**EE378C Information-theoretic Lower Bounds in Data Science** Lecture 12 - 05/05/2021

# Lecture 12: Global Fano's method

*Lecturer: Yanjun Han* *Scribe: Wei-Ning Chen*

In this lecture, we will introduce global Fano's method, which allows us to apply Fano's method without manually constructing the hypotheses or upper bounding the mutual information. Today's plan includes

- covering and packing
- packing under target loss $\Longrightarrow$ separation condition
- covering under KL divergence $\Longrightarrow$ upper bound of mutual information
- examples including nonparametric density estimation, isotonic regression, convex regression, and sparse linear regression.

## 1 Covering and packing

We begin by introducing the notion of covering and packing. Let $(X, d)$ be a metric space and $A \subseteq X$ be a compact set. Then a covering and a packing of $A$ are defined as follows.

**Definition 1** (Covering)**.** *We say that $\{x_1, \cdots, x_n\} \subseteq X$ is an $\varepsilon$-covering of $A$ if $A \subseteq \bigcup_{i=1}^n B(x_i, \varepsilon)$ (i.e. for any $x \in A$, there exists $i \in [n]$ such that $d(x, x_i) \leq \varepsilon$). Moreover, we say*

$$N(\varepsilon) \triangleq N(A, d, \varepsilon) \triangleq \min\{n : \exists\, \varepsilon\text{-covering of } A \text{ with size } n\}$$

*is the covering number of $A$.*

**Definition 2** (Packing)**.** *We say that $\{x_1, \cdots, x_n\} \subseteq A$ is an $\varepsilon$-packing of $A$ if $\left\{B\left(x_i, \frac{\varepsilon}{2}\right) : i \in [n]\right\}$ are pairwise disjoint (i.e. $d(x_i, x_j) > \varepsilon$ for all distinct $i, j$). Moreover, we say*

$$M(\varepsilon) \triangleq M(A, d, \varepsilon) \triangleq \max\{n : \exists\, \varepsilon\text{-packing of } A \text{ with size } n\}$$

*is the packing number of $A$.*

The next lemma gives a basic property of covering and packing number.

**Lemma 3.** *For any compact set $A$ in a metric space $(X, d)$ and $\varepsilon > 0$, we have*

$$M(A, d, 2\varepsilon) \leq N(A, d, \varepsilon) \leq M(A, d, \varepsilon).$$

**Proof** For the first inequality, assume by contradiction $M(A, d, 2\varepsilon) \geq N(A, D, \varepsilon)$. Then by the pigeon-hole principle, at least two points in the $2\varepsilon$-packing belong to the same $\varepsilon$-cover in the covering, but this implies $d(x_i, x_j) \leq d(x_i, c) + d(x_j, c) \leq 2\varepsilon$. For the second inequality, we claim that any maximal $\varepsilon$-packing is an $\varepsilon$-covering. Otherwise, there exists some $x \in A$ such that $d(x, x_i) > \varepsilon$, so one can add $x$ into the the packing and the resulting larger set is still an $\varepsilon$-packing. □

Next, we bound the packing and covering number of $A$ by its volume.

**Lemma 4.** *Let $A \subseteq X = \mathbb{R}^d$ and $\|\cdot\|$ be any norm. Then*

$$\frac{1}{\varepsilon^d}\frac{\mathsf{vol}(A)}{\mathsf{vol}(B)} \leq N(A, \|\cdot\|, \varepsilon) \leq M(A, \|\cdot\|, \varepsilon) \leq \frac{\mathsf{vol}(A + \varepsilon B/2)}{\mathsf{vol}(\varepsilon B/2)},$$

*where $B$ is the unit ball under $\|\cdot\|$ and $A + B \triangleq \{a + b : a \in A,\, b \in B\}$.*

**Proof** The first inequality holds since $A \subseteq \bigcup_{i=1}^N B(x_i, \varepsilon)$, so

$$\mathsf{vol}(A) \le \mathsf{vol}\left(\bigcup_{i=1}^N B(x_i, \varepsilon)\right) \le \varepsilon^d N\left(A, \|\cdot\|, \varepsilon\right) \mathsf{vol}(B).$$

