Minimax Theory for Nonparametric Regression

Advanced Topics in Statistical Learning, Spring 2023

Ryan Tibshirani (based on notes by Larry Wasserman)

1 Introduction

When we are doing theory for nonparametric regression (or really statistical estimation in general), how can we tell if a convergence rate that we can prove for a particular method is "impressive"? Can the analysis be tightened? Or the method itself improved? And even if we carried this out, will such refinements actually lead to a better convergence rate?

The answer to the last question can be provided by minimax theory, which is a set of techniques for characterizing the best worst-case behavior of a procedure over a class of distributions for a particular statistical learning task.

Let $\mathcal{P}$ be a set of distributions, and let $Z_1, \dots, Z_n$ be i.i.d. from $P \in \mathcal{P}$. Let $\theta(P)$ be some functional of P (we will give several concrete examples shortly), and let $\hat{\theta} = \hat{\theta}(Z_1, \dots, Z_n)$ denote an estimator of θ, based on the sample $Z_1, \dots, Z_n$. Given a symmetric nonnegative loss function d (acting over the space in which $\theta(P)$ lies), we define the *minimax risk* over $\mathcal{P}$ with respect to d to be

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[d(\theta(P), \hat{\theta})\big],$$

where the infimum is over all estimators $\hat{\theta}$, and we use the subscript P on the expectation to refer to the fact that we are averaging over the samples $Z_1, \dots, Z_n$ drawn from P, that are used to form $\hat{\theta}$.

This may all look a little obscure. What does the class $\mathcal{P}$ look like for some typical problems? What about the functional $\theta(P)$, and the loss d? Examples will help.

Example: Gaussian mean estimation. As a simple parametric example, suppose that $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$. For $P = N(\theta, 1)$, we can just take our functional to be $\theta(P) = \theta$, the mean. Consider estimating the mean with the squared loss $d(a, b) = (a - b)^2$. The minimax risk is

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

It is implicit notationally that the expectation here is taken over i.i.d. samples $Z_1, \dots, Z_n \sim N(\theta, 1)$, used to fit $\hat{\theta}$.

For parametric models, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $\Theta \subseteq \mathbb{R}^d$, recall that under regularity conditions, the MLE has risk $\lesssim \mathrm{tr}[I(\theta)^{-1}]/n$ at θ, where $I(\theta)$ is the Fisher information matrix (and for typical models this will be of the order d/n). Meanwhile, it can be shown that there is a local minimax lower bound—local in the sense that the sup is taken over a neighborhood around θ—of the same order $\mathrm{tr}[I(\theta)^{-1}]/n$. Thus the MLE is locally minimax. In fact, it is more than this, because this statement can be made to be uniform over all local neighborhoods around all $\theta \in \Theta$. This is due to a general theory developed by Hájek and Le Cam, but we won't cover any of this. We'll focus on nonparametric minimax theory (assuming you've seen parametric minimax theory in previous courses).

Example: nonparametric function estimation at a point, Random-X. Let Q be a fixed distribution on $[0, 1]^d$ (e.g, the uniform distribution), and let $Z_i = (x_i, y_i)$, $i = 1, \dots, n$ be i.i.d. from P, with

$$y_i = f(x_i) + \epsilon_i, \quad x_i \sim Q, \quad \epsilon_i \sim N(0, \sigma^2), \quad \text{and} \quad x_i \perp\!\!\!\perp \epsilon_i, \tag{1}$$

for some fixed $\sigma^2 > 0$. Let $\theta(P) = f$, which is an entire function. Suppose that $\mathcal{P}$ is the set of distributions P of the form (1) for which $f \in \mathcal{F}$, for some class of functions $\mathcal{F}$ on $[0,1]^d$. To study function estimation at a single point—say, the origin—we can take the loss to be $d(\hat{f}, f) = (\hat{f}(0) - f(0))^2$. The minimax risk is

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

The expectation is understood to be with respect to (1), which describes the samples used to fit $\hat{f}$.

