ECE 598 - Fall 2025

Lecture 11: Minimax Lower Bounds VI: Global Fano

2nd October, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we conclude our module on deriving minimax lower bounds by discussing the global Fano method. We begin our discussion by introducing a variational definition of mutual information, which we then use to get an upper bound on the mutual information in terms of KL-covering numbers. This naturally leads to our statement of the global Fano method. We end the lecture with some applications that illustrate how this approach allows us to use off-the-shelf metric entropy results to derive minimax lower bounds.

1 A variational definition of mutual information

Assume that $V \sim \mu$ is some random variable taking values in Θ, and let X have a distribution P_θ conditioned on $V = \theta$; that is,

$$P_X(E) = \int P_\theta(E) d\mu(\theta), \quad \text{for all measurable} \quad E \subset \mathcal{X}.$$

The mutual information between V and X is defined as

$$I(V; X) = D_{\mathrm{KL}}(P_{VX} \parallel P_V P_X) = D_{\mathrm{KL}}(P_{X|V} \parallel P_X \mid P_V).$$

In this section, we will show a variational definition of mutual information.

Lemma 1.1. *Suppose (X, V) are jointly distributed as discussed above. Then, for any distribution Q, we have*

$$I(V; X) \leq D_{KL}(P_{X|V} \parallel Q \mid P_V), \quad \textit{with equality iff} \quad Q = P_X.$$

This implies the following variational definition of mutual information:

$$I(X; V) = \inf_Q D_{KL}(P_{X|V} \parallel Q \mid P_V).$$

Proof. We begin with the conditional relative-entropy definition of mutual information:

$$I(V : X) = D_{\mathrm{KL}}(P_{X|V} \parallel P_X \mid P_V) = \int_\Theta D(P_\theta \parallel P_\mu) d\mu(\theta),$$

where $P_X(E) = P_\mu(E) = \int P_\theta(E) d\mu(\theta)$. Now, let us consider any other distribution Q, and for simplicity, assume that there exists some common dominating measure ν w.r.t. which all P_θ, P_μ,

and Q admit densities. Then, we can write:

$$\begin{aligned}\int_\Theta D(P_\theta \parallel P_\mu) d\mu(\theta) &= \int_\Theta d\mu(\theta) \int_{\mathcal{X}} \log\left(\frac{p_\theta(x)}{p_\mu(x)}\right) p_\theta(x) d\nu(x) \\ &= \int_\Theta d\mu(\theta) \left(\int_{\mathcal{X}} \log\left(\frac{p_\theta(x)}{q(x)}\right) p_\theta(x) d\nu(x) - \int_{\mathcal{X}} \log\left(\frac{p_\mu(x)}{q(x)}\right) p_\theta(x) d\nu(x) \right).\end{aligned}$$

The first term in RHS above is simply $D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V)$. For the second term, we can interchange the order of integration (by appealing to Fubini's theorem) to get

$$\int_\Theta d\mu(\theta) \int_{\mathcal{X}} \log\left(\frac{p_\mu(x)}{q(x)}\right) p_\theta(x) d\nu(x) = \int_{\mathcal{X}} \log\left(\frac{p_\mu(x)}{q(x)}\right) d\nu(x) \underbrace{\int_\Theta p_\theta(x) d\mu(\theta)}_{=p_\mu(x)}.$$

Thus, the second term is simply the relative entropy between the mixture (or the marginal of X) distribution $P_\mu \equiv P_X$, and the distribution Q. That is, we have proved that for any Q, we have

$$I(V;X) = D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V) - D_{\mathrm{KL}}(P_X \parallel Q) \leq D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V).$$

Furthermore, the equality is achieved by choosing $Q = P_X = P_\mu$, which implies the second statement:

$$I(X;V) = \inf_Q D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V).$$

□

Remark 1.2. Note that an important concept in information theory is the capacity of a channel (in this case, the channel is the conditional distribution $P_{X|V}$), defined as

$$C(P_{X|V}) = \sup_{P_V} I(X;V) = \sup_{P_V} \inf_Q D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V),$$

where the second equality follows from Lemma 1.1.

