Institute of Mathematics, EPFL MATH524

Chapter 6: Minimax theory

Lecturer: Rajita Chandak *Spring 2025*

Now that we have expanded our tool set to deal with uniform consistency of estimators, there are two new interesting questions that arise. Lets stick with the KDE as our working example here. We want to answer the following:

1. Can the rate of convergence (established from the uniform consistency result) be improved upon by any other estimators for the Holder class of probability functions, $F(\beta, L)$?
2. What is the best possible rate of convergence for this class of functions?

To answer these questions, we need to first introduce the concept of **minimaxity**:

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}_f[(\hat{f}(x) - f(x))^2] \qquad \forall x,$$

where $\hat{f}$ is any estimator. This statement means we are attempting to identify the estimator that provides the **best** rate of convergence over *all* possible functions for the class of functions defined by $F(\beta, L)$.

Recall that we have previously seen that for the kernel density estimator,

$$\sup_{x \in \mathbb{R}} \sup_{f \in F(\beta, L)} \mathbb{E}_f[(\hat{f}_n(x) - f(x))^2] \leq C n^{-\frac{2\beta}{2\beta+1}}.$$

This is an upper bound on the maximum risk for a specific (kernel) estimator. To complement the upper bound, we would be interested in a lower bound of the type

$$\forall \hat{f}_n : \qquad \sup_{f \in F(\beta, L)} \mathbb{E}_f[(f(x) - \hat{f}_n(x))^2] \geq C \psi_n^2,$$

where ψ_n is some positive sequence that approaches zero as $n \to \infty$ and C is some positive constant. We have so far restricted our analysis to (I)MSE loss metrics. There may be other symmetric loss or *distance* metrics, $d(\cdot, \cdot)$, that can be used to answer similar questions about optimality. i.e., bounds of the form

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[d(f, \hat{f}_n)],$$

where d can be a pointwise or integrated metric. The upper bounds we have come across so far imply the existence of a constant $C < \infty$ such that

$$\limsup_{n \to \infty} \psi_n^{-2} \inf_{T_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f[d(\hat{f}, f)] \leq C. \tag{1}$$

Matching lower bounds would suggest there exists a $c < \infty$ such that

$$\liminf_{n\to\infty} \psi_n^{-2} \inf_{T_n} \sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\hat{f}, f)] \geq c. \tag{2}$$

If we can establish both the lower and upper bounds, we can define the optimal rate of convergence:

Definition 1 (Optimal rate of convergence)**.** *A positive sequence ψ_n is called the **optimal rate of convergence** of estimators on $(\mathcal{F}, d)$ if (1) and (2) hold. Then, the estimator $\hat{f}$ that satisfies*

$$\sup_{f\in\mathcal{F}} \mathbb{E}_f[d(\hat{f}, f)] \asymp \psi_n^2$$

is called the minimax optimal estimator (or the rate-optimal estimator).

We will use tools from empirical process theory to establish general methods for identifying optimal rates of convergence and in particular, by the end of this chapter we will prove the following lower bound:

$$\inf_{\hat{f}} \sup_{f\in F(\beta,L)} \mathbb{E}_f[(\hat{f}(x) - f(x))^2] \geq Cn^{-\frac{2\beta}{2\beta+1}}.$$

This lower bound will allow us to conclude that KDE is minimax-optimal for $F(\beta, L)$ by showing that the KDE achieves the best rate of convergence (up to constants) over all possible estimators.

We will start with more tractable examples to build intuition and develop the tools necessary to prove nonparametric minimaxity for different function classes.

1 Examples

1.1 Parametric model

Starting in the parametric setting gives a smaller, possibly nicer space to work with for establishing notions of minimaxity. Lets start with the class of Gaussian distributions with unknown means. That is, $\mathcal{F} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$. Our objective function will simply be the mean $\theta(F) = \mu$, the only unknown parameter of the function. Consider the least-squares loss function in defining the minimax risk as:

$$R_n = \inf_{\hat{\mu}_n} \sup_{\mu} \mathbb{E}[(\hat{\mu}_n - \mu)^2].$$

For such parametric models, recall that under some regularity conditions, the MLE risk is bounded by $\text{tr}[I(\theta)^{-1}]/n$ at the true parameter θ, where $I(\theta)$ is the Fisher information matrix (and for typical models this will be of the order d/n). It can also be shown that there is a local minimax lower bound (local in the sense that the sup is taken only over a neighborhood around the true θ) of the same order $\text{tr}[I(\theta)^{-1}]/n$. Thus, the MLE is locally minimax. In fact, this bound can be extended to global minimaxity of the MLE by making uniform bound arguments over all local neighborhoods around all $\theta \in \Theta$. This is due to theory developed by Hájek and Le Cam, but we won't go into the technical details of this theory. We'll focus on non-parametric minimax theory that is applicable to the estimators and function classes we have considered so far.

