10-704: Information Processing and Learning **Spring 2015**

Lecture 20: April 2, Minimax Theory & Testing

Lecturer: Akshay Krishnamurthy *Scribe: Siheng Chen*

20.1 Review

20.1.1 Minimax Risk

The minimax risk for class Θ and loss ℓ is

$$\widehat{R}(\Theta) \;=\; \inf_{T} \sup_{\theta \in \Theta} \mathbb{E}_{x \sim P_\theta} \left[\ell \left(T(x), \theta \right) \right],$$

where T is any estimator. The upper bound of the minimax risk is given by designing algorithm and the lower bound of the minimax risk is given by information theoretical techniques.

Testing problems focus on specific loss function $\ell \left(T(x), \theta \right) = \mathbf{1}\{T(x) \neq \theta\}$, so, the minimax risk is

$$\widehat{R}_n(\Theta) \;=\; \inf_{T} \sup_{\theta \in \Theta} \mathbb{P}_{x \sim \theta} \left[T(x) \neq \theta \right].$$

In the previous lecture, we saw that if there are two parameters θ_0 and θ_1, then the minimax task is lower bounded by

$$\begin{aligned} \widehat{R}_n(\{\theta_0, \theta_1\}) &\overset{(a)}{\geq} \frac{1}{2} - \frac{1}{2} \left\| P_{\theta_0}^n - P_{\theta_1}^n \right\|_{TV} \\ &\geq \frac{1}{2} - \frac{1}{2} \sqrt{\frac{1}{2} KL(P_{\theta_0}^n, P_{\theta_1}^n)}, \end{aligned}$$

where (a) follows from Neyman-Pearson lemma. We saw lower bounds for a simple normal mean testing problem.

20.1.2 Neyman-Pearson Lemma

For simple vs. simple tests, the optimal statistics is the likelihood ratio test

$$\Lambda(x) \;=\; \frac{P_0(x)}{P_1(x)}, \quad T(x) = \mathbf{1}\{\Lambda(x) \leq x\},$$

and

$$\frac{1}{2} P_0[T(x) \neq 0] + \frac{1}{2} P_1[T(x) \neq 1] = \frac{1}{2} - \frac{1}{2} \left\| P_0 - P_1 \right\|_{TV}.$$

There are two important ways to use the Neyman-Pearson lemma, both of these are sometimes called Le Cam's method.

1. We can always throw away parameters in the supremum and lower bound the risk:

$$\inf_T \sup_\Theta \mathbb{P}_\theta [\cdot] \geq \inf_T \sup_{\Theta' \subseteq \Theta} \mathbb{P}_\theta [\cdot].$$

   Any problem with $\mathbf{1}\{\cdot\}$ loss can be lower bounded by just choosing two parameters $\theta_0, \theta_1 \in \Theta$ and computing their TV or KL.

2. We can also separate the parameter space into two regions and mix over these sets.

$$\begin{aligned}
\inf_T \sup_\Theta \mathbb{P}_\theta [T(x) \neq \theta] &\geq \inf_T \sup_{j \in \{0,1\}} \sup_{\theta \in \Theta_j} \mathbb{P}_\theta [T(x) \neq j] \\
&\geq \inf_T \{ \frac{1}{2} \mathbb{E}_{\theta \sim \pi_0, x \sim P_\theta} [\mathbf{1}\{T(x) \neq 0\}] + \frac{1}{2} \mathbb{E}_{\theta \sim \pi_1, x \sim P_\theta} [\mathbf{1}\{T(x) \neq 1\}] \} \\
&\geq \frac{1}{2} - \frac{1}{2} \| P_{\pi_0} - P_{\pi_1} \|_{TV},
\end{aligned}$$

   where $P_{\pi_0}(A) = \mathbb{E}_{\theta \sim \pi_0}[P_\theta(A)]$, π_0 is a distribution on Θ_0, and π_1 is a distribution on Θ_1.

   This is important for some problems. By mixing you can make the distributions much closer together to prove stronger lower bounds. But it is often challenging to compute the divergence to mixtures.