Now, assuming that we can interchange the order of $\sup$ and $\inf$ in the definition above (we will discuss conditions for this to be feasible in a few lectures), we get the following:

$$C(P_{X|V}) = \sup_{P_V} \inf_Q D_{\mathrm{KL}}(P_{X|V} \parallel Q \mid P_V) = \inf_Q \sup_\theta D_{\mathrm{KL}}(P_\theta \parallel Q), \tag{1}$$

where in the second equality we used the fact that the objective is linear in P_V, and is thus achieved at an extreme point (i.e., $P_V = \delta_\theta$). The quantity in (1) can be interpreted as the value of a two-player game:

- Player chooses a distribution Q
- Nature or adversary chooses a distribution P_θ, and draws an $X \sim P_\theta$
- The loss incurred by the player (gain achieved by the adversary) is equal to the relative entropy between P_θ and Q.

Then, Lemma 1.1 implies the value of this game is equal to the capacity of the channel $P_{X|V}$, and the optimal action for the player is the output distribution P_μ associated with the capacity achieving input distribution μ^*. This is called the Redundancy-Capacity theorem.

2 Global Fano Method

We now go back to our usual minimax estimation problem. Let $\{P_\theta : \theta \in \Theta\}$ denote a statistical model indexed by a pseudo-metric space (Θ, ρ). For some non-decreasing function $\Phi : [0, \infty) \to [0, \infty)$, the associated minimax estimation risk is defined as

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

To present the global Fano bound, we need to introduce the notions of packing and covering numbers. We present these definitions in the exact context they will be used. We begin by introducing the notion of δ-packing number of the parameter space (Θ, ρ).

Definition 2.1. For any $\delta > 0$, the δ-packing number of (Θ, ρ) is defined as

$$M_\rho(\Theta, \delta) = \sup\{M : \exists\, \theta_1, \ldots, \theta_M \in \Theta;\ \rho(\theta_i, \theta_j) \geq \delta\}.$$

In words, the δ-packing number of (Θ, ρ) is the cardinality of the largest subset of Θ whose elements satisfy a pairwise distance (in terms of ρ) of at least δ.

Next, we introduce the definition of ϵ^2-covering number of a class of probability distributions in terms of relative entropy.

Definition 2.2. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ denote a collection of probability distributions. Then, the ϵ^2-covering number of $\mathcal{P}$ is defined as

$$N_{KL}(\Theta, \epsilon^2) := \inf\{N : \exists\, Q_1, \ldots, Q_N;\ \sup_{\theta \in \Theta} \min_{1 \leq i \leq N} D_{\mathrm{KL}}(P_\theta \parallel Q_i) \leq \epsilon^2\}.$$

In other words, the ϵ^2-covering number of $\mathcal{P}$ is the size of the smallest subset of $\mathcal{P}$, such that for any $P \in \mathcal{P}$, there exists an element in the subset no more than ϵ^2 away from P in relative entropy.

We can now proceed toward the derivation of a global Fano lower bound. The first step is to use Lemma 1.1 to establish an upper bound on the mutual information in terms of the ϵ^2-covering number.

Lemma 2.3. *Let $V \sim \mu$ denote a Θ-valued random variable, and let $X \sim P_\mu$ (i.e., $X|V = \theta \sim P_\theta$). Then, we have the following:*

$$I(X; V) \leq \inf_{\epsilon > 0} \left\{ \epsilon^2 + \log N_{KL}(\Theta, \epsilon^2) \right\}.$$

Remark 2.4. Note that the upper bound in Lemma 2.3 is independent of the choice of the marginal distribution P_V. Hence, it is also a valid upper bound after taking a supremum over all P_V; that is, it is a valid upper bound on the capacity of the channel $(P_{X|V})$.

