Minimax lower bounds

Han Zhang & Ergan Shang

December 18, 2022

Table of Contents

In this chapter,our goal is to derive lower bounds on the estimation error achievable by any procedure, regardless of its computational complexity and/or storage.
Given a class of distributions $\mathcal{P}$, we let θ denote a functional on the space $\mathcal{P}$-that is, a mapping from a distribution P to a parameter $\theta(P)$taking values in some space Ω. Our goal is to estimate $\theta(P)$ based on samples drawn from the unknown distribution P.

Table of Contents

Minimax risks

Suppose that we are given a random variable X drawn according to a distribution $\mathbb{P}$ for which $\theta(\mathbb{P}) = \theta^*$. Our goal is to estimate the unknown quantity θ^* on the basis of the data X. An estimator $\widehat{\theta}$ for doing so can be viewed as a measurable function from the domain $\mathcal{X}$ of the random variable X to the parameter space Ω. In order to assess the quality of any estimator, we let $\rho : \Omega \times \Omega \to [0, \infty)$ be a semi-metric, and we consider the quantity $\rho(\widehat{\theta}, \theta^*)$.
The quantity known as the minimax risk is defined as follows:

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

When the estimator is based on n i.i.d. samples from P, we use $\mathfrak{M}_n$to denote the associated minimax risk.

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

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

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

From estimation to testing

Here we will show how lower bounds of the minimax risk can be obtained via"reduction" to the problem of obtaining lower boundes for the probability of error in a certain testing problem.
Suppose that $\{\theta^1, \ldots, \theta^M\}$ is a $2\delta-$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 $\mathbb{P}_{\theta^j}$that is, a distribution such that$\theta(\mathbb{P}_{\theta^j}) = \theta^j$ and then consider the M-ary hypothesis testing problem defined by the family of distributions $\{\mathbb{P}_{\theta^j}, j = 1, \ldots, M\}$. In particular, we generate a random variable Z by the following procedure:

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

We let Q denote the joint distribution of the pair (Z, J)generated by this procedure. Note that the marginal distribution over Z is given by the uniformly weighted mixture distribution$\bar{Q} := \frac{1}{M} \sum_{j=1}^{M} P_{\theta j}$. 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 functionfor this problem is a mapping $\Psi : \mathcal{Z} \to [M]$, and the associated probability of error is given by $Q[\Psi(Z) \neq J]$, where the probability is taken jointly over the pair (Z, J).

Proposition (15.1)

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

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

where the infimum ranges over test functions.

Remark.
If we choose a value δ^*sufficiently small to ensure that this testing error is at least $1/2$, then we may conclude that $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta^*)$
For a given choice of δ, the other additional degree of freedom is our choice of packing set.

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

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

Note that

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

so we have reduced the problem to lower bounding the quantity $\mathbb{Q}\left[\rho\left(\widehat{\theta}, \theta^J\right) \geq \delta\right]$.
Now observe that any estimator $\widehat{\theta}$ can be used to define a test—namely, via

$$\Psi(Z) := \text{argmin}_{l \in [M]} \rho(\theta^l, \widehat{\theta}). \tag{15.4}$$

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

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

Combined with our earlier argument, we have shown that

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

Finally,we take the infimum twice(both sides).

Some divergence measures

The *total variation(TV) distance*:

$$\|P - Q\|_{TV} := \sup_{A \subseteq \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \nu(dx). \tag{15.5/15.6}$$

where ν is some underlying base measure defining the densities.
The *Kullback-Leibler divergence*:

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

Lemma (15.2)

(Pinsker-Csiszár-Kullback inequality) For all distributions P and Q,

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

The *squared Hellinger distance*:

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

Lemma (15.3)

(Le Cam's inequality) For all distributions P and Q,

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

Let$(P_1, \ldots, P_n)$ be a collection of n probability measures, each defined on X, and let $P^{1:n} = \bigotimes_{i=1}^{n} P_i$ be the product measure defined on $\mathcal{X}^n$. The $Q^{1:n}$ is similar.
Try to express the divergence between$P^{1:n}$and $Q^{1:n}$ in terms of divergences between the individual pairs.

- It is difficult for the TV distance in general.
-

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

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

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

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

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

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

Table of Contents

Bayes error and total variation distance

The simplest type of testing problem, known as a binary hypothesis test, involves only two distributions.
Here $\bar{Q} := \frac{1}{2}P_0 + \frac{1}{2}P_1$.For a given decision rule $\Psi : \mathcal{Z} \to \{0, 1\}$,the associated probability of error is given by

$$Q[\Psi(Z) \neq J] = \frac{1}{2}P_0[\Psi(Z) \neq 0] + \frac{1}{2}P_1[\Psi(Z) \neq 1].$$

We have

$$\inf_\Psi Q[\Psi(Z) \neq J] = \frac{1}{2}\left\{1 - \|P_1 - P_0\|_{TV}\right\}. \tag{15.13}$$

