Preliminary Notes

MINIMAX THEORY

Ariel Boyarsky*†

1. preliminaries

Notation: Suppose $(\Omega, \mathcal{F}, P)$ is a probability space. Let $\mathcal{P}$ be a set of distributions on the probability space and let $X_1, \ldots, X_n$ be a sample from distribution $P \in \mathcal{P}$. Let $\theta(P)$ be some function of P, we will sometimes refer to this "true" parameter as θ_0. For instance this code be the mean of P or some other population parameter. We denote an estimator as $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. We assume that there exists a metric $d(\cdot, \cdot)$ that satisfies the triangle inequality on the space of distributions. Furthermore note that we sometimes use $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. If P is a distribution then p is it's density. The product distribution is given by P^n with density $p^n = \prod_{i=1}^n p(x_i)$. We use the following stochastic convergence notation, $a_n \asymp b_n$ denotes that $0 < a_n/b_n < \infty$ for all large n. Furthermore $a_n = \Omega(b_n)$ means that there exists $C > 0$ such that $a_n \geq C b_n$.

Definition 1 Given a metric d that satisfies triangle inequality we say that the **minimax risk** of an estimator $\hat{\theta}$ 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 taken over all estimators.

Definition 2 We say that the **sample complexity** is,

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

Definition 3 We define the **Kullback-Leibler** distance (divergence, since it's between two distributions) between two distributions P_0 and P_1 with densities p_0 and p_1 as,

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

Example 1 Suppose that $\mathcal{P} = \{N(\theta, 1) : \theta \in \mathbb{R}\}$. Then let $d(a, b) = (a - b)^2$. Then the minimax risk is,

$$R_n = \inf_{\hat{\theta}} \sup_{P} \mathbb{E}_P\left[\left(\hat{\theta} - \theta\right)^2\right]$$

and in fact the sample mean achieves this rate.

Some other distances and measures between distributions that should be noted are,

Definition 4 The **Total Variation metric**,

$$TV(P, Q) = \sup_{A \in \mathcal{B}(\mathbb{R})} |P(A) - Q(A)|$$

Definition 5 The **Hellinger metric,**

$$H(P, Q) = \sqrt{\int (\sqrt{p} - \sqrt{q})^2}$$

Definition 6 The χ^2 **metric** is given by,

$$\chi^2(P, Q) = \int \left(\frac{p}{q} - 1\right)^2 dQ = \int \frac{p^2}{q} - 1$$

*ariel.boyarsky@yale.edu

†Notes based and developed from Larry Wasserman's Nonparametric Statistics course at Carnegie Mellon University.

DEFINITION 7 The **affinity** between two distributions,

$$a(p,q) = \int (p \wedge q)$$

FACT 1 *The following relationships hold,*

1. $TV(P,Q) = \frac{1}{2}\left|\left|P-Q\right|\right|_1 = 1 - a(p,q)$
2. $\frac{1}{2}H^2(P,Q) \leq TV(P,Q) \leq \sqrt{KL(P,Q)} \leq \sqrt{\chi^2(P,Q)}$

2. BOUNDING THE MINIMAX RISK

It is usually very hard to compute the minimax risk. So instead we try to solve for a lower and upper bound,

$$L_n \leq R_n \leq U_n$$

Then if we find that both L_n and U_n decay with the same rate then we have found the minimax rate. For instance if $L_n = cn^{-\alpha}$ and $U_n = Cn^{-\alpha}$ then the minimax rate is $n^{-\alpha}$. Typically we can ignore constants (though not always and sometimes they are important) and so we are usually content with speaking in terms of rates.

Usually the upper bound is easy to find,

$$R_n = \inf_{\hat{\theta}} \sup_{P} \mathbb{E}_P[d(\hat{\theta}, \theta(P)] \leq \sup_{P} \mathbb{E}_P[d(\hat{\theta}, \theta(P))] \equiv U_n \tag{1}$$

Thus, the worst error of any estimator provides an upper bound. Finding the lower bound usually takes some work but there are a few common methods that work in most cases: **Le Cam's method, Fano's method, and Tsybakov's bound.**

2.1. *Lower Bound*

Getting the lower bound requires a series of simplifications be made to the original problem. To get a lower bound we almost always need to use the following theorem which involves the application of several tricks to prove.

THEOREM 1 (Minimax Lower Bound) *Let $M = \{P_1, \ldots, P_N\} \subset \mathcal{P}$ and let $s = \min_{j \neq k} d(\theta_j, \theta_k)$ where $\theta_i = P_i(\theta)$ and let $\psi^* = \arg\min_j d(\hat{\theta}, \theta_j)$. Then,*

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

PROOF: There are three tricks. First, Suppose we choose $M = \{P_1, \ldots, P_N\} \subset \mathcal{P}$. This is a finite set of possible distributions. Then we know,

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

where $\theta_j = P_j(\theta)$. **Note:** This is obvious because the supremum over an infinite set must always be at least as large as the maximum over a finite set that is a subset of the original infinite set. Otherwise the sup of the infinite set would just be the max of the finite set.

But we can simplify this problem further with the second trick. Let $s = \min_{j \neq k} d(\theta_j, \theta_k)$. The Markov inequality gives,

$$\begin{aligned} P[d(\hat{\theta}, \theta) > t] \leq & \frac{\mathbb{E}[d(\hat{\theta}, \theta)]}{t} \\ \implies & \mathbb{E}[d(\hat{\theta}, \theta)] \geq tP[d(\hat{\theta}, \theta) > t] \end{aligned}$$

Now set $t = s/2$ so that,

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

Finally for the third trick for any $\hat{\theta}$ define,

$$\psi^* = \arg\min_j d(\hat{\theta}, \theta_j)$$

Now suppose that $\psi^* \neq j$ in that there exists a closer distribution to our estimator $\hat{\theta}$ then the true parameter θ_j. Suppose instead that $\psi^* = k$. Then,

$$\begin{aligned}
s &\leq d(\theta_j, \theta_k) && \text{Since } s \text{ is the min distance between } \theta\text{s} \\
&\leq d(\theta_j, \hat{\theta}) + d(\hat{\theta}, \theta_k) && \text{Trinagle Inequality} \\
&\leq d(\theta_j, \hat{\theta}) + d(\hat{\theta}, \theta_j) && \text{Since } \psi^* = k \neq j \implies d(\hat{\theta}, \theta_k) \leq d(\hat{\theta}, \theta_j) \\
&= 2d(\theta_j, \hat{\theta})
\end{aligned}$$

So this computation implies that $d(\theta_j, \hat{\theta}) \geq s/2$. And so,

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

This is clear because if the estimator and the truth are greater than $s/2$ apart then there must be a distribution closer to $\hat{\theta}$ that ψ^* then picks up. So that implies that the event in which the distance is greater than $\frac{s}{2}$ contains the event where $\psi^* \neq j$ since $\psi^* \neq j \implies d\left(\hat{\theta}, \theta_j\right) > s/2$. That is,

$$\left\{d\left(\hat{\theta}, \theta_j\right) > s/2\right\} \supset \{\psi^* \neq j\}$$

Now we can substitute this back into the risk,

$$\begin{aligned}
R_n &= \inf_{\hat{\theta}} \sup_P \mathbb{E}_P[d(\hat{\theta}, \theta(P)] \\
&\geq \inf_{\hat{\theta}} \max_{P_j \in M} \mathbb{E}_{P_j}[d(\hat{\theta}, \theta_j] \\
&\geq \frac{s}{2} \inf_{\hat{\theta}} \max_{P_j \in M} P_j\left(d\left(\hat{\theta}, \theta_j\right) > s/2\right) \\
&\geq \frac{s}{2} \inf_{\psi} \max_{P_j \in M} P_j(\psi \neq j)
\end{aligned}$$

Which yields the result in the theorem. *Q.E.D.*

Once we have established Theorem 1 we can begin to apply the methods we mentioned earlier to determine the lower bound.

