturbed density g constructed in Example 15.8 is a special member of this family, generated by the binary vector $\alpha = (1, 1, \ldots, 1)$. Let $\mathbb{P}_\alpha^n$ denote the product distribution on $\mathcal{X}^n$ formed by sampling n times independently from $\mathbb{P}_\alpha$, and define the two classes $\mathcal{P}_0 := \{\mathbb{U}^n\}$ and $\mathcal{P}_1 := \{\mathbb{P}_\alpha^n,\ \alpha \in \{-1, +1\}^m\}$. With these choices, we then have

$$\inf_{\substack{\mathbb{P}_j \in \operatorname{conv}(\mathcal{P}_j) \\ j=0,1}} \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \leq \|\mathbb{U}^n - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{U}^n \,\|\, \mathbb{Q}),$$

where $\mathbb{Q} := 2^{-m} \sum_{\alpha \in \{-1,+1\}^m} \mathbb{P}_\alpha^n$ is the uniformly weighted mixture over all 2^m choices of $\mathbb{P}_\alpha^n$.

In this case, since $\mathbb{Q}$ is not a product distribution, we can no longer apply the decomposition (15.12a) so as to bound the Hellinger distance $H(\mathbb{U}^n \,\|\, \mathbb{Q})$ by a univariate version. Instead, some more technical calculations are required. One possible upper bound is given by

$$H^2(\mathbb{U}^n \,\|\, \mathbb{Q}) \leq n^2 \sum_{j=1}^m \Big(\int_0^1 \phi_j^2(x)\,dx\Big)^2. \tag{15.28}$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

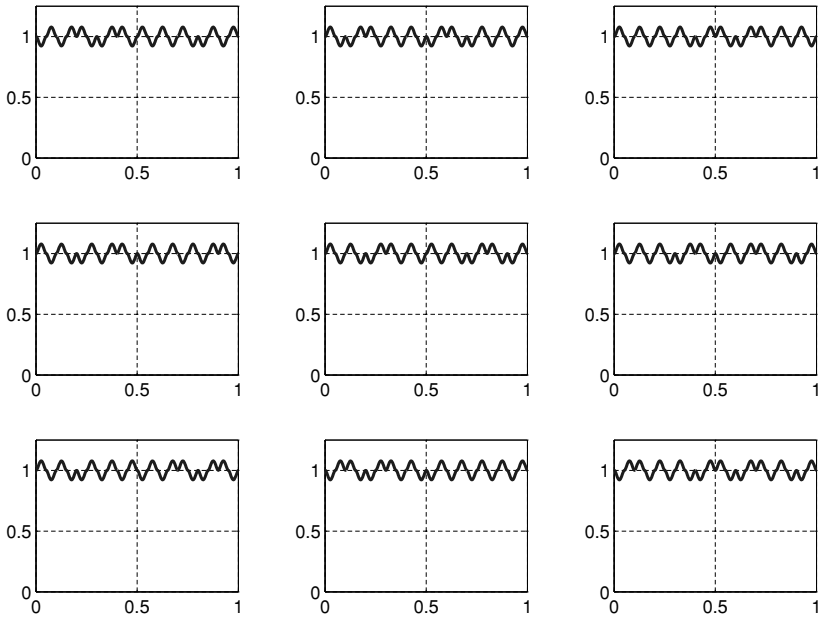

Figure 15.4 Illustration of some densities of the form $f_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j \phi_j(x)$ for different choices of sign vectors $\alpha \in \{-1, 1\}^m$. Note that there are 2^m such densities in total.

See the bibliographic section for discussion of this upper bound as well as related results. If we take the upper bound (15.28) as given, then using the calculations from Example 15.8 —in particular, recall the definition of the constants b_ℓ from equation (15.22)—we find that

$$H^2(\mathbb{U}^n \,\|\, \mathbb{Q}) \leq mn^2 \frac{b_0^2}{m^{10}} = b_0^2 \frac{n^2}{m^9}.$$

Setting $m^9 = 4b_0^2 n^2$ yields that $\|\mathbb{U}^{1:n} - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{U}^{1:n} \,\|\, \mathbb{P}^{1:n}) \leq 1/2$, and hence Lemma 15.9 implies that

$$\sup_{f \in \mathscr{F}_2} \mathbb{E}|\widehat{\theta}(f) - \theta(f)| \geq \delta/4 = \frac{C^2 b_1}{8m^2} \gtrsim n^{-4/9}.$$

Thus, by using the full convex form of Le Cam's method, we have recovered a better lower bound on the minimax risk ($n^{-4/9} \gg n^{-1/2}$). This lower bound turns out to be unimprovable; see the bibliographic section for further discussion. ♣

15.3 Fano's method

In this section, we describe an alternative method for deriving lower bounds, one based on a classical result from information theory known as Fano's inequality.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

15.3.1 Kullback–Leibler divergence and mutual information

Recall our basic set-up: we are interested in lower bounding the probability of error in an M-ary hypothesis testing problem, based on a family of distributions $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^M}\}$. A sample Z is generated by choosing an index J uniformly at random from the index set $[M] := \{1, \ldots, M\}$, and then generating data according to $\mathbb{P}_{\theta^J}$. In this way, the observation follows the *mixture distribution* $\mathbb{Q}_Z = \bar{\mathbb{Q}} := \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$. Our goal is to identify the index J of the probability distribution from which a given sample has been drawn.

Intuitively, the difficulty of this problem depends on the amount of dependence between the observation Z and the unknown random index J. In the extreme case, if Z were actually independent of J, then observing Z would have no value whatsoever. How to measure the amount of dependence between a pair of random variables? Note that the pair (Z, J) are independent if and only if their joint distribution $\mathbb{Q}_{Z,J}$ is equal to the product of its marginals—namely, $\mathbb{Q}_Z \mathbb{Q}_J$. Thus, a natural way in which to measure dependence is by computing some type of divergence measure between the joint distribution and the product of marginals. The *mutual information* between the random variables (Z, J) is defined in exactly this way, using the Kullback–Leibler divergence as the underlying measure of distance—that is

$$I(Z, J) := D(\mathbb{Q}_{Z,J} \,\|\, \mathbb{Q}_Z \, \mathbb{Q}_J). \tag{15.29}$$

By standard properties of the KL divergence, we always have $I(Z, J) \geq 0$, and moreover $I(Z, J) = 0$ if and only if Z and J are independent.

Given our set-up and the definition of the KL divergence, the mutual information can be written in terms of component distributions $\{\mathbb{P}_{\theta^j}, \, j \in [M]\}$ and the mixture distribution $\bar{\mathbb{Q}} \equiv \mathbb{Q}_Z$—in particular as

$$I(Z; J) = \frac{1}{M} \sum_{j=1}^{M} D(\mathbb{P}_{\theta^j} \,\|\, \bar{\mathbb{Q}}), \tag{15.30}$$

corresponding to the mean KL divergence between $\mathbb{P}_{\theta^j}$ and $\bar{\mathbb{Q}}$, averaged over the choice of index j. Consequently, the mutual information is small if the distributions $\mathbb{P}_{\theta^j}$ are hard to distinguish from the mixture distribution $\bar{\mathbb{Q}}$ on average.

15.3.2 Fano lower bound on minimax risk

Let us now return to the problem at hand: namely, obtaining lower bounds on the minimax error. The Fano method is based on the following lower bound on the error probability in an M-ary testing problem, applicable when J is uniformly distributed over the index set:

$$\mathbb{P}[\psi(Z) \neq J] \geq 1 - \frac{I(Z; J) + \log 2}{\log M}. \tag{15.31}$$

When combined with the reduction from estimation to testing given in Proposition 15.1, we obtain the following lower bound on the minimax error:

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Proposition 15.12 *Let $\{\theta^1, \ldots, \theta^M\}$ be a 2δ-separated set in the ρ semi-metric on $\Theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $\{1, \ldots, M\}$, and $(Z \mid J = j) \sim \mathbb{P}_{\theta^j}$. Then for any increasing function $\Phi : [0, \infty) \to [0, \infty)$, the minimax risk is lower bounded as*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z; J) + \log 2}{\log M} \right\}, \tag{15.32}$$

where $I(Z; J)$ is the mutual information between Z and J.

We provide a proof of the Fano bound (15.31), from which Proposition 15.12 follows, in the sequel (see Section 15.4). For the moment, in order to gain intuition for this result, it is helpful to consider the behavior of the different terms of $\delta \to 0^+$. As we shrink δ, then the 2δ-separation criterion becomes milder, so that the cardinality $M \equiv M(2\delta)$ in the denominator increases. At the same time, in a generic setting, the mutual information $I(Z; J)$ will decrease, since the random index $J \in [M(2\delta)]$ can take on a larger number of potential values. By decreasing δ sufficiently, we may thereby ensure that

$$\frac{I(Z; J) + \log 2}{\log M} \leq \frac{1}{2}, \tag{15.33}$$

so that the lower bound (15.32) implies that $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$. Thus, we have a generic scheme for deriving lower bounds on the minimax risk.

In order to derive lower bounds in this way, there remain two technical and possibly challenging steps. The first requirement is to specify 2δ-separated sets with large cardinality $M(2\delta)$. Here the theory of metric entropy developed in Chapter 5 plays an important role, since any 2δ-packing set is (by definition) 2δ-separated in the ρ semi-metric. The second requirement is to compute—or more realistically to upper bound—the mutual information $I(Z; J)$. In general, this second step is non-trivial, but various avenues are possible.

The simplest upper bound on the mutual information is based on the convexity of the Kullback–Leibler divergence (see Exercise 15.3). Using this convexity and the mixture representation (15.30), we find that

$$I(Z; J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D(\mathbb{P}_{\theta^j} \,||\, \mathbb{P}_{\theta^k}). \tag{15.34}$$

Consequently, if we can construct a 2δ-separated set such that all pairs of distributions $\mathbb{P}_{\theta^j}$ and $\mathbb{P}_{\theta^k}$ are close on average, the mutual information can be controlled. Let us illustrate the use of this upper bound for a simple parametric problem.