1.2 Non-parametric models

1.2.1 Estimation with random X

Let Q be a fixed distribution on $[0,1]^d$ (e.g., the uniform distribution), and let $(x_i, y_i) \overset{i.i.d.}{\sim} P$, with

$$y_i = f(x_i) + \varepsilon_i, \quad x_i \sim Q, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \text{and} \quad x_i \perp\!\!\!\perp \varepsilon_i, \tag{3}$$

for some fixed $\sigma^2 > 0$. Let $\theta(P) = f$, the entire regression function. Suppose that $\mathcal{P}$ is the set of regression distributions P of the form (3) for which $f \in \mathcal{F}$, (for example, $\mathcal{F}$ defined on $[0,1]^d$). To study function estimation at a single point (e.g., the origin), take the squared loss, $d(\hat{f}, f) = (\hat{f}(0) - f(0))^2$. The minimax risk is then

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[(\hat{f}(0) - f(0))^2].$$

1.2.2 Estimation with fixed X

Consider the regression model with fixed covariates. That is,

$$y_i = f(x_i) + \varepsilon_i, \quad x_i \text{ fixed}, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \text{and} \quad x_i \perp\!\!\!\perp \varepsilon_i.$$

We can still define the minimax risk as before, where now the expectation is understood to be only with respect to the distribution on y (since x is not random). This requires some notational adjustment because now the y_i are independent but no longer i.i.d. Similarly, we will need to be careful with some of the techniques that will be introduced in the remainder of this chapter, because as written we assume i.i.d. data. In several cases, these adjustments will be straightforward and the minimax risk for the random and fixed covariate models will behave in the same manner. However, there will be some cases in which the two models behave very differently. We'll briefly touch upon this at the end.

1.2.3 Estimation in L_2

Lets go back to (3), but with an L_2 loss function $d(\hat{f}, f) = \|\hat{f} - f\|_{L^2}^2$. This has minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[\int (\hat{f}(x) - f(x))^2 \mathrm{d}Q].$$

1.2.4 Estimation in empirical L_2

Now, consider the L_2 loss function with respect to the empirical distribution of X: $d(\hat{f}, f) = \|\hat{f} - f\|_{L^2(Q_n)}^2 = \|\hat{f} - f\|_n^2 = n^{-1} \sum_i (\hat{f}(x_i) - f(x_i))^2$. This has minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \frac{1}{n} \mathbb{E}[\sum_i (\hat{f}(X_i) - f(X_i))^2].$$

2 Estimation to testing

Recall that, typically, we are interested in the order of the rate of convergence (dependency on n) [1] instead of the full explicit form of the minimax risk R_n (i.e., we usually ignore constants). As it turns out, finding a lower bound on R_n will require a totally different technique than what we have used to derive upper bounds for specific estimators. We will develop a general formulation of the common approach to lower bounds first and then look at some specific examples.

Our first step is to show how lower bounds can be obtained via a "*reduction*" to the problem of obtaining lower bounds for the probability of error in a certain testing problem. We do so by constructing a suitable packing of the parameter space (recall from Chapter 5 our definition of packing numbers).

In this section we focus on the function class of probability distributions, i.e., $\mathcal{F} = \mathcal{P}$. More precisely, suppose that $S = \{P_1, \dots, P_N\} \subseteq \mathcal{P}$ is a 2δ-separated set contained in the space $\theta(\mathcal{F})$, meaning a collection of elements[2] $d(\theta_j, \theta_k) \geq 2\delta$ for all $j \neq k$. The minimax risk can then be bounded from below as:

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\theta(P), \hat{\theta})] \geq \inf_{\hat{\theta}} \max_{P_j \in S} \mathbb{E}_{P_j}[d(\theta_j, \hat{\theta})],$$

where we use the shorthand $\theta_j = \theta(P_j)$. For each θ_j, let us choose some representative distribution P_j-that is, a distribution such that $\theta\left(P_j\right) = \theta_j$ and then consider the N-ary hypothesis testing problem defined by the family of distributions $\left\{P_j, j = 1, \dots, N\right\}$. In particular, we generate a random variable Z by the following procedure:

1. Sample a random integer J from the uniform distribution over the index set $[N] := \{1, \dots, N\}$.
2. Given $J = j$, sample $Z \sim P_{\theta_j}$.

We let Q denote the joint distribution of the pair (Z, J) generated by this procedure. Note that the marginal distribution for Z is given by the uniformly weighted mixture distribution $\overline{Q} := \frac{1}{N} \sum_{j=1}^{N} P_{\theta_j}$. Then, given a sample Z from this mixture distribution, we consider the N-ary hypothesis testing problem of determining the randomly chosen index J. The decision rule for this hypothesis test can be defined as $\psi : \mathcal{Z} \to [N]$, and the associated probability of error is given by $Q(\psi(Z) \neq J)$. In order to control this error probability, we will need to define the following two quantities

$$s = \min_{j \neq k} d(\theta_j, \theta_k)$$
$$\psi^* = \operatorname*{argmin}_j d(\theta_j, \hat{\theta}).$$

[1] We may also be interested in how it depends on auxiliary parameters that define $\mathcal{P}$. For example, in function estimation if $\mathcal{F}$ is a norm ball in some function space, then we may also be interested in how R_n scales with the radius of this ball—and indeed, below, we'll track minimax rates as a function of n and the Lipschitz constant L of the regression function (when $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|$).

[2] Notice we only define the packing number with a weak inequality, as opposed to with the strict inequality $d(\theta_j, \theta_k) > 2\delta$ used in the definition in Chapter 5. This is purely to simplify the calculations later on.

