

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 error guarantees that match the lower bounds up to constant factors. 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. 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 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 $\mathbb{P} \mapsto \theta(\mathbb{P}) \in \Omega$. 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 non-parametric problems, among them non-parametric 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} on the unit interval $[0, 1]$ with differentiable density functions f , we might be interested in 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 on the unit interval $[0, 1]$, we might be interested in estimating the mode $\theta(\mathbb{P}) = \arg \max_{x \in [0, 1]} f(x)$. Thus, the viewpoint of estimating functionals adopted here is somewhat 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 $\hat{\theta}$ 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 $\rho : \Omega \times \Omega \rightarrow [0, \infty)$ be a given error (semi)-metric, and we consider the quantity $\rho(\hat{\theta}, \theta^*)$. Here the quantity θ^* is fixed (but unknown), whereas the quantity $\hat{\theta} \equiv \hat{\theta}(X)$ is a random variable, so that $\rho(\hat{\theta}, \theta^*)$ is random. By taking expectations over the observable X , we obtain the deterministic quantity $\mathbb{E}[\rho(\hat{\theta}, \theta^*)]$. 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, more frequentist in philosophy, 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 $\hat{\theta}$, we compute the worst-case risk $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}[\rho(\hat{\theta}, \theta(\mathbb{P}))]$, and rank estimators according to this ordering. The estimator that is optimal in this sense defines a quantity known as the *minimax risk*—namely,

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

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 $\Phi : [0, \infty) \rightarrow [0, \infty)$ be an increasing function on the non-negative real line, and then defining a slight generalization of the ρ -minimax risk—namely

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

A particularly common choice is $\Phi(t) = t^2$, which can be used to obtain minimax risks for the mean-squared error. With this set-up, we now turn to the primary goal of this chapter: developing methods for obtaining lower bounds on the minimax risk.

■ 15.1.2 From estimation to testing

Having described the notion of minimax risk, we begin our exploration by showing how it can be “reduced” 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 $\{\theta^1, \dots, \theta^M\}$ is 2δ -separated set 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}_{θ^j} , and then consider the the M -ary hypothesis testing problem defined by the family of distributions $\{\mathbb{P}_{\theta^j}, j = 1, \dots, M\}$. In particular, we generate a random variable Z by the following procedure:

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

Note that the size of the packing set $M = M(\delta)$ grows as $\delta \rightarrow 0^+$.

We let \mathbb{Q} denote the joint distribution of the pair (Z, J) given by this procedure. Note that the marginal distribution \mathbb{Q}_Z over Z is given by the uniformly weighted mixture distribution $\mathbb{Q}_Z = \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j}$. 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 $\Psi : \mathcal{Z} \rightarrow [M]$, 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 probability of error as follows:

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

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

where the infimum ranges over test functions.

Note that the left-hand side of the bound (15.3) involves two terms, both of which depend on the choice of δ . As $\delta \rightarrow 0^+$, the size $M \equiv M(2\delta)$ of the 2δ -separated set increases. At least generically, a testing problem with more hypotheses is more difficult, so we should expect that $\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J]$ grows as δ decreases. If we choose a value δ^* sufficiently small to ensure that this testing error is at least $1/2$, then we may conclude that $\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta^*)/2$. The other additional degree of freedom (for a given δ) 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

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

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 $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta]$. Recall that \mathbb{Q} denotes the joint distribution over the pair (Z, J) defined by our construction. Note that

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

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

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

$$\psi(Z) := \arg \min_{j \in [M]} \rho(\theta^j, \hat{\theta}). \quad (15.4)$$

(If there are multiple indices that achieve the minimizing argument, then we break such ties arbitrarily.) Suppose that the true parameter is θ^j : we then claim that the event $\{\rho(\theta^j, \hat{\theta}) < \delta\}$ ensures that the test (15.4) is correct. To see this implication, note that for any other index $k \in [M]$, by applying the triangle inequality we have

$$\rho(\theta^k, \hat{\theta}) \geq \rho(\theta^k, \theta^j) - \rho(\theta^j, \hat{\theta}) > 2\delta - \delta = \delta,$$

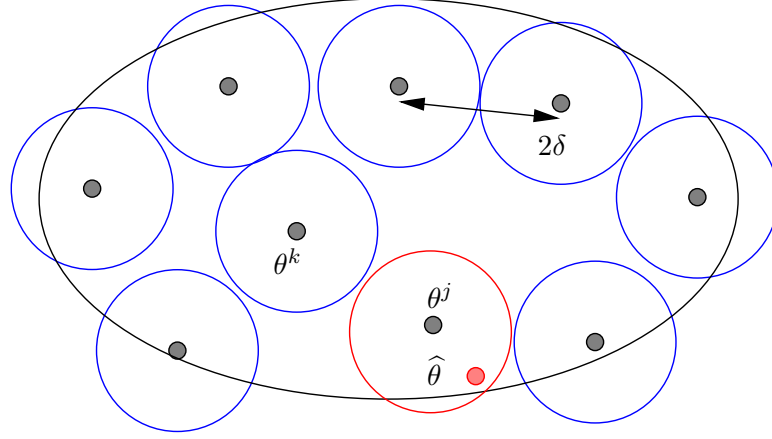


Figure 15-1. Reduction from estimation to testing using a 2δ -separated set in the space Ω in the semi-metric ρ . If an estimator $\hat{\theta}$ satisfies the bound $\rho(\hat{\theta}, \theta^j) < \delta$ whenever the true parameter is θ^j , then it can be used to construct determine the correct index j in the associated testing problem.

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

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

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

Combined with our earlier argument, we have shown that

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

Finally, we may take the the infimum over all estimators $\hat{\theta}$ 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. \square

■ 15.1.3 Some divergence measures

Thus far, we have established a connection between minimax errors 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 type of divergence measures between probability distributions. Three such measures of particular impor-

tance are the total variation (TV) distance, the Kullback-Leibler (KL) divergence, and the Hellinger distance.

Let \mathbb{P} and \mathbb{Q} be two distributions with densities p and q with respect to some underlying base measure ν . Note that this assumption entails 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

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

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

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

corresponding to $1/2$ the $L^1(\nu)$ -norm between the densities. (See Exercise 3.11 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.

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

$$D(\mathbb{Q} \parallel \mathbb{P}) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} \nu(dx). \quad (15.7)$$

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 $D(\mathbb{Q} \parallel \mathbb{P}) \neq D(\mathbb{P} \parallel \mathbb{Q})$). However, it can be used to upper bound the TV distance, as stated in the following classical result:

Lemma 15.1 (Pinsker-Csiszar-Kullback inequality). For all distributions \mathbb{P} and \mathbb{Q} ,

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D(\mathbb{Q} \parallel \mathbb{P})}. \quad (15.8)$$

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.

A third distance that plays an important role in statistical problems is the *squared*

Hellinger distance, given by

$$H^2(\mathbb{P} \parallel \mathbb{Q}) := \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \nu(dx). \quad (15.9)$$

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 $[0, 2]$.

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

Lemma 15.2 (Le Cam's inequality). For all distributions \mathbb{P} and \mathbb{Q} ,

$$\|\mathbb{P} - \mathbb{Q}\|_{\text{TV}} \leq H(\mathbb{P} \parallel \mathbb{Q}) \sqrt{1 - \frac{H^2(\mathbb{P} \parallel \mathbb{Q})}{4}}. \quad (15.10)$$

We work through the proof of this inequality in Exercise 15.4.

Let $(\mathbb{P}_1, \dots, \mathbb{P}_n)$ be a collection of n probability measures, each defined on \mathcal{X} , and let $\mathbb{P}^n = \otimes_{i=1}^n \mathbb{P}_i$ be the product probability defined on \mathcal{X}^n . If we define another product measure \mathbb{Q}^n in a similar manner, then it is natural to ask whether the divergence between \mathbb{P}^n and \mathbb{Q}^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 express the distance $\|\mathbb{P}^n - \mathbb{Q}^n\|_{\text{TV}}$ in terms of the individual distances $\|\mathbb{P}_i - \mathbb{Q}_i\|_{\text{TV}}$. On the other hand, the Kullback-Leibler divergence exhibits a very attractive decoupling property, in that we have

$$D(\mathbb{P}^n \parallel \mathbb{Q}^n) = \sum_{i=1}^n D(\mathbb{P}_i \parallel \mathbb{Q}_i). \quad (15.11)$$

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

$$H^2(\mathbb{P}^n \parallel \mathbb{Q}^n) = 2 - 2 \prod_{i=1}^n \left(1 - \frac{1}{2} H^2(\mathbb{P}_i \parallel \mathbb{Q}_i)\right). \quad (15.12)$$

See Exercises 15.3 and 15.5 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

- 1 and the total variation norm, and use it to develop various lower bounds, culminating
 2 in a general technique known as Le Cam's method.

