140.850: Mathematical Foundation of Big Data Analysis **Spring 2016**

Lecture 8: Minimax Lower Bounds: LeCam, Fano, and Assouad

Lecturer: Fang Han *March 07*

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 Lecturer.*

8.1 Overview on minimax lower bound

This lecture note is focused on introducing the general framework for constructing lower bounds for any given statistical problem. Such a framework, formulated by Lucien LeCam and many others in 1970s-80s, aims to mathematically rigorously understand statistical problems' challenge via a worst-case analysis.

We define a "statistical problem" as follows. It contains three components: (1) a parameter space Θ; (2) a class of probability measures $\{P_\theta, \theta \in \Theta\}$; (3) a semi-distance $d(\cdot,\cdot) : \Theta \times \Theta \to \mathbb{R}$ on Θ. A statistical problem aims to recover θ, measured by $d(\cdot,\cdot)$, given P_θ.

Example 8.1.1. *Throughout this note, we will repeatedly visit the sparse normal mean estimation problem: estimate $\boldsymbol{\theta} \in \Theta_s := \{\boldsymbol{v} \in \mathbb{R}^p : |\text{supp}(\boldsymbol{v})| \le s\}$ based on $P_{\boldsymbol{\theta}} := N_p(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_d)^{\otimes n}$. Here $d(\cdot,\cdot)$ is commonly adopted to be the Euclidean distance on $\mathbb{R}^p$ space.*

Regarding a general statistical problem, we aim to know how hard, at worst, it is to recover any given $\theta \in \Theta$. This is formulated by the *maximum risk* of the estimator on Θ:

$$r(\widehat{\theta}_n) := \sup_{\theta \in \Theta} \mathbb{E}_\theta d^2(\widehat{\theta}_n, \theta).$$

8.1.1 Upper bounding the maximum risk

By the techniques we have learnt so far, usually it is not too hard to upper bound $r(\widehat{\theta}_n)$. For example, in Example 8.1.1, let's consider the MLE (also called the least square estimator):

$$\widehat{\boldsymbol{\theta}}_n := \operatorname*{argmin}_{\boldsymbol{v} \in \Theta_s} \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{v}\|^2.$$

We then have the following theorem.

Theorem 8.1.2 (Sparse normal mean estimation - upper bound)**.** *For Example 8.1.1, we have*

$$\sup_{\boldsymbol{\theta} \in \Theta_s} \mathbb{E}_{\boldsymbol{\theta}} \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2^2 \le C \cdot \frac{\sigma^2 s \log(ep/s)}{n},$$

where C is an absolute constant.

Remark 8.1.3. *Note that the upper bound does not depend on the scale of $\boldsymbol{\theta}$, which is a little bit surprising.*

Proof. For any $\boldsymbol{\theta} \in \Theta_s$, by definition of $\widehat{\boldsymbol{\theta}}_n$, we have

$$\sum_{i=1}^n \|\boldsymbol{X}_i - \widehat{\boldsymbol{\theta}}_n\|^2 \le \sum_{i=1}^n \|\boldsymbol{X}_i - \boldsymbol{\theta}\|^2,$$

which implies

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2^2 \le 2\langle \overline{\boldsymbol{X}} - \boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\rangle \quad \Rightarrow \quad \|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 \le 2 \left\langle \overline{\boldsymbol{X}} - \boldsymbol{\theta}, \frac{\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}}{\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2} \right\rangle$$

where $\overline{\boldsymbol{X}} := \frac{1}{n}\sum \boldsymbol{X}_i$. By the standard uniform consistency argument in EP, using the fact that $|\text{supp}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})| \le 2s$, we can continue writing

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 \le 2 \sup_{\boldsymbol{v} \in B(2s)} |\boldsymbol{v}^T(\overline{\boldsymbol{X}} - \boldsymbol{\theta})| \quad \text{where } B(s) := \{\boldsymbol{v} \in \mathbb{R}^p : |\text{supp}(\boldsymbol{v})| \le s, \|\boldsymbol{v}\|_2 = 1\}.$$