Furthermore, we will need to assume that our chosen distance d satisfies a quasi-triangle inequality,

$$d(\theta, \theta') \leq C\left(d(\theta, \theta'') + d(\theta', \theta'')\right)$$

with some global constant $C > 0$. It should be obvious that any valid metric will satisfy this property with $C = 1$. If $d(x, y) = \|x - y\|_2^2$, then the property is satisfied with $C = 2$.

These quantities can now be used to obtain a lower bound on the minimax risk as shown in the following lemma:

Lemma 1 (Testing lower bound)
Let $S = \{P_1, \ldots, P_N\} \subseteq \mathcal{P}$ be any 2δ-packing set, and $d(\cdot, \cdot)$ be a nonnegative symmetric loss function satisfying the quasi-triangle inequality with some constant $C > 0$. Then,

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\theta(P), \hat{\theta})] \geq \frac{s}{2C} \inf_{\psi} \max_{P_j \in S} P_j(\psi \neq j) \geq \frac{s}{2C} \inf_{\psi} Q(\psi \neq J), \tag{4}$$

where the infimum is over all maps ψ and $s = \min_{j \neq k} d(\theta_j, \theta_k)$.

This result is known as the standard reduction for minimax lower bounds. Finding the tightest lower bound requires a careful selection of the set of distributions in S. If S is too big then s will be small. But if S is too small then $\max_{P_j \in S} P_j(\psi \neq j)$ will be small. Note that the right-hand side of the bound (4) involves two terms, both of which depend on the choice of δ. By construction, $s/2C$ is increasing in δ, so that it is maximized by choosing δ as large as possible. On the other hand, the testing error $Q(\psi(Z) \neq J)$ is defined in terms of a collection of 2δ-separated distributions. As $\delta \to 0$, the underlying testing problem becomes more difficult, and so that, at least in general, we should expect that $Q(\psi(Z) \neq J)$ grows as δ decreases. For a given choice of δ, the other additional degree of freedom is our choice of packing set, and we will see a number of different constructions for this shortly.

Proof of Lemma 1. By Markov's inequality, for each j, and any $t > 0$,

$$\mathbb{E}_{P_j}[d(\theta_j, \hat{\theta})] \geq t P_j(d(\theta_j, \hat{\theta}) \geq t),$$

and thus,

$$R_n \geq t \inf_{\hat{\theta}} \max_{P_j \in S} P_j(d(\theta_j, \hat{\theta}) \geq t).$$

Any value of t will give us a valid lower bound. However, we are interested in finding the value of t that gives us the best lower bound. To find this "best" t, we look at the minimum distance between our contenders θ_j, $j = 1, \ldots, N$.

Suppose that the true parameter is θ_k (i.e., $\psi^* = k \neq j$): we then claim that the event $\left\{d(\theta_k, \hat{\theta}) < \delta\right\}$ ensures that the test ψ^* is correct. In order to see this implication, note that, for any other index $j \in [N]$, an application of the triangle inequality guarantees that

$$s \leq d(\theta_j, \theta_k)$$

$$\leq Cd(\theta_j, \hat{\theta}) + Cd(\theta_k, \hat{\theta})$$
$$\leq 2Cd(\theta_j, \hat{\theta}).$$

Where we used the quasi-triangle inequality for the second inequality, and the fact that $d(\theta_k, \hat{\theta}) \leq d(\theta_j, \hat{\theta})$ in the third line. As a result, we have shown that

$$\psi^* \neq j \Rightarrow d(\theta_j, \hat{\theta}) \geq s/(2C)$$

and so

$$P_j\left(d(\theta_j, \hat{\theta}) \geq \frac{s}{2C}\right) \geq P_j(\psi^* \neq j).$$

Now, plugging in for $t = s/(2C)$, we can derive a the following lower bound:

$$R_n \geq \frac{s}{2C} \inf_{\hat{\theta}} \max_{P_j \in S} P_j(\psi^*(\hat{\theta}) \neq j),$$

Thus, given access to $\hat{\theta}$, it tries to pick out which one of $\{\theta_j\}_{j=1}^N$ it thinks is most likely. We can further lower bound the right-hand side by considering all hypothesis tests based on the data. The full infimum over all tests can only be smaller, from which the claim follows. ■

3 Distance between probability measures

In order to proceed with analyzing the minimax risk quantity for any of the notions from the previous section (or indeed any other minimax risk quantity), we need to capture the distance/divergence between functions from the class $\mathcal{P}$ over which we want to bound the risk. The KL divergence is one of the most common quantities used to describe the variation of functions within a class $\mathcal{P}$.

Definition 2 (KL divergence)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let P and Q be two probability measures on this space. Suppose ν is a σ-finite measure on $(\mathcal{X}, \mathcal{A})$ such that P and Q are both absolutely continuous with respect to ν. i.e., $P \ll \nu$ and $Q \ll \nu$. Let $p = \mathrm{d}P/\mathrm{d}\nu$ and $q = \mathrm{d}Q/\mathrm{d}\nu$ (such a measure ν will always exist because it can always take the trivial form $\nu = P + Q$ by the Lebesgue decomposition theorem). The **Kullback-Leibler (KL) divergence** between the two distributions P, Q is then defined as*

$$\mathrm{KL}(P, Q) = \int \log\left(\frac{dP}{dQ}\right) \mathrm{d}P = \int \log\left(\frac{p(z)}{q(z)}\right) p(z) \mathrm{d}\nu.$$

The following observations of the KL divergence can be made directly from its definition:

Lemma 2 (Properties of KL divergence)
The following properties hold for the KL divergence.