Example: nonparametric function estimation at a point, Fixed-X. Similar to the last example, but now suppose that y_i, $i = 1, \ldots, n$ are independent draws from P, with

$$y_i = f(x_i) + \epsilon_i, \quad x_i \text{ fixed}, \quad \text{and} \quad \epsilon_i \sim N(0, \sigma^2). \tag{3}$$

We can still define the minimax risk as in (2), where now the expectation is understood to be with respect to (3). This requires some notational adjustment in the introductory paragraphs, because now y_i, $i = 1, \ldots, n$ are independent but no longer i.i.d. (this will be true of all Fixed-X models that we'll discuss henceforth). Similarly, we would need to adjust some of the techniques (Le Cam, Fano) that will be introduced below, because as written they assume i.i.d. data. In several cases, these adjustments will be straightforward and the minimax risk for the Random-X and Fixed-X models will behave the same. However, interestingly, in other cases this will not be true, and the minimax risk for the Random-X and Fixed-X models will be very different. We'll discuss this at the end.

Example: nonparametric function estimation in population L^2 norm, Random-X. As in our running example, under the Random-X model (1), consider the loss $d(\hat{f}, f) = \|\hat{f} - f\|^2_{L^2(Q)}$, where recall Q is the input distribution. This yields the minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\bigg[\int (\hat{f}(x) - f(x))^2 \, dQ(x)\bigg], \tag{4}$$

where the expectation is with respect to (1), which describes the samples used to fit $\hat{f}$.

Example: nonparametric function estimation in empirical L^2 norm Random-X. As in our running example, under the Random-X model (1), consider the loss $d(\hat{f}, f) = \|\hat{f} - f\|^2_{L^2(Q)}$, where Q_n denotes the empirical distribution of x_i, $i = 1, \ldots, n$. This yields the minimax risk

$$R_n = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (\hat{f}(x_i) - f(x_i))^2\bigg], \tag{5}$$

where the expectation is with respect to (1), which describes the samples used to fit $\hat{f}$.

1.1 KL divergence

The *Kullback-Leibler divergence* (KL) between two distributions P, Q, having densities p, q, respectively, is

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

KL divergence will play a prominent role in a lot of the calculations that follow. The following elementary fact will be useful for us. For Gaussians, $P = N(\theta, \sigma^2)$ and $Q = N(\mu, \sigma^2)$, we have

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

In general, $\mathrm{KL}(P, Q)$ is nonnegative and zero iff $P = Q$. This one of the properties required of a distance (interpreting "distance" as being an equivalent term to "metric"). Yet KL divergence is not a distance, as it fails each of the other two properties: it is not symmetric, nor does it satisfy the triangle inequality.

Nonetheless, you'll sometimes hear people calling it "KL distance" anyway. There are many other distances on distributions (TV, L^1, Hellinger, χ^2, etc.) as well many relationships known between them, including relationships to KL divergence. We do not review these here, but will simply define other distances and use known relationships as they naturally arise in what follows. See, e.g., Chapter 2.4 of Tsybakov (2009) for a more thorough treatment of distances on distributions.

1.2 Notation

As in the nonparametric regression lectures, for sequences a_n, b_n, we will write $a_n \lesssim b_n$ to mean $a_n = O(b_n)$, and we use $a_n \asymp b_n$ to mean $a_n = O(b_n)$ and $b = O(a_n)$. We also use the abbreviation $a \wedge b = \min\{a, b\}$.

2 Standard reduction

Typically we will not be interested in R_n exactly, but only its dependence on n.[1] Of course, if $\hat{\theta}$ is a particular estimator, then

$$R_n \leq \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[d(\theta(P), \hat{\theta})\big],$$

so if the rate of convergence of $\hat{\theta}$ over the class of distributions $\mathcal{P}$ is (say) n^{-w}, then we learn $R_n \lesssim n^{-w}$.

Finding a lower bound on R_n will require a totally different technique, which we will outline below. But if we can establish that $R_n \gtrsim n^{-w}$, matching the upper bound in rate, then we conclude that $R_n \asymp n^{-w}$ and we consider the case to be closed.

How do we find a lower bound? We reduce the problem to a hypothesis testing problem. We do this because, in (certain simple) hypothesis testing problems, it can be easier to reason about optimality.

The general approach works like this. Fix a finite set of distributions $S = \{P_1, \ldots, P_N\} \subseteq \mathcal{P}$. Then

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

where we abbreviate where $\theta_j = \theta(P_j)$ and $E_j = \mathbb{E}_{P_j}$. By Markov's inequality, for each j, and any $t > 0$,

$$E_j\big[d(\theta_j, \hat{\theta})\big] \geq t P_j\big\{d(\theta_j, \hat{\theta}) \geq t\big\},$$

thus

$$R_n \geq t \cdot \inf_{\hat{\theta}} \max_{P_j \in S} P_j\big\{d(\theta_j, \hat{\theta}) \geq t\big\}. \tag{6}$$

Any value of t will give us a valid lower bound, but to find the "right" value of t, let's look at a calculation involving the minimum gap between distinct θ_j, $j = 1, \ldots, N$.