Example 15.13 (Normal location model via Fano method) Recall from Example 15.4 the normal location family, and the problem of estimating $\theta \in \mathbb{R}$ under the squared error. There we showed how to lower bound the minimax error using Le Cam's method; here let us derive a similar lower bound using Fano's method.

Consider the 2δ-separated set of real-valued parameters $\{\theta^1, \theta^2, \theta^3\} = \{0, 2\delta, -2\delta\}$. Since

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

$\mathbb{P}_{\theta^j} = \mathcal{N}(\theta^j, \sigma^2)$, we have

$$D(\mathbb{P}_{\theta^j}^{1:n} \,||\, \mathbb{P}_{\theta^k}^{1:n}) = \frac{n}{2\sigma^2}(\theta^j - \theta^k)^2 \leq \frac{2n\delta^2}{\sigma^2} \qquad \text{for all } j, k = 1, 2, 3.$$

The bound (15.34) then ensures that $I(Z; J_\delta) \leq \frac{2n\delta^2}{\sigma^2}$, and choosing $\delta^2 = \frac{\sigma^2}{20n}$ ensures that $\frac{2n\delta^2/\sigma^2 + \log 2}{\log 3} < 0.75$. Putting together the pieces, the Fano bound (15.32) with $\Phi(t) = t^2$ implies that

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] \geq \frac{\delta^2}{4} = \frac{1}{80}\frac{\sigma^2}{n}.$$

In this way, we have re-derived a minimax lower bound of the order σ^2/n, which, as discussed in Example 15.4, is of the correct order. ♣

15.3.3 Bounds based on local packings

Let us now formalize the approach that was used in the previous example. It is based on a local packing of the parameter space Ω, which underlies what is called the "generalized Fano" method in the statistics literature. (As a sidenote, this nomenclature is very misleading, because the method is actually based on a substantial weakening of the Fano bound, obtained from the inequality (15.34).)

The local packing approach proceeds as follows. Suppose that we can construct a 2δ-separated set contained within Ω such that, for some quantity c, the Kullback–Leibler divergences satisfy the uniform upper bound

$$\sqrt{D(\mathbb{P}_{\theta^j} \,||\, \mathbb{P}_{\theta^k})} \leq c\sqrt{n}\,\delta \qquad \text{for all } j \neq k. \tag{15.35a}$$

The bound (15.34) then implies that $I(Z; J) \leq c^2 n\delta^2$, and hence the bound (15.33) will hold as long as

$$\log M(2\delta) \geq 2\{c^2 n\delta^2 + \log 2\}. \tag{15.35b}$$

In summary, if we can find a 2δ-separated family of distributions such that conditions (15.35a) and (15.35b) both hold, then we may conclude that the minimax risk is lower bounded as $\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$.

Let us illustrate the local packing approach with some examples.

Example 15.14 (Minimax risks for linear regression) Consider the standard linear regression model $y = \mathbf{X}\theta^* + w$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a fixed design matrix, and the vector $w \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ is observation noise. Viewing the design matrix $\mathbf{X}$ as fixed, let us obtain lower bounds on the minimax risk in the prediction (semi-)norm $\rho_{\mathbf{X}}(\widehat{\theta}, \theta^*) := \frac{\|\mathbf{X}(\widehat{\theta} - \theta^*)\|_2}{\sqrt{n}}$, assuming that θ^* is allowed to vary over $\mathbb{R}^d$.

For a tolerance $\delta > 0$ to be chosen, consider the set

$$\{\gamma \in \text{range}(\mathbf{X}) \mid \|\gamma\|_2 \leq 4\delta\sqrt{n}\},$$

and let $\{\gamma^1, \ldots, \gamma^M\}$ be a $2\delta\sqrt{n}$-packing in the ℓ_2-norm. Since this set sits in a space of dimension $r = \text{rank}(\mathbf{X})$, Lemma 5.7 implies that we can find such a packing with $\log M \geq r \log 2$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

elements. We thus have a collection of vectors of the form $\gamma^j = \mathbf{X}\theta^j$ for some $\theta^j \in \mathbb{R}^d$, and such that

$$\frac{\|\mathbf{X}\theta^j\|_2}{\sqrt{n}} \leq 4\delta, \qquad \text{for each } j \in [M], \tag{15.36a}$$

$$2\delta \leq \frac{\|\mathbf{X}(\theta^j - \theta^k)\|_2}{\sqrt{n}} \leq 8\delta \qquad \text{for each } j \neq k \in [M] \times [M]. \tag{15.36b}$$

Let $\mathbb{P}_{\theta^j}$ denote the distribution of y when the true regression vector is θ^j; by the definition of the model, under $\mathbb{P}_{\theta^j}$, the observed vector $y \in \mathbb{R}^n$ follows a $\mathcal{N}(\mathbf{X}\theta^j, \sigma^2\mathbf{I}_n)$ distribution. Consequently, the result of Exercise 15.13 ensures that

$$D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{P}_{\theta^k}) = \frac{1}{2\sigma^2}\|\mathbf{X}(\theta^j - \theta^k)\|_2^2 \;\leq\; \frac{32n\delta^2}{\sigma^2}, \tag{15.37}$$

where the inequality follows from the upper bound (15.36b). Consequently, for r sufficiently large, the lower bound (15.35b) can be satisfied by setting $\delta^2 = \frac{\sigma^2}{64}\frac{r}{n}$, and we conclude that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\Big[\frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta)\|_2^2\Big] \geq \frac{\sigma^2}{128}\,\frac{\operatorname{rank}(\mathbf{X})}{n}.$$

This lower bound is sharp up to constant pre-factors: as shown by our analysis in Example 13.8 and Exercise 13.2, it can be achieved by the usual linear least-squares estimate. ♣

Let us now see how the upper bound (15.34) and Fano's method can be applied to a non-parametric problem.

Example 15.15 (Minimax risk for density estimation) Recall from equation (15.21) the family $\mathscr{F}_2$ of twice-smooth densities on $[0, 1]$, bounded uniformly above, bounded uniformly away from zero, and with uniformly bounded second derivative. Let us consider the problem of estimating the entire density function f, using the Hellinger distance as our underlying metric ρ.

In order to construct a local packing, we make use of the family of perturbed densities from Example 15.11, each of the form $f_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j \phi_j(x)$, where $\alpha \in \{-1, +1\}^m$ and the function ϕ_j was defined in equation (15.23). Although there are 2^m such perturbed densities, it is convenient to use only a well-separated subset of them. Let $M_H(\frac{1}{4}; \mathbb{H}^m)$ denote the $\frac{1}{4}$-packing number of the binary hypercube $\{-1, +1\}^m$ in the rescaled Hamming metric. From our calculations in Example 5.3, we know that

$$\log M_H(\tfrac{1}{4};\, \mathbb{H}^m) \geq m\, D(\tfrac{1}{4} \,\|\, \tfrac{1}{2}) \geq \frac{m}{10}.$$

(See in particular equation (5.3).) Consequently, we can find a subset $\mathbb{T} \subset \{-1, +1\}^m$ with cardinality at least $e^{m/10}$ such that

$$d_H(\alpha, \beta) = \frac{1}{m}\sum_{j=1}^m \mathbb{I}[\alpha_j \neq \beta_j] \geq 1/4 \qquad \text{for all } \alpha \neq \beta \in \mathbb{T}. \tag{15.38}$$

We then consider the family of $M = e^{m/10}$ distributions $\{\mathbb{P}_\alpha,\, \alpha \in \mathbb{T}\}$, where $\mathbb{P}_\alpha$ has density f_α.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

We first lower bound the Hellinger distance between distinct pairs f_α and f_β. Since ϕ_j is non-zero only on the interval $I_j = [x_j, x_{j+1}]$, we can write

$$\int_0^1 \left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2 dx = \sum_{j=0}^{m-1} \int_{I_j} \left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2 dx.$$

But on the interval I_j, we have

$$\left(\sqrt{f_\alpha(x)} + \sqrt{f_\beta(x)}\right)^2 = 2\left(f_\alpha(x) + f_\beta(x)\right) \leq 4,$$

and therefore

$$\begin{aligned}\int_{I_j} \left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2 dx &\geq \frac{1}{4} \int_{I_j} \left(f_\alpha(x) - f_\beta(x)\right)^2 \\ &\geq \int_{I_j} \phi_j^2(x)\, dx \qquad \text{whenever } \alpha_j \neq \beta_j.\end{aligned}$$

Since $\int_{I_j} \phi_j^2(x)\, dx = \int_0^1 \phi^2(x)\, dx = \frac{b_0}{m^5}$ and any distinct $\alpha \neq \beta$ differ in at least $m/4$ positions, we find that $H^2(\mathbb{P}_\alpha \,\|\, \mathbb{P}_\beta) \geq \frac{m}{4}\frac{b_0}{m^5} = \frac{b_0}{m^4} \equiv 4\delta^2$. Consequently, we have constructed a 2δ-separated set with $\delta^2 = \frac{b_0}{4m^4}$.

