Minimax Lower Bounds II

SDS 391P.6, Spring 2026

Pratik Patil

These notes are a work in progress and are provided as-is for instructional purposes only. They are not (yet) at the level of a scholarly document. In particular, the notes draw from various sources and do not (yet) have sufficient references to the original sources. Additionally, almost surely the notes have errors and they are only probably approximately correct. The notes will be updated regularly as the course progresses. Last updated: 2026-04-20.

1 Motivation and roadmap

In the previous lecture, we developed the general minimax lower-bound strategy:

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

The goal of this lecture is to see that recipe in action.

We will work through three examples:

(1) the normal means model;

(2) dense linear regression;

(3) sparse linear regression.

In each case, the algebra is slightly different, but the logic is always the same:

- build a packing with many points;
- compute a KL bound between any pair;
- choose the packing radius so that the KL diameter is comparable to the log packing size.

This is exactly how minimax lower bounds should feel in practice: not mysterious, but a repeated geometric-information-theoretic template.

2 A KL formula for Gaussian location models

All of our examples today reduce to Gaussian location families, so let us record the KL formula once and for all.

Lemma 2.1 (Gaussian location KL). *Let $P_\mu = \mathcal{N}(\mu, \sigma^2 I_m)$ and $P_\nu = \mathcal{N}(\nu, \sigma^2 I_m)$. Then*

$$D_{\mathrm{KL}}(P_\mu \,\|\, P_\nu) = \frac{\|\mu - \nu\|_2^2}{2\sigma^2}.$$

Proof. This is a standard computation. Since the two distributions have the same covariance matrix, their KL divergence is just the squared Mahalanobis distance between the means divided by 2:

$$D_{\mathrm{KL}}(P_\mu \,\|\, P_\nu) = \frac{1}{2}(\mu-\nu)^\top(\sigma^2 I_m)^{-1}(\mu-\nu) = \frac{\|\mu-\nu\|_2^2}{2\sigma^2}.$$

$\square$

3 Example 1: the normal means model

Consider the normal means model

$$Y = \theta + W, \qquad W \sim \mathcal{N}(0,\sigma^2 I_n), \qquad \theta \in \mathbb{R}^n.$$

We ask for a minimax lower bound on squared Euclidean error:

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \mathbb{E}_\theta \|\widehat{\theta}-\theta\|_2^2.$$

Proposition 3.1 (Normal means lower bound)**.** *There exists a universal constant $c > 0$ such that*

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \mathbb{E}_\theta \|\widehat{\theta}-\theta\|_2^2 \;\geqslant\; c\,\sigma^2 n.$$

Proof. Since $\mathbb{R}^n$ contains every Euclidean ball, it suffices to lower bound the minimax risk over

$$\Theta_\delta := 4\delta\sqrt{n}\, B_2^n.$$

Choose a maximal $2\delta\sqrt{n}$-packing of Θ_δ in Euclidean norm. By the volume-ratio argument from metric entropy, we may choose points $\theta_1,\ldots,\theta_M \in \Theta_\delta$ such that

$$\|\theta_j - \theta_k\|_2 \geqslant 2\delta\sqrt{n} \qquad (j \neq k),$$

and

$$\log M \geqslant n \log 2.$$

Because all packing points lie in $4\delta\sqrt{n}B_2^n$, any pair satisfies

$$\|\theta_j - \theta_k\|_2 \leqslant 8\delta\sqrt{n}.$$

By Lemma 2.1,

$$D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}) = \frac{\|\theta_j-\theta_k\|_2^2}{2\sigma^2} \leqslant \frac{32 n\delta^2}{\sigma^2}.$$

Applying the local-packing Fano bound from the previous lecture with $\rho(\widehat{\theta},\theta) = \|\widehat{\theta}-\theta\|_2$ and $\Phi(t) = t^2$, we get

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \mathbb{E}_\theta \|\widehat{\theta}-\theta\|_2^2 \geqslant n\delta^2 \left(1 - \frac{32n\delta^2/\sigma^2 + \log 2}{n\log 2}\right).$$

Now choose

$$\delta^2 := \frac{\sigma^2 \log 2}{128}.$$

Then

$$\frac{32n\delta^2/\sigma^2 + \log 2}{n\log 2} = \frac{n\log 2/4 + \log 2}{n \log 2} = \frac{1}{4} + \frac{1}{n}.$$

Hence for $n \geqslant 2$,

$$1 - \frac{32n\delta^2/\sigma^2 + \log 2}{n \log 2} \geqslant \frac{1}{4}.$$

Therefore

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^n} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \geqslant \frac{1}{4} n\delta^2 = \frac{\log 2}{512} \sigma^2 n.$$

This proves the claim after absorbing constants. □

Remark 3.2. This lower bound is of the correct order. Indeed, the estimator $\widehat{\theta} = Y$ has risk

$$\mathbb{E}_\theta \|Y - \theta\|_2^2 = \mathbb{E}\|W\|_2^2 = \sigma^2 n.$$

So the minimax risk is actually $\asymp \sigma^2 n$.