Minimum gap calculation. Define

$$s = \min_{j \neq k} d(\theta_j, \theta_k). \tag{7}$$

Given an arbitrary estimator $\hat{\theta}$, define

$$\psi^* = \operatorname*{argmin}_{j=1,\ldots,N} d(\theta_j, \hat{\theta}). \tag{8}$$

Let's assume that d satisfies a quasi-triangle inequality, of the form

$$d(\theta, \theta') \leq C d(\theta, \theta'') + C d(\theta', \theta''), \quad \text{for all } \theta, \theta', \theta'', \tag{9}$$

and a global constant $C > 0$. For example, if d is a metric, then it would satisfy (9) with $C = 1$, and if $d(\theta, \theta') = \|\theta - \theta'\|_2^2$, then it would satisfy it with $C = 2$.

[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.

Now, if $\psi^* \neq j$, then letting $k = \psi^*$, observe that

$$\begin{aligned} s &\leq d(\theta_j, \theta_k) \\ &\leq Cd(\theta_j, \hat{\theta}) + Cd(\theta_k, \hat{\theta}) \\ &\leq 2Cd(\theta_j, \hat{\theta}). \end{aligned}$$

In the second line we use the quasi-triangle inequality, and in the third we use $d(\theta_k, \hat{\theta}) \leq d(\theta_j, \hat{\theta})$ (because $k = \psi^*$). Therefore we have shown that $\psi^* \neq j$ implies that $d(\theta_j, \hat{\theta}) \geq s/(2C)$, and

$$P_j\left\{d(\theta_j, \hat{\theta}) \geq \frac{s}{2C}\right\} \geq P_j(\psi^* \neq j). \tag{10}$$

Back to minimax risk. Backing up, we have shown from (6) and (10), plugging in $t = s/(2C)$, that

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

where we write $\psi^* = \psi^*(\hat{\theta})$ to emphasize its dependence on $\hat{\theta}$. But in fact we can go further. We can ψ^* as defined in (8) as multiple hypothesis test: given access to $\hat{\theta}$, it tries to pick out which one of θ_j it thinks is most likely. We can continue on lower bounding the right-hand side in the last display by considering *all* hypothesis tests that have access to the data (on which the estimator $\hat{\theta}$ is fit). We'll summarize this in a proposition for easy reference.

Proposition 1. *Let $S = \{P_1, \ldots, P_N\} \subseteq \mathcal{P}$ be any finite set, and d be a nonnegative symmetric loss satisfying the quasi-triangle inequality (9) with a constant $C > 0$. Then*

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P\big[d(\theta(P), \hat{\theta})\big] \geq \frac{s}{2C} \cdot \inf_{\psi} \max_{P_j \in S} P_j(\psi \neq j), \tag{11}$$

where s is the minimum gap as in (7), and the infimum is over all maps ψ from the data to $\{1, \ldots, N\}$.

This is called *the standard reduction* for minimax lower bounds. Making the best use of (11) (i.e., getting a tight lower bound) requires carefully crafting $S = \{P_1, \ldots, P_N\}$. 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.

3 Le Cam's method

Le Cam's method is only a short hop away from the standard reduction. Consider just two hypotheses $\theta_0 = \theta(P_0)$ and $\theta_1 = \theta(P_1)$, so that $s = d(\theta_0, \theta_1)$. Let's also start with $n = 1$ so we only have a single observation. Then (11) tells us that

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

Since a maximum is no smaller than an average,

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

The reason that we switched from max testing risk to aggregate testing risk is that, for the latter, we know what optimality looks like: this is given by 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 elementary yet critical fact that $\inf_\psi [P_0(\psi \neq 0) + P_1(\psi \neq 1)] = P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1)$. This is the essence of the Neyman-Pearson lemma.

Now we compute

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

Thus we have shown that

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

Supposing we have n observations, 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

$$R_n \geq \frac{s}{4C}\big[P_0(\psi \neq 0) + P_1(\psi \neq 1)\big] = \frac{s}{4C} \int p_0^n(z) \wedge p_1^n(z)\,dz. \tag{12}$$

The integral on the right-hand side above is often called the *affinity* between p_0^n and p_1^n. Using relationships between affinity, TV distance, and KL divergence gives the set of results summarized in the next theorem.

