10/36-702: Minimax Theory

1 Introduction

When solving a statistical learning problem, there are often many procedures to choose from. This leads to the following question: how can we tell if one statistical learning procedure is better than another? One answer is provided by *minimax theory* which is a set of techniques for finding the minimum, worst case behavior of a procedure.

References: Yu (2008), Tsybakov (2009), van der Vaart (1998), Wasserman (2014).

2 Definitions and Notation

Let $\mathcal{P}$ be a set of distributions and let $X_1, \ldots, X_n$ be a sample from some distribution $P \in \mathcal{P}$. Let $\theta(P)$ be some function of P. For example, $\theta(P)$ could be the mean of P, the variance of P or the density of P. Let $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ denote an estimator. Given a metric d, the *minimax risk* is

$$R_n \equiv R_n(\mathcal{P}) = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \tag{1}$$

where the infimum is over all estimators. The *sample complexity* is

$$n(\epsilon, \mathcal{P}) = \min\Bigl\{n : \ R_n(\mathcal{P}) \leq \epsilon\Bigr\}. \tag{2}$$

Example 1 *Suppose that $\mathcal{P} = \{N(\theta, 1) : \ \theta \in \mathbb{R}\}$ where $N(\theta, 1)$ denotes a Gaussian with mean θ and variance 1. Consider estimating θ with the metric $d(a, b) = (a - b)^2$. The minimax risk is*

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(\widehat{\theta} - \theta)^2]. \tag{3}$$

In this example, θ is a scalar.

Example 2 *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a distribution P. Let $m(x) = \mathbb{E}_P(Y|X = x) = \int y\, dP(y|X = x)$ be the regression function. In this case, we might use the metric $d(m_1, m_2) = \int (m_1(x) - m_2(x))^2 dx$ in which case the minimax risk is*

$$R_n = \inf_{\widehat{m}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[\int (\widehat{m}(x) - m(x))^2 \right]. \tag{4}$$

In this example, θ is a function.

Notation. Recall that the *Kullback-Leibler distance* between two distributions P_0 and P_1 with densities p_0 and p_1 is defined to be

$$\mathsf{KL}(P_0, P_1) = \int \log\left(\frac{dP_0}{dP_1}\right) dP_0 \int \log\left(\frac{p_0(x)}{p_1(x)}\right) p_0(x)dx.$$

The appendix defines several other distances between probability distributions and explains how these distances are related. We write $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. If P is a distribution with density p, the product distribution for n iid observations is P^n with density $p^n(x) = \prod_{i=1}^n p(x_i)$. It is easy to check that $\mathsf{KL}(P_0^n, P_1^n) = n\mathsf{KL}(P_0, P_1)$. For positive sequences a_n and b_n we write $a_n = \Omega(b_n)$ to mean that there exists $C > 0$ such that $a_n \geq Cb_n$ for all large n. $a_n \asymp b_n$ if a_n/b_n is strictly bounded away from zero and infinity for all large n; that is, $a_n = O(b_n)$ and $b_n = O(a_n)$.

3 Bounding the Minimax Risk

Usually, we do not find R_n directly. Instead, we find an upper bound U_n and a lower bound L_n on R_n. To find an upper bound, let $\widehat{\theta}$ be any estimator. Then

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \leq \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \equiv U_n. \tag{5}$$

So the maximum risk of any estimator provides an upper bound U_n. Finding a lower bound L_n is harder. We will consider three methods: the *Le Cam method*, the *Fano method* and *Tsybakov's bound.* If the lower and upper bound are close, then we have succeeded. For example, if $L_n = cn^{-\alpha}$ and $U_n = Cn^{-\alpha}$ for some positive constants c, C and α, then we have established that the *minimax rate of convergence* is $n^{-\alpha}$.

All the lower bound methods involve a the following trick: we reduce the problem to a hypothesis testing problem. It works like this. First, we will choose a finite set of distributions $M = \{P_1, \ldots, P_N\} \subset \mathcal{P}$. Then

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \inf_{\widehat{\theta}} \max_{P_j \in M} \mathbb{E}_j[d(\widehat{\theta}, \theta_j)] \tag{6}$$

where $\theta_j = \theta(P_j)$ and $\mathbb{E}_j$ is the expectation under P_j. Let $s = \min_{j \neq k} d(\theta_j, \theta_k)$. By Markov's inequality,

$$P(d(\widehat{\theta}, \theta) > t) \leq \frac{\mathbb{E}[d(\widehat{\theta}, \theta)]}{t}$$

and so

$$\mathbb{E}[d(\widehat{\theta}, \theta)] \geq tP(d(\widehat{\theta}, \theta) > t).$$

Setting $t = s/2$, and using (6), we have

$$R_n \geq \frac{s}{2} \inf_{\widehat{\theta}} \max_{P_j \in M} P_j(d(\widehat{\theta}, \theta_j) > s/2). \tag{7}$$

Given any estimator $\widehat{\theta}$, define

$$\psi^* = \operatorname*{argmin}_j d(\widehat{\theta}, \theta_j).$$

Now, of $\psi^* \neq j$ then, letting $k = \psi^*$,

$$\begin{aligned} s &\leq d(\theta_j, \theta_k) \leq d(\theta_j, \widehat{\theta}) + d(\theta_k, \widehat{\theta}) \\ &\leq d(\theta_j, \widehat{\theta}) + d(\theta_j, \widehat{\theta}) \ \text{ since } \psi^* \neq j \text{ implies that } d(\widehat{\theta}, \theta_k) \leq d(\widehat{\theta}, \theta) \\ &= 2d(\theta_j, \widehat{\theta}). \end{aligned}$$

So $\psi^* \neq j$ implies that $d(\theta_j, \widehat{\theta}) \geq s/2$. Thus

$$P_j(d(\widehat{\theta}, \theta_j) > s/2) \geq P_j(\psi^* \neq j) \geq \inf_{\psi} P_j(\psi \neq j)$$

where the infimum is over all maps ψ form the data to $\{1, \ldots, N\}$. (We can think of ψ is a multiple hypothesis test.) Thus we have

$$R_n \geq \frac{s}{2} \inf_{\psi} \max_{P_j \in M} P_j(\psi \neq j).$$

We can summarize this as a theorem:

Theorem 3 *Let $M = \{P_1, \ldots, P_N\} \subset \mathcal{P}$ and let $s = \min_{j \neq k} d(\theta_j, \theta_k)$. Then*

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{s}{2} \inf_{\psi} \max_{P_j \in M} P_j(\psi \neq j). \tag{8}$$

Getting a good lower bound involves carefully selecting $M = \{P_1, \ldots, P_N\}$. If M is too big, s will be small. If M is too small, then $\max_{P_j \in M} P_j(\psi \neq j)$ will be small.

4 Lower Bound Method 1: Le Cam

Theorem 4 *Let $\mathcal{P}$ be a set of distributions. For any pair $P_0, P_1 \in \mathcal{P}$,*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{4} \int [p_0^n(x) \wedge p_1^n(x)] dx \geq \frac{\Delta}{8} e^{-n\mathsf{KL}(P_0, P_1)} \tag{9}$$

where $\Delta = d(\theta(P_0), \theta(P_1))$.

Remark: The second inequality is useful, if $\mathsf{KL}(P_0, P_1) = \infty$, since it is usually difficult to compute $\int [p_0^n(x) \wedge p_1^n(x)] dx$ directly. An alternative is

$$\int [p_0^n(x) \wedge p_1^n(x)] dx \geq \frac{1}{2} \left(1 - \frac{1}{2} \int |p_0 - p_1|\right)^{2n}. \tag{10}$$

Corollary 5 *Suppose there exist $P_0, P_1 \in \mathcal{P}$ such that $\mathsf{KL}(P_0, P_1) \leq \log 2/n$. Then*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{16} \tag{11}$$

where $\Delta = d(\theta(P_0), \theta(P_1))$.

Proof. Let $\theta_0 = \theta(P_0)$, $\theta_1 = \theta(P_1)$ and $\Delta = d(\theta_0, \theta_1)$. First suppose that $n = 1$ so that we have a single observation X. An estimator $\widehat{\theta}$ defines a *test statistic* ψ, namely,

$$\psi(X) = \begin{cases} 1 & \text{if } d(\widehat{\theta}, \theta_1) \leq d(\widehat{\theta}, \theta_0) \\ 0 & \text{if } d(\widehat{\theta}, \theta_1) > d(\widehat{\theta}, \theta_0). \end{cases}$$

If $P = P_0$ and $\psi = 1$ then

$$\Delta = d(\theta_0, \theta_1) \leq d(\theta_0, \widehat{\theta}) + d(\theta_1, \widehat{\theta}) \leq d(\theta_0, \widehat{\theta}) + d(\theta_0, \widehat{\theta}) = 2d(\theta_0, \widehat{\theta})$$

and so $d(\theta_0, \widehat{\theta}) \geq \frac{\Delta}{2}$. Hence

$$\mathbb{E}_{P_0}[d(\widehat{\theta}, \theta_0)] \geq \mathbb{E}_{P_0}[d(\widehat{\theta}, \theta_0) I(\psi = 1)] \geq \frac{\Delta}{2}\mathbb{E}_{P_0}[I(\psi = 1)] = \frac{\Delta}{2} P_0(\psi = 1). \tag{12}$$

Similarly,

$$\mathbb{E}_{P_1}[d(\widehat{\theta}, \theta_1)] \geq \frac{\Delta}{2} P_1(\psi = 0). \tag{13}$$

Taking the maximum of (12) and (13), we have

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \max_{P \in \{P_0, P_1\}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{2} \max\Big\{P_0(\psi = 1), P_1(\psi = 0)\Big\}.$$

Taking the infimum over all estimators, we have

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{2}\pi$$

where

$$\pi = \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j). \tag{14}$$

Since a maximum is larger than an average,

$$\pi = \inf_{\psi} \max_{j=0,1} P_j(\psi \neq j) \geq \inf_{\psi} \frac{P_0(\psi \neq 0) + P_1(\psi \neq 1)}{2}.$$

Define the *Neyman-Pearson test*

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

In Lemma 7 below, we show that the sum of the errors $P_0(\psi \neq 0) + P_1(\psi \neq 1)$ is minimized by ψ^*. Now

$$\begin{aligned} P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1) &= \int_{p_1 > p_0} p_0(x)dx + \int_{p_0 > p_1} p_1(x)dx \\ &= \int_{p_1 > p_0} [p_0(x) \wedge p_1(x)]dx + \int_{p_0 > p_1} [p_(x) \wedge p_1(x)]dx = \int [p_0(x) \wedge p_1(x)]dx. \end{aligned}$$

Thus,

$$\frac{P_0(\psi_* \neq 0) + P_1(\psi_* \neq 1)}{2} = \frac{1}{2} \int [p_0(x) \wedge p_1(x)]dx.$$

