ECE 598 - Fall 2025

Lecture 10: Fano's Method for Minimax Lower Bounds

30th September, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we begin our discussion of Fano's method for deriving minimax lower bounds. Like Assouad's method, this technique also proceeds via a reduction to M-ary hypothesis testing. However, unlike Assouad, it requires fewer structural assumptions (no indexing by Hamming cube or separability of loss functions), and thus it is more generally applicable.

1 General Scheme

For simplicity, we present the idea behind this technique in the context of minimax estimation. As always, we let $\mathcal{X}$ denote some observation space, $\{P_\theta : \theta \in \Theta\}$ denotes a statistical model indexed by a psedometric space (Θ, ρ). For some non-decreasing function $\Phi : [0, \infty) \to [0, \infty)$, we define a loss function $L(\theta, \theta') = \Phi \circ \rho(\theta, \theta')$, and the risk associated with any estimator $\widehat{\theta}$ is defined as

$$R(\widehat{\theta}(X), \theta) = \mathbb{E}_{X \sim P_\theta}\left[L(\widehat{\theta}(X), \theta)\right].$$

The minimax risk in this scenario is equal to

$$R^*(\Theta, \Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{X \sim P_\theta}\left[L(\widehat{\theta}(X), \theta)\right].$$

Reduction to M-ary hypothesis testing. For some $M \geq 2$, let $\Theta_M = \{\theta_i : i \in [M]\}$ denote a finite subset of the parameter space, and let $\{P_i \equiv P_{\theta_i} : i \in [M]\}$ denote the corresponding distributions. Furthermore, assume that the subset Θ_M is 2δ-separated in ρ; that is, $\rho(\theta_i, \theta_j) \geq 2\delta$ for all $i \neq j \in [M]$. Then, we have the usual lower bound:

$$R^* \geq \inf_{\widehat{\theta}} \max_{i \in [M]} \mathbb{E}_{X \sim P_i}\left[L(\widehat{\theta}(X), \theta_i)\right] \geq \inf_{\widehat{\theta}} \ \mathbb{E}_V\left[\mathbb{E}_{X \sim P_V}\left[L(\widehat{\theta}(X), \theta_V)\right]\right],$$

where $V \sim P_V$ is any $[M]$-valued random variable. Introduce the term

$$w(\delta) = \inf_{\theta, \theta' : \rho(\theta, \theta') \geq \delta} L(\theta, \theta'),$$

and observe that

$$L(\theta_V, \widehat{\theta}(X)) \geq w(\delta)\, \mathbf{1}\{\rho(\theta_V, \widehat{\theta}(X)) \geq \delta\}. \tag{1}$$

Given any estimator $\widehat{\theta}$, we can turn it into a minimum distance estimator of V as follows:

$$\hat{\psi}(X) = \operatorname*{argmin}_{i \in [M]} \rho(\theta_i, \widehat{\theta}(X)),$$

breaking ties in a deterministic manner. This implies that we can rewrite (1) as

$$L(\theta_V, \widehat{\theta}(X)) \geq w(\delta)\, \mathbf{1}\{\hat{\psi}(X) \neq V\},$$

which leads to

$$R^* \geq w(\delta) \inf_{\hat{V}:\mathcal{X}\to[M]} \mathbb{P}_{X,V}\left(\hat{V}(X) \neq V\right). \tag{2}$$

In other words, we can lower bound the minimax estimation risk with $w(\delta)$ times the probability of error in the M-ary hypothesis testing problem.

Fano's Inequality for M-ary Hypothesis Testing. So we are interested in analyzing the following Markov chain

$$V \longrightarrow X \longrightarrow \hat{V}.$$

Let p_e denote probability of error $\mathbb{P}(\hat{V} \neq V)$ associated with any estimator (or M-ary hypothesis test) $\hat{V}$. Then, Fano's inequality tells us that

$$p_e \log(M-1) + h_2(p_e) \geq H(V|\hat{V}) \geq H(V|X),$$

where h_2 denotes the binary entropy. The second inequality above is due to the data processing inequality. Plugging this into (2), we get the general Fano's lower bound.

