Minimax Lower Bounds I

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

In the previous weeks, we developed a collection of *upper bounds*: concentration inequalities, chaining bounds, VC-based control of empirical processes, and risk bounds for regression and classification procedures. These results tell us what a specific estimator *can do*.

But to understand whether a rate is genuinely good, we also need a theory of *impossibility*. What can *any* estimator do, no matter how cleverly it is designed? What error is unavoidable because of noise, dimensionality, or the size of the model class?

This is the purpose of minimax lower bounds. They answer questions of the form:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\text{loss}(\widehat{\theta}, \theta)].$$

The outer infimum ranges over all estimators. So if we can prove a lower bound on this quantity, then no procedure can beat it.

The main idea of this lecture is surprisingly simple: If estimation were too accurate, then we could solve a hard hypothesis testing problem.

So we reduce estimation to testing, and then lower bound the probability of testing error. This is the conceptual core behind Le Cam's method, Fano's method, Assouad's method, and many related arguments.

In this lecture, we focus on the Fano route:

$$\text{estimation} \longrightarrow \text{multiple testing} \longrightarrow \text{mutual information} \longrightarrow \text{minimax lower bound}.$$

The next lecture will apply this framework to several familiar models, including normal means and linear regression.

2 Decision-theoretic setup

We work with a statistical model

$$\{P_\theta : \theta \in \Theta\},$$

where Θ is the parameter space and P_θ is the distribution of the data under parameter θ. An estimator is any measurable function $\widehat{\theta} = \widehat{\theta}(Z)$ of the observed data Z.

Definition 2.1 (Risk and minimax risk)**.** Let ρ be a semi-metric on Θ, and let $\Phi : [0,\infty) \to [0,\infty)$ be increasing. The risk of an estimator $\widehat{\theta}$ at parameter θ is

$$\mathcal{R}(\widehat{\theta},\theta) := \mathbb{E}_\theta\big[\Phi(\rho(\widehat{\theta},\theta))\big].$$

The associated minimax risk is

$$\mathfrak{M}(\Theta;\Phi\circ\rho) := \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} \mathcal{R}(\widehat{\theta},\theta).$$

The most common choices in this course are:

- $\Phi(t) = t$, which gives an absolute-error type loss;
- $\Phi(t) = t^2$, which gives a squared-error loss;
- $\rho(\widehat{\theta},\theta) = \|\widehat{\theta}-\theta\|_2$, for Euclidean parameter estimation;
- $\rho(\widehat{\theta},\theta) = \|X(\widehat{\theta}-\theta)\|_2/\sqrt{n}$, for prediction error in linear regression.

Remark 2.2. The minimax point of view is deliberately pessimistic: we judge an estimator by its worst-case performance over the model class. This is exactly what makes minimax lower bounds powerful: they rule out all procedures at once.

3 From estimation to testing

The first step is to reduce estimation to a finite testing problem. The geometry is organized by a packing set.

Definition 3.1 (2δ-separated set)**.** A finite subset $\{\theta_1,\dots,\theta_M\} \subset \Theta$ is called 2δ-separated in the semi-metric ρ if

$$\rho(\theta_j,\theta_k) \geqslant 2\delta \qquad \text{for all } j \neq k.$$

Given such a set, we build a testing problem as follows:

(1) Draw J uniformly from $[M] := \{1,\dots,M\}$.

(2) Given $J = j$, draw $Z \sim P_{\theta_j}$.

Let Q denote the joint law of (Z,J).

In words: nature first picks one of the packing points uniformly at random, then generates data from the corresponding distribution. If we can estimate θ_j accurately from Z, then we should also be able to identify j.

Proposition 3.2 (From estimation to testing)**.** *Let $\{\theta_1,\dots,\theta_M\} \subset \Theta$ be 2δ-separated. Then for every increasing $\Phi : [0,\infty) \to [0,\infty)$,*

$$\mathfrak{M}(\Theta;\Phi\circ\rho) \;\geqslant\; \Phi(\delta)\cdot \inf_{\psi} Q\big(\psi(Z) \neq J\big),$$

where the infimum is over all testing functions $\psi : \mathcal{Z} \to [M]$.

Proof. Fix any estimator $\widehat{\theta} = \widehat{\theta}(Z)$. Since Φ is increasing,

$$\mathbb{E}_{\theta_j}\big[\Phi(\rho(\widehat{\theta}, \theta_j))\big] \geqslant \Phi(\delta)\, P_{\theta_j}\big(\rho(\widehat{\theta}, \theta_j) \geqslant \delta\big) \qquad \text{for every } j \in [M].$$

Therefore

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\Phi(\rho(\widehat{\theta}, \theta))\big] \geqslant \frac{1}{M} \sum_{j=1}^{M} \Phi(\delta)\, P_{\theta_j}\big(\rho(\widehat{\theta}, \theta_j) \geqslant \delta\big).$$

Now define the nearest-neighbor decoder

$$\psi(Z) \in \operatorname*{arg\,min}_{k \in [M]} \rho(\widehat{\theta}(Z), \theta_k),$$

with ties broken arbitrarily.