Theorem 1 (Le Cam's lower bound)**.** *Let $P_0, P_1 \in \mathcal{P}$, and let d be a nonnegative symmetric loss satisfying the quasi-triangle inequality (9) with a constant $C > 0$. Then*

$$R_n \geq \frac{d(\theta_0, \theta_1)}{4C}\big[1 - \mathrm{TV}(P_0^n, P_1^n)\big], \tag{13}$$

where $\mathrm{TV}(P, Q) = \frac{1}{2}\int |p(z) - q(z)|\,dz$ denotes the total variation distance between distributions P, Q with densities p, q. We also have the further lower bound

$$R_n \geq \frac{d(\theta_0, \theta_1)}{8C} e^{-n\mathrm{KL}(P_0, P_1)}. \tag{14}$$

The lower bounds in (13) and (14) simply come from (12), combined with the following facts about affinity, TV distance, and KL divergence of distributions P, Q with densities p, q.

- $\int p(z) \wedge q(z)\,dz = 1 - \mathrm{TV}(P, Q)$.
- $\int p(z) \wedge q(z) \geq \frac{1}{2} e^{-\mathrm{KL}(P,Q)}$.
- $\mathrm{KL}(P^n, Q^n) = n \cdot \mathrm{KL}(P, Q)$.

A useful corollary of Le Cam's KL bound (14) is the following.

Corollary 1. *Under the same conditions on d 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 $R_n \geq d(\theta_0, \theta_1)/(16C)$.*

3.1 Example: Lipschitz function estimation at a point, Random-X

We can demonstrate the utility of Le Cam's method by considering a nonparametric regression model of the form (1). For simplicity, let's take the input distribution to be uniform, $Q = \mathrm{Unif}([0,1]^d)$, 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 $[0,1]^d$, and consider pointwise risk at the origin, in squared loss, as in (2).

Recall that in this context, $\theta_0 = f_0(0)$ and $\theta_1 = f_1(0)$, where f_0, f_1 are functions on $[0,1]^d$, and they are required to be Lipschitz in order for $P_0, P_1 \in \mathcal{P}$. Let's just fix $f_0 = 0$ (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 specify later. It is not hard to verify that f_1 is L-Lipschitz continuous. We compute

$$\begin{aligned}
\mathrm{KL}(P_0, P_1) &= \int_{[0,1]^d} \int p_0(x,y) \log\left(\frac{p_0(x,y)}{p_1(x,y)}\right) dy\, dx \\
&= \int_{[0,1]^d} \int p_0(y|x) \log\left(\frac{p_0(y|x)}{p_1(y|x)}\right) dy\, dx \\
&= \int_{[0,1]^d} \int \phi(y) \log\left(\frac{\phi(y)}{\phi(y - f_1(x))}\right) dy\, dx \\
&= \int_{[0,1]^d} \mathrm{KL}\big(N(0,1), N(f_1(x),1)\big)\, dx \\
&= \frac{1}{2} \int_{[0,1]^d} f_1(x)^2\, dx \\
&= \frac{L^2 h^2}{2} \int_{[0,1]^d} K(x/h)^2\, dx \\
&\leq \frac{L^2 h^{2+d} \|K\|_2^2}{2}.
\end{aligned}$$

In the second line, we use that $p_0(x) = p_1(x) = 1$ for all x; in the third, we use ϕ for the standard normal density; in the fourth, we recognize the inner integral as a KL divergence between $N(0,1)$ and $N(f_1(x),1)$; 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 let $h = ((2\log 2)/(L^2 n \|K\|_2^2))^{1/(2+d)}$. Then $\mathrm{KL}(P_0, P_1) \leq (\log 2)/n$, so by Corollary 1 (where we note that squared loss satisfies the quasi-triangle inequality (9) with $C = 2$):

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

Meanwhile, kNN regression or kernel smoothing can be shown to achieve the same pointwise rate, which means we have found a tight lower bound.

3.2 Example: Lipschitz function estimation at a point, Fixed-X

Suppose we change the problem setting from the previous subsection to a Fixed-X model, now assuming (3) instead of (1). Then y_i, $i = 1, \ldots, n$ are independent but no longer i.i.d. Thankfully, 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, $j = 0, 1$ in (12), (13) with $P_{j1} \times \cdots \times P_{jn}$, $j = 0, 1$, whose densities are $(P_{j1} \times \cdots \times P_{jn})(z) = \prod_{i=1}^n p_{ji}(z_i)$, $j = 0, 1$, and then the lower bounds would still hold. The KL bound (14) similarly becomes

$$R_n \geq \frac{d(\theta_0, \theta_1)}{8C} e^{-\sum_{i=1}^n \mathrm{KL}(P_{0i}, P_{1i})}. \tag{15}$$

Using an analogous construction to that from the last subsection, 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 grow with 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, $i = 1, \dots, n$ are on a regular lattice in $[0,1]^d$, which is a typical assumption in Fixed-X lower bounds.