1. $\mathrm{KL}(P, Q) \geq 0$ *and* $\mathrm{KL}(P, Q) = 0$ *iff* $P = Q$.
2. $\mathrm{KL}(\cdot, \cdot)$ *is* ***not*** *a distance. (see:* $\mathrm{KL}(P, Q) \neq \mathrm{KL}(Q, P)$*).*

3. *If* $P = \otimes_{i=1}^n P_i$ *and* $Q = \otimes_{i=1}^n Q_i$, $\mathrm{KL}(P, Q) = \sum_i \mathrm{KL}(P_i, Q_i)$.

The following fact will be useful for us. For two Gaussian densities, $P = N(\theta, \sigma^2)$ and $Q = N(\mu, \sigma^2)$, we have

$$\mathrm{KL}(P, Q) = \frac{(\theta - \mu)^2}{2\sigma^2}.$$

There are many other notions of distances on distributions (TV, L_1, Hellinger, χ^2, etc.) that have relationships to allow moving between the different metrics, including relationships to KL divergence. We will not review these here, but will simply define other distances and use known relationships as they naturally arise. Some of these definitions and properties will be explored in the exercises.

4 Le Cam's method

Le Cam's method is an application of this general idea for $N = 2$. This should also help us build intuition for what to expect when we move to a more general N case. Consider two hypotheses: $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, so that $s = d(\theta_0, \theta_1)$.

Theorem 1 (Le Cam's lower bound)
Let $P_0, P_1 \in \mathcal{P}$, *and let* $d(\cdot, \cdot)$ *be a nonnegative, symmetric loss function satisfying the quasi-triangle inequality with some constant* $C > 0$. *Then,*

$$R_n \geq \frac{d(\theta_0, \theta_1)}{8C} e^{-n\mathrm{KL}(P_0, P_1)}.$$

We also have

$$R_n \geq \frac{d(\theta_0, \theta_1)}{4C}[1 - \mathrm{TV}(P_0^n, P_1^n)],$$

where $\mathrm{TV}(P, Q) = \frac{1}{2}\int |p(z) - q(z)|\mathrm{d}z$ *denotes the total variation distance between distributions* P, Q *with densities* p, q.

The lower bounds in Theorem 1 require the following facts about affinity, TV distance, and KL divergence of distributions P, Q with densities p, q.

- $\int p(z) \wedge q(z)\mathrm{d}z = 1 - \mathrm{TV}(P, Q)$ (*Scheffé's Theorem, see exercise sheet 7*).
- $\int p(z) \wedge q(z)\mathrm{d}z \geq \frac{1}{2}e^{-\mathrm{KL}(P,Q)}$ (*see exercise sheet 8*).
- $\mathrm{KL}(P^n, Q^n) = n\mathrm{KL}(P, Q)$.

Proof of Theorem 1. For simplicity, we will start with $n = 1$ (i.e., only one sample). Then, by Lemma 1,

$$R_n \geq \frac{s}{2C} \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j).$$

By the property that $\max_i X_i \geq \bar{X}_n$,

$$R_n \geq \frac{s}{4C} \inf_{\psi} \left[P_0(\psi \neq 0) + P_1(\psi \neq 1) \right].$$

The reason we use the average for the lower bound is the following result (which is derived from the Neyman-Pearson test):

$$\psi_*(z) = \begin{cases} 0 & \text{if } p_0(z) \geq p_1(z) \\ 1 & \text{if } p_0(z) < p_1(z) \end{cases}.$$

We will use (without proof) the fact that (which follows from the Neyman-Pearson lemma)

$$\inf_{\psi}[P_0(\psi \neq 0) + P_1(\psi \neq 1)] = P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1).$$

Now we compute

$$\begin{aligned} P_0\left(\psi_* \neq 0\right) + P_1\left(\psi_* \neq 1\right) &= \int_{p_1 > p_0} p_0(z)\mathrm{d}z + \int_{p_0 \geq p_1} p_1(z)\mathrm{d}z \\ &= \int_{p_1 > p_0} p_0(z) \wedge p_1(z)\mathrm{d}z + \int_{p_0 \geq p_1} p_0(z) \wedge p_1(z)\mathrm{d}z \\ &= \int p_0(z) \wedge p_1(z)\mathrm{d}z. \end{aligned}$$

Thus,

$$R_n \geq \frac{s}{2C} \frac{P_0\left(\psi_* \neq 0\right) + P_1\left(\psi_* \neq 1\right)}{2} = \frac{s}{4C} \int p_0(z) \wedge p_1(z)\mathrm{d}z.$$

Now, if we have $n > 1$ i.i.d. samples, we replace p_0 and p_1 with $p_0^n(z) = \prod_{i=1}^n p_0(z_i)$ and $p_1^n(z) = \prod_{i=1}^n p_1(z_i)$, and by the same arguments, we have

$$R_n \geq \frac{s}{4C}[P_0(\psi \neq 0) + P_1(\psi \neq 1)] = \frac{s}{4C} \int p_0^n(z) \wedge p_1^n(z)\mathrm{d}z. \tag{5}$$