2.2. *Le Cam's Method*

Perhaps the most widely known and used approach is **Le Cam's Method.** This approach finds the lower bound by averaging over two possible draws $\mathcal{P}$.

THEOREM 2 *Let $\mathcal{P}$ be a set of distributions. For any pair $P_0, P_1 \in \mathcal{P}$,*

$$\inf_{\hat{\theta}} \sup_P \mathbb{E}_P[d(\hat{\theta}, \theta(P)] \geq \frac{s}{4} \int [p_0^n(x) \wedge p_1^n(x)]\, dx = \frac{s}{4}\left[1 - TV\left(P_0^n, P_1^n\right)\right] \tag{2}$$

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

$$\inf_{\hat{\theta}} \sup_P \mathbb{E}_P[d(\hat{\theta}, \theta(P)] \geq \frac{s}{8} e^{-nKL(P_0, P_1)} \geq \frac{s}{8} e^{-n\chi^2(P_0^n, P_1^n)} \tag{3}$$

and,

$$\inf_{\hat{\theta}} \sup_P \mathbb{E}_P[d(\hat{\theta}, \theta(P)] \geq \frac{s}{8}\left(1 - \frac{1}{2}\int |p_0 - p1|\right)^{2n\downarrow}$$

COROLLARY 1 *Suppose $P_0, P_1 \in \mathcal{P}$ such that $KL(P_0, P_1) \leq \log 2/n$. Then,*

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

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

To prove these results we need the following two lemmas. The first lemma establishes that when ψ^* is given by the Newman-Pearson test then we are able to bound the probability of ψ^* "making a mistake" from below. Notice how in Theorem 1 we use the fact that ψ^* chooses the wrong parameter as a way of getting a lower bound.

LEMMA 1 *Let ψ^*be the Neyman-Pearson test. Recall this means,*

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

Then for any test ψ,

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

PROOF: Notice that $p_0 > p_1$ when $\psi^* = 0$ and that $p_0 < p_1$ when $\psi^* = 1$. So,

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

Q.E.D.

The next lemma that we prove allows us to form a relationship between the affinity of two distributions and their Kullback-Leibler divergence. This is necessary in achieving the result shown in equation 3 in theorem 2.

LEMMA 2 *For any distributions P and Q we have,*

$$\int p \wedge q \geq \frac{1}{2} e^{-KL(P,Q)}$$

PROOF: First notice that $(a \wedge b) + (a \vee b) = a + b$. So,

$$\int p \wedge q + \int p \vee q = 2$$

Then,

$$\begin{aligned}
2\int p\wedge q &\geq 2\int p\wedge q-\left(\int p\wedge q\right)^2 \\
&= \int p\wedge q\left[2-\int p\wedge q\right] && \text{Factoring out } \int p\wedge q \\
&= \int p\wedge q\int p\vee q && \text{From Above} \\
&\geq \left(\int\sqrt{(p\wedge q)\,(p\vee q)}\right)^2 && \text{Cauchy-Schwartz} \\
&= \left(\int\sqrt{pq}\right)^2 \\
&= \exp(2\log\int\sqrt{pq}) \\
&= \exp(2\log\int p\sqrt{q/p}) \\
&\geq \exp(2\int p\log\sqrt{q/p}) && \text{Jensen's Inequality} \\
&= \exp\left(\int p\log\sqrt{q/p}+p\log\sqrt{q/p}\right) \\
&= \exp\left(\int p\log\frac{q}{p}\right)=\exp\left(\int p\log\left(\frac{p}{q}\right)^{-1}\right) \\
&= e^{-KL(P,Q)}
\end{aligned}$$

Yielding the result. *Q.E.D.*

Now we are ready to prove Le Cam's theorem.

PROOF OF THEOREM 2: Let $\theta_0=\theta(P_0),\theta_1=\theta(P_1)$ and $s=d(\theta_0,\theta_1)$. Let us suppose $n=1$. Then we have,

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

where

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

Now notice that a maximum of a set is always greater then it's average so we can write,

$$\pi\geq\inf_{\psi}\frac{P_1\left(\psi\neq 1\right)+P_0\left(\psi\neq 0\right)}{2}$$

Now notice that in Lemma 1 we know that the Neyman-Pearson test, ψ^*, will minimize the numerator. So,

$$\begin{aligned}
P_1\left(\psi\neq 1\right)+P_0\left(\psi\neq 0\right) &= \int_{\{p_1>p_0\}}p_0(x)dx+\int_{\{p_1<p_0\}}p_1(x)dx \\
&= \int_{p_1>p_0}\left[p_1\wedge p_0\right]dx+\int_{\{p_1<p_0\}}\left[p_1\wedge p_0\right]dx \\
&= \int\left[p_1\wedge p_0\right]dx
\end{aligned}$$

So,

$$\frac{P_1\left(\psi^*\neq 1\right)+P_0\left(\psi^*\neq 0\right)}{2}=\frac{1}{2}\int\left[p_1\wedge p_0\right]dx$$

Thus with a general n we have,

$$R_n\geq\frac{s}{4}\int\left[p_1^n\wedge p_0^n\right]dx$$

Then using the characterization of the affinity as the exponential Kullback-Leibler from lemma 2 and the fact that $KL(P^n, Q^n) = nKL(P, Q)$ we have,

$$R_n \geq \frac{s}{8} e^{-nKL(P_0, P_1)}$$

From here the other results are just functions of the distances as related to the KL divergence. *Q.E.D.*

THEOREM 3 (General Le Cam) *Suppose $P, Q_1, \ldots, Q_N$ are 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 $q = \frac{1}{N} \sum_{j=1}^{N} q_j$.

EXAMPLE 2 Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ and $X_i \sim Unif[0, 1]$. Consider a nonparametric regression,

$$Y_i = m(X_i) + \epsilon_i$$

and let $\epsilon_i \sim N(0, 1)$. Suppose that m has Lipschitz smoothness. That is,

$$m \in \mathcal{M} = \{m : |m(y) - m(x)| \leq L |x - y|, x, y \in [0, 1]\}$$

So the joint distribution is given by $p(x, y) = p(x)p(y|x) = \phi(y - m(x))$ where ϕ is the normal pdf. We want to know how well we can estimate m. First define, $d(\theta, \theta') = |\theta - \theta'|$. Next set $m_0(x) = 0$ for all x and take the parameter of interest to be $\theta = m(0)$. However we could choose any parameter and our choice of m_0 is a kind of baseline on which we will apply Le Cam's method. Next let us define,

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

This construction will ensure Lipschitz continuity but it also very close (but not too close) to m_0 which will make our computation better. Notice,