We then employ a covering net argument as in Lecture note #4. In particular, for any $\mathbf{a} \in \mathbb{R}^q$, $\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2 \le \epsilon$, and $\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\| = 1$, we have

$$|(\boldsymbol{v}_1^T \mathbf{a})^2 - (\boldsymbol{v}_2^T \mathbf{a})^2| \le 2\epsilon \sup_{\|\boldsymbol{v}\|_2 = 1} (\boldsymbol{v}^T \mathbf{a})^2.$$

This implies

$$\sup_{\|\boldsymbol{v}\|_2 = 1} |\boldsymbol{v}^T \mathbf{a}|^2 \le (1 - 2\epsilon)^{-1} \sup_{\boldsymbol{v} \in \mathcal{N}_\epsilon} |\boldsymbol{v}^T \mathbf{a}|^2.$$

Here $\mathcal{N}_\epsilon$ is the ϵ-net on $\mathbb{S}^{q-1}$, with the cardinality smaller than $(1 + 2/\epsilon)^q$. This further implies

$$\begin{aligned} P_{\boldsymbol{\theta}}(\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 \ge 2\sqrt{2}t) &\le P_{\boldsymbol{\theta}}(\sup_{\boldsymbol{v} \in B(2s)} |\boldsymbol{v}^T(\overline{\boldsymbol{X}} - \boldsymbol{\theta})| \ge \sqrt{2}t) \\ &\le \binom{p}{2s} 9^{2s} P_{\boldsymbol{\theta}}(|\boldsymbol{v}^T(\overline{\boldsymbol{X}} - \boldsymbol{\theta})| \ge t) \\ &\le 2\left(\frac{9p}{2s}\right)^{2s} \exp(-nt^2/2\sigma^2) \end{aligned}$$

where the last inequality is due to the fact that $\overline{\boldsymbol{X}} - \boldsymbol{\theta} \sim N_p(\mathbf{0}, \sigma^2 \mathbf{I}_d/n)$ under $P_{\boldsymbol{\theta}}$, implying $\boldsymbol{v}^T(\overline{\boldsymbol{X}} - \boldsymbol{\theta}) \sim N_q(0, \sigma^2/n)$. This implies

$$\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\|_2 = O_P(\sqrt{s\log(ep/s)/n}).$$

The expectation version is by $\mathbb{E}X = \int_{t>0} P(X \ge t)dt$ for any positive r.v. X. □

8.1.2 A general reduction scheme

In this section, we introduce a general framework to calculate lower bounds for any given statistical problem. This section largely follows Tsybakov's wonderful book "Introduction to Nonparametric Estimation".

Our aim is to lower bound the minimax error:

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta d^2(\widetilde{\theta}_n, \theta),$$

where $\widetilde{\theta}_n$ is any measurable statistic on P_θ for any given θ.

(1) The first step is to reduce the expectation bounds to probability. This is via the Markov's inequality:

$$\mathbb{E}_\theta d(\widetilde{\theta}_n, \theta) \geq A P_\theta(d(\widetilde{\theta}_n, \theta) \geq A).$$

Therefore, as long as we can find a A such that $\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq C > 0$ for some absolute constant C, we have

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta d^2(\widetilde{\theta}_n, \theta) \geq C^2 A^2.$$

(2) The second step is to find a worst-case parameter space $\{\theta_0, \theta_1, \ldots, \theta_M\} \in \Theta$ of infinite number of elements (θ_0 is made to be the reference value). This is usually the key step. By the property of infimum, we must have

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq \inf_{\widetilde{\theta}_n} \sup_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A).$$

(3) The third step is to make $\theta_0, \ldots, \theta_M$ uniformly distinguishable, so that we could transfer the estimation problem to a testing problem. In detail, we must have to construct $\theta_0, \ldots, \theta_M$ such that

$$d(\theta_j, \theta_k) \geq 2A, \quad \text{for any } k \neq j.$$

Accordingly, for any estimator $\widetilde{\theta}_n$, if there exists θ_k such that $d(\widetilde{\theta}_n, \theta_k) \leq d(\widetilde{\theta}_n, \theta_j)$, then we must have $d(\widetilde{\theta}_n, \theta_j) \geq A$ (otherwise $d(\theta_j, \theta_k) \leq d(\widetilde{\theta}_n, \theta_k) + d(\widetilde{\theta}_n, \theta_k) \leq 2A$). In other words, we must have

$$P_{\theta_j}(d(\widetilde{\theta}_n, \theta_j) \geq A) \geq P_{\theta_j}(\Psi^* \neq j), \quad \text{for } j = 0, \ldots, M,$$

where Ψ^* is the *minimum distance test*:

$$\Psi^* := \operatorname*{argmin}_{0 \leq j \leq M} d(\widetilde{\theta}_n, \theta_j).$$

