Statistics 210B Lecture 27 Notes

Daniel Raban

April 28, 2022

1 Methods for Proving Minimax Lower Bounds

1.1 Recap: Testing lemma and divergence measures for minimax lower bounds

We have been studying minimax lower bounds. We have a semi-meric $\rho : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$ and a 2δ-separated set $\{\theta^1, \dots, \theta^M\} \subseteq \Theta$. In our testing situation, we have the joint distribution

$$Q : \begin{cases} J \sim \mathrm{Unif}(\{1, 2, \dots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

We have an increasing function Φ, as well. We proved the following result:

Proposition 1.1 (From estimation to testing). *Let Ψ be increasing and $\{\theta^1, \dots, \theta^M\}$ be 2δ-separated for $\delta > 0$. Then*

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}(\psi(Z) \neq J).$$

We also defined the total variation distance the K-L divergence, and the Hellinger distance

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx,$$

$$D(\mathbb{P} \,||\, \mathbb{Q}) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \, dx,$$

$$\mathbb{H}^2(\mathbb{P} \,||\, \mathbb{Q}) \int_{\mathcal{X}} (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx.$$

These had the following relationships:

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\tfrac{1}{2} D(\mathbb{P} \,||\, \mathbb{Q})},$$

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq \sqrt{\mathbb{H}^2(\mathbb{P} \,||\, \mathbb{Q})} \underbrace{\sqrt{1 - \frac{\mathbb{H}^2(\mathbb{P} \,||\, \mathbb{Q})}{4}}}_{\leq 1}.$$

$$\mathbb{H}^2(\mathbb{P} \,||\, \mathbb{Q}) \leq \frac{1}{2} D(\mathbb{P} \,||\, \mathbb{Q}).$$

1.2 Le Cam's two points method

Take $M = 2$. Then $J \sim \mathrm{Unif}(\{0, 1\})$, and $Z \mid J = j \sim \mathbb{P}_j$, and $\overline{\mathbb{Q}} = \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$. We claim that

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}(1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}}).$$

Proof. For any ψ, we can find an A such that

$$\psi(x) = \begin{cases} 1 & x \in A \\ 0 & x \in A^c. \end{cases}$$

Then

$$\begin{aligned} \mathbb{Q}(\psi(Z) = J) &= \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\mathbb{P}_0(A^c) \\ &= \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2}. \end{aligned}$$

If we take the supremum over all ψ, we get

$$\begin{aligned} \sup_{\psi} \mathbb{Q}(\psi(Z) = J) &= \sup_{A} \frac{1}{2}(\mathbb{P}_1(A) - \mathbb{P}_0(A)) + \frac{1}{2} \\ &= \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}} + \frac{1}{2} \end{aligned}$$

The probability of the bad event is then

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2} - \frac{1}{2}\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}. \qquad \square$$

This gives the following theorem.

Theorem 1.1 (Le Cam's two points lower bound)**.** *For all $\delta > 0$ and $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ with $\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$,*

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2}(1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}).$$

For the generalization to Le Cam's convex hull method, read chapter 15.2.2 in Wainwright's textbook.

Example 1.1 (Gaussian location family, $d = 1$)**.** Our model is $\mathcal{P} = \{\mathbb{P}_\theta = N(\theta, \sigma^2) : \theta \in \mathbb{R}\}$, where σ is known. We have the semimetric $\rho(\theta', \theta)) = |\theta' - \theta|$ and $\Phi(t) = t^2$. Our sample is $X_{1:n} \sim \mathbb{P}_\theta^n$. The true minimax risk is $\mathcal{M}_n = \frac{\sigma^2}{n}$. Here is a lower bound by Le Cam's method:

Consider $\mathbb{P}_{2\delta}$ and $\mathbb{P}_0$, so $\rho(2\delta, 0) \geq 2\delta$. Then

$$\mathcal{M}_n(\theta(\mathcal{P}); |\theta - \theta'|^2) \geq \frac{\delta^2}{2}(1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}),$$

where the n only appears in the bound as the fact that the measures are product measures. We want to lower bound $1 - \|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$ by $1/2$. We have by Pinsker's inequality and the tensorization property of K-L divergence

$$\begin{aligned}
\|\mathbb{P}_{2\delta}^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 &\leq \frac{1}{2} D(\mathbb{P}_{2\delta}^n \mid\mid \mathbb{P}_0^n) \\
&= \frac{1}{2} n D(\mathbb{P}_{2\delta} \mid\mid \mathbb{P}_0) \\
&= \frac{1}{2} n \frac{(2\delta)^2}{2\sigma^2} \\
&= \frac{n\delta^2}{\sigma^2}.
\end{aligned}$$

Now choose $\frac{n\delta_n^2}{\sigma^2} = \frac{1}{2}$, so $\delta_n^2 = \frac{\sigma^2}{2n}$. Then $\|\mathbb{P}_{2\delta_n}^n - \mathbb{P}_0^n\|_{\mathrm{TV}} \leq \frac{1}{2}$, and we get the minimax lower bound

$$\mathcal{M}_n \geq \frac{\delta_n^2}{2} \cdot \frac{1}{2} = \frac{\sigma^2}{16n}.$$

Up to constants, this is sharp.