Thus we have shown that

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{4} \int [p_0(x) \wedge p_1(x)]dx.$$

Now suppose we have n observations. Then, replacing p_0 and p_1 with $p_0^n(x) = \prod_{i=1}^n p_0(x_i)$ and $p_1^n(x) = \prod_{i=1}^n p_1(x_i)$, we have

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{4} \int [p_0^n(x) \wedge p_1^n(x)]dx.$$

In Lemma 7 below, we show that $\int p \wedge q \geq \frac{1}{2} e^{-\mathsf{KL}(P,Q)}$. Since $\mathsf{KL}(P_0^n, P_1^n) = n\mathsf{KL}(P_0, P_1)$, we have

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{8} e^{-n\mathsf{KL}(P_0,P_1)}.$$

The result follows. □

Lemma 6 *Let ψ^* be the Neyman-Pearson test. For any test ψ,*

$$P_0(\psi = 1) + P_1(\psi = 0) \geq P_0(\psi^* = 1) + P_1(\psi^* = 0).$$

Proof. Recall that $p_0 > p_1$ when $\psi^* = 0$ and that $p_0 < p_1$ when $\psi^* = 1$. So

$$\begin{aligned} P_0(\psi = 1) + P_1(\psi = 0) &= \int_{\psi=1} p_0(x)dx + \int_{\psi=0} p_1(x)dx \\ &= \int_{\psi=1,\psi^*=1} p_0(x)dx + \int_{\psi=1,\psi^*=0} p_0(x)dx + \int_{\psi=0,\psi^*=0} p_1(x)dx + \int_{\psi=0,\psi^*=1} p_1(x)dx \\ &\geq \int_{\psi=1,\psi^*=1} p_0(x)dx + \int_{\psi=1,\psi^*=0} p_1(x)dx + \int_{\psi=0,\psi^*=0} p_1(x)dx + \int_{\psi=0,\psi^*=1} p_0(x)dx \\ &= \int_{\psi^*=1} p_0(x)dx + \int_{\psi^*=0} p_1(x)dx \\ &= P_0(\psi^* = 1) + P_1(\psi^* = 0). \end{aligned}$$

□

Lemma 7 *For any P and Q, $\int p \wedge q \geq \frac{1}{2} e^{-\mathsf{KL}(P,Q)}$.*

Proof. First note that $\int \max(p,q) + \int \min(p,q) = 2$. Hence

$$\begin{aligned} 2\int p \wedge q &\geq \left[2 - \int p \wedge q\right] \int p \wedge q = \int p \wedge q \int p \vee q \\ &\geq \left(\int \sqrt{(p \wedge q)\ (p \vee q)}\right)^2 \geq \left(\int \sqrt{pq}\right)^2 = \exp\left(2 \log \int \sqrt{pq}\right) \\ &= \exp\left(2 \log \int p \sqrt{q/p}\right) \geq \exp\left(2 \log \int p \log \sqrt{\frac{q}{p}}\right) = e^{-\mathsf{KL}(P,Q)} \end{aligned}$$

where we used Jensen's inequality in the last inequality. □

Example 8 *Consider data $(X_1,Y_1),\ldots,(X_n,Y_n)$ where $X_i \sim \text{Uniform}(0,1)$, $Y_i = m(X_i) + \epsilon_i$ and $\epsilon_i \sim N(0,1)$. Assume that*

$$m \in \mathcal{M} = \left\{ m:\ |m(y) - m(x)| \leq L|x-y|, \quad \text{for all } x, y \in [0,1] \right\}.$$

So $\mathcal{P}$ is the set of distributions of the form $p(x,y) = p(x)p(y|x) = \phi(y - m(x))$ where $m \in \mathcal{M}$.

How well can we estimate $m(x)$ at some point x? Without loss of generality, let's take $x = 0$ so the parameter of interest is $\theta = m(0)$. Let $d(\theta_0,\theta_1) = |\theta_0 - \theta_1|$. Let $m_0(x) = 0$ for all x. Let $0 \leq \epsilon \leq 1$ and define

$$m_1(x) = \begin{cases} L(\epsilon - x) & 0 \leq x \leq \epsilon \\ 0 & x \geq \epsilon. \end{cases}$$

Then $m_0, m_1 \in \mathcal{M}$ and $\Delta = |m_1(0) - m_0(0)| = L\epsilon$. The KL distance is

$$\begin{aligned} \mathsf{KL}(P_0,P_1) &= \int_0^1 \int p_0(x,y) \log\left(\frac{p_0(x,y)}{p_1(x,y)}\right) dydx \\ &= \int_0^1 \int p_0(x) p_0(y|x) \log\left(\frac{p_0(x)p_0(y|x)}{p_1(x)p_1(y|x)}\right) dydx \\ &= \int_0^1 \int \phi(y) \log\left(\frac{\phi(y)}{\phi(y - m_1(x))}\right) dydx \\ &= \int_0^\epsilon \int \phi(y) \log\left(\frac{\phi(y)}{\phi(y - m_1(x))}\right) dydx \\ &= \int_0^\epsilon \mathsf{KL}(N(0,1), N(m_1(x),1)) dx. \end{aligned}$$

Now, $\mathsf{KL}(N(\mu_1,1), N(\mu_2,1)) = (\mu_1 - \mu_2)^2/2$. So

$$\mathsf{KL}(P_0,P_1) = \frac{L^2}{2} \int_0^\epsilon (\epsilon - x)^2 dx = \frac{L^2 \epsilon^3}{6}.$$

Let $\epsilon = (6 \log 2/(L^2 n))^{1/3}$. Then, $\mathsf{KL}(P_0, P_1) = \log 2/n$ and hence, by Corollary 5,

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{\Delta}{16} = \frac{L\epsilon}{16} = \frac{L}{16}\left(\frac{6\log 2}{L^2 n}\right)^{1/3} = \left(\frac{c}{n}\right)^{1/3}. \tag{15}$$

It is easy to show that the regressogram (regression histogram) $\widehat{\theta} = \widehat{m}(0)$ has risk

$$\mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \leq \left(\frac{C}{n}\right)^{1/3}.$$

Thus we have proved that

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \asymp n^{-\frac{1}{3}}. \tag{16}$$

The same calculations in d dimensions yield

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \asymp n^{-\frac{1}{d+2}}. \tag{17}$$

On the squared scale we have

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d^2(\widehat{\theta}, \theta(P))] \asymp n^{-\frac{2}{d+2}}. \tag{18}$$

Similar rates hold in density estimation.

5 Lower Bound Method II: Fano

For metrics like $d(f, g) = \int (f - g)^2$, Le Cam's method will usually not give a tight bound. Instead, we use Fano's method. Instead of choosing two distributions P_0, P_1, we choose a finite set of distributions $P_1, \ldots, P_N \in \mathcal{P}$.

We start with Fano's lemma.

Lemma 9 (Fano) *Let $X_1, \ldots, X_n \sim P$ where $P \in \{P_1, \ldots, P_N\}$. Let Z be any function of $X_1, \ldots, X_n$ taking values in $\{1, \ldots, N\}$. Let $\beta = \max_{j \neq k} \mathsf{KL}(P_j, P_k)$. Then*

$$\frac{1}{N}\sum_{j=1}^{N} P_j(Z \neq j) \geq \left(1 - \frac{n\beta + \log 2}{\log N}\right).$$

Proof: See Lemma 32 in the appendix. $\Box$.

Now we can state and prove the Fano minimax bound.

Theorem 10 *Let $F = \{P_1, \ldots, P_N\} \subset \mathcal{P}$. Let $\theta(P)$ be a parameter taking values in a metric space with metric d. Then*

$$\inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \ \mathbb{E}_P \left(d\Big(\widehat{\theta}, \theta(P)\Big)\right) \geq \frac{\alpha}{2} \ \left(1 - \frac{n\beta + \log 2}{\log N}\right) \tag{19}$$

where

$$\alpha = \min_{j\neq k} d\left(\theta(P_j), \theta(P_k)\right), \tag{20}$$

and

$$\beta = \max_{j\neq k} \mathsf{KL}(P_j, P_k). \tag{21}$$

Corollary 11 (Fano Minimax Bound) *Suppose there exists $F = \{P_1, \ldots, P_N\} \subset \mathcal{P}$ such that $N \geq 16$ and*

$$\beta = \max_{j\neq k} \mathsf{KL}(P_j, P_k) \leq \frac{\log N}{4n}. \tag{22}$$

Then

$$\inf_{\widehat{\theta}} \max_{P\in\mathcal{P}} \ \mathbb{E}_P \left[d\Big(\widehat{\theta}, \theta(P)\Big)\right] \geq \frac{\alpha}{4}. \tag{23}$$

Proof. Let $\widehat{\theta}$ be any estimator and let $Z = \operatorname{argmin}_{j\in\{1,\ldots,N\}} d(\widehat{\theta}, \theta_j)$. For any $j \neq Z$, $d(\widehat{\theta}_n(X), \theta_j) \geq \alpha/2$. Hence,

$$\mathbb{E}_{P_j} \left(d(\widehat{\theta}_n, \theta(P_j))\right) \geq \left(\frac{\alpha}{2}\right) P_j(Z \neq j)$$

and so

$$\max_j \mathbb{E}_{P_j} d(\widehat{\theta}_n, \theta(P_j)) \geq \frac{\alpha}{2} \max_j P_j(Z \neq j) \geq \frac{\alpha}{2}\frac{1}{N}\sum_{j=1}^N P_j(Z \neq j).$$

By Fano's lemma,

$$\frac{1}{N}\sum_{j=1}^N P_j(Z \neq j) \geq \left(1 - \frac{n\beta + \log 2}{\log N}\right).$$

Thus,

$$\inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \ \mathbb{E}_P \left(d\Big(\widehat{\theta}, \theta(P)\Big)\right) \geq \inf_{\widehat{\theta}} \max_{P\in\mathcal{F}} \ \mathbb{E}_P \left(d\Big(\widehat{\theta}, \theta(P)\Big)\right) \geq \frac{\alpha}{2} \ \left(1 - \frac{n\beta + \log 2}{\log N}\right). \tag{24}$$

□

6 Lower Bound Method III: Tsybakov's Bound

This approach is due to Tsybakov (2009).