Similar calculations to the last subsection can be used to show

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

From (15), we learn that if we set $h \asymp (L^2 n)^{-1/(2+d)}$, then we get

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

just as in the Random-X setting.

4 Fano's method

When we move from a pointwise loss to an integrated loss, such as population or empirical L^2 loss, Le Cam's method—which only allows us to construct a pair of hypotheses that are hard to distinguish—is usually insufficient.

Recall, however, that the standard reduction (11) was based on an arbitrarily large but finite set $S = \{P_1, \dots, P_N\} \subseteq \mathcal{P}$. 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

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

Now *Fano's inequality*, a well-known result in information theory, tells us that for any ψ,

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

where

$$\beta = \max_{j \neq k} \mathrm{KL}(P_j, P_k). \tag{16}$$

Putting this together gives the following result.

Theorem 2 (Fano's lower bound)**.** *Let $P_1, \dots, P_N \in \mathcal{P}$, and let d be a nonnegative symmetric loss satisfying the quasi-triangle inequality (9) with a constant $C > 0$. Then*

$$R_n \geq \frac{s}{2C}\left(1 - \frac{n\beta + \log 2}{\log N}\right), \tag{17}$$

where s is the minimum d-gap as in (7), and β is the maximum KL-gap as in (16).

Corollary 2. *Under the same conditions on d 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 $R_n \geq s/(8C)$.*

There are many more methods for constructing lower bounds than just the Le Cam and Fano methods. We won't cover these, but see, e.g., Yu (1997); Yang and Barron (1999), as well as Chapter 2.7 of Tsybakov (2009), for other techniques.

4.1 Varshamov-Gilbert lemma

To use Fano's method or Tsybakov's method, we need to construct a finite class of distributions. Often we will use set of the form $\{P_\omega : \omega \in \Omega\}$, where

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

which is called a hypercube. There are 2^m elements in Ω. For $\omega, \nu \in \Omega$, their *Hamming distance* is

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

One "problem" with a hypercube, in terms of using it to index distributions that we will construct, is that some pairs P_ω, P_ν might be very close together which will make the minimum d-gap, which recall is given in (7), too small. 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 $\Omega' \subseteq \Omega$ having nearly the same number of elements as Ω, but where each pair P_ω, P_ν is far apart in Hamming distance, for $\omega, \nu \in \Omega'$ with $\omega \neq \nu$. The technique for constructing such a pruned hypercube is given to us by what is known as the *Varshamov-Gilbert lemma*.

Lemma 1 (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*

1. *$N \geq 2^{m/8}$; and*
2. *$H(\omega^j, \omega^k) \geq m/8$ for each $j \neq k$.*

This is a standard result in information theory and its proof is somewhat interesting because it involves randomization and Hoeffding's inequality, but we won't cover it here. See, e.g., Chapter 2.6 in Tsybakov (2009).

4.2 Example: Lipschitz function estimation in population L^2 norm

We now demonstrate the utility of Fano's method by consdering the same problem setup as in Section 3.1 but with the squared L^2 loss

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

As before, 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 $0 < \int K(x)^2\, dx < \infty$. For an integer $r > 0$ to be specified later, define grid points

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

where we abbreviate $[r] = \{1, \dots, r\}$. 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 they have non-overlapping supports. Now just enumerate these functions as $g_1, \dots, g_m$, for $m = r^d$, and define