3 ■ 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 $\frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. For a given decision rule $\psi : \mathcal{Z} \rightarrow \{0, 1\}$, the associated probability of error is given by

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

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 $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}}$, as previously defined in equation (15.5)—more precisely, we have

$$\inf_{\psi} \mathbb{P}[\psi(Z) \neq J] = \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}}. \quad (15.13)$$

- 4 Note that the worst-case value of the Bayes risk is $\frac{1}{2}$, achieved when $\mathbb{P}_1 = \mathbb{P}_0$ so that the
 5 hypotheses are completely indistinguishable. At the other extreme, the best-case Bayes
 6 risk is 0, achieved when $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\text{TV}} = 1$. This latter equality occurs, for instance,
 7 when \mathbb{P}_0 and \mathbb{P}_1 have disjoint supports.

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} —in particular, any rule ψ is uniquely determined by the set $A = \{x \in \mathcal{X} \mid \psi(x) = 1\}$. Thus, we have

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

- 8 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

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

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

10

- 11 **Example 15.1** (Gaussian location family). For a fixed variance σ^2 , let \mathbb{P}_{θ} be the
 12 distribution of a $\mathcal{N}(\theta, \sigma^2)$ variable; if we let the mean θ vary, it defines the Gaussian
 13 location family $\{\mathbb{P}_{\theta}, \theta \in \mathbb{R}\}$. Here we consider the problem of estimating θ under the

squared error $(\theta' - \theta)^2$ using a collection $Z = (Y_1, \dots, Y_n)$ of n i.i.d. samples. 1

Let \mathbb{P}_θ^n denote the product distribution over these n samples. By an application of the result of Exercise 15.9, the Kullback-Leibler divergence between \mathbb{P}_θ^n and $\mathbb{P}_{\theta'}^n$ takes the form

$$D(\mathbb{P}_\theta^n \parallel \mathbb{P}_{\theta'}^n) = \frac{n}{2\sigma^2}(\theta - \theta')^2.$$

Consequently, using Lemma 15.1, we find that $\|\mathbb{P}_\theta^n - \mathbb{P}_{\theta'}^n\|_{\text{TV}} \leq \frac{\sqrt{n}}{2\sigma}|\theta - \theta'|$. 2

Accordingly, if we take a pair θ and θ' such that $|\theta - \theta'| = 2\delta := \frac{\sigma}{\sqrt{n}}$, then we are guaranteed $\|\mathbb{P}_\theta^n - \mathbb{P}_{\theta'}^n\|_{\text{TV}} \leq \frac{1}{2}$, and hence, using the lower bound (15.14) with $\Phi(t) = t^2$,

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \frac{1}{8} \frac{\sigma^2}{n}.$$

Although the pre-factor $1/8$ is not optimal, the scaling σ^2/n is sharp. For instance, the sample mean $\tilde{\theta}_n := \frac{1}{n} \sum_{i=1}^n Y_i$ satisfies the bound $\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\tilde{\theta}_n - \theta)^2] = \frac{\sigma^2}{n}$. 3

♣ 5

The n^{-1} mean-squared error in this example 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: 6

Example 15.2 (Uniform shift family). Let us consider the uniform shift family, in which for each $\theta \in \mathbb{R}$, the distribution \mathbb{U}_θ is uniform over the interval $[\theta, \theta + 1]$. We let \mathbb{U}_θ^n denote the product distribution of n i.i.d. samples from \mathbb{U}_θ . In this case, Lemma 15.1 is no longer useful in controlling the total variation norm, since the Kullback-Leibler divergence between \mathbb{U}_θ and $\mathbb{U}_{\theta'}$ is infinite whenever $\theta \neq \theta'$. Accordingly, we need to use an alternative distance measure: here we illustrate the use of the Hellinger distance. 12

We begin by computing the Hellinger distance between \mathbb{U}_θ and $\mathbb{U}_{\theta'}$. By symmetry, we may assume without loss of generality that $\theta < \theta'$. If $\theta' > \theta + 1$, then we have $H^2(\mathbb{U}_\theta \parallel \mathbb{U}_{\theta'}) = 2$. Otherwise, when $\theta' \in (\theta, \theta + 1]$, we have 13

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

Consequently, if we take a pair θ, θ' such that $|\theta' - \theta| = 2\delta := \frac{1}{4n}$, then the relation (15.12) guarantees 14

$$\frac{1}{2} H^2(\mathbb{U}_\theta^n \parallel \mathbb{U}_{\theta'}^n) = 1 - \left(1 - \frac{1}{4n}\right)^n \leq \frac{1}{4}$$

In conjunction with Lemma 15.2, we find that

$$\|\mathbb{U}_\theta^n - \mathbb{U}_{\theta'}^n\|_{\text{TV}} \leq \frac{1}{2} \sqrt{1 - \frac{1}{16}} \leq \frac{1}{2}.$$

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

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

- 1 The significant aspect of this lower bound is the faster n^{-2} rate, to be contrasted
 2 with the n^{-1} rate in the regular situation. In fact, this n^{-2} rate is optimal for the
 3 uniform location model, achieved for instance by the estimator $\tilde{\theta} = \min\{Y_1, \dots, Y_n\}$;
 4 see Exercise 15.6 for details. ♣

- 5 We now turn to the use of two-class lower bound for a non-parametric problem.
 6 Although it does lead to a non-trivial lower bound, it is not a sharp result, unlike in
 7 the previous examples. Later, we will develop Le Cam's refinement of the two-class
 8 approach so as to obtain sharp rates.

9

Example 15.3 (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

$$\mathcal{F}_2([0, 1]) = \{f : [0, 1] \rightarrow [c_0, c_1] \mid \|f''\|_\infty \leq c_2, \int_0^1 f(x) dx = 1\}. \quad (15.15)$$

- 10 that are uniformly bounded above and below, and have a uniformly bounded sec-
 11 ond derivative. Suppose that we are interested in estimating the quadratic functional
 12 $f \mapsto \theta(f) := \int_0^1 (f'(x))^2 dx$. Note that $\theta(f)$ provides a measure of the “smoothness” of
 13 the density: it is zero for the uniform density, and becomes large for densities with
 14 more erratic behavior. Estimation of such quadratic functionals arises in a variety of
 15 applications; see the bibliographic section for further discussion.

Let \mathbb{U} denote the uniform distribution on $[0, 1]$; its density is given by $u(x) = 1$ for all $x \in [0, 1]$ and so belongs to \mathcal{F}_2 . We derive a lower bound by comparing \mathbb{U} to another distribution \mathbb{Q} , constructed by suitably perturbing the uniform distribution. In order to construct this perturbation, let $g : [0, 1] \rightarrow \mathbb{R}$ be a fixed twice differentiable function such that $\|g\|_\infty \leq 1/2$,

$$\int_0^1 g(x) dx = 0, \quad \text{and} \quad b_j := \int_0^1 (g^{(j)}(x))^2 dx > 0 \quad \text{for } j = 0, 1.$$

Now divide the unit interval $[0, 1]$ into m subintervals $[x_j, x_{j+1}]$ where $x_j = \frac{j}{m}$, for

$j = 0, \dots, m-1$. For a suitably small constant $C > 0$, define the shifted and rescaled functions

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

We then consider the distribution \mathbb{Q} with density $q(x) = 1 + \sum_{j=1}^m g_j(x)$, and note that $q \in \mathcal{F}_2$ as long as C is chosen sufficiently small.

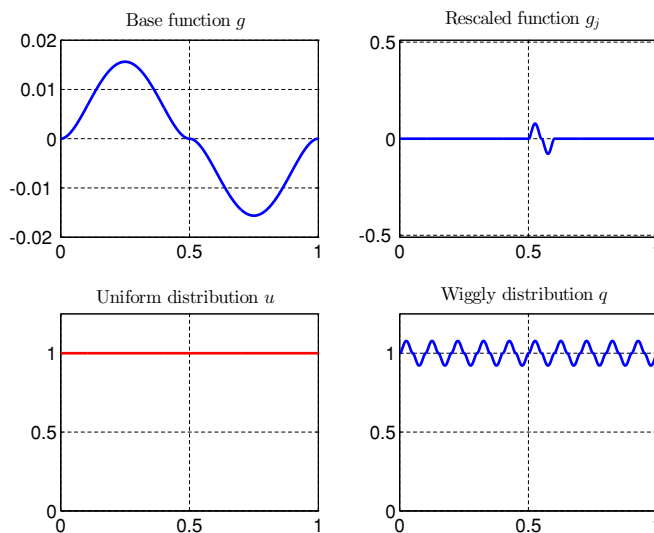


Figure 15-2. Illustration of the construction of the density q . Upper left panel: an example of a base function g . Upper right: function g_j is a rescaled and shifted version of g . Lower left: original uniform distribution u . Lower right: final density q is the superposition of u with the sum of the shifted functions $\{g_j\}_{j=1}^m$.