Proof. Let $\{Q_1, \ldots, Q_N\}$ denote an ϵ^2-covering of $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $N = N_{KL}(\Theta, \epsilon^2)$. Then, Lemma 1.1 implies the following with $\bar{Q} = (1/N) \sum_{i=1}^N Q_i$:

$$I(X; V) \leq D_{\mathrm{KL}}(P_{X|V} \parallel \bar{Q} \mid P_V) = \int_\Theta d\mu(\theta) \int_{\mathcal{X}} \log \left(\frac{N p_\theta(x)}{\sum_i q_i(x)} \right) p_\theta(x) d\nu(x),$$

where we have again assumed for simplicity that there exists a common dominating measure ν on $\mathcal{X}$. Now, observe that for any p_θ, there exists a q_{i^*} such that $D_{\mathrm{KL}}(P_\theta \parallel Q_{i^*}) \leq \epsilon^2$, which implies

$$\int_{\mathcal{X}} \log\left(\frac{N p_\theta(x)}{\sum_i q_i(x)}\right) p_\theta(x) d\nu(x) \leq \log N + \int_{\mathcal{X}} \log\left(\frac{p_\theta(x)}{q_{i^*}(x)}\right) d\nu(x) \leq \log N + \epsilon^2.$$

In the first inequality, we have simply used the fact that the densities are nonnegative, and thus $q_{i^*}(x) \leq \sum_i q_i(x)$, while the second inequality uses the ϵ^2-covering assumption. This gives us the required upper bound

$$I(X;V) \leq \int_\Theta \left(\log N + \epsilon^2\right) d\mu(\theta) = \log N + \epsilon^2.$$

Since this statement is true for an arbitrary $\epsilon > 0$, the inequality remains valid on taking an infimum over all $\epsilon > 0$. □

Remark 2.5. When working with n i.i.d. observations, the bound of Lemma 2.3 specializes to

$$I(V; X^n) \leq \inf_{\epsilon > 0} \left\{\log N_{KL}(\Theta, \epsilon^2) + n\epsilon^2\right\}.$$

Note that here N_{KL} is the covering number of the distribution class $\{P_\theta : \theta \in \Theta\}$, and the observations are assumed to be drawn from $P_\theta^{\otimes n}$.

We now have all the components to state the Global Fano lower bound.

Theorem 2.6. *Fix any $\epsilon, \delta > 0$. Then, the minimax risk of the estimation problem introduced in this section satisfies*

$$R^*(\Theta, \Phi \circ \rho) \geq \Phi(\delta)\left(1 - \frac{\epsilon^2 + \log N_{KL}(\Theta, \epsilon^2) + \log 2}{\log M_\rho(\Theta, 2\delta)}\right).$$

When working with n i.i.d. observations drawn $(X_1, \ldots, X_n)$, the minimax risk satisfies

$$R_n^*(\Theta, \Phi \circ \rho) \geq \Phi(\delta)\left(1 - \frac{n\epsilon^2 + \log N_{KL}(\Theta, \epsilon^2) + \log 2}{\log M_\rho(\Theta, 2\delta)}\right).$$

Proof. The proof essentially follows immediately by plugging in the mutual information upper bound in the usual Fano lower bound. We present all the details below for completeness.

For a fixed $\delta > 0$, let $M \equiv M_\rho(\Theta, 2\delta)$ denote the 2δ-packing number of (Θ, ρ), and let Θ_M denote such a packing set. Let V denote a random variable with uniform distribution on Θ_M. Then, we have the following:

$$R^*(\Theta, \Phi \circ \rho) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta} R(\widehat{\theta}, \theta) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_M} R(\widehat{\theta}, \theta) \geq \inf_{\widehat{\theta}} \mathbb{E}_V[R(\theta, V)],$$

where $V \sim \mathrm{Uniform}(\Theta_M)$. The last term above can be further lower bounded by

$$\inf_{\widehat{\theta}} \mathbb{E}_V[R(\theta, V)] \geq \Phi(\delta) \inf_{\Psi : \mathcal{X} \to \Theta_M} \mathbb{P}\left(\Psi(X) \neq V\right),$$

where we used the 2δ-packing property of the set Θ_M. Since we have assumed that V is uniformly distributed over the finite set Θ_M, we can use the mutual information version of standard Fano's inequality to get the lower bound:

$$R^*(\Theta, \Phi \circ \rho) \geq \Phi(\delta) \left(1 - \frac{I(V;X) + \log 2}{\log M}\right).$$