Theorem 12 (Tsybakov 2009) *Let $X_1, \ldots, X_n \sim P \in \mathcal{P}$. Let $\{P_0, P_1, \ldots, P_N\} \subset \mathcal{P}$ where $N \geq 3$. Assume that P_0 is absolutely continuous with respect to each P_j. Suppose that*

$$\frac{1}{N}\sum_{j=1}^{N} \mathrm{KL}(P_j, P_0) \leq \frac{\log N}{16}.$$

Then

$$\inf_{\widehat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\widehat{\theta}, \theta(P))] \geq \frac{s}{16}$$

where

$$s = \max_{0\leq j<k\leq N} d(\theta(P_j), \theta(P_k)).$$

Proof. Let $X = (X_1, \ldots, X_n)$ and let $\psi \equiv \psi(X) \in \{0, 1, \ldots, N\}$. Fix $\tau > 0$ and define

$$A_j = \left\{ \frac{dP_0}{dP_j} \geq \tau \right\}.$$

Then

$$\begin{aligned}
P_0(\psi \neq 0) &= \sum_{j=1}^{N} P_0(\psi = j) \geq \sum_{j=1}^{N} P_0\Big(\psi = j \bigcap A_j\Big)\\
&= \sum_{j=1}^{N} \frac{P_0\Big(\psi = j \bigcap A_j\Big)}{P_j\Big(\psi = j \bigcap A_j\Big)} P_j\Big(\psi = j \bigcap A_j\Big)\\
&\geq \tau \sum_{j=1}^{N} P_j\Big(\psi = j \bigcap A_j\Big)\\
&\geq \tau \sum_{j=1}^{N} P_j\Big(\psi = j\Big) - \tau \sum_{j=1}^{N} P_j\Big(A_j^c\Big)\\
&= \tau N \left(\frac{1}{N}\sum_{j=1}^{N} P_j\Big(\psi = j\Big)\right) - \tau N \left(\frac{1}{N}\sum_{j=1}^{N} P_j\Big(A_j^c\Big)\right)\\
&= \tau N (p_0 - a)
\end{aligned}$$

where

$$p_0 = \frac{1}{N}\sum_{j=1}^{N} P_j\Big(\psi = j\Big), \quad a = \frac{1}{N}\sum_{j=1}^{N} P_j\left(\frac{dP_0}{dP_j} < \tau\right).$$

Hence,

$$\begin{aligned}
\max_{0\le j\le N} P_j(\psi\neq j) &= \max\left\{ P_0(\psi\neq 0),\, \max_{1\le j\le N} P_j(\psi\neq j)\right\}\\
&\ge \max\left\{ \tau N(p_0-a),\; \max_{1\le j\le N} P_j(\psi\neq j)\right\}\\
&\ge \max\left\{ \tau N(p_0-a),\; \frac{1}{N}\sum_{j=1}^{N} P_j(\psi\neq j)\right\}\\
&= \max\left\{ \tau N(p_0-a),\; 1-p_0\right\}\\
&\ge \min_{0\le p\le 1}\max\left\{ \tau N(p-a), 1-p\right\} = \frac{\tau N(1-a)}{1+\tau N}\\
&= \left(\frac{\tau N}{1+\tau N}\right)\left[\frac{1}{N}\sum_{j=1}^{N} P_j\left(\frac{dP_0}{dP_j}\ge \tau\right)\right].
\end{aligned}$$

In Lemma 13 below we show that

$$\frac{1}{N}\sum_{j=1}^{N} P_j\left(\frac{dP_0}{dP_j}\ge \tau\right)\ge 1-\frac{a_*+\sqrt{a_*/2}}{\log(1/\tau)}$$

where $\alpha_* = N^{-1}\sum_j K(P_j,P_0)$. Choosing $\tau = 1/\sqrt{N}$ we get

$$\begin{aligned}
\max_{0\le j\le N} P_j(\psi\neq j) &\ge \frac{\sqrt{N}}{1+\sqrt{N}}\left(1-\frac{a_*+\sqrt{a_*/2}}{\log(1/\tau)}\right)\\
&= \frac{\sqrt{N}}{1+\sqrt{N}}\left(1-\frac{2(a_*+\sqrt{a_*/2})}{\log N}\right)\\
&\ge \frac{1}{2}\left(1-\frac{1}{4}\right) = \frac{3}{8}.
\end{aligned}$$

By Theorem 3,

$$R_n \ge \frac{s}{2}\frac{3}{8}\ge \frac{s}{8}.$$

□

Lemma 13 *Let $a_* = N^{-1}\sum_j K(P_j,P_0)$. Then,*

$$\frac{1}{N}\sum_{j=1}^{N} P_j\left(\frac{dP_0}{dP_j}\ge \tau\right)\ge 1-\frac{a_*+\sqrt{a_*/2}}{\log(1/\tau)}.$$

Proof. Note that

$$P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) = P_j\left(\frac{dP_j}{dP_0} \leq \frac{1}{\tau}\right)$$
$$= 1 - P_j\left(\log\frac{dP_j}{dP_0} \geq \log\left(\frac{1}{\tau}\right)\right).$$

By Markov's inequality,

$$P_j\left(\log\frac{dP_j}{dP_0} \geq \log\left(\frac{1}{\tau}\right)\right) \leq P_j\left(\left[\log\frac{dP_j}{dP_0}\right]_+ \geq \log\left(\frac{1}{\tau}\right)\right) \leq \frac{1}{\log(1/\tau)}\int\left[\log\frac{dP_j}{dP_0}\right]_+ dP_j.$$

According to Pinsker's second inequality (see Thheorem 29 in the appendix and Tsyabakov Lemma 2.5),

$$\int\left[\log\frac{dP_j}{dP_0}\right]_+ dP_j \leq K(P_j,P_0) + \sqrt{K(P_j,K_0)/2}.$$

So

$$P_j\left(\log\frac{dP_j}{dP_0} \geq \log\left(\frac{1}{\tau}\right)\right) \geq 1 - \frac{1}{\log(1/\tau)}\left[K(P_j,P_0) + \sqrt{K(P_j,K_0)/2}\right].$$

Using Jensen's inequality,

$$\frac{1}{N}\sum_j\sqrt{K(P_j,P_0)} \leq \sqrt{\frac{1}{N}\sum_j K(P_j,P_0)} = \sqrt{a_*}.$$

So

$$\frac{1}{N}\sum_{j=1}^N P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) \geq 1 - \frac{1}{\log(1/\tau)}\frac{1}{N}\sum_j K(P_j,P_0) - \frac{1}{\log(1/\tau)}\frac{1}{N}\sum_j\sqrt{K(P_j,P_0)/2}$$
$$\geq 1 - \frac{a_*}{\log(1/\tau)} - \frac{\sqrt{a_*/2}}{\log(1/\tau)}.$$

□

7 Hypercubes

To use Fano's method or Tsyabkov's method, we need to construct a finite class of distributions $\mathcal{F}$. Sometimes we use a set of the form

$$\mathcal{F} = \Big\{P_\omega :\ \omega \in \Omega\Big\}$$

where

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

which is called a hypercube. There are $N = 2^m$ distributions in $\mathcal{F}$. For $\omega, \nu \in \Omega$, define the *Hamming distance* $H(\omega, \nu) = \sum_{j=1}^{m} I(\omega_k \neq \nu_k)$.

One problem with a hypercube is that some pairs $P, Q \in \mathcal{F}$ might be very close together which will make $\alpha = \min_{j \neq k} d\left(\theta(P_j), \theta(P_k)\right)$ small. This will result in a poor lower bound. We can fix this problem by pruning the hypercube. That is, we can find a subset $\Omega' \subset \Omega$ which has nearly the same number of elements as Ω but such that each pair $P, Q \in \mathcal{F}' = \Big\{ P_\omega : \ \omega \in \Omega' \Big\}$ is far apart. We call Ω' a *pruned hypercube.* The technique for constructing Ω' is the *Varshamov-Gilbert lemma.*

Lemma 14 *(Varshamov-Gilbert) Let* $\Omega = \Big\{ \omega = (\omega_1, \ldots, \omega_N) : \ \omega_j \in \{0,1\} \Big\}$*. Suppose that* $N \geq 8$*. There exists* $\omega^0, \omega^1, \ldots, \omega^M \in \Omega$ *such that (i)* $\omega^0 = (0, \ldots, 0)$*, (ii)* $M \geq 2^{N/8}$ *and (iii)* $H(\omega^{(j)}, \omega^{(k)}) \geq N/8$ *for* $0 \leq j < k \leq M$*. We call* $\Omega' = \{\omega^0, \omega^1, \ldots, \omega^M\}$ *a pruned hypercube.*

Proof. Let $D = \lfloor N/8 \rfloor$. Set $\omega^0 = (0, \ldots, 0)$. Define $\Omega_0 = \Omega$ and $\Omega_1 = \{\omega \in \Omega : \ H(\omega, \omega^0) > D\}$. Let ω^1 be any element in Ω_1. Thus we have eliminated $\{\omega \in \Omega : \ H(\omega, \omega^0) \leq D\}$. Continue this way recursively and at the j^{th} step define $\Omega_j = \{\omega \in \Omega_{j-1} : \ H(\omega, \omega^{j-1}) > D\}$ where $j = 1, \ldots, M$. Let n_j be the number of elements eliminated at step j, that is, the number of elements in $A_j = \{\omega \in \Omega_j : \ H(\omega, \omega^{(j)}) \leq D\}$. It follows that

$$n_j \leq \sum_{i=0}^{D} \binom{N}{i}.$$

The sets $A_0, \ldots, A_M$ partition Ω and so $n_0 + n_1 + \cdots + n_M = 2^N$. Thus,

$$(M+1) \sum_{i=0}^{D} \binom{N}{i} \geq 2^N.$$

Thus

$$M + 1 \geq \frac{1}{\sum_{i=0}^{D} 2^{-N} \binom{N}{i}} = \frac{1}{\mathbb{P}\left(\sum_{i=1}^{N} Z_i \leq \lfloor m/8 \rfloor\right)}$$

where $Z_1, \ldots, Z_N$ are iid Bernoulli (1/2) random variables. By Hoeffding's inequaity,

$$\mathbb{P}\left(\sum_{i=1}^{N} Z_i \leq \lfloor m/8 \rfloor\right) \leq e^{-9N/32} < 2^{-N/4}.$$

Therefore, $M \geq 2^{N/8}$ as long as $N \geq 8$. Finally, note that, by construction, $H(\omega^j, \omega^k) \geq D + 1 \geq N/8$. □