Theorem 1.1. *Suppose $\{P_i \equiv P_{\theta_i} : i \in [M]\}$ denote a collection of distributions with associated parameters $\{\theta_i : i \in [M]\}$ satisfying $\rho(\theta_i, \theta_j) \geq 2\delta$ for all $i \neq j \in [M]$. Then, with V denoting any $[M]$-valued random variable, and $X|(V = v) \sim P_v$, we have the following:*

$$R^*(\Theta, L) \geq w(\delta)\left(\frac{H(V|X) - \log 2}{\log M}\right) \quad \text{where} \quad w(\delta) = \inf_{\theta,\theta':\rho(\theta,\theta')\geq\delta} L(\theta, \theta'). \tag{3}$$

For the special case of $V \sim \mathrm{Uniform}([M])$, then we can get the following lower bound:

$$\mathcal{R}(\mathcal{P}, L) \geq w(\delta)\left(1 - \frac{I(V;X) + \log 2}{\log M}\right). \tag{4}$$

Proof. The bound in (3) follows from (2) using the fact that $h_2(p_e) \leq \log 2$, $\log(M-1) \leq \log M$ and on rearranging. To obtain (4), simply use the fact that when $V \sim \mathrm{Uniform}([M])$, then $H(V) = \log M$, and thus $I(V;X) = H(V) - H(V|X) = \log M - H(V|X)$. □

Thus, in order to apply Theorem 1.1 to establish minimax rates for estimation problems, we need to do the following:

(a) Identify a collection of M distributions which are pairwise 2δ separated

(b) For some $[M]$-valued r.v. V, we need a lower bound on $H(V|X)$. Or with $V \sim \mathrm{Uniform}([M])$, we need an upper bound on $I(V;X)$.

We discuss some details of step (a) in Section. 1.1. For step (b), a convenient approach is to use the relative entropy based definition of mutual information. In particular, we know that

$$I(V;X) = D_{\mathrm{KL}}(P_{XV} \parallel P_X \times P_V) = D_{\mathrm{KL}}(P_{X|V} \parallel P_X | P_V)$$

When $V \sim \mathrm{Uniform}([M])$, we have $P_X = (1/M)\sum_{i=1}^M P_i$, which implies that

$$\begin{aligned} I(V;X) = D_{\mathrm{KL}}(P_{X|V} \parallel P_X \mid P_V) &= \frac{1}{M}\sum_{i=1}^M D_{\mathrm{KL}}(P_i \parallel P_X) \\ &\leq \frac{1}{M^2}\sum_{i=1}^M \sum_{j=1}^M D_{\mathrm{KL}}(P_i \parallel P_j) && \text{(convexity of } D_{\mathrm{KL}}) \\ &\leq \max_{i,j \in [M]} D_{\mathrm{KL}}(P_i \parallel P_j). \end{aligned}$$

Thus, it suffices to construct a collection of distributions whose worst case relative entropy over all pairs can be controlled. We summarize this discussion in our next result.

Corollary 1.2. *Suppose* $V \sim \mathrm{Uniform}[M]$*, and let* $\{P_i : i \in [M]\}$ *denote a collection of distributions in* $\mathcal{P}$ *satisfying*

$$\rho(\theta_i, \theta_j) \geq 2\delta, \quad D_{KL}(P_i \parallel P_j) \leq f(\delta), \quad \text{and} \quad \log M \geq 2(f(\delta) + \log 2), \tag{5}$$

for some function $f : (0,\infty) \to (0,\infty)$*. Then, we have*

$$\mathcal{R}(\mathcal{P}, L) \geq \frac{w(\delta)}{2}.$$

Remark 1.3. (5) states the properties of the "hard subset of problems" that characterize the minimax rate of the estimation problem: we should be able to identify sufficiently many distributions ($\log M \geq 2(f(\delta) + \log 2)$) that are well separated in parameters space ($\rho(\theta_i, \theta_j)$), but are difficult to distinguish statistically ($D_{\mathrm{KL}}(P_i \parallel P_j) \leq f(\delta)$).