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

Then let us compute the KL divergence,

$$\begin{aligned}
KL(P_0, P_1) &= \int_0^1 \int p_0 \log \frac{p_0}{p_1} dy dx \\
&= \int_0^1 \int p_0(x) p_0(y|x) \log \frac{p_0(x) p_0(y|x)}{p_1(x) p_1(y|x)} dy dx \\
&= \int_0^1 \int \phi(y - m_0(x)) \log \frac{\phi(y - m_0(x))}{\phi(y - m_1(x))} dy dx \\
&= \int_0^1 \int \phi(y) \log \frac{\phi(y)}{\phi(y - m_1(x))} dy dx && \text{Since } m_0(x) = 0 \forall x \\
&= \int_0^\epsilon \int \phi(y) \log \frac{\phi(y)}{\phi(y - m_1(x))} dy dx \\
&= \int_0^\epsilon KL(N(0, 1), N(m_1(x), 1)) dx
\end{aligned}$$

Notice how the last equality follows by noticing that the inner integral is simply measuring the KL distance between two normal distributions. Furthermore we know that, $KL(N(\mu_1, 1), N(\mu_2, 1)) = \frac{(\mu_1 - \mu_2)^2}{2}$ such that,

$$KL(P_0, P_1) = \int_0^\epsilon \frac{m_1(x)^2}{2} dx = \frac{1}{2} \int (L(\epsilon - x))^2 dx = \frac{L^2}{2} \int (\epsilon - x)^2 dx = \frac{L^2 \epsilon^3}{6}$$

Now recall corollary 1 which says if we can get the KL distance to be less than or equal to $\log 2/n$ then we can bound R_n from below byty $\frac{s}{2}$. With this goal in mind we can try letting $\epsilon = \left(6\log 2/L^2 n\right)^{1/3}$ which yields,

$$KL(P_0, P_1) = \frac{L^2\left(6\log 2/L^2 n\right)}{6} = \frac{\log 2}{n}$$

Thus for some c,

$$R_n \geq \frac{s}{16} = \frac{L\epsilon}{16} = \frac{L}{16}\left(\frac{6\log 2}{L^2 n}\right)^{1/3} = \left(\frac{c}{n}\right)^{1/3}$$

Thus we have the lower bound rate. Furthermore notice that by bounding from above by the MSE we can get an upper bound of $\left(\frac{C}{n}\right)^{1/3}$ hence,

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

2.3. *Fano's Method*

Suppose instead we are interested in minimizing some integrated metric. For instance the L_2 metric. Then it turns out that Le Cam's method may not give the best lower bound. Instead we can use Fano's method which in a way generalized Le Cam's method from looking at just two distributions to a finite set, $P_1, \ldots, P_N \in \mathcal{P}$ of distributions.

The method begins with Fano's Inequality which is a widely used theorem in statistics, information theory, and computer science.

LEMMA 3 *(Fano Inequality) Let $X_1, \ldots, X_n \sim P$ where $P \in \{P_1, \ldots, P_N\} \subset \mathcal{P}$. Let ψ be any function of $X_1, \ldots, X_n$ that takes values in $\{1, \ldots, N\}$. Let $\beta = \max_{j\neq k} KL(P_j, P_k)$. Then.*

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

To prove this lemma we need to develop some extra tools.

DEFINITION 8 For $0 < p < 1$ we define the **entropy** $h(p) = -p\log p - (1-p)\log(1-p)$ and note that $0 < h(\dot{p}) < \log 2$.

DEFINITION 9 Suppose (Y, Z) are random variables taking values in $\{1, \ldots, N\}$ with joint distribution $P_{Y,Z} = P((Y, Z) \in D)$ then their **mutual information** is defined as,

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

where $H(Y) = -\sum_j P[Y = j]\log(P[Y = j])$ is the entropy of Y and $H(Y|Z)$ is the entropy of Y conditioned on Z.

FACT 2 *For any measurable function h we have that $I(Y, h(Z)) \leq I(Y, Z)$.*

FACT 3 *Suppose Y uniform on $\{1, \ldots, N\}$, then $H(Y) = -\sum_j P[Y = j]\log(P[Y = j]) = -\sum_j \frac{1}{N}\log\left(\frac{1}{N}\right) = -\sum_j \frac{1}{N}(-\log(N)) = \log N$.*

LEMMA 4 *Let Y be a random variable taking values in $\{1, \ldots, N\}$. Let $\{P_1, \ldots, P_N\} \subset \mathcal{P}$ be a set of distributions. Let X be drawn from P_j for some $j \in \{1, \ldots, N\}$. Then $P(X \in A|Y = j) = P_j(A)$. Let $Z = g(X)$ be an estimate of Y taking values in $\{1, \ldots, N\}$. Then,*

$$H(Y|Z) \leq P(Z \neq Y)\log(N - 1) + h(P(Z = Y))$$

PROOF: Let $E = \mathbb{I}_{\{Z \neq Y\}}$. Then,

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

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

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

And, $H(E|X) \leq H(E) = h(P(Z = Y))$. So,

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

Q.E.D.

PROOF OF FANO'S INEQUALITY, LEMMA 3: . Let us assume that $n = 1$. Then the general case will follow by noting that $KL(P^n, Q^n) = nKL(P, Q)$. Suppose Y is uniform on $\{1, \ldots, N\}$. Assume that if $Y = j$ then $X \sim P_j$. Then P is the joint distribution of Y and X,

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

Sine $P(Y = j) = 1/N$. Then,

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

Then from Lemma 4,

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

Then we have,

$$\begin{aligned} P(Z \neq Y) \log (N - 1) &\geq H(Y|Z) - \log 2 = H(Y) - I(Y, Z) - \log 2 \\ &= \log N - I(Y, Z) - \log 2 \geq \log N - \beta - \log 2 \end{aligned}$$

Where we apply Fact 2 to get the last inequality,

$$I(Y, Z) \leq I(Y, X) = \frac{1}{N} \sum_{j=1}^{N} KL(P_j, \bar{P}) \leq \frac{1}{N^2} \sum_{j,k}^{N} KL(P_j, P_k) \leq \beta$$

where $\bar{P} = N^{-1} \sum_{j=1}^{N} P_j$. Then the result follows from,

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

Q.E.D.

Now with Fano's Inequality established we can derive Fano's minimax lower bound.

THEOREM 4 (Fano's Minimax Bound) *Let $F = \{P_1, \ldots, P_N\} \subset \mathcal{P}$. Let $\theta(P)$ be a parameter taking values in a metric space equipped with a metric $d(\cdot, \cdot)$. Then,*

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

where $s = \min_{j \neq k} d(\theta(P_j), \theta(P_k))$ and $\beta = \max_{j \neq k} KL(P_j, P_k)$.

PROOF: Recall that we can write the minimax bound as,

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

Then since the max is always bigger than an average of a finite set we have,

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

Now applying Fano's Inequality we have,

$$R_n \geq \frac{s}{2} \frac{1}{N} \sum_{j=1}^{N} \left(1 - \frac{n\beta + \log 2}{\log N}\right) = \frac{s}{2}\left(1 - \frac{n\beta + \log 2}{\log N}\right)$$

Q.E.D.

COROLLARY 2 *Suppose* $F = \{P_1, \ldots, P_N\} \subset \mathcal{P}$ *and* $N \geq 16$. *If,*

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

Then,

$$R_n \geq \frac{s}{4}$$

2.4. *Tsybakov's Method*

A newer approach that is essentially a simplification of Fano's method is Tsybakov's bound.

THEOREM 5 *Let* $X_1, \ldots, X_n \sim P \in \mathcal{P}$. *Then let* $\{P_0, P_1, \ldots, P_N\} \subset \mathcal{P}$ *where* $N \geq 3$. *Then if ,*

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

Then,

$$R_n \geq \frac{s}{16}$$

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

PROOF: See Appendix of Wasserman's Minimax Notes. *Q.E.D.*

2.5. *Hypercubes*

Notice that when we are using Fano's and Tsybakov's methods we must construct finite sets of distributions, call this $\mathcal{F}$. That is,

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

where

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

which we call a hypercube. Thus there are $N = 2^m$ distributions in $\mathcal{F}$. We also want to define a difference metrics between the sequnces $\omega, \nu \in \Omega$.

