ECE 598 - Fall 2025

Lecture 8: Method for Minimax Lower Bounds III

23rd September, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we will consider a new technique of deriving minimax lower bounds in statistical estimation problems, often called *Assouad's* method.

The idea in this scheme is a reduction from estimation problem (or more generally, some decision-making task) to that of *multiple binary hypothesis testing* problems. Equivalently, it reduces the general estimation problem to that of estimating a parameter in the *Hamming cube* of appropriate dimension. This reduction is obtained under a "separability assumption", that states that the loss function under consideration can be lower-bounded in terms of the Hamming distance between two associated binary vectors.

One important class of applications well-suited to this approach is *interactive or adaptive sensing* problems. For the non-adaptive case, there are generally no benefits (in terms of rates) to using Assouad's method over some version of Fano's method that we will study next week.

1 General Scheme

We work in the usual setting for minimax estimation. Let $\mathcal{X}$ denote some observation space, and let $\{P_\theta : \theta \in \Theta\}$ represent a statistical model/experiment with an associated parameter space (Θ, ρ), endowed with a pseudo-metric ρ. Furthermore, let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, denote a non-increasing function. Then, the minimax risk associated with this scenario be defined as

$$R(\Theta, L) \coloneqq \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{X \sim P_\theta} \left[L\big(\widehat{\theta}(X), \theta\big) \right], \quad \text{where} \quad L(\theta, \theta') = \Phi \circ \rho(\theta, \theta').$$

Recall, that $\widehat{\theta}$ represents both deterministic and possibly randomized estimators $P_{W|X}$.

Separability Condition. Assouad's method relies on a so-called *Hamming separability* condition that can be summarized in terms of the following two points:

- There exists a collection of distribution $\mathcal{V}_d \subset \mathcal{P}(\mathcal{X})$, with $\mathcal{V}_d = \{P_v : v \in \mathcal{H}_d\}$, with associated parameters $\{\theta_v : v \in \mathcal{H}_d\}$. Here $\mathcal{H}_d = \{-1, +1\}^d$ denotes the *Hamming cube* in d-dimensions.
- There exists an *encoding function* $E : \Theta \to \mathcal{H}_d$, that maps elements of Θ to the Hamming cube. Furthermore, this encoding map is such that $E(\theta_v) = v$ for all $v \in \mathcal{H}_d$.

Then, for any $\theta \in \Theta$ and $v \in \mathcal{H}_d$, we have

$$L\left(\theta, \theta_v\right) \geq 2\delta \sum_{j=1}^{d} \mathbf{1}\{E(\theta)[j] \neq v[j]\} = 2\delta \, d_H\left(E(\theta), v\right), \tag{1}$$

where $d_H(\cdot, \cdot)$ is the Hamming metric.

Example 1.1. *The condition* (1) *if often satisfied when $\Theta = \mathbb{R}^d$, ρ is the metric induced by an ℓ_p norm, and $\Phi(t) = |t|^p$ for $p \geq 1$. For example, consider the encoding map, $E : \mathbb{R}^d \to \mathcal{H}_d$, that is defined as follows:*

$$E(\theta)[j] = \operatorname{sign}(\theta[j]) = \begin{cases} +1, & \text{if } \theta[j] \geq 0, \\ -1, & \text{if } \theta[j] < 0. \end{cases}$$

We can think of E as the one-bit quantizer of the parameter θ. Define $\theta_v = \delta\, v$ for some $\delta > 0$ and for all $v \in \mathcal{H}_d$. Then, for any $\theta \in \Theta = \mathbb{R}^d$, and $v \in \mathcal{H}_d$, we have

$$\begin{aligned} \Phi \circ \rho(\theta, \theta_v) = \|\theta - \theta_v\|_p^p = \sum_{j=1}^d |\theta[j] - \theta_v[j]|^p \geq \delta^p \sum_{j=1}^d \mathbf{1}\{E(\theta)[j] \neq v[j]\} \\ = \delta^{q/p} d_H(E(\theta), v). \end{aligned}$$

In other words, for all values of $p \geq 1$, the problem described above satisfies a δ^p-Hamming-separability condition.

To state the key idea of this method, we need to introduce some notation. Recall that we have a collection of distributions $\{P_v : v \in \mathcal{H}_d\}$. Suppose $V \sim \mathrm{Uniform}(\mathcal{H}_d)$ is a random variable that is uniformly distributed over the Hamming cube, and suppose our observation $X \sim P_V$. That is, the conditional distribution of X given $V = v$ is P_v for every $v \in \mathcal{H}_d$. The unconditional distribution of X is the uniform mixture of all P_v, with $v \in \mathcal{H}_d$. We will use P_{+j} (resp. P_{-j}) to denote the conditional distribution of X, given that $V[j] = +1$ (resp. $V[j] = -1$).