4 Example 2: dense linear regression

Consider the fixed-design Gaussian linear model

$$Y = X\beta + W, \qquad W \sim \mathcal{N}(0, \sigma^2 I_n), \qquad \beta \in \mathbb{R}^d,$$

where $X \in \mathbb{R}^{n \times d}$ is deterministic. The natural loss is the prediction semi-norm

$$\rho_X(\widehat{\beta}, \beta) := \frac{\|X(\widehat{\beta} - \beta)\|_2}{\sqrt{n}}.$$

Proposition 4.1 (Linear regression lower bound)**.** *Let $r = \operatorname{rank}(X)$. There exists a universal constant $c > 0$ such that*

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|X(\widehat{\beta} - \beta)\|_2^2\right] \geqslant c\,\frac{\sigma^2 r}{n}.$$

Proof. The parameter that matters for prediction is not β itself, but the mean vector

$$\theta = X\beta \in \operatorname{range}(X) \subset \mathbb{R}^n.$$

So this is a normal means problem, but restricted to the r-dimensional subspace $\operatorname{range}(X)$.

Fix $\delta > 0$ and consider the set

$$\Theta_\delta := 4\delta\sqrt{n}\, B_2^n \cap \operatorname{range}(X).$$

Since Θ_δ lies in an r-dimensional Euclidean subspace, the same volume-ratio argument gives a $2\delta\sqrt{n}$-packing $\theta_1, \ldots, \theta_M \subset \Theta_\delta$ with

$$\log M \geqslant r \log 2$$

and pairwise distances satisfying

$$2\delta\sqrt{n} \leqslant \|\theta_j - \theta_k\|_2 \leqslant 8\delta\sqrt{n} \qquad (j \neq k).$$

Each $\theta_j \in \operatorname{range}(X)$, so choose $\beta_j \in \mathbb{R}^d$ such that $X\beta_j = \theta_j$. Under parameter β_j, the observation distribution is

$$P_{\beta_j} = \mathcal{N}(X\beta_j, \sigma^2 I_n) = \mathcal{N}(\theta_j, \sigma^2 I_n).$$

Hence by Lemma 2.1,

$$D_{\mathrm{KL}}(P_{\beta_j} \,\|\, P_{\beta_k}) = \frac{\|\theta_j - \theta_k\|_2^2}{2\sigma^2} \leqslant \frac{32n\delta^2}{\sigma^2}.$$

Applying Fano with the semi-metric ρ_X and squared loss $\Phi(t) = t^2$, we obtain

$$\inf_{\widehat{\beta}} \sup_{\beta\in\mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|X(\widehat{\beta}-\beta)\|_2^2\right] \geqslant \delta^2 \left(1 - \frac{32n\delta^2/\sigma^2 + \log 2}{r\log 2}\right).$$

Choose

$$\delta^2 := \frac{\sigma^2 r \log 2}{128n}.$$

Then

$$\frac{32n\delta^2/\sigma^2 + \log 2}{r\log 2} = \frac{r\log 2/4 + \log 2}{r\log 2} = \frac{1}{4} + \frac{1}{r}.$$

So for $r \geqslant 2$, the parenthesis is bounded below by an absolute constant. Thus

$$\inf_{\widehat{\beta}} \sup_{\beta\in\mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|X(\widehat{\beta}-\beta)\|_2^2\right] \geqslant c\,\frac{\sigma^2 r}{n}$$

for a universal constant $c > 0$. □

Remark 4.2. This lower bound matches the upper bound achieved by ordinary least squares, up to constants. So the prediction-error rate $\sigma^2\,\mathrm{rank}(X)/n$ is minimax optimal.

5 A sparse packing lemma

To study sparse models, we need a packing set that is simultaneously

- *large*, so that $\log M$ is large;
- *sparse*, so that every point lies in the parameter class of interest.