DEFINITION 10 The **Hamming metric** is given by $H(\omega, \nu) = \sum_{j=1}^{m} \mathbb{I}_{\{\omega_j \neq \nu_j\}}$. Not to be confused with entropy above or the Hellinger metric.

When we are picking distributions in the hypercube a major problem that may come up is if we just pick $P, Q \in \mathcal{F}$ there is a good chance they are so close together that s will be too small to get a good bound. Imagine them varying by just one coordinate ω_k. So instead we want to "prune" the hypercube so that all the distributions are the right distance apart. We can do this with the following lemma.

LEMMA 5 (Varshamov-Gilbert) *Let $\Omega = \{\omega = (\omega_1, \dots, \omega_m) : \omega_i \in [0,1], i = 1, \dots, m\}$. Suppose that $m \geq 8$. There exists $\omega^1, \dots, \omega^N \in \Omega$ such that,*

1. $\omega^0 = (0, \dots, 0)$
2. $N \geq 2^{m/8}$
3. $H(\omega^j, \omega^k) \geq m/8$ *for all* $0 \leq j < k \leq N$ *where* H *is the Hamming metric.*

We call $\Omega' = \{\omega^0, \dots, \omega^N\}$ *a pruned hypercube.*

PROOF: Let $D = \lfloor m/8 \rfloor$ and set $\omega^0 = (0, \dots, 0)$. 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 already eliminated any ω in which $H(\omega^0, \omega) \leq D$. Now we just continue this process recursively such that at the j-th step we define $\Omega_j = \{\omega \in \Omega_{j-1} : H(\omega, \omega^{j-1}) > D\}$. At this point we have already satisfied 1 and 3. We can show 2 but a little combonotorics. Now let n_j be the number of elements eliminated at step j it follows that,

$$n_j \leq \sum_{i=1}^{D} \binom{m}{i}$$

and define,

$$A_j = \{\omega \in \Omega_j : H(\omega, \omega^j) \leq D\}$$

so that $n_j = |A_j|$ .Then we have that $A_0, \dots, A_N$ partition Ω. This means that $n_0 + \cdots + n_N = 2^m$. Then we have,

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

Then,

$$N + 1 \geq \frac{1}{\sum_{i=0}^{D} 2^{-m} \binom{m}{i}} = \frac{1}{P[\sum_{i=1}^{m} Z_i \leq D]}$$

Then $Z_1, \dots, Z_m$ are Bernoulli-1/2 random variables and by Hoeffding's inequality we have,

$$P[\sum_{i=1}^{m} Z_i \leq D] \leq e^{-9m/32} < 2^{-m/4}$$

But then $N \geq 2^{m/8}$when $m \geq 8$. *Q.E.D.*

Another approach to getting a lower bound using Hypercubes is Assouad's Lemma.

LEMMA 6 (Assouad's Lemma) *Let $\{P_\omega : \omega \in \Omega\}$ be the set of distributions indexed by ω and let $\theta(P)$ be some parameter. Then for any $p > 0$ and any metric d that satisfies the triangle inequality we have,*

$$\max_{\omega \in \Omega} \mathbb{E}_\omega \left(d^p(\hat{\theta}, \theta(P_\omega)) \right) \geq \frac{N}{2^{p+1}} \left(\min_{\{\omega,\nu : H(\omega,\nu) \neq 0\}} \frac{d^p(\theta(P_\omega), \theta(P_\nu))}{h(\omega, \nu)} \right) \left(\min_{\{\omega,\nu : H(\omega,\nu)=1\}} ||P_\omega \wedge P_\nu|| \right)$$

3. EXAMPLES AND FURTHER RESULTS

In this section we provide a variety of examples in which we compute minimax bounds and introduce other related results.

3.1. *Parametric Likelihood*

It is well known that the typical MLE under weak regularity conditions is minimax. That is,

$$R(\theta, \hat{\theta}) = Var_\theta(\hat{\theta}) + \left(b(\hat{\theta}) \right)^2 \approx Var(\hat{\theta})$$

Where $Var(\hat{\theta}) \approx \frac{1}{nI(\theta)}$ where $I(\theta)$ is the Fisher information. Since the bias is usually of order $O(n^{-2})$. Thus,

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

For any other estimator θ' we can show $R(\theta, \theta') \geq R(\theta, \hat{\theta})$. In the d-dimensional case we have, $R = O(d/n)$. In particular this is summarized by a famous theorem of Hajek and Le Cam.

DEFINITION 11 We say that the family of distributions $\{P_\theta : \theta \in \Theta\}$ with density given by p_θ is **differentiable in quadratic mean** if there exists a functional l'_θ such that,

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

DEFINITION 12 We say that a function g is **bowl-shaped** if the *hypograph* which is given by the $\{x : g(x) \leq c\}$ is convex and symmetric about the origin.

THEOREM 6 (Hajek and Le Cam Theorem) *Suppose $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ with non-singular Fisher information matrix I_θ. Furthermore suppose ψ is differentiable at θ. Then $\psi(\hat{\theta}_n)$ where the MLE estimator $\hat{\theta}_n$ is asymptotically locally uniformly minimax in the sense that for any estimator T_n and any bowl-shaped loss function, l, we have that,*

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

where $\mathcal{I}$is the class of all finite subsets of $\mathbb{R}^k$ and $U \sim N(0, \dot{\psi_\theta} I_\theta^{-1} \dot{\psi_\theta}^T)$.

PROOF SKETCH: The idea behind the proof is to note that the left hand side is bigger than,

$$R = \lim_{k \to \infty} \liminf_{n \to \infty} \sup_{h \in I_k} \mathbb{E}_{\theta + h/\sqrt{n}} \left[l\left(\sqrt{n} \left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}} \right) \right) \right) \right]$$

Where we place the rationals in some order and let I_k be the first few vectors in this sequence. Then there is some subsequence of $\{n\}$ $\{n_k\}$ such that,

$$\lim_{k \to \infty} \sup \mathbb{E}_{\theta + h/\sqrt{n}} \left[l\left(\sqrt{n_k} \left(T_{n_k} - \psi\left(\theta + \frac{h}{\sqrt{n_k}} \right) \right) \right) \right]$$

Then to simplify the problem if we apply tightness and lower semi-continuity then we know there exists a convergent subsequence $\{n_k\}$ so that $\sqrt{n}\left(T_n - \psi(\theta)\right)$ converge in law and that because ψ is differentiable $\sqrt{n}\left(T_n - \psi(\theta + h/\sqrt{n})\right)$ also converges in law. So,

$$\sup_{h \in \mathbb{R}^k} \mathbb{E}_h l(T - \dot{\psi}_\theta h) \geq \mathbb{E}_{\nu} l(\dot{\psi}_\theta \mathbb{X}) = \mathbb{E}[l(U)]$$

Where we apply Le Cam's third lemma to $\left(\sqrt{n}(T_n - \psi(\theta)), \frac{1}{\sqrt{n}} \sum \dot{l}_\theta(X_i) \right)$ to get the limiting distribution. Then, we combine these two results to get,

$$\liminf_{n \to \infty} \mathbb{E}_{\theta + h/\sqrt{n}} \left[l\left(\sqrt{n} \left(T_n - \psi\left(\theta + \frac{h}{\sqrt{n}} \right) \right) \right) \right] \geq \mathbb{E}[l(U)]$$

And,

$$R \geq \sup_{h \in \mathbb{Q}^k} \mathbb{E}[l(U)] = \sup_{h \in \mathbb{Q}^k} \mathbb{E}_h l(T - \dot{\psi}_\theta h)$$

Then the result follows by noticing this will not change when replacing rationals with reals. *Q.E.D.*

3.2. *Estimating Smooth Densities*

In this section we show how to use the general strategy to derive the minimax rate for the estimation of a smooth density. Let $\mathcal{F}$ be all the probability densities f on $[0,1]$ such that,

$$0 < c_0 \leq f(x) \leq c_1 < \infty$$

$$|f''(x)| \leq c_2 < \infty$$

Suppose we observe $X_1, \dots, X_n \sim P$ where P has density $f \in \mathcal{F}$. Let us use the squared Hellinger metric, $d^2(f,g) = \int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx$ as a loss function.