Next we upper bound the pairwise KL divergence. By construction, we have $f_\alpha(x) \geq 1/2$ for all $x \in [0, 1]$, and thus

$$\begin{aligned}D(\mathbb{P}_\alpha \,\|\, \mathbb{P}_\beta) &\leq \int_0^1 \frac{\left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2}{f_\alpha(x)}\, dx \\ &\leq 2 \int_0^1 \left(\sqrt{f_\alpha(x)} - \sqrt{f_\beta(x)}\right)^2 dx \leq \frac{4b_0}{m^4}, \end{aligned} \tag{15.39}$$

where the final inequality follows by a similar sequence of calculations. Overall, we have established the upper bound $D(\mathbb{P}_\alpha^n \,\|\, \mathbb{P}_\beta^n) = nD(\mathbb{P}_\alpha \,\|\, \mathbb{P}_\beta) \leq 4b_0\frac{n}{m^4} = 4n\delta^2$. Finally, we must ensure that

$$\log M = \frac{m}{10} \geq 2\left\{4n\delta^2 + \log 2\right\} = 2\left\{4b_0\frac{n}{m^4} + \log 2\right\}.$$

This equality holds if we choose $m = \frac{n^{1/5}}{C}$ for a sufficiently small constant C. With this choice, we have $\delta^2 \asymp m^{-4} \asymp n^{-4/5}$, and hence conclude that

$$\sup_{f \in \mathscr{F}_2} H^2(\widehat{f} \,\|\, f) \gtrsim n^{-4/5}.$$

This rate is minimax-optimal for densities with two orders of smoothness; recall that we encountered the same rate for the closely related problem of nonparametric regression in Chapter 13. ♣

As a third example, let us return to the high-dimensional parametric setting, and study minimax risks for the problem of sparse linear regression, which we studied in detail in Chapter 7.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Example 15.16 (Minimax risk for sparse linear regression) Consider the high-dimensional linear regression model $y = \mathbf{X}\theta^* + w$, where the regression vector θ^* is known *a priori* to be sparse, say with at most $s < d$ non-zero coefficients. It is then natural to consider the minimax risk over the set

$$\mathbb{S}^d(s) := \mathbb{B}_0^d(s) \cap \mathbb{B}_2(1) = \left\{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \, \|\theta\|_2 \leq 1\right\} \tag{15.40}$$

of s-sparse vectors within the Euclidean unit ball.

Let us first construct a $1/2$-packing of the set $\mathbb{S}^d(s)$. From our earlier results in Chapter 5 (in particular, see Exercise 5.8), there exists a $1/2$-packing of this set with log cardinality at least $\log M \geq \frac{s}{2} \log \frac{d-s}{s}$. We follow the same rescaling procedure as in Example 15.14 to form a 2δ-packing such that $\|\theta^j - \theta^k\|_2 \leq 4\delta$ for all pairs of vectors in our packing set. Since the vector $\theta^j - \theta^k$ is at most $2s$-sparse, we have

$$\sqrt{D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{P}_{\theta^k})} = \frac{1}{\sqrt{2}\sigma} \|\mathbf{X}(\theta^j - \theta^k)\|_2 \leq \frac{\gamma_{2s}}{\sqrt{2}\sigma} 4\delta,$$

where $\gamma_{2s} := \max_{|T|=2s} \sigma_{\max}(\mathbf{X}_T)/\sqrt{n}$. Putting together the pieces, we see that the minimax risk is lower bounded by any $\delta > 0$ for which

$$\frac{s}{2} \log \frac{d-s}{s} \geq 128 \frac{\gamma_{2s}^2}{\sigma^2} n\delta^2 + 2 \log 2.$$

As long as $s \leq d/2$ and $s \geq 10$, the choice $\delta^2 = \frac{\sigma^2}{400\gamma_{2s}^2} s \log \frac{d-s}{s}$ suffices. Putting together the pieces, we conclude that in the range $10 \leq s \leq d/2$, the minimax risk is lower bounded as

$$\mathfrak{M}(\mathbb{S}^d(s); \|\cdot\|_2) \gtrsim \frac{\sigma^2}{\gamma_{2s}^2} \frac{s \log \frac{ed}{s}}{n}. \tag{15.41}$$

The constant obtained by this argument is not sharp, but this lower bound is otherwise unimprovable: see the bibliographic section for further details. ♣

15.3.4 Local packings with Gaussian entropy bounds

Our previous examples have also used the convexity-based upper bound (15.34) on the mutual information. We now turn to a different upper bound on the mutual information, applicable when the conditional distribution of Z given J is Gaussian.

Lemma 15.17 *Suppose J is uniformly distributed over $[M] = \{1, \ldots, M\}$ and that Z conditioned on $J = j$ has a Gaussian distribution with covariance $\mathbf{\Sigma}^j$. Then the mutual information is upper bounded as*

$$I(Z; J) \leq \frac{1}{2} \left\{ \log \det \operatorname{cov}(Z) - \frac{1}{M} \sum_{j=1}^{M} \log \det(\mathbf{\Sigma}^j) \right\}. \tag{15.42}$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

This upper bound is a consequence of the maximum entropy property of the multivariate Gaussian distribution; see Exercise 15.14 for further details. In the special case when $\mathbf{\Sigma}^j = \mathbf{\Sigma}$ for all $j \in [M]$, it takes on the simpler form

$$I(Z; J) \leq \frac{1}{2} \log \left(\frac{\det \operatorname{cov}(Z)}{\det(\mathbf{\Sigma})} \right). \tag{15.43}$$

Let us illustrate the use of these bounds with some examples.

Example 15.18 (Variable selection in sparse linear regression) Let us return to the model of sparse linear regression from Example 15.16, based on the standard linear model $y = \mathbf{X}\theta^* + w$, where the unknown regression vector $\theta^* \in \mathbb{R}^d$ is s-sparse. Here we consider the problem of lower bounding the minimax risk for the problem of variable selection—namely, determining the support set $S = \{j \in \{1, 2, \ldots, d\} \mid \theta^*_j \neq 0\}$, which is assumed to have cardinality $s \ll d$.

In this case, the problem of interest is itself a multiway hypothesis test—namely, that of choosing from all $\binom{d}{s}$ possible subsets. Consequently, a direct application of Fano's inequality leads to lower bounds, and we can obtain different such bounds by constructing various ensembles of subproblems. These subproblems are parameterized by the pair (d, s), as well as the quantity $\theta_{\min} = \min_{j \in S} |\theta^*_j|$. In this example, we show that, in order to achieve a probability of error below $1/2$, any method requires a sample size of at least

$$n > \max \left\{ 8 \frac{\log(d + s - 1)}{\log(1 + \frac{\theta^2_{\min}}{\sigma^2})}, \; 8 \frac{\log \binom{d}{s}}{\log(1 + s \frac{\theta^2_{\min}}{\sigma^2})} \right\}, \tag{15.44}$$

as long as $\min \left\{ \log(d + s - 1), \log \binom{d}{s} \right\} \geq 4 \log 2$.

For this problem, our observations consist of the response vector $y \in \mathbb{R}^n$ and design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. We derive lower bounds by first conditioning on a particular instantiation $\mathbf{X} = \{x_i\}_{i=1}^n$ of the design matrix, and using a form of Fano's inequality that involves the mutual information $I_{\mathbf{X}}(y; J)$ between the response vector y and the random index J with the design matrix $\mathbf{X}$ held fixed. In particular, we have

$$\mathbb{P}[\psi(y, \mathbf{X}) \neq J \mid \mathbf{X} = \{x_i\}_{i=1}^n] \geq 1 - \frac{I_{\mathbf{X}}(y; J) + \log 2}{\log M},$$

so that by taking averages over $\mathbf{X}$, we can obtain lower bounds on $\mathbb{P}[\psi(y, \mathbf{X}) \neq J]$ that involve the quantity $\mathbb{E}_{\mathbf{X}}[I_{\mathbf{X}}(y; J)]$.

Ensemble A: Consider the class $M = \binom{d}{s}$ of all possible subsets of cardinality s, enumerated in some fixed way. For the ℓth subset S^ℓ, let $\theta^\ell \in \mathbb{R}^d$ have values $\theta_{\min}$ for all indices $j \in S^\ell$, and zeros in all other positions. For a fixed covariate vector $x_i \in \mathbb{R}^d$, an observed response $y_i \in \mathbb{R}$ then follows the mixture distribution $\frac{1}{M} \sum_{\ell=1}^M \mathbb{P}_{\theta^\ell}$, where $\mathbb{P}_{\theta^\ell}$ is the distribution of a $\mathcal{N}(\langle x_i, \theta^\ell \rangle, \sigma^2)$ random variable.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

By the definition of mutual information, we have

$$\begin{aligned} I_{\mathbf{X}}(y; J) &= H_{\mathbf{X}}(y) - H_{\mathbf{X}}(y \mid J) \\ &\overset{\text{(i)}}{\leq} \left[\sum_{i=1}^{n} H_{\mathbf{X}}(y_i) \right] - H_{\mathbf{X}}(y \mid J) \\ &\overset{\text{(ii)}}{=} \sum_{i=1}^{n} \{ H_{\mathbf{X}}(y_1) - H_{\mathbf{X}}(y_1 \mid J) \} \\ &= \sum_{i=1}^{n} I_{\mathbf{X}}(y_i; J), \end{aligned} \tag{15.45}$$

where step (i) follows since independent random vectors have larger entropy than dependent ones (see Exercise 15.4), and step (ii) follows since $(y_1, \ldots, y_n)$ are independent conditioned on J. Next, applying Lemma 15.17 repeatedly for each $i \in [n]$ with $Z = y_i$, conditionally on the matrix $\mathbf{X}$ of covariates, yields

$$I_{\mathbf{X}}(y; J) \leq \frac{1}{2} \sum_{i=1}^{n} \log \frac{\operatorname{var}(y_i \mid x_i)}{\sigma^2}.$$

Now taking averages over $\mathbf{X}$ and using the fact that the pairs (y_i, x_i) are jointly i.i.d., we find that

$$\mathbb{E}_{\mathbf{X}} \left[I_{\mathbf{X}}(y; J) \right] \leq \frac{n}{2} \mathbb{E}[\log \frac{\operatorname{var}(y_1 \mid x_1)}{\sigma^2}] \leq \frac{n}{2} \log \frac{\mathbb{E}_{x_1}[\operatorname{var}(y_1 \mid x_1)]}{\sigma^2},$$

where the last inequality follows Jensen's inequality, and concavity of the logarithm.