Therefore, we have successfully transferred the estimation problems to testing problems:

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq \inf_{\Psi} \sup_{0 \leq j \leq M} P_{\theta_j}(\Psi \neq j),$$

The testing problem on the righthand side has been very nice to deal with. In particular, when $M = 1$, we have recovered the two-class hypothesis testing problem, which by intuition we know Nayman-Pearson Lemma could get in to give a sharp lower bound (actually, this is one of the major motivations to bring information theory to statistics: MLE is exactly the projection estimator regarding the K-L divergence!). The lower bound for multiple (but finite) hypothesis testing problem is constructed by LeCam. As you can imagine, it is still based on likelihood ratios.

8.2 LeCam's approach

In the following, we drop the absolute continuous issue. Please check Tsybakov's book for the complete version.

Theorem 8.2.1. *Let $P_0, P_1, \ldots, P_M$ be probability measures on $(\mathcal{X}, \mathcal{A})$. We then have*

$$\inf_{\Psi} \sup_{0 \leq j \leq M} P_j(\Psi \neq j) \geq \sup_{\tau > 0} \frac{\tau M}{1 + \tau M} \left[\frac{1}{M} \sum_{j=1}^{M} P_j\left(\frac{dP_0}{dP_j} \geq \tau \right) \right].$$

Proof. Let's write Ψ be the test taking values $\{0, 1, \dots, M\}$. Let

$$T_j := \left\{ \frac{dP_0}{dP_j} \geq \tau \right\}.$$

We then have

$$\begin{aligned} P_0(\Psi \neq 0) &= \sum_{j=1}^M P_0(\Psi = j) = \sum_{j=1}^M \int I(\Psi_j) dP_0 = \sum_{j=1}^M \int I(\Psi_j) \frac{dP_0}{dP_j} dP_j \geq \sum_{j=1}^M \tau P_j(\{\Psi = j\} \cap T_j) \\ &\geq \tau M \left(\frac{1}{M} \sum_{j=1}^M P_j(\Psi \neq j) \right) - \tau \sum_{j=1}^M P_j(T_j^C) = \tau M(p_0 - \alpha), \end{aligned}$$

where

$$p_0 := \frac{1}{M} \sum_{j=1}^M P_j(\Psi = j) \quad \text{and} \quad \alpha := \frac{1}{M} \sum_{j=1}^M P_j \left(\frac{dP_0}{dP_j} < \tau \right).$$

Accordingly, we have

$$\begin{aligned} &\max_{0 \leq j \leq M} P_j(\Psi \neq j) = \max \left\{ P_0(\Psi \neq 0), \max_{1 \leq j \leq M} P_j(\Psi \neq j) \right\} \geq \max \left\{ \tau M(p_0 - \alpha), \frac{1}{M} \sum_{j=1}^M P_j(\Psi \neq j) \right\} \\ &= \max\{\tau M(p_0 - \alpha), 1 - p_0\} \geq \min_{0 \leq p \leq 1} \max\{\tau M(p - \alpha), 1 - p\} = \frac{\tau M(1 - \alpha)}{1 + \tau M}. \end{aligned}$$

This completes the proof. □

This then gives us the desired result.

Theorem 8.2.2 (LeCam)**.** *Assume $\Theta = \{\theta_0, \theta_1, \dots, \theta_M\}$ is constructed such that*

(1) $d(\theta_j, \theta_k) \geq 2A > 0$ for any $0 \leq j < k \leq M$;

(2) there exists $\tau > 0$ and $0 < \alpha < 1$ such that

$$\frac{1}{M} \sum_{j=1}^M P_j \left(\frac{dP_0}{dP_j} \geq \tau \right) \geq 1 - \alpha.$$

We then have

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq \frac{\tau M}{1 + \tau M}(1 - \alpha).$$

8.2.1 Elementary information theory

Given Theorem 8.2.2, what is left is to determine the value of $P_j \left(\frac{dP_0}{dP_j} \geq \tau \right)$. This is, naturally, the problem of determining the distance between two probability measures P_0 and P_j, which plays a key role in information theory.