Example 15 *Consider data* $(X_1,Y_1),\ldots,(X_n,Y_n)$ *where* $X_i \sim \text{Uniform}(0,1)$, $Y_i = f(X_i) + \epsilon_i$ *and* $\epsilon_i \sim N(0,1)$. *(The assumption that* X *is uniform is not crucial.) Assume that* f *is in the Holder class* $\mathcal{F}$ *defined by*

$$\mathcal{F} = \left\{ f:\ |f^{(\ell)}(y) - f^{(\ell)}(x)| \leq L|x-y|^{\beta-\ell}, \ \text{ for all } x,y \in [0,1] \right\}$$

where $\ell = \lfloor \beta \rfloor$. $\mathcal{P}$ *is the set of distributions of the form* $p(x,y) = p(x)p(y|x) = \phi(y - m(x))$ *where* $f \in \mathcal{F}$. *Let* Ω' *be a pruned hypercube and let*

$$\mathcal{F}' = \left\{ f_\omega(x) = \sum_{j=1}^m \omega_j \phi_j(x) : \omega \in \Omega' \right\}$$

where $m = \lceil cn^{\frac{1}{2\beta+1}} \rceil$, $\phi_j(x) = Lh^\beta K((x - X_j)/h)$, *and* $h = 1/m$. *Here,* K *is any sufficiently smooth function supported on* $(-1/2, 1/2)$. *Let* $d^2(f,g) = \int (f-g)^2$. *Some calculations show that, for* $\omega, \nu \in \Omega'$,

$$d(f_\omega, f_\nu) = \sqrt{H(\omega,\nu)} L h^{\beta + \frac{1}{2}} \int K^2 \geq \sqrt{\frac{m}{8}} L h^{\beta+\frac{1}{2}} \int K^2 \geq c_1 h^\beta.$$

We used the Varshamov-Gilbert result which implies that $H(\omega,\nu) \geq m/8$. *Furthermore,*

$$\mathsf{KL}(P_\omega, P_\nu) \leq c_2 h^{2\beta}.$$

To apply Corollary 11, we need

$$\mathsf{KL}(P_\omega, P_\nu) \leq \frac{\log N}{4n} = \frac{\log 2^{m/8}}{4n} = \frac{m}{32n} = \frac{1}{32nh}.$$

This holds if we set $h = (c/n)^{1/(2\beta+1)}$. *In that case,* $d(f_\omega, f_\nu) \geq c_1 h^\beta = c_1 (c/n)^{\beta/(2\beta+1)}$. *Corollary 11 implies that*

$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\widehat{f}, f)] = \Omega(n^{-\frac{\beta}{2\beta+1}}).$$

It follows that

$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \int (f - \widehat{f})^2 \geq n^{-\frac{2\beta}{2\beta+1}}.$$

It can be shown that there are kernel estimators that achieve this rate of convergence. (The kernel has to be chosen carefully to take advantage of the degree of smoothness β*.) A similar calculation in* d *dimensions shows that*

$$\inf_{\widehat{f}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \int (f - \widehat{f})^2 \geq n^{-\frac{2\beta}{2\beta+d}}.$$

8 Further Examples

8.1 Parametric Maximum Likelihood

For parametric models that satisfy weak regularity conditions, the maximum likelihood estimator is approximately minimax. Consider squared error loss which is squared bias plus variance. In parametric models with large samples, it can be shown that the variance term dominates the bias so the risk of the mle $\widehat{\theta}$ roughly equals the variance:[1]

$$R(\theta,\widehat{\theta}) = \mathrm{Var}_\theta(\widehat{\theta}) + \mathrm{bias}^2 \approx \mathrm{Var}_\theta(\widehat{\theta}). \tag{25}$$

The variance of the mle is approximately $\mathrm{Var}(\widehat{\theta}) \approx \frac{1}{nI(\theta)}$ where $I(\theta)$ is the *Fisher information.* Hence,

$$nR(\theta,\widehat{\theta}) \approx \frac{1}{I(\theta)}. \tag{26}$$

For any other estimator θ', it can be shown that for large n, $R(\theta,\theta') \geq R(\theta,\widehat{\theta})$.

Here is a more precise statement, due to Hájek and Le Cam. The family of distributions $(P_\theta :\ \theta \in \Theta)$ with densities $(P_\theta :\ \theta \in \Theta)$ is *differentiable in quadratic mean* if there exists ℓ'_θ such that

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \ell'_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2). \tag{27}$$

Theorem 16 (Hájek and Le Cam) *Suppose that $(P_\theta :\ \theta \in \Theta)$ is differentiable in quadratic mean where $\Theta \subset \mathbb{R}^k$ and that the Fisher information I_θ is nonsingular. Let ψ be differentiable. Then $\psi(\widehat{\theta}_n)$, where $\widehat{\theta}_n$ is the mle, is asymptotically, locally, uniformly minimax in the sense that, for any estimator T_n, and any bowl-shaped ℓ,*

$$\sup_{I\in\mathcal{I}} \liminf_{n\to\infty} \sup_{h\in I} \mathbb{E}_{\theta+h/\sqrt{n}} \ell \left(\sqrt{n} \left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}} \right) \right) \right) \geq \mathbb{E}(\ell(U)) \tag{28}$$

where $\mathcal{I}$ is the class of all finite subsets of $\mathbb{R}^k$ and $U \sim N(0, \psi'_\theta I_\theta^{-1} (\psi'_\theta)^T)$.

For a proof, see van der Vaart (1998). Note that the right hand side of the displayed formula is the risk of the mle. In summary: in well-behaved parametric models, with large samples, the mle is approximately minimax. There is a crucial caveat: these results break down when the number of parameters is large.

[1] Typically, the squared bias is order $O(n^{-2})$ while the variance is of order $O(n^{-1})$.

8.2 Estimating a Smooth Density

Here we use the general strategy to derive the minimax rate of convergence for estimating a smooth density. (See Yu (2008) for more details.)

Let $\mathcal{F}$ be all probability densities f on $[0,1]$ such that

$$0 < c_0 \leq f(x) \leq c_1 < \infty, \quad |f''(x)| \leq c_2 < \infty.$$

We observe $X_1, \ldots, X_n \sim P$ where P has density $f \in \mathcal{F}$. We will use the squared Hellinger distance $d^2(f,g) = \int_0^1 (\sqrt{f(x)} - \sqrt{g(x)})^2 dx$ as a loss function.

Upper Bound. Let $\widehat{f}_n$ be the kernel estimator with bandwidth $h = n^{-1/5}$. Then, using bias-variance calculations, we have that

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f \left(\int (\widehat{f}(x) - f(x))^2 dx \right) \leq C n^{-4/5}$$

for some C. But

$$\int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx = \int \left(\frac{f(x) - g(x)}{\sqrt{f(x)} + \sqrt{g(x)}} \right)^2 dx \leq C' \int (f(x) - g(x))^2 dx \qquad (29)$$

for some C'. Hence $\sup_f \mathbb{E}_f(d^2(f, \widehat{f}_n)) \leq C n^{-4/5}$ which gives us an upper bound.

Lower Bound. For the lower bound we use Fano's inequality. Let g be a bounded, twice differentiable function on $[-1/2, 1/2]$ such that

$$\int_{-1/2}^{1/2} g(x)dx = 0, \int_{-1/2}^{1/2} g^2(x)dx = a > 0, \int_{-1/2}^{1/2} (g'(x))^2 dx = b > 0.$$

Fix an integer m and for $j = 1, \ldots, m$ define $x_j = (j - (1/2))/m$ and

$$g_j(x) = \frac{c}{m^2} g(m(x - x_j))$$

for $x \in [0,1]$ where c is a small positive constant. Let $\mathcal{M}$ denote the Varshamov-Gilbert pruned version of the set

$$\left\{ f_\tau = 1 + \sum_{j=1}^{m} \tau_j g_j(x) : \ \tau = (\tau_1, \ldots, \tau_m) \in \{-1, +1\}^m \right\}.$$

For $f_\tau \in \mathcal{M}$, let f_τ^n denote the product density for n observations and let $\mathcal{M}_n = \left\{ f_\tau^n : \ f_\tau \in \mathcal{M} \right\}$. Some calculations show that, for all τ, τ',

$$\mathsf{KL}(f_\tau^n, f_{\tau'}^n) = n \mathsf{KL}(f_\tau, f_{\tau'}) \leq \frac{C_1 n}{m^4} \equiv \beta. \qquad (30)$$

By Lemma 14, we can choose a subset F of $\mathcal{M}$ with $N = e^{c_0 m}$ elements (where c_0 is a constant) and such that

$$d^2(f_\tau, f_{\tau'}) \geq \frac{C_2}{m^4} \equiv \alpha \tag{31}$$

for all pairs in F. Choosing $m = cn^{1/5}$ gives $\beta \leq \log N/4$ and $d^2(f_\tau, f_{\tau'}) \geq \frac{C_2}{n^{4/5}}$. Fano's lemma implies that

$$\max_j \mathbb{E}_j d^2(\widehat{f}, f_j) \geq \frac{C}{n^{4/5}}.$$

Hence the minimax rate is $n^{-4/5}$ which is achieved by the kernel estimator. Thus we have shown that $R_n(\mathcal{P}) \asymp n^{-4/5}$.

This result can be generalized to higher dimensions and to more general measures of smoothness. Since the proof is similar to the one dimensional case, we state the result without proof.

Theorem 17 *Let $\mathcal{Z}$ be a compact subset of $\mathbb{R}^d$. Let $\mathcal{F}(p, C)$ denote all probability density functions on $\mathcal{Z}$ such that*

$$\int \sum \left| \frac{\partial^p}{\partial z_1^{p_1} \cdots \partial z_d^{p_d}} f(z) \right|^2 dz \leq C$$

where the sum is over al $p_1, \ldots, p_d$ such that $\sum_j p_j = p$. Then there exists a constant $D > 0$ such that

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}(p,C)} \mathbb{E}_f \int (\widehat{f}_n(z) - f(z))^2 dz \geq D \left(\frac{1}{n}\right)^{\frac{2p}{2p+1}}. \tag{32}$$

The kernel estimator (with an appropriate kernel) with bandwidth $h_n = n^{-1/(2p+d)}$ achieves this rate of convergence.

8.3 Minimax Classification

Let us now turn to classification. We focus on some results of Yang (1999), Tsybakov (2004), Mammen and Tsybakov (1999), Audibert and Tsybakov (2005) and Tsybakov and van de Geer (2005).

The data are $Z = (X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \{0, 1\}$. Recall that a classifier is a function of the form $h(x) = I(x \in G)$ for some set G. The classification risk is

$$R(G) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y \neq I(X \in G)) = \mathbb{E}(Y - I(X \in G))^2. \tag{33}$$

The optimal classifier is $h^*(x) = I(x \in G^*)$ where $G^* = \{x : \; m(x) \geq 1/2\}$ and $m(x) = \mathbb{E}(Y|X = x)$. We are interested in how close $R(G)$ is to $R(G^*)$. Following Tsybakov (2004) we define

$$d(G, G^*) = R(G) - R(G^*) = 2 \int_{G \Delta G^*} \left| m(x) - \frac{1}{2} \right| dP_X(x) \tag{34}$$

where $A\Delta B = (A \cap B^c) \cup (A^c \cup B)$ and P_X is the marginal distribution of X.