20.2 Information Theoretic Connections and Fano's Method

One way to think about Le Cam's method is as a channel decoding problem. Given a channel $\Theta \to X$, we send $\Theta \in \{0, 1\}$, and you see the samples $X \sim P_\theta$. If P_0 is close to P_1, then you will have a high decoding error, because when P_0 close to P, $H(\theta|X)$ is big. Earlier in the class we saw another result this form, which is Fanno's lemma.

Consider a Markov chain $\Theta \to X \to T$. Let $P_e = \mathbb{P}[T \neq \Theta]$, for any test/decoder T:

$$\begin{aligned}
h(P_e) + P_e \log(|\Theta| - 1) &\geq H(\Theta|X), \\
&\text{or,} \\
P_e &\geq \frac{H(\Theta|X) - \log 2}{\log(|\Theta| - 1)},
\end{aligned}$$

where $P_e = \mathbb{P}_{\theta \sim unif, x \sim P_\theta}[T(x) \neq \theta]$. Using the identities from earlier in the course, there are many equivalent ways to state this inequality:

$$\inf_T \sup_\Theta P_e \geq 1 - \frac{I(\Theta; X) + \log 2}{\log |\Theta|} = 1 - \frac{\mathbb{E}_{\theta \sim \pi}[KL(P_\theta || P_\pi)]}{\log |\Theta|}.$$

This is the *global Fano's method*.

We can weaken the mixture representation of KL to obtain the *local Fano method*,

$$I(\Theta; X) = \int \pi(\theta) P_\theta(X) \log \left(\frac{\pi(\theta) P_\theta(X)}{\pi(\theta) \int \pi(\theta) P_\theta(X)} \right) = \mathbb{E}_{\theta \sim \pi} [KL(P_\theta || P_\pi)] \leq \mathbb{E}_{\theta, \theta' \sim \pi} [KL(P_\theta || P_{\theta'})].$$

If we have M hypothesis $\theta_1, \cdots, \theta_M$, then we obtain

$$\begin{aligned}
\inf_T \sup_{j \in [M]} P_{\theta_j}[T(x) \neq j] &\geq \inf_T \frac{1}{M} \sum_{j=1}^M P_{\theta_j}[T(x) \neq j] \\
&\geq 1 - \frac{\frac{1}{M^2} \sum_{i,j} KL(P_{\theta_i} || P_{\theta_j}) + \log 2}{\log M}.
\end{aligned}$$

Example: testing for nonzero in a 1-sparse vector in $\mathbb{R}^d$, $k \ll d$.

$$H_v : \ x_1^n \sim \mathcal{N}(\mu v, 1), \tag{20.1}$$

where $v \in \{0,1\}^d$, with only 1 nonzero component. There are d hypothesis and each one has $KL(P_i^n || P_j^n) = 2n\mu^2$. The local Fano method then gives

$$R_n(\Theta) \geq 1 - \frac{2n\mu^2 + \log 2}{\log d},$$

which is bounded away from zero if

$$\mu \leq \sqrt{\frac{\log d}{n}}.$$

Note that this rate is achieved for this problem by the test that takes the largest coordinate of $\bar{X}$.

$$T(X^n) = \arg\max_j \bar{X}_j.$$

By Gaussian tail bound and union bound, we know that

$$\mathbb{P}[\forall j, |\bar{X}_j - \mu_j| \geq \epsilon] \leq 2d \exp\{-2n\epsilon^2\},$$

or, with probability $\geq 1 - \epsilon$:

$$\forall j, |\bar{X}_j - \mu_j| \geq \sqrt{\frac{\log(2d/\epsilon)}{2n}}.$$

The estimated coordinate $\hat{j}$ agrees with the true one $j^\star$ if:

$$\begin{aligned} \bar{X}_{j^\star} &\geq \bar{X}_k, \ \forall \ k \\ \bar{X}_{j^*} - \mu_{j^*} + \mu_{j^*} - \mu_k + \mu_k &\geq \bar{X}_k \\ \mu_{j^*} - \mu_k &\geq \bar{X}_k - \mu_k + \mu_{j^*} - \bar{X}_{j^*} \\ \mu &\geq 2\sqrt{\frac{\log(2d/\epsilon)}{2n}}. \end{aligned}$$

so that if $\mu = \omega(\sqrt{\frac{\log(d)}{n}})$, this estimator has success probability tending to 1.