The quantity is known as the *Bayes risk*.
Then as a result of Proposition 15.1,for any pair of distributions $P_0, P_1 \in \mathcal{P}$ such that $\rho(\theta(P_0), \theta(P_1)) \geq 2\delta$, we have

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2}\left\{1 - \|P_1 - P_0\|_{TV}\right\}. \tag{15.14}$$

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

Let us apply the two-point Le Cam bound (15.14) with the distributions P_0^n and P_θ^n. We set $\theta = 2\delta$, for some δ to be chosen later in the proof, which ensures that the two means are 2δ-separated. In order to apply the two-point Le Cam bound, we need to bound the total variation distance $\|P_\theta^n - P_0^n\|_{TV}$.

From the second-moment bound in Exercise 15.10(b), we have

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

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

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

and

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

Although the pre-factors $1/12$ and $1/24$ are not optimal, the scalings $\sigma/\sqrt{n}$ and σ^2/n are sharp.

Example (15.5)(Uniform location family) Let us consider the uniform location family, in which, for each $\theta \in \mathbb{R}$, the distribution $\mathbb{U}_\theta$ is uniform over the interval $[\theta, \theta + 1]$. We let $\mathbb{U}_\theta^n$ denote the product distribution of n i.i.d. samples from $\mathbb{U}_\theta$. In this case, it is not possible to use Lemma 15.2 to control the total variation norm, since the Kullback-Leibler divergence between $\mathbb{U}_\theta$ and $\mathbb{U}_{\theta'}$ is infinite whenever $\theta \neq \theta'$. Accordingly, we need to use an alternative distance measure: in this example, we illustrate the use of the Hellinger distance (see equation (15.9)).
Given a pair $\theta, \theta' \in \mathbb{R}$, let us compute the Hellinger distance between $\mathbb{U}_\theta$ and $\mathbb{U}_{\theta'}$. By symmetry, it suffices to consider the case $\theta' > \theta$. If $\theta' > \theta + 1$, then we have $H^2\left(\mathbb{U}_\theta \| \mathbb{U}_{\theta'}\right) = 2$. Otherwise, when $\theta' \in (\theta, \theta + 1]$, we have

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

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

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

In conjunction with Lemma 15.3, we find that

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

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

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

Le Cam's method is also useful for various nonparametric problems.
An important quantity in the Le Cam approach to such problems is the Lipschitz constant of the functional θ with respect to the Hellinger norm, given by

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

Corollary (15.6)

(Le Cam for functionals) For any increasing function Φon the non-negative real line and any functional$\theta : \mathscr{F} \to \mathbb{R}$, we have

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

Proof.
Setting $\epsilon^2 = \frac{1}{4n}$, choose a pair f, g that achieve the supremum defining $\omega(1/(2\sqrt{n}))$. By a combination of Le Cam's inequality (Lemma 15.3) and the decoupling property (15.12b) for the Hellinger distance, we have

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

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

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

as claimed.

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

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

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

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

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

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

Observe that

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

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

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

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

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

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

Le Cam's convex hull method

Le Cam's method is an elegant generalization of this idea, one which allows us to take the convex hulls of two classes of distributions. Consider two subsets $\mathcal{P}_0$ and $\mathcal{P}_1$of $\mathcal{P}$ that are 2δ-separated, in the sense that

$$\rho(\theta(P_0), \theta(P_1)) \geq 2\delta \text{ for all } P_0 \in \mathcal{P}_0 \text{ and } P_1 \in \mathcal{P}_1. \tag{15.25}$$

Lemma (15.9)

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

$$sup_{p\in\mathcal{P}} E_P\left[\rho(\widehat{\theta}, \theta(P))\right] \geq \frac{\delta}{2} sup_{P_0\in conv(\mathcal{P}_0), P_1\in conv(\mathcal{P}_1)} \left\{1 - \|P_0 - P_1\|_{TV}\right\}. \tag{15.26}$$

Proof.

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

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

We then have

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

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

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

By the triangle inequality, we have

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

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

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

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

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

which completes the proof.

Example(15.10)(Sharpened bounds for Gaussian location family) In Example 15.4, we used a two-point form of Le Cam's method to prove a lower bound on mean estimation in the Gaussian location family. A key step was to upper bound the TV distance $\|P_\theta^n - P_0^n\|_{TV}$between the n-fold product distributions based on the Gaussian models $N(\theta, \sigma^2)$ and $N(0, \sigma^2)$ respectively.