Upper Bound: Suppose $\hat{f}_n$ is a Kernel estimator with bandwidth $n^{-1/5}$. Then using bias-variance calculation we get,

$$\sup_f \mathbb{E}_f \left(\int \left(\hat{f} - f\right)^2 \right) dx \leq Cn^{-4/5}$$

So that,

$$\int_0^1 \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx = \int_0^1 \left(\frac{f(x) - g(x)}{\sqrt{f(x)} - \sqrt{g(x)}} \right)^2 dx \leq C_1 \int \left(\hat{f(x)} - f(x)\right)^2 dx$$

So we have $R_n \leq Cn^{-4/5}$.

Lower Bound: We will use Fano's method to get the lower bound. Suppose that g is a bounded twice differentiable function on $[-1/2, 1/2]$ such that,

$$\int_{-1/2}^{1/2} g(x)dx = 0 \text{ and } \int_{-1/2}^{1/2} g^2(x)dx = a > 0 \text{ and } \int_{-1/2}^{1/2} \left(g'(x)\right)^2 dx = b > 0$$

Fix an integer m and for $j = 1, \dots, m$ define $x_j = (j - 0.5)/m$ and,

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

for $x \in [0,1]$ and $c > 0$. Then let $\mathcal{M}$ denote the Varshamov-Gilbert pruned version of the set,

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

$f_\tau \in \mathcal{M}$ let f_τ^n be the product density. Consider,

$$nKL(f_\tau, f_{\tau'}) = n \int f_\tau \log \frac{f_\tau}{f_{\tau'}} \leq \frac{C_1 n}{m^4} = \beta$$

We know by Varshamov Gilbert that there are $N = 2^{m/8}$. Now if we choose $m = cn^{9/5}$ and plug this into β we get $\beta = \frac{C_1 n}{c^4 n^{4/5}}$ and so we can apply the corollary of Fano's minimax bound to get,

$$R_n \geq \frac{C}{n^{4/5}} \implies R_n \asymp \frac{C}{n^{4/5}}$$

3.3. *Normal Means and Pinsker*

Recall the definition of a bowl shaped function. We say a loss function is bowl-shaped if $l(\theta, \theta') = g(\theta - \theta')$ for some bowl-shaped function g.

THEOREM 7 *Suppose the random vector X has a normal distribution. The measure zero unique estimator that is minimax for every bowl-shaped loss function is the sample mean $\bar{X}_n$.*

EXAMPLE 3 (Normal Means Problem) Let $X_j = \theta_j + \epsilon_j/\sqrt{n}$ and we want to estimate $\theta = (\theta_1, \dots, \theta_n)$ with loss function $l(\hat{\theta}, \theta) = \sum_{j=1}^n \left(\hat{\theta_j} - \theta_j\right)^2$ and let $\epsilon_j \sim N(0, \sigma^2)$ for each $j = 1, \dots, n$.

REMARK 1 There is a lot of similarity between normal means and nonparametric estimation. For instance consider the model $Z_i = f(i/n) + \delta_i$ where $\delta_i \sim N(0,1)$ and then expand f using a basis expansion so that: $f(x) = \sum_j \theta_j \psi_j(x)$. An estimate of θ_j is then $X_j = \frac{1}{n} \sum_{i=1}^n Z_i \psi_i(i/n)$ and then $X_j \approx N(\theta_j, \sigma^2/n)$.

In fact in these kinds of estimators with normal random variables we have the following minimax theorem.

DEFINITION 13 We say that a parametric estimator $\hat{\theta}$ is minimax if,

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\hat{\theta}} \sup_{\theta} R(\theta, \hat{\theta})$$

DEFINITION 14 The **James-Stein** estimator is given by,

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

THEOREM 8 (Pinsker) *1. Suppose $\Theta_n = \mathbb{R}^n$ then $R_n = \sigma^2$ and $\hat{\theta} = X = (X_1, \ldots, X_n)$ is minimax.*

2. If the parameter "sieve space" is given by $\Theta_n = \left\{\theta : \sum_{j=1}^{n} \theta_j^2 \leq C^2\right\}$, then

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

(a) This implies that the James-Stein estimator $\hat{\theta}_{JS}$ is asymptotically minimax since,

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

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

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

where $a_j^2 = (\pi j)^{2p}$. Then,

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

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

$$\frac{\sigma^2}{n} \sum_{j} a_j \left(\mu - a_j\right)_+ = C^2$$

The Pinsker estimator is the estimation of a function by smoothing. In fact the Sobelov ellipsoid corresponds to smooth functions. The point here is that if we want to estiamte a smooth function at a minimax rate wee need to shrink the data.

3.4. *Hypothesis Testing*

Let $Y_1, \ldots, Y_n \sim P$ where $P \in \mathcal{P}$ and let $P_0 \in \mathcal{P}$. We want to test,

$$H_0 : P = P_0 \text{ vs. } H_1 : P \neq P_0$$

Recall that a size α test is a function $\phi(y_1, \ldots, y_n) \in \{0, 1\}$ and $P_0^n(\phi = 1) \leq \alpha$. Let Φ be the α level set based on n observations where $0 < \alpha < 1$ is fixed. Our goal is to find the minimax type II error,

$$\beta_n(\epsilon) = \inf_{\phi\in\Phi_n} \sup_{P\in\mathcal{P}(\epsilon)} P^n(\phi = 0)$$

where we define $P(\epsilon) = \{P \in \mathcal{P} : d(P_0, P) > \epsilon\}$. The **minimax testing rate** is then,

$$\epsilon_n = \inf\{\epsilon : \beta_n(\epsilon) \leq \delta\}$$

Lower Bound: To get a lower bound let us define a new distribution Q,

$$Q(A) = \int P^n(A) d\mu(P)$$

where μ is any distribution whose support is in P. Let μ be uniform on $P_1, \ldots, P_N$ then,

$$Q(A) = \frac{1}{N} \sum_j P_j^n(A)$$

Now let us define a likelihood ratio,

$$L_n = \frac{dQ}{dP_0^n} = \int \frac{q(y^n)}{p_0(y^n)} d\mu(q) = \int \prod_j \frac{q(y_j)}{p_0(y_j)} d\mu(p)$$

LEMMA 7 *Let $0 < \delta < 1 - \delta$. If,*

$$\mathbb{E}_0\left[L_n^2\right] \leq 1 + 4(1 - \alpha - \delta)^2$$

then $\beta_n(\epsilon) \geq \delta$.