On the other hand, the last inequality holds since $\bigcup_{i=1}^M B(x_i, \varepsilon/2) \subseteq A + \frac{\varepsilon B}{2}$ and the balls are disjoint, so again we have

$$\mathsf{vol}\left(A + \frac{\varepsilon B}{2}\right) \ge \mathsf{vol}\left(\bigcup_{i=1}^M B(x_i, \varepsilon)\right) = M\left(A, \|\cdot\|, \varepsilon\right) \mathsf{vol}(\varepsilon B/2).$$

□

For the special case where $A = rB$, we have

$$\left(\frac{r}{\varepsilon}\right)^d \le N\left(rB, \|\cdot\|, \varepsilon\right) \le M\left(rB, \|\cdot\|, \varepsilon\right) \le \left(1 + \frac{2r}{\varepsilon}\right)^d,$$

so $\log M(\varepsilon) \asymp \log N(\varepsilon) \asymp d \log\left(1 + \frac{r}{\varepsilon}\right)$. Notice that this result only holds when $B$ is the unit ball under the same norm as the metric space. The next theorem characterizes the metric entropy of $\ell_p$ ball under $\|\cdot\|_q$ norm.

**Theorem 5** ([Schütt, 1984, Guedon and Litvak, 2000])**.** *For $0 < p < q \le \infty$ and dimension $d$,*

$$\log N\left(B_p, \|\cdot\|_q, \varepsilon\right) \asymp_{p,q} \begin{cases} \varepsilon^{-\frac{pq}{q-p}} \log\left(d \varepsilon^{\frac{pq}{q-p}}\right), & \text{if } \varepsilon \gtrsim d^{1/q-1/p} \\ d \log\left(1/\left(d \varepsilon^{\frac{pq}{q-p}}\right)\right), & \text{if } \varepsilon \lesssim d^{1/q-1/p}. \end{cases}$$

**Theorem 6** ([Artstein et al., 2004])**.** *Let $N(A, B)$ be the smallest number of translations of $B$ that cover $A$. For convex symmetric body $A$ and $B = \varepsilon B_2$, there exist $\alpha, \beta > 0$ such that*

$$\beta^{-1} \log N\left(B_2, \alpha^{-1} \varepsilon A^\circ\right) \le \log N(A, \varepsilon B_2) \le \beta \log N(B_2, \alpha \varepsilon A^\circ),$$

*where $A^\circ = \{y : \sup_{x \in A} \langle x, y \rangle \le 1\}$ is the polar body of $A$.*

## 1.1 Example: Gaussian mean estimation under $L_p$ loss

Let $X_1, ., ., ., .X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, I_d)$ with unknown $\theta \in \mathbb{R}^d$. The goal is to estimate $\theta$ under general $L_p$ loss $L(\theta, T) = \|T - \theta\|_p$, $p \in [1, \infty]$.

**Claim 7.** *Let $R^*_{n,d,p}$ be the minimax risk. Then*

$$R^*_{n,d,p} \asymp_p \begin{cases} \sqrt{\frac{d}{n}}, & \text{if } 1 \le p \le 2, \\ \frac{d^{1/p}}{\sqrt{n}}, & \text{if } 2 < p < \infty, \\ \sqrt{\frac{\log d}{n}}, & \text{if } p = \infty. \end{cases}$$

**Proof** For $p = 2$, the result follows from the standard two-point method (or LAM theorem). For $1 < p < 2$, the results also hold trivially since $\|\cdot\|_p \le \|\cdot\|_q$ for $1 \le q \le p$. Next, we apply packing to GLM for $p > 2$.