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

$$\left\|\overline{\mathbb{P}} - \mathbb{P}_0^n\right\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{\sqrt{n}\theta}{\sigma}\right)^4} - 1\right\} = \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{2\sqrt{n}\hat{\sigma}}{\sigma}\right)^4} - 1\right\}.$$

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

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

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

Table of Contents

Kullback-Leibler divergence and mutual information

The difficulty of the testing problem before depends on the amount of dependence between the observation Z and the unknown random index J. Try to measure the divergence between the joint distribution $Q_{Z,J}$and the product of marginals$Q_Z Q_J$
The mutual information between the random variables $(Z, J)$is defined as follows(using the Kullback-Leibler divergence)

$$I(Z, J) := D(Q_{Z,J} \| Q_Z Q_J). \tag{15.29}$$

Given our set-up and the definition of the KL divergence, with $\bar{Q} = Q_Z := \frac{1}{M} \sum_{j=1}^{M} P_{\theta^j}$ we have

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

Fano lower bound on minimax risk

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:

$$P[\Psi(Z) \neq J] \geq 1 - \frac{I(Z;J) + log2}{logM}. \tag{15.31}$$

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

Proposition (15.12)

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

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z; J) + log2}{logM} \right\}, \tag{15.32}$$

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

Remark.

By decreasing δ suffficiently,we may thereby ensure that

$$\frac{I(Z; J) + log2}{logM} \leq \frac{1}{2}, \tag{15.33}$$

So that the lower bound (15.32)implies that $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$.

Two technical and possibly challenging steps.

- The first requirement is to specify δ-separated sets with large cardinality $M(2\delta)$. Here the theory of metric entropy developed in Chapter 5 plays an important role, since any 2δ-packing set is (by definition) $2\delta-$separated in the ρsemi-metric.
- The second requirement is to upper bound the mutual information$I(Z; J)$.

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

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

Example(15.13)(Normal location model via Fano method) Recall from Example 15.4 the normal location family, and the problem of estimating $\theta \in \mathbb{R}$ under the squared error. There we showed how to lower bound the minimax error using Le Cam's method; here let us derive a similar lower bound using Fano's method.
Consider the 2δ-separated set of real-valued parameters$\{\theta^1, \theta^2, \theta^3\} = \{0, 2\delta, -2\delta\}$.Since $\mathbb{P}_{\theta^j} = \mathcal{N}\left(\theta^j, \sigma^2\right)$, we have

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

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

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

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

Bounds based on local packings

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \frac{\left\|\mathbf{X}\theta^j\right\|_2}{\sqrt{n}} &\leq 4\delta, \qquad \text{for each } j \in [M], \\ 2\delta \leq \frac{\left\|\mathbf{X}\left(\theta^j - \theta^k\right)\right\|_2}{\sqrt{n}} &\leq 8\delta \qquad \text{for each } j \neq k \in [M] \times [M]. \end{aligned} \tag{15.36}$$

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

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

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

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

Local packings with Gaussian entropy bounds

Let us now formalize the approach that was used in the previous example. We now turn to a different upper bound on the mutual information, applicable when the conditional distribution of Z given J is Gaussian.

Lemma (15.17)

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

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

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

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

Yang-Barron version of Fano's method

In this section, we develop an alternative upper bound on the mutual information. It is particularly useful for nonparametric problems, since it obviates the need for constructing a local packing.

Lemma (15.21)

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

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

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

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

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

Since the upper bound (15.52) holds for any distribution $\mathbb{Q}$, we are free to choose it: in particular, we let $\{\gamma^1, \ldots, \gamma^N\}$ be an ϵ-covering of Ω in the square-root KL pseudo-distance,and then set $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{P}_{\gamma^k}$. By construction, for each θ^j with $j \in [M]$, we can find some γ^k such that $D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\gamma^k}\right) \leq \epsilon^2$. Therefore, we have

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

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

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

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

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

- First, choose $\epsilon_n > 0$ such that

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

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

- Second choose the largest $\delta_n > 0$that satisfies the lower bound

$$log M(\delta_n; \rho, \Omega) \geq 4\epsilon_n^2 + 2log2. \tag{15.53b}$$

Table of Contents

Appendix: Basic background in information theory

Definition (15.14)

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

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

when this integral is finite.

Similiarly we get the discrete entropy

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

When $\mathcal{X}$ is a finite set, it satisfies the upper bound $H(Q) \leq log|\mathcal{X}|$, with equality achieved when Q is uniform over $\mathcal{X}$.

Definition (15.25)

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

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

Some elementary properties of entropy and mutual information:

- $$H(X|Y) \leq H(X). \tag{15.60a}$$
- $$H(X, Y) = H(Y) + H(X|Y). \tag{15.60b}$$

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

The proof of Fano bound(15.31)

We here introduce the shorthand notation $q_e = \mathbb{P}[\psi(Z) \neq J]$,
$h(q_e) = -q_e \log q_e - (1 - q_e) \log (1 - q_e)$ denoting the binary entropy.
With this notation,we give following inequality:

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

We claim with this inequality right we can get Fano bound.

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

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

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

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

where the inequality follows since conditioning can only reduce entropy.

By the definition of conditional entropy, we have

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

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

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

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

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

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