Chapter 5

Global Fano Method

In this chapter, we extend the techniques of Chapter 2.4 on Fano’s method (the local Fano method) to a more global construction. In particular, we show that, rather than constructing a local packing, choosing a scaling \( \delta > 0 \), and then optimizing over this \( \delta \), it is actually, in many cases, possible to prove lower bounds on minimax error directly using packing and covering numbers (metric entropy and packing entropy). The material in this chapter is based on a paper of Yang and Barron [5].

5.1 A mutual information bound based on metric entropy

To begin, we recall the classical Fano inequality, which says that for any Markov chain \( V \to X \to \hat{V} \), where \( V \) is uniform on the finite set \( \mathcal{V} \), we have

\[
P(\hat{V} \neq V) \geq 1 - \frac{I(V;X) + \log 2}{\log(|\mathcal{V}|)}.
\]

(Recall Corollary 2.11.) Thus, there are two ingredients in proving lower bounds on the error in a hypothesis test: upper bounding the mutual information and lower bounding the size \( |\mathcal{V}| \). Here, we state a proposition doing the former.

Before stating our result, we require a bit of notation. First, we assume that \( V \) is drawn from a distribution \( \mu \), and conditional on \( V = v \), assume the sample \( X \sim P_v \). Then a standard calculation (or simply the definition of mutual information; recall equation (2.4.4)) gives that

\[
I(V;X) = \int D_{\text{KL}}(P_v || \mathcal{P}) \, d\mu(v), \quad \text{where} \quad \mathcal{P} = \int P_v \, d\mu(v).
\] (5.1.1)

Now, we show how to connect this mutual information quantity to a covering number of a set of distributions.

Assume that for all \( v \), we have \( P_v \in \mathcal{P} \), where \( \mathcal{P} \) is a collection of distributions. In analogy with Definition 2.4.1 we say that the collection of distributions \( \{Q_i\}_{i=1}^N \) form an \( \epsilon \)-cover of \( \mathcal{P} \) in KL-divergence if for all \( P \in \mathcal{P} \), there exists some \( i \) such that \( D_{\text{KL}}(P || Q_i) \leq \epsilon^2 \). With this, we may define the KL-covering number of the set \( \mathcal{P} \) as

\[
N_{\text{kl}}(\epsilon, \mathcal{P}) := \inf \left\{ N \in \mathbb{N} \mid \exists Q_i, i = 1, \ldots, N, \sup_{P \in \mathcal{P}} \min_i D_{\text{kl}}(P || Q_i) \leq \epsilon^2 \right\}, \quad (5.1.2)
\]

where \( N_{\text{kl}}(\epsilon, \mathcal{P}) = +\infty \) if no such cover exists. With definition (5.1.2) in place, we have the following proposition.
Proposition 5.1. Under conditions of the preceding paragraphs, we have

\[ I(V; X) \leq \inf_{\epsilon > 0} \left\{ \epsilon^2 + \log N_{\text{KL}}(\epsilon, P) \right\}. \quad (5.1.3) \]

Proof. First, we claim that

\[
\int D_{\text{KL}}(P_v \| P) \, d\mu(v) \leq \int D_{\text{KL}}(P_v \| Q) \, d\mu(v) \tag{5.1.4}
\]

for any distribution \( Q \). Indeed, briefly, we have

\[
\int D_{\text{KL}}(P_v \| P) \, d\mu(v) = \int \int \frac{dP_v}{dP} \, d\mu(v) = \int \int \left[ \log \frac{dP_v}{dQ} + \log \frac{dQ}{dP} \right] d\mu(v)
\]

\[
= \int \int D_{\text{KL}}(P_v \| Q) \, d\mu(v) + \int \int \frac{d\mu(v) dP_v}{dP} \log \frac{dQ}{dP}
\]

\[
= \int D_{\text{KL}}(P_v \| Q) \, d\mu(v) - D_{\text{KL}}(P \| \overline{P}) \leq \int D_{\text{KL}}(P_v \| Q) \, d\mu(v),
\]

so that inequality (5.1.4) holds. By carefully choosing the distribution \( Q \) in the upper bound (5.1.4), we obtain the proposition.

Now, assume that the distributions \( Q_i, i = 1, \ldots, N \) form an \( \epsilon^2 \)-cover of the family \( P \), meaning that

\[
\min_{i \in [N]} D_{\text{KL}}(P_v \| Q_i) \leq \epsilon^2 \quad \text{for all } P \in \mathcal{P}.
\]