The integral on the right-hand side above is often called the **affinity** between p_0^n and p_1^n. The proof of both parts of the statement follows essentially by applying the corresponding relationship between the loss metric and the affinity integral to the bound in (5) with $S = \{P_0, P_1\}$ and $s = d(\theta_0, \theta_1)$. i.e.,

$$\begin{aligned} R_n &\geq \frac{d(\theta_0, \theta_1)}{4C}[P_0(\psi \neq 0) + P_1(\psi \neq 1)] \\ &= \frac{d(\theta_0, \theta_1)}{4C} \int p_0^n(z) \wedge p_1^n(z)\mathrm{d}z \\ &\geq \frac{d(\theta_0, \theta_1)}{8C} e^{-\mathrm{KL}(P_0^n, P_1^n)} = \frac{d(\theta_0, \theta_1)}{8C} e^{-n\mathrm{KL}(P_0, P_1)} \end{aligned}$$

and similarly,

$$R_n \geq \frac{d(\theta_0, \theta_1)}{4C}[P_0(\psi \neq 0) + P_1(\psi \neq 1)]$$

$$= \frac{d(\theta_0, \theta_1)}{4C} \int p_0^n(z) \wedge p_1^n(z) \mathrm{d}z$$
$$= \frac{d(\theta_0, \theta_1)}{4C} (1 - TV(P_0^n, P_1^n))$$

■

A useful corollary of Le Cam's KL bound in Theorem 1 is the following.

Corollary 1
Under the same conditions on $d(\cdot, \cdot)$ as in Theorem 1, suppose there exists $P_0, P_1 \in \mathcal{P}$ such that $\mathrm{KL}(P_0, P_1) \leq (\log 2)/n$. Then $R_n \geq d(\theta_0, \theta_1)/(16C)$.

Example 1 (KDE lower bound)**.** *We can demonstrate the applicability of Le Cam's method by considering our KDE example. Let's start by defining our problem in terms of the minimaxity setup we have introduced in this chapter. For simplicity, consider the input distribution to be uniform, $Q = \mathrm{Unif}\left([0,1]^d\right)$, and just take $\sigma^2 = 1$. Consider $\mathcal{F} = C^1(L; [0,1])^d$, the space of functions that are L-Lipschitz continuous on $[0,1]^d$. i.e., there exists a positive constant L such that*

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|.$$

and consider pointwise risk at the $x = 0$, in squared loss,

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}[(\hat{f}(0) - f(0))^2].$$

Recall that in this context, $\theta_0 = f_0(0)$ and $\theta_1 = f_1(0)$, where f_0, f_1 are functions in $\mathcal{F}$. Let's suppose $f_0 = 0$ (the zero function). Let K be any 1-Lipschitz function supported on the unit ℓ_2 ball $\{x : \|x\|_2 \leq 1\}$, such that $K(0) = 1$ and $0 < \int K(x)^2 dx < \infty$. Then let $f_1(x) = LhK(x/h)$, for a value $h > 0$ that we will choose later. It should be clear from the construction that f_1 is L-Lipschitz continuous. Let's now compute the KL divergence.

$$\begin{aligned}
\mathrm{KL}(P_0, P_1) &= \int_{[0,1]^d} \int p_0(x,y) \log(\frac{p_0(x,y)}{p_1(x,y)}) \mathrm{d}y \mathrm{d}x \\
&= \int_{[0,1]^d} \int p_0(y \mid x) \log(\frac{p_0(y \mid x)}{p_1(y \mid x)}) \mathrm{d}y \mathrm{d}x \\
&= \int_{[0,1]^d} \int \phi(y) \log(\frac{\phi(y)}{\phi(y - f_1(x))}) \mathrm{d}y \mathrm{d}x \\
&= \int_{[0,1]^d} \mathrm{KL}(\mathcal{N}(0,1), \mathcal{N}(f_1(x), 1)) \mathrm{d}x \\
&= \frac{1}{2} \int_{[0,1]^d} f_1(x)^2 \mathrm{d}x \\
&= \frac{L^2 h^2}{2} \int_{[0,1]^d} K(x/h)^2 \mathrm{d}x
\end{aligned}$$

$$\leq \frac{L^2 h^{2+d} \|K\|_2^2}{2}.$$

In the second line, we use the fact that $p_0(x) = p_1(x) = 1$ for all x; in the fifth, we use the closed-form expression for the KL divergence between normals; and in the sixth and seventh, we recall the definition of f_1 and use variable substitution to compute the integral, denoting $\|K\|_2^2 = \int K(x)^2 dx$. Now, by setting $h = ((2\log 2)/(L^2 n \|K\|_2^2))^{1/(2+d)}$, $\mathrm{KL}(P_0, P_1) \leq (\log 2)/n$.

Then, by Corollary 1,

$$\begin{aligned}\inf_{\hat{f}} \sup_{f \in C^1(L;[0,1]^d)} \mathbb{E}[(\hat{f}(0) - f(0))^2] &\geq \frac{f_1(0)^2}{32} \\ &= \frac{L^2 h^2}{32} \\ &\asymp L^{2d/(2+d)} n^{-2/(2+d)}\end{aligned}$$

This means we have found a tight lower bound and KDE achieves the point-wise minimax optimal rate of convergence.