There are two common types of classifiers. The first type are *plug-in classifiers* of the form $\widehat{h}(x) = I(\widehat{m}(x) \geq 1/2)$ where $\widehat{m}$ is an estimate of the regression function. The second type are *empirical risk minimizers* where $\widehat{h}$ is taken to be the h that minimizes the observed error rate $n^{-1}\sum_{i=1}^n (Y_i \neq h(X_i))$ as h varies over a set of classifiers $\mathcal{H}$. Sometimes one minimizes the error rate plus a penalty term.

According to Yang (1999), the classification problem has, under weak conditions, the same order of difficulty (in terms of minimax rates) as estimating the regression function $m(x)$. Therefore the rates are given in Example 28. According to Tsybakov (2004) and Mammen and Tsybakov (1999), classification is easier than regression. The apparent discrepancy is due to differing assumptions.

To see that classification error cannot be harder than regression, note that for any $\widehat{m}$ and corresponding $\widehat{G}$

$$\begin{aligned} d(G,\widehat{G}) &= 2\int_{G\Delta\widehat{G}} \left|m(x) - \tfrac{1}{2}\right| dP_X(x) && (35)\\ &\leq 2\int |\widehat{m}(x) - m(x)| dP_X(x) \leq 2\sqrt{\int (\widehat{m}(x) - m(x))^2 dP_X(x)} && (36)\end{aligned}$$

so the rate of convergence of $d(G, G^*)$ is at least as fast as the regression function.

Instead of putting assumptions on the regression function m, Mammen and Tsybakov (1999) put an entropy assumption on the set of *decision sets* $\mathcal{G}$. They assume

$$\log N(\epsilon, \mathcal{G}, d) \leq A\epsilon^{-\rho} \tag{37}$$

where $N(\epsilon, \mathcal{G}, d)$ is the smallest number of balls of radius ϵ required to cover $\mathcal{G}$. They show that , if $0 < \rho < 1$, then there are classifiers with rate

$$\sup_P \mathbb{E}(d(\widehat{G}, G^*)) = O(n^{-1/2}) \tag{38}$$

independent of dimension d. Moreover, if we add the margin (or low noise) assumption

$$\mathbb{P}_X\left(0 < \left|m(X) - \tfrac{1}{2}\right| \leq t\right) \leq Ct^\alpha \quad \text{for all } t > 0 \tag{39}$$

we get

$$\sup_P \mathbb{E}(d(\widehat{G}, G^*)) = O\left(n^{-(1+\alpha)/(2+\alpha+\alpha\rho)}\right) \tag{40}$$

which can be nearly $1/n$ for large α and small ρ. The classifiers can be taken to be plug-in estimators using local polynomial regression. Moreover, they show that this rate is minimax. We will discuss classification in the low noise setting in more detail in another chapter.

8.4 Estimating a Large Covariance Matrix

Let $X_1, \ldots, X_n$ be iid Gaussian vectors of dimension d. Let $\Sigma = (\sigma_{ij})_{1 \leq i,j \leq d}$ be the $d \times d$ covariance matrix for X_i. Estimating Σ when d is large is very challenging. Sometimes we can take advantage of special structure. Bickel and Levina (2008) considered the class of *covariance matrices* Σ whose entries have polynomial decay. Specifically, $\Theta = \Theta(\alpha, \epsilon, M)$ is all covariance matrices Σ such that $0 < \epsilon \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq 1/\epsilon$ and such that

$$\max_j \sum_i \left\{ |\sigma_{ij}| : \ |i - j| > k \right\} \leq M k^{-\alpha}$$

for all k. The loss function is $\|\widehat{\Sigma} - \Sigma\|$ where $\| \cdot \|$ is the operator norm

$$\|A\| = \sup_{x: \ \|x\|_2 = 1} \|Ax\|_2.$$

Bickel and Levina (2008) constructed an estimator that that converges at rate $(\log d/n)^{\alpha/(\alpha+1)}$. Cai, Zhang and Zhou (2009) showed that the minimax rate is

$$\min \left\{ n^{-2\alpha/(2\alpha+1)} + \frac{\log d}{n}, \frac{d}{n} \right\}$$

so the Bickel-Levina estimator is not rate minimax. Cai, Zhang and Zhou then constructed an estimator that is rate minimax.

8.5 Semisupervised Prediction

Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ for a classification or regression problem. In addition, suppose we have extra unlabelled data $X_{n+1}, \ldots, X_N$. Methods that make use of the unlabeled are called *semisupervised methods*. We discuss semisupervised methods in another Chapter.

When do the unlabeled data help? Two minimax analyses have been carried out to answer that question, namely, Lafferty and Wasserman (2007) and Singh, Nowak and Zhu (2008). Here we briefly summarize the results of the latter.

Suppose we want to estimate $m(x) = \mathbb{E}(Y|X = x)$ where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Let p be the density of X. To use the unlabelled data we need to link m and p in some way. A common assumption is the *cluster assumption*: m is smooth over clusters of the marginal $p(x)$. Suppose that p has clusters separated by a amount γ and that m is α smooth over each cluster. Singh, Nowak and Zhu (2008) obtained the following upper and lower minimax bounds as γ varies in 6 zones which we label I to VI. These zones relate the size of γ and the number of unlabeled points:

γ	semisupervised upper bound	supervised lower bound	unlabelled data help?
I	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	NO
II	$n^{-2\alpha/(2\alpha+d)}$	$n^{-2\alpha/(2\alpha+d)}$	NO
III	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	YES
IV	$n^{-1/d}$	$n^{-1/d}$	NO
V	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	YES
VI	$n^{-2\alpha/(2\alpha+d)}$	$n^{-1/d}$	YES

The important message is that there are precise conditions when the unlabeled data help and conditions when the unlabeled data do not help. These conditions arise from computing the minimax bounds.

8.6 Graphical Models

Elsewhere in the book, we discuss the problem of estimating graphical models. Here, we shall briefly mention some minimax results for this problem. Let X be a random vector from a multivariate Normal distribution P with mean vector μ and covariance matrix Σ. Note that X is a random vector of length d, that is, $X = (X_1, \ldots, X_d)^T$. The $d \times d$ matrix $\Omega = \Sigma^{-1}$ is called the precision matrix. There is one node for each component of X. The undirected graph associated with P has no edge between X_j and X_j if and only if $\Omega_{jk} = 0$. The edge set is $E = \{(j,k) : \ \Omega_{jk} \neq 0\}$. The graph is $G = (V, E)$ where $V = \{1, \ldots, d\}$ and E is the edge set. Given a random sample of vectors $X^1, \ldots, X^n \sim P$ we want to estimate G. (Only the edge set needs to be estimated; the nodes are known.)

Wang, Wainwright and Ramchandran (2010) found the minimax risk for estimating G under zero-one loss. Let $\mathcal{G}_{d,r}(\lambda)$ denote all the multivariate Normals whose graphs have edge sets with degree at most r and such that

$$\min_{(i,j)\in E} \frac{|\Omega_{jk}|}{\sqrt{\Omega_{jj}\Omega_{kk}}} \geq \lambda.$$

The sample complexity $n(d, r, \lambda)$ is the smallest sample size n needed to recover the true graph with high probability. They show that for any $\lambda \in [0, 1/2]$,

$$n(d,r,\lambda) > \max\left\{ \frac{\log\binom{d-r}{2} - 1}{4\lambda^2},\ \frac{\log\binom{d}{r} - 1}{\frac{1}{2}\left(\log\left(1 + \frac{r\lambda}{1-\lambda}\right) - \frac{r\lambda}{1+(r-1)\lambda}\right)} \right\}. \tag{41}$$

Thus, assuming $\lambda \approx 1/r$, we get that $n \geq Cr^2 \log(d-r)$.

8.7 Deconvolution and Measurement Error

A problem has seems to have received little attention in the machine learning literature is *deconvolution.* Suppose that $X_1, \ldots, X_n \sim P$ where P has density p. We have seen that the minimax rate for estimating p in squared error loss is $n^{-\frac{2\beta}{2\beta+1}}$ where β is the assumed amount of smoothness. Suppose we cannot observe X_i directly but instead we observe X_i with error. Thus, we observe $Y_1, \ldots, Y_n$ where

$$Y_i = X_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{42}$$

The minimax rates for estimating p change drastically. A good account is given in Fan (1991). As an example, if the noise ϵ_i is Gaussian, then Fan shows that the minimax risk satisfies

$$R_n \geq C \left(\frac{1}{\log n}\right)^{\beta}$$

which means that the problem is essentially hopeless.

Similar results hold for nonparametric regression. In the usual nonparametric regression problem we observe $Y_i = m(X_i) + \epsilon_i$ and we want to estimate the function m. If we observe $X_i^* = X_i + \delta_i$ instead of X_i then again the minimax rates change drastically and are logarithmic of the δ_i's are Normal (Fan and Truong 1993). This is known as *measurement error* or *errors in variables.*

This is an interesting example where minimax theory reveals surprising and important insight.

8.8 Normal Means

Perhaps the best understood cases in minimax theory involve normal means. First suppose that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where σ^2 is known. A function g is *bowl-shaped* if the sets $\{x : g(x) \leq c\}$ are convex and symmetric about the origin. We will say that a loss function ℓ is bowl-shaped if $\ell(\theta, \widehat{\theta}) = g(\theta - \widehat{\theta})$ for some bowl-shaped function g.

Theorem 18 *The unique*[2] *estimator that is minimax for every bowl-shaped loss function is the sample mean* $\overline{X}_n$.

For a proof, see Wolfowitz (1950).

Now consider estimating several normal means. Let $X_j = \theta_j + \epsilon_j/\sqrt{n}$ for $j = 1, \ldots, n$ and suppose we and to estimate $\theta = (\theta_1, \ldots, \theta_n)$ with loss function $\ell(\widehat{\theta}, \theta) = \sum_{j=1}^n (\widehat{\theta}_j - \theta_j)^2$. Here, $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. This is called the *normal means problem.*

[2] Up to sets of measure 0.

There are strong connections between the normal means problem and nonparametric learning. For example, suppose we want to estimate a regression function $f(x)$ and we observe data $Z_i = f(i/n) + \delta_i$ where $\delta_i \sim N(0, \sigma^2)$. Expand f in an othonormal basis: $f(x) = \sum_j \theta_j \psi_j(x)$. An estimate of θ_j is $X_j = \frac{1}{n}\sum_{i=1}^n Z_i \, \psi_j(i/n)$. It follows that $X_j \approx N(\theta_j, \sigma^2/n)$. This connection can be made very rigorous; see Brown and Low (1996).

The minimax risk depends on the assumptions about θ.