We claim that

$$\{\psi(Z) \neq J\} \subseteq \{\rho(\widehat{\theta}(Z), \theta_J) \geqslant \delta\}.$$

Indeed, if $\rho(\widehat{\theta}(Z), \theta_J) < \delta$ and $\psi(Z) \neq J$, then

$$\rho(\widehat{\theta}(Z), \theta_{\psi(Z)}) \leqslant \rho(\widehat{\theta}(Z), \theta_J) < \delta.$$

By the triangle inequality,

$$\rho(\theta_J, \theta_{\psi(Z)}) \leqslant \rho(\theta_J, \widehat{\theta}(Z)) + \rho(\widehat{\theta}(Z), \theta_{\psi(Z)}) < 2\delta,$$

contradicting the 2δ-separation of the packing set. Hence

$$P_{\theta_j}\big(\rho(\widehat{\theta}, \theta_j) \geqslant \delta\big) \geqslant P_{\theta_j}\big(\psi(Z) \neq j\big).$$

Averaging over j gives

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\Phi(\rho(\widehat{\theta}, \theta))\big] \geqslant \Phi(\delta)\, Q(\psi(Z) \neq J).$$

Finally, take the infimum over all estimators $\widehat{\theta}$. Each estimator induces some decoder ψ, so

$$\mathfrak{M}(\Theta; \Phi \circ \rho) \geqslant \Phi(\delta) \inf_{\psi} Q(\psi(Z) \neq J). \qquad \square$$

Remark 3.3. This proposition is the key conceptual step. Once we have a well-separated packing, the minimax problem is reduced to a pure testing problem. Everything from now on is about showing that this testing problem is intrinsically hard.

4 Mutual information and Fano's inequality

To lower bound the testing error, we use an information-theoretic quantity.

Definition 4.1 (Mutual information)**.** Let $Q_{Z,J}$ be the joint law of (Z, J), and let Q_Z, Q_J be its marginals. The mutual information between Z and J is

$$I(Z; J) := D_{\mathrm{KL}}(Q_{Z,J} \,\|\, Q_Z \otimes Q_J).$$

It measures how much observing Z tells us about the hidden index J. If Z and J are independent, then $I(Z;J) = 0$. If Z determines J almost perfectly, then $I(Z;J)$ is large.

In our finite-mixture setup, if $J \sim \mathrm{Unif}([M])$ and $\bar{P} := \frac{1}{M}\sum_{j=1}^{M} P_{\theta_j}$, then

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, \bar{P}). \tag{1}$$

Theorem 4.2 (Fano's inequality)**.** *Assume that $J \sim \mathrm{Unif}([M])$. Then for every decoder $\psi : \mathcal{Z} \to [M]$,*

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

Proof. Let $V := \mathbf{1}\{\psi(Z) \neq J\}$, so that $Q(V = 1) = Q(\psi(Z) \neq J)$. Since $\psi(Z)$ is a function of Z, we have

$$H(J \mid Z) \leqslant H(J \mid \psi(Z)).$$

Now expand $H(J \mid \psi(Z))$ by adjoining the error indicator V:

$$H(J \mid \psi(Z)) \leqslant H(V, J \mid \psi(Z)) = H(V \mid \psi(Z)) + H(J \mid V, \psi(Z)).$$

The first term is at most $\log 2$, since $V \in \{0,1\}$. For the second term:

- on the event $\{V = 0\}$, we have $J = \psi(Z)$, so there is no uncertainty and $H(J \mid V = 0, \psi(Z)) = 0$;
- on the event $\{V = 1\}$, the index J can take at most $M - 1$ values, so $H(J \mid V = 1, \psi(Z)) \leqslant \log(M-1)$.