If we construct hypotheses in $\{\theta : \|\theta\|_2 \le r\}$ (i.e. $\mathsf{supp}(\pi_\theta) = \{\theta : \|\theta\|_2 \le r\}$), then the following mutual information upper bound holds:

$$I(\theta; X) = \inf_{P_{X^n}} \mathrm{E}_{\pi_\theta}\left[D_{\mathsf{KL}}\left(\mathcal{N}\left(\theta, I_d\right)^{\otimes n} \middle\| P_{X^n}\right)\right] \le \mathrm{E}_{\pi_\theta}\left[D_{\mathsf{KL}}\left(\mathcal{N}\left(\theta, I_d\right)^{\otimes n} \middle\| \mathcal{N}\left(0, I_d\right)^{\otimes n}\right)\right] \le \frac{nr^2}{2}.$$

We then choose the hypotheses to be uniform distributed on the maximized $\varepsilon$-packing in $\{\theta : \|\theta\|_2 \le r\}$, and thus by Fano's inequality,

$$\forall \delta > 0,\ R^*_{n,d,p} \gtrsim \delta \left(1 - \frac{nr^2/2 + \log 2}{\log M\left(rB_2, \|\cdot\|_p, \delta\right)}\right).$$

Finally, choosing $(\delta, r)$ to be $r \asymp \sqrt{d/n}$, $\delta \asymp d^{1/p}/\sqrt{n}$ for $2 < p < \infty$, and $r \asymp \delta \asymp \sqrt{(\log d)/n}$ for $p = \infty$ yields the desired results. □

## 2 Upper bounding $I(\theta; X)$ via covering

In the previous example, we upper bound the mutual information by restricting $\|\theta\|_2 < r$. The natural next question is: if $\theta \in \Theta$ almost surely, can we find an upper bound on $I(\theta, X)$ in a systematic way? It turns out that the mutual information can be upper bounded by KL covering number (which is the *channel capacity* in information theory).

**Theorem 8** (KL covering number)**.** *For $\varepsilon > 0$, let $N_{\mathsf{KL}}(\varepsilon) = N_{\mathsf{KL}}(\Theta, \varepsilon)$ be the smallest integer $n$ such that there exist distributions $Q_1, ..., Q_n$ on $\mathcal{X}$ such that $\sup_{\theta\in\Theta} \min_{i\in[n]} D_{\mathsf{KL}}(P_\theta \| Q_i) \le \varepsilon^2$. Then*

$$I(\theta; X^n) \le \inf_{\varepsilon > 0} \left(n\varepsilon^2 + \log N_{\mathsf{KL}}(\varepsilon)\right).$$

**Proof** According to the golden formula of mutual information,

$$I\left(\theta; X^n\right) \le \mathrm{E}_\theta \left[D_{\mathsf{KL}}\left(P_\theta^{\otimes n} \middle\| \frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}\right)\right],$$

where we choose $\{Q_i : i \in [n]\}$ to be a minimal KL covering. Then we have

$$D_{\mathsf{KL}}\left(P_\theta^{\otimes n} \middle\| \frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}\right) = \mathrm{E}_{P_\theta}\left[\log\left(\frac{P_\theta^{\otimes n}(x^n)}{N^{-1}Q_i^{\otimes n}(x^n)}\right)\right] \le \mathrm{E}_{P_\theta}\left[\min_{i\in[N]} \log\left(\frac{P_\theta^{\otimes n}(x^n)}{Q_i^{\otimes n}(x^n)}\right) + \log N\right],$$

which implies

$$I(\theta; X^n) \le \mathrm{E}_\theta\left[\log N + n \min_{i\in[N]} D_{\mathsf{KL}}(P_\theta \| Q_i)\right] \le \log N_{\mathsf{KL}}(\varepsilon) + n\varepsilon^2,$$

where the last inequality holds since $\{Q_i\}$ is a KL covering on $\Theta$. □

We summarize the general steps of applying global Fano's method.