Theorem 1.2. *Suppose the 2δ-Hamming separation condition is satisfied. Then, we have the following:*

$$R(\Theta, L) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{X \sim P_\theta} \left[\Phi \circ \rho \left(\widehat{\theta}(X), \theta \right) \right] \geq \delta \sum_{j=1}^d \left(1 - TV(P_{+j}, P_{-j}) \right). \tag{2}$$

Remark 1.3. Informally, this result says the following: the minimax risk can be lower-bounded by the minimax risk over a subset of parameters $\{\theta_v = \delta v : v \in \mathcal{H}_d\}$. This lower minimax risk cannot be smaller than the average risk, with the parameter $V \sim \mathrm{Uniform}(\mathcal{H}_d)$. Finally, the average risk can be further lower-bounded by the error of d hypothesis testing problems about the sign of the j^{th} coordinate, with $j \in [d]$.

The separability condition essentially allows us to independently apply the two-point method to each coordinate of V, resulting in the bound stated in (2).

Proof of Theorem 1.2. The starting point of the proof is the usual step that bounds the maximum with the average. In this instance, we consider $V \sim \mathrm{Uniform}(\mathcal{H}_d)$, and observe that

$$\begin{aligned} R(\Theta, L) &\geq \inf_{\widehat{\theta}} \mathbb{E}_V \left[\mathbb{E}_{X \sim P_V} \left[\Phi \circ \left(\widehat{\theta}(X), \theta(P) \right) \right] \right] && (\max \geq \text{average}) \\ &\geq 2\delta \inf_{\widehat{\theta}} \mathbb{E}_V \left[\mathbb{E}_{X \sim P_V} \left[d_H \left(E(\widehat{\theta}(X)), V \right) \right] \right] && (\text{Hamming-Separation}) \end{aligned}$$

Now, the estimator $\widehat{\theta}$ only influences the last expression through its composition with E. Hence, we can get a further lower bound by taking an infimum over all $\psi : \mathcal{X} \to \mathcal{H}_d$

$$R(\Theta, L) \geq 2\delta \inf_{\psi:\mathcal{X}\to\mathcal{H}_d} \mathbb{E}_V \left[\mathbb{E}_{X\sim P_V}\left[d_H\left(\psi(X), V\right)\right]\right].$$

On expanding the RHS, we get

$$\begin{aligned}
R(\Theta, L) &\geq 2\delta \inf_{\psi} \frac{1}{2^d} \sum_{v\in\mathcal{H}_d} \sum_{j=1}^{d} P_v(\psi(X)[j] \neq v[j]) \\
&\geq 2\delta \sum_{j=1}^{d} \inf_{\psi_j:\mathcal{X}\to\{\pm 1\}} \frac{1}{2^d} \left(\sum_{v:v[j]=+1} P_v\left(\psi_j(X) = -1\right) + \sum_{v:v[j]=-1} P_v\left(\psi_j(X) = +1\right) \right) \\
&= 2\delta \sum_{j=1}^{d} \inf_{\psi_j} \frac{1}{2} \left(P_{+j}\left(\psi_j(X) = -1\right) + P_{-j}\left(\psi_j(X) = +1\right)\right) \qquad (3)
\end{aligned}$$

In the last equality, we used the fact that for any event G, we have

$$\begin{aligned}
P_{+j}(G) &= \mathbb{P}\left(G \mid V[j] = +1\right) = \frac{\mathbb{P}\left(G \cap \{V[j] = +1\}\right)}{\mathbb{P}(V[j] = +1)} \\
&= 2\mathbb{P}\left(G \cap \{V[j] = +1\}\right) && \text{(since } \mathbb{P}(V[j] = +1) = 1/2\text{)} \\
&= 2 \sum_{v:v[j]=+1} \mathbb{P}\left(G \cap \{V = v\}\right) \\
&= 2 \sum_{v:v[j]=+1} \mathbb{P}(V = v) P_v(G) \\
&= 2\, \frac{1}{2^d} \sum_{v:v[j]=+1} P_v(G). && \text{(since } V \sim \text{Uniform}(\mathcal{H}_d)\text{)}
\end{aligned}$$

A exactly similar calculation gives us the P_{-j} term in (3).