The final step is to use Lemma 2.3 to upper bound $I(V;X)$ with $\epsilon^2 + \log N_{KL}(\Theta, \epsilon^2)$ to conclude the proof. □

To summarize, Theorem 2.6 suggests the following general recipe for deriving minimax lower bounds (Duchi, 2025, Section 12.2.22) (assuming n i.i.d. observations):

- Obtain an upper bound for $N_{KL}(\Theta, \epsilon^2)$ and a lower bound for $M_\rho(\Theta, 2\delta)$.
- Choose a value of ϵ_n such that $n\epsilon_n^2 \asymp \log N_{KL}(\Theta, \epsilon_n^2)$
- Choose a value of δ_n for which $\log M_\rho(\Theta, 2\delta) \geq 2\left(\log 2 + 4n\epsilon_n^2\right)$
- With these choices, we get the following lower bound on the minimax risk: $R_n^*(\Theta, \Phi \circ \rho) \geq \Phi(\delta_n)/2$.

The main benefit of this approach is that in many settings, there exist well-known tight approximations of the packing and covering numbers of parameter spaces which can be used directly to derive lower bounds. In particular, in many examples, we have $D_{\mathrm{KL}}(P_\theta \parallel P_{\theta'}) \leq \kappa\rho(\theta, \theta)^2$, and we can directly employ the covering and packing number bounds for (Θ, ρ), as we illustrate in the examples next.

3 Examples

In this section, we look at some examples in which the method described in the previous section is applicable.

3.1 Gaussian Location Model

Suppose we have n i.i.d. observations $X_1, \ldots, X_n$ drawn from $N(\theta, \sigma^2 I_d)$, with the parameter set $\Theta = B_2(R) \coloneqq \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$, $\rho(\theta, \theta') = \|\theta - \theta'\|_2$ and $\Phi(t) = t^2$. Then, observe the following:

- With $P_\theta = N(\theta, \sigma^2 I_d)$, we have $D_{\mathrm{KL}}(P_\theta \parallel P_{\theta'}) = \frac{\|\theta - \theta'\|^2}{2\sigma^2} = \frac{\Phi \circ \rho(\theta, \theta')}{2\sigma^2}$. This implies the following:

$$N_{KL}(\Theta, \epsilon^2) = N_\rho(\Theta, \sqrt{2}\sigma\epsilon) \leq C_d \left(\frac{R}{\sqrt{2}\sigma\epsilon}\right)^d$$

- Similarly, with standard volume arguments, we have the following

$$M_\rho(\Theta, 2\delta) \geq c_d \left(\frac{R}{2\delta}\right)^d$$

- The next step is to choose an ϵ_n such that

$$n\epsilon_n^2 \geq \log N_{KL}(\Theta, \epsilon_n^2).$$

Ignoring the constants, a sufficient condition for this is to ensure $n\epsilon_n^2 \geq d \log\left(\frac{R}{\sigma\epsilon_n}\right)$. By selecting $\epsilon_n \asymp \sqrt{d/n}$, we get that $\log N_{KL}(\Theta, \epsilon_n^2) + n\epsilon_n^2 \lesssim d \log(R\sqrt{n}/\sigma\sqrt{d})$. Thus, with this choice we have

$$I(V; X^n) + \log 2 \lesssim d \log(R\sqrt{n}/\sigma\sqrt{d}).$$

- Next, we choose $\delta_n \asymp \sigma\sqrt{d/n}$, and use the lower bound on $M_\rho(\Theta, \delta_n)$ to ensure that $M_\rho(\Theta, 2\delta_n) \geq 2(\log 2 + I(V; X^n))$. With this choice we get the minimax lower bound of the order

$$R_n^*(\Theta, \Phi \circ \rho) \gtrsim \Phi\left(\sigma\sqrt{\frac{d}{n}}\right) \asymp \frac{\sigma^2 d}{n}.$$