Example 2 (Lipschitz function, fixed X)**.** *Suppose we now look at the fixed-X regression model. Then $y_i, i = 1, \ldots, n$ are independent but no longer i.i.d.. It turns our that very few changes will be required to amend the arguments given above with Le Cam's method in the i.i.d. case. Careful inspection shows that we must only replace P_j^n, $j = 0, 1$ with $P_{j1} \times \cdots \times P_{jn}$, $j = 0, 1$, whose densities are $\prod_{i=1}^n p_{ji}(z_i)$, $j = 0, 1$, and then the lower bounds would still hold. The KL bound from Theorem 1 simply becomes*

$$R_n \geq \frac{d(\theta_0, \theta_1)}{8C} e^{-\sum_{i=1}^n \mathrm{KL}(P_{0i}, P_{1i})}$$

Using an analogous construction to that from the random-X setting, we define $f_0 = 0$ and $f_1(x) = LhK(x/h)$, where K is 1-Lipschitz, supported on the unit ball, with $K(0) = 1$, and now satisfies $\|K\|_n^2 = \frac{1}{n}\sum_{i=1}^n K(x_i)^2 = c$ for some $0 < c < \infty$ that does not depend on n.

Satisfying this last requirement, which requires us to construct K so that we have precise control over its empirical norm, is easiest to do when $x_i, i = 1, \ldots, n$ are on a regular lattice in $[0, 1]^d$, which is a typical assumption in fixed-X lower bounds. Similar calculations to the previous example can be used to show

$$\frac{1}{n}\sum_{i=1}^n \mathrm{KL}(P_{0i}, P_{1i}) = \frac{L^2 h^2}{2n} \sum_{i=1}^n K(x_i/h) \lesssim L^2 h^{2+d}.$$

If we set $h \asymp (L^2 n)^{-1/(2+d)}$, then we get

$$\inf_{\hat{f}} \sup_{f \in C^1(L;[0,1]^d)} \mathbb{E}[(\hat{f}(0) - f(0))^2] \gtrsim f_1(0)^2 \asymp L^2 h^2 \asymp L^{2d/(2+d)} n^{-2/(2+d)},$$

just as in the random-X setting.

However, Le Cam will not always provide useful bounds. Lets take the following regression example.

Example 3 (A bad choice for Le Cam)**.** *Consider the regression model*

$$Y_i = f(i/n) + \varepsilon_i$$

where $f \in \mathcal{H}(1,1) = \Theta$ (Holder class). We can show the following risk upper bound for the L_∞ loss:

$$\limsup_{n\to\infty} \sup_{\mathcal{H}(1,1)} \mathbb{E}[\|\hat{f} - f\|_\infty^2] \leq \left(\frac{\log n}{n}\right)^{1/3}$$

Lets consider the following pair of hypotheses:

$$\theta_0 = f_0(x) \equiv 0 \qquad \textit{and} \qquad \theta_1 = f_1(x) = (2\pi n)^{-1}\sin(2\pi n x).$$

Notice that for evaluating at the points of the regression, $f_0(i/n) = f_1(i/n) = 0, \forall i$. This implies that $P_0 = P_1$ for the observed data $Y_1, \ldots, Y_n$. It can be shown using the Neyman-Pearson bound from the proof of Le Cam that $p_{e,2} \geq 1/2$. Then, with the L_∞ loss,

$$d(f_0, f_1) = \|f_0 - f_1\|_\infty = (2\pi n)^{-1}.$$

Then we can repeat the steps of lower bounding the rish with $s = (4\pi n)^{-1}$ to show

$$\liminf_{n\to\infty} n^{-2} R_n \geq c.$$

Clearly this does not match the upper bound. And so, the question remains as to whether the upper bound is also rate-optimal. It turns out that we need to use a more involved multiple hypothesis lower bound in order to prove the rate-optimality. We will come back to this example later.

5 Fano's method

Intuitively, it should be clear that Le Cam's method - which only allows us to construct a pair of hypotheses - will likely be insufficient. Recall, however, that the standard reduction in Lemma 1 was based on an arbitrarily large but finite set $S = \{P_1, \ldots, P_N\} \subseteq \mathcal{P}$. Like we did in the derivation of Le Cam's method, we can use the fact that a maximum is no smaller than an average, which gives

$$R_n \geq \frac{s}{2C} \inf_\psi \frac{1}{N} \sum_{j=1}^{n} P_j(\psi \neq j) = \frac{s}{2C} \inf_\psi Q(\psi \neq J).$$

Now, in order to understand the interpretation of this bound, we need to understand a simple, yet well-known result from information theory, known as *Fano's inequality*. The idea of behind this inequality comes from thinking about the data as being generated from the two step-process outlined

in Section 2. Recall that when we think of the data Z as being generated from a mixture distribution $\bar{Q}$ where the true distribution is uniformly randomly chosen from N possible values, we are trying to use the information in Z to correctly identify the true mixture component. Then, we can think of the difficulty of the minimaxity problem as largely dependent on the relationship between Z and J.

