Minimax

김지수 (Jisu KIM)

인공지능을 위한 이론과 모델링, 2023 가을학기

The lecture note is a minor modification of the lecture notes from Prof. Larry Wasserman's "Statistical Machine Learning".

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.

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 $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ denote an estimator. Given a metric d, the *minimax risk* is

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

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

$$n(\epsilon, \mathcal{P}) = \min\Big\{ n : \ R_n(\mathcal{P}) \leq \epsilon \Big\}.$$

Example. 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

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[(\hat{\theta} - \theta)^2]. \tag{1}$$

In this example, θ is a scalar.

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

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

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 $a_n \geq C b_n$ 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 $\hat{\theta}$ be any estimator. Then

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \leq \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \equiv U_n. \tag{3}$$

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

$$R_n = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \inf_{\hat{\theta}} \max_{P_j \in M} \mathbb{E}_j[d(\hat{\theta}, \theta_j)] \tag{4}$$

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,

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

and so

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

Setting $t = s/2$, and using (4), we have

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

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

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

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

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

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

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

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

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

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 Distances

We will need some distances between distributions. Specifically,

Total Variation	$TV(P,Q)$	$=$	$\sup_A \lvert P(A)-Q(A)\rvert$
L_1	$\lVert P-Q\rVert_1$	$=$	$\int \lvert p-q\rvert$
Kullback-Leibler	$KL(P,Q)$	$=$	$\int p\log(p/q)$
χ^2	$\chi^2(P,Q)$	$=$	$\int \left(\frac{p}{q}-1\right)^2 dQ = \int \frac{p^2}{q}-1$
Hellinger	$H(P,Q)$	$=$	$\sqrt{\int(\sqrt{p}-\sqrt{q})^2}$.

We also define the *affinity* between P and Q by

$$a(P,Q) = \int (p \wedge q).$$

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

Theorem. *The following relationships hold:*

1. $\mathsf{TV}(P,Q) = \frac{1}{2}\lVert P-Q\rVert_1 = 1-a(P,Q)$. *(Scheffés Theorem.)*
2. $\mathsf{TV}(P,Q) = P(A)-Q(A)$ *where* $A = \{x:\ p(x) > q(x)\}$.
3. $0 \le H(P,Q) \le \sqrt{2}$.
4. $H^2(P,Q) = 2(1-a(P,Q))$.
5. $a(P,Q) \ge \frac{1}{2}a^2(P,Q) = \frac{1}{2}\left(1-\frac{H^2(P,Q)}{2}\right)^2$. *(Le Cam's inequalities.)*
6. $\frac{1}{2}H^2(P,Q) \le \mathsf{TV}(P,Q) = \frac{1}{2}\lVert P-Q\rVert_1 \le H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}}$.
7. $\mathsf{TV}(P,Q) \le \sqrt{\mathsf{KL}(P,Q)/2}$. *(Pinsker's inequality.)*
8. $\int(\log dP/dQ)_+ dP \le \mathsf{KL}(P,Q) + \sqrt{\mathsf{KL}(P,Q)/2}$.
9. $a(P,Q) \ge \frac{1}{2}e^{-\mathsf{KL}(P,Q)}$.
10. $\mathsf{TV}(P,Q) \le H(P,Q) \le \sqrt{\mathsf{KL}(P,Q)} \le \sqrt{\chi^2(P,Q)}$.

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

Theorem. *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. $a(P^n,Q^n) \ge \frac{1}{2}a^2(P^n,Q^n) = \frac{1}{2}\left(1-\frac{1}{2}H^2(P,Q)\right)^{2n}$.
3. $a(P^n,Q^n) \ge \left(1-\frac{1}{2}\lVert P-Q\rVert_1\right)^n$.
4. $\mathsf{KL}(P^n,Q^n) = n\mathsf{KL}(P,Q)$.