1. Fix some $\theta > 0$ and $\Theta_0 \subseteq \Theta$. Find a $\delta$-packing of $\Theta_0$ under the following (pseudo-)metric:

$$d(\theta_1, \theta_2) \triangleq \min_{a\in\mathcal{A}} L(\theta_1, a) + L(\theta_2, a).$$

2. Fix some $\varepsilon > 0$, find the KL covering of $\Theta_0$.

3. Apply Fano's method yields

$$R_n^* \ge \frac{\delta}{2}\left(1 - \frac{\log N_{\mathsf{KL}}\left(\Theta_0, \varepsilon\right) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, d, \delta)}\right).$$

4. Finally choose the parameters appropriately to make the above bound as large as possible.

**Remark** In step 4, a rule of thumb is first picking $\varepsilon$ to make the term $\log N_{\mathsf{KL}}(\Theta_0, \varepsilon) + n\varepsilon^2$ as small as possible, and then picking $\delta$ to make the second term $\left(1 - \frac{\log N_{\mathsf{KL}} + n\varepsilon^2 + \log 2}{\log M}\right) > \frac{1}{2}$.

# 3 More examples

In this section, we provide several examples to demonstrate the power of global Fano's method. Although all the minimax rates are indeed tight, we will only prove the lower bounds part. We refer the readers to the references for the proof of the upper bounds.

## 3.1 Example I: nonparametric density estimation

Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} f$ with $f$ supported on $[0,1]^d$ and being Hölder smooth with smoothness parameter $s \in [0, \infty)$. The goal is to estimate the density $f$ under $L_p$ loss, with $p \in [1, \infty)$.

**Claim 9.** *The minimax risk $R^*_{n,s,d,p}$ has the following rate:*

$$R^*_{n,s,d,p} \asymp n^{-\frac{s}{2s+d}}.$$

**Proof** In order to lower bound the minimax rate, we apply global Fano's inequality. To begin with, we need a bound on metric entropy of the Hölder ball $\mathcal{H}^s_d$:

**Theorem 10** (Kolmogorov-Tikhomirov)**.**

$$\log N\left(\mathcal{H}^s_d, \|\cdot\|_p, \varepsilon\right) \asymp \varepsilon^{-d/s}.$$

Then we control the KL covering number as follows. Let $\mathcal{F}_L \triangleq \left\{f : f \geq \frac{1}{2} \text{ everywhere}\right\}$. Then

$$D_{\mathsf{KL}}(f \| f') \leq \chi^2(f, f') \leq 2 \|f - f'\|_2^2,$$

so $\log N_{\mathsf{KL}}(\mathcal{H}^s_d \cap \mathcal{F}_L, \varepsilon) \asymp \log N(\mathcal{H}^s_d \cap \mathcal{F}_L, \|\cdot\|_2, \varepsilon)$. We also compute the $L_p$ packing $\log M(\mathcal{H}^s_d \cap \mathcal{F}_L, \|\cdot\|_2, \delta)$ and apply Fano's inequality:

$$\begin{aligned} R^*_{n,s,d,p} &\gtrsim \delta\left(1 - \frac{\log N(\mathcal{H}^s_d \cap \mathcal{F}_L, \|\cdot\|_2, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\mathcal{H}^s_d \cap \mathcal{F}_L, \|\cdot\|_2, \delta)}\right) \\ &\geq \delta\left(1 - \frac{c'\varepsilon^{-d/s} + n\varepsilon^2 + \log 2}{c\delta^{-d/s}}\right), \end{aligned}$$

where the last inequality is due to Theorem 10 (we have used a slightly stronger result that the same covering/packing entropy result holds for $\mathcal{H}^s_d \cap \mathcal{F}_L$). Finally, by choosing $\varepsilon \asymp \delta \asymp n^{-\frac{s}{2s+d}}$, we arrive at the desired result. □

## 3.2 Example II: isotonic regression

Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} P_X$ with bounded density on $[0,1]$, and $Y_i \sim \mathcal{N}(f(X_i), 1)$. Moreover, assume $f : [0,1] \to [0,1]$ is non-decreasing. The goal is to estimate $f$ under $L_p$ loss, i.e. $L(f, T) = \|f - T\|_p$ with $p \in [1, \infty)$.