Let's take the most extreme relationship to see precisely what this implies. If $Z \perp J$, then Z contains no information about J and furthermore, the joint density of Z and J splits into the product of their marginals. If the joint density looks very different from the product of the marginals, we could argue that Z and J are unlikely to be independent of each other and therefore, Z may contain more information about J than if the joint density looked more like the product of the marginals. This naturally fits into the concept of *distance* or *divergence* between two distributions as a metric for identifying how difficult the problem is to solve through the amount of information Z contains about J. In **Information Theory** this concept is called *mutual information* as is defined as

$$I(Z; J) = \mathrm{KL}(Q_{ZJ}, Q_Z Q_J).$$

In our particular case, since we assume that J follows a uniform distribution over N values, we can use the third property of KL from Lemma 2 to further decompose the mutual information into

$$I(Z; J) = \frac{1}{N} \sum_j \mathrm{KL}(P_j, \bar{Q}).$$

Now, we can see that the mutual information is small only if the $P_j, j = 1, \ldots, N$ are hard to distinguish from each other. Or in other words, if the elements of the packing set are too close to each other, i.e., δ is too small. This quantification of the packing set will help us establish a lower bound that accounts for the trade-off when more elements are added to S.

In particular, Fano's inequality tells us that for any ψ,

$$\frac{1}{N} \sum_{j=1}^{n} P_j(\psi \neq j) \geq 1 - \frac{I(Z; J) + \log 2}{\log N} \geq 1 - \frac{n\beta + \log 2}{\log N}$$

where $\beta = \max_{j \neq k} \mathrm{KL}(P_j, P_k)$, is also known as the maximum KL-gap. Putting this together with the general bound from Lemma 1 gives the following result.

Theorem 2 (Fano's lower bound)
Let $P_1, \ldots, P_N \in \mathcal{P}$, and let $d(\cdot, \cdot)$ be a nonnegative symmetric loss satisfying the quasi-triangle inequality for some $C > 0$. Then

$$R_n \geq \frac{s}{2C} \left(1 - \frac{I(Z; J) + \log 2}{\log N}\right) \geq \frac{s}{2C} \left(1 - \frac{n\beta + \log 2}{\log N}\right)$$

where s is the minimum distance, and β is the maximum KL-gap.

While the second inequality is one way of controlling mutual information, another method that may also be useful employs convexity of the KL divergence:

$$I(Z; J) \leq \frac{1}{N^2} \sum_{j,k} \mathrm{KL}(P_j, P_k).$$

Finally, we state a simple corollary of Fano's method.

Corollary 2
Under the same conditions on $d(\cdot,\cdot)$ as in Theorem 2, suppose there exists $P_1,\dots,P_N \in \mathcal{P}$ such that $N \geq 4$ and $\beta \leq (\log N)/(4n)$. Then $R_n \geq s/(8C)$.

5.1 Selecting hypothesis classes

The bound in Theorem 2 depends largely on selecting a reasonable δ. Here we will briefly touch upon how this can be done to ensure the lower bound remains non-trivial. We will use sets of the form $S = \{P_\omega : \omega \in \Omega\}$, where

$$\Omega = \{0,1\}^m = \{\omega = (\omega_1,\dots,\omega_m) : \omega_i \in \{0,1\}, i = 1,\dots,m\}.$$

Ω is also known as a hypercube. There are 2^m elements in Ω. For $\omega, \nu \in \Omega$, define the Hamming distance:

$$H(\omega,\nu) = \sum_{i=1}^{m} \mathbf{1}(\omega_i \neq \nu_i)$$

One "problem" with a hypercube (in terms of using it to index distributions), is that some pairs P_ω, P_ν might be very close together which will make the minimum d-gap (what we call s) too small (in a relative sense). This will result in a poor lower bound.

We can try to fix this problem by pruning the hypercube. That is, we will seek some subset $\Omega' \subseteq \Omega$ covering nearly all the elements of Ω, but where each pair P_ω, P_ν is 'far apart' in Hamming distance, for $\omega, \nu \in \Omega'$ with $\omega \neq \nu$. It may help to think of this as an approach to selecting the elements of the packing set with respect to the Hamming distance instead of in some L_p norm. The technique for constructing such a subset is outlined by the Varshamov-Gilbert lemma.

Lemma 3 (Varshamov-Gilbert)
Let $\Omega = \{0,1\}^m$, where $m \geq 8$. Then there exists a pruned hypercube $\Omega' = \{\omega_1,\dots,\omega_N\} \subseteq \Omega$ such that

- *$N \geq 2^{m/8}$, and*
- *$H\left(\omega_j, \omega_k\right) \geq m/8$ for each $j \neq k$.*

For our purposes we will take this result as a given. See (Tsybakov, 2009, Chapter 2.6) for a complete proof of this lemma. Lets now see how we can combine this information theoretic result with Fano's method to establish an integrated-loss minimax bound.