Therefore

$$H(J \mid \psi(Z)) \leqslant \log 2 + Q(V = 1)\log(M-1).$$

Since J is uniform on $[M]$, $H(J) = \log M$. Using $I(Z;J) = H(J) - H(J \mid Z)$, we get

$$I(Z;J) \geqslant \log M - \log 2 - Q(V = 1)\log(M-1).$$

Rearranging,

$$Q(V = 1) \geqslant \frac{\log M - \log 2 - I(Z;J)}{\log(M-1)}.$$

Finally, since $\log(M-1) \leqslant \log M$, the weaker but simpler lower bound

$$Q(V = 1) \geqslant 1 - \frac{I(Z;J) + \log 2}{\log M}$$

follows. □

Combining Fano with Proposition 3.2, we obtain the standard minimax lower bound:

Proposition 4.3 (Fano lower bound on minimax risk)**.** *Let $\{\theta_1, \ldots, \theta_M\} \subset \Theta$ be 2δ-separated. Then for every increasing $\Phi : [0,\infty) \to [0,\infty)$,*

$$\mathfrak{M}(\Theta; \Phi \circ \rho) \;\geqslant\; \Phi(\delta)\left(1 - \frac{I(Z;J) + \log 2}{\log M}\right).$$

In particular, if

$$I(Z;J) + \log 2 \leqslant \frac{1}{2}\log M,$$

then

$$\mathfrak{M}(\Theta; \Phi \circ \rho) \geqslant \frac{1}{2}\,\Phi(\delta).$$

5 How to control the mutual information

Proposition 4.3 reduces the problem to two tasks:

(1) construct a large 2δ-packing, so that $\log M$ is large;

(2) show that the induced distributions P_{θ_j} are close enough that $I(Z;J)$ is small.

A convenient way to bound $I(Z;J)$ is through pairwise Kullback–Leibler divergences.

Lemma 5.1 (Average pairwise KL bound)**.** *Let $\bar{P} = \frac{1}{M}\sum_{j=1}^{M} P_{\theta_j}$. Then*

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, \bar{P}) \;\leqslant\; \frac{1}{M^2}\sum_{j,k=1}^{M} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}).$$

Consequently,

$$I(Z;J) \leqslant \max_{j \neq k} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}).$$

Proof. Write p_j for a density of P_{θ_j} and $\bar{p} = \frac{1}{M}\sum_{k=1}^{M} p_k$. Then

$$D_{\mathrm{KL}}(P_{\theta_j} \,\|\, \bar{P}) = \int p_j \log \frac{p_j}{\bar{p}}.$$

Since $-\log$ is convex,

$$-\log\Big(\frac{1}{M}\sum_{k=1}^{M}\frac{p_k}{p_j}\Big) \leqslant \frac{1}{M}\sum_{k=1}^{M}\Big(-\log\frac{p_k}{p_j}\Big) = \frac{1}{M}\sum_{k=1}^{M}\log\frac{p_j}{p_k}.$$

Multiplying by p_j and integrating yields

$$D_{\mathrm{KL}}(P_{\theta_j} \,\|\, \bar{P}) \leqslant \frac{1}{M}\sum_{k=1}^{M} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}).$$

Averaging over j proves the first claim; the second is immediate. □

This leads to the standard "local packing" version of Fano's method.

Proposition 5.2 (Local packing form of Fano)**.** *Suppose that for some 2δ-separated set $\{\theta_1, \dots, \theta_M\} \subset \Theta$, all pairwise KL divergences satisfy*

$$D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}) \leqslant D \qquad \text{for all } j \neq k.$$

Then for every increasing Φ,

$$\mathfrak{M}(\Theta; \Phi \circ \rho) \;\geqslant\; \Phi(\delta)\left(1 - \frac{D + \log 2}{\log M}\right).$$

In particular, if $D + \log 2 \leqslant \frac{1}{2} \log M$*, then*

$$\mathfrak{M}(\Theta; \Phi \circ \rho) \geqslant \frac{1}{2} \Phi(\delta).$$

Remark 5.3. This is the practical recipe used in most examples:

(1) choose a packing radius δ;

(2) build a packing set of size M;

(3) upper bound the KL diameter D;

(4) choose δ so that D is of the same order as $\log M$.

The next lecture is mostly about learning how to execute these four steps in familiar models.

6 Look ahead

Let us summarize the roadmap we now have.

1. Packing: Build a well-separated finite subset of the parameter space. This is a packing problem, so metric entropy enters naturally.
2. Information: Show that the induced distributions are statistically hard to distinguish. This is usually done by computing or bounding pairwise KL divergences.
3. Fano: If the KL diameter is small compared to $\log M$, then the testing error stays bounded away from zero.
4. From testing to estimation: A nontrivial testing error implies a nontrivial minimax estimation error.

In the next lecture, we will run this recipe in three central examples:

- normal means;
- dense linear regression;
- sparse linear regression.

These examples will recover the same rates that we saw earlier from upper bounds, showing that those rates are minimax optimal.

Source material

Parts of this lecture are based on references: Wainwright (2019), in addition to the author's accumulated experience working on related topics.

References

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.