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

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

For $\omega, \nu \in \Omega$, note that by the non-overlapping supports property,

$$\begin{aligned}\int_{[0,1]^d} (f_\omega(x) - f_\nu(x))^2 \, dx &= \int_{[0,1]^d} \bigg(\sum_{i=1}^m (\omega_i - \nu_i) g_i(x) \bigg)^2 \, dx \\ &= H(\omega, \nu) \cdot L^2 h^2 \int_{[0,1]^d} K\bigg(\frac{x}{h}\bigg)^2 \, dx \\ &= H(\omega, \nu) \cdot L^2 h^{2+d} \|K\|_2^2, \end{aligned} \tag{18}$$

where $H(\omega, \nu)$ is the Hamming distance between ω, ν, and $\|K\|_2^2 = \int K(x)^2 \, dx$. A similar calculation to that done in the pointwise loss case shows that for the hypotheses P_ω, P_ν corresponding to the regression functions f_ω, f_ν, respectively,

$$\begin{aligned}\mathrm{KL}(P_\omega, P_\nu) &= \frac{1}{2} \int_{[0,1]^d} (f_\omega(x) - f_\nu(x))^2 \, dx \\ &= H(\omega, \nu) \cdot L^2 h^{2+d} \|K\|_2^2 / 2, \end{aligned} \tag{19}$$

with the calculation for the second line just following like that for (18).

At this point we apply the Varshamov-Gilbert lemma to produce a pruned hypercube $\Omega' = \{\omega^1, \ldots, \omega^N\} \subseteq \Omega = \{0,1\}^d$, with cardinality $N \geq 2^{m/8}$, such that $H(\omega^j, \omega^k) \geq m/8$ 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 (18) and the lower bound on the Hamming distance over distinct pairs in Ω',

$$s = \min_{j \neq k} \|f_{\omega^j} - f_{\omega^k}\|_2^2 \geq m L^2 h^{2+d} \|K\|_2^2 / 8 = c L^2 r^{-2}.$$

for a constant $c > 0$. Meanwhile, from (19), and the trivial upper bound on the Hamming distance of m,

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

Finally, it is time to choose the grid side length r. 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

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

which leads us to choose $r = \lceil c' (L^2 n)^{1/(2+d)} \rceil$ for another constant $c' > 0$. Corollary 2 then tells us (using again that squared loss satisfies the quasi-triangle inequality (9) with $C = 2$) that

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

Recall, we know from our earlier nonparametric regression lecture that kNN regression and kernel smoothing each achieve the above rate in squared L^2 norm, so we know that our lower bound is tight.

Furthermore, a similar calculation is possible (and the same rate holds) for the Fixed-X case, but we skip the details.

5 Cautionary note

The Fixed-X minimax rate is not always the same as the Random-X rate. In fact, you have to be careful how you set up minimax estimation problems, because in some cases, you'll get trivial answers.

Let's take the example of nonparametric regression over the Sobolev class $\mathcal{F} = W^{m,2}(L; [0,1]^d)$, which we write to mean the set of functions f on $[0,1]^d$ that are m times weakly differentiable with

$$\int_{[0,1]^d} \sum_{|\alpha|=m} [D^\alpha f(x)]^2 \, dx \leq L^2.$$

When $2m > d$, standard calculations can be used to show that (assuming for simplicity at the moment that L does not grow with n):

$$\inf_{\hat{f}} \sup_{f \in W^{m,2}(L;[0,1]^d)} \mathbb{E}\bigg[\int_{[0,1]^d} (\hat{f}(x) - f(x))^2 \, dx \bigg] \asymp n^{-2s/(2s+d)},$$

in both the Random-X and Fixed-X models. Note that this generalizes all results for Hölder classes, including Lipschitz classes, that we have seen in lectures thus far.

When $2m \leq d$, in a Random-X model, it can still be shown that the minimax rate in squared L^2 loss over $W^{m,2}(L; [0,1]^d)$ is still as above, $n^{-2s/(2s+d)}$. However, when $2m \leq d$ and we use a Fixed-X model it turns out that the minimax rate is a constant—meaning that the sup risk of any estimator over $W^{m,2}(L; [0,1]^d)$ cannot go to zero!

The underlying issue here is similar to what we encountered in the splines lecture. In what we referred to there as the subcritical regime, $2m \leq d$, we cannot really talk about point evaluation in a Sobolev space (as the point evaluation operator is not continuous). Fortunately, in a Random-X model, the randomness in the inputs x_i, $i = 1, \ldots, n$ sort of finesses this problem, in a sense—which is itself fairly remarkable. In a Fixed-X model, however, obtaining knowledge of $f(x_i)$, $i = 1, \ldots, n$ doesn't help you reason about what f looks like on the rest of the domain $[0,1]^d$, for $f \in W^{m,2}(L; [0,1]^d)$. (This is the case even if you know that its Sobolev seminorm is zero!) And this makes consistent estimation, in the sense of sup risk, impossible.

There's more to say about this general phenomenon but to avoid going on a huge tangent, we'll save that for another day (another lecture?).

References

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

Bin Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.