It remains to upper bound the variance term. Since the random vector y_1 follows a mixture distribution with M components, we have

$$\begin{aligned} \mathbb{E}_{x_1}[\operatorname{var}(y_1 \mid x_1)] \ \leq \mathbb{E}_{x_1}[\mathbb{E}[y_1^2 \mid x_1]] &= \mathbb{E}_{x_1}\Big[x_1^{\mathrm{T}} \{ \frac{1}{M} \sum_{j=1}^{M} \theta^j \otimes \theta^j \} x_1 + \sigma^2 \Big] \\ &= \operatorname{trace}\Big(\frac{1}{M} \sum_{j=1}^{M} (\theta^j \otimes \theta^j) \Big) + \sigma^2. \end{aligned}$$

Now each index $j \in \{1, 2, \ldots, d\}$ appears in $\binom{d-1}{s-1}$ of the total number of subsets $M = \binom{d}{s}$, so that

$$\operatorname{trace}\Big(\frac{1}{M} \sum_{j=1}^{M} \theta^j \otimes \theta^j \Big) = d \frac{\binom{d-1}{s-1}}{\binom{d}{s}} \theta_{\min}^2 = s \theta_{\min}^2.$$

Putting together the pieces, we conclude that

$$\mathbb{E}_{\mathbf{X}}[I_{\mathbf{X}}(y; J)] \leq \frac{n}{2} \log \left(1 + \frac{s \theta_{\min}^2}{\sigma^2} \right),$$

and hence the Fano lower bound implies that

$$\mathbb{P}[\psi(y, \mathbf{X}) \neq J] \geq 1 - \frac{\frac{n}{2} \log(1 + \frac{s\theta_{\min}^2}{\sigma^2}) + \log 2}{\log \binom{d}{s}},$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

from which the first lower bound in equation (15.44) follows as long as $\log \binom{d}{s} \geq 4 \log 2$, as assumed.

Ensemble B: Let $\bar{\theta} \in \mathbb{R}^d$ be a vector with $\theta_{\min}$ in its first $s-1$ coordinates, and zero in all remaining $d-s+1$ coordinates. For each $j = 1, \ldots, d$, let $e_j \in \mathbb{R}^d$ denote the jth standard basis vector with a single one in position j. Define the family of $M = d - s + 1$ vectors $\theta^j := \bar{\theta} + \theta_{\min} e_j$ for $j = s, \ldots, d$. By a straightforward calculation, we have $\mathbb{E}[Y \mid x] = \langle x, \gamma \rangle$, where $\gamma := \bar{\theta} + \frac{1}{M}\theta_{\min} e_{s \to d}$, and the vector $e_{s \to d} \in \mathbb{R}^d$ has ones in positions s through d, and zeros elsewhere. By the same argument as for ensemble A, it suffices to upper bound the quantity $\mathbb{E}_{x_1}[\mathrm{var}(y_1 \mid x_1)]$. Using the definition of our ensemble, we have

$$\mathbb{E}_{x_1}[\mathrm{var}(y_1 \mid x_1)] = \sigma^2 + \mathrm{trace}\left\{ \frac{1}{M} \sum_{j=1}^{M} (\theta^j \otimes \theta^j - \gamma \otimes \gamma) \right\} \leq \sigma^2 + \theta_{\min}^2. \tag{15.46}$$

Recall that we have assumed that $\log(d - s + 1) > 4 \log 2$. Using Fano's inequality and the upper bound (15.46), the second term in the lower bound (15.44) then follows. ♣

Let us now turn to a slightly different problem, namely that of lower bounds for principal component analysis. Recall from Chapter 8 the spiked covariance ensemble, in which a random vector $x \in \mathbb{R}^d$ is generated via

$$x \stackrel{\mathrm{d}}{=} \sqrt{\nu} \xi \theta^* + w. \tag{15.47}$$

Here $\nu > 0$ is a given signal-to-noise ratio, θ^* is a fixed vector with unit Euclidean norm, and the random quantities $\xi \sim \mathcal{N}(0, 1)$ and $w \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent. Observe that the d-dimensional random vector x is zero-mean Gaussian with a covariance matrix of the form $\mathbf{\Sigma} := \mathbf{I}_d + \nu(\theta^* \otimes \theta^*)$. Moreover, by construction, the vector θ^* is the unique maximal eigenvector of the covariance matrix $\mathbf{\Sigma}$.

Suppose that our goal is to estimate θ^* based on n i.i.d. samples of the random vector x. In the following example, we derive lower bounds on the minimax risk in the squared Euclidean norm $\|\widehat{\theta} - \theta^*\|_2^2$. (As discussed in Chapter 8, recall that there is always a sign ambiguity in estimating eigenvectors, so that in computing the Euclidean norm, we implicitly assume that the correct direction is chosen.)

Example 15.19 (Lower bounds for PCA) Let $\{\Delta^1, \ldots, \Delta^M\}$ be a $1/2$-packing of the unit sphere in $\mathbb{R}^{d-1}$; from Example 5.8, for all $d \geq 3$, there exists such a set with cardinality $\log M \geq (d-1)\log 2 \geq d/2$. For a given orthonormal matrix $\mathbf{U} \in \mathbb{R}^{(d-1)\times(d-1)}$ and tolerance $\delta \in (0, 1)$ to be chosen, consider the family of vectors

$$\theta^j(\mathbf{U}) = \sqrt{1 - \delta^2} \begin{bmatrix} 1 \\ 0_{d-1} \end{bmatrix} + \delta \begin{bmatrix} 0 \\ \mathbf{U}\Delta^j \end{bmatrix} \qquad \text{for } j \in [M], \tag{15.48}$$

where 0_{d-1} denotes the $(d-1)$-dimensional vector of zeros. By construction, each vector $\theta^j(\mathbf{U})$ lies on the unit sphere in $\mathbb{R}^d$, and the collection of all M vectors forms a $\delta/2$-packing set. Consequently, we can lower bound the minimax risk by constructing a testing problem based on the family of vectors (15.48). In fact, so as to make the calculations clean, we construct one testing problem for each choice of orthonormal matrix $\mathbf{U}$, and then take averages over a randomly chosen matrix.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Let $\mathbb{P}_{\theta^j(\mathbf{U})}$ denote the distribution of a random vector from the spiked ensemble (15.47) with leading eigenvector $\theta^* := \theta^j(\mathbf{U})$. By construction, it is a zero-mean Gaussian random vector with covariance matrix

$$\mathbf{\Sigma}^j(\mathbf{U}) := \mathbf{I}_d + \nu\Big(\theta^j(\mathbf{U}) \otimes \theta^j(\mathbf{U})\Big).$$

Now for a fixed $\mathbf{U}$, suppose that we choose an index $J \in [M]$ uniformly at random, and then drawn n i.i.d. samples from the distribution $\mathbb{P}_{\theta^j(\mathbf{U})}$. Letting $Z_1^n(\mathbf{U})$ denote the samples thus obtained, Fano's inequality then implies that the testing error is lower bounded as

$$\mathbb{P}[\psi(Z_1^n(\mathbf{U})) \neq J \mid \mathbf{U}] \geq 1 - \frac{I(Z_1^n(\mathbf{U}); J) + \log 2}{d/2}, \tag{15.49}$$

where we have used the fact that $\log M \geq d/2$. For each fixed $\mathbf{U}$, the samples $Z_1^n(\mathbf{U})$ are conditionally independent given J. Consequently, following the same line of reasoning leading to equation (15.45), we can conclude that $I(Z_1^n(\mathbf{U}); J) \leq nI(Z(\mathbf{U}); J)$, where $Z(\mathbf{U})$ denotes a single sample.

Since the lower bound (15.49) holds for each fixed choice of orthonormal matrix $\mathbf{U}$, we can take averages when $\mathbf{U}$ is chosen uniformly at random. Doing so simplifies the task of bounding the mutual information, since we need only bound the averaged mutual information $\mathbb{E}_{\mathbf{U}}[I(Z(\mathbf{U}); J)]$. Since $\det(\mathbf{\Sigma}^j(\mathbf{U})) = 1 + \nu$ for each $j \in [M]$, Lemma 15.17 implies that

$$\begin{aligned} \mathbb{E}_{\mathbf{U}}[I(Z(\mathbf{U}); J) &\leq \tfrac{1}{2}\Big\{\mathbb{E}_{\mathbf{U}} \log\det(\operatorname{cov}(Z(\mathbf{U}))) - \log(1+\nu)\Big\} \\ &\leq \tfrac{1}{2}\Big\{\log\det \underbrace{\mathbb{E}_{\mathbf{U}}(\operatorname{cov}(Z(\mathbf{U})))}_{:=\mathbf{\Gamma}} - \log(1+\nu)\Big\}, \end{aligned} \tag{15.50}$$

where the second step uses the concavity of the log-determinant function, and Jensen's inequality. Let us now compute the entries of the expected covariance matrix $\mathbf{\Gamma}$. It can be seen that $\Gamma_{11} = 1 + \nu - \nu\delta^2$; moreover, using the fact that $\mathbf{U}\Delta^j$ is uniformly distributed over the unit sphere in dimension $(d-1)$, the first column is equal to

$$\Gamma_{(2\to d),1} = \nu\delta\sqrt{1-\delta^2}\,\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_{\mathbf{U}}[\mathbf{U}\Delta^j] = 0.$$

Letting $\mathbf{\Gamma}_{\text{low}}$ denote the lower square block of side length $(d-1)$, we have

$$\mathbf{\Gamma}_{\text{low}} = \mathbf{I}_{d-1} + \frac{\delta^2\nu}{M}\sum_{j=1}^{M}\mathbb{E}\Big[(\mathbf{U}\Delta^j) \otimes (\mathbf{U}\Delta^j)\Big] = \Big(1 + \frac{\delta^2\nu}{d-1}\Big)\mathbf{I}_{d-1},$$

again using the fact that the random vector $\mathbf{U}\Delta^j$ is uniformly distributed over the sphere in dimension $d-1$. Putting together the pieces, we have shown that $\mathbf{\Gamma} = \operatorname{blkdiag}(\Gamma_{11}, \mathbf{\Gamma}_{\text{low}})$, and hence

$$\log\det\mathbf{\Gamma} = (d-1)\log\Big(1 + \frac{\nu\delta^2}{d-1}\Big) + \log(1 + \nu - \nu\delta^2).$$