Let \( p_v \) and \( q_i \) denote the densities of \( P_v \) and \( Q_i \) with respect to some fixed base measure on \( \mathcal{X} \) (the choice of base measure does not matter). Then defining the distribution \( Q = (1/N) \sum_i^N Q_i \), we obtain for any \( v \) that in expectation over \( X \sim P_v \),

\[
D_{\text{KL}}(P_v \| Q) = \mathbb{E}_{P_v} \left[ \log \frac{p_v(X)}{q(X)} \right] = \mathbb{E}_{P_v} \left[ \log \frac{p_v(X)}{\sum_{i=1}^N q_i(X)} \right]
\]

\[
= \log N + \mathbb{E}_{P_v} \left[ \log \frac{p_v(X)}{\sum_{i=1}^N q_i(X)} \right] \leq \log N + \mathbb{E}_{P_v} \left[ \log \frac{p_v(X)}{\max_i q_i(X)} \right]
\]

\[
\leq \log N + \min_i \mathbb{E}_{P_v} \left[ \log \frac{p_v(X)}{q_i(X)} \right] = \log N + \min_i D_{\text{KL}}(P_v \| Q_i).
\]

By our assumption that the \( Q_i \) form a cover, this gives the desired result, as \( \epsilon \geq 0 \) was arbitrary, as was our choice of the cover. \( \square \)

By a completely parallel proof, we also immediately obtain the following corollary.

Corollary 5.2. Assume that \( X_1, \ldots, X_n \) are drawn i.i.d. from \( P_v \) conditional on \( V = v \). Let \( N_{\text{KL}}(\epsilon, P) \) denote the KL-covering number of a collection \( \mathcal{P} \) containing the distributions (over a single observation) \( P_v \) for all \( v \in \mathcal{V} \). Then

\[
I(V; X_1, \ldots, X_n) \leq \inf_{\epsilon \geq 0} \left\{ n\epsilon^2 + \log N_{\text{KL}}(\epsilon, \mathcal{P}) \right\}.
\]
With Corollary 5.2 and Proposition 5.1 in place, we thus see that the global covering numbers in KL-divergence govern the behavior of information.

We remark in passing that the quantity (5.1.3), and its i.i.d. analogue in Corollary 5.2, is known as the index of resolvability, and it controls estimation rates and redundancy of coding schemes for unknown distributions in a variety of scenarios; see, for example, Barron [1] and Barron and Cover [2]. It is also similar to notions of complexity in Dudley’s entropy integral (cf. Dudley [3]) in empirical process theory, where the fluctuations of an empirical process are governed by a tradeoff between covering number and approximation of individual terms in the process.

5.2 Minimax bounds using global packings

There is now a four step process to proving minimax lower bounds using the global Fano method. Our starting point is to recall the Fano minimax lower bound in Proposition 2.12, which begins with the construction of a set of points \( \{ \theta(P_v) \}_{v \in V} \) that form a 2\( \delta \)-packing of a set \( \Theta \) in some \( \rho \)-semimetric. With this inequality in mind, we perform the following four steps:

(i) **Bound the packing entropy.** Give a lower bound on the packing number of the set \( \Theta \) with 2\( \delta \)-separation (call this lower bound \( M(\delta) \)).

(ii) **Bound the metric entropy.** Give an upper bound on the KL-metric entropy of the class \( \mathcal{P} \) of distributions containing all the distributions \( P_v \), that is, an upper bound on \( \log N_{\text{kl}}(\epsilon, \mathcal{P}) \).

(iii) **Find the critical radius.** Noting as in Corollary 5.2 that with \( n \) i.i.d. observations, we have

\[
I(V; X_1, \ldots, X_n) \leq \inf_{\epsilon \geq 0} \left\{ n\epsilon^2 + \log N_{\text{kl}}(\epsilon, \mathcal{P}) \right\},
\]

we now balance the information \( I(V; X^n) \) and the packing entropy \( \log M(\delta) \). To that end, we choose \( \epsilon_n \) and \( \delta > 0 \) at the critical radius, defined as follows: choose the any \( \epsilon_n \) such that

\[
n\epsilon_n^2 \geq \log N_{\text{kl}}(\epsilon_n, \mathcal{P}),
\]

and choose the largest \( \delta_n > 0 \) such that

\[
\log M(\delta_n) \geq 4n\epsilon_n^2 + 2 \log 2 \geq 2N_{\text{kl}}(\epsilon_n, \mathcal{P}) + 2n\epsilon_n^2 + 2 \log 2 \geq 2 \left( I(V; X^n) + \log 2 \right).
\]

(We could have chosen the \( \epsilon_n \) attaining the infimum in the mutual information, but this way we need only an upper bound on \( \log N_{\text{kl}}(\epsilon, \mathcal{P}) \).)

(iv) **Apply the Fano minimax bound.** Having chosen \( \delta_n \) and \( \epsilon_n \) as above, we immediately obtain that for the Markov chain \( V \to X^n \to \hat{V} \),

\[
\mathbb{P}(V \neq \hat{V}) \geq 1 - \frac{I(V; X_1, \ldots, X_n) + \log 2}{\log M(\delta_n)} \geq 1 - \frac{1}{2} = \frac{1}{2},
\]

and thus, applying the Fano minimax bound in Proposition 2.12 we obtain

\[
\mathcal{M}_n(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2} \Phi(\delta_n).
\]
5.3 Example: non-parametric regression

In this section, we flesh out the outline in the prequel to show how to obtain a minimax lower bound for a non-parametric regression problem directly with packing and metric entropies. In this example, we sketch the result, leaving explicit constant calculations to the dedicated reader. Nonetheless, we recover an analogue of Theorem 4.4 on minimax risks for estimation of 1-Lipschitz functions on [0,1].