**Claim 11.** *The minimax risk $R^*_{n,p}$ has the following rate:*

$$R^*_{n,p} \asymp n^{-\frac{1}{3}}.$$

**Proof** We first upper bound the KL covering number. Observe that as long as $P_X$ upper bounded from above,

$$D_{\mathsf{KL}}\left(P_f \,\|P'_f\right) = \frac{1}{2}\left\|f - f'\right\|^2_{L_2(P_X)} \lesssim \left\|f - f'\right\|^2_2,$$

so $\log N_{\mathsf{KL}}\left(\mathcal{F}_M, \varepsilon\right) \lesssim \log N\left(\mathcal{F}_M, \|\cdot\|_2, \varepsilon\right)$, where $\mathcal{F}_M \triangleq \left\{f : [0,1]^d \to [0,1] : f \text{ is non-decreasing}\right\}$. Therefore applying global Fano's method yields

$$R^*_{n,p} \succeq \delta\left(1 - \frac{\log N\left(\mathcal{F}_M, \|\cdot\|_2, \varepsilon\right) + n\varepsilon^2 + \log 2}{\log M\left(\mathcal{F}_M, \|\cdot\|_p, \delta\right)}\right) \geq \delta\left(1 - \frac{c\varepsilon^{-1} + n\varepsilon^2 + \log 2}{c'\delta^{-1}}\right),$$

where the last inequality is due to the following bound on the covering number:

**Theorem 12.**

$$\log N\left(\mathcal{F}_M, \|\cdot\|_p, \varepsilon\right) \asymp \frac{1}{\varepsilon}.$$

Finally picking $\varepsilon \asymp \delta \asymp n^{-1/3}$ yields the desired result. □

### 3.3 Example III: convex regression

Let $X_1, ..., X_n \overset{\text{i.i.d.}}{\sim} P_X$ with bounded density on $[0,1]^d$ and $Y_i \sim \mathcal{N}(f(X_i), 1)$. Further assume that $f : [0,1]^d \to [0,1]$ is convex. The goal again is to estimate $f$ under the $L_p$ loss.

**Claim 13.** *The minimax risk $R^*_{n,d,p}$ has the following rate:*

$$R^*_{n,d,p} \asymp n^{-\frac{2}{4+d}}.$$

**Proof** The proof is the same as in Example 3.2, except that now we need a bound on the covering number of $\mathcal{F}_C \triangleq \left\{f : [0,1]^d \to [0,1] : f \text{ is convex}\right\}$ (instead of on $\mathcal{F}_M$).

**Theorem 14.** *Let $\mathcal{F}_C$ be defined as above. Then*

$$\log N\left(\mathcal{F}_C, \|\cdot\|_p, \varepsilon\right) \asymp \varepsilon^{-\frac{d}{2}}.$$

Applying the Theorem 14, together with Fano's inequality, we obtain

$$R^*_{n,p} \gtrsim \delta\left(1 - \frac{c\varepsilon^{-d/2} + n\varepsilon^2 + \log 2}{c'\delta^{-d/2}}\right).$$

Optimizing over $\varepsilon$ and $\delta$ gives the desired result. □

**Remark** In fact, $N\left(\mathcal{F}_C, \|\cdot\|_p, \varepsilon\right)$ highly depends on the domain of $f$. If the domain of $f$ is not a polytope, than the metric entropy may be much higher. For instance, let $\tilde{\mathcal{F}}_C \triangleq \{f : \text{unit ball} \to [0,1] : f \text{ is convex}\}$, then

$$\log N\left(\tilde{\mathcal{F}}_C, \|\cdot\|_p, \varepsilon\right) \asymp \max\{\varepsilon^{-1/2}, \varepsilon^{-(d-1)}\}.$$

### 3.4 Example IV: sparse linear regression/prediction

Consider the sparse regression model where the design matrix $X \in \mathbb{R}^{n\times d}$, and the response $Y \sim \mathcal{X}(X\theta, I_n)$ with unknown $\theta$. Further more, assume $\|\theta\|_q \leq R$ for some $q \in (0,1)$. The goal is to minize the estimation error $L_{\mathsf{est}}(\theta,T) = \|T-\theta\|_p$ with $p \in [1,\infty]$ or the prediction error $L_{\mathsf{pre}}(\theta,T) = \|X(T-\theta)\|_2/\sqrt{n}$.