Combining our earlier bound (15.50) with the elementary inequality $\log(1+t) \leq t$, we find

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

that

$$\begin{aligned} 2\mathbb{E}_{\mathbf{U}}\big[I(Z(\mathbf{U}); J)\big] &\leq (d-1)\log\Big(1 + \frac{\nu\delta^2}{d-1}\Big) + \log\Big(1 - \frac{\nu}{1+\nu}\delta^2\Big) \\ &\leq \Big(\nu - \frac{\nu}{1+\nu}\Big)\delta^2 \\ &= \frac{\nu^2}{1+\nu}\delta^2. \end{aligned}$$

Taking averages over our earlier Fano bound (15.49) and using this upper bound on the averaged mutual information, we find that the minimax risk for estimating the spiked eigenvector in squared Euclidean norm is lower bounded as

$$\mathfrak{M}(\text{PCA}; \mathbb{S}^{d-1}, \|\cdot\|_2^2) \gtrsim \min\Big\{\frac{1+\nu}{\nu^2}\frac{d}{n}, 1\Big\}.$$

In Corollary 8.7, we proved that the maximum eigenvector of the sample covariance achieves this squared Euclidean error up to constant pre-factors, so that we have obtained a sharp characterization of the minimax risk. ♣

As a follow-up to the previous example, we now turn to the sparse variant of principal components analysis. As discussed in Chapter 8, there are a number of motivations for studying sparsity in PCA, including the fact that it allows eigenvectors to be estimated at substantially faster rates. Accordingly, let us now prove some lower bounds for variable selection in sparse PCA, again working under the spiked model (15.47).

Example 15.20 (Lower bounds for variable selection in sparse PCA) Suppose that our goal is to determine the scaling of the sample size required to ensure that the support set of an s-sparse eigenvector θ^* can be recovered. Of course, the difficulty of the problem depends on the minimum value $\theta_{\min} = \min_{j \in S} |\theta^*_j|$. Here we show that if $\theta_{\min} \gtrsim \frac{1}{\sqrt{s}}$, then any method requires $n \gtrsim \frac{1+\nu}{\nu^2} s \log(d - s + 1)$ samples to correctly recover the support. In Exercise 15.15, we prove a more general lower bound for arbitrary scalings of $\theta_{\min}$.

Recall our analysis of variable selection in sparse linear regression from Example 15.18: here we use an approach similar to ensemble B from that example. In particular, fix a subset S of size $s-1$, and let $\varepsilon \in \{-1, 1\}^d$ be a vector of sign variables. For each $j \in S^c := [d] \setminus S$, we then define the vector

$$[\theta^j(\varepsilon)]_\ell = \begin{cases} \frac{1}{\sqrt{s}} & \text{if } \ell \in S, \\ \frac{\varepsilon_j}{\sqrt{s}} & \text{if } \ell = j, \\ 0 & \text{otherwise.} \end{cases}$$

In Example 15.18, we computed averages over a randomly chosen orthonormal matrix $\mathbf{U}$; here instead we average over the choice of random sign vectors ε.

Let $\mathbb{P}_{\theta^j(\varepsilon)}$ denote the distribution of the spiked vector (15.47) with $\theta^* = \theta^j(\varepsilon)$, and let $Z(\varepsilon)$ be a sample from the mixture distribution $\frac{1}{M}\sum_{j \in S^c} \mathbb{P}_{\theta^j(\varepsilon)}$. Following a similar line of calculation as Example 15.19, we have

$$\mathbb{E}_\varepsilon[I(Z(\varepsilon); J)] \leq \tfrac{1}{2}\Big\{\log\det(\mathbf{\Gamma}) - \log(1+\nu)\Big\},$$

where $\mathbf{\Gamma} := \mathbb{E}_\varepsilon[\text{cov}(Z(\varepsilon))]$ is the averaged covariance matrix, taken over the uniform distribution over all Rademacher vectors. Letting $\mathbf{E}_{s-1}$ denote a square matrix of all ones with

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

side length $s-1$, a straightforward calculation yields that $\mathbf{\Gamma}$ is a block diagonal matrix with $\mathbf{\Gamma}_{SS} = \mathbf{I}_{s-1} + \frac{\nu}{s}\mathbf{E}_{s-1}$ and $\mathbf{\Gamma}_{S^cS^c} = (1 + \frac{\nu}{s(d-s+1)})\,\mathbf{I}_{d-s+1}$. Consequently, we have

$$\begin{aligned}
2\mathbb{E}_\varepsilon[I(Z(\varepsilon); J)] &\leq \log\Big(1 + \nu\frac{s-1}{s}\Big) + (d-s+1)\log\Big(1 + \frac{\nu}{s\,(d-s+1)}\Big) - \log(1+\nu) \\
&= \log\Big(1 - \frac{\nu}{1+\nu}\frac{1}{s}\Big) + (d-s+1)\log\Big(1 + \frac{\nu}{s\,(d-s+1)}\Big) \\
&\leq \frac{1}{s}\Big\{-\frac{\nu}{1+\nu} + \nu\Big\} \\
&= \frac{1}{s}\,\frac{\nu^2}{1+\nu}.
\end{aligned}$$

Recalling that we have n samples and that $\log M = \log(d-s-1)$, Fano's inequality implies that the probability of error is bounded away from zero as long as the ratio

$$\frac{n}{s\log(d-s+1)}\,\frac{\nu^2}{1+\nu}$$

is upper bounded by a sufficiently small but universal constant, as claimed. ♣

15.3.5 Yang–Barron version of Fano's method

Our analysis thus far has been based on relatively naive upper bounds on the mutual information. These upper bounds are useful whenever we are able to construct a local packing of the parameter space, as we have done in the preceding examples. In this section, we develop an alternative upper bound on the mutual information. It is particularly useful for nonparametric problems, since it obviates the need for constructing a local packing.

Lemma 15.21 (Yang–Barron method) *Let $N_{\mathrm{KL}}(\epsilon; \mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ in the square-root KL divergence. Then the mutual information is upper bounded as*

$$I(Z; J) \leq \inf_{\epsilon > 0}\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P})\}. \tag{15.51}$$

Proof Recalling the form (15.30) of the mutual information, we observe that for any distribution $\mathbb{Q}$, the mutual information is upper bounded by

$$I(Z; J) = \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta^j} \,\|\, \bar{\mathbb{Q}}) \overset{\text{(i)}}{\leq} \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{Q}) \leq \max_{j=1,\ldots,M} D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{Q}), \tag{15.52}$$

where inequality (i) uses the fact that the mixture distribution $\bar{\mathbb{Q}} := \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\theta^j}$ minimizes the average Kullback–Leibler divergence over the family $\{\mathbb{P}_{\theta^1}, \ldots, \mathbb{P}_{\theta^M}\}$—see Exercise 15.11 for details.

Since the upper bound (15.52) holds for any distribution $\mathbb{Q}$, we are free to choose it: in particular, we let $\{\gamma^1, \ldots, \gamma^N\}$ be an ϵ-covering of Ω in the square-root KL pseudo-distance,

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

and then set $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{P}_{\gamma^k}$. By construction, for each θ^j with $j \in [M]$, we can find some γ^k such that $D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{P}_{\gamma^k}) \leq \epsilon^2$. Therefore, we have

$$\begin{aligned} D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{Q}) &= \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta_j}}{\frac{1}{N}\sum_{\ell=1}^{N} d\mathbb{P}_{\gamma^\ell}}\right] \\ &\leq \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta_j}}{\frac{1}{N} d\mathbb{P}_{\gamma^k}}\right] \\ &= D(\mathbb{P}_{\theta^j} \,\|\, \mathbb{P}_{\gamma^k}) + \log N \\ &\leq \epsilon^2 + \log N. \end{aligned}$$

Since this bound holds for any choice of $j \in [M]$ and any choice of $\epsilon > 0$, the claim (15.51) follows. □

In conjunction with Proposition 15.12, Lemma 15.21 allows us to prove a minimax lower bound of the order δ as long as the pair $(\delta, \epsilon) \in \mathbb{R}_+^2$ are chosen such that

$$\log M(\delta; \rho, \Omega) \geq 2\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P}) + \log 2\}.$$

Finding such a pair can be accomplished via a two-step procedure:

(A) First, choose $\epsilon_n > 0$ such that

$$\epsilon_n^2 \geq \log N_{\mathrm{KL}}(\epsilon_n; \mathcal{P}). \tag{15.53a}$$

Since the KL divergence typically scales with n, it is usually the case that ϵ_n^2 also grows with n, hence the subscript in our notation.

(B) Second, choose the largest $\delta_n > 0$ that satisfies the lower bound

$$\log M(\delta_n; \rho, \Omega) \geq 4\epsilon_n^2 + 2\log 2. \tag{15.53b}$$

As before, this two-step procedure is best understood by working through some examples.

Example 15.22 (Density estimation revisited) In order to illustrate the use of the Yang–Barron method, let us return to the problem of density estimation in the Hellinger metric, as previously considered in Example 15.15. Our analysis involved the class $\mathscr{F}_2$, as defined in equation (15.21), of densities on $[0, 1]$, bounded uniformly above, bounded uniformly away from zero, and with uniformly bounded second derivative. Using the local form of Fano's method, we proved that the minimax risk in squared Hellinger distance is lower bounded as $n^{-4/5}$. In this example, we recover the same result more directly by using known results about the metric entropy.

For uniformly bounded densities on the interval $[0, 1]$, the squared Hellinger metric is sandwiched above and below by constant multiples of the $L^2([0, 1])$-norm:

$$\|p - q\|_2^2 := \int_0^1 \big(p(x) - q(x)\big)^2 \, dx.$$

Moreover, again using the uniform lower bound, the Kullback–Leibler divergence between any pair of distributions in this family is upper bounded by a constant multiple of the squared Hellinger distance, and hence by a constant multiple of the squared Euclidean distance. (See

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

equation (15.39) for a related calculation.) Consequently, in order to apply the Yang–Barron method, we need only understand the scaling of the metric entropy in the L^2-norm. From classical theory, it is known that the metric entropy of the class $\mathscr{F}_2$ in L^2-norm scales as $\log N(\delta; \mathscr{F}_2, \|\cdot\|_2) \asymp (1/\delta)^{1/2}$ for $\delta > 0$ sufficiently small.