Of note, there is a track in probability to rethink everything in statistics from the perspective of the information theory. One motivating example is Stein's method, which is a flexible way in determining the rate for weak convergence in weak topology. People of interest should read Ramon von Handel's note (`https://www.princeton.edu/~rvan/ORF570.pdf`) as well as John Duchi's (`http://stanford.edu/class/stats311/Lectures/full_notes.pdf`).

Nevertheless, let's restrict our interest to studying $P_j\left(\frac{dP_0}{dP_j} \geq \tau\right)$. To this end, let's introduce the following concepts.

Definition 8.2.3. *For any two probability measures P and Q on $(\mathcal{X}, \mathcal{A})$, suppose ν is a σ-filed measure on $(\mathcal{X}, \mathcal{A})$ satisfying $P \ll \nu$ and $Q \ll \nu$. We define $p = dP/d\nu$ and $q = dQ/d\nu$, and*

- *f-divergence:*
$$D_f(P, Q) := \int f\left(\frac{dP}{dQ}\right) dQ.$$
- *Hellinger distance (H^2 as a special case of f-divergence by taking $f(x) = (\sqrt{x} - 1)^2$):*
$$H(P, Q) := \left(\int (\sqrt{p} - \sqrt{q})^2 d\nu\right)^{1/2} = \left(\int [\sqrt{dP} - \sqrt{dQ}]^2\right)^{1/2}.$$
The last equality is the famous Sheffe's theorem.
- *Total variation distance (taking $f(x) = |x - 1|/2$):*
$$V(P, Q) := \sup_{T \in \mathcal{A}} |P(T) - Q(T)| = \sup_{T \in \mathcal{A}} \left|\int_T (p - q) d\nu\right| = \frac{1}{2}\int |p - q| d\nu.$$
- *Kullback-Leibler (K-L) divergence (taking $f(x) = x \log x$):*
$$K(P, Q) = \int \log \frac{dP}{dQ} dP \quad \text{for } P \ll Q.$$
- *χ^2 divergence (by taking $f(x) = (x - 1)^2$):*
$$\xi^2(P, Q) := \int \left(\frac{dP}{dQ} - 1\right)^2.$$

There are a bunch of useful inequalities within the f-divergence family. However, due to the scope limit, let's focus on the one of the most useful, Pinsker's inequalities.

Lemma 8.2.4 (Pinsker's inequalities). *(1) $V(P, Q) \leq \sqrt{K(P, Q)/2}$.*

(2) If $P \ll Q$, then
$$\int \left|\log \frac{dP}{dQ}\right| dP := \int_{pq > 0} p \left|\log \frac{p}{q}\right| d\nu \leq K(P, Q) + \sqrt{2K(P, Q)},$$
and
$$\int \left(\log \frac{dP}{dQ}\right)_+ dP \leq K(P, Q) + \sqrt{K(P, Q)/2},$$
where $a_+ := \max(a, 0)$.

Proof. Left as a good exercise. □

8.2.2 Back to Theorem 8.2.2

Using the Pinsker's inequalities, we are now well equipped to provide a more user-friendly version of LeCam's theorem.

Theorem 8.2.5. *Assume that $M \geq 2$ and $\Theta = \{\theta_0, \dots, \theta_M\}$ satisfies*

(1) $d(\theta_j, \theta_k) \geq 2A > 0$ for any $0 \leq j < k \leq M$;

(2) $P_j \ll P_0$ and

$$\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0) \leq \alpha \log M,$$

with $0 < \alpha < 1/8$ (the upper bound $1/8$ is to make sure the final lower bound is larger than 0).