Let us control the total variation distance via the Hellinger distance. Since the perturbation $q(x) - 1 = \sum_{j=1}^m g_j(x)$ is uniformly bounded, a Taylor series expansion of $\sqrt{1+t}$ around 1 yields such that

$$\begin{aligned} \frac{1}{2} H^2(\mathbb{U} \parallel \mathbb{Q}) &= 1 - \int_0^1 \sqrt{1 + (q(x) - 1)} dx \leq \int_0^1 \left\{ -\frac{\sum_{j=1}^m g_j(x)}{2} + c \left(\sum_{j=1}^m g_j(x) \right)^2 \right\} dx \\ &= c \sum_{j=1}^m \int_0^1 g_j^2(x) dx = c b_0 \frac{1}{m^4}, \end{aligned}$$

where $c > 0$ is a universal constant. Consequently, using the decomposition (15.12), we

have $\frac{1}{2}H^2(\mathbb{U}^n \parallel \mathbb{Q}^n) \leq 1 - 1(1 - \frac{cb_0}{m^4})^n$. Setting $m^4 = 4cb_0n$ ensures that $\frac{1}{2}H^2(\mathbb{U}^n \parallel \mathbb{Q}^n) \leq 1 - (1 - \frac{1}{4n})^n \leq 1/4$, and hence $\|\mathbb{U}^n - \mathbb{Q}^n\|_{\text{TV}} \leq 1/2$. On the other hand, we have $\theta(u) = 0$ and

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

so with the specified choice of m , we have $|\theta(q) - \theta(u)| \geq \frac{K}{\sqrt{n}}$. for some universal constant K independent of n . Consequently, the lower bound (15.14) implies that

$$\sup_{f \in \mathcal{F}_2} \mathbb{E}[|\hat{\theta}(f) - \theta(f)|] \gtrsim n^{-1/2} \quad (15.16)$$

- 1 This lower bound, while valid, is *not optimal*—there is no estimator that can achieve
 2 error of the order $n^{-1/2}$ uniformly over \mathcal{F}_2 . Indeed, the optimal lower bound is $n^{-4/9}$;
 3 it can be obtained via an extension of the basic two-point technique, as we describe in
 4 the next section. ♣

5 ■ 15.2.2 Le Cam's method

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

$$\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta \quad \text{for all } \mathbb{P}_0 \in \mathcal{P}_0 \text{ and } \mathbb{P}_1 \in \mathcal{P}_1. \quad (15.17)$$

Lemma 15.3 (Le Cam). For any 2δ -separated classes of distributions \mathcal{P}_0 and \mathcal{P}_1 contained with \mathcal{P} , and any estimator $\hat{\theta}$, we have

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\hat{\theta}, \theta(\mathbb{P}))] \geq \frac{\delta}{2} \sup_{\substack{\mathbb{P}_0 \in \text{conv}(\mathcal{P}_0) \\ \mathbb{P}_1 \in \text{conv}(\mathcal{P}_1)}} \{1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}}\}. \quad (15.18)$$

Proof. For any estimator $\hat{\theta}$ and any pair of distributions $\mathbb{P}_j \in \mathcal{P}_j$, $j = 0, 1$, let us define

the random variables $V_j(\hat{\theta}) = \frac{1}{2\delta} \inf_{\mathbb{P}_j \in \mathcal{P}_j} \rho(\hat{\theta}, \theta(\mathbb{P}_j))$. We then have

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

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

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

By the triangle inequality, we have

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

Taking infima over \mathcal{P}_j and recalling the definition of $V_j(\hat{\theta})$, we obtain the lower bound $V_0(\hat{\theta}) + V_1(\hat{\theta}) \geq 1$. Since $V_j(\hat{\theta}) \geq 0$ 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

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

which completes the proof. □ 1

In Example 15.3, we investigated the problem of estimating the quadratic functional $f \mapsto \theta(f) = \int_0^1 (f'(x))^2 dx$ over the class \mathcal{F}_2 from equation (15.15). 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. 2
3
4
5

Example 15.4 (Optimal bounds for quadratic functionals). For each $\alpha \in \{-1, +1\}^m$, define the distribution \mathbb{Q}_α with density $q_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j g_j(x)$. Note that our previous choice of q from Example 15.3 was a special member of this family, with $\alpha = (1, 1, \dots, 1)$. Let \mathbb{Q}_α^n denote the product distribution on \mathcal{X}^n formed by sampling n times independently from \mathbb{Q}_α , and define the two classes $\mathcal{P}_0 = \{\mathbb{U}^n\}$ and $\mathcal{P}_1 = \{\mathbb{Q}_\alpha^n, \alpha \in \{-1, +1\}^m\}$. With these choices, we then have

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

where $\bar{\mathbb{Q}}^n := 2^{-m} \sum_{\alpha \in \{-1, +1\}^m} \mathbb{Q}_\alpha^n$ is the uniformly weighted mixture over all 2^m choices of \mathbb{Q}_α^n . 6
7

Unfortunately, the mixture distribution $\bar{\mathbb{Q}}^n$ is no longer a product distribution,

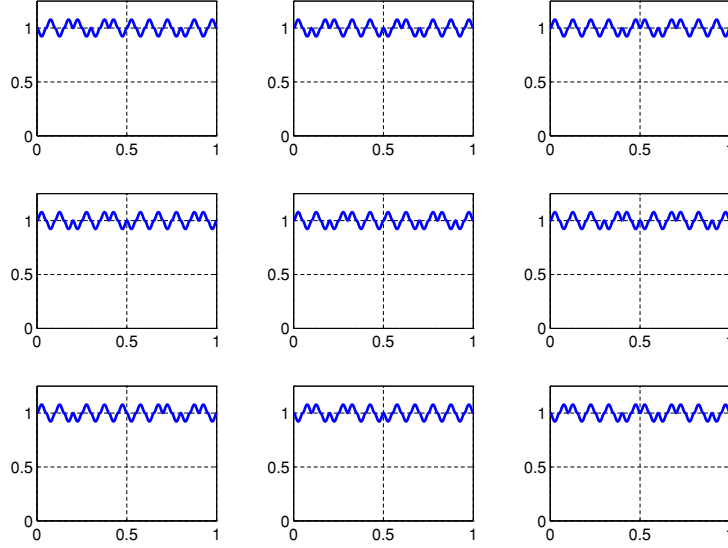


Figure 15-3. Illustration of some densities of the form $q_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j g_j(x)$ for different choices of sign vectors $\alpha \in \{-1, 1\}^m$. Note that there are 2^m such densities in total.

so that it cannot be reduced to a univariate problem by applying the decomposition (15.12). In this case, some more technical calculations are required, but one possible upper bound is given by

$$H^2(\mathbb{U}^n \parallel \bar{\mathbb{Q}}^n) \leq n^2 \sum_{j=1}^m \left(\int_0^1 g_j^2(x) dx \right)^2 \quad (15.19)$$

See the bibliographic section for discussion of this upper bound as well as related results. If we take the upper bound (15.19) as given, then using the calculations from Example 15.3, we find that

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

Setting $m^9 = 4b_0^2 n^2$ yields that $\|\mathbb{U}^n - \bar{\mathbb{Q}}^n\|_{\text{TV}} \leq H(\mathbb{U}^n \parallel \mathbb{P}^n) \leq 1/2$, and hence Lemma 15.3 implies that

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

- 1 Thus, by using the full convex form of Le Cam's method, we have recovered a tighter
- 2 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.



1

■ 15.3 Fano's method

2

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.

4

■ 15.3.1 Kullback-Leibler divergence and mutual information

5

Recall our basic set-up: we are interested in lower bounding the probability of error in an M -ary hypothesis testing problem, in which we observe a sample Z and would like to identify the index J of the probability distribution from which it is 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 it would have no value whatsoever.

11

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 the joint distribution $\mathbb{P}_{Z,J}$ is equal to the product of its marginals—namely, $\mathbb{P}_Z \mathbb{P}_J$. 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:

$$I(Z, J) := D(\mathbb{P}_{Z,J} \| \mathbb{P}_Z \mathbb{P}_J). \quad (15.20)$$

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.