Step A: Given n i.i.d. samples, the square-root Kullback–Leibler divergence is multiplied by a factor of $\sqrt{n}$, so that the inequality (15.53a) can be satisfied by choosing $\epsilon_n > 0$ such that

$$\epsilon_n^2 \succsim \left(\frac{\sqrt{n}}{\epsilon_n}\right)^{1/2}.$$

In particular, the choice $\epsilon_n^2 \asymp n^{1/5}$ is sufficient.

Step B: With this choice of ϵ_n, the second condition (15.53b) can be satisfied by choosing $\delta_n > 0$ such that

$$\left(\frac{1}{\delta_n}\right)^{1/2} \succsim n^{2/5},$$

or equivalently $\delta_n^2 \asymp n^{-4/5}$. In this way, we have a much more direct re-derivation of the $n^{-4/5}$ lower bound on the minimax risk. ♣

As a second illustration of the Yang–Barron approach, let us now derive some minimax risks for the problem of nonparametric regression, as discussed in Chapter 13. Recall that the standard regression model is based on i.i.d. observations of the form

$$y_i = f^*(x_i) + \sigma w_i, \qquad \text{for } i = 1, 2, \ldots, n,$$

where $w_i \sim \mathcal{N}(0, 1)$. Assuming that the design points $\{x_i\}_{i=1}^n$ are drawn in an i.i.d. fashion from some distribution $\mathbb{P}$, let us derive lower bounds in the $L^2(\mathbb{P})$-norm:

$$\|\widehat{f} - f^*\|_2^2 = \int_{\mathcal{X}} \left[\widehat{f}(x) - f^*(x)\right]^2 \mathbb{P}(dx).$$

Example 15.23 (Minimax risks for generalized Sobolev families) For a smoothness parameter $\alpha > 1/2$, consider the ellipsoid $\ell^2(\mathbb{N})$ given by

$$\mathcal{E}_\alpha = \Big\{(\theta_j)_{j=1}^\infty \mid \sum_{j=1}^\infty j^{2\alpha}\, \theta_j^2 \leq 1\Big\}. \tag{15.54a}$$

Given an orthonormal sequence $(\phi_j)_{j=1}^\infty$ in $L^2(\mathbb{P})$, we can then define the function class

$$\mathscr{F}_\alpha := \left\{ f = \sum_{j=1}^\infty \theta_j \phi_j \mid (\theta_j)_{j=1}^\infty \in \mathcal{E}_\alpha \right\}. \tag{15.54b}$$

As discussed in Chapter 12, these function classes can be viewed as particular types of reproducing kernel Hilbert spaces, where α corresponds to the degree of smoothness. For

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

any such function class, we claim that the minimax risk in squared $L^2(\mathbb{P})$-norm is lower bounded as

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{F}_\alpha} \mathbb{E}[\|\widehat{f} - f\|_2^2] \gtrsim \min\left\{1, \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\right\}, \tag{15.55}$$

and here we prove this claim via the Yang–Barron technique.

Consider a function of the form $f = \sum_{j=1}^{\infty} \theta_j \phi_j$ for some $\theta \in \ell^2(\mathbb{N})$, and observe that by the orthonormality of $(\phi_j)_{j=1}^{\infty}$, Parseval's theorem implies that $\|f\|_2^2 = \sum_{j=1}^{\infty} \theta_j^2$. Consequently, based on our calculations from Example 5.12, the metric entropy of $\mathscr{F}_\alpha$ scales as $\log N(\delta; \mathscr{F}_\alpha, \|\cdot\|_2) \asymp (1/\delta)^{1/\alpha}$. Accordingly, we can find a δ-packing $\{f^1, \ldots, f^M\}$ of $\mathscr{F}_\alpha$ in the $\|\cdot\|_2$-norm with $\log M \gtrsim (1/\delta)^{1/\alpha}$ elements.

Step A: For this part of the calculation, we first need to upper bound the metric entropy in the KL divergence. For each $j \in [M]$, let $\mathbb{P}_{f^j}$ denote the distribution of y given $\{x_i\}_{i=1}^n$ when the true regression function is f^j, and let $\mathbb{Q}$ denote the n-fold product distribution over the covariates $\{x_i\}_{i=1}^n$. When the true regression function is f^j, the joint distribution over $(y, \{x_i\}_{i=1}^n)$ is given by $\mathbb{P}_{f^j} \times \mathbb{Q}$, and hence for any distinct pair of indices $j \neq k$, we have

$$\begin{aligned} D(\mathbb{P}_{f^j} \times \mathbb{Q} \,\|\, \mathbb{P}_{f^k} \times \mathbb{Q}) = \mathbb{E}_x\big[D(\mathbb{P}_{f^j} \,\|\, \mathbb{P}_{f^k})\big] &= \mathbb{E}_x\Big[\frac{1}{2\sigma^2} \sum_{i=1}^n (f^j(x_i) - f^k(x_i))^2\Big] \\ &= \frac{n}{2\sigma^2} \|f^j - f^k\|_2^2. \end{aligned}$$

Consequently, we find that

$$\log N_{\mathrm{KL}}(\epsilon) = \log N\Big(\frac{\sigma\sqrt{2}}{\sqrt{n}}\epsilon; \mathscr{F}_\alpha, \|\cdot\|_2\Big) \precsim \big(\frac{\sqrt{n}}{\sigma\epsilon}\big)^{1/\alpha},$$

where the final inequality again uses the result of Example 5.12. Consequently, inequality (15.53a) can be satisfied by setting $\epsilon_n^2 \asymp \big(\frac{n}{\sigma^2}\big)^{\frac{1}{2\alpha+1}}$.

Step B: It remains to choose $\delta > 0$ to satisfy the inequality (15.53b). Given our choice of ε_n and the scaling of the packing entropy, we require

$$(1/\delta)^{1/\alpha} \geq c \left\{\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}} + 2\log 2\right\}. \tag{15.56}$$

As long as n/σ^2 is larger than some universal constant, the choice $\delta_n^2 \asymp \big(\frac{\sigma^2}{n}\big)^{\frac{2\alpha}{2\alpha+1}}$ satisfies the condition (15.56). Putting together the pieces yields the claim (15.55). ♣

In the exercises, we explore a number of other applications of the Yang–Barron method.

15.4 Appendix: Basic background in information theory

This appendix is devoted to some basic information-theoretic background, including a proof of Fano's inequality. The most fundamental concept is that of the *Shannon entropy*: it is a

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

functional on the space of probability distributions that provides a measure of their dispersion.

Definition 15.24 Let $\mathbb{Q}$ be a probability distribution with density $q = \frac{d\mathbb{Q}}{d\mu}$ with respect to some base measure μ. The Shannon entropy is given by

$$H(\mathbb{Q}) := -\mathbb{E}[\log q(X)] = -\int_{\mathcal{X}} q(x) \log q(x) \mu(dx), \tag{15.57}$$

when this integral is finite.

The simplest form of entropy arises when $\mathbb{Q}$ is supported on a discrete set $\mathcal{X}$, so that q can be taken as a probability mass function—hence a density with respect to the counting measure on $\mathcal{X}$. In this case, the definition (15.57) yields the discrete entropy

$$H(\mathbb{Q}) = -\sum_{x \in \mathcal{X}} q(x) \log q(x). \tag{15.58}$$

It is easy to check that the discrete entropy is always non-negative. Moreover, when $\mathcal{X}$ is a finite set, it satisfies the upper bound $H(\mathbb{Q}) \leq \log |\mathcal{X}|$, with equality achieved when $\mathbb{Q}$ is uniform over $\mathcal{X}$. See Exercise 15.2 for further discussion of these basic properties.

An important remark on notation is needed before proceeding: Given a random variable $X \sim \mathbb{Q}$, one often writes $H(X)$ in place of $H(\mathbb{Q})$. From a certain point of view, this is abusive use of notation, since the entropy is a functional of the distribution $\mathbb{Q}$ as opposed to the random variable X. However, as it is standard practice in information theory, we make use of this convenient notation in this appendix.

Definition 15.25 Given a pair of random variables (X, Y) with joint distribution $\mathbb{Q}_{X,Y}$, the conditional entropy of $X \mid Y$ is given by

$$H(X \mid Y) := \mathbb{E}_Y[H(\mathbb{Q}_{X|Y})] = \mathbb{E}_Y\left[\int_{\mathcal{X}} q(x \mid Y) \log q(x \mid Y) \mu(dx)\right]. \tag{15.59}$$

We leave the reader to verify the following elementary properties of entropy and mutual information. First, conditioning can only reduce entropy:

$$H(X \mid Y) \leq H(X). \tag{15.60a}$$

As will be clear below, this inequality is equivalent to the non-negativity of the mutual information $I(X; Y)$. Secondly, the joint entropy can be decomposed into a sum of singleton and conditional entropies as

$$H(X, Y) = H(Y) + H(X \mid Y). \tag{15.60b}$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

This decomposition is known as the chain rule for entropy. The conditional entropy also satisfies a form of chain rule:

$$H(X, Y \mid Z) = H(X \mid Z) + H(X \mid Y, Z). \tag{15.60c}$$

Finally, it is worth noting the connections between entropy and mutual information. By expanding the definition of mutual information, we see that

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \tag{15.60d}$$

By replacing the joint entropy with its chain rule decomposition (15.60b), we obtain

$$I(X; Y) = H(Y) - H(Y \mid X). \tag{15.60e}$$

With these results in hand, we are now ready to prove the Fano bound (15.31). We do so by first establishing a slightly more general result. Introducing the shorthand notation $q_e = \mathbb{P}[\psi(Z) \neq J]$, we let $h(q_e) = -q_e \log q_e - (1 - q_e) \log(1 - q_e)$ denote the binary entropy. With this notation, the standard form of Fano's inequality is that the error probability in any M-ary testing problem is lower bounded as

$$h(q_e) + q_e \log(M - 1) \geq H(J \mid Z). \tag{15.61}$$

To see how this lower bound implies the stated claim (15.31), we note that

$$H(J \mid Z) \stackrel{\text{(i)}}{=} H(J) - I(Z; J) \stackrel{\text{(ii)}}{=} \log M - I(Z; J),$$

where equality (i) follows from the representation of mutual information in terms of entropy, and equality (ii) uses our assumption that J is uniformly distributed over the index set. Since $h(q_e) \leq \log 2$, we find that

$$\log 2 + q_e \log M \geq \log M - I(Z; J),$$

which is equivalent to the claim (15.31).

