# Statistics 210B Lecture 26 Notes

Daniel Raban

April 25, 2022

## 1 Introduction to Minimax Lower Bounds

### 1.1 Minimax risk and methods of obtaining lower bounds

In the last few lectures, we were talking about upper bounds for error of statistical estimators. Now we will prove some lower bounds, which tell us that for a certain number of samples, you cannot have vanishing estimation error.

In statistical decision theory, we have a class of distributions $\mathcal{P}$ and a parameter/function of distributions $\theta : \mathcal{P} \to \Theta$. If this is a one to one mapping, we write $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. Then we have **statistical estimators**, which are mappings $\widehat{\theta} : \mathcal{X} \to \Theta$. Suppose there is a **semimetric**[1] $\rho(\theta, \theta') : \Theta \times \Theta \to \mathbb{R}$, such as

$$\rho(\theta, \theta') = \|\theta - \theta'\|_2, \qquad \rho(f, f') = \|f - f'\|_{L^2}.$$

If $\Phi : [0, \infty) \to [0, \infty)$ is increasing, the **risk** is

$$R(\widehat{\theta}; \theta(P)) = \mathbb{E}_{X \sim P}[\Phi(\rho(\widehat{\theta}(X); \theta(P)))].$$

In this framework, the **loss function** is $\ell = \Phi \circ \rho$.

**Definition 1.1.** The **minimax risk** with $n$ samples is

$$\mathcal{M}_n(\theta(\mathcal{P}), \Phi \circ \rho) = \inf_{\widehat{\theta}: \mathcal{X} \to \Theta} \sup_{P \in \mathcal{P}} R(\widehat{\theta}; \theta(P))$$

The inf and the sup mean that we are taking the best estimator for the worst model.

(a) If $R(\widehat{\theta})$ achieves $\mathcal{M}_n$, it is good enough.

(b) If $R(\widehat{\theta}) \gg \mathcal{M}_n$, we should either find a better estimator or a sharper lower bound.

[1] For a semimetric, we may allow $\theta \neq \theta'$ to still have $\rho(\theta, \theta') = 0$.

**Example 1.1.** Let $\Theta = \mathbb{R}^d$ with $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d), \theta \in \mathbb{R}^d$, where $\sigma^2$ is known. Our sample is $(x_i)_{i\in[n]} \overset{\text{iid}}{\sim} \mathbb{P}_\theta$, so $x_{1:n} \sim \mathbb{P}_\theta^n$. Our metric is $\rho(\theta, \theta') = \|\theta - \theta'\|_2$, and we pick $\Phi(t) = t^2$. Consider the estimator $\widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^n x_i$. Then

$$R(\widehat{\theta}_n; P_\theta) = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n x_i - \theta\right\|_2^2\right] = \sigma^2\frac{d}{n},$$

which tells us that

$$\mathcal{M}_n \leq \sigma^2\frac{d}{n}.$$

However, we can prove the same value as a lower bound. Our goal in this lecture and the next is to show that $\mathcal{M}_n \geq c\sigma^2\frac{d}{n}$ for some constant $c$.

**Remark 1.1.** Here are some methods of showing lower bounds for estimation error, some of which we have already seen.

(a) Bayesian decision theory: $\mathcal{M}_n$ is the Bayes risk of the least favorable prior.

(b) Cramer-Rao lower bound: For unbiased estimators, there is a lower bound given in terms of the Fisher information. If this does not depend on the Fisher information, then it is a minimax lower bound.

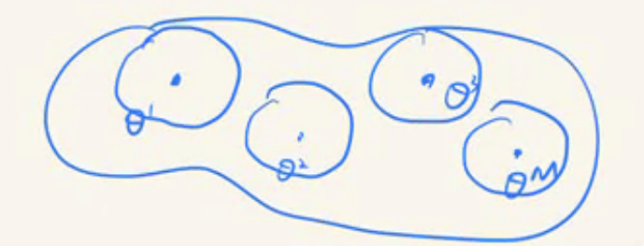
(c) Bayes Cramer-Rao (Van-Tree's inequality): This gives a "local minimax" lower bound.

(d) Reduction to a testing problem: We will study this now. We first need some tools from information theory.

## 1.2 Reduction to an $M$-ary testing problem

The idea is to find a testing problem easier than the estimation problem. A lower bound for the testing problem will imply a lower bound for estimation.

Step 1: Construct a $2\delta$-separated set of $\Theta$ in the $\rho$-metric.