1.1 Construction of 2δ separated distribution sets

In most applications, in order to construct the collection of 2δ separated disrtibutions $\mathcal{V} = \{P_v : v \in [M]\}$, we will follow the following approach:

- We will construct a collection of distributions associated with an appropriate covering set of the parameter space Θ (or a small region of the parameter space).
- Then, as needed, we will scale this set with a parameter δ.

The following two results will be useful in constructing the packing sets.

Lemma 1.4. *Let $B(0,1) = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$ denote the ℓ_2 unit ball in $\mathbb{R}^d$. Then, there exists a set $\mathcal{V} \subset B(0,1)$ such that*

- *For any $v \neq v' \in \mathcal{V}$, we have $\|v - v'\|_2 \geq 1/2$.*
- *The cardinality of $\mathcal{V}$ satisfies $|\mathcal{V}| \geq 2^d$.*

Proof. The proof of this statement follows by standard "volume arguments". In particular, suppose $\mathcal{V}$ is a maximal set with the property that $\|v - v'\|_2 \geq 1/2$. Hence, for any $w \in B(0,1)$ there must exist a $v \in \mathcal{V}$ such that $\|v - w\|_2 < 1/2$. For if this were not true, then we could append this w to $\mathcal{V}$ while maintaining the separation condition, thus violating the "maximal" assumption on $\mathcal{V}$. Another way of stating this condition is that

$$\cup_{v \in \mathcal{V}} B(v, 1/2) \supset B(0,1) \qquad \Longrightarrow \qquad \mathrm{Vol}\left(\cup_{v \in \mathcal{V}} B(v, 1/2)\right) \geq \mathrm{Vol}\left(B(0,1)\right),$$

where Vol denote the usual volume (i.e., Lebesgue measure) of the sets involved. Since the Lebesgue measure is translation invariant and subadditive, we know that $\mathrm{Vol}(B(v,1/2)) = \mathrm{Vol}(B(0,1/2))$ for all $v \in \mathcal{V}$, and

$$|\mathcal{V}|\, C_d (1/2)^d = |\mathcal{V}|\, \mathrm{Vol}(B(0,1/2)) \geq \mathrm{Vol}\left(\cup_{v \in \mathcal{V}} B(v, 1/2)\right) \geq \mathrm{Vol}\left(B(0,1)\right) = C_d 1^d,$$

for some d-dependent constant C_d. This leads to the required conclusion that $|\mathcal{V}| \geq 2^d$. □

Another important case is when we need to construct packing sets of the Hamming cube $\mathcal{H}_d = \{-1, +1\}^d$.

Lemma 1.5 (Gilbert-Varshamov)**.** *There exists a subset $\mathcal{V} \subset \mathcal{H}_d = \{-1, +1\}^d$ satisfying*

- *For any $v \neq v' \in \mathcal{V}$, we have $d_H(v, v') \geq d/4$.*
- *The cardinality of $\mathcal{V}$ satisfies $|\mathcal{V}| \geq e^{d/8}$.*

In other words there exists a subset of $\mathcal{H}_d$ of size growing exponentially in d with pairwise separation at least $d/4$ in Hamming metric.

Proof. Assume that $\mathcal{V}$ is a maximal $d/4$-packing set in Hamming metric. Then, as before, we have

$$\cup_{v \in \mathcal{V}} B_H(v, d/4) \supset \mathcal{H}_d,$$

where $B_H(v, d/4) = \{v' \in \mathcal{H}_d : d_H(v, v') \leq d/4\}$. We can also show that $|B_H(v, d/4)|$ is independent of v, which means that for some fixed arbitrary element v_0 of $\mathcal{V}$, we have $|\mathcal{V}||B_H(v_0, d/4)| \geq |\mathcal{H}_d| = 2^d$. To conclude the proof, we will derive an appropriate upper bound on $|B_H(v_0, d/4)$.