12

13

When (as in our case) the random variable J is discrete, taking values over a finite index set $\{1, \dots, M\}$, then the mutual information has an alternative interpretation in terms of the *finite mixture distribution*

$$\bar{\mathbb{P}} := \frac{1}{M} \sum_{j=1}^M \mathbb{P}^j, \quad (15.21)$$

where \mathbb{P}^j denotes the conditional distribution of Z given $J = j$. Note that $\bar{\mathbb{P}}$ corresponds to the marginal distribution of Z . From the definition of the KL divergence, it is easy to see that the mutual information can be written in terms of $\bar{\mathbb{P}}$ —in particular as

$$I(Z; J) = \frac{1}{M} \sum_{j=1}^M D(\mathbb{P}^j \| \bar{\mathbb{P}}), \quad (15.22)$$

1 corresponding to the mean KL divergence between \mathbb{P}^j and $\bar{\mathbb{P}}$, averaged over the choice
 2 of index j . Consequently, the mutual information is small if the typical conditional
 3 distribution \mathbb{P}^j is hard to distinguish from the mixture distribution $\bar{\mathbb{P}}$.

4 ■ 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:

$$\mathbb{P}[\psi(Z) \neq J] \geq 1 - \frac{I(Z; J) + \log 2}{\log M}. \quad (15.23)$$

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

7 **Proposition 15.2.** Let $\{\theta^1, \dots, \theta^M\}$ 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 $\{1, \dots, M\}$, and $(Z \mid J = j) \sim \mathbb{P}_{\theta^j}$. Then for any increasing function $\Phi : [0, \infty) \rightarrow [0, \infty)$, the minimax risk is lower bounded as

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z; J) + \log 2}{\log M} \right\}, \quad (15.24)$$

8 where $I(Z; J)$ is the mutual information between Z and J .
 9

We provide a proof of the Fano bound (15.23) (and hence Proposition 15.2) in Section 15.3.5 to follow. For the moment, in order to gain intuition for this result, it is helpful to consider the behavior of the different terms of $\delta \rightarrow 0^+$. 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

$$\frac{I(Z; J) + \log 2}{\log M} \leq \frac{1}{2}, \quad (15.25)$$

10 so that the lower bound (15.24) implies that $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$. Thus, we have a
 11 generic scheme for deriving lower bounds on the minimax risk.

12 In order to derive lower bounds in this way there remain two technical and possi-
 13 bly challenging steps. The first requirement is to specify 2δ -separated sets with large
 14 cardinality $M(2\delta)$. Here the theory of metric entropy developed in Chapter 5 plays
 15 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 (15.22), we find that

$$I(Z; J) \leq \frac{1}{M^2} \sum_{j,k=1}^M D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\theta^k}). \quad (15.26)$$

Consequently, if we can construct a 2δ -separated set such that all pairs of distributions \mathbb{P}^j and \mathbb{P}^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 15.5 (Normal location model via Fano method). Recall from Example 15.1 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 we 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 $\mathbb{P}_{\theta^j} \sim \mathcal{N}(\theta^j, \sigma^2)$, we have

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

The bound (15.26) 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 (15.24) with $\Phi(t) = t^2$ implies that

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

In this way, we have re-derived a minimax lower bound of the order σ^2/n , which as discussed in Example 15.1, 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 somewhat misleading, because the method is actually based on a substantial weakening of the Fano bound, obtained from the inequality (15.26).)

The local packing approach proceeds as follows. Suppose that we can construct a

2δ -packing of Ω such that, for some quantity c , the Kullback-Leibler divergences satisfy the uniform upper bound

$$\sqrt{D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\theta^k})} \leq c\sqrt{n}\delta \quad \text{for all } j \neq k. \quad (15.27a)$$

The bound (15.26) then implies that $I(Z; J) \leq c^2\delta^2$, and hence the bound (15.25) will hold as long as

$$\log M(2\delta) \geq 2\{c^2n\delta^2 + \log 2\}. \quad (15.27b)$$

- 1 In summary, if we can find a 2δ -separated family of distributions such that condi-
- 2 tions (15.27a) and (15.27b) both hold, then we may conclude that $\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$.
- 3
- 4 Let us illustrate the local packing approach with some examples.

5 **Example 15.6** (Minimax risks for linear regression). Consider the standard linear
 6 regression model $y = \mathbf{X}\theta^* + w$ where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a fixed design matrix, and the vector
 7 $w \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ is observation noise. Viewing the design matrix \mathbf{X} as fixed, let us obtain
 8 lower bounds on the minimax risk in the prediction (semi)-norm $\rho_{\mathbf{X}}(\hat{\theta}, \theta^*) := \frac{\|\mathbf{X}(\hat{\theta} - \theta^*)\|_2}{\sqrt{n}}$,
 9 assuming that θ^* is allowed to vary over \mathbb{R}^d .

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

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

and let $\{\gamma^1, \dots, \gamma^M\}$ be a 2δ -packing in the ℓ_2 -norm. Since this set sits in a space of dimension $r = \text{rank}(\mathbf{X})$, Lemma 5.2 implies that we can find such a packing with $\log M \geq r \log 2$ elements. We thus have a collection of vectors of the form $\gamma^j = \frac{\mathbf{X}}{\sqrt{n}}\theta^j$ for some $\theta^j \in \mathbb{R}^d$, and such that $\|\gamma^j\|_2 = \frac{\|\mathbf{X}\theta^j\|_2}{\sqrt{n}} \leq 4\delta$, and

$$2\delta \leq \frac{\|\mathbf{X}(\theta^j - \theta^k)\|_2}{\sqrt{n}} \leq 8\delta \quad \text{for each } j \neq k \in [M]. \quad (15.28)$$

Let \mathbb{P}^j denote the distribution of y when θ^j is the true regression vector. Noting that $y \sim \mathcal{N}(\mathbf{X}\theta^j, \sigma^2 \mathbf{I}_n)$, the result of Exercise 15.9 ensures that

$$\sqrt{D(\mathbb{P}^j \parallel \mathbb{P}^k)} = \frac{\sqrt{n}}{\sqrt{2}\sigma} \|\mathbf{X}(\theta^j - \theta^k)\|_2 \leq \frac{8n}{\sqrt{2}\sigma} \delta, \quad (15.29)$$

where the inequality follows from the upper bound (15.28). Consequently, the lower

bound (15.27b) is satisfied by setting $\delta^2 = \frac{\sigma^2}{64} \frac{r}{n}$, and we conclude that

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E} \left[\frac{1}{n} \|\mathbf{X}(\hat{\theta} - \theta)\|_2^2 \right] \geq \frac{\sigma^2}{128} \frac{\text{rank}(\mathbf{X})}{n}.$$



1

Let us now see how the upper bound (15.26) and Fano's method can be applied to a non-parametric problem.

2

3

Example 15.7 (Minimax risk for density estimation). Recall from equation (15.15) the family \mathcal{F}_2 of twice-smooth densities on $[0, 1]$, bounded uniformly above, and 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 .

4

5

6

7

8

In order to construct a local packing, we make use of the family of perturbed densities from Example 15.4, each of the form $q_\alpha(x) = 1 + \sum_{j=1}^m \alpha_j g_j(x)$, where $\alpha \in \{-1, +1\}^m$ and g_j was previously defined. Although there are 2^m such perturbed densities, it is convenient to use only a well-separated subset of them. Let $M_H(1/4; \mathbb{H}^m)$ denote the $\frac{1}{4}$ -packing number of the binary hypercube $\{-1, +1\}^m$ in the rescaled Hamming metric. From our calculations in Example 5.2, we know that

$$\log M_H\left(\frac{1}{4}; \mathbb{H}^m\right) \geq m D\left(\frac{1}{4} \parallel \frac{1}{2}\right) \geq \frac{m}{10}.$$

(See in particular equation (5.6).) Consequently, we can find a subset $\mathbb{T} \subset \{-1, +1\}^m$ with cardinality at least $e^{m/10}$ such that

$$d_H(\alpha, \beta) = \frac{1}{m} \sum_{j=1}^m \mathbb{I}[\alpha_j \neq \beta_j] \geq 1/4 \quad \text{for all } \alpha \neq \beta \in \mathbb{T}. \quad (15.30)$$