Here is the problem with Le Cam's method. If we take $\theta \in \mathbb{R}^d$ with $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d)$ for $d \geq 2$, then we will get the lower bound

$$\mathcal{M}_n \geq \frac{\sigma^2}{16n},$$

even though the actual minimax risk is $\mathcal{M}_n = \sigma^2 \frac{d}{n}$.

1.3 Mutual information

Here, we will develop some tools for Fano's method, which is a sharper method for lower bounding the minimax risk. Suppose we have two random variables $(X, Y) \sim \mathbb{P}_{X,Y}$. We want a measure of their dependence/independence (not the same as correlation). If X is independent of Y, we have

$$\mathbb{P}_{X,Y} = \mathbb{P}_X \times \mathbb{P}_Y = \int_{\mathcal{Y}} \mathbb{P}_{X,Y}(x, y)\, dy \times \int_{\mathcal{X}} P_{X,Y}(x, y)\, dx.$$

To get a measure of independence, we should look at the distance between these two objects:

$$D\left(\mathbb{P}_{X,Y}, \int_{\mathcal{Y}} \mathbb{P}_{X,Y}(x,y)\, dy \times \int_{\mathcal{X}} P_{X,Y}(x,y)\, dx\right).$$

Definition 1.1. The **mutual information** between X and Y is

$$I(X;Y) := D(\mathbb{P}_{X,Y} \mid\mid \mathbb{P}_X \times \mathbb{P}_Y).$$

Remark 1.1. The mutual information is always ≥ 0. Although the K-L divergence is not symmetric, we have $I(X;Y) = I(X;Y)$.

If X and Y are independent, $I(X;Y) = 0$, and if $Y = f(X)$, the mutual information is maximized.

Recall that

$$Q : \begin{cases} J \sim \mathrm{Unif}(\{1, 2, \dots, M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Then

$$\begin{aligned} I(J;Z) &= D(\mathbb{Q}_{2,J} \mid\mid \mathbb{Q}_2 \times \mathbb{Q}_J) \\ &= \frac{1}{M} \sum_{j=1}^{M} D(\mathbb{P}_{\theta^j} \mid\mid ba\mathbb{Q}), \end{aligned}$$

where

$$\overline{Q} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}.$$

Suppose $\theta^j = \theta$ for all j. Then $I(J;Z) = 0$. Conversely, if the θ^j are far away from each other, then $I(J;Z)$ will be large.

Here are two upper bounds of $I(J;Z)$ we will now prove:

Proposition 1.2.

$$I(J;Z) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D(\mathbb{P}_{\theta^j} \mid\mid \mathbb{P}_{\theta^k}) \leq \max_{j,k} D(\mathbb{P}_{\theta^j} \mid\mid \mathbb{P}_{\theta^k}).$$

Lemma 1.1 (Yang-Barron's bound)**.** *Let $N_{\mathrm{KL}}(\varepsilon; \mathcal{P})$ be an ε-cover of $\mathcal{P}$ in $\sqrt{D_{\mathrm{KL}}}$. Then*

$$I(Z;J) \leq \inf_{\varepsilon > 0} \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon; \mathcal{P})$$

1.4 Fano's inequality

Let

$$Q : \begin{cases} J \sim \mathrm{Unif}(\{1,2,\ldots,M\}) \\ Z \mid J = j \sim \mathbb{P}_{\theta^j}. \end{cases}$$

Lemma 1.2.

$$\inf_{\psi} \mathbb{Q}(\psi(Z) \neq J) \geq 1 - \frac{I(Z;J) + \log 2}{\log M}.$$

The proof is in Section 15.4 and requires some ideas such as the entropy. This does not require any restriction on the $\mathbb{P}_{\theta^j}$. This lower bound gives us

Proposition 1.3. *Let $\{\theta^1, \ldots, \theta^M\}$ be 2δ-separated in the semimetric ρ. Then*

$$\mathcal{M}_n(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(Z;J) + \log 2}{\log M}\right).$$

When using this lower bound, we will find δ_n such that

$$1 - \frac{I(Z;J) + \log 2}{\log M} \geq \frac{1}{2}.$$

Then we will get

$$\mathcal{M}_n \geq \frac{1}{2}\Phi(\delta_n).$$

So we need to upper bound $I(Z;J)$.