So we require $\rho(\theta^i, \theta^j) \geq 2\delta$ for all $i \neq j$. This is the same as a packing, except we allow $\geq$ instead of $>$. If our separated set is $\{\theta^1, \theta^2, \ldots, \theta^M\}$, we get $\{\mathbb{P}_{\theta^1}, \mathbb{P}_{\theta^2}, \ldots, \mathbb{P}_{\theta^M}\}$.

Step 2: Sample $(J, Z) \in [M] \times \mathcal{X}$. The joint distribution is

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

Step 3: Let $\mathbb{Q}$ be the joint distribution of $(J, Z)$. Then the marginal distribution of $Z$ is

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

Our testing problem is that we want to find a $\psi : \mathcal{X} \to [M]$ such that $\mathbb{Q}(\psi(Z) \neq J)$ is small. If $M = 2$, this is standard binary hypothesis testing. The testing error is

$$\mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}[\underbrace{\mathbb{P}_{\theta^1}(\psi(Z) \neq 1)}_{\text{Type I error}} + \underbrace{\mathbb{P}_{\theta^2}(\psi(Z) \neq 2)}_{\text{Type II error}}].$$

This is different from the traditional hypothesis testing setup in that instead of fixing the Type I error and minimizing the Type II error, we want to minimize the average of these errors.

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

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

This works for all $\delta > 0$, so we can pick the $\delta$ which gives the best lower bound. In general, $\Phi(\delta)$ is increasing with $\delta$, but the testing error $\inf_\psi \mathbb{Q}(\psi(Z) \neq J)$ is decreasing with $\delta$. We can choose $\delta = \delta_n$ such that $\inf_\psi \mathbb{Q}(\psi(Z) \neq J) = \frac{1}{2}$; any constant would work here. Then the minimax lower bound will be

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

*Proof.* Fix $P$ and $\widehat{\theta}$. By Markov's inequality,

$$\begin{aligned} \mathbb{E}[\Phi(\rho(\widehat{\theta}, \theta))] &\geq \Phi(\delta)\mathbb{P}(\Phi(\rho(\widehat{\theta}, \theta)) \geq \Phi(\delta)) \\ &= \Phi(\delta)\mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta). \end{aligned}$$

We now want to relate this probability with the testing error. We have

$$\sup_{P \in \mathcal{P}} \mathbb{P}(\rho(\widehat{\theta}, \theta) \geq \delta) \geq \sup_{\theta \in \{\theta^1, \ldots, \theta^M\}} \mathbb{P}_\theta(\rho(\widehat{\theta}, \theta) \geq \delta)$$

$$\geq \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}_{\theta^j}(\rho(\widehat{\theta},\theta^j)\geq\delta)$$
$$=\mathbb{Q}(\rho(\widehat{\theta},\theta^J)\geq\delta).$$

Define a test $\psi$ via $\widehat{\theta}$: Let

$$\psi(z)=\arg\min_{L\in[M]}\rho(\widehat{\theta}(Z),\theta).$$

This gives the $\theta^j$ which is the closest to our estimate $\widehat{\theta}(Z)$. With this definition,

$$\{\psi(Z)\neq J\}\subseteq\{\rho(\widehat{\theta}(Z),\theta^j)\geq\delta\}.$$

This means we can lower bound the above $\mathbb{Q}$ probability:

$$\inf_{\widehat{\theta}}\mathbb{Q}(\rho(\widehat{\theta}(Z),\theta^j)\geq\delta)\geq\inf_{\psi}\mathbb{Q}(\psi(Z)\neq J). \qquad \square$$

How do we choose $\{\theta^1,\ldots,\theta^M\}$? Moreover, how do we lower bound $\inf_\psi\mathbb{Q}(\psi(Z)\neq J)$? Here are two general methods.

1. $M=2$: Le Cam's method
   - Two points method
   - Convex hull method
2. $M\geq 3$:
   - Assoaud's method
   - Fano's method

Le Cam's method is the most classical one, so we will start with it. Fano's method is the most important and useful method for high-dimensional models.

### 1.3 Some divergence measures

Here are some basic tools for these methods. Let $\mathbb{P},\mathbb{Q}$ be two probability distributions on $\mathcal{X}$. How can we measure their distance?

**Definition 1.2.** The **total variation distance** is

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}}:=\sup_{A\subseteq\mathcal{X}}|\mathbb{P}(A)-\mathbb{Q}(A)|=\frac{1}{2}\int_{\mathcal{X}}|p(x)-q(x)|\,d\mu(x),$$

where $p,q$ are the densities of $\mathbb{P},\mathbb{Q}$, if they exist.