Let $S_1, \ldots, S_d$ denote i.i.d. $\mathrm{Bernoulli}(1/2)$ random variables, using which we can define the random variable V such that

$$V[j] = v_0[j] \oplus S_j, \quad \text{for all} \quad j \in [d].$$

It is easy to verify that V is uniformly distributed over the Hamming cube. Hence, we have

$$\mathbb{P}\left(V \in B_H(v_0, d/4)\right) = \frac{|B_H(v_0, d/4)|}{2^d} = \mathbb{P}\left(\sum_{j=1}^d S_j \leq d/4\right) = \mathbb{P}\left(\sum_{j=1}^d S_j \geq 3d/4\right).$$

The last inequality uses the fact that $S_j \overset{d}{=} 1 - S_j$ for all $j \in [d]$. An application of Chernoff bound then implies

$$\frac{|B_H(v_0, d/4)|}{2^d} \leq e^{-\lambda 3d/4}\mathbb{E}\left[e^{\lambda \sum_{i=1}^d S_i}\right] = e^{-\lambda 3d/4}\left(\frac{1+e^\lambda}{2}\right)^d.$$

It turns out that the upper bound is optimized at $\lambda = \log 3$, giving the bound

$$|B_H(v_0, d/4)| \leq 4^d 3^{-3d/4}.$$

Plugging this into the bound $|\mathcal{V}| \geq 2^d/|B_H(v_0, d/4)|$, gives us

$$|\mathcal{V}| \geq 2^d 3^{3d/4} 4^{-d} = \exp\left(d \log\left(\frac{2 \times 3^{3/4}}{4}\right)\right) \geq e^{d/8}.$$

This completes the proof. □

2 Nonparametric regression in sup norm

Let $\Theta \equiv \Theta(C, \gamma)$ denote the set of all Hölder smooth functions (with parameters $C \in (0, \infty)$ and $\gamma \in (0, 1]$) supported on the unit interval $[0, 1]$; that is, for every $x, x' \in [0, 1]$, and $\theta \in \Theta$, we have

$$|\theta(x) - \theta(x')| \leq C|x - x'|^\gamma.$$

Suppose we have n observations of the form

$$Y_i = \theta(x_i) + \sigma\varepsilon_i, \quad \text{where} \quad x_i = i/n,\ (\varepsilon_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, 1).$$

For simplicity, we have assumed a "fixed design" (with $x_i = i/n$), but the same idea can be extended to the case of $X_i \sim P_X$ as well, for sufficiently regular P_X.

The observation space in this problem is $\mathcal{X} = \prod_{i=1}^n [0, 1] \times \mathbb{R}$. The statistical model $\{P_\theta : \theta \in \Theta(C, \gamma)\}$ consists of distributions of the form

$$P_\theta \equiv P_{\theta, X^n, Y^n} = \otimes_{i=1}^n \left(\delta_{x_i} \times N(\theta(x_i), \sigma^2)\right).$$

Thus, given the observations Y^n, our goal is to construct an estimate $\widehat{\theta}(Y^n)$ that achieves a small worst-case risk

$$R_n(\widehat{\theta}, \Theta) = \sup_{\theta \in \Theta} \mathbb{E}\left[\|\widehat{\theta}(Y^n) - \theta\|_2^2\right].$$

This corresponds to $L(\theta, \theta') = \Phi \circ \rho(\theta, \theta')$ with ρ denoting the metric induced by the L^2-norm (w.r.t. to the Lebesgue measure on $[0, 1]$), and Φ denotes the mapping $t \mapsto t^2$.

For simplicity, we will focus on the case of $\gamma = 1$, which corresponds to Lipschitz functions.

2.1 Converse

To establish a converse result using Fano's method, we need to construct an appropriate "hard" class of problems. To do this, we will proceed as follows:

- For some integer m to be decided later, partition the unit interval into m equal sub-intervals $I_i = [i - 1/n, i/n)$ for $i \in [n-1]$, and $I_n = [1 - 1/n, 1]$.

- Let us consider a bump function $\phi : [0,1] \to [0,1]$ defined as

  $$\phi(u) = \begin{cases} 3u, & \text{if } u \in [0, 1/3), \\ 1, & \text{if } u \in [1/3, 2/3), \\ 1 - 3u, & \text{if } u \in [2/3, 1]. \end{cases}$$

  We can verify that this function is Lipschitz with parameter $C = 3$.