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

9

10

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

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

But on the interval I_j , we have $(\sqrt{q_\alpha(x)} + \sqrt{q_\beta(x)})^2 = 2(q_\alpha(x) + q_\beta(x)) \leq 4$ and

therefore

$$\int_{I_j} (\sqrt{q_\alpha(x)} - \sqrt{q_\beta(x)})^2 dx \geq \frac{1}{4} \int_{I_j} (q_\alpha(x) - q_\beta(x))^2 \geq \int_{I_j} g_j^2(x) dx \quad \text{whenever } \alpha_j \neq \beta_j.$$

- 1 Since $\int_{I_j} g_j^2(x) dx = \int_0^1 g^2(x) dx = \frac{b_0}{m^5}$ and any distinct $\alpha \neq \beta$ differ in at least $m/4$
 2 positions, we find that $H^2(\mathbb{Q}_\alpha \| \mathbb{Q}_\beta) \geq \frac{m}{4} \frac{b_0}{m^5} = \frac{b_0}{m^4} \equiv 4\delta^2$. Consequently, we have
 3 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 $q_\alpha(x) \geq 1/2$ for all $x \in [0, 1]$, and thus

$$D(\mathbb{Q}_\alpha \| \mathbb{Q}_\beta) \leq \int_0^1 \frac{(\sqrt{q_\alpha(x)} - \sqrt{q_\beta(x)})^2}{q_\alpha(x)} dx \leq 2 \int_0^1 (\sqrt{q_\alpha(x)} - \sqrt{q_\beta(x)})^2 dx \leq \frac{4b_0}{m^4}, \quad (15.31)$$

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

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

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

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

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

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

Example 15.8 (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

$$\mathbb{S}^d(s) := \mathbb{B}_0^d(s) \cap \mathbb{B}_2(1) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\} \quad (15.32)$$

- 10 of s -sparse vectors within the Euclidean unit ball.

Let us first construct $1/2$ -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 15.6 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


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

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

$$\mathfrak{M}(\mathbb{S}^d(s); \|\cdot\|_2) \gtrsim \frac{\sigma^2}{\gamma_{2s}^2} s \log \frac{ed}{s}. \quad (15.33)$$

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 (15.26) 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 15.4. Suppose J is uniformly distributed over $[M] = \{1, \dots, M\}$ and that Z conditioned on $J = j$ has a Gaussian distribution with covariance Σ^j . Then the mutual information is upper bounded as

$$I(Z; J) \leq \frac{1}{2} \log \det \text{cov}(Z) - \frac{1}{2M} \sum_{j=1}^M \log \det(\Sigma^j). \quad (15.34)$$

This upper bound is a consequence of the maximum entropy property of the multivariate Gaussian distribution; see Exercise 15.10 for further details. In the special case when

$\Sigma^j = \Sigma$ for all $j \in [M]$, it takes on the simpler form

$$I(Z; J) \leq \frac{1}{2} \log \left(\frac{\det \text{cov}(Z)}{\det(\Sigma)} \right). \quad (15.35)$$

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

2

3 **Example 15.9** (Variable selection in sparse linear regression). Let us return to the
 4 model of sparse linear regression from Example 15.8, based on the standard linear
 5 model $y = \mathbf{X}\theta^* + w$, where the unknown regression vector $\theta^* \in \mathbb{R}^d$ is s -sparse. Here
 6 consider the problem of lower bounding the minimax risk for the problem of variable
 7 selection—namely, determining the support set $S = \{j \in \{1, 2, \dots, d\} \mid \theta_j^* \neq 0\}$, a set
 8 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 possible $\binom{d}{s}$ possible subsets. Consequently, we can use Fano's inequality to derive lower bounds, and we do so by constructing various ensembles of sub-problems. These sub-problems 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 in order to achieve a probability of error less than $1/2$, any method requires a sample size of at least

$$n > \max \left\{ 8 \frac{\log(d + s - 1)}{\log(1 + \frac{\theta_{\min}^2}{\sigma^2})}, \quad 8 \frac{\log \binom{d}{s}}{\log(1 + s \frac{\theta_{\min}^2}{\sigma^2})} \right\}. \quad (15.36)$$

We derive these 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 $I_x(Y; J)$ with the covariates fixed. In particular, we have

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

9 so that by taking averages over \mathbf{X} , we can obtain lower bounds that involve the quantity
 10 $\mathbb{E}_{\mathbf{X}}[I_{\mathbf{X}}(Y; J)]$.

Ensemble A: Consider the class of $M = \binom{d}{s}$ all possible subsets of cardinality s , enumerated in some fixed way. For the ℓ^{th} subset S^ℓ , let $\theta^\ell \in \mathbb{R}^d$ have values θ_{\min} for all indices $j \in S^\ell$, and zeroes in all other positions. With a fixed covariate vector x , a given observation Y then has the mixture distribution $\frac{1}{M} \sum_{\ell=1}^M \mathbb{P}_{\theta^\ell}$, where \mathbb{P}_{θ^ℓ} is the distribution of a $\mathcal{N}(\langle x, \theta^\ell \rangle, \sigma^2)$ random variable. Since the n samples are independent, we have

$$I_x(Y_1, \dots, Y_n; J) = \sum_{i=1}^n \text{var}_x(Y_i; J) \leq \frac{n}{2} \sum_{i=1}^n \log \frac{\text{var}(Y_i \mid x_i)}{\sigma^2}$$

where the inequality follows by applying Lemma 15.4 with $Z = Y_i$, conditioned on $X_i = x_i$. Taking averages over \mathbf{X} and using the i.i.d. nature of the samples along with the concavity of the logarithm, we have

$$\mathbb{E}_{\mathbf{X}}[I_{\mathbf{X}}(Y_1, \dots, Y_n; J)] \leq \frac{n}{2} \log \frac{\mathbb{E}_{X_1}[\text{var}(Y_1 | X_1)]}{\sigma^2}.$$

It remains to upper bound the variance term. Since Y_1 is a mixture distribution with M components, we have

$$\begin{aligned} \mathbb{E}_X[\text{var}(Y_1 | X_1)] &\leq \mathbb{E}_X[\mathbb{E}[Y_1^2 | X_1]] = \mathbb{E}_X\left[X_1^T \left\{ \frac{1}{M} \sum_{j=1}^M \theta^j \otimes \theta^j \right\} X_1 + \sigma^2\right] \\ &= \text{trace} \left(\frac{1}{M} \sum_{j=1}^M (\theta^j \otimes \theta^j) \right) + \sigma^2. \end{aligned}$$

Now each index $j \in \{1, 2, \dots, d\}$ appears in $\binom{d-1}{s-1}$ of the total number of subsets $M = \binom{d}{s}$, so that $\text{trace} \left(\frac{1}{M} \sum_{j=1}^M \theta^j \otimes \theta^j \right) = d \frac{\binom{d-1}{s-1}}{\binom{d}{s}} \theta_{\min}^2 = s \theta_{\min}^2$. Putting together the pieces, we conclude that

$$\mathbb{E}_{\mathbf{X}}[I_{\mathbf{X}}(Y_1, \dots, Y_n; J)] \leq \frac{n}{2} \log \left(1 + \frac{s \theta_{\min}^2}{\sigma^2} \right),$$


and hence the Fano lower bound implies that

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

from which the first lower bound follows as long as $\frac{\log 2}{\log \binom{d}{s}} \leq 1/4$. 1

Ensemble B: Let $\tilde{\theta} \in \mathbb{R}^d$ be a vector with θ_{\min} in its first $s-1$ co-ordinates, and zero in all remaining $d-s+1$ co-ordinates. For each $j = 1, \dots, d$, let $e_j \in \mathbb{R}^d$ denote the j^{th} standard basis vector with a single one in position j . Define the family of $M = d-s+1$ vectors $\theta^j := \tilde{\theta} + \theta_{\min} e_j$ for $j = s, \dots, d$. By a straightforward calculation, we have $\mathbb{E}[Y | x] = \langle x, \gamma \rangle$, where $\gamma := \tilde{\theta} + \frac{1}{M} \theta_{\min} e_{s \rightarrow d}$, and the vector $e_{s \rightarrow d} \in \mathbb{R}^d$ has ones in positions s through d . By the same argument as for ensemble A, it suffices to upper bound the quantity $\mathbb{E}_X[\text{var}(Y_1 | X_1)]$. Using the definition of our ensemble, we have

$$\mathbb{E}_{X_1}[\text{var}(Y_1 | X_1)] = \sigma^2 + \text{trace} \left\{ \frac{1}{M} \sum_{j=1}^M (\theta^j \otimes \theta^j - \gamma \otimes \gamma) \right\} \leq \sigma^2 + \theta_{\min}^2. \quad (15.37)$$

1 Let us suppose that $d - s + 1 > 16$, which ensures that $\frac{\log 2}{\log(d-s+1)} < 1/4$. Using Fano's
 2 inequality and the upper bound (15.37), the second component (15.36) of the lower
 3 bound 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

$$x \stackrel{d}{=} \sqrt{\nu} \xi \theta^* + w. \quad (15.38)$$

4 Here $\nu > 0$ is a given signal-to-noise ratio, θ^* is a unit norm vector, and the random
 5 quantities $\xi \sim \mathcal{N}(0, 1)$ and $w \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent. By construction, the random
 6 vector x is zero-mean Gaussian with a covariance matrix of the form $\mathbf{I}_d + \nu(\theta^* \otimes \theta^*)$,
 7 and our goal is to estimate the leading eigenvector θ^* of this covariance matrix based on
 8 n i.i.d. samples of x . In the following example, we derive lower bounds on the minimax
 9 risk in the squared Euclidean norm $\|\hat{\theta} - \theta^*\|_2^2$. (As discussed in Chapter 8, recall that
 10 there is always a sign ambiguity in estimating eigenvectors, so that in computing the
 11 Euclidean norm, we implicitly assume that the correct direction is chosen.)