It remains to prove the lower bound (15.61). Define the $\{0, 1\}$-valued random variable $V := \mathbb{I}[\psi(Z) \neq J]$, and note that $H(V) = h(q_e)$ by construction. We now proceed to expand the conditional entropy $H(V, J \mid Z)$ in two different ways. On one hand, by the chain rule, we have

$$H(V, J \mid Z) = H(J \mid Z) + H(V \mid J, Z) = H(J \mid Z), \tag{15.62}$$

where the second equality follows since V is a function of Z and J. By an alternative application of the chain rule, we have

$$H(V, J \mid Z) = H(V \mid Z) + H(J \mid V, Z) \leq h(q_e) + H(J \mid V, Z),$$

where the inequality follows since conditioning can only reduce entropy. By the definition of conditional entropy, we have

$$H(J \mid V, Z) = \mathbb{P}[V = 1] H(J \mid Z, V = 1) + \mathbb{P}[V = 0] H(J \mid Z, V = 0).$$

If $V = 0$, then $J = \psi(Z)$, so that $H(J \mid Z, V = 0) = 0$. On the other hand, if $V = 1$, then we

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

know that $J \neq \psi(Z)$, so that the conditioned random variable $(J \mid Z, V = 1)$ can take at most $M - 1$ values, which implies that

$$H(J \mid Z, V = 1) \leq \log(M - 1),$$

since entropy is maximized by the uniform distribution. We have thus shown that

$$H(V, J \mid Z) \leq h(q_e) + \log(M - 1),$$

and combined with the earlier equality (15.62), the claim (15.61) follows.

15.5 Bibliographic details and background

Information theory was introduced in the seminal work of Shannon (1948; 1949); see also Shannon and Weaver (1949). Kullback and Leibler (1951) introduced the Kullback–Leibler divergence, and established various connections to both large-deviation theory and testing problems. Early work by Lindley (1956) also established connections between information and statistical estimation. Kolmogorov was the first to connect information theory and metric entropy; in particular, see appendix II of the paper by Kolmogorov and Tikhomirov (1959). The book by Cover and Thomas (1991) is a standard introductory-level text on information theory. The proof of Fano's inequality given here follows their book.

The parametric problems discussed in Examples 15.4 and 15.5 were considered in Le Cam (1973), where he described the lower bounding approach now known as Le Cam's method. In this same paper, Le Cam also shows how a variety of nonparametric problems can also be treated by this method, using results on metric entropy. The paper by Hasminskii (1978) used the weakened form of the Fano method, based on the upper bound (15.34) on the mutual information, to derive lower bounds on density estimation in the uniform metric; see also the book by Hasminskii and Ibragimov (1981), as well as their survey paper (Hasminskii and Ibragimov, 1990). Assouad (1983) developed a method for deriving lower bounds based on placing functions at vertices of the binary hypercube. See also Birgé (1983; 1987; 2005) for further refinements on methods for deriving both lower and upper bounds. The chapter by Yu (1996) provides a comparison of both Le Cam's and Fano's method, as well Assouad's method (Assouad, 1983). Examples 15.8, 15.11 and 15.15 follow parts of her development. Birgé and Massart (1995) prove the upper bound (15.28) on the squared Hellinger distance; see theorem 1 in their paper for further details. In their paper, they study the more general problem of estimating functionals of the density and its first k derivatives under general smoothness conditions of order α. The quadratic functional problem considered in Examples 15.8 and 15.11 correspond to the special case with $k = 1$ and $\alpha = 2$. The refined upper bound on mutual information from Lemma 15.21 is due to Yang and Barron (1999). Their work showed how Fano's method can be applied directly with global metric entropies, as opposed to constructing specific local packings of the function class, as in the local packing version of Fano's method discussed in Section 15.3.3.

Guntuboyina (2011) proves a generalization of Fano's inequality to an arbitrary f-divergence. See Exercise 15.12 for further background on f-divergences and their properties. His result reduces to the classical Fano's inequality when the underlying f-divergence is the Kullback–Leibler divergence. He illustrates how such generalized Fano bounds can be used to derive minimax bounds for various classes of problems, including covariance estimation.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Lower bounds on variable selection in sparse linear regression using the Fano method, as considered in Example 15.18, were derived by Wainwright (2009a). See also the papers (Reeves and Gastpar, 2008; Fletcher et al., 2009; Akcakaya and Tarokh, 2010; Wang et al., 2010) for further results of this type. The lower bound on variable selection in sparse PCA from Example 15.20 was derived in Amini and Wainwright (2009); the proof given here is somewhat more streamlined due to the symmetrization with Rademacher variables.

The notion of minimax risk discussed in this chapter is the classical one, in which no additional constraints (apart from measurability) are imposed on the estimators. Consequently, the theory allows for estimators that may involve prohibitive computational, storage or communication costs to implement. A more recent line of work has been studying constrained forms of statistical minimax theory, in which the infimum over estimators is suitably restricted (Wainwright, 2014). In certain cases, there can be substantial gaps between the classical minimax risk and their computationally constrained analogs (e.g., Berthet and Rigollet, 2013; Ma and Wu, 2013; Wang et al., 2014; Zhang et al., 2014; Cai et al., 2015; Gao et al., 2015). Similarly, privacy constraints can lead to substantial differences in the classical and private minimax risks (Duchi et al., 2014, 2013).

15.6 Exercises

Exercise 15.1 (Alternative representation of TV norm) Show that the total variation norm has the equivalent variational representation

$$\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}} = 1 - \inf_{f_0 + f_1 \geq 1} \left\{ \mathbb{E}_0[f_0] + \mathbb{E}_1[f_1] \right\},$$

where the infimum runs over all non-negative measurable functions, and the inequality is taken pointwise.

Exercise 15.2 (Basics of discrete entropy) Let $\mathbb{Q}$ be the distribution of a discrete random variable on a finite set $\mathcal{X}$. Letting q denote the associated probability mass function, its Shannon entropy has the explicit formula

$$H(\mathbb{Q}) \equiv H(X) = -\sum_{x \in \mathcal{X}} q(x) \log q(x),$$

where we interpret $0 \log 0 = 0$.

(a) Show that $H(X) \geq 0$.
(b) Show that $H(X) \leq \log |\mathcal{X}|$, with equality achieved when X has the uniform distribution over $\mathcal{X}$.

Exercise 15.3 (Properties of Kullback–Leibler divergence) In this exercise, we study some properties of the Kullback–Leibler divergence. Let $\mathbb{P}$ and $\mathbb{Q}$ be two distributions having densities p and q with respect to a common base measure.

(a) Show that $D(\mathbb{P} \,\|\, \mathbb{Q}) \geq 0$ with equality if and only if the equality $p(x) = q(x)$ holds $\mathbb{P}$-almost everywhere.
(b) Given a collection of non-negative weights such that $\sum_{j=1}^m \lambda_j = 1$, show that

$$D(\sum_{j=1}^m \lambda_j \mathbb{P}_j \,\|\, \mathbb{Q}) \leq \sum_{j=1}^m \lambda_j D(\mathbb{P}_j \,\|\, \mathbb{Q}) \tag{15.63a}$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

and

$$D(\mathbb{Q} \,\|\, \sum_{j=1}^{m} \lambda_j \mathbb{P}_j) \leq \sum_{j=1}^{m} \lambda_j D(\mathbb{Q} \,\|\, \mathbb{P}_j). \tag{15.63b}$$

(c) Prove that the KL divergence satisfies the decoupling property (15.11a) for product measures.

Exercise 15.4 (More properties of Shannon entropy) Let (X, Y, Z) denote a triplet of random variables, and recall the definition (15.59) of the conditional entropy.

(a) Prove that conditioning reduces entropy—that is, $H(X \mid Y) \leq H(X)$.
(b) Prove the chain rule for entropy:

$$H(X, Y, Z) = H(X) + H(Y \mid X) + H(Z \mid Y, X).$$

(c) Conclude from the previous parts that

$$H(X, Y, Z) \leq H(X) + H(Y) + H(Z),$$

so that joint entropy is maximized by independent variables.

Exercise 15.5 (Le Cam's inequality) Prove the upper bound (15.10) on the total variation norm in terms of the Hellinger distance. (*Hint:* The Cauchy–Schwarz inequality could be useful.)

Exercise 15.6 (Pinsker–Csiszár–Kullback inequality) In this exercise, we work through a proof of the Pinsker–Csiszár–Kullback inequality (15.8) from Lemma 15.2.

(a) When $\mathbb{P}$ and $\mathbb{Q}$ are Bernoulli distributions with parameters $\delta_p \in [0, 1]$ and $\delta_q \in [0, 1]$, show that inequality (15.8) reduces to

$$2\,(\delta_p - \delta_q)^2 \leq \delta_p \log \frac{\delta_p}{\delta_q} + (1 - \delta_p) \log \frac{1 - \delta_p}{1 - \delta_q}. \tag{15.64}$$

Prove the inequality in this special case.
(b) Use part (a) and Jensen's inequality to prove the bound in the general case. (*Hint:* Letting p and q denote densities, consider the set $A := \{x \in \mathcal{X} \mid p(x) \geq q(x)\}$, and try to reduce the problem to a version of part (a) with $\delta_p = \mathbb{P}[A]$ and $\delta_q = \mathbb{Q}[A]$.)