- For any $j \geq 1$, define the scaled-and-shifted bump function ϕ_j as

  $$\phi_j(x) = \phi\left(m\left(x - \frac{j-1}{m}\right)\right).$$

  This function is supported on the interval I_j, and can be verified to be Lipschitz continuous with parameter $3m$. Finally, we define the finite subset $\Theta_m = \{\theta_j : j \in [m]\}$ as

  $$\theta_j(x) = h\phi_j(x) = h\phi\left(m\left(x - \frac{j-1}{m}\right)\right), \quad \text{with} \quad h > 0, \quad 3mh \leq C.$$

- Each function θ_j in Θ_m is a scaled-and-shifted bump function that lies in Θ, and also satisfies the separation condition

  $$\rho(\theta_i, \theta_j) = \|\theta_i - \theta_j\|_\infty = h, \quad \text{for all} \quad i \neq j \in [m].$$

  This simply follows from the fact that θ_i and θ_j have disjoint supports for $i \neq j$. Thus, we have constructed a 2δ-separated set with $\delta = h/2$.

- To apply Fano's method, we now need to control the statistical distinguishability via the pairwise relative entropy of the associated distributions. Observe that for any θ_j, θ_k, the distributions P_j, P_k have the form

  $$P_{\theta_j} = \otimes_{i=1}^n \left(\delta_{x_i} \times N(\theta_j(x_i), \sigma^2)\right), \quad \text{and} \quad P_{\theta_k} = \otimes_{i=1}^n \left(\delta_{x_i} \times N(\theta_k(x_i), \sigma^2)\right).$$

  Thus, the relative entropy between them, which we denote by D_{jk} is completely governed by the ℓ_2 norm between the "mean-vectors" $\mu_j = [\theta_j(1/n), \ldots, \theta_j(n/n)]$ and $\mu_k = [\theta_k(1/n), \ldots, \theta_k(n/n)]$; that is,

  $$D_{jk} = D_{\mathrm{KL}}(P_{\theta_j} \parallel P_{\theta_k}) = \sum_{i=1}^n \frac{(\theta_j(x_i) - \theta_k(x_i))^2}{2\sigma^2} = \frac{1}{2\sigma^2}\|\mu_j - \mu_k\|_2^2.$$

Since θ_j and θ_k have disjoint supports and take the maximum value of h, we have

$$D_{jk} \leq \frac{1}{2\sigma^2} h^2 \times 2 \times \left(\frac{n}{m} + 1\right) \leq \frac{4h^2 n}{2\sigma^2 m}.$$

Plugging this into the maximum relative entropy form of Fano's lower bound, we get

$$R_n^* \geq \frac{h}{2}\left(1 - \frac{2h^2 n/\sigma^2 m + \log 2}{\log m}\right).$$

- To complete the proof, we need to choose an appropriate value of m and h. For a fixed m, suppose we choose h to ensure that

  $$\frac{2h^2 n}{\sigma^2 m} = \frac{1}{2}\log m \qquad \Longrightarrow \qquad h = \frac{\sigma}{2}\sqrt{\frac{m \log m}{n}}.$$

  Assuming m is large enough to ensure that $\log m > 4 \log 2$ (i.e., $m \geq 16$), then with this h, the lower bound becomes

  $$R_n^* \geq \frac{\sigma}{4}\sqrt{\frac{m \log m}{n}} \times \left(1 - \frac{1}{2} - \frac{1}{4}\right) = \frac{\sigma}{16}\sqrt{\frac{m \log m}{n}}.$$

- It remains to make a choice of the parameter m. We want to choose m to be as large as possible while ensuring that the constraint on the Lipschitz constants of θ_j for $j \in [m]$ are satisfied; that is,

  $$3mh = 3m\frac{\sigma}{2}\sqrt{\frac{m \log m}{n}} \leq C \qquad \Longrightarrow \qquad m^3 \leq \frac{4C^2 n}{\log m \sigma^2}$$

  Without taking the constants into account, a suitable value of m turns out to be

  $$m^* \asymp \left(\frac{n}{\log n}\right)^{1/3}, \quad \text{which implies} \quad R_n^* \gtrsim \left(\frac{\log n}{n}\right)^{1/3}.$$

Remark 2.1. Observe that the loss function that we considered does not have the additive structure required to ensure the Hamming separability condition. As a result Assoaud's method would yield a suboptimal lower bound in this problem, as it would not be able to capture the logarithmic factor.