5 Lower Bound Method 1: Le Cam

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

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{s}{4} \int [p_0^n(x) \wedge p_1^n(x)] dx = \frac{s}{4}\Big[1 - \mathrm{TV}(P_0^n, P_1^n)\Big]$$

where $s = d(\theta(P_0), \theta(P_1))$. We also have:

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{s}{8} e^{-n\mathsf{KL}(P_0, P_1)} \geq \frac{s}{8} e^{-n\chi^2(P_0, P_1)}$$

and

$$\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \geq \frac{s}{8}\left(1 - \frac{1}{2}\int |p_0 - p_1|\right)^{2n}.$$

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

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

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

Proof. Let $\theta_0 = \theta(P_0)$, $\theta_1 = \theta(P_1)$ and $s = d(\theta_0, \theta_1)$. First suppose that $n = 1$ so that we have a single observation X. From Theorem 3,

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

where

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

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

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

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

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

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

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

In Lemma 5 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

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

The other results follow from the inequalities on the distances. □

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

Proof. First note that, since $(a \vee b) + (a \wedge b) = a + b$, we have

$$\int (p \vee q) + \int (p \wedge q) = 2. \tag{5}$$

Hence

$$\begin{aligned}
2\int p \wedge q &\geq 2\int p \wedge q - \left(\int p \wedge q\right)^2 = \left(\int p \wedge q\right) \left[2 - \int p \wedge q\right] \\
&= \left(\int p \wedge q\right) \left(\int p \vee q\right) \text{ from (5)} \\
&\geq \left(\int \sqrt{(p \wedge q)\ (p \vee q)}\right)^2 \quad \text{Cauchy} - \text{Schwartz} \\
&= \left(\int \sqrt{pq}\right)^2 = \exp\left(2 \log \int \sqrt{pq}\right) \\
&= \exp\left(2 \log \int p\sqrt{q/p}\right) \geq \exp\left(2 \int p \log \sqrt{\frac{q}{p}}\right) = e^{-\mathsf{KL}(P,Q)}
\end{aligned}$$

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

Example. 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 $s = |m_1(0) - m_0(0)| = L\epsilon$. 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,

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

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

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

Thus we have proved that

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

The same calculations in d dimensions yield

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

On the squared scale we have

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

Similar rates hold in density estimation.

There is a more general version of Le Cam's lemma that is sometimes useful.

Lemma. *Let $P, Q_1, \ldots, Q_N$ be distributions such that $d(\theta(P), \theta(Q_j)) \geq s$ for all j. Then*

$$\inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))] \geq \frac{s}{4}\int (p^n \wedge q^n)$$

where $q = \frac{1}{N}\sum_j q_j$.

Example. Let

$$Y_i = \theta_i + \frac{1}{\sqrt{n}} Z_i, \quad i = 1, \ldots, d$$

where $Z_1, Z_2, \ldots, Z_d \sim N(0,1)$ and $\theta = (\theta_1, \ldots, \theta_d) \in \Theta$ where $\Theta = \{\theta \in \mathbb{R}^d : \ ||\theta||_0 \leq 1\}$. Let $P = N(0, n^{-1}I)$. Let Q_j have mean 0 expect that j^{th} coordinate has mean $\sqrt{a\log d/n}$ where $0 < a < 1$. Let $q = \frac{1}{N}\sum_j q_j$. Some algebra (good homework question!) shows that $\chi^2(q,p) \to 0$ as $d \to \infty$. By the generalized Le Cam lemma, $R_n \geq a\log d/n$ using squared error loss. We can estimate θ by thresholding (Bonferroni). This gives a matching upper bound.

6 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 (Fano)**.** *Let $X_1, \ldots, X_n \sim P$ where $P \in \{P_1, \ldots, P_N\}$. Let ψ 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*

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

Now we can state and prove the Fano minimax bound.

Theorem. *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*

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

where

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

and

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

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

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

Then

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

Proof. From Theorem 3,

$$R_n \geq \frac{s}{2} \inf_{\psi} \max_{P_j \in F} P_j(\psi \neq j) \geq \frac{s}{2}\frac{1}{N}\sum_{j=1}^{N} P_j(\psi \neq j),$$

where the latter is due to the fact that a max is larger than an average. By Fano's lemma,

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

Thus,

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

□

7 Lower Bound Method III: Tsybakov's Bound

This approach is due to Tsybakov (2009).