3.2 Nonparametric Regression

Consider a problem of nonparametric regression with random design (we can also consider the fixed design case, but random design simplifies matters). In this case, our observation $\boldsymbol{X} = \{(U_i, Y_i) : 1 \leq i \leq n\}$, with

$$Y_i = \theta(U_i) + \sigma\varepsilon_i.$$

As before, we assume that Θ consists of (C, γ) Hölder continuous functions, and each $\varepsilon_i \sim N(0, 1)$. So each distribution P_θ in our case is a product $\otimes_{i=1}^n(\text{Uniform}([0, 1]^d) \times N(\theta(U_i), \sigma^2))$, and we observe that

$$D_{\text{KL}}(P_\theta \parallel P'_\theta) = \frac{n}{2\sigma^2}\|\theta - \theta'\|_{L^2}^2, \tag{2}$$

where $L^2 \equiv L^2(P_U)$ denotes the L^2 norm w.r.t. the uniform distribution on $[0, 1]^d$. Suppose our goal is to estimate the regression function θ in terms of squared L^2 norm (i.e., $\rho(\theta, \theta') = \|\theta - \theta'\|_{L^2}$ and $\Phi(t) = t^2$). Then, we have the following:

- The covering number $N_{KL}(\Theta, \epsilon^2)$ for any $\epsilon > 0$ is of the order $N_\rho(\Theta, \sigma\epsilon/\sqrt{n})$. This is known to satisfy

$$\log N_{KL}(\Theta, \epsilon^2) \asymp \log N_\rho(\Theta, \sigma\epsilon) \asymp \left(\frac{C\sqrt{n}}{\sigma\epsilon_n}\right)^{d/\gamma}.$$

- The 2δ packing number is of a similar order

$$\log M_\rho(\Theta, 2\delta) \asymp \left(\frac{C}{\delta}\right)^{d/\gamma}$$

- We need to select ϵ_n to balance the two terms in the upper bound on $I(V; \boldsymbol{X})$; that is,

$$\epsilon_n : \quad \epsilon_n^2 \asymp \left(\frac{C\sqrt{n}}{\sigma\epsilon_n}\right)^{d/\gamma} \quad \Longrightarrow \epsilon_n \asymp \left(\frac{\sigma}{\sqrt{n}}\right)^{-1/(1+2\gamma/d)}.$$

  This leads to a bound on $I(V; \boldsymbol{X}) \asymp (\sqrt{n}/\sigma)^{2d/(d+2\gamma)}$.

- Hence, an appropriate choice of δ_n to make $(I(V; \boldsymbol{X}) + \log 2)/\log(M_\rho(\Theta, 2\delta_n))$ is

$$\delta_n : \quad \left(\frac{C}{\delta_n}\right)^{d/\gamma} \asymp \left(\frac{\sqrt{n}}{\sigma}\right)^{2d/(d+2\gamma)}, \quad \Longrightarrow \delta_n \asymp \sigma^{\frac{2\gamma}{d+2\gamma}} n^{-\frac{\gamma}{d+2\gamma}}.$$

- With this choice, we know that the minimax risk is of the order $\asymp \Phi(\delta_n) \asymp \delta_n^2$:

$$R_n^*(\Theta, \Phi \circ \rho) \gtrsim \left(\frac{\sigma^2}{n}\right)^{\frac{2\gamma}{d+2\gamma}}.$$

This example illustrates how the global Fano method simplifies the task of deriving minimax lower bounds in many instances. Unlike the (local Fano) method of the previous lecture, we do not have to manually construct a class of "hard problems" in an ad-hoc manner. However, this simplification sometimes comes at the cost of a loss in tightness; for example, the global Fano method fails to capture the poly-log factor in the case of regression in sup norm (Exercise: try it.), but local Fano does.

Remark 3.1. When working with fixed design, the key difference will be in the relative entropy expression (2), where the L^2-norm will be replaced by an empirical L^2-norm. In that case, we can use the Hölder continuity of θ to get an extra approximation term between the empirical and population L^2 norms.

References

J. Duchi. Lecture Notes for "Statistics 311 / EE 377: Information Theory and Statistics". Online lecture notes, Stanford University, 2025. URL `https://stanford.edu/class/ee377/book.html`. Accessed: 2025-10-01.