Exercise 15.7 (Decoupling for Hellinger distance) Show that the Hellinger distance satisfies the decoupling relation (15.12a) for product measures.

Exercise 15.8 (Sharper bounds for Gaussian location family) Recall the normal location model from Example 15.4. Use the two-point form of Le Cam's method and the Pinsker–Csiszár–Kullback inequality from Lemma 15.2 to derive the sharper lower bounds

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\widehat{\theta} - \theta|] \geq \frac{1}{8} \frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2] \geq \frac{1}{16} \frac{\sigma^2}{n}.$$

Exercise 15.9 (Achievable rates for uniform shift family) In the context of the uniform shift family (Example 15.5), show that the estimator $\widetilde{\theta} = \min\{Y_1, \ldots, Y_n\}$ satisfies the bound $\sup_{\theta \in \mathbb{R}} \mathbb{E}[(\widetilde{\theta} - \theta)^2] \leq \frac{2}{n^2}$.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Exercise 15.10 (Bounds on the TV distance)

(a) Prove that the squared total variation distance is upper bounded as

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left\{ \int_{\mathcal{X}} \frac{p^2(x)}{q(x)} \nu(dx) - 1 \right\},$$

where p and q are densities with respect to the base measure ν.

(b) Use part (a) to show that

$$\|\mathbb{P}_{\theta,\sigma}^n - \mathbb{P}_{0,\sigma}^n\|_{\mathrm{TV}}^2 \leq \tfrac{1}{4}\left\{ e^{\left(\frac{\sqrt{n}\theta}{\sigma}\right)^2} - 1 \right\}, \tag{15.65}$$

where, for any $\gamma \in \mathbb{R}^n$, we use $\mathbb{P}_{\gamma,\sigma}^n$ to denote the n-fold product distribution of a $\mathcal{N}(\gamma, \sigma^2)$ variate.

(c) Use part (a) to show that

$$\|\bar{\mathbb{P}} - \mathbb{P}_{0,\sigma}^n\|_{\mathrm{TV}}^2 \leq \tfrac{1}{4}\left\{ e^{\frac{1}{2}\left(\frac{\sqrt{n}\theta}{\sigma}\right)^4} - 1 \right\}, \tag{15.66}$$

where $\bar{\mathbb{P}} = \frac{1}{2}\mathbb{P}_{\theta,\sigma}^n + \frac{1}{2}\mathbb{P}_{-\theta,\sigma}^n$ is a mixture distribution.

Exercise 15.11 (Mixture distributions and KL divergence) Given a collection of distributions $\{\mathbb{P}_1, \ldots, \mathbb{P}_M\}$, consider the mixture distribution $\bar{\mathbb{Q}} = \frac{1}{M}\sum_{j=1}^M \mathbb{P}_j$. Show that

$$\frac{1}{M}\sum_{j=1}^M D(\mathbb{P}_j \,\|\, \bar{\mathbb{Q}}) \leq \frac{1}{M}\sum_{j=1}^M D(\mathbb{P}_j \,\|\, \mathbb{Q})$$

for any other distribution $\mathbb{Q}$.

Exercise 15.12 (f-divergences) Let $f\colon \mathbb{R}_+ \to \mathbb{R}$ be a strictly convex function. Given two distributions $\mathbb{P}$ and $\mathbb{Q}$ (with densities p and q, respectively), their f-divergence is given by

$$D_f(\mathbb{P} \,\|\, \mathbb{Q}) := \int q(x) f\Big(p(x)/q(x)\Big) \nu(dx). \tag{15.67}$$

(a) Show that the Kullback–Leibler divergence corresponds to the f-divergence defined by $f(t) = t \log t$.

(b) Compute the f-divergence generated by $f(t) = -\log(t)$.

(c) Show that the squared Hellinger divergence $H^2(\mathbb{P} \,\|\, \mathbb{Q})$ is also an f-divergence for an appropriate choice of f.

(d) Compute the f-divergence generated by the function $f(t) = 1 - \sqrt{t}$.

Exercise 15.13 (KL divergence for multivariate Gaussian) For $j = 1, 2$, let $\mathbb{Q}_j$ be a d-variate normal distribution with mean vector $\mu_j \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma}_j \succ 0$.

(a) If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, show that

$$D(\mathbb{Q}_1 \,\|\, \mathbb{Q}_2) = \tfrac{1}{2}\langle \mu_1 - \mu_2, \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2) \rangle.$$

(b) In the general setting, show that

$$D(\mathbb{Q}_1 \,\|\, \mathbb{Q}_2) = \tfrac{1}{2}\left\{ \langle \mu_1 - \mu_2, \boldsymbol{\Sigma}_2^{-1}(\mu_1 - \mu_2) \rangle + \log \frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} + \operatorname{trace}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - d \right\}.$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Exercise 15.14 (Gaussian distributions and maximum entropy) For a given $\sigma > 0$, let $\mathcal{Q}_\sigma$ be the class of all densities q with respect to Lebesgue measure on the real line such that $\int_{-\infty}^{\infty} xq(x) = 0$, and $\int_{-\infty}^{\infty} q(x)x^2\,dx \leq \sigma^2$. Show that the maximum entropy distribution over this family is the Gaussian $\mathcal{N}(0, \sigma^2)$.

Exercise 15.15 (Sharper bound for variable selection in sparse PCA) In the context of Example 15.20, show that for a given $\theta_{\min} = \min_{j \in S} |\theta^*_j| \in (0, 1)$, support recovery in sparse PCA is not possible whenever

$$n < c_0 \frac{1+\nu}{\nu^2} \frac{\log(d-s+1)}{\theta^2_{\min}}$$

for some constant $c_0 > 0$. (*Note:* This result sharpens the bound from Example 15.20, since we must have $\theta^2_{\min} \leq \frac{1}{s}$ due to the unit norm and s-sparsity of the eigenvector.)

Exercise 15.16 (Lower bounds for sparse PCA in ℓ_2-error) Consider the problem of estimating the maximal eigenvector θ^* based on n i.i.d. samples from the spiked covariance model (15.47). Assuming that θ^* is s-sparse, show that any estimator $\widehat{\theta}$ satisfies the lower bound

$$\sup_{\theta^* \in \mathbb{B}_0(s) \cap \mathbb{S}^{d-1}} \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] \geq c_0 \frac{\nu+1}{\nu^2} \frac{s \log\left(\frac{ed}{s}\right)}{n}$$

for some universal constant $c_0 > 0$. (*Hint:* The packing set from Example 15.16 may be useful to you. Moreover, you might consider a construction similar to Example 15.19, but with the random orthonormal matrix $\mathbf{U}$ replaced by a random permutation matrix along with random sign flips.)

Exercise 15.17 (Lower bounds for generalized linear models) Consider the problem of estimating a vector $\theta^* \in \mathbb{R}^d$ with Euclidean norm at most one, based on regression with a fixed set of design vectors $\{x_i\}_{i=1}^n$, and responses $\{y_i\}_{i=1}^n$ drawn from the distribution

$$\mathbb{P}_\theta(y_1, \ldots, y_n) = \prod_{i=1}^n \left[h(y_i) \exp\left(\frac{y_i \langle x_i, \theta \rangle - \Phi(\langle x_i, \theta \rangle)}{s(\sigma)} \right) \right],$$

where $s(\sigma) > 0$ is a known scale factor, and $\Phi \colon \mathbb{R} \to \mathbb{R}$ is the cumulant function of the generalized linear model.

(a) Compute an expression for the Kullback–Leibler divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ involving Φ and its derivatives.

(b) Assuming that $\|\Phi''\|_\infty \leq L < \infty$, give an upper bound on the Kullback–Leibler divergence that scales quadratically in the Euclidean norm $\|\theta - \theta'\|_2$.

(c) Use part (b) and previous arguments to show that there is a universal constant $c > 0$ such that

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}\left[\|\widehat{\theta} - \theta\|_2^2\right] \geq \min\left\{1,\ c\,\frac{s(\sigma)}{L\eta^2_{\max}} \frac{d}{n}\right\},$$

where $\eta_{\max} = \sigma_{\max}(\mathbf{X}/\sqrt{n})$ is the maximum singular value. (Here as usual $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix with x_i as its ith row.)

(d) Explain how part (c) yields our lower bound on linear regression as a special case.

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press

Exercise 15.18 (Lower bounds for additive nonparametric regression) Recall the class of additive functions first introduced in Exercise 13.9, namely

$$\mathscr{F}_{\text{add}} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,\middle|\, f = \sum_{j=1}^{d} g_j \text{ for } g_j \in \mathscr{G} \right\},$$

where $\mathscr{G}$ is some fixed class of univariate functions. In this exercise, we assume that the base class has metric entropy scaling as $\log N(\delta; \mathscr{G}, \|\cdot\|_2) \asymp (\frac{1}{\delta})^{1/\alpha}$ for some $\alpha > 1/2$, and that we compute $L^2(\mathbb{P})$-norms using a product measure over $\mathbb{R}^d$.

(a) Show that

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{F}_{\text{add}}} \mathbb{E}[\|\widehat{f} - f\|_2^2] \gtrsim d \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

By comparison with the result of Exercise 14.8, we see that the least-squares estimator is minimax-optimal up to constant factors.

(b) Now consider the sparse variant of this model, namely based on the sparse additive model (SPAM) class

$$\mathscr{F}_{\text{spam}} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \,\middle|\, f = \sum_{j \in S} g_j \text{ for } g_j \in \mathscr{G}, \text{ and a subset } |S| \leq s \right\}.$$

Show that

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{F}_{\text{spam}}} \mathbb{E}[\|\widehat{f} - f\|_2^2] \gtrsim s \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}} + \sigma^2 \frac{s \log\left(\frac{ed}{s}\right)}{n}.$$

https://doi.org/10.1017/9781108627771.015 Published online by Cambridge University Press