Example 15.10 (Lower bounds for PCA). Let $\{\Delta^1, \dots, \Delta^M\}$ be a $1/2$ -packing of the unit sphere in \mathbb{R}^{d-1} ; from Example 5.4, 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

$$\theta^j(\mathbf{U}) = \sqrt{1 - \delta^2} \begin{bmatrix} 1 \\ 0_{d-1} \end{bmatrix} + \delta \begin{bmatrix} 0 \\ \mathbf{U} \Delta^j \end{bmatrix}, \quad \text{for } j \in [M], \quad (15.39)$$

12 where 0_{d-1} is a vector of $d-1$ zeroes. By construction, each vector $\theta^j(\mathbf{U})$ lies on the
 13 sphere in \mathbb{R}^d , and the collection of all M vectors forms $\delta/2$ -packing set. Consequently,
 14 we can construct a testing problem based on the vectors (15.39) so as to lower bound
 15 the minimax risk. In fact, so as to make the calculations clean, we construct one
 16 testing problem for each choice of orthonormal matrix \mathbf{U} , and then take averages over
 17 a randomly chosen matrix.

More precisely, let $\mathbb{P}_{\theta^j(\mathbf{U})}$ denote the distribution of a random vector from the spiked ensemble (15.38) with leading eigenvector $\theta^* := \theta^j(\mathbf{U})$. By construction, it is a zero-mean Gaussian random vector with covariance

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

Now let $Z(\mathbf{U})$ follow the mixture distribution $\frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j(\mathbf{U})}$. Given an i.i.d. sample $Z^n(\mathbf{U})$ of n such vectors, Fano's inequality implies that the testing error is lower

bounded as

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

where we have used the fact that $\log M \geq d/2$. 1

Since this lower bound 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 $\det(\Sigma^j(\mathbf{U})) = 1 + \nu$ for each $j \in [M]$, Lemma 15.4 implies that

$$\begin{aligned} \mathbb{E}_{\mathbf{U}}[I(Z(\mathbf{U}); J)] &\leq \frac{1}{2} \mathbb{E}_{\mathbf{U}} \log \det(\text{cov}(Z(\mathbf{U}))) - \frac{1}{2} \log(1 + \nu) \\ &\leq \frac{1}{2} \log \det \mathbb{E}_{\mathbf{U}}(\text{cov}(Z(\mathbf{U}))) - \frac{1}{2} \log(1 + \nu), \end{aligned} \quad (15.40)$$

where the second step uses the concavity of the log determinant function. Let us now compute the matrix $\mathbf{\Gamma} := \mathbb{E}_{\mathbf{U}}[\text{cov}(Z(\mathbf{U}))]$. On one hand, we have $\Gamma_{11} = 1 + \nu - \nu\delta^2$, and

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

using the fact that $\mathbf{U} \Delta^j$ is uniformly distributed over the unit sphere in dimension $(d - 1)$. Letting $\mathbf{\Gamma}_{\text{low}}$ denote the lower block of side length $(d - 1)$, we have

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

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


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

Combining with inequality (15.40) and applying the elementary inequality $\log(1+t) \leq t$, we find that

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

Combined with the averaged version of Fano's inequality, we conclude that the minimax risk for estimating the spiked eigenvector in squared Euclidean norm is lower bounded as

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

1 In Corollary 8.1, we proved that the maximum eigenvector of the sample covariance
 2 achieves this squared Euclidean error (up to constant pre-factors), so that we have
 3 obtained a sharp characterization of the minimax rate. 

4 As a follow-up to the previous example, let us now investigate lower bounds for
 5 variable selection in sparse PCA, again working under the spiked model (15.38).

6 **Example 15.11** (Lower bounds for variable selection in sparse PCA). Suppose that our
 7 goal is to determine the scaling of the sample size required to ensure that the support set
 8 of an s -sparse eigenvector θ^* can be recovered. Of course, the difficulty of the problem
 9 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
 10 any method requires $n \gtrsim \frac{\nu^2}{1+\nu} s \log(d - s + 1)$ samples to correctly recover the support.
 11 In Exercise 15.11, we prove a more general lower bound for arbitrary scalings of θ_{\min} .

Our analysis is based on an ensemble similar to Ensemble B previously used in Example 15.9 for analyzing variable selection in sparse linear regression. 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

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

12 As with the previous example (in which we averaged over an orthonormal matrix \mathbf{U}),
 13 here we will average over the choice of sign vectors ε . Let $\mathbb{P}_{\theta^j(\varepsilon)}$ denote the distribution
 14 of the spiked vector (15.38) with $\theta^* = \theta^j(\varepsilon)$, and let $Z(\varepsilon)$ be a sample from the mixture
 15 distribution $\frac{1}{M} \sum_{j \in S^c} \mathbb{P}_{\theta^j(\varepsilon)}$.

Following a similar line of calculation as Example 15.10, we have

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


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 side length $s - 1$, a straightforward calculation yields that $\mathbf{\Gamma}$ is a block diagonal

matrix with $\mathbf{\Gamma}_{SS} = \frac{\nu}{s} \mathbf{E}_{s-1}$ and $\mathbf{\Gamma}_{S^c S^c} = \frac{\nu}{s(d-s+1)} \mathbf{I}_{d-s+1}$. Consequently, we have

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

Recalling that we have n i.i.d. 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, which establishes the claim. 

■ 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 non-parametric problems, since it obviates the need for constructing a local packing. The reader should recall the multi-way hypothesis testing problem, and the mutual information between the observations Z and the label J .

Lemma 15.5 (Yang-Barron method). Let $N_{\text{KL}}(\epsilon; \mathcal{P})$ denote the ϵ -covering number of \mathcal{P} in the square-root KL divergence. Then the mutual information is upper bounded as

$$I(Z; J) \leq \inf_{\epsilon > 0} \{ \epsilon^2 + \log N_{\text{KL}}(\epsilon; \mathcal{P}) \}. \quad (15.41)$$

Proof. Recalling the form (15.22) of the mutual information, we observe that for any distribution \mathbb{Q} ,

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

where the inequality follows from the fact (see Exercise 15.7) that the mixture distribution $\bar{\mathbb{P}} = \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{\theta^j}$ minimizes the average Kullback-Leibler divergence over the family $\{\mathbb{P}_{\theta^1}, \dots, \mathbb{P}_{\theta^M}\}$. Consequently, we have $I(Z; J) \leq \max_{j=1, \dots, M} D(\mathbb{P}_{\theta^j} \parallel \mathbb{Q})$, where the distribution \mathbb{Q} is free to be chosen.

In particular, let $\{\gamma^1, \dots, \gamma^N\}$ be a ϵ -covering of Ω in the square-root KL pseudo-distance, and set $\mathbb{Q} = \frac{1}{N} \sum_{k=1}^N \mathbb{P}_{\gamma^k}$. By construction, for each $\theta^j, j = 1, \dots, M$, we can find some γ^k such that $D(\mathbb{P}_{\theta^j} \parallel \mathbb{P}_{\gamma^k}) \leq \epsilon^2$. Therefore, we have

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

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

In conjunction with Proposition 15.2, Lemma 15.5 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

$$\log M(\delta; \rho, \Omega) \geq 2\{\epsilon^2 + \log N_{\text{KL}}(\epsilon; \mathcal{P}) + \log 2\}. \quad (15.42)$$

A typical approach is based on a two step procedure:

(A) First, choose $\epsilon_n > 0$ such that $\epsilon_n^2 \geq \log N_{\text{KL}}(\epsilon_n; \mathcal{P})$. Since the KL divergence typically scales with n , it is typical that ϵ_n^2 also grows with n .

