10-704: Information Processing and Learning **Fall 2016**

Lecture 21: Nov 14

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

21.1 Minimax Risk and Le Cam's lower bound

The minimax risk for class Θ and loss ℓ is

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

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

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

In the previous lecture, we saw that if there are two parameters θ_0 and θ_1, then Le Cam's method shows that the minimax task is lower bounded by

$$\begin{aligned} R_n(\{\theta_0,\theta_1\}) &\overset{(a)}{\geq} \frac{1}{2} - \frac{1}{2}\left\|P^n_{\theta_0} - P^n_{\theta_1}\right\|_{TV} \\ &\geq \frac{1}{2} - \frac{1}{2}\sqrt{\frac{1}{2}KL(P^n_{\theta_0}, P^n_{\theta_1})}, \end{aligned}$$

We saw lower bounds for a simple normal mean testing problem.

We also saw that we can use Le Cam's method for composite hypothesis tests using the following two tricks:

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

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

   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.

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

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.

21.2 Neyman-Pearson Lemma

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

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

and

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

Proof: In last class, we saw that for any deterministic test $T : \mathcal{X} \to \{0,1\}$ with acceptance region $A = \{x \in \mathcal{X} : T(x) = 1\}$

$$\mathbb{P}_0(T\neq 0) + \mathbb{P}_1(T\neq 1) = \mathbb{P}_0(A) + \mathbb{P}_1(A^c) = 1 - \mathbb{P}_1(A) + \mathbb{P}_0(A) \tag{21.1}$$

The result follows by noticing that this is minimized if A is the region where $P_0(x) \le P_1(x)$. ■

21.3 Information Theoretic Connections and Fano's Method

Another way to think of minimax testing 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. Fano's inequality characterizes this relationship and can be used for proving minimax lower bounds for multiple hypothesis tests.

Consider the Markov chain $\theta \to X \to T$. Let $P_e = \mathbb{P}[T \neq \theta]$, for any test/decoder T Fano's inequality implies that

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

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

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

since

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

This is the *global Fano's method*.

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

$$\mathbb{E}_{\theta\sim\pi}\left[KL(P_\theta||P_\pi)\right] \;\le\; \mathbb{E}_{\theta,\theta'\sim\pi}\left[KL(P_\theta||P_{\theta'})\right].$$

The last step follows from Jensen's inequality since KL divergence is convex in the second argument.

In this case, if we have M hypothesis $\theta_1, \cdots, \theta_M$, then we obtain (here $[M] = 1, \ldots, M$)

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

21.4 Application to testing for nonzero in a 1-sparse vector in $\mathbb{R}^d$

$$H_v : \; X_1^n \overset{iid}{\sim} \mathcal{N}(\mu v, 1), \tag{21.2}$$

where $v \in \{0,1\}^d$, with only 1 nonzero component. There are d hypothesis and each pair 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

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

Note that this rate is achieved for this problem by the largest coordinate of $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

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

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 $\geq 1 - \delta$:

$$\forall j, |\bar{X}(j) - \mu(j)| \leq \sqrt{\frac{\log(2d/\delta)}{2n}}.$$

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

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

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.

Also, there are many techniques for proving lower bounds, like Le Cam, 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.

21.5 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 semi-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 log loss, things we have seen before.

21.5.1 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:*

$$R_n(\Theta, \Phi \circ \rho) \geq \Phi(\delta) \inf_\Psi \mathbb{P}_{j \sim unif([M]), X_1^n \sim P_{\theta_j}} \left[\Psi(X_1^n) \neq j\right].$$

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

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

Now, define the test $\Psi(T) = \arg\min_j \rho(T, \theta_j)$. If $\rho(T, \theta_j) < \delta$, then $\Psi(T) = j$ by 2δ separation and triangle inequality since

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

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 Le Cam to Lower Bound P_e in Testing Problems. We saw how to do this earlier in this lecture and in the previous lecture.