2.2 Achievability

A simple piecewise constant estimator can achieve the optimal rate for this problem. The idea is to use the same partition we used in the converse derivation and define the estimator $\widehat{\theta}$ as

$$\widehat{\theta}(x) = \frac{1}{n_j}\sum_{i : x_i \in I_j} Y_i, \;\; \text{if } x \in I_j, \quad \text{where} \quad n_j = |\{i : x_i \in I_j\}|.$$

The error at any point $x \in I_j$ of this estimator satisfies

$$|\widehat{\theta}(x) - \theta(x)| = \left| \frac{1}{n_j} \sum_{i: x_i \in I_j} (\theta(x_i) + \sigma \varepsilon_i) - \theta(x) \right| \leq \sigma \bar{\varepsilon}_j + |\bar{\mu}_j - \theta(x)|\,,$$

where we use the notation

$$\bar{\varepsilon}_j = \frac{1}{n_j} \sum_{i: x_i \in I_j} \varepsilon_j, \sim N(0, \frac{1}{n_j}), \quad \text{and} \quad \bar{\mu}_j = \frac{1}{n_j} \sum_{i: x_i \in I_j} \theta(x_i).$$

Hence, the estimation risk is upper bound by

$$\mathbb{E}_\theta \left[\|\widehat{\theta}(X^n) - \theta\|_\infty \right] \leq \sigma \mathbb{E} \left[\max_{j \in [m]} \bar{\varepsilon}_j \right] + \max_j \sup_{x \in I_j} |\bar{\mu}_j - \theta(x)|. \tag{6}$$

Now, a standard fact about Gaussian random variables $Z_1, \ldots, Z_m \overset{i.i.d.}{\sim} N(0, a^2)$ is that

$$\mathbb{E}[\max_j |Z_j|] \leq a\sqrt{2 \log 2m}, \qquad \implies \qquad \mathbb{E}[\max_j |\bar{\varepsilon}_j|] \leq \sqrt{\frac{2 \log 2m}{n_j}} \asymp \sqrt{\frac{2m \log 2m}{n}}.$$

The second term in (6) is controlled by the Lipschitz property of θ:

$$\max_j \sup_{x \in I_j} |\bar{\mu}_j - \theta(x)| \leq C \frac{1}{m}.$$

Combining these two facts, we get the following upper bound on the minimax risk of $\widehat{\theta}$,

$$R_n(\widehat{\theta}, \Theta) = \sup_{\theta \in \Theta} R_n(\widehat{\theta}, \theta) \leq \frac{C}{m} + \sqrt{\frac{2m \log 2m}{n}}.$$

Choosing $m \asymp (\log n/n)^{1/3}$ then gives us the required minimax rate

$$R_n(\widehat{\theta}, \Theta) \lesssim \left(\frac{\log n}{n} \right)^{1/3}.$$

Remark 2.2. Observe the we used the fact that the noise is Gaussian only while obtaining an upper bound on $\mathbb{E}[\max_j |\bar{\varepsilon}_j|]$, which relies on the fact that the Gaussian random variables have a light tail. Hence, the same argument carries through to the case of a larger family of light tailed noise distributions called sub-Gaussian distributions. Furthermore, for this family, our results can be easily extended to hold with high probability (rather than in expectation). The corresponding high probability lower bound derivation however, will require more powerful versions of Fano's inequality.

Remark 2.3. While our focus in this lecture was on considering the simple case of $\gamma \leq 1$, we note that very similar ideas can be applied for functions with higher order smoothness $\gamma > 1$. The only difference is that we have to employ more powerful estimators such as local polynomial (instead of constant) estimators, kernel smoothing etc.