**Definition 1.3.** The **Kullback-Leibler divergence** is

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

There is a more general definition of the K-L divergence that does not require $\mathbb{Q}, \mathbb{P}$ to have densities with respect to Lebesgue measure. This is not a distance because $D(\mathbb{Q} \,||\, \mathbb{P}) \neq D(\mathbb{P} \,||\, \mathbb{Q})$, but it has distance-like properties, such as $D(\mathbb{P} \,||\, \mathbb{Q}) \geq 0$ with $D(\mathbb{P} \,||\, \mathbb{Q}) = 0$ iff $\mathbb{P} = \mathbb{Q}$.

**Definition 1.4.** The **Hellinger distance** is

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

Here are some relationships between these notions of distance:

**Proposition 1.2** (Pinsker's inequality)**.**

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

**Proposition 1.3** (Le Cam's inequality)**.**

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

**Proposition 1.4.**

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

We will see that the TV distance is related to the testing error for a binary testing situation. On the other hand, the KL-divergence and Hellinger distance have good tensorization properties: If we let

$$\mathbb{P}^{1:n} = \mathbb{P}_1 \times \mathbb{P}_2 \times \cdots \times \mathbb{P}_n, \qquad \mathbb{Q}^{1:n} = \mathbb{Q}_1 \times \mathbb{Q}_2 \times \cdots \times \mathbb{Q}_n,$$

then

$$D(\mathbb{P}^{1:n} \,||\, \mathbb{Q}^{1:n}) = \sum_{i=1}^{n} D(\mathbb{P}_i \,||\, \mathbb{Q}_i),$$

$$\frac{1}{2}\mathbb{H}^2(\mathbb{P}^{1:n} \,||\, \mathbb{Q}^{1:n}) = 1 - \prod_{i=1}^{n} \left(1 - \frac{1}{2}\mathbb{H}^2(\mathbb{P}_i \,||\, \mathbb{Q}_i)\right).$$

**Example 1.2** (Gaussian distribution)**.** For a Gaussian distribution, we have the density

$$p_\theta = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \qquad \theta \in \mathbb{R}.$$

The K-L divergence is

$$\begin{aligned}
D(\mathbb{P}_\theta \,||\, \mathbb{P}_{\theta'}) &= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \log \frac{\exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right)}{\exp\left(-\frac{(x-\theta')^2}{2\sigma^2}\right)}\, dx \\
&= \mathbb{E}_{X\sim\mathbb{P}_\theta}\left[-\frac{(X-\theta)^2}{2\sigma^2} + \frac{(X-\theta')^2}{2\sigma^2}\right] \\
&= \frac{(\theta')^2}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} + \frac{1}{\sigma^2}\,\mathbb{E}_{X\sim\mathbb{P}_\theta}[(\theta-\theta')X] \\
&= \frac{(\theta')^2}{2\sigma^2} - \frac{\theta^2}{2\sigma^2} + \frac{1}{\sigma^2}(\theta-\theta')\theta \\
&= \frac{(\theta-\theta')^2}{2\sigma^2}.
\end{aligned}$$

Using Pinsker's inequality and the tensorization property of the K-L divergence,

$$\begin{aligned}
\|\mathbb{P}_\theta^n - \mathbb{P}_{\theta'}^n\|_{\mathrm{TV}} &\le \sqrt{\frac{1}{2} D(\mathbb{P}_\theta^n \,||\, \mathbb{P}_{\theta'}^n)} \\
&\le \sqrt{\frac{n}{2} D(\mathbb{P}_\theta \,||\, \mathbb{P}_{\theta'})} \\
&\le \sqrt{\frac{n(\theta-\theta')^2}{4\sigma^2}}.
\end{aligned}$$

We can also calculate the Hellinger distance

$$\mathbb{H}^2(\mathbb{P}_\theta \,||\, \mathbb{P}_{\theta'}) = 1 - \exp\left(-\frac{(\theta-\theta)^2}{8\sigma^2}\right).$$

More generally, for $\theta \in \mathbb{R}^d$ and $\mathbb{P}_\theta = N(\theta, \sigma^2 I_d)$, we get

$$D(p_\theta \,||\, p_{\theta'}) = \frac{\|\theta-\theta\|_2^2}{2\sigma^2},$$

$$\mathbb{H}^2(\mathbb{P}_\theta \,||\, \mathbb{P}_{\theta'}) = 1 - \exp\left(-\frac{\|\theta-\theta\|_2^2}{8\sigma^2}\right).$$