Theorem 1 *For the 1 sparse recovery problem, the minimax rate is:*

$$\mu \asymp \sqrt{\frac{\log d}{n}}.$$

Actually the same rate holds for the k-sparse problem, but it is slightly less obvious.

Some takeaway messages are: by *discretizing*, we can look at a few close hypotheses, and this could lower the KL much more than entropy. Also, there are many techniques, like Neyman-Pearson, local and global Fano just for testing problems. It is important to know about all of these techniques because some are better for some problems.

20.3 Estimation Problem

Now let's turn to estimation problems, or more general losses. We write:

$$R_n(\Theta) \;=\; \inf_T \sup_\Theta \mathbb{E}\left[\Phi \circ \rho(T(X), \Theta)\right]$$

where $\rho : \Theta \times \Theta \to \mathbb{R}_+$ is a metric, $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a non-decreasing function with $\Phi(0) = 0$.

Example: $\rho(\Theta, \Theta') = |\Theta - \Theta'|$ and $\Phi(t) = t^2$, so we are looking at mean square error. This can also cover things like classification performance, excess risk, things we have seen before.

20.3.1 Proving lower bounds

Step 1: Discretization. Fix a $\delta > 0$, and find a large set of parameters $\Theta' = \{\theta_i\}_{i=1}^M \subseteq \Theta$, such that

$$\rho(\theta_i, \theta_j) \geq 2\delta, \quad \forall\, i \neq j.$$

This set is called a 2δ packing in the ρ-metric.

Step 2: Reduce to Testing. Consider $j \sim$ uniform($[M]$) and $X \sim P_{\theta_j}$. Now if you cannot differentiate between θ_i and some other θ, you will certainly make error $\Phi(\delta)$ in the estimation problem. More formally:

Proposition 1 *Let $\{\theta_j\}_{j=1}^M$ be a 2δ-packing in the ρ metric. Then:*

$$R_n(\Theta, \Phi \circ \rho) \geq \Phi(\delta) \inf_\Psi \mathbb{P}_{j \sim unif([M]), x^n \sim P_{\theta_j}} \left[\Psi(x^n) \neq j\right].$$

Proof: Fix an estimator T. For any fixed θ, we have

$$\mathbb{E}[\Psi(\rho(T, \theta))] \geq \mathbb{E}[\Psi(\delta)\mathbf{1}\{\rho(T, \theta) \geq \delta\}] = \Psi(\delta)\mathbb{P}[\rho(T, \theta) \geq \delta].$$

Now, let $\Psi(\widehat{T}) = \arg\min_j \rho(T, \theta_j)$. If $\rho(T, \theta_j) < \delta$, then $\Psi(T) = j$ by 2δ separation triangle inequality,

$$\rho(T, \theta_k) \geq \rho(\theta_j, \theta_k) - \rho(T, \theta_k) > 2\delta - \delta = \delta.$$

The converse of this statement is that if $\Psi(T) \neq v$, then $\rho(T, \theta_v) \geq \delta$.

$$\sup_{\theta \in \Theta} \mathbb{P}[\rho(T, \theta) \geq \delta] \geq \frac{1}{M} \sum_{j=1}^M \mathbb{P}_j[\rho(T, \theta_j) \geq \delta] = \frac{1}{M} \sum_{j=1}^M \mathbb{P}_j[\Psi(T) \neq j].$$

Now take an inf over all T, Ψ. ■

Step 3: Use Fano or Neyman Pearson to Lower Bound P_e in Testing Problems. We saw how to do this earlier in this lecture and in the previous lecture.