**Claim 15.** *Under appropriate conditions,*

$$R^*_{n,d,p,q,R}(\mathsf{estimation}) \asymp R^{q/p}\left(\frac{\log d}{n}\right)^{\frac{p-q}{2p}}, \text{ and } R^*_{n,d,q,R}(\mathsf{prediction}) \asymp R^{q/2}\left(\frac{\log d}{n}\right)^{\frac{2-q}{4}}.$$

**Proof** Although the second result seems to be a special case of the first result by setting $p = 2$, these results impose very different assumptions on the design matrix $X$.

**Estimation error: mild assumption on** $X$ First, applying the $L_p$ packing bound Theorem 5 on $B_q(R)$ yields

$$\log M\left(B_q(R), \|\cdot\|_p, \delta\right) \asymp \left(\frac{R}{\delta}\right)^{\frac{pq}{p-q}} \log d, \text{ if } \delta \gg Rd^{1/p-1/q}.$$

Next, we bound the KL covering number. Observe that $D_{\mathsf{KL}}(P_\theta \| P_{\theta'}) = \|X(\theta-\theta')\|_2^2/2$, and from approximation theory, we have

$$\log N(X \cdot B_q(R), \|\cdot\|_2, \varepsilon) \lesssim \log N\left(B_q(R), \|\cdot\|_2, \varepsilon/\sqrt{n}\right),$$

provided that $\|\|X\|\|_{1\to 2} = \max_{j\in[d]} \|X_j\|_2 = O(\sqrt{n})$. Therefore applying global Fano's method, we obtain

$$R^*_{n,d,p,q,R} \gtrsim \delta\left(1 - \frac{\log N(B_q(R), \|\cdot\|_2, \varepsilon/\sqrt{n}) + \varepsilon^2 + \log 2}{\log M\left(B_q(R), \|\cdot\|_p, \delta\right)}\right).$$

Finally, applying Theorem 5 and optimizing over $\delta$ and $\varepsilon$ yield the first result.

**Prediction error: strong assumption on** $X$ To lower bound the prediction error, we use the same KL covering bound. On the other hand, the packing becomes $\log M\left(X \cdot B_q(R), \|\cdot\|_p\, \delta\right)$. However, since we need a lower bound on the packing entropy, we need a stronger assumption on $X$:

$$\forall\, \theta, \theta' \in B_q(R),\ \|X(\theta-\theta')\|_2 \geq \kappa \cdot \sqrt{n}\,\|\theta-\theta'\|_2 - (\text{lower order terms})^1.$$

With this additional assumption, we have

$$\log M(X \cdot B_q(R), \|\cdot\|_p, \delta) \gtrsim \log M(B_q(R), \|\cdot\|_2, \delta/\sqrt{n}).$$

Similarly, applying Fano's inequality and Theorem 5 and optimizing over $\varepsilon$ and $\delta$ give us the second result. □

# References

[Artstein et al., 2004] Artstein, S., Milman, V., and Szarek, S. J. (2004). Duality of metric entropy. *Annals of mathematics*, pages 1313–1328.

[Guedon and Litvak, 2000] Guedon, O. and Litvak, A. (2000). Euclidean projections of a p-convex body. In *Geometric aspects of functional analysis*, pages 95–108. Springer.

[Schütt, 1984] Schütt, C. (1984). Entropy numbers of diagonal operators between symmetric banach spaces. *Journal of approximation theory*, 40(2):121–128.

[1] This is also known as a restricted eigenvalue (RE) condition.