PROOF: Since $P_0^n(\phi = 1) \leq \alpha$ for each ϕ it follows that $P_0^n(\phi = 0) \geq 1 - \alpha$, we have

$$\begin{aligned}
\beta_n(\epsilon) &= \inf_{\phi \in \Phi_n} \sup_{P \in \mathcal{P}(\epsilon)} P^n(\phi = 0) \geq \inf_{\phi} Q(\phi = 0) \geq \inf_{\phi} \left(P_0^n(\phi = 0) + [Q(\phi = 0) - P_0^n(\phi = 0)]\right) \\
&\geq (1 - \alpha + [Q(\phi = 0) - P_0^n(\phi = 0)]) \geq 1 - \alpha - \sup_A |Q(A) - P_0^n(A)| \\
&= 1 - \alpha - \frac{1}{2} ||Q - P_0^n||_1 = 1 - \alpha - \frac{1}{2} \int_\Omega |Q(\omega) - P_0^n(\omega)| \, dP_0(\omega)
\end{aligned}$$

Then,

$$\begin{aligned}
||Q - P_0^n||_1 &= \int |L_n(y^n) - 1| \, dP_0(y^n) = \mathbb{E}_0[|L_n(y^n) - 1|] \leq \sqrt{\mathbb{E}_0[L_n^2] - 1} = \sqrt{4(1 - \alpha - \delta)^2} \\
&= 2\sqrt{(1 - \alpha - \delta)^2} = 2 |\alpha + \delta - 1|
\end{aligned}$$

So,

$$\beta_n(\epsilon) \geq 1 - \alpha + 2\delta \implies \beta_n(\epsilon) \geq \delta$$

Q.E.D.

Upper Bound: Notice for any size α test ϕ we have,

$$\beta_n(\epsilon) \leq \sup_{P \in \mathcal{P}(\epsilon)} P^n(\phi = 0)$$

3.4.1. *Testing Densities*

Let us now provide an example of how this may work when testing densities. Let $Y_1, \ldots, Y_n \sim P$ where $Y_i \in [0,1]^d$ and let p_0 be the uniform density. Let us test,

$$H_0 : P = P_0$$

Define,

$$\mathcal{P}(\epsilon, s, L) = \left\{ f : \int |f_0 - f| \geq \epsilon \right\} \cap \mathcal{H}_s(L)$$

where $f \in \mathcal{H}_s(L)$ implies it inside a Holder defined as for all x, y,

$$\left| f^{(t-1)}(y) - f^{(t-1)}(x) \right| \leq L |x - y|^t$$

and $\left|\left| f^{(t-1)} \right|\right|_\infty$ for all t.

THEOREM 9 *Fix $\delta > 0$ and define ϵ_n by $\beta(\epsilon_n) = \delta$. Then, there exists $c_1, c_2 > 0$ such that,*

$$c_1 n^{\frac{-2s}{4s+d}} \leq \epsilon_n \leq c_2 n^{\frac{-2s}{4s+d}}$$

PROOF: We begin by dividing the space into $k = n^{2/(4s+d)}$ equal sized bins. We let η be a sequence of Radamacher random variables. That is, η has density given by $f(k) = \begin{cases} 1/2 & k = 1 \\ 1/2 & k = -1 \\ 0 & o.w. \end{cases}$. The let ψ be a smooth function such that $\int \psi = 0$ and $\int \psi^2 = 1$. For the j-th bin B_j let ψ_j be the function ψ rescaled and recentered to be supported on B_j with $\int \psi_j^2 = 1$. Define,

$$p_\eta(y) = p_0(y) + \gamma \sum_j \eta_j \psi_j(y)$$

where $\gamma = cn^{-(2s+d)/(4s+d)}$. It may be verified p_η is a density in $\mathcal{H}_s(L)$ and that $\int |p_0 - p_\eta| \geq \epsilon$. Let N be the number Rademacher sequences. Then,

$$L_n = \frac{1}{N} \sum_\eta \prod_i \frac{p_\eta(Y_i)}{p_0(Y_i)}$$

With this defined we want to verify the inequality in the above lemma. With this satisfied we achieve the lower bound $\beta_n(\epsilon) = \inf_{\phi \in \Phi_n} \sup_{P \in \mathcal{P}(\epsilon)} P^n(\phi = 0) \geq \delta$. So,

$$\begin{aligned} L_n^2 &= \frac{1}{N^2} \sum_\eta \sum_\nu \prod_i \frac{p_\eta(Y_i) p_\nu(Y_i)}{p_0(Y_i) p_0(Y_i)} \\ &= \frac{1}{N^2} \sum_\eta \sum_\nu \prod_i \frac{p_0(y) + \gamma \sum_j \eta_j \psi_j(y)}{p_0(Y_i)} \times \frac{p_0(y) + \gamma \sum_j \nu_j \psi_j(y)}{p_0(Y_i)} \\ &= \frac{1}{N^2} \sum_\eta \sum_\nu \prod_i \left(1 + \frac{\gamma \sum_j \eta_j \psi_j(y)}{p_0(Y_i)}\right) \times \left(1 + \frac{\gamma \sum_j \nu_j \psi_j(y)}{p_0(Y_i)}\right) \end{aligned}$$

Taking the expected value over $Y_1, \ldots, Y_n$ and using the fact the ψ_j are constructed to be orthonormal,

$$\mathbb{E}_0[L_n^2] = \frac{1}{N^2} \sum_\eta \sum_\nu \left(1 + \gamma^2 \sum_j \eta_j \nu_j\right)^n \leq \frac{1}{N^2} \sum_\eta \sum_\nu \exp\left(n\gamma^2 \sum_j \eta_j \nu_j\right)$$

Thus $\mathbb{E}_0\left[L_n^2\right] \leq \mathbb{E}_{\eta,\nu} e^{n\langle \eta, \nu \rangle}$ using the inner product $\langle \eta, \nu \rangle = \gamma^2 \sum_j \eta_j \nu_j$. Hence,

$$\begin{aligned} \mathbb{E}_0[L_n^2] \leq \mathbb{E}_{\eta,\nu} e^{n\langle \eta, \nu \rangle} &= \prod_j \mathbb{E} e^{n\eta_j \nu_j} \\ &= \prod_j \cosh\left(n\rho_j^2\right) \leq \prod_j \left(1 + n^2 \rho_j^4\right) \leq \prod_j e^{n^2\gamma^4} = e^{kn^2\gamma^4} \leq C_0 \end{aligned}$$

Hence we can apply the lower bound from the previous lemma such that, $\beta_n\left(\mathcal{P}\left(\epsilon, L\right)\right) \geq \delta$. *Q.E.D.*

4. BAYESIAN RISK

It is also possible to determine the minimax risk is to use a Bayes estimator. In this section we assume a parametric family of estimators $\{p(x;\theta) : \theta \in \Theta\}$ and the goal is to estimate θ. Then we write the risk $R(\theta, \hat{\theta}_n)$ and the maximum risk is $\sup_{\theta \in \Theta} R(\theta, \hat{\theta}_n)$.

DEFINITION 15 Let Q be the prior distribution for θ. The **Bayes risk** (w.r.t Q) is defined as,

$$B_Q(\hat{\theta}_n) = \int R(\theta, \hat{\theta}_n) dQ(\theta)$$

DEFINITION 16 The **Bayes estimator with respect to Q** is the estimator $\bar{\theta}_n$ that minimized $B_Q(\hat{\theta}_n)$.