The following is the standard combinatorial ingredient.

Lemma 5.1 (Sparse Varshamov–Gilbert packing)**.** *Assume* $10 \leqslant s \leqslant d/2$. *Then there exists a set*

$$\mathcal{V} \subset \{0,1\}^d$$

such that:

(i) every $v \in \mathcal{V}$ *has exactly* s *ones;*

(ii) for all distinct $v, v' \in \mathcal{V}$,

$$H(v,v') \geqslant \frac{s}{2},$$

where $H(\cdot,\cdot)$ *denotes Hamming distance;*

(iii)

$$\log|\mathcal{V}| \geqslant c\,s\log\left(\frac{ed}{s}\right)$$

for a universal constant $c > 0$.

Remark 5.2. This lemma is a constant-weight refinement of the usual Varshamov–Gilbert construction. For minimax lower bounds in sparse models, it is exactly what produces the combinatorial factor $s\log(ed/s)$.

6 Example 3: sparse linear regression under orthonormal design

Now consider the linear model

$$Y = X\beta + W, \qquad W \sim \mathcal{N}(0, \sigma^2 I_n),$$

under the orthonormal design condition

$$X^\top X = nI_d.$$

Assume the parameter is s-sparse:

$$\beta \in \mathcal{T}_d(s) := \{\beta \in \mathbb{R}^d : \|\beta\|_0 \leqslant s,\ \|\beta\|_2 \leqslant 1\}.$$

We study squared ℓ_2-error:

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{T}_d(s)} \mathbb{E}_\beta \|\widehat{\beta} - \beta\|_2^2.$$

Because $X^\top X = nI_d$, the transformed observation

$$Z := \frac{1}{n} X^\top Y$$

satisfies

$$Z = \beta + \frac{1}{n} X^\top W.$$

Moreover,

$$\frac{1}{n} X^\top W \sim \mathcal{N}\Big(0, \frac{\sigma^2}{n} I_d\Big),$$

so this is exactly a normal means model with effective noise level $\sigma/\sqrt{n}$.

Proposition 6.1 (Sparse lower bound)**.** *Assume* $10 \leqslant s \leqslant d/2$*. Then there exists a universal constant* $c > 0$ *such that*

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{T}_d(s)} \mathbb{E}_\beta \|\widehat{\beta} - \beta\|_2^2 \geqslant c\, \frac{\sigma^2 s \log(ed/s)}{n},$$

as long as the right-hand side is at most a sufficiently small constant.

Proof. Let $\mathcal{V}$ be the set from Lemma 5.1. For each $v \in \mathcal{V}$, define

$$\beta^v := \frac{4\delta}{\sqrt{s}}\, v.$$

If $\delta \leqslant 1/4$, then

$$\|\beta^v\|_0 = s, \qquad \|\beta^v\|_2 = 4\delta \leqslant 1,$$

so $\beta^v \in \mathcal{T}_d(s)$.

Now consider two distinct elements $v, v' \in \mathcal{V}$. Their squared Euclidean distance is

$$\|\beta^v - \beta^{v'}\|_2^2 = \frac{16\delta^2}{s}\, H(v, v').$$

By the Hamming separation $H(v, v') \geqslant s/2$,

$$\|\beta^v - \beta^{v'}\|_2^2 \geqslant 8\delta^2, \qquad \text{so} \qquad \|\beta^v - \beta^{v'}\|_2 \geqslant 2\delta.$$

Thus $\{\beta^v : v \in \mathcal{V}\}$ is a 2δ-separated subset of $\mathcal{T}_d(s)$.

Next, because each vector has s ones, the Hamming distance is at most $2s$, so

$$\|\beta^v - \beta^{v'}\|_2^2 = \frac{16\delta^2}{s}\, H(v, v') \leqslant 32\delta^2.$$

Since the transformed observation Z follows

$$Z \sim \mathcal{N}\Big(\beta, \frac{\sigma^2}{n} I_d\Big),$$

Lemma 2.1 yields

$$D_{\mathrm{KL}}(P_{\beta^v} \,\|\, P_{\beta^{v'}}) = \frac{n}{2\sigma^2}\|\beta^v - \beta^{v'}\|_2^2 \leqslant \frac{16n\delta^2}{\sigma^2}.$$

Now apply the local-packing Fano bound with $\rho(\widehat{\beta}, \beta) = \|\widehat{\beta} - \beta\|_2$ and $\Phi(t) = t^2$:

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{T}_d(s)} \mathbb{E}_\beta \|\widehat{\beta} - \beta\|_2^2 \geqslant \delta^2 \left(1 - \frac{16n\delta^2/\sigma^2 + \log 2}{\log |\mathcal{V}|}\right).$$

Using $\log |\mathcal{V}| \geqslant c_1 s \log(ed/s)$, it suffices to choose

$$\delta^2 = c_2\, \frac{\sigma^2 s \log(ed/s)}{n}$$

with $c_2 > 0$ small enough that

$$16n\delta^2/\sigma^2 + \log 2 \leqslant \frac{1}{2} \log |\mathcal{V}|.$$

This gives

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathcal{T}_d(s)} \mathbb{E}_\beta \|\widehat{\beta} - \beta\|_2^2 \geqslant \frac{1}{2}\,\delta^2 \asymp \frac{\sigma^2 s \log(ed/s)}{n}.$$

□

Remark 6.2. This lower bound matches the familiar sparse-regression upper rate

$$\frac{\sigma^2 s \log(ed/s)}{n},$$

so the logarithmic factor $\log(ed/s)$ is not an artifact of the proof of the upper bound. It is the correct minimax price of not knowing the support.

7 A brief word on more general sparse designs

The previous example used the orthonormal design condition $X^\top X = nI_d$, which reduces the problem to sparse normal means. For a more general design matrix, lower bounds depend on sparse eigenvalue quantities.

For instance, if one wants a lower bound on parameter error $\|\widehat{\beta} - \beta\|_2^2$, then the KL calculation involves upper sparse singular values of X. If one wants a lower bound on prediction error $\|X(\widehat{\beta} - \beta)\|_2^2/n$, then one also needs a lower sparse singular-value condition to ensure that the packing is separated in prediction norm. The resulting rates are again of order

$$\frac{\sigma^2 s \log(ed/s)}{n},$$

up to design-dependent constants.

So the orthonormal design case already captures the essential combinatorics; general designs add a layer of geometry through restricted singular values.

8 Yang–Barron and the role of metric entropy

So far, our lower bounds have been based on *explicit packings*. This is often enough in finite-dimensional parametric problems. For nonparametric classes, however, explicit packing constructions can be awkward. In that setting, it is useful to have a direct connection between mutual information and metric entropy.

Theorem 8.1 (Yang–Barron, informal form)**.** *Let $\mathcal{P} = \{P_{\theta_1}, \dots, P_{\theta_M}\}$ be a finite family of distributions. Then the mutual information can be bounded as*

$$I(Z; J) \leqslant \inf_{\varepsilon > 0} \Big\{ \varepsilon^2 + \log N_{\mathrm{KL}}(\varepsilon, \mathcal{P}) \Big\},$$

where $N_{\mathrm{KL}}(\varepsilon, \mathcal{P})$ is the covering number of $\mathcal{P}$ in a suitable square-root KL metric.

We will not prove this here. Its message is the important part:

mutual information can be controlled by metric entropy.

This is the lower-bound analogue of Dudley's entropy integral and related upper-bound methods.

Remark 8.2. For smooth nonparametric classes, Yang–Barron lets one derive minimax lower bounds directly from entropy growth rates. Combined with the upper bounds from our nonparametric regression lectures, this is how one proves that rates such as

$$n^{-2\alpha/(2\alpha+1)} \quad \text{or} \quad n^{-2s/(2s+d)}$$

are not just achievable, but minimax optimal.

9 Takeaways

Let us collect the big ideas from the two minimax lectures:

- Fano: If the mutual information is much smaller than $\log M$, then the testing error stays bounded away from zero.
- Packing versus KL diameter: Good lower bounds come from a balance between

$$\log M \qquad \text{and} \qquad \max_{j \neq k} D_{\mathrm{KL}}(P_{\theta_j} \,\|\, P_{\theta_k}).$$

- Matching lower and upper bounds: For the models treated here:

$$\text{normal means: } \sigma^2 n, \qquad \text{dense linear prediction: } \frac{\sigma^2 \operatorname{rank}(X)}{n}, \qquad \text{sparse regression: } \frac{\sigma^2 s \log(ed/s)}{n}.$$

  These are the correct minimax rates, up to constants.

This is one of the most satisfying points in a course: the upper and lower bounds meet, so the theory closes!

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.