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

$$\log M(\delta_n; \rho, \Omega) \geq 4\epsilon_n^2 + 2\log 2. \quad (15.43)$$

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

Example 15.12 (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 15.7. Our analysis involved the class \mathcal{F}_2 (as defined in equation (15.15)) of densities on $[0, 1]$, bounded uniformly above, and 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. 1
2

For uniformly bounded densities on $[0,1]$, the squared Hellinger metric is sandwiched above and below by constant multiples of the squared Euclidean distance

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

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 equation (15.31) 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. 3
4
5
6
7
8

From classical theory, it is known that the metric entropy of the class \mathcal{F}_2 in Euclidean distance scales as $\log N(\delta; \mathcal{F}_2, \|\cdot\|_2) \asymp (1/\delta)^{1/2}$. 9
10

Step A: Given n i.i.d. samples, the square-root Kullback-Leibler divergence is multiplied by a factor of \sqrt{n} , so that Step A amounts to choosing $\epsilon_n > 0$ such that

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

This condition can be satisfied by setting $\epsilon_n^2 \asymp n^{1/5}$. 11

Step B: With this choice of ϵ_n , the condition in Step B can be satisfied by choosing $\delta_n > 0$ such that

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

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. 12
13

♣ 14

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

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

where $w_i \sim \mathcal{N}(0,1)$. We assume that the design points $\{x_i\}_{i=1}^n$ are drawn in an

i.i.d. fashion from some distribution \mathbb{P} , and we provide lower bounds in the $L^2(\mathbb{P})$ -norm

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

Example 15.13 (Minimax risks for generalized Sobolev families). For a smoothness parameter $\alpha > 1/2$, consider the ellipsoid $\ell^2(\mathbb{N})$ given by

$$\mathcal{E}_\alpha = \left\{ (\theta_j)_{j=1}^\infty \mid \sum_{j=1}^\infty j^{2\alpha} \theta_j^2 \leq 1 \right\}. \quad (15.44)$$

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

$$\mathcal{F}_\alpha := \left\{ f = \sum_{j=1}^\infty \theta_j \phi_j \mid (\theta_j)_{j=1}^\infty \in \mathcal{E}_\alpha \right\}. \quad (15.45)$$

As discussed in Chapter 12, many of these function classes correspond to particular types of reproducing kernel Hilbert spaces, where α corresponds to the degree of smoothness. For any such function class, we claim that the minimax risk in squared $L^2(\mathbb{P})$ norm is lower bounded as

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_\alpha} \mathbb{E}[\|\hat{f} - f\|_2^2] \gtrsim \min\left\{1, \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}\right\}, \quad (15.46)$$

- 1 and we do so via the Yang-Barron technique.
- 2 For any function of the form $f = \sum_{j=1}^\infty \theta_j \phi_j$ where $\theta \in \ell^2(\mathbb{N})$, Parseval's theo-
- 3 rem ensures that $\|f\|_2^2 = \sum_{j=1}^\infty \theta_j^2$. Consequently, based on our calculations from Exam-
- 4 ple 5.8, we have $\log N(\delta; \mathcal{F}_\alpha, \|\cdot\|_2) \asymp (1/\delta)^{1/\alpha}$. Accordingly, we can find a δ -packing
- 5 $\{f^1, \dots, f^M\}$ of \mathcal{F}_α 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 distance. For each $j \in [M]$, let \mathbb{Q}^j denote the distribution of y given $\{x_i\}_{i=1}^n$ when the true regression function is f^j , and let \mathbb{P}^n denote the product distribution over the covariates. When the true regression function is f^j , the joint distribution over $(y, \{x_i\}_{i=1}^n)$ is given by $\mathbb{Q}^j \times \mathbb{P}^n$, and hence for any distinct pair of indices $j \neq k$, we have

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

Consequently, we find that

$$\log N_{\text{KL}}(\epsilon) = \log N\left(\frac{\sigma\sqrt{2}}{\sqrt{n}}\epsilon; \mathcal{F}_\alpha, \|\cdot\|_2\right) \stackrel{(i)}{\lesssim} \left(\frac{\sqrt{n}}{\sigma\epsilon}\right)^{1/\alpha},$$

where the final inequality again uses the result of Example 5.8. Consequently, the condition in Step A can be satisfied by setting $\epsilon_n^2 \asymp \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}}$.

Step B: It remains to choose $\delta > 0$ so that $\log M(\delta) \geq 4\epsilon_n^2 + 2\log 2$. Given our choice of ϵ_n and the scaling of the packing entropy, we require

$$(1/\delta)^{1/\alpha} \geq c \left\{ \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}} + 2\log 2 \right\}. \quad (15.47)$$

As long as n/σ^2 is larger than some universal constant, the choice $\delta_n^2 \asymp \left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ satisfies the condition (15.47). Putting together the pieces yields the claim (15.46).



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

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

Definition 15.1. 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

$$H(\mathbb{Q}) := -\mathbb{E}[\log q(X)] = -\int_{\mathcal{X}} q(x) \log q(x) \mu(dx), \quad (15.48)$$

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 (15.48) yields the discrete entropy

$$H(\mathbb{Q}) = -\sum_{x \in \mathcal{X}} q(x) \log q(x). \quad (15.49)$$

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 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 $H(X)$ is non-random; however, as it is standard practice in information theory and convenient, we make use of it in this appendix.

Definition 15.2. Given a pair of random variables (X, Y) with joint distribution $\mathbb{Q}_{X,Y}$ the conditional entropy of $X | Y$ is given by

$$H(X | Y) := \mathbb{E}_Y[H(\mathbb{Q}_{X|Y})] = \mathbb{E}_Y\left[\int_{\mathcal{X}} q(x | Y) \log q(x | Y) \mu(dx)\right]. \quad (15.50)$$

As an exercise, it is worthwhile verifying the following facts. First, conditioning can only reduce entropy:

$$H(X | Y) \leq H(X). \quad (15.51a)$$

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

$$H(X, Y) = H(Y) + H(X | Y). \quad (15.51b)$$

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

$$H(X, Y | Z) = H(X | Z) + H(X | Y, Z). \quad (15.51c)$$

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

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \quad (15.51d)$$

By replacing the joint entropy with its chain rule decomposition (15.51b), we obtain

$$I(X; Y) = H(Y) - H(Y | X). \quad (15.51e)$$

With these results in hand, we are now ready to prove the Fano bound (15.23), which we do 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

$$h(q_e) + q_e \log(M - 1) \geq H(J | Z). \quad (15.52)$$

To see how this lower bound implies the stated claim (15.23), we note that

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

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 (15.23). 1

It remains to prove the lower bound (15.52). 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 $H(V, J | Z)$ in two different ways. On one hand, by chain rule, we have

$$H(V, J | Z) = H(J | Z) + H(V | J, Z) = H(J | Z), \quad (15.53)$$

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

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

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

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

If $V = 0$, then $J = \psi(Z)$, so that $H(J | Z, V = 0) = 0$. On the other hand, if $V = 1$, then we know that $J \neq \psi(Z)$, so that the conditioned random variable $(J | Z, V = 1)$ can take at most $M - 1$ values, which implies that

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

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

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

1 and combined with the earlier equality (15.53), the claim (15.52) follows.

2 ■ 15.4 Bibliographic details and background

3 Information theory was introduced in the seminal work of Shannon [SW49, Sha48,
4 Sha49]. Kullback and Leibler [KL51] introduced the Kullback-Leibler divergence, and
5 established various connections both to large deviations theory and testing problems.
6 Early work by Lindley [Lin56] also established connections between information and sta-
7 tistical estimation. Kolmogorov was the first to connect information theory and metric
8 entropy; in particular, see Appendix II of the paper by Kolmogorov and Tikhomirov [KT59].
9 The book by Cover and Thomas is a standard introductory level text on information
10 theory [CT91]. The proof of Fano's inequality given here follows their book.

11 The parametric problems discussed in Examples 15.1 and 15.2 were considered in Le
12 Cam [LC73], where he described the lower bounding approach now known as Le Cam's
13 method. In this same paper, Le Cam also shows how a variety of non-parametric
14 problems can also be treated by this method, using results on metric entropy. The
15 paper by Hasminskii [Has78] used the weakened form of the Fano bound to derive lower
16 bounds on density estimation in the uniform metric; see also the book by Hasminskii
17 and Ibragimov [HI81], as well as their survey paper [HI90]. Assouad [Ass83] developed a
18 method for deriving lower bounds based on placing functions at the vertices of the binary
19 hypercube. See also Birgé [Bir83, Bir87, Bir05] for further refinements on methods for
20 deriving both lower and upper bounds.