21.6 Normal Means Estimation in ℓ_2

Let $X_1^n \sim \mathcal{N}(v, I), v \in \mathbb{R}^d$. 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 $\theta_u = \delta u \in \mathbb{R}^d$ for some $\delta > 0$, so that

$$\|\theta_u - \theta_{u'}\|_2 = \delta \|u - u'\|_2 \geq \frac{\delta}{2}. \tag{21.3}$$

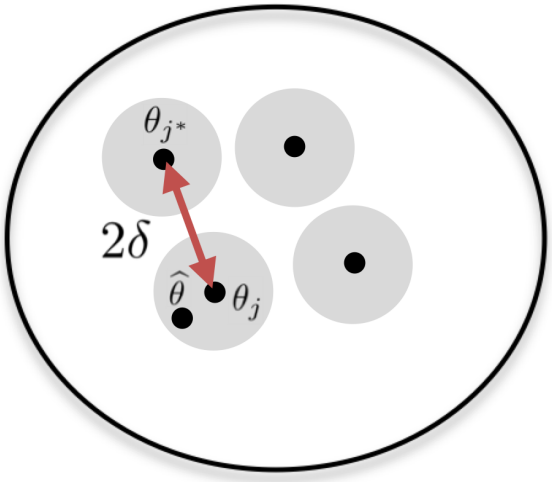

Figure 21.1: If you get θ_j instead of θ_{j^*}, then your estimate $\widehat{\theta}$ must be far from θ_{j^*}.

Also notice that since u, u' lie in the unit ball, $\|\theta_u - \theta_{u'}\| \leq \delta$. so the KL between each pair of $\theta_u, \theta_{u'}$ is

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

so the Fano's Lemma gives

$$\inf_T \frac{1}{M} \sum_{j=1}^{M} P_{\theta_j}[T(X_1^n) \neq j] \geq 1 - \frac{n\delta^2/2 + \log 2}{d \log 2},$$

thus, lower bound is

$$\begin{aligned} R_n(\Theta, \|\cdot\|_2^2) &\geq \left(\frac{\delta}{4}\right)^2 \left[\inf_T \mathbb{E}_j \mathbb{P}_{\theta_j}[T(X_1^n) \neq j]\right] \\ &\geq \left(\frac{\delta^2}{16}\right)\left(1 - \frac{n\delta^2/2 + \log 2}{d \log 2}\right) \end{aligned}$$

Now we can choose δ, set it to $\delta^2 = d \log 2/(2n)$. Then, for $d \geq 2$ [1], $R_n \geq cd/n$ for some constant $c > 0$. This is the right parametric rate for this problem.

21.7 Strong data processing inequalities

How can we leverage these lower bound techniques to new settings that arise in modern learning problems? One approach is to use *strong data processing inequalities*, as modern learning settings can be thought of as a classical problem with some transformation to the data, i.e.

$$\text{parameter} \rightarrow \text{classical data} \rightarrow \text{new data} \tag{21.4}$$

$$\theta \rightarrow X \rightarrow Z \tag{21.5}$$

[1] For $d = 1$ the problem reduces to testing two simple hypothesis for which we can use Le Cam's method.

Example: Local Differentially private channel: Channel $X \to Z$ must be differentially private for each data point, i.e. for each data point X_i we have distribution $Q(Z|X)$ s.t.

$$\sup_S \sup_{x,x' \in \mathcal{X}} \frac{Q(Z_i \in S | X_i = x)}{Q(Z_i \in S | X_i = x')} \leq exp(\alpha). \tag{21.6}$$

We would like to leverage existing technology to get lower bound in these settings for learning with Z. Clearly we can use data processing inequality, where we get $I(\theta, X) \geq I(\theta, Z)$. But this bound is quite loose. Thus we are interested in strong data processing inequalities, where suppose we have channel $\theta \to X \to Z$, and $Q(Z|X)$ is the distribution of $Z|X$ with certain property, we want to show that $I(\theta; Z) \leq f(Q) I(\theta; X)$, where $f(Q) \ll 1$, which yields a much tighter lower bound.

In the next class, we will see that $(\alpha, 0)$ differentially private learning leads to α^2 contraction in KL divergence, which means the effective sample size goes from n to $n\alpha^2$. This means that if we had n samples in the differentially private setting, it is as if we only had $n\alpha^2$ samples in the classical setting. So we need more samples in the new setting to learn well.