A simple upper bound is given by

$$I(J;Z) = \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{P}_{\theta^j} \mid\mid \frac{1}{M}\sum_{\ell=1}^{M} \mathbb{P}_{\theta^\ell})$$

$$\leq \frac{1}{M^2}\sum_{j,\ell=1}^{M} D(\mathbb{P}_{\theta^j} \mid\mid \mathbb{P}_{\theta^\ell})$$

Where we have used Jensens's inequality to show that the K-L divergence is convex in the second argument.

$$\leq \max_{j,\ell} D(\mathbb{P}_{\theta^j} \mid\mid \mathbb{P}_{\theta^\ell})$$

Example 1.2 (Gaussian location family, $d \geq 2$)**.** Our model is $\mathcal{P} = \{\mathbb{P}_\theta = n(\theta, \sigma^2 I_d) : \theta \in \mathbb{R}^d\}$, where σ is known. Our semimetric is $\rho(\theta', \theta) = \|\theta' - \theta\|_2$ with $\Phi(t) = t^2$. The true minimax risk is

$$\mathcal{M}_n = \inf_{\widehat{\theta}} \sup_{\theta} \mathbb{E}[\|\widehat{\theta} - \theta\|_2^2] = \sigma^2 \frac{d}{n}.$$

The lower bound by Fano's method gives

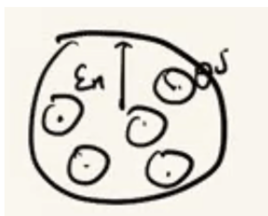
$$\begin{aligned}\mathcal{M}_n &\geq \Phi(\delta)\left(1 - \frac{I(Z;J)+\log 2}{\log M}\right)\\ &\geq \Phi(\delta)\left(1 - \frac{\max_{j,k} D(\mathbb{P}^n_{\theta^j} \mid\mid \mathbb{P}^n_{\theta^k}) + \log 2}{\log M}\right)\end{aligned}$$

Our goal is to find the largest $\delta_n, M, \{\theta^1, \ldots, \theta^M\}$ such that

(a) $\|\theta^j - \theta^k\|_2 \geq 2\delta_n$

(b)
$$\frac{\max_{j,k} D(\mathbb{P}^n_{\theta^j} \mid\mid \mathbb{P}^n_{\theta^k}) + \log 2}{\log M} \leq \frac{1}{2}.$$

Here is our construction: Let $\varepsilon_n = \sigma\sqrt{\frac{d}{n}}$ and $\delta_n = \frac{1}{100}\varepsilon_n = \frac{1}{100}\sigma\sqrt{\frac{d}{n}}$. Let $\{\theta^1, \ldots, \theta_M\}$ be a maximal $2\delta_n$ packing of $B(0, \varepsilon_n) = \{\theta \in \mathbb{R}^d : \theta\|_2 \leq \varepsilon_n\}$.



By a volume argument, we can get upper and lower bounds of M:

$$\log M \asymp d \log\left(c\frac{\varepsilon_n}{\delta_n}\right) \asymp c \cdot d.$$

To upper bound the K-L divergence on top, we have

$$\begin{aligned}\max_{j,k} D(\mathbb{P}^n_{\theta^j} \mid\mid \mathbb{P}^n_{\theta^k}) &= n \max_{j,k} D(\mathbb{P}_{\theta^j} \mid\mid \mathbb{P}_{\theta^k})\\ &= n \max_{j,k} \frac{n\|\theta^j - \theta^k\|_2^2}{2\sigma^2}\\ &\leq \frac{n\varepsilon_n^2}{2\sigma^2}\\ &= c \cdot d\end{aligned}$$

Our quantities only depend on the ratio between ε_n and δ_n, so we can adjust the constant in front of δ_n to get the desired upper bound of $\frac{1}{2}$.

We then get

$$\mathcal{M}_n \geq \Phi(\delta_n)\frac{1}{2} = \frac{1}{2} \cdot \left(\frac{1}{100}\right)^2 \sigma^2 \frac{d}{n} = c\sigma^2\frac{d}{n}.$$

1.5 Yang-Barron's method

The bound on $I(J; Z)$ by the max of the K-L divergences is generally only good when we have a parametric problem. For nonparametric problems, we want to use a better bound.

Lemma 1.3 (Yang-Barron's bound)**.** *Let $N_{\mathrm{KL}}(\varepsilon; \mathcal{P})$ be an ε-cover of $\mathcal{P}$ in $\sqrt{D_{\mathrm{KL}}}$. Then*

$$I(Z; J) \leq \inf_{\varepsilon > 0} \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon; \mathcal{P})$$

To apply this bound, we have two steps:

1. Choose $\varepsilon_n > 0$ such that
$$\varepsilon_n^2 \geq \log N_{\mathrm{KL}}(\varepsilon_n; \mathcal{P}).$$

2. Choose the largest $\delta_n > 0$ such that
$$\log M(\delta_n; \rho, \Omega) \geq 4\varepsilon_n^2 + 2\log 2.$$