DEFINITION 17 The **posterior density** is given by,

$$q(\theta|X^n) = \frac{p(X_1,\dots,X_n;\theta)q(\theta)}{\int p(X_1,\dots,X_n;\theta)q(\theta)d\theta}$$

LEMMA 8 *The Bayes risk can be written as,*

$$\int \left(\int L(\theta,\hat{\theta}_n)q(\theta|x_1,\dots,x_n)d\theta \right) \int p(x_1,\dots,x_n;\theta)q(\theta)d\theta$$

REMARK 2 From this lemma it follows that we can find the minimax estimator $\hat{\theta}$ by just minimizing the inner integral, $\int L(\theta,\hat{\theta}_n)q(\theta|x_1,\dots,x_n)d\theta$.

EXAMPLE 4 Suppose $L(\theta,\hat{\theta}_n) = \left(\theta - \hat{\theta}_n\right)^2$. Then the Bayes estimator is the posterior mean $\bar{\theta}_Q = \int \theta q(\theta|x_1,\dots,x_n)d\theta$.

THEOREM 10 *Let $\hat{\theta}_n$ be an estimator. Suppose,*

1. *The risk function $R(\theta,\hat{\theta}_n)$ is constant as a function of θ.*
2. *$\hat{\theta}_n$ is the Bayes estimator for some prior Q*

Then, $\hat{\theta}_n$ is minimax.

PROOF: Suppose by way of contradiction that $\hat{\theta}_n$ is not minimax. Then there is some other estimator θ' such that,

$$\sup_{\theta\in\Theta} R(\theta,\theta') < \sup_{\theta\in\Theta} R(\theta,\hat{\theta}_n)$$

Then,

$$\begin{aligned} B_Q(\theta') &= \int R(\theta,\theta')dQ(\theta) \\ &\leq \sup_{\theta\in\Theta} R(\theta,\theta') \\ &< \sup_{\theta\in\Theta} R(\theta,\hat{\theta}_n) \\ &= \int R(\theta,\hat{\theta}_n)dQ(\theta) \\ &= B_Q(\hat{\theta}_n) \end{aligned}$$

But then this implies that $B_Q(\theta') \leq B_Q(\hat{\theta}_n)$ which would be a contradiction since $\hat{\theta}_n$ cannot then be the Bayes estimator otherwise it would minimize B_Q. *Q.E.D.*

EXAMPLE 5 We show that the sample mean is minimax for the Normal model. Let $X \sim N_p(\theta, I)$ be a multivariate normal. Let $L(\theta,\hat{\theta_n}) = \left|\left|\hat{\theta}_n - \theta\right|\right|^2$. Assign $Q = N(0, c^2 I)$. Then the posterior is,

$$N(\frac{c^2 x}{1+c^2}, \frac{c^2}{1+c^2} I)$$

The posterior mean $\tilde{\theta} = c^2 X/\left(1+c^2\right)$. The Bayes risk is then $B_Q(\theta) = \frac{pc^2}{1+c^2}$. Furthermore for any other estimator θ^* then,

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

So $R(\Theta) \geq pc^2/\left(1+c^2\right) \implies R(\Theta) \geq p$. But the risk $\hat{\theta}_n = X$ is p and so $\hat{\theta}_n = X$ is minimax.

5. NONPARAMETRIC MAXIMUM LIKELIHOOD

DEFINITION 18 Let $H(\epsilon) = \log N(\epsilon)$ where $N(\epsilon)$ is the smallest number of balls of size ϵ needed to cover $\mathcal{P}$ in the Hellinger metric. We call $H(\epsilon)$ the **Hellinger entropy** of $\mathcal{P}$.

Then the **Le Cam** equation is given by,

$$H(\epsilon_n) = n\epsilon_n^2 \tag{4}$$

Sometimes the minimax rate may be obtained by solving this equation. This is the case with the nonparametric maximum likelihood estimator in which we want to estimate a density function using maximum likelihood. We use the Hellinger metric $H(P,Q) = h(p,q) = \sqrt{\frac{1}{2}\int\left(\sqrt{p}-\sqrt{q}\right)^2}$ as our loss function. Let $\mathcal{P}$ be the space of probability density functions and let us assume this is a infinite dimensional space. We need the following assumptions:

ASSUMPTION 1 *We assume there exists* $0 < c_1 < c_2 < \infty$ *such that* $c_1 < p(x) < c_2$ *for all* x *and all* $p \in \mathcal{P}$.

REMARK 3 This assumption is very strong and is not needed but it will help simplify the proof.

ASSUMPTION 2 *We assume there exists* $\alpha > 0$ *such that,*

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

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

REMARK 4 This assumption implies that the global entropy is of the same order as the local entropy the right hand side.

ASSUMPTION 3 *Let* $\sqrt{n}\epsilon_n \to \infty$ *as* $n \to \infty$ *where* $H(\epsilon_n) \asymp n\epsilon_n^2$.

REMARK 5 This assumption implies that the convergence rate is slower than $O(n^{-1/2})$ which is normal in a nonparametric setting.

EXAMPLE 6 An example of a class that would satisfy assumptions 1, 2, and 3 is,

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

REMARK 6 There are also some important implications of these assumptions. In particular assumption 1 gives us,

$$\begin{aligned} KL(p,q) \leq \chi^2(p,q) &= \int \frac{(p-q)^2}{p} \leq \frac{1}{c_1}\int (p-q)^2 && \text{Using } c_1 < p \\ &= \frac{1}{c_1}\int \left(\sqrt{p}-\sqrt{q}\right)^2\left(\sqrt{p}-\sqrt{q}\right)^2 \\ &\leq \frac{4c_2}{c_1}\int \left(\sqrt{p}-\sqrt{q}\right)^2 = Ch^2(p,q) && \text{Setting } C = \frac{4c_2}{c_1} \end{aligned}$$

Now let ϵ_n solve the Le Cam equation 4. That is,

$$\epsilon_n = \min\left\{\epsilon : H\left(\frac{\epsilon}{\sqrt{2C}}\right) \leq \frac{n\epsilon^2}{16C}\right\} \tag{5}$$

Then we will show that ϵ_n is the minimax rate.

We now proceed to prove an upper bound. In particular let $\mathcal{P} = \{p_1, \ldots, p_N\}$ be a $\epsilon_n/\sqrt{2C}$ covering of the set where $N = N(\epsilon_n/\sqrt{2C})$. The set $\mathcal{P}_n$ will be a **sieve space** in that it will approximate $\mathcal{P}$ better and better as $n \to \infty$. Let $\hat{p}$ be the sieve maximum likelihood estimator defined by,

$$\hat{p} = \arg\max_{p \in \mathcal{P}_n} L(p)$$

where $L(p) = \prod_{i=1}^n p(X_i)$ is the likelihood function. The sieve approach is essentially a type of regularization and allows us to maximize over a finite dimensional space greatly simplifying the problem.

Upper Bound:

Achieving an upper bound can be done with the following inequality from Wong and Shen 1995.