Theorem 19 (Pinsker) *1. If $\Theta_n = \mathbb{R}^n$ then $R_n = \sigma^2$ and $\widehat{\theta} = X = (X_1, \ldots, X_n)$ is minimax.*

2. If $\Theta_n = \{\theta : \ \sum_j^n \theta_j^2 \leq C^2\}$ then

$$\liminf_{n\to\infty} \inf_{\widehat{\theta}} \sup_{\theta\in\Theta_n} R(\widehat{\theta}, \theta) = \frac{\sigma^2 C^2}{\sigma^2 + C^2}. \tag{43}$$

Define the James-Stein estimator

$$\widehat{\theta}_{\rm JS} = \left(1 - \frac{(n-2)\sigma^2}{\frac{1}{n}\sum_{j=1}^n X_j^2}\right) X. \tag{44}$$

Then

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_n} R(\widehat{\theta}_{\rm JS}, \theta) = \frac{\sigma^2 C^2}{\sigma^2 + C^2}. \tag{45}$$

Hence, $\widehat{\theta}_{\rm JS}$ is asymptotically minimax.

3. Let $X_j = \theta_j + \epsilon_j$ for $j = 1, 2, \ldots,$ where $\epsilon_j \sim N(0, \sigma^2/n)$.

$$\Theta = \left\{\theta : \ \sum_{j=1}^\infty \theta_j^2 a_j^2 \leq C^2\right\} \tag{46}$$

where $a_j^2 = (\pi j)^{2p}$. Let R_n denote the minimax risk. Then

$$\min_{n\to\infty} n^{\frac{2p}{2p+1}} R_n = \left(\frac{\sigma}{\pi}\right)^{\frac{2p}{2p+1}} C^{\frac{2}{2p+1}} \left(\frac{p}{p+1}\right)^{\frac{2p}{2p+1}} (2p+1)^{\frac{1}{2p+1}}. \tag{47}$$

Hence, $R_n \asymp n^{-\frac{2p}{2p+1}}$. An asymptotically minimax estimator is the Pinsker estimator *defined by $\widehat{\theta} = (w_1 X_1, w_2 X_2, \ldots,)$ where $w_j = [1 - (a_j/\mu)]_+$ and μ is determined by the equation*

$$\frac{\sigma^2}{n}\sum_j a_j(\mu - a_j)_+ = C^2.$$

The set Θ in (46) is called a *Sobolev ellipsoid.* This set corresponds to smooth functions in the function estimation problem. The Pinsker estimator corresponds to estimating a function by smoothing. The main message to take away from all of this is that minimax estimation under smoothness assumptions requires shrinking the data appropriately.

9 Adaptation

The results in this chapter provide minimax rates of convergence and estimators that achieve these rates. However, the estimators depend on the assumed parameter space. For example, estimating a β-times differential regression function requires using an estimator tailored to the assumed amount of smoothness to achieve the minimax rate $n^{-\frac{2\beta}{2\beta+1}}$. There are estimators that are *adaptive*, meaning that they achieve the minimax rate without the user having to know the amount of smoothness. See, for example, Chapter 9 of Wasserman (2006) and the references therein.

10 The Bayesian Connection

Another way to find the minimax risk and to find a minimax estimator is to use a carefully constructed Bayes estimator. In this section we assume we have a parametric family of densities $\{p(x;\theta):\ \theta\in\Theta\}$ and that our goal is to estimate the parameter θ. Since the distributions are indexed by θ, we can write the risk as $R(\theta,\widehat{\theta}_n)$ and the maximum risk as $\sup_{\theta\in\Theta} R(\theta,\widehat{\theta}_n)$.

Let Q be a prior distribution for θ. The *Bayes risk (with respect to Q)* is defined to be

$$B_Q(\widehat{\theta}_n) = \int R(\theta,\widehat{\theta}_n)dQ(\theta) \tag{48}$$

and the *Bayes estimator with respect to Q* is the estimator $\overline{\theta}_n$ that minimizes $B_Q(\widehat{\theta}_n)$. For simplicity, assume that Q has a density q. The posterior density is then

$$q(\theta|X^n) = \frac{p(X_1\ldots,X_n;\theta)q(\theta)}{m(X_1,\ldots,X_n)}$$

where $m(x_1,\ldots,x_n)=\int p(x_1,\ldots,x_n;\theta)q(\theta)d\theta$.

Lemma 20 *The Bayes risk can be written as*

$$\int\left(\int L(\theta,\widehat{\theta}_n)q(\theta|x_1,\ldots,x_n)d\theta\right)m(x_1,\ldots,x_n)dx_1\cdots dx_n.$$

It follows from this lemma that the Bayes estimator can be obtained by finding $\widehat{\theta}_n = \widehat{\theta}_n(x_1\ldots,x_n)$ to minimize the inner integral $\int L(\theta,\widehat{\theta}_n)q(\theta|x_1,\ldots,x_n)d\theta$. Often, this is an easy calculation.

Example 21 *Suppose that $L(\theta, \widehat{\theta}_n) = (\theta - \widehat{\theta}_n)^2$. Then the Bayes estimator is the posterior mean $\overline{\theta}_Q = \int \theta \; q(\theta|x_1, \ldots, x_n) d\theta$.*

Now we link Bayes estimators to minimax estimators.

Theorem 22 *Let $\widehat{\theta}_n$ be an estimator. Suppose that (i) the risk function $R(\theta, \widehat{\theta}_n)$ is constant as a function of θ and (ii) $\widehat{\theta}_n$ is the Bayes estimator for some prior Q. Then $\widehat{\theta}_n$ is minimax.*

Proof. We will prove this by contradiction. Suppose that $\widehat{\theta}_n$ is not minimax. Then there is some other estimator θ' such that

$$\sup_{\theta \in \Theta} R(\theta, \theta') < \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}_n). \tag{49}$$

Now,

$$\begin{aligned}
B_Q(\theta') &= \int R(\theta, \theta') dQ(\theta) && \text{definition of Bayes risk} \\
&\leq \sup_{\theta \in \Theta} R(\theta, \theta') && \text{average is less than sup} \\
&< \sup_{\theta \in \Theta} R(\theta, \widehat{\theta}_n) && \text{from (49)} \\
&= \int R(\theta, \widehat{\theta}_n) dQ(\theta) && \text{since risk is constant} \\
&= B_Q(\widehat{\theta}_n) && \text{definition of Bayes risk.}
\end{aligned}$$

So $B_Q(\theta') < B_Q(\widehat{\theta}_n)$. This is a contradiction because $\widehat{\theta}_n$ is the Bayes estimator for Q so it must minimize B_Q. $\square$

Example 23 *Let $X \sim \text{Binomial}(n, \theta)$. The mle is X/n. Let $L(\theta, \widehat{\theta}_n) = (\theta - \widehat{\theta}_n)^2$. Define*

$$\widehat{\theta}_n = \frac{\frac{X}{n} + \sqrt{\frac{1}{4n}}}{1 + \sqrt{\frac{1}{n}}}.$$

Some calculations show that this is the posterior mean under a Beta(α, β) prior with $\alpha = \beta = \sqrt{n/4}$. By computing the bias and variance of $\widehat{\theta}_n$ it can be seen that $R(\theta, \widehat{\theta}_n)$ is constant. Since $\widehat{\theta}_n$ is Bayes and has constant risk, it is minimax.

Example 24 *Let us now show that the sample mean is minimax for the Normal model. Let $X \sim N_p(\theta, I)$ be multivariate Normal with mean vector $\theta = (\theta_1, \ldots, \theta_p)$. We will prove that*

$\widehat{\theta}_n = X$ is minimax when $L(\theta, \widehat{\theta}_n) = \|\widehat{\theta}_n - \theta\|^2$. Assign the prior $Q = N(0, c^2 I)$. Then the posterior is

$$N\left(\frac{c^2 x}{1+c^2}, \frac{c^2}{1+c^2} I\right). \tag{50}$$

The Bayes risk $B_Q(\widehat{\theta}_n) = \int R(\theta, \widehat{\theta}_n) dQ(\theta)$ is minimized by the posterior mean $\widetilde{\theta} = c^2 X/(1+c^2)$. Direct computation shows that $B_Q(\widetilde{\theta}) = pc^2/(1+c^2)$. Hence, if θ^* is any estimator, then

$$\frac{pc^2}{1+c^2} = B_Q(\widetilde{\theta}) \leq B_Q(\theta^*) = \int R(\theta^*, \theta) dQ(\theta) \leq \sup_\theta R(\theta^*, \theta).$$

This shows that $R(\Theta) \geq pc^2/(1+c^2)$ for every $c > 0$ and hence

$$R(\Theta) \geq p. \tag{51}$$

But the risk of $\widehat{\theta}_n = X$ is p. So, $\widehat{\theta}_n = X$ is minimax.

11 Nonparametric Maximum Likelihood and the Le Cam Equation

In some cases, the minimax rate can be found by finding ϵ to solve the equation

$$H(\epsilon_n) = n\epsilon_n^2$$

where $H(\epsilon) = \log N(\epsilon)$ and $N(\epsilon)$ is the smallest number of balls of size ϵ in the Hellinger metric needed to cover $\mathcal{P}$. $H(\epsilon)$ is called the Hellinger entropy of $\mathcal{P}$. The equation $H(\epsilon) = n\epsilon^2$ is known as the *Le Cam equation*. In this section we consider one case where this is true. For more general versions of this argument, see Shen and Wong (1995), Barron and Yang (1999) and Birgé and Massart (1993).

Our goal is to estimate the density function using maximum likelihood. The loss function is Hellinger distance. Let $\mathcal{P}$ be a set of probability density functions. We have in mind the nonparametric situation where $\mathcal{P}$ does not correspond to some finite dimensional parametric family. Let $N(\epsilon)$ denote the Hellinger covering number of $\mathcal{P}$. We will make the following assumptions:

(A1) We assume that there exist $0 < c_1 < c_2 < \infty$ such that $c_1 \leq p(x) \leq c_2$ for all x and all $p \in \mathcal{P}$.

(A2) We assume that there exists $a > 0$ such that

$$H(a\epsilon, \mathcal{P}, h) \leq \sup_{p \in \mathcal{P}} H(\epsilon, B(p, 4\epsilon), h)$$

where $B(p,\delta) = \{q:\ h(p,q) \leq \delta\}$.

(A3) We assume $\sqrt{n}\epsilon_n \to \infty$ as $n \to \infty$ where $H(\epsilon_n) \asymp n\epsilon_n^2$.