Example 4 (Lipschitz functions, L_2 norm)**.** *Consider the random-X regression setting (3) with the squared L_2 loss:*

$$d(\hat{f}, f) = \|\hat{f} - f\|_2^2 = \int_{[0,1]^d} (\hat{f}(x) - f(x))^2 \mathrm{d}x$$

As before, let K be a 1-Lipschitz function supported on the unit ℓ_2 ball $\{x : \|x\|_2 \leq 1\}$, such that $K(0) = 1$ and $0 < \int K(x)^2 \mathrm{d}x < \infty$. For an integer $r > 0$ (we will set the value later), define the grid

points

$$x_\alpha = \left(\frac{\alpha_1 - 1/2}{r}, \ldots, \frac{\alpha_r - 1/2}{r} \right) \in [0,1]^d, \quad \text{for } \alpha \in [r]^d,$$

where $[r] = \{1, \ldots, r\}$. *Let* $h = 1/(2r)$ *and define the functions*

$$g_\alpha(x) = LhK\left(\frac{x - x_\alpha}{h} \right), \quad \text{for } \alpha \in [r]^d.$$

It is straightforward to check that each g_α *is L-Lipschitz, and that each of the* g_α*'s have non-overlapping support. Now, enumerate these functions as* $g_1, \ldots, g_m$, *for* $m = r^d$, *and define*

$$f_\omega(x) = \sum_{i=1}^{m} \omega_i g_i(x), \quad \text{for } \omega \in \{0,1\}^m.$$

In other words, we construct each hypothesis f_ω *by adding together some finite subset of the locally-supported kernels* $g_1, \ldots, g_m$, *indexed by* ω.

For $\omega, \nu \in \Omega$, *since the functions* g_α *have non-overlapping support,*

$$\begin{aligned} \int_{[0,1]^d} (f_\omega(x) - f_\nu(x))^2 \mathrm{d}x &= \int_{[0,1]^d} \left(\sum_{i=1}^{m} (\omega_i - \nu_i)\, g_i(x) \right)^2 \mathrm{d}x \\ &= H(\omega, \nu) L^2 h^2 \int_{[0,1]^d} K\left(\frac{x}{h} \right)^2 \mathrm{d}x \\ &= H(\omega, \nu) L^2 h^{2+d} \|K\|_2^2 \end{aligned}$$

where $H(\omega, \nu)$ *is the Hamming distance, and* $\|K\|_2^2 = \int K(x)^2 \mathrm{d}x$. *A similar calculation to the pointwise loss case shows that for the hypotheses* P_ω, P_ν *corresponding to the regression functions* f_ω, f_ν, *respectively,*

$$\begin{aligned} \mathrm{KL}\left(P_\omega, P_\nu\right) &= \frac{1}{2} \int_{[0,1]^d} (f_\omega(x) - f_\nu(x))^2 \mathrm{d}x \\ &= H(\omega, \nu) \cdot L^2 h^{2+d} \|K\|_2^2 / 2. \end{aligned}$$

At this point we can apply the Varshamov-Gilbert lemma to define the hypercube subset $\Omega' = \left\{ \omega^1, \ldots, \omega^N \right\} \subseteq \Omega = \{0,1\}^d$, *with cardinality* $N \geq 2^{m/8}$, *such that* $H\left(\omega^j, \omega^k\right) \geq m/8$ *for each* $j \neq k$. *Then for each* $j = 1, \ldots, N$, *denote by* P_j *the distribution corresponding to the regression function* f_{ω_j}. *Observe that, from the integrated loss and the lower bound on the Hamming distance over distinct pairs in* Ω',

$$s = \min_{j \neq k} \|f_{\omega_j} - f_{\omega_k}\|_2^2 \geq \frac{m}{8} L^2 h^{2+d} \|K\|_2^2 = cL^2 r^{-2}$$

Meanwhile, by the KL distance and the trivial upper bound on the Hamming distance of m,

$$\beta = \max_{j \neq k} \mathrm{KL}(P_j, P_k) \leq \frac{m}{2} L^2 h^{2+d} \|K\|_2^2 = 4cL^2 r^{-2}.$$

Now, we would like to have $\beta \leq (\log N)/(4n)$ *in order to be able to apply Corollary 2. Recalling that* $N \geq 2^{m/8}$*, we have* $\log N \geq (\log 2)m/8 = (\log 2)r^d/8$*, so we want*

$$4cL^2 r^{-2} \leq (\log 2) r^d/(16n),$$

which implies we must choose $r = \lceil c'(L^2 n)^{1/(2+d)} \rceil$ *for some constant* $c' > 0$*. Corollary 2 then tells us (using the fact that squared loss satisfies the quasi-triangle inequality with* $C = 2$ *) that*

$$\begin{aligned} \inf_{\hat{f}} \sup_{f \in C^1(L;[0,1]^d)} \mathbb{E}[\int_{[0,1]^d} (\hat{f}(x) - f(x))^2 \mathrm{d}x] &\geq \frac{s}{16} \\ &= \frac{cL^2 r^{-2}}{16} \\ &\asymp L^{2d/(2+d)} n^{-2/(2+d)} \end{aligned}$$

A similar calculation is possible (and the same rate holds) for the fixed-X case, but we will skip the details.

Cautionary note: The fixed-X minimax rate is not always the same as the random-X rate. In fact, often it is necessary to be careful in setting up the minimax estimation problems, because in some cases, the answers may be trivial. Of course there is much more that can be discussed based on approaches to minimaxity for different function classes, but we leave this tangent here for now.

There are many more methods for constructing lower bounds than just the Le Cam and Fano methods. We won't cover these, but the interested reader can see Yu (1997) and (Tsybakov, 2009, Chapter 2.7) for other techniques.

References

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Yu, B. (1997). Assouad, fano and le cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics, Springer-Verlag, New York*, pages 423–435.