Picking up from (3), we have

$$\begin{aligned}
R(\Theta, L) &\geq \delta \sum_{j=1}^{d} \inf_{\psi_j} \left(1 - \left(P_{+j}(\psi_j(X) = +1) + P_{-j}(\psi_j(X) = +1)\right)\right) \\
&= \delta \sum_{j=1}^{d} \left(1 - \sup_{\psi_j} \left(P_{+j}(\psi_j(X) = +1) - P_{-j}(\psi_j(X) = +1)\right)\right) \\
&= \delta \sum_{j=1}^{d} \left(1 - \sup_{G\in\mathcal{F}_{\mathcal{X}}} \left(P_{+j}(G) - P_{-j}(G)\right)\right) \\
&= \delta \sum_{j=1}^{d} \left(1 - TV(P_{+j}, P_{-j})\right).
\end{aligned}$$

Note that in the second equality above, we have used the fact that we can restrict our attention to non-randomized hypothesis tests ψ_j, and thus a sup over all ψ_j is equivalent to a sup over all measurable sets $G \in \mathcal{F}_{\mathcal{X}}$. This completes the proof. □

Some Simplifications. In order to apply this technique to a given problem, we will need to get some control over the terms $TV(P_{+j}, P_{-j})$ for $j \in [d]$. Both P_{+j} and P_{-j} are uniform mixtures of 2^{d-1} distributions each, which implies

$$\left\| \frac{1}{2^{d-1}} \sum_{v:v[j]=+1} P_v - \sum_{u:u[j]=-1} P_u \right\|_{TV} \leq \frac{1}{2^{d-1}} \frac{1}{2^{d-1}} \sum_{v:v[j]=1} \sum_{u:u[j]=-1} TV(P_v, P_u)$$

due to convexity of the total-variation distance in each of its arguments. A simple, but sometimes too crude, idea is to use the bound $TV(P_v, P_u) \leq \max_{u,v} TV(P_v, P_u)$, to get

$$R(\Theta, L) \geq \delta \left(1 - \max_{\substack{v:v[j]=+1 \\ u:u[j]=-1}} TV(P_v, P_u) \right). \tag{4}$$

Another approach uses the Cauchy-Schwarz inequality to get

$$\begin{aligned} \sum_{j=1}^{d} (1 - TV(P_{+j}, P_{-j})) &= d - \sum_{j=1}^{d} 1 \times TV(P_{+j}, P_{-j}) \\ &\geq d - \sqrt{d} \left(\sum_{j=1}^{d} TV(P_{+j}, P_{-j})^2 \right)^{1/2} \\ &\geq d \left(1 - \left(\frac{1}{d2^d} \sum_{j=1}^{d} \sum_{v \in \mathcal{H}_d} TV(P_{v,+j}, P_{v,-j})^2 \right)^{1/2} \right). \end{aligned} \tag{5}$$

In the last inequality, we used the convexity of TV distance, and the notation $P_{v,+j}$ is equal to P_v if $v[j] = +1$, otherwise it is equal to $P_{v'}$, where $v'[j] = +1$ and $v'[i] = v[i]$ for all $i \neq j$.

Finally, often it may be easier to work with other distance/divergence measures, and we can use the following inequalities to replace the occurrence of every $TV(\cdot, \cdot)$:

$$TV(P, Q)^2 \leq H^2(P, Q) \leq D_{\mathrm{KL}}(P \parallel Q) \leq \log\left(1 + \chi^2(P \parallel Q)\right).$$

Plugging either of these into (2), (4), or (5) might result in a more tractable form depending on the problem structure.

2 Application to Linear Regression with Adaptive Sensing

Consider a regression problem with the parameter space $\Theta = \mathbb{R}^d$, and let ρ denote the ℓ_2 norm with $\Phi(t) = t^2$. Now suppose we have n observations of the form

$$Y_t = \langle A_t, \theta \rangle + \sigma \varepsilon_t, \quad \text{where} \quad A_t \text{ is } \mathcal{F}_{t-1}\text{-measurable with } \|A_t\|_2 \leq 1, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, 1).$$

In other words, we collect n observations via adaptive measurements under a power constraint. The goal is to use $\boldsymbol{X} = (Y_1, \ldots, Y_n)$ to construct an estimate of the unknown parameter θ. The minimax risk in this case is

$$R_n \equiv R_n(\Theta, L) = \inf_{\widehat{\theta}, \mathcal{A}} \sup_{\theta \in \Theta} \|\widehat{\theta}(\boldsymbol{X}) - \theta\|_2^2,$$

where $\mathcal{A}$ denotes the adaptive sensing policy. In order to apply Assouad's method to this problem, we will select the encoding map: $E : \Theta \to \mathcal{H}_d$, such that $E(\theta)[j] = \mathrm{sign}(\theta[j])$ for all $j \in [d]$. Then, from Example 1.1, we know that with $\theta_v = \delta v$ for all $v \in \mathcal{H}_d$, we have

$$\|\theta - \theta_v\|_2^2 \geq \delta^2 d_H(E(\theta), v).$$

Hence, we have a δ^2-Hamming separation condition. Now, in this problem, we will use $P_\theta \equiv P_{\theta,\mathcal{A}}$ to denote the joint distribution of the random variables:

$$(A_1, Y_1, A_2, Y_2, \ldots, A_n, Y_n) \sim P_\theta \equiv P_{\theta,\mathcal{A}}.$$

Recall that $\mathcal{A}$ represents the adaptive sensing strategy, which is a sequence of conditional distributions $P_{A_t|A^{t-1},Y^{t-1}}$ for $t \in [n]$.