Assumption (A1) is a very strong and is made only to make the proofs simpler. Assumption (A2) says that the local entropy and global entropy are of the same order. This is typically true in nonparametric models. Assumption (A3) says that the rate of convergence is slower than $O(1/\sqrt{n})$ which again is typical of nonparametric problems. An example of a class $\mathcal{P}$ that satisfies these conditions is

$$\mathcal{P} = \left\{ p : [0,1] \to [c_1, c_2] :\ \int_0^1 p(x)dx = 1,\ \ \int_0^1 (p''(x))^2 dx \leq C^2 \right\}.$$

Thanks to (A1) we have,

$$\begin{aligned} \mathsf{KL}(p,q) &\leq \chi^2(p,q) = \int \frac{(p-q)^2}{p} \leq \frac{1}{c_1}\int (p-q)^2 \\ &= \frac{1}{c_1}\int (\sqrt{p}-\sqrt{q})^2(\sqrt{p}+\sqrt{q})^2 \\ &\leq \frac{4c_2}{c_1}\int (\sqrt{p}-\sqrt{q})^2 = Ch^2(p,q) \end{aligned} \tag{52}$$

where $C = 4c_2/c_1$.

Let ϵ_n solve the Le Cam equation. More precisely, let

$$\epsilon_n = \min\left\{ \epsilon :\ H\left(\frac{\epsilon}{\sqrt{2C}}\right) \leq \frac{n\epsilon^2}{16C} \right\}. \tag{53}$$

We will show that ϵ_n is the minimax rate.

Upper Bound. To show the upper bound, we will find an estimator that achieves the rate. Let $\mathcal{P}_n = \{p_1, \ldots, p_N\}$ be an $\epsilon_n/\sqrt{2C}$ covering set where $N = N(\epsilon_n/\sqrt{2C})$. The set $\mathcal{P}_n$ is an approximation to $\mathcal{P}$ that grows with sample size n. Such a set is called *sieve.* Let $\widehat{p}$ be the mle over $\mathcal{P}_n$, that is, $\widehat{p} = \text{argmax}_{p\in\mathcal{P}_n}\mathcal{L}(p)$ where $\mathcal{L}(p) = \prod_{i=1}^n p(X_i)$ is the likelihood function. We call $\widehat{p}$, a *sieve maximum likelihood estimator*. It is crucial that the estimator is computed over $\mathcal{P}_n$ rather than over $\mathcal{P}$ to prevent overfitting. Using a sieve is a type of regularization. We need the following lemma.

Lemma 25 (Wong and Shen) *Let p_0 and p be two densities and let $\delta = h(p_0, p)$. Let $Z_1, \ldots, Z_n$ be a sample from p_0. Then*

$$\mathbb{P}\left(\frac{\mathcal{L}(p)}{\mathcal{L}(p_0)} > e^{-n\delta^2/2}\right) \leq e^{-n\delta^2/4}.$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\frac{\mathcal{L}(p)}{\mathcal{L}(p_0)} > e^{-n\delta^2/2}\right) &= \mathbb{P}\left(\prod_{i=1}^n \sqrt{\frac{p(Z_i)}{p_0(Z_i)}} > e^{-n\delta^2/4}\right) \leq e^{n\delta^2/4}\mathbb{E}\left(\prod_{i=1}^n \sqrt{\frac{p(Z_i)}{p_0(Z_i)}}\right)\\
&= e^{n\delta^2/4}\left(\mathbb{E}\left(\sqrt{\frac{p(Z_i)}{p_0(Z_i)}}\right)\right)^n = e^{n\delta^2/4}\left(\int \sqrt{p_0\, p}\right)^n\\
&= e^{n\delta^2/4}\left(1 - \frac{h^2(p_0,p)}{2}\right)^n = e^{n\delta^2/4}\exp\left(n\log\left(1 - \frac{h^2(p_0,p)}{2}\right)\right)\\
&\leq e^{n\delta^2/4}e^{-nh^2(p_0,p)/2} = e^{-n\delta^2/4}.
\end{aligned}$$

□

In what follows, we use $c, c_1, c_2, \ldots,$ to denote various positive constants.

Theorem 26 $\sup_{P\in\mathcal{P}} \mathbb{E}_p(h(p,\widehat{p})) = O(\epsilon_n)$.

Proof. Let p_0 denote the true density. Let p_* be the element of $\mathcal{P}_n$ that minimizes $\mathsf{KL}(p_0, p_j)$. Hence, $\mathsf{KL}(p_0, p_*) \leq Cd^2(p_0, p_*) \leq C(\epsilon_n^2/(2C)) = \epsilon_n^2/2$. Let

$$B = \{p \in \mathcal{P}_n : \ d(p_*, p) > A\epsilon_n\}$$

where $A = 1/\sqrt{2C}$. Then

$$\begin{aligned}
\mathbb{P}(h(\widehat{p}, p_0) > D\epsilon_n) &\leq \mathbb{P}(h(\widehat{p}, p_*) + h(p_0, p_*) > D\epsilon_n) \leq \mathbb{P}(h(\widehat{p}, p_*) + \frac{\epsilon}{\sqrt{2C}} > D\epsilon_n)\\
&= \mathbb{P}(h(\widehat{p}, p_*) > A\epsilon_n) = \mathbb{P}(\widehat{p} \in B) \leq \mathbb{P}\left(\sup_{p\in B}\frac{\mathcal{L}(p)}{\mathcal{L}(p_*)} > 1\right)\\
&\leq \mathbb{P}\left(\sup_{p\in B}\frac{\mathcal{L}(p)}{\mathcal{L}(p_*)} > e^{-n\epsilon_n^2(A^2/2+1)}\right)\\
&\leq \mathbb{P}\left(\sup_{p\in B}\frac{\mathcal{L}(p)}{\mathcal{L}(p_*)} > e^{-n\epsilon_n^2 A^2/2}\right) + \mathbb{P}\left(\sup_{p\in B}\frac{\mathcal{L}(p_0)}{\mathcal{L}(p_*)} > e^{n\epsilon_n^2}\right)\\
&\equiv P_1 + P_2.
\end{aligned}$$

Now

$$P_1 \leq \sum_{p\in B}\mathbb{P}\left(\frac{\mathcal{L}(p)}{\mathcal{L}(p_*)} > e^{-n\epsilon_n^2 A^2/2}\right) \leq N(\epsilon/\sqrt{2C})e^{-n\epsilon_n^2 A^2/4} \leq e^{n\epsilon_n^2/(16C)}$$

where we used Lemma 25 and the definition of ϵ_n. To bound P_2, define $K_n = \frac{1}{n}\sum_{i=1}^n \log \frac{p_0(Z_i)}{p_*(Z_i)}$. Hence, $\mathbb{E}(K_n) = \mathsf{KL}(p_0, p_*) \leq \epsilon^2/2$. Also,

$$\begin{aligned} \sigma^2 &\equiv \mathsf{Var}\left(\log \frac{p_0(Z)}{p_*(Z)}\right) \leq \mathbb{E}\left(\log \frac{p_0(Z)}{p_*(Z)}\right)^2 \leq \log\left(\frac{c_2}{c_1}\right)\mathbb{E}\left(\log \frac{p_0(Z)}{p_*(Z)}\right) \\ &= \log\left(\frac{c_2}{c_1}\right)\mathsf{KL}(p_0, p_*) \leq \log\left(\frac{c_2}{c_1}\right)\frac{\epsilon^2}{2} \equiv c_3\epsilon_n^2 \end{aligned}$$

where we used (52). So, by Bernstein's inequality,

$$\begin{aligned} P_2 &= \mathbb{P}(K_n > \epsilon_n^2) = \mathbb{P}(K_n - \mathsf{KL}(p_0, p_*) > \epsilon_n^2 - \mathsf{KL}(p_0, p_*)) \\ &\leq \mathbb{P}\left(K_n - \mathsf{KL}(p_0, p_*) > \frac{\epsilon_n^2}{2}\right) \leq 2\exp\left(-\frac{n\epsilon_n^4}{8\sigma^2 + c_4\epsilon_n^2}\right) \\ &\leq 2\exp\left(-c_5 n\epsilon_n^2\right). \end{aligned}$$

Thus, $P_1 + P_2 \leq \exp\left(-c_6 n\epsilon_n^2\right)$. Now,

$$\begin{aligned} \mathbb{E}(h(\widehat{p}, p_0)) &= \int_0^{\sqrt{2}} \mathbb{P}(h(\widehat{p}, p_0) > t)dt \\ &= \int_0^{D\epsilon_n} \mathbb{P}(h(\widehat{p}, p_0) > t)dt + \int_{D\epsilon_n}^{\sqrt{2}} \mathbb{P}(h(\widehat{p}, p_0) > t)dt \\ &\leq D\epsilon_n + \exp\left(-c_6 n\epsilon_n^2\right) \leq c_7\epsilon_n. \end{aligned}$$

□

Lower Bound. Now we derive the lower bound.

Theorem 27 *Let ϵ_n be the smallest ϵ such that $H(a\epsilon) \geq 64C^2 n\epsilon^2$. Then*

$$\inf_{\widehat{p}} \sup_{P\in\mathcal{P}} \mathbb{E}_p(h(p, \widehat{p})) = \Omega(\epsilon_n).$$

Proof. Pick any $p \in \mathcal{P}$. Let $B = \{q : \; h(p, q) \leq 4\epsilon_n\}$. Let $F = \{p_1, \ldots, p_N\}$ be an ϵ_n packing set for B. Then

$$N = \log P(\epsilon_n, B, h) \geq \log H(\epsilon_n, B, h) \geq \log H(a\epsilon_n) \geq 64C^2 n\epsilon^2.$$

Hence, for any $P_j, P_k \in F$,

$$\mathsf{KL}(P_j^n, P_k^n) = n\mathsf{KL}(P_j, P_k) \leq Cnh^2(P_j, P_k) \leq 16Cn\epsilon_n^2 \leq \frac{N}{4}.$$

It follows from Fano's inequality that

$$\inf_{\widehat{p}} \sup_{p\in\mathcal{P}} \mathbb{E}_p h(p,\widehat{p}) \geq \frac{1}{4}\min_{j\neq k} h(p_j,p_k) \geq \frac{\epsilon_n}{4}$$

as claimed. □

In summary, we get the minimax rate by solving

$$H(\epsilon_n) \asymp n\epsilon_n^2.$$

Now we can use the Le Cam equation to compute some rates:

Example 28 *Here are some standard examples:*

Space	*Entropy*	*Rate*
Sobolev α	$\epsilon^{-1/\alpha}$	$n^{-\alpha/(2\alpha+1)}$
Sobolev α *dimension* d	$\epsilon^{-d/\alpha}$	$n^{-\alpha/(2\alpha+d)}$
Lipschitz α	$\epsilon^{-d/\alpha}$	$n^{-\alpha/(2\alpha+d)}$
Monotone	$1/\epsilon$	$n^{-1/3}$
Besov $B^{\alpha}_{p,q}$	$\epsilon^{-d/\alpha}$	$n^{-\alpha/(2\alpha+d)}$
Neural Nets	*see below*	*see below*
m*-dimensional parametric model*	$m\log(1/\epsilon)$	(m/n)

In the neural net case we have $f(x) = c_0 + \sum_i c_i\sigma(v_i^T x + b_i)$ *where* $\|c\|_1 \leq C$, $\|v_i\| = 1$ *and* σ *is a step function or a Lipschitz sigmoidal function. Then*

$$\left(\frac{1}{\epsilon}\right)^{1/2+1/d} \leq H(\epsilon) \leq \left(\frac{1}{\epsilon}\right)^{1/2+1/(2d)} \log(1/\epsilon) \tag{54}$$

and hence

$$n^{-(1+2/d)/(2+1/d)}(\log n)^{-(1+2/d)(1+1/d)/(2+1/d)} \leq \epsilon_n \leq (n/\log n)^{-(1+1/d)/(2+1/d)}. \tag{55}$$

12 Summary

Minimax theory allows us to state precisely the best possible performance of any procedure under given conditions. The key tool for finding lower bounds on the minimax risk is Fano's inequality. Finding an upper bound usually involves finding a specific estimator and computing its risk.