LEMMA 9 (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 an ~~iid~~ sample from p_0. Then,*

$$\Pr[\frac{L(p)}{L(p_0)} > e^{-n\delta^2/2}] \leq e^{-n\delta^{2/4}}$$

PROOF: We compute,

$$\begin{aligned}
\Pr[\frac{L(p)}{L(p_0)} > e^{-n\delta^2/2}] &= \Pr[\prod_{i=1}^n \sqrt{\frac{p(Z_i)}{p_0(Z_i)}}\sqrt{\frac{p(Z_i)}{p_0(Z_i)}} > e^{-n\delta^2/2}] \\
&= \Pr[\prod_{i=1}^n \sqrt{\frac{p(Z_i)}{p_0(Z_i)}} > e^{-n\delta^2/4}] \\
&\leq e^{n\delta^2/4}\mathbb{E}[\prod_{i=1}^n \sqrt{\frac{p(Z_i)}{p_0(Z_i)}}] && \text{Markov} \\
&= e^{n\delta^2/4}\left(\mathbb{E}[\left(\sqrt{\frac{p(Z_i)}{p_0(Z_i)}}\right)]\right)^n \\
&= e^{n\delta^2/4}\left(\int\left(\sqrt{p(Z_i)p_0(Z_i)}\right)]\right)^n && \text{Noting by indep/density } \int\sqrt{p/q} = \int q\sqrt{p/q} \\
&= e^{n\delta^2/4}\left(1 - \frac{h^2(p_0,p)}{2}\right)^n && h^2(P,Q) = 2(1 - \int\sqrt{pq}) \\
&= e^{n\delta^2/4}\exp\left(n\log\left(1 - \frac{h^2(p_0,p)}{2}\right)\right) \\
&\leq e^{n\delta^2/4}e^{-nh^2(p_0,p)/2} = e^{n\delta^2/4 - n\delta^2/2} && \text{First order Taylor} \\
&= e^{-n\delta^2/4}
\end{aligned}$$

Q.E.D.

Now we are ready to compute the upper bound.

THEOREM 11 $\sup_{P \in \mathcal{P}} \mathbb{E}_p\left(h(p, \hat{p})\right) = O(\epsilon_n)$.

PROOF: Suppose p_0 is the true density. Let $p' \in \mathcal{P}_n$ be the element of $\mathcal{P}_n$ that minimizes the KL divergence. That is,

$$p' = \arg\min_{p \in \mathcal{P}_n} KL(p_0, p)$$

Thus,

$$KL(p_0, p') \leq Ch^2(p_0, p') \leq C\left(\epsilon_n^2/2C\right) = \epsilon_n^2/2 \tag{6}$$

Since $N = N(\epsilon_n/\sqrt{2C})$ so that $h(p, q) \leq \sqrt{\epsilon_n^2/2C}$. Let,

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

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

$$\begin{aligned}\Pr[h(\hat{p}, p_0) > \epsilon_n] &\leq \Pr[h(\hat{p}, p_0) + h(p_0, p') > \epsilon_n] \leq \Pr[h(\hat{p}, p_0) + \epsilon_n/\sqrt{2C} > \epsilon_n] \\ &= \Pr[h(\hat{p}, p_0) > A\epsilon_n] = \Pr(\hat{p} \in B) \leq \Pr[\sup_{p \in B} \frac{L(p)}{L(p')} > 1] \\ &\leq \Pr[\sup_{p \in B} \frac{L(p)}{L(p')} > e^{-n\epsilon_n^2(A^2/2+1)}] \\ &\leq \Pr[\sup_{p \in B} \frac{L(p)}{L(p')} > e^{-n\epsilon_n^2(A^2/2)}] + \Pr[\sup_{p \in B} \frac{L(p)}{L(p')} > e^{n\epsilon_n^2}] \equiv P_1 + P_2\end{aligned}$$

Next,

$$P_1 \leq \sum_{p \in B} \Pr[\frac{L(p)}{L(p')} > e^{-n\epsilon_n^2(A^2/2)}] \leq N(\epsilon/\sqrt{2C}) e^{-n\epsilon_n^2 A^2/4} \leq e^{\frac{n\epsilon_n^2}{16C}}$$

Using Lemma (9) with $\delta = \epsilon_n A$ multiplied by the number of Balls needed to cover $\mathcal{P}$ to get the second inequality. And then noting that this covering number is $e^{\frac{n\epsilon^2}{16C}}$ as per the definition of epsilon in equation (5).

Now to bound P_2 define $K_n = \frac{1}{2}\sum_{i=1}^n \log \frac{p_0(Z_i)}{p'(Z_i)}$ as a kind of KL analogue such that $\mathbb{E}[K_n] = KL(p_0, p') \leq \epsilon^2/2$ from equation (6). Furthermore,

$$\begin{aligned}\sigma^2 \equiv Var\left(\log \frac{p_0(Z)}{p'(Z)}\right) &\leq \mathbb{E}[\left(\log \frac{p_0(Z)}{p'(Z)}\right)^2] \leq \log\left(\frac{c_2}{c_1}\right) \mathbb{E}[\left(\log \frac{p_0(Z)}{p'(Z)}\right)] \\ &= \log\left(\frac{c_2}{c_1}\right) KL(p_0, p') \leq \log\left(\frac{c_2}{c_1}\right) \frac{\epsilon_n^2}{2} \equiv c_3\epsilon_n^2\end{aligned}$$

Now we apply the non mean zero Bernstein's inequality to P_2 under log,

$$\begin{aligned}P_2 &= \Pr[\sup_{p \in B} \frac{L(p)}{L(p')} > e^{n\epsilon_n^2}] = \Pr[K_n > \epsilon_n^2] = \Pr[K_n - KL(p_0, p' > \epsilon_n^2 - KL(p_0, p')] \\ &\leq \Pr[K_n - KL(p_0, p') > \frac{\epsilon_n^2}{2}] = \Pr[K_n - \mathbb{E}[K_n] > \frac{\epsilon_n^2}{2}] \leq e^{\frac{-\epsilon_n^4}{8\sigma^2 + c_4\epsilon_n^2}} \\ &\leq \exp(-c_5\epsilon_n^2)\end{aligned}$$

Then $P_1 + P_2 \leq \exp(-c_6 n\epsilon_n^2)$. So finally,

$$\begin{aligned}\mathbb{E}(h(p_0, \hat{p})) &= \int_0^1 \Pr[h(p_0, \hat{p}) > t] dt \\ &= \int_0^{\epsilon_n} \Pr[h(p_0, \hat{p}) > t] dt + \int_{\epsilon_n}^1 \Pr[h(p_0, \hat{p}) > t] dt \\ &\leq \epsilon_n + \exp\left(-c_6 n\epsilon_n^2\right) \leq c_7\epsilon_n\end{aligned}$$

Hence we have a bound from above. *Q.E.D.*

Lower Bound:
The lower bound is an application of Fano's method.

THEOREM 12 *Let ϵ_n be the smallest ϵ such that $H(\alpha\epsilon) \geq 64C^2 n\epsilon^2$. Then,*

$$\inf_{\hat{p}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[h(p, \hat{p})] = \Omega(\epsilon_n)$$

Recall $a = \Omega(\epsilon_n)$ means that $a \geq C\epsilon_n$ for some $C > 0$.

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

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

Then for $P_j, P_k \in F$ we have,

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

So since we can bound the KL divergence we can just use the corollary to Fano's method to get,

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

Q.E.D.

Together with the upper bound we have,

$$R_n \asymp \epsilon_n^2$$

We can now use Le Cam's equation to compute rates for different spaces.

EXAMPLE 7 Neural Networks

In the neural net we have that $f(x) = c_0 + \sum_i c_i \sigma(v_i^T x + b_i)$ where $||c||_1 \leq C$, $||v_i|| \leq 1$, and σ is a step function or a sigmoidal function that is Lipschitz. Le Cam's equation is,

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

Notice,

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

Solving this for ϵ gives the rate.