We use the standard non-parametric regression setting, where our observations \(Y_i\) follow the independent noise model (4.1.1), that is, \(Y_i = f(X_i) + \varepsilon_i\). Letting \(\mathcal{F} := \{f : [0, 1] \to \mathbb{R}, \ f(0) = 0, \ f \text{ is Lipschitz}\}\) be the family of 1-Lipschitz functions with \(f(0) = 0\), we have

**Proposition 5.3.** There exists a universal constant \(c > 0\) such that

\[
M_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[ \|\hat{f}_n - f\|_\infty \right] \geq c \left( \frac{\sigma^2}{n} \right)^{1/3} ,
\]

where \(\hat{f}_n\) is constructed based on the \(n\) independent observations \(f(X_i) + \varepsilon_i\).

The rate in Proposition 5.3 is sharp to within factors logarithmic in \(n\); a more precise analysis of the upper and lower bounds on the minimax rate yields

\[
M_n(\mathcal{F}, \|\cdot\|_\infty) := \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[ \|\hat{f}_n - f\|_\infty \right] \asymp \left( \frac{\sigma^2 \log n}{n} \right)^{1/3} .
\]

See, for example, Tsybakov [4] for a proof of this fact.

**Proof** Our first step is to note that the covering and packing numbers of the set \(\mathcal{F}\) in the \(\ell_\infty\) metric satisfy

\[
\log N(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \log M(\delta, \mathcal{F}, \|\cdot\|_\infty) \asymp \frac{1}{\delta} . \quad (5.3.1)
\]

To see this, fix some \(\delta \in (0, 1)\) and assume for simplicity that \(1/\delta\) is an integer. Define the sets \(E_j = [\delta(j - 1), \delta j)\), and for each \(v \in \{-1, 1\}^{1/\delta}\) define \(h_v(x) = \sum_{j=1}^{1/\delta} v_j 1 \{x \in E_j\}\). Then define the function \(f_v(t) = \int_0^t h_v(t) dt\), which increases or decreases linearly on each interval of width \(\delta\) in [0,1]. Then these \(f_v\) form a \(2\delta\)-packing and a \(2\delta\)-cover of \(\mathcal{F}\), and there are \(2^{1/\delta}\) such \(f_v\). Thus the asymptotic approximation (5.3.1) holds. **TODO: Draw a picture**

Now, if for some fixed \(x \in [0, 1]\) and \(f, g \in \mathcal{F}\) we define \(P_f\) and \(P_g\) to be the distributions of the observations \(f(x) + \varepsilon\) or \(g(x) + \varepsilon\), we have that

\[
D_{\text{kl}}(P_f \| P_g) = \frac{1}{2\sigma^2} (f(X_i) - g(X_i))^2 \leq \frac{\|f - g\|^2_{\infty}}{2\sigma^2} ,
\]

and if \(P_f^n\) is the distribution of the \(n\) observations \(f(X_i) + \varepsilon_i, \ i = 1, \ldots, n\), we also have

\[
D_{\text{kl}}(P_f^n \| P_g^n) = \sum_{i=1}^n \frac{1}{2\sigma^2} (f(X_i) - g(X_i))^2 \leq \frac{n}{2\sigma^2} \frac{\|f - g\|^2_{\infty}}{2}. \quad 52
\]
In particular, this implies the upper bound
\[
\log N_{\text{kl}}(\epsilon, P) \lesssim \frac{1}{\sigma \epsilon}
\]
on the KL-metric entropy of the class \( P = \{ P_f : f \in \mathcal{F} \} \), as \( \log N(\delta, \mathcal{F}, \|\cdot\|_{\infty}) \propto \delta^{-1} \). Thus we have completed steps (i) and (ii) in our program above.

It remains to choose the critical radius in step (iii), but this is now relatively straightforward: by choosing \( \epsilon_n \asymp (1/\sigma n)^{1/3} \), and whence \( n \epsilon_n^2 \asymp (n/\sigma^2)^{1/3} \), we find that taking \( \delta \asymp (\sigma^2/n)^{1/3} \) is sufficient to ensure that \( \log N(\delta, \mathcal{F}, \|\cdot\|_{\infty}) \gtrsim \delta^{-1} \geq 4n\epsilon_n^2 + 2 \log 2 \). Thus we have
\[
\mathcal{M}_n(\mathcal{F}, \|\cdot\|_{\infty}) \gtrsim \delta_n \cdot \frac{1}{2} \gtrsim \left( \frac{\sigma^2}{n} \right)^{1/3}
\]
as desired.
Bibliography