13 Bibliographic remarks

There is a vast literature on minimax theory however much of it is scattered in various journal articles. Some texts that contain minimax theory include Tsybakov (2009), van de Geer (2000), van der Vaart (1998) and Wasserman (2006).

14 Appendix

14.1 Metrics For Probability Distributions

Minimax theory often makes use of various metrics for probability distributions. Here we summarize some of these metrics and their properties.

Let P and Q be two distributions with densities p and q. We write the distance between P and Q as either $d(P, Q)$ or $d(p, q)$ whichever is convenient. We define the following distances and related quantities.

Total variation	$\mathsf{TV}(P, Q) = \sup_A \lvert P(A) - Q(A)\rvert$
L_1 distance	$d_1(P, Q) = \int \lvert p - q\rvert$
L_2 distance	$d_2(P, Q) = \sqrt{\int \lvert p - q\rvert^2}$
Hellinger distance	$h(P, Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$
Kullback-Leibler distance	$\mathsf{KL}(P, Q) = \int p \log(p/q)$
χ^2	$\chi^2(P, Q) = \int (p - q)^2/p$
Affinity	$\lVert P \wedge Q\rVert = \int p \wedge q = \int \min\{p(x), q(x)\}dx$
Hellinger affinity	$A(P, Q) = \int \sqrt{pq}$

There are many relationships between these quantities. These are summarized in the next two theorems. We leave the proofs as exercises.

Theorem 29 *The following relationships hold:*

1. $\mathsf{TV}(P, Q) = \frac{1}{2}d_1(P, Q) = 1 - \lVert p \wedge q\rVert$. *(Scheffés Theorem.)*
2. $\mathsf{TV}(P, Q) = P(A) - Q(A)$ *where* $A = \{x :\ p(x) > q(x)\}$.
3. $0 \leq h(P, Q) \leq \sqrt{2}$.
4. $h^2(P, Q) = 2(1 - A(P, Q))$.
5. $\lVert P \wedge Q\rVert = 1 - \frac{1}{2}d_1(P, Q)$.
6. $\lVert P \wedge Q\rVert \geq \frac{1}{2}A^2(P, Q) = \frac{1}{2}\left(1 - \frac{h^2(P,Q)}{2}\right)^2$. *(Le Cam's inequalities.)*

7. $\frac{1}{2}h^2(P,Q) \le \mathsf{TV}(P,Q) = \frac{1}{2}d_1(P,Q) \le h(P,Q)\sqrt{1 - \frac{h^2(P,Q)}{4}}$.
8. $\mathsf{TV}(P,Q) \le \sqrt{\mathsf{KL}(P,Q)/2}$. *(Pinsker's inequality.)*
9. $\int (\log dP/dQ)_+ dP \le \mathsf{KL}(P,Q) + \sqrt{\mathsf{KL}(P,Q)/2}$.
10. $\|P \wedge Q\| \ge \frac{1}{2}e^{-\mathsf{KL}(P,Q)}$.
11. $\mathsf{TV}(P,Q) \le h(P,Q) \le \sqrt{\mathsf{KL}(P,Q)} \le \sqrt{\chi^2(P,Q)}$.

Let P^n denote the product measure based on n independent samples from P.

Theorem 30 *The following relationships hold:*

1. $h^2(P^n,Q^n) = 2\left(1 - \left(1 - \frac{h^2(P,Q)}{2}\right)^n\right)$.
2. $\|P^n \wedge Q^n\| \ge \frac{1}{2}A^2(P^n,Q^n) = \frac{1}{2}\left(1 - \frac{1}{2}h^2(P,Q)\right)^{2n}$.
3. $\|P^n \wedge Q^n\| \ge \left(1 - \frac{1}{2}d_1(P,Q)\right)^n$.
4. $\mathsf{KL}(P^n,Q^n) = n\mathsf{KL}(P,Q)$.

14.2 Fano's Lemma

For $0 < p < 1$ define the entropy $h(p) = -p\log p - (1-p)\log(1-p)$ and note that $0 \le h(p) \le \log 2$. Let (Y,Z) be a pair of random variables each taking values in $\{1,\ldots,N\}$ with joint distribution $P_{Y,Z}$. Then the mutual information is defined to be

$$I(Y;Z) = \mathsf{KL}(P_{Y,Z}, P_Y \times P_Z) = H(Y) - H(Y|Z) \tag{56}$$

where $H(Y) = -\sum_j \mathbb{P}(Y=j)\log\mathbb{P}(Y=j)$ is the entropy of Y and $H(Y|Z)$ is the entropy of Y given Z. We will use the fact that $I(Y;h(Z)) \le I(Y;Z)$ for an function h.

Lemma 31 *Let Y be a random variable taking values in $\{1,\ldots,N\}$. Let $\{P_1,\ldots,P_N\}$ be a set of distributions. Let X be drawn from P_j for some $j \in \{1,\ldots,N\}$. Thus $P(X \in A|Y=j) = P_j(A)$. Let $Z = g(X)$ be an estimate of Y taking values in $\{1,\ldots,N\}$. Then,*

$$H(Y|X) \le \mathbb{P}(Z \ne Y)\log(N-1) + h(\mathbb{P}(Z=Y)). \tag{57}$$

We follow the proof from Cover and Thomas (1991).

Proof. Let $E = I(Z \ne Y)$. Then

$$H(E,Y|X) = H(Y|X) + H(E|X,Y) = H(Y|X)$$

since $H(E|X,Y)=0$. Also,

$$H(E,Y|X)=H(E|X)+H(Y|E,X).$$

But $H(E|X)\le H(E)=h(P(Z=Y))$. Also,

$$\begin{aligned}H(Y|E,X) &= P(E=0)H(Y|X,E=0)+P(E=1)H(Y|X,E=1)\\ &\le P(E=0)\times 0+h(P(Z=Y))\log(N-1)\end{aligned}$$

since $Y=g(X)$ when $E=0$ and, when $E=1$, $H(Y|X,E=1)$ by the log number of remaining outcomes. Combining these gives $H(Y|X)\le \mathbb{P}(Z\ne Y)\log(N-1)+h(\mathbb{P}(Z=Y))$. □

Lemma 32 *(Fano's Inequality) Let $\mathcal{P}=\{P_1,\ldots,P_N\}$ and $\beta=\max_{j\ne k}\mathsf{KL}(P_j,P_k)$. For any random variable Z taking values on $\{1,\ldots,N\}$,*

$$\frac{1}{N}\sum_{j=1}^{N}P_j(Z\ne j)\ge\left(1-\frac{n\beta+\log 2}{\log N}\right).$$

Proof. For simplicity, assume that $n=1$. The general case follows since $\mathsf{KL}(P^n,Q^n)=n\mathsf{KL}(P,Q)$. Let Y have a uniform distribution on $\{1,\ldots,N\}$. Given $Y=j$, let X have distribution P_j. This defines a joint distribution P for (X,Y) given by

$$P(X\in A,Y=j)=P(X\in A|Y=j)P(Y=j)=\frac{1}{N}P_j(A).$$

Hence,

$$\frac{1}{N}\sum_{j=1}^{N}P(Z\ne j|Y=j)=P(Z\ne Y).$$

From (57),

$$\begin{aligned}H(Y|Z) &\le P(Z\ne Y)\log(N-1)+h(P(Z=Y))\le P(Z\ne Y)\log(N-1)+h(1/2)\\ &= P(Z\ne Y)\log(N-1)+\log 2.\end{aligned}$$

Therefore,

$$\begin{aligned}P(Z\ne Y)\log(N-1) &\ge H(Y|Z)-\log 2=H(Y)-I(Y;Z)-\log 2\\ &= \log N-I(Y;Z)-\log 2\ge\log N-\beta-\log 2. \qquad (58)\end{aligned}$$

The last inequality follows since

$$I(Y;Z)\le I(Y;X)=\frac{1}{N}\sum_{j=1}^{N}\mathsf{KL}(P_j,\overline{P})\le\frac{1}{N^2}\sum_{j,k}^{N}\mathsf{KL}(P_j,P_k)\le\beta \qquad (59)$$

where $\overline{P} = N^{-1}\sum_{j=1}^{N} P_j$ and we used the convexity of K. Equation (58) shows that

$$P(Z \neq Y)\log(N-1) \geq \log N - \beta - \log 2$$

and the result follows. □

14.3 Assouad's Lemma

Assouad's Lemma is another way to get a lower bound using hypercubes. Let

$$\Omega = \Big\{\omega = (\omega_1, \ldots, \omega_N) :\ \omega_j \in \{0,1\}\Big\}$$

be the set of binary sequences of length N. Let $\mathcal{P} = \{P_\omega :\ \omega \in \Omega\}$ be a set of 2^N distributions indexed by the elements of Ω. Let $h(\omega,\nu) = \sum_{j=1}^{N} I(\omega_j \neq \nu_j)$ be the *Hamming distance* between $\omega, \nu \in \Omega$.

Lemma 33 *Let $\{P_\omega :\ \omega \in \Omega\}$ be a set of distributions indexed by ω and let $\theta(P)$ be a parameter. For any $p > 0$ and any metric d,*

$$\max_{\omega\in\Omega} E_\omega\left(d^p(\widehat{\theta}, \theta(P_\omega))\right) \geq \frac{N}{2^{p+1}}\left(\min_{\substack{\omega,\nu\\ h(\omega,\nu)\neq 0}} \frac{d^p(\theta(P_\omega), \theta(P_\nu))}{h(\omega,\nu)}\right)\left(\min_{\substack{\omega,\nu\\ h(\omega,\nu)=1}} \|P_\omega \wedge P_\nu\|\right). \quad (60)$$

For a proof, see van der Vaart (1998) or Tsybakov (2009).