Theorem (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

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

where

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

8 Hypercubes

To use Fano's method or Tsybakov'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 $s = \min_{j \neq k} d\left(\theta(P_j), \theta(P_k)\right)$ 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 (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. 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 $\omega, \nu \in \Omega'$,

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

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 6, we need to have

$$\mathsf{KL}(P_\omega, P_\nu) \leq \frac{\log N}{4n}.$$

Now

$$\frac{\log N}{4n} = \frac{\log 2^{m/8}}{4n} = \frac{m}{32n} = \frac{1}{32nh}.$$

So 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 6 implies that

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

It follows that

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

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

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

9 Further Examples

9.1 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 $\hat{\theta}$ roughly equals the variance:[1]

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

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

$$nR(\theta, \hat{\theta}) \approx \frac{1}{I(\theta)}.$$

For any other estimator θ', it can be shown that for large n, $R(\theta, \theta') \geq R(\theta, \hat{\theta})$. For d-dimensional vectors we have $R(\theta, \hat{\theta}) \approx |I(\theta)|^{-1}/n = O(d/n)$.

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

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \ell'_\theta \sqrt{p_\theta} \right)^2 d\mu = o(\|h\|^2). \tag{6}$$

Theorem (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 $\psi(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the mle, is asymptotically, locally, uniformly minimax in the sense that, for any estimator T_n, and any bowl-shaped ℓ,*

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

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

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

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.

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

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

for some C. But

$$\int (\sqrt{f(x)} - \sqrt{g(x)})^2 dx = \int \left(\frac{f(x) - g(x)}{\sqrt{f(x)} + \sqrt{g(x)}} \right)^2 dx \leq C' \int (f(x) - g(x))^2 dx, \tag{7}$$

for some C'. Hence $\sup_f \mathbb{E}_f(d^2(f, \hat{f}_n)) \leq C n^{-4/5}$ 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',

$$\mathsf{KL}(f_\tau^n, f_{\tau'}^n) = n \mathsf{KL}(f_\tau, f_{\tau'}) \leq \frac{C_1 n}{m^4} \equiv \beta. \tag{8}$$

By Lemma 8, 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

$$d^2(f_\tau, f_{\tau'}) \geq \frac{C_2}{m^4} \equiv \alpha \tag{9}$$

for all pairs in F. Choosing $m = c n^{1/5}$ 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

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

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. *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

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

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

9.3 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

$$R(G) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y \neq I(X \in G)) = \mathbb{E}(Y - I(X \in G))^2. \tag{10}$$

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

$$d(G, G^*) = R(G) - R(G^*) = 2 \int_{G \Delta G^*} \left| m(x) - \frac{1}{2} \right| dP_X(x) \tag{11}$$

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 $\hat{h}(x) = I(\hat{m}(x) \geq 1/2)$ where $\hat{m}$ is an estimate of the regression function. The second type are *empirical risk minimizers* where $\hat{h}$ 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 **??**. 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 $\hat{m}$ and corresponding $\hat{G}$

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

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

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

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

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

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

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

we get

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

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.

9.4 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 $\|\hat{\Sigma} - \Sigma\|$ 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.

9.5 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.

9.6 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 $\lambda \in [0, 1/2]$,

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

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

9.7 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

$$Y_i = X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

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.

9.8 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 $\ell(\theta, \hat{\theta}) = g(\theta - \hat{\theta})$ for some bowl-shaped function g.

Theorem. *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 $\ell(\hat{\theta}, \theta) = \sum_{j=1}^n (\hat{\theta}_j - \theta_j)^2$. Here, $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$. This is called the *normal means problem*.

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

[2]Up to sets of measure 0.

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

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

Define the James-Stein estimator

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

Then

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

Hence, $\hat{\theta}_{\text{JS}}$ *is asymptotically minimax.*

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

$$\Theta = \left\{\theta : \sum_{j=1}^{\infty} \theta_j^2 a_j^2 \leq C^2\right\} \tag{12}$$

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

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

Hence, $R_n \asymp n^{-\frac{2p}{2p+1}}$. *An asymptotically minimax estimator is the* Pinsker estimator *defined by* $\hat{\theta} = (w_1 X_1, w_2 X_2, \ldots,)$ *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 (12) 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.

10 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.

11 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.