Using Theorem 1.2, we get that

$$R_n \geq \frac{\delta^2}{2} \sum_{j=1}^{d} (1 - TV(P_{+j}, P_{-j})) \overset{\text{Pinsker's}}{\geq} \frac{\delta^2}{2} \sum_{j=1}^{d} \left(1 - \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P_{+j} \parallel P_{-j})}\right). \tag{6}$$

The next step is to get an upper bound on the relative entropy between P_{+j} and P_{-j}. To do this, we need to introduce the following terms:

$$v^{(j)} = \begin{cases} v[i], & \text{if } i \neq j \\ -v[i], & \text{if } i = j, \end{cases} \quad v^{(+j)} = \begin{cases} v[i], & \text{if } i \neq j \\ +1, & \text{if } i = j, \end{cases} \quad v^{(-j)} = \begin{cases} v[i], & \text{if } i \neq j \\ -1, & \text{if } i = j, \end{cases}$$

Then, we have

$$\begin{aligned} D_{\mathrm{KL}}(P_{+j} \parallel P_{-j}) &= D_{\mathrm{KL}}\left(\frac{1}{2^{d-1}} \sum_{v:v[j]=+1} P_v \parallel \frac{1}{2^{d-1}} \sum_{v:v[j]=-1} P_v\right) \\ &= D_{\mathrm{KL}}\left(\frac{1}{2^d} \sum_{v \in \mathcal{H}_d} P_{v^{(+j)}} \parallel \frac{1}{2^d} \sum_{v \in \mathcal{H}_d} P_{v^{(-j)}}\right) \end{aligned}$$

The last term is simply the relative entropy between two mixture distributions, and due to the joint-convexity property, it implies

$$D_{\mathrm{KL}}(P_{+j} \parallel P_{-j}) \leq \frac{1}{2^d} \sum_{v \in \mathcal{H}_d} D_{\mathrm{KL}}(P_{v^{(+j)}} \parallel P_{v^{(-j)}}) \leq \max_{v \in \mathcal{H}_d} D_{\mathrm{KL}}(P_{v^{(+j)}} \parallel P_{v^{(-j)}}).$$

Now, let $f_{\sigma^2}(y, \mu)$ denote the cdf of a Normal distribution with mean μ and variance σ^2 at y, and observe that

$$\begin{aligned} D_{\mathrm{KL}}(P_{v^{(+j)}} \parallel P_{v^{(-j)}}) &= \mathbb{E}_{P_{v^{(+j)}}}\left[\log\left(\prod_{t=1}^{n} \frac{p_{A_t|H_{t-1}} f_{\sigma^2}(Y_t, \langle \theta_{v^{(+j)}}, A_t \rangle)}{p_{A_t|H_{t-1}} f_{\sigma^2}(Y_t, \langle \theta_{v^{(-j)}}, A_t \rangle))}\right)\right] \\ &= \mathbb{E}_{P_{v^{(+j)}}}\left[\sum_{t=1}^{n} \log\left(\frac{f_{\sigma^2}(Y_t, \langle \theta_{v^{(+j)}}, A_t \rangle)}{f_{\sigma^2}(Y_t, \langle \theta_{v^{(-j)}}, A_t \rangle))}\right)\right] \\ &= \sum_{t=1}^{n} \mathbb{E}_{P_{v^{(+j)}}}\left[\mathbb{E}_{P_{v^{(+j)}}}\left[\frac{\langle \theta_{v^{(+j)}} - \theta_{v^{(-j)}}, A_t \rangle^2}{2\sigma^2}\middle| A_t\right]\right] \leq n\frac{4\delta^2}{2\sigma^2}, \end{aligned} \tag{7}$$

where in the last inequality we used Cauchy-Schwarz to get $\langle \theta_{v(+j)} - \theta_{v(-j)}, A_t \rangle^2 \leq \|\theta_{v(+j)} - \theta_{v(-j)}\|^2 \|A_t\|^2$, and then the power constraint $\|A_t\| \leq 1$ (almost surely).

Using (7) in (6), we get

$$R_n \geq \frac{d\,\delta^2}{2}\left(1 - \frac{2n\delta^2}{\sigma^2}\right).$$

Finally, we complete the proof by selecting $\delta^2 = \sigma^2/4n$ which gives us

$$R_n \geq \frac{\sigma^2 d}{16n}.$$

Remark 2.1. Note that this result tells us that in a minimax sense, adaptive sensing does not lead to an improvement in estimation performance without further assumptions or structure. For example, as an extreme case, suppose we relaxed the power constraint, and assumed that the true θ has only one non-zero element, then this task reduces to a noisy binary testing problem. In this case, adaptive methods can lead to significantly better error.