Example 1 (Normal Means Estimation in ℓ_2) *Let $x_1^n \sim \mathcal{N}(v, 1), v \in \mathbb{R}^d$. The goal is to have $\mathbb{E}_{x_1^n} \|T(X_1^n - v)\|_2^2$ small. Let U be a $1/2$ packing of the unit ball in $\mathbb{R}^d$. Note that the unit ball in d dimensions has a packing of size at least 2^d in the ℓ_2 metric. For each $u \in U$, let $\theta_u = \delta_u \in \mathbb{R}^d$, so that*

$$\|\theta_u - \theta_{u'}\|_2 = \delta \|u - u'\|_2 \geq \frac{\delta}{2}. \tag{20.2}$$

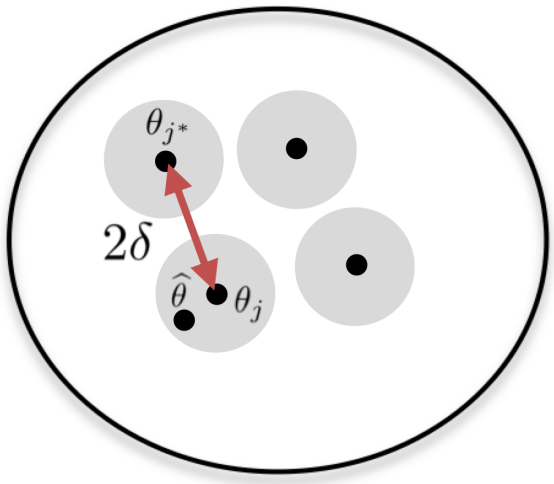

Figure 20.1: If you get θ_k instead of θ_{j^*}, then your estimate $\widehat{\theta}$ must be far from θ_{j^*}.

By triangle inequality, $\|\theta_u - \theta_{u'}\| \leq 2\delta$. so the KL between each pair of $\theta_u, \theta_{u'}$ is

$$KL\{P_{\theta_u} || P_{\theta_{u'}}\} \leq 2n\delta^2,$$

so the Fano's Lemma gives

$$\inf_T \frac{1}{M} \sum_{j=1}^{M} P_{\theta_j}[T(x_1^n \neq j)] \neq 1 - \frac{2n\delta^2 + \log 2}{d \log 2},$$

thus, lower bound is

$$\begin{aligned} R_n(\Theta, \|\cdot\|_2^2) &\geq \left(\frac{\delta}{4}\right)^2 \left[\inf_T \mathbb{E}_j \mathbb{P}_{\theta_j}[T(X^n) \neq j]\right] \\ &\geq \left(\frac{\delta^2}{16}\right)\left(1 - \frac{2n\delta^2 + \log 2}{d \log 2}\right) \end{aligned}$$

Now we can choose δ, set it to $\delta^2 = d \log 2/(\delta n)$. Then, $R_n \geq cd/n$. This is the right rate for this problem.

20.3.2 Metric Entropy

The size of the parameter space shows the difficulty of an estimation problem. This shows up in packing and covering numbers which, as we saw, play a role in our minimax lower bounds.

Definition 1 *A δ-covering of the set $\mathcal{X}$ with metric ρ is a set $\{x_1, x_2, \cdots, x_N\}$ that satisfies, for any $x \in \mathcal{X}$, there exists some $i \in \{1, 2, \cdots, N\}$, such that $d(x, x_i) \leq \delta$. The δ-covering number of $\mathcal{X}$ is*

$$N(\delta, \mathcal{X}, d) \;=\; \inf\{\; |\mathcal{X}_1| \;:\; \textit{there exists a } \delta - \textit{covering of } \mathcal{X}_1 \textit{ of } \mathcal{X}\}.$$

Metric entropy of the set $\mathcal{X}$ is the logarithm of the covering number, $\log N(\delta, \mathcal{X}, d)$.

Definition 2 *A δ-packing of the set $\mathcal{X}$ with metric ρ is a set $\{x_1, x_2, \cdots, x_N\}$, such that $d(x_i, x_j) \geq \delta$, for all $i \neq j$. The δ-packing number of $\mathcal{X}$ is*

$$M(\delta, \mathcal{X}, d) \;=\; \sup\{\; |\mathcal{X}_1| \;:\; \textit{there exists a } \delta - \textit{packing of } \mathcal{X}_1 \textit{ of } \mathcal{X}\}.$$

The packing number and covering number satisfy the following relationships:

$$M(2\delta, \mathcal{X}, d) \leq N(\delta, \mathcal{X}, d) \leq M(\delta, \mathcal{X}, d).$$