21 The chapter by Yu [Yu96] provides a comparison of both Le Cam's and Fano's
22 method, as well as a closely related method due to Assouad [Ass83], not discussed
23 here. Examples 15.3, 15.4 and 15.7 follow parts of her development. Birgé and Mas-
24 sart [BM95] prove the upper bound (15.19) on the squared Hellinger distance; see
25 Theorem 1 in their paper for further details. In their paper, they study the more gen-
26 eral problem of estimating functionals of the density and its first k derivatives under
27 general smoothness conditions of order α . The quadratic functional problem considered
28 in Examples 15.3 and 15.4 correspond to the special case with $k = 1$ and $\alpha = 2$.

29 The refined upper bound on mutual information from Lemma 15.5 is due to Yang
30 and Barron [YB99]. Their work showed how Fano's method can be applied directly with
31 global metric entropies, as opposed to constructing specific local packings of the function
32 class, as in the local packing version of Fano's method discussed in Section 15.3.3.

33 Guntuboyina [Gun11] proves a generalization of Fano's inequality to an arbitrary
34 f -divergence. See Exercise 15.8 for further background on f -divergences and their
35 properties. His result reduces to the classical Fano's inequality when the underlying
36 f -divergence is the Kullback-Leibler divergence. He illustrates how such generalized
37 Fano bounds can be used to derive minimax bounds for various classes of problems,
38 including covariance estimation.

Lower bounds on variable selection in sparse linear regression using the Fano method, as considered in Example 15.9, were first derived by Wainwright [Wai09a]. See also the papers [FRG09, WWR10, AT10, RG08] for further results of this type. The lower bound on variable selection in sparse PCA from Example 15.11 was derived in Amini and Wainwright [AW09]; 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 [Wai14]. In certain cases, there can be substantial gaps between the classical minimax risk and their computationally-constrained analogues [BR13, MW13, ZWJ14].

■ 15.5 Exercises

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

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

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

Exercise 15.2 (Properties of discrete entropy). Let \mathbb{Q} be a probability distribution over a discrete set \mathcal{X} with finite cardinality. Letting q denote the associated probability mass function, its entropy has the explicit formula $H(\mathbb{Q}) = -\sum_{x \in \mathcal{X}} q(x) \log q(x)$, where we interpret $0 \log 0 = 0$.

(a) Show that $H(\mathbb{Q}) \geq 0$.

(b) Show that $H(\mathbb{Q}) \leq \log |\mathcal{X}|$, with equality achieved when \mathbb{Q} is 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.

(a) Show that $D(\mathbb{Q} \parallel \mathbb{P}) \geq 0$ with equality if and only if the equality $q(x) = p(x)$ holds \mathbb{Q} -almost everywhere.

(b) Given a collection of non-negative weights such that $\sum_{j=1}^m \lambda_j = 1$, show that

$$D\left(\sum_{j=1}^m \lambda_j \mathbb{P}_j \parallel \mathbb{Q}\right) \leq \sum_{j=1}^m \lambda_j D(\mathbb{P}_j \parallel \mathbb{Q}), \quad \text{and} \quad (15.54a)$$

$$D(\mathbb{Q} \parallel \sum_{j=1}^m \lambda_j \mathbb{P}_j) \leq \sum_{j=1}^m \lambda_j D(\mathbb{Q} \parallel \mathbb{P}_j). \quad (15.54b)$$

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

3 **Exercise 15.4** (Le Cam's inequality). Prove the upper bound (15.10) of the total vari-
4 ation norm in terms of the Hellinger distance. (*Hint*: The Cauchy-Schwarz inequality
5 could be useful.)

6 **Exercise 15.5** (Decoupling for Hellinger distance). Show that the Hellinger distance
7 satisfies the decoupling relation (15.12) for product measures.

8 **Exercise 15.6** (Achievable rates for uniform shift family). In the context of the uniform
9 shift family (Example 15.2), show that the estimator $\tilde{\theta} = \min\{Y_1, \dots, Y_n\}$ satisfies the
10 bound $\sup_{\theta \in \mathbb{R}} \mathbb{E}[(\tilde{\theta} - \theta)^2] \leq \frac{2}{n^2}$.

Exercise 15.7 (Mixture distributions and KL divergence). Given a collection of dis-
tributions $\{\mathbb{P}^1, \dots, \mathbb{P}^M\}$, consider the mixture distribution $\bar{\mathbb{P}} = \frac{1}{M} \sum_{j=1}^M \mathbb{P}^j$. Show that

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

11 for any other distribution \mathbb{Q} .

Exercise 15.8 (f -divergences). Let $f : \mathbb{R}_+ \rightarrow \mathbb{R}$ 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

$$D_f(\mathbb{P} \parallel \mathbb{Q}) := \int q(x) f(p(x)/q(x)) \mu(dx) \quad (15.55)$$

12 (a) Show that the Kullback-Leibler divergence corresponds to the f -divergence de-
13 fined by $f(t) = t \log t$.

(b) Specify a choice of f that generates the $L^2(\mu)$ distance

$$\|p - q\|_\mu^2 = \int (p(x) - q(x))^2 \mu(dx).$$

- (c) Compute the f -divergence generated by $f(t) = -\log(t)$. 1
- (d) Show that the f -divergence associated with $f(t) = (\sqrt{t} - 1)^2$ is the squared Hellinger divergence $H^2(\mathbb{P} \parallel \mathbb{Q})$. 2
3
- (e) Compute the f -divergence generated by the function $f(t) = 1 - \sqrt{t}$. 4

Exercise 15.9 (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 $\Sigma_j \succ 0$. 5
6

- (a) If $\Sigma_1 = \Sigma_2 = \Sigma$, show that

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

- (b) In the general setting, show that

$$D(\mathbb{Q}_1 \parallel \mathbb{Q}_2) = \frac{1}{2} \left\{ \langle \mu_1 - \mu_2, \Sigma_2^{-1} (\mu_1 - \mu_2) \rangle + \log \frac{\det(\Sigma_2)}{\det(\Sigma_1)} + \text{trace}(\Sigma_2^{-1} \Sigma_1) - d \right\}.$$

Exercise 15.10 (Gaussian distributions and maximum entropy). For a given $\sigma > 0$, let \mathcal{Q}_σ 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 $\mathcal{N}(0, \sigma^2)$. 7
8
9
10

Exercise 15.11 (Sharper bound for variable selection in sparse PCA). In the context of Example 15.11, show that support recovery in sparse PCA is not possible whenever

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

for a sufficiently small but universal constant c_0 . (*Note:* This result is a strict generalization of the bound from Example 15.11, since we must have $\theta_{\min}^2 \leq \frac{1}{s}$ due to the unit norm and s -sparsity of the eigenvector.) 11
12
13

Exercise 15.12 (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 (15.38). Assuming that θ^* is s -sparse, show that any estimator $\hat{\theta}$ satisfies the lower bound

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

- 1 for some universal constant $c_0 > 0$. (*Hints:* The packing set from Example 15.8 may be
 2 useful to you. Moreover, you might consider a construction similar to Example 15.10,
 3 but with the random orthonormal matrix \mathbf{U} replaced by a random permutation matrix
 4 along with random sign flips.)

Exercise 15.13 (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 fixed design vectors $\{x_i\}_{i=1}^n$ and responses $\{y_i\}_{i=1}^n$ drawn from the distribution

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

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

- 7 (a) Compute an expression for the Kullback-Leibler divergence between \mathbb{P}_θ and $\mathbb{P}_{\theta'}$
 8 involving Φ and its derivatives.
- 9 (b) Assuming that $\|\Phi''\|_\infty \leq L < \infty$, give an upper bound on the Kullback-Leibler
 10 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

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

- 11 where $\eta_{\max} = \sigma_{\max}(\mathbf{X}/\sqrt{n})$ is the maximum singular value. (Here as usual
 12 $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix with x_i as its i^{th} row.)

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

Exercise 15.14 (Lower bounds for additive non-parametric regression). Recall the class of additive functions first introduced in Exercise 13.8, namely

$$\mathcal{F}_{\text{additive}} = \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f = \sum_{j=1}^d g_j \text{ for } g_j \in \mathcal{G} \right\},$$

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

(a) Show that

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

By comparison with the result of Exercise 14.7, we see that the least-squares estimator is minimax-optimal up to constant factors. 1
2

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

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

Show that

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