We then have

$$\inf_{\widetilde{\theta}_n} \sup_{\theta\in\Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

Proof. In Theorem 8.2.2, for any $0 < \tau < 1$, we have

$$P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) = P_j\left(\frac{dP_j}{dP_0} \leq \frac{1}{\tau}\right) = 1 - P_j\left(\log\frac{dP_j}{dP_0} > \log\frac{1}{\tau}\right) \geq 1 - \frac{1}{\log(1/\tau)}\int\left(\log\frac{dP_j}{dP_0}\right)_+ dP_j,$$

where the last inequality is due to Markov's inequality. We then can employ the second Pinsker's inequality to derive

$$P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) \geq 1 - \frac{1}{\log 1/\tau}\left[K(P_j, P_0) + \sqrt{K(P_j, P_0)/2}\right].$$

By Jensen's inequality and Condition (2),

$$\frac{1}{M}\sum_{j=1}^{M}\sqrt{K(P_j, P_0)} \leq \left(\frac{1}{M}\sum_{j=1}^{M} K(P_j, P_0)\right)^{1/2} \leq \sqrt{\alpha \log M}.$$

Accordingly, we have

$$\frac{1}{M}\sum_{j=1}^{M} P_j\left(\frac{dP_0}{dP_j} \geq \tau\right) \geq 1 - \frac{1}{\log 1/\tau}(\alpha \log M + \sqrt{\alpha \log M/2}),$$

which, combined with Theorem 8.2.2, yields

$$\inf_{\widetilde{\theta}_n} \sup_{\theta\in\Theta} P_\theta(d(\widetilde{\theta}_n, \theta) \geq A) \geq \frac{\tau M}{1+\tau M}\left(1 - \frac{1}{\log 1/\tau}(\alpha \log M + \sqrt{\alpha \log M/2})\right).$$

Picking $\tau = 1/\sqrt{M}$ minimizes the above term and finalizes the proof. □

8.2.3 Application to Example 8.1.1

Let's then show the upper bound in Theorem 8.1.2 is rate-optimal in the minimax sense via using Theorem 8.2.5.

Theorem 8.2.6. *Regarding the statistical problem in Example 8.1.2, we have*

$$\inf_{\widetilde{\theta}_n} \sup_{\theta \in \Theta_s} \mathbb{E}_\theta \|\widetilde{\theta}_n - \theta\|_2^2 \gtrsim \sigma^2 s \log(ep/s)/n.$$

To prove the result, we need a strong result in combinatorics. In detail, let

$$\Omega := \{\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m), \omega_i \in \{0,1\}\} = \{0,1\}^m.$$

We then have the Varshamov-Gilbert bound.

Lemma 8.2.7 (Varshamov-Gilbert bound)**.** *Let $m > 8$. Then there exists a subset $\{\omega^{(0)}, \ldots, \omega^{(M)}\}$ of Ω such that $\omega^{(0)} = (0, \ldots, 0)$ and*

$$\|\omega^{(j)} - \omega^{(k)}\|_0 \geq \frac{m}{8} \quad \text{for } 0 \leq j < k \leq M$$

(Here $\|\cdot\|_0$ represents the Hamming distance) and

$$M \geq 2^{m/8}.$$

Proof of Theorem 8.2.6. To prove this theorem, we use a variation of the Varshamov-Gilbert bound. We construct the parameter space Θ^0 as follows

$$\theta_0 = \mathbf{0}, \quad [\theta_{T_i}]_j = a \cdot I(j \in T_i),$$

where $T_i \in \mathcal{T}$, which includes all s distinct elements in $\{1, \ldots, p\}$ that pairwise have overlaps at, at most, $s/8$ position, and $a > 0$ is a value to be determined later. Using a variation of the Varshamov-Gilbert bound (a proof using probabilistic method like the proof of random projection. Maybe I will cover it in class), we can prove

$$\log M := \log |\Theta^0| \gtrsim s \log(p/s).$$

Furthermore, for Gaussian distribution, we know

$$KL(P_j, P_0) = \frac{1}{2}\|\boldsymbol{\theta}_j\|_2^2 = nsa^2/2\sigma^2$$

and

$$\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_k\|_2^2 \gtrsim sa^2.$$

Accordingly, picking $a^2 \asymp \sigma^2 \log(p/s)/n$ completes the proof. □

8.3 Fano and Assouad

I do not believe I will have time to introduce them, but I can promise you they are very useful. Actually, they are developed to handle the cases LeCam's method cannot or is very difficult to handle. People of interest should read Duchi's note.